Fresenius J Anal Chem (1990) 338:363–367

Fresenius' Journal of Analytical Chemistry

Analytical problem-solving, reference materials, and multivariate quality control: A chemometrics approach

Robert R. Meglen

Center for Environmental Sciences, Laboratory for Chemometrics, University of Colorado at Denver, 1200 Larimer St. Box 136, Denver, CO 80204, USA

Summary. Breakthroughs in sensor technology have augmented the chemist's measurement repertoire by introducing new kinds of detectors with improved selectivity and the capacity to perform simultaneous multi-species measurements. Thus, the electronic revolution has qualitatively and quantitatively changed the data matrices to which the analyst/problem-solver has access. The new chemical subdiscipline of chemometrics is developing powerful mathematical and statistical data analysis tools to exploit the electronic windfall and enhance data interpretation. Principal component analysis and graphical procedures have been used to examine the multivariate suitability of current reference materials in matching the concentration ranges and matrices for various food analyses. Principal component analysis has been useful in developing and exploring quality control information for the routine analysis laboratory.

Introduction

The past decade has been revolutionary to the measurement sciences. Microprocessor control and computer acquisitiȯn of data have increased the rate at which data are obtained. It is now possible to obtain large numbers of measurements in a fraction of the time previously required for much smaller efforts. Breakthroughs in sensor technology have augmented the chemist's measurement repertoire by introducing new kinds of detectors with improved selectivity and the capacity to perform simultaneous multi-species measurements. Thus, the electronic revolution has qualitatively and quantitatively changed the data matrices to which the analyst/problem-solver has access. In addition, the increasing demand for multi-species accuracy assessment is challenging the capacity of reference materials to provide validation. The new chemical subdiscipline of chemometrics is developing powerful mathematical and statistical data analysis tools (instruments of reasoning) to exploit the electronic windfall, enhance data interpretation and strengthen our inferences. The approach of data affluence requires re-examination of the analyst's tools and the analytical chemist's role in problem-solving. The analytical chemist is no longer just a supplier of data, but has assumed an increased role in converting data into useful information.

Analytical chemists who have participated in the measurement revolution have also lead the search for improved methods to exploit the power of increased measurement capacity and to interpret the large data matrices that they are producing. Unfortunately data tables containing large numbers of measurements yield slowly to traditional data analysis tools because these techniques are limited to handling one or two variables at-a-time. Furthermore, while human beings are very good at recognizing patterns, they are generally limited to dealing with two or three dimensions at a time. The p-variable problem requires a p-dimensional examination of the data. During the past decade an increasing number of data-burdened researchers [1–6] have "rediscovered" the power of multivariate statistical methods to enhance their exploration of large data matrices and multi-dimensional problems. The underlying principle or philosophy which has characterized much of the work in this area is that what makes problems complex is the existence of many interacting variables and that the one variable at-a-time has inherent limitations. A new chemical subdiscipline, chemometrics, has been developed to exploit multivariate techniques for chemical problem solving.

Chemometrics is the chemical discipline that uses applied math and statistical methods to (1) design or select optimal measurement procedures and experiments, and (2) to provide maximum chemical information by analyzing chemical data. While chemometricians concern themselves with all steps in chemical problem solving [7–9], this discussion will be limited to techniques that enhance data interpretation.

We begin by emphasizing that *data are not information.* As indicated earlier, data are easy to collect. Any machine can be programmed to mindlessly acquire numbers. What one really needs is information. We may view a database as a domain that requires probes and tools to extract relevant information. Just like the measurement process itself, appropriate instruments of reasoning need to be applied to the data interpretation task. The tools should serve in two capacities; to summarize the data, and to assist in interpretation. The objectives of summarizing are simply to show the data. That is, to provide a means by which the totality of the database can be viewed. We wish to present many numbers in a small space. In addition, the summary should make the large data sets coherent. This means that it should be possible to compare numbers of different magnitudes without a bias. The summary should provide a means to explore relationships among numbers from different measurement domains, geological, chemical, biological, atmospheric, and even social and economic characteristics. The objectives of the interpretive aids are to reveal the data at several levels of detail. Exploring the fuzzy data picture sometimes requires a "wide-angle lens" to view its totality. Other times it requires a "close-up lens" to focus on fine detail. The tools that we apply in this process should provide this flexibility.

