Chemometrics: Theory and Application

Bruce R. Kowalski

ACS Symposium Series 52

Chemometrics: Theory and Application

Bruce R. Kowalski, EDITOR
University of Washington

A symposium sponsored by the
Division of Computers in Chemistry
at the 172nd Meeting of the
American Chemical Society,
San Francisco, Calif.,
Sept. 2, 1976.

ACS SYMPOSIUM SERIES **52**

AMERICAN CHEMICAL SOCIETY
WASHINGTON, D. C. 1977

Library of Congress CIP Data

Chemometrics.

(ACS symposium series; 52 ISSN 0097-6156)

Includes bibliographical references and index.
1. Chemistry—Data processing—Congresses.
I. Kowalski, Bruce R., 1942– II. American Chemical Society. Division of Computers in Chemistry. III. Series: American Chemical Society. ACS symposium series; 52.

QD39.3.E46C48 542'.8 77-9088
ISBN 0-8412-0379-2 ACSMC8 52 1–288

PRINTED IN THE UNITED STATES OF AMERICA

ACS Symposium Series

FOREWORD

The ACS Symposium Series was founded in 1974 to provide a medium for publishing symposia quickly in book form. The format of the Series parallels that of the continuing Advances in Chemistry Series except that in order to save time the papers are not typeset but are reproduced as they are submitted by the authors in camera-ready form. As a further means of saving time, the papers are not edited or reviewed except by the symposium chairman, who becomes editor of the book. Papers published in the ACS Symposium Series are original contributions not published elsewhere in whole or major part and include reports of research as well as reviews since symposia may embrace both types of presentation.

PREFACE

During the mid-1800s a number of events led scientists to seek a relationship among the chemical elements that were known at the time. About five years before Mendeleev's publication of what has been called the first periodic table, John Newlands published his periodic table which was not accepted by the scientific community, and was not to be recognized as an achievement until the Royal Society belatedly awarded him the Davy Medal five years after it had similarly honored Mendeleev. The remarkable discovery of Newlands was the repetitive pattern of properties of the elements when they were arranged according to Cannizzaro's new atomic weights. Thus Newlands, studying a collection of objects (elements) via the properties of each object, applied unsupervised pattern recognition to a problem of multivariate analysis long before computers and pattern recognition were invented. His approach to the discovery of the periodicity of the elements would make him one of the early chemometricians, if not the first.

Modern chemistry, as a physical science, studies chemical systems by obtaining information through the use of a variety of measurement systems. On the whole, the measurement systems available to chemists are quite sophisticated and generate data that are accurate and precise. Psychology, as a social science, studies human systems, also by making measurements. However, the data generated by psychology's measurement techniques are comparatively imprecise and inaccurate, and sometimes even nonmetric in nature. As a result of this problem experimental psychologists are eager to discover new mathematical and statistical methods to extract useful information from their observations. The area of psychology concerned with the design of experiments and the interpretation of observations is called psychometrics, and the journal *Psychometrika* has been published since 1936.

Psychology is not the only science formally searching for better methods of information extraction. Biometrics and econometrics are formal areas of study in biology and economics. In June 1974 the Chemometrics Society was founded in Seattle, Washington during an informal gathering of chemists. In a published letter to prospective chemometricians (*Journal of Chemical Information and Computer Sciences* (1975) **15**, 201), chemometrics is defined as the development and application of mathematical and statistical methods to extract useful chemical information from chemical measurements. Modern chemistry has ventured outside the controlled environment of the laboratory to tackle difficult problems with chemical measurements. This, combined with the proliferation of computers in chemical laboratories, has prompted a demand for new and improved methods to design and control experiments and to analyze the wealth of data that can be generated.

"Chemometrics: Theory and Application" represents a sampling of the work of chemometricians and does not constitute an all-inclusive review of that field. With one major exception and some minor ones this volume represents the content of a symposium under the same title presented by the Division of Computers in Chemistry at the 172nd National Meeting of the American Chemical Society, August 29 to September 3, 1976 in San Francisco. The major exception is the inclusion of a contributon by Sjöström and Wold. S. Wold was an invited speaker who unfortunately was unable to attend the meeting. The minor exceptions amount to differences in scope and emphasis between the papers delivered at the San Francisco meeting and those found in this book.

