Information Theory as Applied to Chemical Analysis

KAREL ECKSCHLAGER

Institute of Inorganic Chemistry
of Czechoslovak
Academy of Sciences, Prague

VLADIMÍR ŠTĚPÁNEK

Environmental Research Centre,
Prague

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS
New York / Chichester / Brisbane / Toronto

Library of Congress Cataloging in Publication Data

Eckschlager, Karel.
Information theory as applied to chemical analysis.

(Chemical analysis; v. 53)
"A Wiley-Interscience publication."
Includes index.
1. Chemistry, Analytic—Mathematics. 2. Information theory. I. Štěpánek, Vladimír, joint author. II. Title. III. Series.
QD75.3.E36 543'.001'5192 79-1405
ISBN 0-471-04945-X

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

In teaching analytical chemistry, in researching new methodologies of analytical chemistry, and in doing practical analytical work, there is need for individual analytical methods that will rapidly and economically provide necessary information about the composition of an analyzed sample. The concept of "information" allows us to derive quantities that facilitate evaluation (and classification) of analytical methods, make them more objective, and serve as objective functions in optimizing analytical procedures.

This book discusses the fundamentals of the current understanding of information concepts in analytical chemistry and aims at presenting most of their applications described so far. This objective appears in the arrangement of the text. Our intention has been to make the monograph self-contained as far as the employed theory is concerned. The mathematical chapters require only knowledge of calculus and of combinatorial counting techniques. Instead of a formal presentation of the concepts we have tried to explain them without including most of the proofs. We have emphasized notation because much confusion has arisen from imprecision in its use.

The introductory chapter is followed by Chapter 2, in which the basic ideas of understanding analytical chemistry as a process of obtaining information are explained. Since messages bearing information are random, randomness must be the basis of every realistic analysis. In this light probability theory is the only tool that can be used to pursue it. Chapter 3, therefore, is devoted to the fundamental concepts of this discipline. Information theory, which has developed from probability theory, is the theoretical study of information measures and information transmissions. It is discussed in Chapter 4. The first goal of this theory is to propose measures of information content and information gain. Since analytical methods are verified by statistical means using experimental results and various statistical techniques are adopted in the text, we have included a chapter (Chapter 5) dealing with some methods of statistical inference. The information access to the evaluation of various procedures in analytical chemistry is examined in the last chapter, where a survey of recent results appears. The results are discussed and possibilities of practical use are shown. Some information measures, introduced in this chapter and illustrated on examples, are specific for the evaluation in chemical analysis.

Each chapter is divided into sections and sometimes subsections. Examples are intermingled with the text. References are given at the end of each chapter in the sequence in which they appear in the text. Some of the references are not cited in the text.

We wish to thank those who stimulated the writing of this book. We feel indebted to Professor J. D. Winefordner of the University of Florida, Gainesville, for his interest in this monograph for the Chemical Analysis Series. We thank him and Professor J. P. Elving, the other editor of the series, for their help and encouragement during the preparation of the book. We appreciate suggestive discussions of some theoretical problems with Dr. I. Vajda. We also thank Mrs. J. Vyvialová for the remarkable speed and accuracy with which she typed the manuscript.

KAREL ECKSCHLAGER
VLADIMÍR ŠTĚPÁNEK

Prague
June 1979

GLOSSARY

CONVENTIONS

We have used Greek letters (except in a few cases for historical reasons) to represent random variables and the corresponding Roman letters to represent their values. Thus, x are values of a random variable ξ. Similarly, a clear distinction is drawn between parameters and estimates of these parameters calculated from observations. Thus, θ denotes a general parameter of a probability distribution and $T = T(x_1, x_2, \ldots, x_n)$ is its estimate calculated from a sample of size n.

Observations of one random variable are indexed by one subscript. If we deal with two or more random variables, observations are indexed by two subscripts; that is, y_{ij} is the jth observation of the random variable η_i.

Several special conventions are used, primarily in Chapters 3 and 4.

