JOURNAL OF MULTIVARIATE ANALYSIS 1, 316–346 (1971)

An Overview of Multivariate Data Analysis

A. P. DEMPSTER

Harvard University

A cross section of basic yet rapidly developing topics in multivariate data analysis is surveyed, emphasizing concepts required in facing problems of practical data analysis while de-emphasizing technical and mathematical detail. Aspects of data structure, logical structure, epistemic structure, and hypothesis structure are examined. Exponential families as models, problems of interpretation, parameters, causality, computation, and data cleaning and missing values are discussed.

1. INTRODUCTION

Over the past quarter century, the technology of computation has experienced a revolutionary development which continues unabated, so that sophisticated analyses of large and complex data sets are becoming rapidly more feasible, while repeated analyses of small-to-moderate sized data sets can be virtually instantaneous from table top terminals in the statistician's office. The hope has often been expressed that the technological revolution in computation would bring in its wake a comparable revolution in data analysis techniques, so that the science of data analysis would jump to a discernibly higher level of organization, power, and beauty. Mathematical statistics could then have a regeneration of its own as it faced the mathematical problems growing out of the new formulations of data analysis. Progress towards the projected new era has so far been limited, but many areas are visibly taking shape, and the next quarter century is full of promise as a period in which notable developments will appear. This paper concerns directions of conceptual development which, in the view of the author, will generate important components of the desired forward push.

Received October 20, 1970.

AMS 1970 subject classification: Primary 62P99; secondary 6202.

Key words and phrases: Data analysis; data structures; exponential families; multivariate methods; statistical computing.

Statisticians will increasingly require concepts adequate to frame analyses of complex highly multivariate data sets. But many academic statisticians have tended to define multivariate analysis narrowly, excluding even such obviously multivariate data types as factorial experiments, contingency tables, and time series. A preferable viewpoint would be to start with ordinary "univariate" data as the simplest case of multivariate data—relating one substantive variable (like weight) to an indexing variable (labelling the animals weighed)—and to place no rigid limits on the varieties of data types to be called multivariate. The standard types of textbooks of multivariate analysis (for example, [2, 8, 33]) present basic and elegant techniques built around multiple linear regression, correlations including canonical correlations, multiple linear discriminant analysis, and principal component analysis. The narrow range here reflects categorizations more natural for mathematical statistics than for applied statistics. It is encouraging that the direct predecessors [27, 28] of the new *Journal of Multivariate Analysis* embrace a broad outlook.

Theorists of multivariate analysis clearly need to venture away from multivariate normal models. One might also hope for less emphasis on technical problems within specific theories of inference, whether frequentist or Bayesian, and whether decision-oriented or conclusion-oriented. Instead, attention should be directed towards questions posed by data, for example: It is plausible to ignore certain possible relations? How should one sift through large arrays of interaction-type parameters to find those which are likely to be reproducible? What simple systems of functional forms of distributions are likely to be adequate to fit the data without being overly rich? What are the criteria which measure fit? Can simplifying structure be acceptably introduced into complex models, as by specifying a prior distribution of parameters which itself has estimable parameters? What quantities of importance does a certain model predict? What quantities can be robustly estimated, and what others, such as properties of tails, are sensitive to distribution assumptions? How are wild values to be detected and handled? Eventually, a more utilitarian and catholic viewpoint may lead to new categorizations of the subject of statistical analysis, and thence to new emphases in theoretical statistics.

The data analyst envisaged here is a professional who sits in a central position among investigators with data, theoretical statisticians and computer specialists. Among these traditional scientific types, he should serve the objectives of communication and integration. His direct contributions are to the development of new techniques and to the accumulated experience in the use of many techniques. For this, he needs the wisdom to evaluate proposed techniques along dimensions of efficiency and resistance to error, both statistical and computational, and along the dimension of relevance to the substantive scientific enterprise involved.

2. Aspects of Data Structure

A primary need is for a framework which permits adequate classification and description of a data set for data analytic purposes. The framework sketched below goes beyond the logical or network aspects of data structure (see Section 2.1), to include basic ways of thinking about data which help bring out meaning (Section 2.2), and to include formal mathematically expressed hypotheses (Section 2.3). The three levels of data structure are ordered in the sense that the second draws on the first while the third draws on the first two. Together they are conceived as providing the basic ingredients for the analyst's model of his data.