As data become easier and less expensive to acquire, there is little doubt that the chemist will be forced to rely more heavily on the computer and new mathematical and statistical analysis methods. Likewise, as instruments become more complex with multiple outputs as well as multiple inputs, the computer will assume a greater role in instrument control as well as such tasks as fault detection and even fault correction. The works of the authors found in this book clearly demonstrate that chemists are indeed interested and active in the search for better measurement system control and optimization and measurement analysis methods.

University of Washington
Seattle, Washington
March 7, 1977

BRUCE R. KOWALSKI

CONTENTS

1

Advances in the Application of Optimization Methodology in Chemistry

STANLEY N. DEMING
Department of Chemistry, University of Houston, Houston, TX 77004

STEPHEN L. MORGAN
Department of Chemistry, University of South Carolina, Columbia, SC 29208

Many chemical measurement processes can be viewed as *systems* (*1*) consisting of *inputs*, *transforms*, and *outputs* (see Figure 1). The *primary input to a chemical measurement process* is a *sample*, some *property* of which is to be assigned a numerical value (*2*). Examples of specific properties that might be measured are the percentage of iron in an ore, the concentration of calcium in a patient's blood serum, and the parts per million of hydrocarbons in urban air.

In addition to the primary input, many secondary inputs (or *factors*) can have an effect upon the numerical value that is eventually assigned to the property of interest. These additional factors include temperature, reagent amount, wavelength, time, homogeneity, and the presence of interfering substances. If the numerical result of the measurement process is to be a precise representation of the property of interest, it is clearly important that the more significant of these factors must be controlled. As Mandel has stated, "The development of a method of measurement is to a large extent the discovery of the most important environmental factors and the setting of tolerances for the variation of each one of them" (*3*,*4*). Ideally, the method should be "rugged" against uncontrolled changes in the environmental factors so that the tolerances can be broad.

It is often convenient to classify factors as *known* or *unknown*, and *controlled* or *uncontrolled*. A further classification results if it is noted at what point a factor enters the measurement scheme (see Figure 2); specifically, is the factor associated with the measurement process itself (*e.g.*, temperature, reagent amount) or is it instead associated with the sample (*e.g.*, homogeneity, presence of interfering substances)?

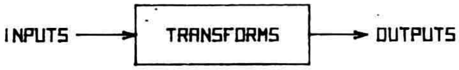

Figure 1. Systems view of the measurement process

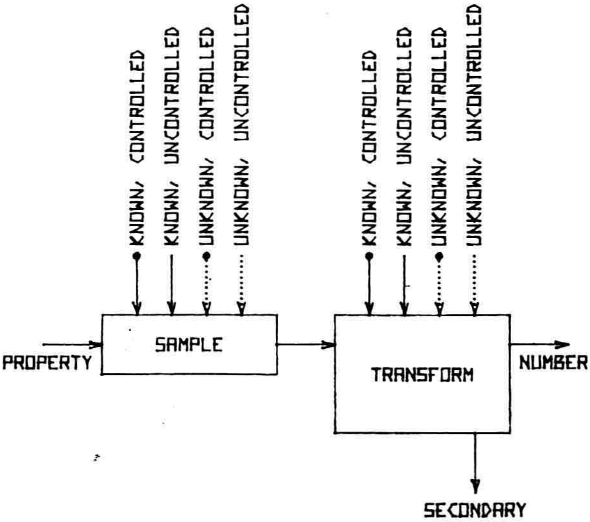

Figure 2. Expanded view of the measurement process

This latter distinction is often important in the development of analytical methods. If a known environmental factor is associated with the measurement process itself, then it is usually possible to control that factor during both the development and implementation of the method; thus, by sufficiently close control, the factor's influence on the numerical result can be minimized. If a known environmental factor is associated with the sample, it might not be possible to control that factor when the method is actually implemented. It is, however, usually possible to control the factor during the development of the method in such a way that the range of values normally encountered for that factor can be simulated. By this mechanism, the effect of a factor associated with the sample can be assessed, and the method can be developed so as to minimize the effect of this normally uncontrolled factor.

