Chemometrics: Views and Propositions†

B. R. KOWALSKI

Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195

Received August 6, 1975

Science is hungry for new tools, and, when one is discovered, history often repeats itself. The interest and support rise rapidly at first. Then, if the use of the tool is not somewhat conservative, the tool is often misused. The reasons behind the misuse are complex and numerous but include pressures to publish in quantity and the desire to execute application "firsts". Misuse of a tool often leads to disenchantment with it, and, coupled with other factors (finite funds, etc.), progress will decline and then oscillate for a time. Eventually, if the tool is really useful, real progress can proceed.

Pattern recognition is a collection of problem solving methods that, in my opinion, collectively comprise a tool that can be applied to a wide range of information extraction application which includes several in the field of chemistry. It is, perhaps, no better or worse than several other collections of data analysis methods (standard statistics, optimization, etc.) and, like these many other collections, can be extremely useful in doing something that is sorely needed in chemistry: extracting useful chemical information from large amounts of measurements or raw data. I, and others, prefer to assemble any and all methods that can be used to extract useful chemical information from raw data under the general title of "chemometrics". With this assemblance in mind, I now present some of my views on the present and future of chemometrics (emphasis on pattern recognition).

RESEARCH DIRECTIONS

Since chemists have only played a small role in the development of data analysis methods, it may be difficult to see how method development can be ideally called chemistry. It is true, however, that some new methods are developed by chemists to handle the special needs associated with chemical data. In these rare but necessary cases, the research must stand up to scrutiny from mathematicians, engineers, and statisticians, as well as chemists. Data analysis needs within chemistry do not always enjoy the highest priority of mathematicians and other method developers. Therefore, I believe that if new needs are truly warranted, chemists should be allowed to develop methods and to be given credit for a service to chemistry. The eminent chemist, Dr. George S. Hammond, recognized the problem and a possible solution. In a condensation of his award address upon receiving the ACS Award in Chemical Education, Dr. Hammond states that [*J. Chem. Educ.*, **51**, 559 (1974)]:

> "Unfortunately, the expansively inclined information theorists tend to turn most of their attention to modeling complex biological and social systems. This is fine, but it leaves unattended the potentially more manageable matrix of chemical information. This is a pity because one of our problems is that we already have far more chemical information than anyone knows how to manage. The digest of current knowledge and evaluation of the prospects of proposed research for extending the power of the knowledge bank could probably be upgraded by use of techniques of information theorists. Fortunately, this is beginning to happen even before I give my official permission. Some interesting work is beginning to appear, introducing the subject of 'Pattern Recognition in Chemistry'. Obviously it will become possible, by shrewd use of computers, to look at many variable problems in chemical reactivity and seek hidden correlative patterns. I can see, for example, that all of the work that I did some years ago seeking and interpreting linear free energy relationships was a very crude exercise in pattern recognition. Obviously, the linear analytical methods that I used so laboriously are now totally outdated. If I were to reenter the field using pattern recognition techniques, I could probably extract far richer results from old data and be much smarter in planning to gather new information."

As to what constitutes new and viable research in the chemical application of chemometric methods, I feel that the emphasis should be placed on the application. This, of course, is a difficult philosophical question. Is the determination of a new set of wave functions that gives the best energies for molecular orbitals more important that the discovery of a new ion rearrangement mechanism in the ion source of a mass spectrometer? Questions such as these must, of course, be answered in some context.