Graphical techniques are particularly useful in addressing these data analysis objectives. They are useful at summarizing because tables of numbers become "pictures". They are non-theoretical, i.e. they do not require detailed models for effective use. Graphical tools are also advantageous because their interpretation requires little formal statistical training. Conventional numerical statistical procedures formulate hypotheses about anticipated behavior and focus attention on the expected. Graphical data analysis tools are powerful because they tend to focus attention on the unexpected. It is important to emphasize that while the graphical approach to exploratory data analysis has certain advantages over the numerical procedures, the empirical approach described here should be viewed as complementary to the more robust treatments that statistical methodologies afford. The fact that we cannot, by ordinary graphical techniques, construct simple representations for multidimensional data suggests that a dimension reduction technique must be applied to the data if we are to exploit the power of the graphical approach.

Methods: principal component and factor analysis

Two closely related techniques, principal component analysis and factor analysis, are used to reduce the dimensionality of multivariate data [10–12]. Factor analysis attempts to explain the correlation among a large set of variables in terms of a small number of underlying factors. Factor analysis begins with the assumption that the data come from a specific model where underlying factors satisfy certain assumptions. In this technique the emphasis is on transforming the underlying factors to the observed variables in order to enhance the interpretability of the data. If the factor model is incorrectly formulated or the assumptions are not met, then factor analysis will give erroneous results. Factor analysis has been successfully used in many chemical problems where adequate understanding of the system permits good initial model formulations [13]. Principal component analysis is similar to factor analysis in many respects. However, it employs a mathematical transformation of the original data with no assumptions about the form of the covariance matrix. The aim of this procedure is to determine a few linear combinations of the original variables which can be used to summarize the data set without losing much information. The remaining discussion is based upon this method of reduction and its use in summarizing and displaying complex data sets.

As indicated earlier, the key feature of many complex systems is that many variables interact with one another. Principal component analysis quantifies the variable interactions by computing the matrix of correlations for the whole database. The matrix of correlations is decomposed (factored) into two matrices by the mathematical tool of eigenanalysis. The scores matrix and loadings matrix provide a means by which one may derive the best, mutually independent axes (dimensions) that describe the data set. These axes are the so-called principal components. They are linear combinations of the original variables that arise out of the natural associations among the variables. They do not require the analyst to make any assumption about the data/variable structure. The utility of constructing a new set of axes to describe the data is that most of the total variance (information) in the data set may be concentrated into a few derived variables. This means that instead of having to depict the data on dozens of bivariate plots prepared from the original sample measurements, we can compute the location or principal component score of each of the observations in the new data space. Thus, we may depict most of the information on just a few two dimensional principal component score plots. This process may be viewed as projecting the original data from its multi-dimensional representation down to two dimensions. As with any projection, information is lost; but this technique maximizes the retention of information and quantifies the amount of information contained within each projection. In most chemical systems it is possible to depict 80–90% of the total information in less than a half dozen plots. While the information about relationships between the objects (samples) is obtained from the scores matrix, quantitative information about relationships/interactions among the variables is contained in the loadings matrix.

The second interpretive aid provided by the principal component analysis consists of interpreting the principal components. Recall that the principal components arise out of the natural associations among the variables and that they consist of linear combinations of the original variables. These variable groupings permit us to generalize behaviors into latent variables or features. By examining the contribution that each of the original variables makes to the linear combination we can begin to explore the "mechanisms" that define the data structure. These contributions are called the loadings. When several variables have large loadings on a feature they may be identified as being associated. From this association one may infer chemical or physical interactions that may be interpreted in a mechanistic sense. A small loading of a variable on a feature indicates that the variable is not associated with the other variables that comprise the latent variable; and, that it is unimportant in making distinctions along this dimension. The key element of this procedure is that we have developed a quantitative scale for characteristics which were not explicitly measured. The data have suggested how the variables may be grouped into latent features that summarize system behavior.