2.1. *Logical structure.* Any data set can be represented as a list of values of variables where each value must be tagged by identifiers for the variable, for the type of unit, and for the specific unit described by the value. For example, the value 121 may refer to the result of an I.Q. test devised for 8-year old children generally and applied to John Doe in particular. Usually, however, a data set can be represented much more compactly than in the list form with every value tagged by three pieces of information. Compactness is possible because the data have a *logical structure* defined by interrelationships among the variables and units involved.

Understanding the logical structure of a data set provides more than a guide to efficient physical representation of the data. Such understanding is also a fundamental prerequisite for understanding at the deeper level of substantive scientific information conveyed by data.

The practice in statistics and in most fields of application has been to treat each data set on an *ad hoc* basis, i.e., to obtain a grasp of the logical structure and to use this knowledge in data representation and in motivating analyses, but still to leave the understanding of structure implicit in the sense of not relating the particular structure to a highly developed, inclusive, multifaceted and formal typology of data structures. Indeed, in the present state of the art, only a rudimentary description of the required typology is available. Improved descriptions may yield dividends in uncovering important directions of progress for multivariate data analysis.

The basic concept is *variable.* Variables can be conceived as substantive variables or indexing variables (cf. Section 2.2). Usually a given variable is viewed as predominantly one or the other, but a dual viewpoint is always possible. Thus, I.Q. is usually regarded as a substantive variable related to educational psychology, while the names of 8-year old children are mainly values of an indexing variable. But the latter could convey information about national origins, and thus become substantive, while I.Q. could be regarded formally

as a device for stratifying individuals, for example, to construct a frequency distribution, which is more akin to indexing individuals than to evaluating them in some meaningful way. Since considerations of logical structure relate primarily to the indexing or stratifying aspect of variable conception, a variable should be viewed in Section 2.1 as a logical device which groups units into categories, viz., categories defined by a common value of the variable.

It is clear that an important piece of information about each individual variable is the mathematical space in which the variable takes values. The chief instances are (a) dichotomy, such as YES or NO, (b) nonordered polytomy, such as 4 different chemical drug treatments, (c) ordered polytomy, such as GOOD or FAIR or POOR, (d) integer response (usually nonegative, often counts), and (e) continuous response, at least with enough digits recorded to make the assumption of continuity an adequate approximation. Mixtures of these types also appear. The description of the logical structure of a given set of data begins naturally with a list of all the variables involved, together with a description of the space of possible values for each variable.

A suggested second element in the description is a *tree structure* where each node corresponds to one of a complete set of k variables (including indexing variables) and where the k nodes are interconnected by a set of $k - 1$ directed lines. It should be emphasized that each node corresponds to a variable as a concept, not to a specific value or data point. Figures 1a,b,c,d illustrate typical variables. A variable with no branch directed into its node may be called a *root* variable. Most often, a single data set has only one root variable which is the indexing variable of the individuals or units or atoms described by the remaining variables. The nodes corresponding to the remaining nonroot variables each have an entering branch coming from the node corresponding to the indexing variable of the set of units to which the variable applies. The standard example of multivariate analysis, as illustrated in Fig. 1a has a single type of unit measured on a set of substantive variables. Figure 1b illustrates how there can be more

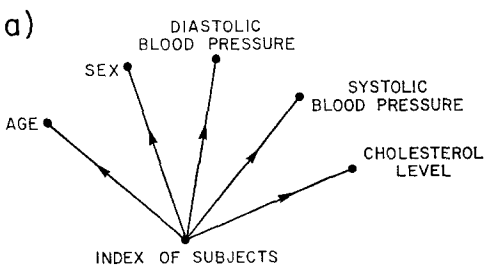

FIG. 1a. Tree structure for standard multivariate sample.

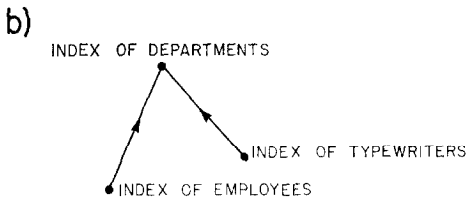

FIG. 1b. Tree structure with more than one root variable.

