Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

ENTROPY DATA ANALYSIS

Bush Jones
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803 U.S.A.

ABSTRACT

Entropy Data Analysis (EDA) is a new approach to the analysis of data. It embodies a framework which encompasses many classical statistical concepts; yet it goes beyond the classical framework both in the generality of its concepts and the correctness of its results. This paper provides an introduction to Entropy Data Analysis. Technical details can be found in the referenced papers.

1. INTRODUCTION

Entropy Data Analysis employs Shannon's concept of information to measure and analyze the information in a multivariate data set. Traditional analytic measures, such as interactions and effects, are easily obtained as well as entropy theoretic measures. There is a major difference between statistics and Entropy Data Analysis that goes beyond the capability to compute additional measures. Entropy Data Analysis is a coherent and "correct" form of analysis which does not contaminate results with assumed models or algorithms which introduce extraneous information. It uses the true measure of information, and it never adds or subtracts from the amount of information present in the data. Information is simply reorganized into a concise and meaningful form. Figure 1. depicts this process.

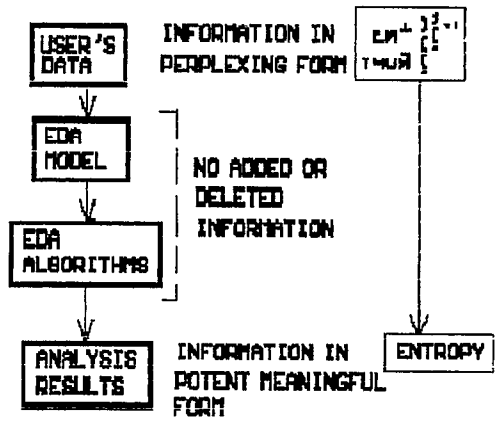

FIGURE 1.

2. PARAMETRIC OR NONPARAMETRIC?

Statistics is composed of two branches: parametric and nonparametric statistics. Parametric statistics deals with a parameter spac which defines a family of distributions as the parameters in the functional form of the distribution vary over the parameter space. In nonparametric statistics one makes no use of functional forms or parameters of such forms. EDA makes use of a functional form and its corresponding parameters, so it is not nonparametric. The parameters are number sum values which simply say "if you sum a set of numbers, you get a number which is their sum." Such simple sums (or simple truths) comprise the entropy data analysis model, and they exactly capture the information in the data (as defined in the sense of Shannon's information theory). The parameters are readily computed, zeroing in on a distribution which has precisely the information content of the data. Of foremost importance is that this process of distribution identification defines or breaks down the distribution in terms of its isolated information theoretic components (factors). Thus, what is critical to entropy data analysis is not distribution identification, but distribution description in terms of information theoretic components -- true components of interactions. The "fitting" process consists of selecting the minimal subset of the parameters which will reproduce the distribution to a specified tolerance. This is not parametric analysis as known in statistics.

Parametric statistics assumes one of a very limited number of distributions (often normal distributions or linear models in multivariate analysis), impels parameter values to accomodate the model, appends assumptions, and attempts to justify the whole process. Entropy data analysis does not begin by selecting a distribution; rather its equations allow for all possible distributions. In these equations are described all possible behaviors (no matter how nonlinear or complex) broken down into parametric components. It is the computation of the parameters themselves which identify the distribution from among all possible distributions, and these parameters are precisely correct for the given data. Note the parameters are not forced to a distribution -- they are in fact

the true parameters for the system (in the total data measurement sense) which produced the data -- and they decide the distribution from among all possible distributions. Of course, the results of the data analysis and the decisions made are dominated by the distribution employed.

3. THE BASIC MECHANISM OF EDA

EDA works with factors. If one picks a subset of the variables, and assigns a value (or cluster value) to each variable in the subset, then the combination that of values that results is called a factor.

EDA begins by forming a system that is in a state of complete entropy (or formlessness) as measured by the standard entropy equation of physics or information theory. It selects that factor (from among all possible system factors) that will do the most to shape this entropy system into whatever whatever system produced the given data. It then adds this factor to the entropy system. EDA continues ssuccessively selecting and adding factors to construct a system until the constructed system has the information of the data system to within a user specified tolerance. The constructed system is thus composed of a minimal number of parameters which accurately reproduce the behavior of the data system.

Each time a factor is added to the system, precisely the amount of information (as measured by Shannon's information theory) that exists in the data on this factor is added -- no more, no less. This is accomplished by keeping entropy at a maximum while adding the information on the factor from the data system to the constructed system. The constructed system is an unbiased reconstruction of the system produced the given data. This is true regardless of the complexities inherent in the original system.

Any form of error in the data can be construed to be a component of the total system that is captured by this approach. The constructed system is correct for all possible systems that could have produced the data.

The approach is so potently effective at concisely capturing system behavior because it employs the true measure of information, and because of the application of maximum entropy as each factor is added. It is well known in entropy theory that the application of this principle yields the most likely system from the information being employed. Each factor is selected based on on optimal information content, and that information is optimally incorporated into the system.

REFERENCES

Jones, B. (1985). "Determination of unbiased reconstructions." IJGS 10.

Jones, B. (1985). "A greedy algorithm for a generalization of the reconstruction problem." IJGS 11.

Jones, B. (1985). "Reconstructability analysis of general functions." IJGS 11.

Jones, B. (1986). "K-systems analysis versus classical multivariate analysis." IJGS 12.

Klir, G. and Cavallo, R. (1981). "Reconstructability analysis: evaluation of reconstruction hypotheses." International Journal of General Systems (IJGS) 7.

AUTHOR'S BIOGRAPHY

Bush Jones is a professor of computer science at Louisiana State University. He received the B.S., M.S., and Ph.D. from Southern Methodist University. He has worked extensively in industry as an applied mathematician and computer scientist. He is a member of ACM and SGSR, and he is president of a corporation which produces data analysis software.

Bush Jones
Computer Science Department
Louisiana State University
Baton Rouge, LA 70803
(504) 388-1495