Having described this mathematical tool we shall now examine how it can be used to explore the structure of a multivariate database. The importance of multipurpose biological-reference materials for accuracy assessment have been described [14–16]. Adequate accuracy assessment requires reference materials (RM's) that approximate the sample matrix being investigated and that the analytes (for which the RM's are certified) approximate the chemical form and concentration range present in the unknown samples. Many existing reference materials have been used by analysts for single species accuracy assessment in food and nutritional studies. It was often possible to obtain RM's which could fulfill the key requirements of matrix and concentration similarity for one analyte at a time. However, the increased use of multi-species analytical methods has revealed the limitations that the available RM's afford for multi-species accuracy assessment. The need for a wider variety of RM's for inorganic analysis of foods has been described [16]. Using principal component analysis Wolf and Ihnat illustrated the problem of adequately representing a variety of food matrices in multi-species analysis. A similar analysis is provided here to illustrate the utility of principal component analysis in the analysis of multivariate data.

Results and discussion

Average concentrations for nine inorganic nutrient concentrations for twenty-four food groups [17] were obtained from the literature [18]. The elemental concentrations for nine certified reference materials [19] were also tabulated and subjected to principal component analysis as outlined earlier and detailed in the literature [20–22]. The first principal component accounts for 58% of the original variance and consists mainly of a relationship to magnesium, potassium, calcium, iron and phosphorus. The loadings indicate that while each variable contributes to every principal component, different groups of correlated variables are major contributors to the principal component. The second principal component, consisting mainly of contributions from molybdenum, copper and zinc, accounts for approximately 14% of the original variance in the database. The third principal component consists, mainly, of the variance contributed by sodium content in the food (accounting for approximately 11% of the total data variance).

An important step in the exploratory analysis consists of transforming the original data into principal component scores. The scores indicate where each original data point (food group or RM) in 9-variable space lies along the new compound principal component axes. Figure 1 shows a plot of scores from the first two principal components. The cumulative eigenvalues of these two axes indicate that 72% (58% + 14%) of the total variance in the database may be viewed on this single bivariate plot. Individually the original 9 variables contain about 11% of the total variance. Using the multivariate approach, we have significantly reduced the dimensionality of the database and compressed the variance into a smaller number of axes for graphical examination. Disregarding, for the moment, the differences between the foods and the RM's, one can see that the scatterplot is not homogeneous. There are regions of densely clustered points suggesting that certain foods are characterized by similar chemical compositions and other points appear to be isolated, indicating unique chemical compositions. A dense cluster in the lower left indicates that these points correspond to lower concentrations of magnesium, potassium, calcium, iron and phosphorus. This cluster is also seen to be slightly below average along the vertical axis (principal component two). Along the vertical axis we see that these points are characterized by lower concentrations of molybdenum, zinc, and copper. (The variables in parentheses indicate variables with only minor contributions). Note how the 24 food groups (open circles) dominate this region of the plot, while the reference materials appear to "avoid" this region. This indicates that in principal component one variables the RM's appear to substantially exceed the concentrations of the foods categories, while the RM's are similar in the principal component two variables. Only bovine liver is substantially different in principal component two variables (Mo, Cu, and Zn). Thus, from this plot, we may conclude that while the matrices represented by these reference materials may adequately approximate the foods, the concentration ranges for some analytes are not equally well representing the foods.

Figure 2 shows the first and third principal component scores plotted. This plot is the second most information-rich plot. It depicts 69% (58% + 11%) of the variance in the database. In this plot the vertical axis represents mainly the elemental sodium composition. Viewing the data in these dimensions suggest no substantial dissimilarities between the foods and reference materials is found in the sodium domain. The preponderance of RM's located to the right along the horizontal axis (as seen already in Fig. 1) again illustrates the dissimilarity between the foods and reference materials in major inorganic compositions. Figure 3 shows the same data plotted in all three principal component dimensions. (The size of the circles indicates in pseudo-perspective the distance to the observer. Only the RM's are labeled for clarity). Note that in this three dimensional perspective, the RM's and foods are seen to be very dissimilar.