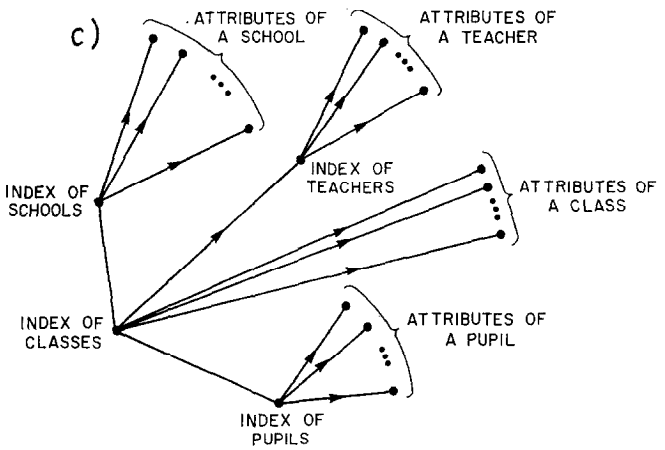

FIG. 1c. Hierarchical tree structure.

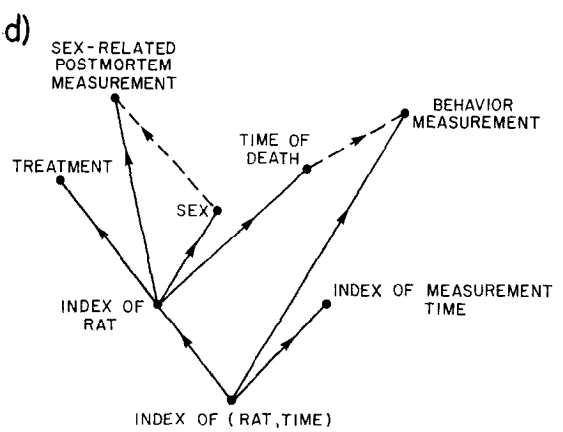

FIG. 1d. Tree structure with an artificial root variable and with variable conditioning.

than one root variable and, correspondingly, more than one branch entering a nonroot variable. If a nonroot variable is conceived as an indexing variable whose possible values or levels are susceptible to further classification, then several branches may emerge from the node of an indexing variable, as illustrated in Fig. 1c. These hierarchical structures are familiar to practitioners of the analysis of variance, where the tree structure implies a list of possibly meaningful mean squares which can be computed from the data.

Figure 1d illustrates two kinds of issue which may complicate the use of a tree structure of variables. First, the choice of what to identify as units and as variables is not necessarily unambiguous. A common example involves time as an index variable. If an animal is weighed at successive points of time, the structure may be conceived as a string of different weight variables defined over the indices of individual animals. An alternative which may often be more streamlined is to create a new indexing variable of (animal, time) pairs, so that animal index, time index, and weight become three variables defined over the new indexing variable. The second issue concerns what may be called a *conditioned* or *filtered* variable, viz., a variable defined only over a subset of the values of an indexing variable, where the subset is determined by the values of one or more variables along directed branches growing out of the same indexing variable. The examples shown in Fig. 1d relate to certain variables being unavailable after death and certain variables being available only on one sex. The device suggested for denoting a conditioned variable is a second type of directed branch proceeding from the conditioning variable to the conditioned variable, represented as a dotted line in Fig. 1d.

It is interesting that the conditioned variable concept introduces a type of nesting in which variables are nested within categories of other variables, while the hierarchical tree structure illustrated in Figs. 1c and 1d introduces nesting in which units are nested within entities of a different type. These two types of nesting are logically distinct and should not be confused with each other.

The two aspects of variable structure described above, viz., the aspect of space of values and the aspect of tree structure, are nonspecific in the sense of applying to many possible data sets involving the same variables rather than applying specifically to a given data set whose structure is desired as a prelude to analysis. Some tightening of the structure may be possible if it is meant to apply only to a specific data set. Thus, some or even most possible values of a certain variable may not appear in the data set. Also, there may be *de facto* variable conditioning, as when certain information was not recorded or was lost on certain blocks of data.