Symbols

$A = \dfrac{\sigma}{\sigma_0}$	Coefficient of precision making
a	Estimate of the regression coefficient α
a_0, a_k	Coefficients of the Fourier series
α	Probability of a type I error (the significance level); regression coefficient
A_P	Plane-discrimination capability
A_V	Space-discrimination capability
$B = \dfrac{\mu - \mu_0}{\sigma_0}$	Coefficient of modification of the results
b	Half-width of a peak; estimate of the regression coefficient β
b_c	Slope of a calibration curve

b_0, b_k	Coefficients of the Fourier series
β	Probability of a type II error; a regression coefficient
$C(p,p_0)$	Specific information price
c_i	Concentration (content) of the ith component
d_i	Estimate of the mean error of the determination of the ith component
δ	Mean error
$\delta^{(0)}$	Mean error related to blank correction
$E[\ ]$	Expectation of the quantity in brackets
e_i, ε_i	Effectivity coefficients
η_i	Signal intensity for the ith component
F	Random variable
F_α	100α-percentage value of the F-distribution
$f_i(X_i)$	Function of the true value of the ith component (known or found by calibration)
$g_i(y_i)$	Function of the signal intensity of the ith component (inverse of the function f_i)
$H(x_1,x_2,\ldots,x_n)$	Joint distribution function of n random variables
H, H_0	Statistical hypotheses
i	Subscript i, index of the component or of a random variable
$I(p,p_0)$, $I(q\|\|p)$	Information content; the gain of information
$J(p,p_0)$	Information flow
j	Subscript j, index of a parallel determination
K	Tolerance coefficient
k	Number of (determined) components
$L(p,p_0)$	Information performance
ln, log	Logarithms
λ	Parameter of the Poisson distribution
M	Total number of components in an analyzed sample
m	Number of components detected by a given qualitative test; weighed amount of a sample; number of signals

μ_i	Expected value of the random variable ξ_i, $\mu_i = E[\xi_i]$
N	Number of elements of a finite set
N_1	Number of possible discriminations of the substance
N_2	Number of discriminated concentrations
$n(A)$	Number of sample points of the event A
n_p	Number of parallel determinations
n_s	Number of determinations from which an estimate of the standard deviation is calculated
$\binom{n}{k}$	Number of combinations of n elements taken k at a time
ν	Number of degrees of freedom
$\Omega_{i-1,i}$	Overlap between the $(i-1)$th and the ith peaks
$P(A)$	Probability of the event A
$P(A\|B)$	Conditional probability of event A given event B
p_i	Probability of a discrete random variable, $p_i = P\{\xi = x_i\}$
$p_0(x)$	Probability density of an a priori distribution
$p(x)$	Probability density of an a posteriori distribution
q_i	Probability of a discrete random variable, $q_i = P\{\eta = y_i\}$
q_0	Ratio of the peak symmetry
q	Ratio of the peak shape (for symmetrical peaks only)
$\mathcal{R}_1$	Set of real numbers
ρ	Information redundance
S	Set of all possible outcomes of a random experiment
S_i	Sensitivity of determination of the ith component
s	Estimate of the standard deviation
s_y	Estimate of the standard deviation of the signal intensity

σ	Standard deviation of a random variable
σ^2	Variance of a random variable
σ_i^2	Variance of the proper determination of the *i*th component
σ_0^2	Variance of the blank correction; variance of an a priori probability distribution
T	Period in a Fourier series
t	Student's random variable
$t_\alpha(n)$	100α-percentage value of the Student distribution with n degrees of freedom
t_ν	Percentage value of the Student distribution for $\alpha = 0.038794$ and $\nu = n_s - 1$ degrees of freedom
t_A	Time of the duration of an analysis
τ_A	Cost of an analysis
θ	General parameter of a probability distribution
$V[\xi]$	Variance of the random variable ξ
X_i	True content of the *i*th component in an analyzed sample
$x_{i,j}$	Value of the result of analysis of the *i*th component and of the *j*th determination
$x_i^{(c)}$	Amount of the component to be determined obtained from the calibration curve
$\bar{x}_i$	Mean of parallel determinations of the *i*th component
x_U	Upper limit of a tolerance interval
x_L	Lower limit of a tolerance interval
ξ_i	Results of analysis of the *i*th component
y_i	Value of the signal intensity for the *i*th component
$\bar{y}_i$	Mean signal intensity of the *i*th component.
$y_{\min}$	Least intensity of the signal distinguishable from zero
$y_{\max}$	Maximum intensity of the signal
z_i	Signal position for the *i*th component
z_α	100α-percentage value of the standardized normal distribution

$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$	
$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n$	
$\in$	Element of
$\cap$	Intersection of sets
$\cup$	Union of sets

CONTENTS

CHAPTER

1

INTRODUCTION

Analytical chemistry, the scientific discipline representing theoretical grounds for chemical and physicochemical analyses of the composition of matter, has in the last few decades been characterized by increasing interest in problems of more general validity. In the first period of development of this scientific discipline, "analytical chemistry" usually was understood to be a collection of procedures to be carried out when performing an analysis. In 1977, W. Fresenius presented this definition of analytical chemistry (paraphrased from [1]): "Analytical chemistry is the science of acquiring information on material systems and interpreting it with regard to its exploitation, employing the methods of natural science." In this definition obtaining information is said to be the objective of chemical and instrumental analyses. We can then define the analysis itself as a process of obtaining information about the chemical composition of matter. From it also follows the importance of the use of concepts and methods of information theory for analytical chemistry.