This example illustrates how this mathematical transformation has permitted us to exploit the interpretive power

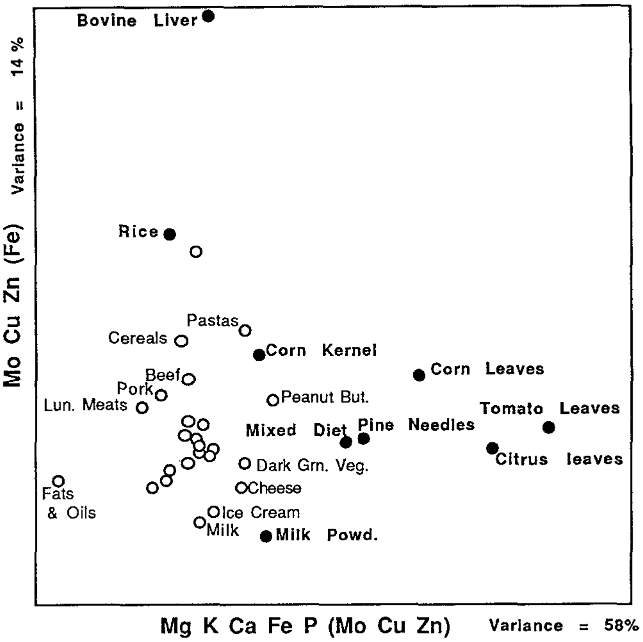

Fig. 1. Principal component plot showing relative locations of certified reference materials. ○ 24 food groups; ● reference materials; 9 inorganics; total variance 72%

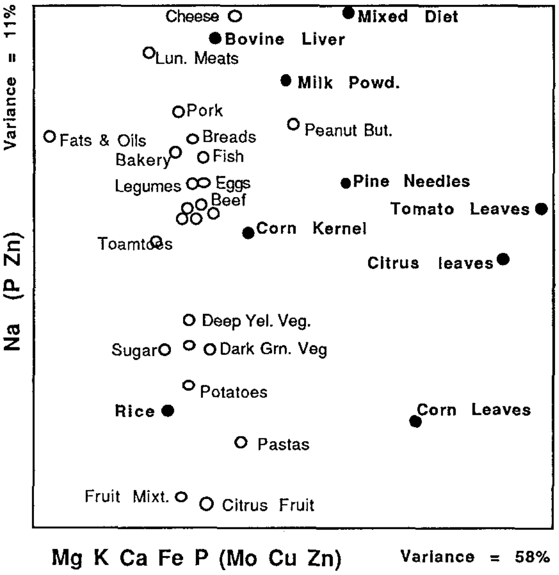

Fig. 2. Principal component plot showing relative locations of certified reference materials. ○ 24 food groups; ● reference materials; 9 inorganics; total variance 69%

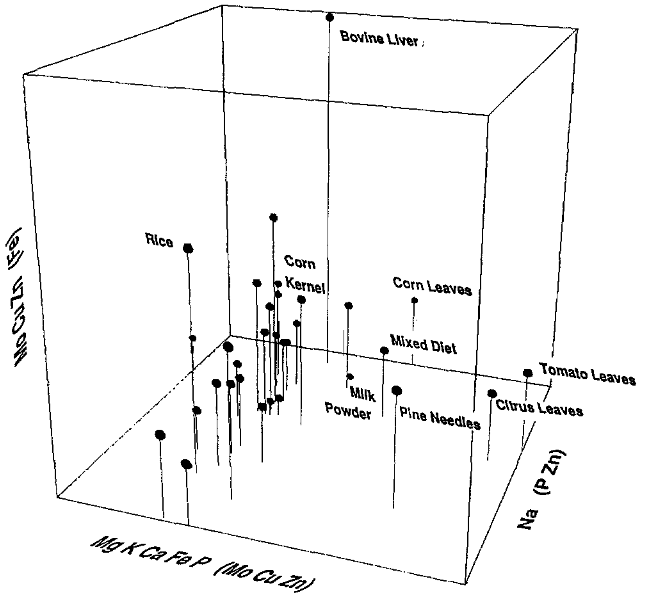

Fig. 3. Plot showing three principal components characterizing foods and reference materials, (only reference materials are labelled)