A third aspect of logical data structure is *balance*. Balance refers quite directly to a specific data set. The terms balanced or partially balanced are often used in connection with designs involving blocking and factorial structures, but the

definitions given are usually highly specific to narrow situations or they are somewhat loose. A path to a definition which is both precise and general is to associate balance with the recognition of groups of symmetries under the permutation of levels of certain variables. For example, an array of R rows and C columns with n observations in each of the RC cells has symmetry under the $n!$ permutations of index levels within each cell and also under the $R!C!$ permutations of cells. If the number of observations n_{ij} in row i and column j varies, then there may be balance only within cells, unless n_{ij} depends only on i or on j, in which cases different groups of symmetries come into play. Another example of balance, although not often referred to as such, is the familiar multivariate data matrix of n rows and p columns giving the values of p variables on each of n individuals. The logical structure of such a data matrix is clearly invariant under all $n!$ permutations of the individuals. It is also invariant under certain permutations of variables, for example all $p!$ permutations of variables of the same type such as continuous measurements.

Balance makes possible the efficient storage of data as multiway arrays where the labeling of individual values of variables can be represented very compactly. Balance also introduces simple structure into mathematical models and related theory, such as when balance introduces simplifying orthogonalities into the models for analysis of variance. The benefits include simpler theory of inference, simpler computational algorithms for data analysis, and simpler interpretation of the results of analysis. The price is mathematical analysis of many special cases, for there are many kinds and degrees of symmetry possible in a reasonably complex data structure, and detailed development of techniques to take advantage of the symmetries is correspondingly varied. Finally, it may be noted that the concept of missing value is closely related to that of balance. From the viewpoint of logical data structure alone there is no reason to tag any particular absent observation as missing unless it destroys symmetry. An important area of concern about missing values in practical data analysis centers around repairing the damage from destroyed symmetry (cf., Section 7).

There is at present no polished typology of logical data structures for multivariate data analysis. The preceding discussion may raise more questions than it answers, but it will have served its purpose if it draws attention to the need for a more systematic and extensive foundation at this fundamental level. Increased cooperation among data analysts and computer scientists in this area would clearly be beneficial.

2.2. *Epistemic structure.* The rational process by which any data set contributes to a particular field of knowledge depends strongly on preconceptions and understanding associated with that field. Some of this previous knowledge is quite specific to the field, such as knowledge of how and why an I.Q. test was

devised. Other aspects are general across many fields and so become parts of a general scientific outlook. These latter aspects are central to data analysis and statistics, especially when they facilitate inference beyond the immediate facts of a particular data set to more general circumstances.

Three varieties of *a priori* knowledge will be surveyed briefly. The first of these was introduced above, namely the knowledge that the value or level of a certain variable contains substantive information within some recognized field of inquiry, as opposed to being simply an index variable. The second variety concerns a distinction between free variation and fixed variation. The third refers to symmetry conditions on *a priori* knowledge, and leads directly into probabilistic concepts. Each of these varieties represents a type of knowledge which comes with a data set, from its scientific context, and which is *not* empirical in the sense that the information in the data itself does not reinforce or refute the preconceptions.

As argued in Section 2.1, any variable is capable of substantive interpretation, although certain variables, such as names, are usually accepted mainly as indexing variables. The possible dual interpretation corresponds to the mathematical duality between a function as a meaningful entity in itself and the set of values of a function which formally resemble index values. The role of duality in the mathematics of multivariate analysis is stressed in [8].

Some variables are regarded as free in the sense of reflecting natural variation or experimental variation in response to controlled conditions. Other variables are regarded as fixed, sometimes because their values have been deliberately set, as in experimentation, but often because the context of the data set suggests that the variable should be seen as having a determining or causal role in the phenomenon under study rather than being a direct measure on the phenomenon. The most common statistical example appears in multiple regression analysis where there is a free or dependent variable to be related to a set of fixed or independent variables. A similar distinction appears throughout multivariate analysis, in analysis of variance, and in contingency table analysis, and the distinction is incorporated in turn into formal hypotheses or models (cf., Section 4). It is clear that substantive variables can be free or fixed. Index variables are most often taken to be fixed, since names are usually assigned, but sometimes even pure index variables can be reponse variables, as when a name conveys the winner of a race. Note especially that whether or not a variable is free or fixed cannot be deduced from the logical structure of the data, nor from the quantitative or qualitative information in the data. The distinction between free and fixed need not be entirely firm in a given situation, but it is conceptually important, and must be made on genuine *a priori* grounds.