As an analytical chemist, I am concerned about the enormous number of analytical measurements that are not used to advantage because of poor experimental design. Several years ago I became a student of pattern recognition and decided that a useful tool for chemical applications would be a large computer program consisting of *several* pattern recognition methods, statistical procedures, and a host of utility routines. Unfortunately, the "founders of pattern recognition" could not provide such a system, so I decided that it was worth the time and effort to build one. Upon using the system, three important points became clear. First, display techniques and pattern recognition methods could solve problems that were not amenable to solution by other methods. Second, *n*-dimensional measurement analysis, which is the primary advantage of pattern recognition, is a considerable advance over the analysis of simple measurement by measurement graphs. This point is extremely important as most of science has been bogged down in low-dimensional studies which severely limit the examination of multidimensional problems. Third, my limits of expertise in the possible chemical applications and my limited time were not making efficient use of the programming system we developed. The solution to the problem was to place the tool in the hands of other chemists. We are currently involved with distributing copies of ARTHUR (our collection of tested pattern recognition program developed with the aid of NSF funds) to many scientists around the world. This will serve to apply the tool to a much wider range of applications and, we predict, stimulate a steady

† Presented in the workshop on "Chemical Applications of Pattern Recognition", National Science Foundation, Washington, D.C., May 1–2, 1975.

growth of novel chemical applications. A more sophisticated system of interactive graphics/pattern recognition programs (developed with ONR funds) will also be released to qualified users in the future.

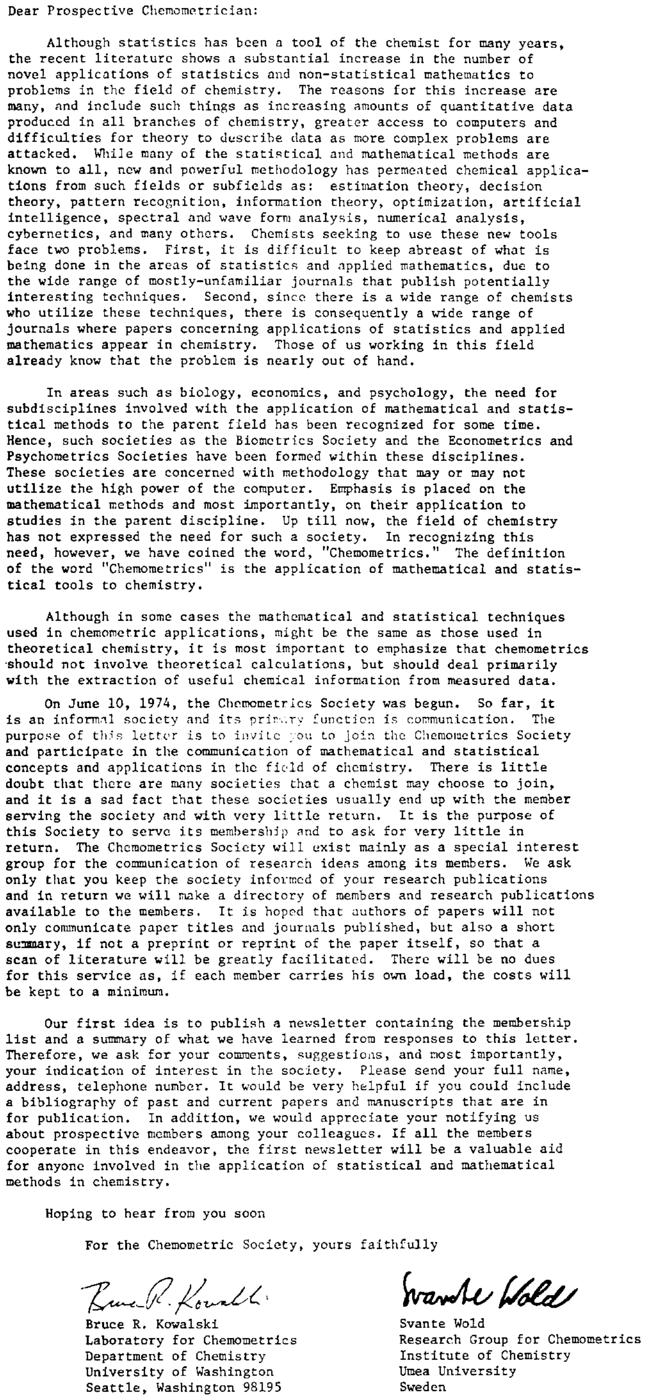
I must refuse to list the areas of chemistry to which pattern recognition is applicable. In the few short months that ARTHUR has been in the hands of other chemists, I am amazed to see it applied to areas I never dreamed of as possibilities. With the proper level of funding (discussed below) and the help of a nonbiased community of reviewers, I predict a most significant period of growth for chemometric methods.