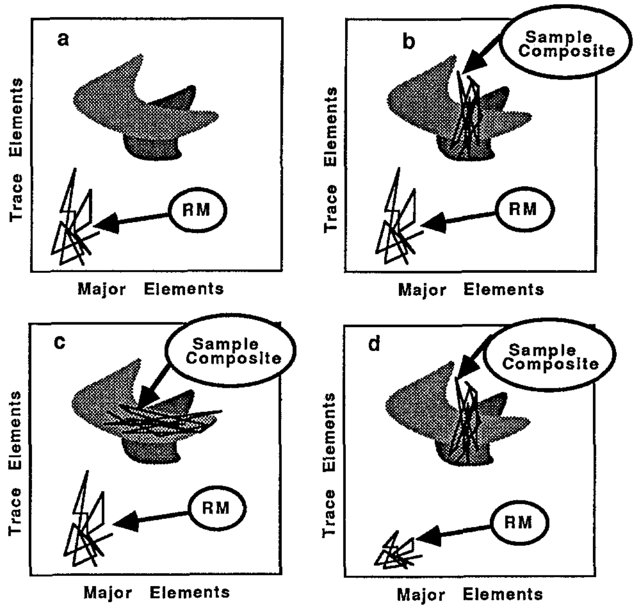

Fig. 4a – d. Example plots showing the use of principal components for examining quality control data

inherent in graphical display. It has enabled us to objectively select the most information-rich plots from the multitude of possibilities. It has shown how we may examine the structure of the data (to see the relationships among the observations/samples) in the most parsimonious fashion. The approach has permitted us to graphically depict the relationships between the foods and candidate reference materials. The utility of these plots can be extended to representations in the time domain for use in quality control. In many laboratories it is common to monitor analytical performance by repeatedly analyzing a reference material. Figure 4 illustrates how principal component plotting could be used for continuous performance surveillance. Figure 4a shows a hypothetical case in which two groups of foods (schematically shown as shaded regions) have been analyzed over a period of time. The repeated multi-species analysis of a RM is schematically shown as line segments connecting points in time-ordered sequence. In this case the envelope of variability for the RM is located at some distance from the actual samples, indicating a poor match between RM and sample compositions. The larger variability along the trace element axis relative to the major element axis indicates a difference in analytical precision. Figure 4b illustrates how a repeated analysis of a sample composite may be used to track analytical performance in the multi-species region occupied by the real samples. Note that while the RM and sample composite occupy different concentration regions, the variabilities in trace element and major element space are similar. Figure 4c illustrates a case in which the sample composite's variability differs from the RM. This variability pattern suggests a poor matrix match between sample and RM. Similarly, Figure 4d illustrates another example of how the pattern of variability may suggest a poor match between RM and sample matrix. Whenever a RM differs from the samples in concentration and matrix it is advisable to use two quality control references. This method of plotting shows how multi-species data may be used to detect and quantify the similarity of RMs and samples in quality control procedures.

A more general application of principal components analysis to multivariate quality control procedures is currently under investigation in our laboratory. In these studies we are examining the automated acquisition of quality control information from gas-chromatography/mass spectrometry, atomic absorption and inductively coupled plasma emission spectrometry. An automated procedure based on principal component analysis of samples and requiring no quality control samples is currently under investigation. The objective of this research is to develop an instrument recalibration strategy that minimizes instrument time devoted to quality control samples and can be utilized with automated computer controlled instrumentation.