Most standard data reduction techniques are anchored in a judgment that certain units or levels of an indexing variable are to be treated symmetrically.

The computation of sample moments or the display of a sample frequency distribution are prime examples where an implicit *a priori* judgment has been made to treat evenly among the sample individuals. When several indexing variables appear in a hierarchical structure, separate symmetry conditions may be applied at several levels of the hierarchy. The term *exchangeable* is used in the theory of personal probability for models which treat symmetrically all subsets of any given size from a set of units, and the same term can be used consistently in a broader data analytic context to describe the *a priori* symmetry judgments discussed above. An analogous symmetry judgment of *stationarity* is often imposed on time series data. Again, the term is usually applied to specific probability models for time series data, but is appropriate at a looser level of assumption to mean that any time stretch of a given length would be regarded and treated *a priori* like any other time stretch of the same length. Similar *isotropicity* assumptions can be made for data indexed by position in physical space.

The symmetries which define exchangeability, stationarity, and isotropicity should be viewed as idealized properties of prior knowledge associated with the idealized logical structure of a given data set. A very basic type of knowledge is the uncertain knowledge of answers to factual questions, this being the type of knowledge which the theory of probability aspires to deal with. Accordingly it is to be expected that the probability models which accompany the analysis of any data set will share the symmetry judgments deemed appropriate to the data set. The restrictions thus imposed on probability models do not entirely determine the specific mathematical forms of the models, but they powerfully restrict these forms, often to mixtures of random samples from hypothetical populations. The specific forms are hypotheses (cf., Section 2.3) which, unlike the symmetry judgments themselves, can and should be tested out on the data. For example, if multiple measurements are made on a sample of 100 human subjects, exchangeability together with the theory of probability says that the 100 subjects can be viewed as a random sample from some population, whence hypotheses about the form of that population become conceptually meaningful as questions about a hypothetical probability mechanism which could have generated the data. The original symmetry assessment is completely *a priori*, however, since any multivariate sample distribution is a conceivable random sample from some distribution. It may be possible to refute the hypothesis that sample individuals were presented in a random order, but not the judgment that, as an unordered set, any subset of n individuals was *a priori* equivalent to any other subset of n individuals.

2.3. *Hypothesis structure.* After logical and epistemic structures are in place, specific mathematical hypotheses are often introduced to motivate and guide analyses.

Many mathematical models describe approximate deterministic relations. Examples include the linear models which often accompany discussions of the analysis of variance, and the quadratic relations appearing in models for factor analysis where both factor scores and factor loadings must be estimated. Heuristic procedures for fitting such structural models find effective use in exploratory data analysis.

The traditional models of mathematical statistics are probability models. Typically, repeated observations on a vector of variables are regarded as drawn from a multivariate distribution. The distributions generally have smoothness properties, and may depend smoothly on hypothesized parameters and on the values of observables as well. Probability models can give greater precision and descriptiveness to structural models, for example, by fitting an error distribution which provides a formal tool for quantitatively assessing deviations from a simple deterministic model. Together with probability models comes the availability of formal tools of statistical inference for testing fit, estimating parameters, and making uncertain predictions. The tension between data analysis without and with probability is explored in Section 3.

In Section 4, the discussion turns to an increasingly used general class of models based on exponential families of distributions. A broad review of mathematical models which have found substantial use in multivariate data analysis would be a worthwhile but very lengthy task, and is beyond the scope of this overview.

3. How Exploratory?

Data analysis proceeds in an exploratory mode or a supportive mode. In the former, the data analyst attempts to pry into the essence of a data set by examination from many angles, using graphs, charts, tables, scatterplots, etc. His tools may be categories of more or less well-tried and well-researched summary statistics, such as moments, quantiles, correlations, etc. Or he may use heuristic algorithms to fit rough models, as in clustering and scaling techniques. In the supportive mode, formal tools of inference are used to assess the plausibility and precision of formal hypotheses.