FUNDING AND THE REVIEW PROCESS

Although academic researchers are free to choose their research topics, they require adequate funds and the opportunity to publish their findings. In this respect, the directors of research sections in funding agencies and the editors of research journals control, to a large extent, the input and output of research programs. A researcher who cannot publish his work is really in the same limiting situation as a researcher who cannot obtain adequate funding, and vice versa.

The chemical pattern recognizers have enjoyed a period of adequate funding and have had little trouble publishing their results. This meeting represents an opportunity to see how far we have come and where we are going. My view is admittedly prejudiced. Chemometrics has just begun to play a significant role in chemistry. I am confident that it will continue to grow in the long run, but I am quite concerned about the short-term period that we now face. Since agency directors and editors are few in number, they have an enormous responsibility and must rely on proposal and manuscript reviewers. If these reviewers are objective and fair, then I believe useful research will be funded and important papers will be published. Up until now, chemical pattern recognizers have been treated more than fairly. There is a fraction of scientists, however, who feel research in chemical pattern recognition has been less than ideal. To a certain extent, their criticism is warranted because there are poor papers in the chemical literature dealing with pattern recognition. When these scientists turn into reviewers, chemical pattern recognizers are in for a difficult time. I cannot defend chemical pattern recognition armed only with the past literature. I certainly will defend the need for more advanced chemometrics in the future.

I have personally met with an enormous amount of encouragement and only a small amount of discouragement from the community of chemists. Industrial chemists have real problems to solve, and the encouragement from them has been unanimous and considerable. A few academic chemists have vested interests in their own research projects and, unfortunately, reject new research if it threatens the supremacy of their own. It has been my experience that there are three different groups of scientists that criticize chemical pattern recognition. First, there are those nonchemical pattern recognizers (engineers, computer scientists, etc.) that feel there is a misunderstanding of the techniques used by chemists. They are correct to a degree, but their criticism does not cover all of chemical pattern recognition. Several nonchemical pattern recognizers are delighted that their methods are finding application in chemistry. Next there are chemists who are disappointed with the results thus far and undecided about the future. Again, this criticism is justified to a degree for reasons mentioned earlier. Third, there are chemists who are disappointed with past results and extrapolate their disappointment into the future. When I have had the opportunity to meet with group-one types, I try to obtain constructive criticism so as to improve our understanding. With the second group, I sympathize with their disappointment and try to instill optimism by reviewing the need for better data analysis methods and discussing our current applications of chemometrics (forensic identification, on-line spectral analysis, structure–function studies, etc). The third group seems to have one thing in common: a lack of understanding of the concepts and techniques of pattern recognition. When possible, I try to discuss concepts and techniques at a fundamental level.



Dear Prospective Chemometrician:

Although statistics has been a tool of the chemist for many years, the recent literature shows a substantial increase in the number of novel applications of statistics and non-statistical mathematics to problems in the field of chemistry. The reasons for this increase are many, and include such things as increasing amounts of quantitative data produced in all branches of chemistry, greater access to computers and difficulties for theory to describe data as more complex problems are attacked. While many of the statistical and mathematical methods are known to all, new and powerful methodology has permeated chemical applications from such fields or subfields as: estimation theory, decision theory, pattern recognition, information theory, optimization, artificial intelligence, spectral and wave form analysis, numerical analysis, cybernetics, and many others. Chemists seeking to use these new tools face two problems. First, it is difficult to keep abreast of what is being done in the areas of statistics and applied mathematics, due to the wide range of mostly-unfamiliar journals that publish potentially interesting techniques. Second, since there is a wide range of chemists who utilize these techniques, there is consequently a wide range of journals where papers concerning applications of statistics and applied mathematics appear in chemistry. Those of us working in this field already know that the problem is nearly out of hand.