The primary output from a chemical measurement process is the numerical value of the property of interest in the sample. But many other, secondary outputs (or responses) might also be important: examples include cost per measurement, sensitivity to interfering substances, and linearity of the assigned numerical value vs. the property being measured. Thus, the development of a method of measurement can be more than the discovery of the most important environmental factors and the setting of tolerances for the variation of each one of them; it can also be the adjustment or optimization of the most important controllable environmental factors so as to achieve the best possible compromise among the many different responses (5).

The "advances" reported here illustrate the use of classical experimental designs in conjunction with optimization techniques to automatically produce a chemical measurement process possessing desirable performance characteristics (6).

Automated Continuous Flow System

Automated continuous flow methods of chemical analysis (7) have become widely accepted as reliable means of carrying out a large number of determinations in a short period of time with minimal analyst interaction. In the future, many existing continuous flow methods will need to be improved and many new continuous flow methods will need to be developed both to meet the more exacting requirements of established disciplines, as well as to fulfill the growing demands of relatively new areas such as environmental and clinical chemistries (8).

The instrument used in this work is built around standard Technicon AutoAnalyzer-II continuous flow components and a Hewlett-Packard 9830A computer. Many of the operations normally carried out by the analyst are under direct computer control. These operations include starting and stopping a tray of samples, acquiring digitized absorbance values from the colorimeter, and controlling the flow rate of individual reagents. This latter operation is accomplished by using individual peristaltic pumps for each reagent line; each peristaltic pump is driven by a stepping motor which can be made to turn at a rate that will deliver the desired flow. Computer options include: 16K bytes of read/write memory; thermal page printer; plotter; dual-platter disc; and read-only-memories for input/output, matrix, and string operations, and for advanced programming capability. A 32-bit serial, bidirectional, time multiplexed interface is used to communicate information between the instrument and computer.

Chemical System

The concentration of calcium in blood serum can be determined by dialysis of calcium ion into a recipient stream followed by reaction with the complexing agent cresolphthalein complexone in basic solution (9). Figure 3 is a diagram of the flow scheme used in this work.

Before dialysis, the serum sample is mixed with a solution containing hydrochloric acid (HCL-B), 8-hydroxyquinoline (8HQ-B), and water (used as a diluent to make up a fixed total flow). During dialysis, the calcium is transferred to a recipient stream containing hydrochloric acid (HCL-A), 8-hydroxyquinoline (8HQ-A), cresolphthalein complexone (CPC), and water. Diethylamine (DEA) is added to make the solution basic and the absorbance of the colored product is measured at 570 nm.

Figure 4 is a systems view of the continuous flow method for calcium. Six controllable factors associated with the measurement process have an influence either upon the number that is assigned to the calcium concentration, or upon some of the secondary outputs, or both. These factors are HCL-B, 8HQ-B, HCL-A, 8HQ-A, CPC, and DEA. Two uncontrollable factors that are associated with the sample are the concentrations of magnesium and protein in the serum. Magnesium is

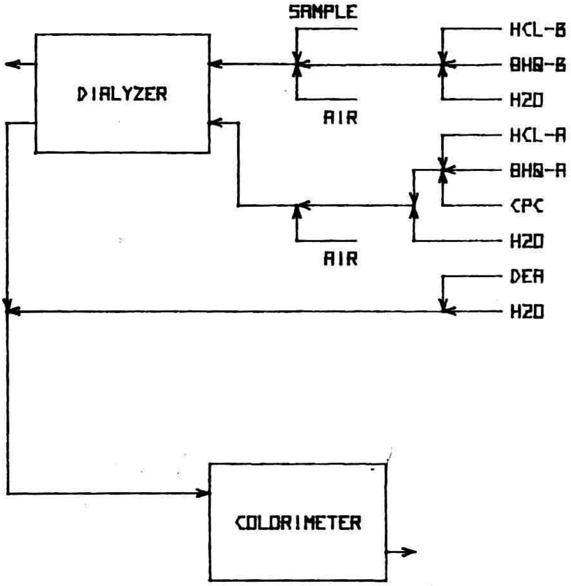

Figure 3. Flow scheme for continuous flow determination of calcium in blood serum