Mathematical statisticians have long concentrated their efforts mainly on the supportive side. In reaction to this one-sidedness, and also to the controversial character of the concepts of inference, some data analysts have claimed that a careful exploratory analysis can turn up everything of interest in a data set, rendering supportive techniques unnecessary. In general, however, there are benefits to be drawn from regarding the two modes as complementary and mutually reinforcing.

On the one hand, a too quick adoption of the supportive mode may lock the data analyst into a spectrum of hypotheses which would be instantly ruled out by casual inspection of a few plots of marginal distributions or plots of empirical relationships based on marginal summary statistics. On the other hand, it is very easy to discern apparently interesting empirical relationships, especially when many variables are sifted in search of relationships. Supportive techniques which can discount at least some of the biases introduced by data snooping would appear to be minimal supportive adjuncts to most exploratory techniques.

A more ambitious defender of supportive techniques could argue that the fitting of precise mathematical models to data, with its clear logical separation of sample and population concepts, is the key feature raising statistics to the level of a science, somewhat analogous to physics, in which exact and sophisticated mathematics is a central part. This is not to deprecate the more intuitive and empirical exploratory side, but rather to suggest that the two modes operating in resonance, as do experimental and theoretical physics, create a living science of considerable range and power. Data exploration suggests hypotheses. Formalized hypotheses in turn suggest under mathematical analysis new quantities to compute which may be illuminating in the same way that more naive data exploration can be illuminating. For example, the recent developments in likelihood methods for fitting log linear models to contingency table data provide natural ways to associate specific interaction terms with individual combinations of factors (cf., [18] and references cited therein). Having estimates of such interaction terms, it becomes natural to plot these estimates against the analogs of main effect estimates, with the aim of discerning special structure or relationships which might be unsuspected *a priori* and which would not be visible in a large multiway array of raw data. The mutual relations among probabilistic fitting procedures and data exploration procedures defines a very large and promising field for research in multivariate data analysis.

To a nonstatistician it may appear that the distinctions drawn here between exploratory and supportive methods are exaggerated In terms of the three aspects of data structure sketched in Sections 2.1, 2.2, and 2.3, both modes assume a common understanding of logical and epistemic structures, and they differ mainly in their approach to hypothesis structure. But even the simplest and most descriptive exploratory analysis is guided by hunches which are naturally conceived as imprecisely formulated hypotheses. The difference lies mainly in the degree of mathematical precision associated with hypotheses. Nevertheless, the difference of degree eventually becomes a difference of kind, as access to powerful mathematical and inferential tools, if used with intelligence and sensitivity becomes a highly significant factor in statistical methodology.

The explorative side of multivariate data analysis has given rise to classes of techniques which delve into data structure at the basic level of attempting

to discern logical structure not directly apparent in the data. The oldest such class is factor analysis, directed at finding important variables which are linear combinations of observed variables, where the criterion of choice generally depends on a principal component analysis; for example, if a group of variables is highly intercorrelated, then a linear combination can be found which explains much of the common variation underlying the correlation. Such combinations are candidates for factors. Forty years of cumulative development have made factor analysis a vast topic to review. A start may be made from [6, 21].

Computer technology has been an obvious catalyst in the development of various close relatives of factor analysis. In the latter, one draws on correlation coefficients which are inverse measures of distance between pairs of variables. In cluster analysis [19, 20, 22, 25, 37, 43], distances between pairs of individuals are used to group individuals into like categories, or more generally to produce hierarchical tree structures, with individuals at the tips of the smallest branches and with close individuals growing out of relatively far out common branches. Multidimensional scaling [5, 29–31, 35, 36] also relies on distances, often directly elicited by asking subjects to assess relative distances among pairs of entities. Again the objective is to place these entities, which may be attributes (e.g., colors) or individuals (e.g., nations), in spaces of a reasonably small number of dimensions.