Conclusions

The exploratory data analysis procedure described here is designed to uncover three main aspects of data; anomalous samples or measurements, significant relationships among the measured variables, significant relationships or groupings among the samples. Exploratory data analysis is an iterative process in which a wide variety of tools are employed [14]. The three primary tools used in this approach are factor analysis, principal component analysis, and cluster analysis. We have limited our discussion to principal component analysis, the most powerful technique. Additional information on ancillary techniques may be found in references [7, 8, 11, 13]. Other examples [21, 22] are available in the literature cited. Other equally powerful data analysis tools are available to examine large complex databases [2, 3, 7, 8, 13]. We have attempted to show that multi-species instrumentation generates large databases that require new approaches for data interpretation. The database is a domain that requires probes and tools to extract relevant information. Like the measurement process itself, appropriate instruments of reasoning must be assembled if data are to be fully exploited. As the scientific investigations become more complex, it becomes increasingly important to apply techniques that are as interpretationally sophisticated as the measurement instruments. The techniques described here provide one means by which scientists may keep pace with the growing analysis task.

References

1. Kowalski BR (1977) Chemometrics: Theory and applications, ACS symposium series no 52. American Chemical Society, Washington DC
2. Harper AM, Duewer DL, Kowalski BR, Fasching JL (1977) ARTHUR and experimental data analysis: The heuristic use of an algorithm. In: Kowalski BR (ed) Chemometrics: Theory and applications. ACS Symposium series no 52, American Chemical Society, Washington DC, p 14–52
3. Albano C, Dunn III W, Edlund U, Johansson E, Norden B, Sjostrum M, Wold S (1978) Anal Chim Acta 103:429–443
4. Varmuza K (1980) Anal Chim Acta 144:227–240
5. Erickson GA, Jochum C, Gerlach RO, Kowalski BR (1980) Applied pattern recognition: what to do with data after the measurements have been made, 65th Annual Meeting of the American Association for Cereals, paper no 99
6. Massart DL, Kaufman L, Coomans D (1980) Anal Chim Acta 122:347–355
7. Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York
8. Massart DL, Vandeginste BGM, Deming SN, Mischotte Y, Kaufman L (1988) Chemometrics: A textbook. Elsevier, Amsterdam
9. Deming SN, Morgan AL (1987) Experimental design: A chemometric approach. Elsevier, Amsterdam
10. Cooley WW, Lohnes PR (1971) Multivariate data analysis. Wiley, New York
11. Davis JC (1986) Statistics and data analysis in geology. 2nd edn. Wiley, New York
12. Gorsuch RL (1974) Factor analysis. W. B. Saunders, Philadelphia
13. Malinowski ER, Howery DG (1973) Factor analysis in chemistry. Wiley-Interscience, New York
14. Iynegar V, Wolf WR, Tanner J (1988) Fresenius Z Anal Chem 332:549–551
15. Wolf WR, Ihnat M (1985) Preparation of a total daily diet reference material (TDD-1) In: Wolf W (ed) Biological reference materials: availability, uses, and need for validation of nutrient measurement. Wiley, New York, p 179–193
16. Wolf WR, Ihnat M (1985) Evaluation of available certified biological reference materials for inorganic nutrient analysis. In: Wolf W (ed) Biological reference materials: availability, uses, and need for validation of nutrient measurement. Wiley, New York, p 89–105
17. Tsongas TA, Meglen RR, Walravens PA, Chappell WR (1980) Am J Clin Nut 33:1103–1107
18. Watt BK, Merrill AL (1975) Composition of foods: raw, processed, prepared, agric. Handbook no 8. U.S. Dept of Agriculture
19. NIST (1990) Standard reference materials catalog, 1990–91. U.S. Dept. of Commerce, NIST special publication 260
20. Meglen RR (1988) Chemometr Intell Lab Syst 3:17–29
21. Meglen RR, Sistko RJ (1985) Evaluating data quality in large data bases using pattern-recognition techniques. In: Breen, Robinson (eds) Environmental applications of chemometrics, ACS symposium series, no 292. ACS, Washington DC, pp 17–33
22. Meglen RR, Erickson GA (1983) Application of pattern recognition to the evaluation of contamination from oil shale retorting. In: Francis, Auerbach (eds) Environment and solid wastes. Butterworths, Boston, pp 369–381

Received June 22, 1990