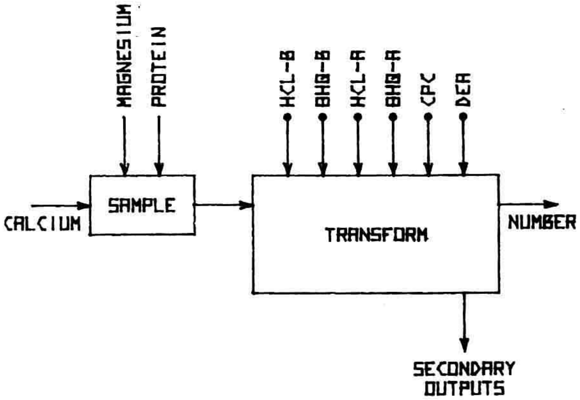

Figure 4. Systems view of continuous flow determination of calcium in blood serum

an interfering factor because it can codialyze with calcium and contribute to the measured absorbance by forming a colored Mg-CPC complex (10). Protein is an interferent because it is thought to contribute to a Donnan-type equilibrium in the dialysis process (11) and because it chemically binds calcium.

Objectives

The objectives of the work presented here were to

- increase the sensitivity of absorbance response with respect to serum calcium concentration (that is, maximize the slope $\delta A/\delta[CA]$);
- decrease the sensitivity of absorbance response with respect to serum magnesium concentration (*i.e.* minimize $|\delta A/\delta[Mg]|$);
- decrease the sensitivity of absorbance response with respect to serum protein concentration (minimize $|\delta A/\delta[\text{protein}]|$);
- maintain a relatively low blank absorbance (in this work, the blank absorbance was considered acceptably low if it was less than the absorbance above the reagent blank produced by a 15 mg dl^{-1} Ca standard); and
- maintain good linearity of absorbance with respect to serum calcium concentration (in this work, the standard deviation of residuals from a model first order in calcium was used to assess linearity).

Samples

A set of 20 serum samples were prepared by diluting one part reference serum with one part saline solution containing a specified amount of calcium ion, magnesium ion, and bovine serum albumin (protein). Each serum sample thus prepared corresponded to a treatment combination in the non-central composite experimental design (12) given in Table I.

Objective Function

Calcium, magnesium, and protein effects were assessed by fitting a full second-order polynomial model to data obtained from the set of 20 samples:

$$A = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \sum_{j=1}^{3} \beta_{ij} x_i x_j \qquad (1)$$

where A is the absorbance above the reagent blank, and x_1, x_2, and x_3 correspond to calcium, magnesium, and

Table I

Experimental Design for Samples

Sample	Calcium[a]	Magnesium[a]	Protein[a]
7	-1	-1	-1
3	-1	-1	+1
14	-1	+1	-1
6	-1	+1	+1
4	+1	-1	-1
8	+1	-1	+1
11	+1	+1	-1
19	+1	+1	+1
13	0	0	-2
5	0	0	+2
2	0	+2	0
16	0	+3	0
17	-3	0	0
18	-3	0	0
10	-2	0	0
15	-2	0	0
1	+2	0	0
12	+2	0	0
9	0	0	0
20	0	0	0

[a] coded levels are:

	-3	-2	-1	0	+1	+2	+3
calcium, mg dl^{-1}:	5.8	7.8	9.8	11.8	13.8	15.8	-
magnesium, mg dl^{-1}:	-	-	1.0	3.0	5.0	7.0	9.0
protein, g dl^{-1}:	-	3.8	4.8	5.8	6.8	7.8	-

protein levels, respectively. The parameter estimates $\underline{b}_1$, $\underline{b}_2$, and $\underline{b}_3$ are thus measures of the sensitivity with respect to calcium, magnesium, and protein levels, respectively.

