PATTERN RECOGNITION APPROACH TO DATA INTERPRETATION

Diane D. Wolff
University of Arizona
Tucson, Arizona

and

Michael L. Parsons
Department of Chemistry
Arizona State University
Tempe, Arizona

Plenum Press • New York and London

Library of Congress Cataloging in Publication Data

Wolff, Diane D., 1953–
Pattern recognition approach to data interpretation.

Bibliography: p.
Includes index.
1. Statistics—Data processing. 2. Experimental design—Data processing. I. Parsons, M. L. (Michael Loewen), 1940– . II. Title. III. Title: Data interpretation.
QA276.4.W64 1983 519.5'028'54 83-9624
ISBN 0-306-41302-7

A Division of Plenum Publishing Corporation
233 Spring Street, New York, N.Y. 10013

Printed in the United States of America

This book is sincerely dedicated to
Paul, Doug, and Ginger,
without whom the finished product
would never have become a reality.

PREFACE

An attempt is made in this book to give scientists a detailed working knowledge of the powerful mathematical tools available to aid in data interpretation, especially when confronted with large data sets incorporating many parameters. A minimal amount of computer knowledge is necessary for successful applications, and we have tried conscientiously to provide this in the appropriate sections and references. Scientific data are now being produced at rates not believed possible ten years ago. A major goal in any scientific investigation should be to obtain a critical evaluation of the data generated in a set of experiments in order to extract whatever useful scientific information may be present. Very often, the large number of measurements present in the data set does not make this an easy task.

The goals of this book are thus fourfold. The first is to create a useful reference on the applications of these statistical pattern recognition methods to the sciences. The majority of our discussions center around the fields of chemistry, geology, environmental sciences, physics, and the biological and medical sciences. In Chapter IV a section is devoted to each of these fields. Since the applications of pattern recognition techniques are essentially unlimited, restricted only by the outer limitations of the human imagination, we realize that this chapter will serve more as an illustration of the almost boundless diversity of problems and approaches in each field than a comprehensive review of published applications. We feel, however, that this goal has been well addressed by much of the scientific pattern recognition literature available today, so our time spent on it is minimized.

A second goal is to introduce the scientist to the techniques currently available. In Chapter II, a brief introduction to the mathematical foundations for each technique is given and references are provided for those interested in more detailed mathematical discussions and explanations. Particular stress is given to the interpretation and applications of the pattern recognition tool. Statistical results can be quite meaningless unless the correct questions are asked, and the data handled in an appropriate manner. Each technique has its advantages and disadvantages, and these are the focus of the discussions in these chapters. A basic knowledge of statistics is assumed, and references are listed for details beyond the scope of our present treatment.

Thirdly, we feel that there exists considerable need for a book that stresses the "how-to-do-it" approach. Many statistical books suffer the disadvantage of presenting detailed mathematical explanations of techniques with no suggestions about how best

to approach specific real problems, and no references to appropriate source material. In Chapter III, approaches to problems utilizing three major statistical packages, SPSS, BMDP, and ARTHUR are considered. The availability of each package is given in Appendix V and a "how-to" step-by-step explanation and comparison of each program within these packages is provided in the chapters. The purpose of these chapters is to motivate the scientist to utilize these powerful, readily available pattern recognition techniques with his or her own experimental data. The authors will feel this book successful if we manage to motivate our readers to do so with some degree of regularity.

There remains last, but first, Chapter I. Choices are always the most difficult problem with the many statistical approaches to data analysis. In this chapter we attempt to answer questions about data choices and organization, about tool choices and their applications, about the most appropriate questions the investigator might ask of the data, and finally, about the philosophies of interpreting results.

This is by no means meant to be a comprehensive mathematically oriented statistics book. We view it more as an active tool to aid in one very important step of scientific laboratory work—namely, data interpretation.

CONTENTS

II
PATTERN RECOGNITION APPROACH TO DATA ANALYSIS

III
IMPLEMENTATION

IV
NATURAL SCIENCE APPLICATIONS

I

PHILOSOPHICAL CONSIDERATIONS AND COMPUTER PACKAGES

I.1. PHILOSOPHICAL CONSIDERATIONS

INTRODUCTION

Scientific research has become an area of enormous data production. Reams of data are routinely generated as the trend towards quantification within the sciences has increased. The need for sound mathematical methods for analyzing these data is crucial. Often data are continually produced without stopping for such analyses. The result can be the production of large amounts of inferior data. Studying the mathematical patterns underlying the data can often help to determine the best next step in the analysis and to draw meaningful conclusions from the data already gathered. Moreover, such studies may reveal that better experimental designs can be devised and implemented effectively. Also, underlying properties of the data, not directly measurable, but related to the data being produced, may be studied, and predictions related to the scientific content of the data, and future data, become possible.