In the present state of the art, it is difficult to evaluate procedures of clustering, scaling and factoring, except by reference to specific applications where the results can be judged meaningful relative to outside criteria, such as recognizing that a clustering technique groups obviously similar individuals. Supportive techniques are badly needed, and these in turn require formulation of acceptable probabilistic models and reliable inference procedures for these models. As argued in Section 5.1, it is a difficult task to match highly multivariate data sets to models with the right degree of complexity so that both adequate fit and informative inferences can be secured.

4. Exponential Families as Models

Multivariate data analysis needs a large and flexible class of hypothetical distributions of free variables indexed by the values of fixed variables. From this class, appropriate subfamilies would be chosen for fitting to specific data sets. The class of exponential models described below contains many subfamilies in common use. For example, the large body of techniques based on multivariate normal population models [2, 8, 33] derive from one basic subfamily, and the growing literature as referenced in [18] for contingency table analysis based on log linear models attests to the importance of another subfamily. Other sub-

families will surely be developed for applications. The many theoretical properties in common across subfamilies make it natural to consider the class as a unit.

The data analytic attitude to models is empirical rather than theoretical. In some kinds of modeling, physical theories or other established theoretical considerations may lead to specific parametric forms, as generally holds, for example, in the field of applied probability. When detailed theoretical understanding is unavailable, a more empirical attitude is natural, so that the estimation of parameters in models should be seen less as attempts to discover underlying truth and more as data calibrating devices which make it easier to conceive of noisy data in terms of smooth distributions and relations. Exponential families are viewed here as intended for use in the empirical mode. With a given data set, a variety of models may be tried on, and one selected on the grounds of looks and fit.

A particular subfamily will refer to a space S, in which a point represents a possible outcome for a set of free or response variables, and a second space R labeling possible outcomes for a set of fixed variables. The first stage of definition provides a family of distributions over S. The second stage specifies how the distribution over S varies with position in R.

Suppose that distributions over S are to be represented by densities relative to some familiar measure, where the two common examples are Lebesgue measure over continuous m-space and counting measure over discrete space. Suppose that $Y_1, Y_2, \ldots, Y_q$ represent q real-valued measurable functions over S. Then a typical family of exponential densities has the form

$$f(s) = \exp(\alpha + \alpha_1 Y_1(s) + \cdots + \alpha_q Y_q(s)) \tag{4.1}$$

for $s \in S$. Here $\alpha_1, \alpha_2, \ldots, \alpha_q$ should be regarded as free parameters, although certain choices may not produce a density with finite integral over S, and are, therefore, not permissible. Also,

$$\alpha = \alpha(\alpha_1, \alpha_2, \ldots, \alpha_q) \tag{4.2}$$

defines the normalizing factor $\exp(\alpha)$ which makes the density integrate to 1.

The functions $Y_1, Y_2, \ldots, Y_q$ may or may not be functionally related and may or may not uniquely determine a point in S. For example, if S is the space R^m with general coordinates $Z_1, Z_2, \ldots, Z_m$, then the family of multivariate normal distributions is determined by

$$[Y_1, Y_2, \ldots, Y_q] = [Z_1, Z_2, \ldots, Z_m, Z_1^2, Z_2^2, \ldots, Z_m^2, Z_1Z_2, Z_1Z_3, \ldots, Z_{\overline{m-1},m}], \tag{4.3}$$

where

$$q = m + m(m + 1)/2. \tag{4.4}$$

In this case, $Y_1, Y_2, \ldots, Y_q$ are functionally related and overdetermine the space. An extreme example arises if S is a space of M points and $q = 0$, so that only the uniform distribution over the M points is included. For an intermediate example, suppose that S consists of points (i, j, k) for $i = 1, 2, \ldots, P, j = 1, 2, \ldots, Q$, and $k = 1, 2, \ldots, M$, and suppose that the Y_i are P row indicators and Q column indicators, so that $q = P + Q$. The corresponding distribution of the triple (i, j, k) may be characterized as three independent multinomials, arbitrary over i and j but uniform over k. In this example, the row and column indicators $Y_1, Y_2, \ldots, Y_{P+Q}$ underdetermine the space S, but nevertheless have a redundancy.