In areas such as biology, economics, and psychology, the need for subdisciplines involved with the application of mathematical and statistical methods to the parent field has been recognized for some time. Hence, such societies as the Biometrics Society and the Econometrics and Psychometrics Societies have been formed within these disciplines. These societies are concerned with methodology that may or may not utilize the high power of the computer. Emphasis is placed on the mathematical methods and most importantly, on their application to studies in the parent discipline. Up till now, the field of chemistry has not expressed the need for such a society. In recognizing this need, however, we have coined the word, "Chemometrics." The definition of the word "Chemometrics" is the application of mathematical and statistical tools to chemistry.

Although in some cases the mathematical and statistical techniques used in chemometric applications, might be the same as those used in theoretical chemistry, it is most important to emphasize that chemometrics should not involve theoretical calculations, but should deal primarily with the extraction of useful chemical information from measured data.

On June 10, 1974, the Chemometrics Society was begun. So far, it is an informal society and its primary function is communication. The purpose of this letter is to invite you to join the Chemometrics Society and participate in the communication of mathematical and statistical concepts and applications in the field of chemistry. There is little doubt that there are many societies that a chemist may choose to join, and it is a sad fact that these societies usually end up with the member serving the society and with very little return. It is the purpose of this Society to serve its membership and to ask for very little in return. The Chemometrics Society will exist mainly as a special interest group for the communication of research ideas among its members. We ask only that you keep the society informed of your research publications and in return we will make a directory of members and research publications available to the members. It is hoped that authors of papers will not only communicate paper titles and journals published, but also a short summary, if not a preprint or reprint of the paper itself, so that a scan of literature will be greatly facilitated. There will be no dues for this service as, if each member carries his own load, the costs will be kept to a minimum.

Our first idea is to publish a newsletter containing the membership list and a summary of what we have learned from responses to this letter. Therefore, we ask for your comments, suggestions, and most importantly, your indication of interest in the society. Please send your full name, address, telephone number. It would be very helpful if you could include a bibliography of past and current papers and manuscripts that are in for publication. In addition, we would appreciate your notifying us about prospective members among your colleagues. If all the members cooperate in this endeavor, the first newsletter will be a valuable aid for anyone involved in the application of statistical and mathematical methods in chemistry.

Hoping to hear from you soon

For the Chemometric Society, yours faithfully

Svante Wold

Bruce R. Kowalski
Laboratory for Chemometrics
Department of Chemistry
University of Washington
Seattle, Washington 98195

Svante Wold
Research Group for Chemometrics
Institute of Chemistry
Umea University
Sweden

Figure 1.

Reviewers of manuscripts should not interpret editorial policy. Rather, they should concern themselves with the quality of the manuscript. If a paper presents the results of an application of pattern recognition to a chemical problem, one reviewer should make sure that the chemical problem really exists. The approach used must be novel and offer demonstrated advantage over existing approaches. Another reviewer should make sure that the pattern recognition techniques are understood and properly applied by the authors. All too often, the reviewers themselves do not understand the techniques. Reviews should always be constructive so as to lead the authors to a more polished publication and improve the quality of the literature. Nonchemists should be asked to review papers, but editors must weigh their criticism carefully because, for instance, many mathematicians consider all applications to be trivial extension of their work.

Reviewing a manuscript is considerably easier than reviewing a research proposal because more objectivity can be applied to the former task. Sometimes it is clear that a principal investigator (PI) is in "left field" and should not have written a proposal. Many of the proposals I have seen, however, are basically sound, and it is very difficult to predict the final results of two or three years of research before it has begun. An error in judgment can deprive chemistry of an important advance or, on the other hand, waste limited funds.