This book will introduce the noncomputer-oriented scientist, as well as those familiar with computer usage, to some techniques available within most scientific communities for such data analyses. Through this study, better data collection methods may be revealed, and the time invested in the creation of inferior data or in suboptimal use of good data may also be minimized.

DATA CONSIDERATIONS

To apply mathematical analyses to scientific data, one must first thoroughly understand the scientific aspect of the problem—what are the goals for the experiments and associated data collection? It is critical to approach the data analysis problem without losing the scientific content underlying the data. This points to the importance of interaction between the scientist and the computer. Underlying structures in the data are

often too complex to be unraveled by the scientist working alone. Hand calculations become impossible. But the ease of access to "number crunching" by today's sophisticated computers is often the cause of excessive data collection and its associated interpretation abuse. The computer cannot tell the difference between proper and improper applications supplied by the investigator, but can only blindly follow the computational algorithms.

Such data analysis may encompass anything from a single set of computations to a complex sequence of steps, each suggesting further procedures. The results of one step may point towards other useful analyses, or single out the ones already taken that may be unnecessary. But a complete and informed interaction between the scientist and the computer is necessary for optimal results.

The mathematical tools for these data analyses are not new. The fundamental ideas are related to the multivariate *normal distributions* derived in the 1930s. But the computer technology has had to catch up with the scientists' abilities, and it is only very recently that the two have become a compatible pair.

PATTERN RECOGNITION APPROACH

This data analysis problem is often referred to as a *pattern recognition* approach. It is the goal of such a study to identify patterns in the experimental data produced in an investigation, and to draw intelligent conclusions from them. For example, let us assume that the scientist is interested in environmental problems and more specifically in environmental air pollution. Assume that this project concentrates on the determination of the causes for the changes in visibility in a large metropolitan area from day to day. The scientist must first determine what data to consider. Without a prior knowledge of the exact causes, the scientist should generally still be able to use his or her scientific expertise to determine probable causes of this visibility variation. To begin with, broad general causes may be considered. The experimenter might start thinking about what general parameters could be important. For instance, weather, particulate concentration, or gas concentrations in the area might be considered.

To expand on these, he or she must then consider more specific knowledge of each. The weather almost certainly plays an important role in visibility. But is it the humidity, solar radiation, precipitation, wind speed, none of these, or a complex combination of all of these (and possibly others) that affect it? Are total organics, trace metal concentrations, sulfides, chlorides, or a complex combination of these important in the pollution itself? Without undertaking critical data analyses and attempting to recognize underlying patterns at each step of the study, much data may be collected that has some scientific importance but is useless in answering the scientist's original question.

In the computer-oriented pattern recognition approach each type of measurement made is considered a variable. In the above example, humidity, solar radiation, and each such measurement is a variable. Mathematically, each variable (or type of measurement) can be considered a dimension in space. For example, if at given times of the day the humidity and visibility were measured, we would consider it a two-dimensional anal-

ysis problem, and plots in two-dimensional space with values of humidity versus visibility could be made. Most problems involve many more variables than this, and therefore become problems in multivariable analyses. This is where the computer becomes an invaluable tool. The scientist has trouble visualizing more than a two- or three-dimensional space, but statistical computations are just as valid in *n*-dimensional space. Mathematical algorithms determine how *n*-dimensional calculations will be performed, but it is the job of the scientist to find the meaning of such calculations. For instance, the scientist can measure visibility, humidity, and solar radiation at given times on given days. A plot of visibility versus solar radiation, or visibility versus humidity, can be made, but usually such simple relationships are not adequate to describe the reality of the problem. What if visibility were really related by some complex mathematical function to both humidity and solar radiation? The best the scientist could do would be to generate enough measurements of the three at given times of the day and week to try to recognize patterns in the data. Computer-implemented multivariate analysis (more appropriately called "pattern recognition") can aid the scientist in this endeavor. Definitions for pattern recognition terminology are given in Appendix I, and the terms are italicized when first used in the text.