During the optimization stage, the following objective function was maximized:

$$F = \underline{b}_1 - |\underline{b}_2| - |\underline{b}_3| \tag{2}$$

Boundary conditions were specified for $|\underline{b}_2|$ and $|\underline{b}_3|$ to keep them less than 10% of the value of $\underline{b}_1$. A boundary condition was also specified for the absorbance of the reagent blank (see Objectives). No boundary was placed on the linearity.

Optimization

The six controllable factors associated with the measurement process itself (HCL-B, 8HQ-B, HCL-A, 8HQ-A, CPC, and DEA) were varied according to the rules of a variable size sequential simplex algorithm (13, 14). The simplex technique has been used previously for laboratory optimizations and is described in the literature (13-20).

After evaluating a predetermined number of vertexes (*i.e.*, after carrying out 25 different treatment combinations of factors associated with the measurement process itself, each time running the tray of 20 samples, fitting Equation 1, and evaluating Equation 2), the objective function had increased by approximately 67% over the worst vertex in the initial simplex. Most of this increase came about because the calcium sensitivity was increased by approximately the same percentage. Magnesium sensitivity increased slightly and protein sensitivity decreased slightly; both were kept well within the established boundary conditions. The baseline response was kept low. Linearity suffered somewhat.

Mapping

After presumably optimum conditions were established by the simplex technique, a Box-Behnken design (21) was evaluated in the region of the suspected optimum to establish tolerances for the six controllable factors and their interactions. The design is a highly fractional, three-level factorial design and is described in the literature (21); it contained 54 treatment combinations, each of which required measuring the 20 serum samples and fitting Equation 1.

The results of the mapping study were used to fit six equations of the form

$$R = \alpha_0 + \sum_{i=1}^{6} \alpha_i y_i + \sum_{i=1}^{6} \sum_{j=1}^{6} \alpha_{ij} y_i y_j \qquad (3)$$

where R is one of the six responses considered (objective function, calcium sensitivity, magnesium sensitivity, protein sensitivity, absorbance of reagent blank, or linearity) and the y_1's correspond to the six controllable factors HCL-B, 8HQ-B, HCL-A, 8HQ-A, CPC, and DEA.

Discussion

The data acquired during the mapping study contains information relating six responses to six factors. The results thus contain information about 36 single-factor effects and 90 two-factor interactions. Only a small subset of this information will be discussed in this paper, the apparent effects of the two factors HCL-B and 8HQ-B on the four responses F, b_1, b_2, and b_3.

Figure 5 is a contour plot of the objective function calculated using Equation 3 *vs.* HCL-B and 8HQ-B,

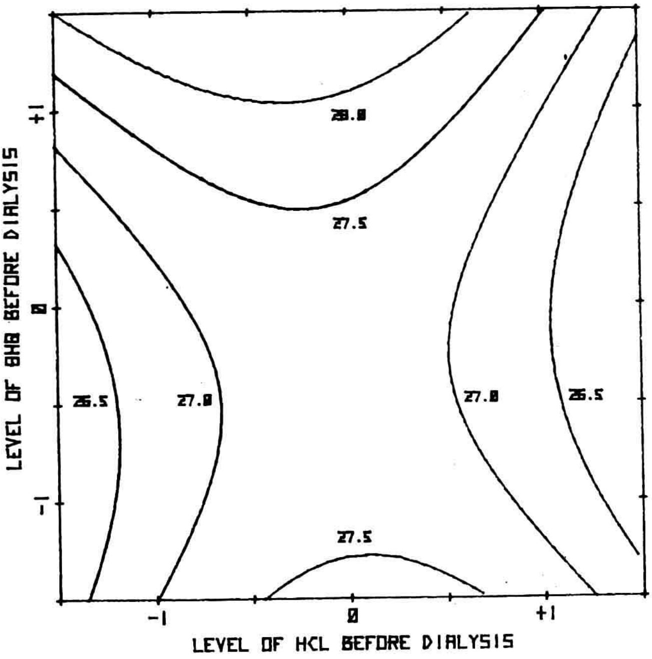

Figure 5. Objective function contours vs. HCL-B and 8HQ-B. Numbers are F × 1000 (see Equation 2).