As I have stated, chemical pattern recognizers have enjoyed a period of adequate funding. How then should funds be distributed in the future? First, proposals should be revised by pattern recognition experts in order to determine whether or not the PI has an adequate understanding of the techniques to be applied. This is a most important step because misuse of the techniques leads to poor results, general confusion in the literature, and, finally, disenchantment with the techniques. Since pattern recognition allows the study of chemical systems where multiple measurements can be made and analyzed efficiently, these studies should enjoy a higher priority for funds. Real innovation will come in the design and application of measurement systems that generate complex, rather than simple data output. Several measurements can be made and pattern recognition used for data reduction. For example, measurement systems can be designed using pattern recognition, which find patterns of enzyme levels in clinical samples that indicate various health problems.

I also believe that the National Science Foundation should continue to support research aimed at developing more advanced chemometric methods. Funding should also continue for *novel* applications of these methods. I do not believe that the permeations of advanced mathematical and statistical methodology into chemistry is proceeding at a rapid enough pace. In many cases chemistry is behind areas such as psychology, economics, and biology in the application of new and powerful methods.

COMMUNICATION

I see no problem in defining the appropriate avenues of communication of the results of chemometric studies. Papers describing new methodology that could be used outside of chemistry should be sent to journals such as *Pattern Recognition* or one of the IEEE journals. Another useful avenue is the newly reorganized *Journal of Chemical Information and Computer Science*. Applications of chemometric methods should go to the proper audience of chemists. *Analytical Chemistry* should receive those papers dealing with applications to analytical problems of a general nature. Applications to solve a problem of interest to a narrower audience (i.e., forensic chemists) should go to the primary journal most read by that audience.

The real problem for chemometricians is keeping abreast of the many new and useful tools that are available. The letter shown in Figure 1 was written to help solve this problem as well as to provide a forum of chemometricians and a channel to statisticians, applied mathematicians, etc. Your presence at this meeting qualifies you to respond to this letter if you wish.

The Scope of Structural Isomerism[1]

DENNIS H. SMITH

Department of Chemistry, Stanford University, Stanford, California 94305

Received July 9, 1975

The variations in the number of structural isomers of organic compounds as a function of variation in atom type and degree of unsaturation[†] are discussed. These results provide quantitative measures which can be used to rationalize and extend intuitions of chemists on the scope of structural isomerism.

The fundamental concept of structural isomerism has been supported largely by intuitive ideas throughout the history of chemistry. Until recently there has been no systematic solution to the problem of specifying either complete sets of structural isomers for a given empirical formula or subsets based on constraints derived from a variety of sources (e.g., chemical isolation procedures, spectroscopic data). Problem-solving in this area, from student problems in introductory organic chemistry to complex problems of molecular structure elucidation, has been relegated to patient doodling with pencil and paper.

Central to the concept of structural isomerism is its scope, or the number *and* identity of each possible isomer. This problem has not been ignored by chemists and mathematicians. A recent review of uses of graph theory in chemistry by Rouvray[2] summarizes past attempts to treat mathematically various problems of isomerism, with examples cited. With the exception of Lederberg's approach to generation of acyclic isomers[3] and recent work by Sasaki and coworkers[4] and Balaban,[5] the work summarized by Rouvray yields only the number, not the identities of isomers.[11]

The DENDRAL algorithm for generation of acyclic isomers[3] provides a systematic approach to the study of structural isomerism in acyclic molecules. More recently, an algorithm[7] and a computer program based on this algorithm[8] have been developed for exhaustive generation of structural isomers, inclusive of isomers containing cyclic and acyclic components. With the scope of structural isomerism amenable to treatment by systematic procedures, we can now consider detailed questions about isomerism.

[†] The term unsaturation, or degree of unsaturation, is used in this paper to mean the number of rings plus double bonds, or "double bond equivalents."