The pattern recognition approach is founded on a few basic assumptions. The first is that there *are* underlying patterns in the data; in the present case, such a variable as visibility is somehow related to solar radiation and humidity. The scientist may not know this as a fact. The best experiment, then, is to make measurements that the scientist believes are necessary to understand the problem. At least some of these experimentally determined variables must be relevant to the problem for pattern recognition techniques to be successful. If the variables are indeed related, not only will pattern recognition techniques lead to the exposure of the underlying relationships, but it will also help reveal which variables are most critical in understanding the problem. The scientist can therefore usually determine which measurements are most important in solving the problem, possibly thereby saving time and money.

PATTERN RECOGNITION QUESTIONS

Pattern recognition can be applied to a much larger variety of problems than the example given above. The critical step in the pattern recognition analysis is often that of formulating the questions to be asked, and the success of this formulation usually determines the success of the outcome. One question may be directed towards the relationship between various measurements for predicting a specific property, such as that of visibility. A second question may be directed towards fitting the data to a given mathematical model. A third approach may consist of separating the data into various *groupings* (either defined or naturally occurring ones) with the intention of determining the underlying patterns of these groupings, what the groupings are, whether the groupings are statistically valid, or what causes the groupings. It may be possible to predict to which groups unclassified samples belong according to the measurements made. For example, assume that our environmental scientist now becomes interested in water pol-

lution. And let us further assume that the researcher would like to characterize bodies of water as either (1) too polluted to support fish life, or (2) able to support it. Now that two groupings or *categories* have been created, the question may be asked: What determines these groupings? Again, scientific expertise is necessary to determine what specific measurements could be important. The application of pattern recognition techniques will then help him or her decide which actually are the key parameters. A second question that might be asked is whether a certain body of water is able to support life (or asked in a different way, does this body of water belong to group 1 or 2?). The investigator would then make those measurements deemed to be important, using pattern recognition techniques to determine into which group that particular body of water best fits. Alternatively, it may be desirable to check the validity of the groupings, and one might, in the course of these studies, find a group of lakes that can support certain more hearty life forms, but not other fragile types. In addition, the investigator may find regrouping necessary as the data are studied in more detail. New questions may arise and therefore new answers may be sought. Many times one question will lead to the realization that more data are necessary. And the study of such data may result in the formulation of further questions. There is no one single correct approach. Each step can lead to a variety of subsequent lines of inquiry, but the results of each should be observed and then analyzed before the next step is taken. This will keep the process in line, optimize the analytical procedures, and therefore minimize the time spent on the problem.

PATTERN RECOGNITION NOMENCLATURE

One of the problems of *multivariate* statistical *analyses* is the inconsistency in the nomenclature used by various scientists. In this book, the nomenclature generally follows the descriptions by Kowalski. A brief example of such usage follows.

An *object* or *case* is an entity for which a list of characterizing parameters are made. The set of characterizing parameters is known as a *data vector* or *pattern* for the object or case. Each parameter is called a measurement, *variable,* or *feature.* The objects can be grouped according to some defined criteria of similarity into categories, or groupings. If the categories are known, and the goal of the analysis is to find the causes of such groupings, the study is referred to as supervised analysis (or *supervised learning*). If the major goal is to determine these groupings, the process is called unsupervised analysis (or *unsupervised learning*). Whether supervised or unsupervised learning is used to characterize objects and study relationships among them, it is called the *training step* since the scientist (through use of the computer) is being "trained" to recognize either defined or natural categories. It then becomes possible to add objects to the analysis whose membership in the groups is not known. This process, called the *test step,* is used to predict their inclusion in or exclusion from the groups under study. If the scientist uses test cases whose membership is known, but not given to the computer as information, a test of the prediction power of the pattern recognition analysis would be possible.

For example, returning to the example of the water environmentalist, his or her two categories are those waters where life can survive, and where it cannot. The objects of inquiry are the various lakes. Each measurement (for example, temperature, algae content, etc.), is a variable or feature. The set of measurements for a given lake is called the data vector or pattern for that lake. To determine what causes the two groupings, techniques involving supervised learning are undertaken. To determine whether any other groups (or another totally different grouping) exists, unsupervised learning can be used. Both are methods in the training step. If additional lakes about which the life functions are unknown are then considered, their investigation is called the test step and the additional lakes are considered the test set.