The development of the concept of analytical chemistry from the original idea of a collection of working recipes to the present "science of acquiring information" was of course quite slow.

An important development was H. Kaiser's use of the probabilistic point of view in 1936 in connection with the poorly reproducible emission spectrometry results available at that time [2]. The basis of the probabilistic point of view resides in the idea that the result of the determination of the unknown content of a component is a random variable which has a probability distribution. This idea enabled evaluation of analytical results and methods by the use of statistical methods. This became quite common in analytical practice during the 10 to 15 years following the first papers by Kaiser [2]. Statistical evaluation of analytical results and methods has spread primarily with the development of instrumental methods, especially trace analysis. There have also appeared probability definitions of purely analytical concepts: for example, Kaiser has presented a statistical definition of the detection and determination limits and J. D. Winefordner and later L. A. Currie have extended it. In the literature of trace analysis, concepts from communication theory are commonly used (e.g., signal, signal-to-noise ratio, detection of signals, etc.). In connection with introducing automation and machine processing of analytical data, an "automation in analysis" work team was established in Lindau at the beginning of the 1970s, and its members have contributed many new approaches and given a number of useful definitions. This group, in their wide and generally directed activities, have also been concerned with questions of the use of information theory in analytical chemistry. Here again, the probabilistic approach is evident, since the concept of information is narrowly linked with the probability distributions of appropriate random variables. Although Kaiser was not the first to be concerned with the possibilities of using information theory in analytical chemistry, his lecture delivered at the University of Georgia in 1969 and published one year later [3] had fundamental importance for the extension of the theoretical point of view of information theory in the literature of analytical chemistry. A certain gain in the understanding and practical use of the information content of analytical results has come with the introduction of the divergence measure. This has substantially more general validity than have measures formerly transferred from communication theory [4].

In the last one or two decades new theoretical and often mathematically expressed approaches to analytical problems have appeared in the literature of analytical chemistry. These new approaches are often accompanied by the introduction of new technology for practical analyses. This is primarily done by transferring basic theoretical notions from other fields of science and

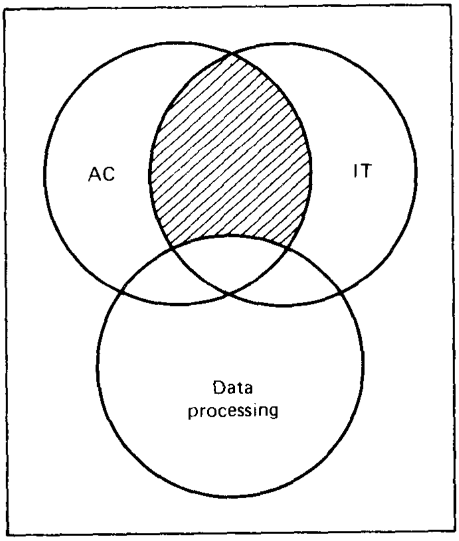

Figure 1.1 Venn diagram. AC, analytical chemistry; IT, information theory.

adapting them for the needs of analytical chemistry. In this way analytical chemistry has become a multidisciplinary field: information theory emerging from probability theory today forms a part. Information theory had its origins in the 1920s, when R. A. Fisher introduced some basic definitions; it spread extensively in the United States after 1945, through the work of N. Wiener, C. E. Shannon, S. Kullback, and others.

In this book, we will be concerned with the use of information theory to describe, evaluate, and optimize processes of obtaining information in analytical chemistry. We will not pay attention to automation and data processing, although information theory can be utilized in these fields as well. The sphere that is the object of our interest is shown by the Venn diagram of Figure 1.1.

All information theory is based on the notion that information and uncertainty are synonymous. Since probability theory is the mathematical study of uncertainty, it is fundamental to information theory. Without it, the central notions of information theory could not be pursued.

References

1. W. Fresenius, *Reviews on Anal. Chem.* p. 11, Akad. Kiadó, Budapest, 1977.
2. H. Kaiser, *Z. Tec. Phys.* **17**, 219, 227 (1936).
3. H. Kaiser, *Anal. Chem.* **42**, No. 2, 24A; No. 4, 26A (1970).
4. K. Eckschlager, *Z. Anal. Chem.* **277**, 1 (1975).