Now assume that at least the general questions to be asked can be determined and that data, at least some of which are pertinent to the problem, can be collected. The next step is to determine how to make this data computer-compatible (see Appendix III). The computer does not know what kind of data it is receiving. Therefore, it must be coded into computer language in some form meaningful to the computer's statistical analysis programs.

VARIABLE CODING

Variables may take on certain levels of measurement, which are determined by the methods of data collection and the inherent precision of these methods. Four major levels of data measurement exist. The first (and probably the most often encountered in the sciences) is the continuous measurement, which assumes that any value within a certain range of values is possible for the analysis. An example would be the measurement of the percent of copper in a sample, knowing that the values must range between zero and 100%. For a set of copper analyses, a subset of this theoretical range is usually present. As with all continuous measurements, the zero value is defined and meaningful, and distances between values make sense, i.e., 20% is twice 10%. In actuality, the numbers possible for the percent copper in the sample are limited in accuracy by the precision of the experimental measurement.

The second level of measurement groups data into coarsely defined regions due to the semiquantitative or limited accuracy of the measuring technique. Each group has a unique position with respect to the other data, but specific values of distances are not known. The zero point is not usually defined. An example would be to group the percent of copper in a sample as either (a) $<25\%$, (b) 25%–50%, (c) 50%–75%, or (d) $>75\%$. Now the order is still important, but distances [for example, group (c) is not twice group (a)] are not explicitly defined.

A third type of measurement is binary coded data consisting of those variables where dichotomies, or only two possible values, exist. An example would be a group of samples designated simply as either containing copper or having no copper. This is often the situation with a "present versus absent" dichotomy, and it is often coded as 0 = absent and 1 = present. In this case, neither distances nor rank order apply.

The fourth type of measurement refers to unordered groups. The numbers assigned

are meaningless, and are used only as names or labels. An example would be coding the colors red, blue, green, and yellow as 1, 3, 2, and 4. Distances are meaningless and mathematical properties, such as adding and subtracting, are invalid. These could be replaced by four individual binary variables (i.e., the first being red, where 0 = absent and 1 = present; the second being blue, where 0 = absent and 1 = present; etc.). This, although mathematically valid, can become quite cumbersome.

Many of the statistical tools to be used in pattern recognition are affected by the level of measurement for the data. Many utilize statistical assumptions for their bases, and the validity of these assumptions in the data determine the validity of the results obtained. Most multivariate statistical analysis tools employed in pattern recognition are *parametric* tools. This means that the distribution of the variable values is known, or can be estimated. Further, a multivariate normal distribution is often assumed (see Appendix II). Tests to study the validity of this assumption will be discussed later in the text. Continuous levels of data measurements usually cause no problems. Usually each variable can be assumed to be normally distributed, and the combination of the variables jointly is also normally distributed.

The problems usually occur when the data takes on either a grouped or binary level of measurement. Pattern recognition techniques must be applied with extreme care to these variables, and interpretations should be approached cautiously. *Nonparametric statistics* should be strongly considered in these cases. Appendix VII describes this topic. The properties of these measurement types do not follow normal mathematical rules, and must be handled with care. The computer will not automatically check the underlying statistical assumptions during the coding of the data. It can only follow the program's algorithms, regardless of the data type given, and it is therefore critical for the scientist to check these types of things early on in the statistical analysis.

CATEGORIZATION OF DATA

The categories present in a data base can also be divided into four types similar to those for variables. The first, the continuous category, is, for example, like the lake eutrophication problem. It is possible to use binary categories as in the example above, i.e., whether a certain lake can or cannot support life. But in reality, it is probably more a continuous type category, one where at one extreme we have pristine conditions, and at the other extreme, totally stagnant water that can support no life. All values in between are possible. And distances are again meaningful, since if 20% of all fish life dies per year, this is a value twice that if only 10% does. Grouped data would be possible as well (0–25% die, 25%–50%, 50%–75%, >75%). Unordered categories bearing no distance-type relationships are also possible if the similarity of lakes is defined according to which variety of fish live and which die in each.

There may exist a combination of types of variables within any of these categories in a given pattern recognition analysis. Again, it is important to realize the statistical validity of the assumptions of each.