The model (4.2) can be extended by the introduction of a set of fixed variables $[X_1, X_2, \ldots, X_p]$ defined over a space R. Each response $s \in S$ is paired with an observable vector $[X_1, X_2, \ldots, X_p]$ which influences or at least correlates with the response. The extended model is defined by

$$f(s \mid X_1, X_2, \ldots, X_p) = \exp\left(\alpha + \sum_{i=1}^{q} \sum_{j=1}^{p} \phi_{ij} X_j Y_i(s)\right) \tag{4.5}$$

which amounts to (4.2) with

$$\alpha_i = \sum_{j=1}^{p} \phi_{ij} X_j \tag{4.6}$$

for $i = 1, 2, \ldots, q$. It is understood in (4.5) that α depends on the ϕ_{ij} and X_j through (4.6) and (4.2). The extended model is restrictive in the sense that the distribution over S is permitted to depend only on linear functions of $X_1, X_2, \ldots, X_p$. One can, however, go a long way with linearity, as shown by the central importance of least-squares multiple regression analysis. The extended model is general in the sense that it provides a broad generalization of the standard normal model underlying multiple regression analysis.

The normal model, in a form suitable for q simultaneous predictors, is equivalent to a conditional distribution of $Y_1, Y_2, \ldots, Y_q$ given $X_1, X_2, \ldots, X_p$ as characterized by conditioning an original multivariate normal distribution of $Y_1, Y_2, \ldots, Y_q, X_1, X_2, \ldots, X_p$ taken jointly. There is, of course, no reason to restrict the use of the conditional model to cases where $X_1, X_2, \ldots, X_p$ could have been regarded as normally distributed. The assumption is essentially that the means of the normal variables $Y_1, Y_2, \ldots, Y_q$ depend linearly on $X_1, X_2, \ldots, X_p$ while their covariances do not. Another special model is the

multinomial logit [4]. Here the $Y_1, Y_2, \ldots, Y_q$ variables are indicator variables for a finite set of categories defining S. A theoretical umbrella including these two special cases is already quite large, but the scope for including other distributions of the form (4.2) is also large. For example, the normal model above can be extended by allowing both first- and second-moment structures to depend linearly on $X_1, X_2, \ldots, X_p$. And the multinomial logit is easily extendable to models where the finite set of categories form an m-way array so that suitably restricted log linear models can be used as the family of distribution over S.

Exponential models have attractive mathematical properties making it convenient to consider likelihood type analyses, which in any case are nearly obligatory from the theory of inference. In particular, a Fisherian style maximum likelihood analysis is a natural beginning point. Some relevant mathematical properties will now be sketched as a preliminary to a discussion of some technical aspects of fitting by maximum likelihood.

The data consist of repeated observations on $[Y_1, Y_2, \ldots, Y_q, X_1, X_2, \ldots, X_p]$. Denote n observations by $[\mathbf{Y}^{(l)}, \mathbf{X}^{(l)}]$ for $l = 1, 2, \ldots, n$, where $\mathbf{Y}$ and $\mathbf{X}$ denote general vectors $[Y_1, Y_2, \ldots, Y_q]$ and $[X_1, X_2, \ldots, X_p]$, respectively. Rewriting (4.5) in the form

$$\log f = \alpha(\boldsymbol{\phi}, \mathbf{X}) + \mathbf{Y}\boldsymbol{\phi}\mathbf{X}^T, \tag{4.7}$$

one finds that the log likelihood is

$$L(\boldsymbol{\phi}) = \sum_{l=1}^{n} (\alpha(\boldsymbol{\phi}, \mathbf{X}^{(l)}) + \mathbf{Y}^{(l)}\boldsymbol{\phi}\mathbf{X}^{(l)T}), \tag{4.8}$$

where the dependence of $L(\boldsymbol{\phi})$ on the data has been suppressed for notational convenience.

As will be seen below, consideration of $L(\boldsymbol{\phi})$ suggests that, alongside the matrix $\boldsymbol{\phi}$ of exponential parameters of the model, it is illuminating to consider a corresponding matrix of *moment parameters*

$$\boldsymbol{\theta} = \sum_{l=1}^{n} \boldsymbol{\theta}^{(l)} = \sum_{l=1}^{n} E^{(l)}(\mathbf{Y}^{(l)T}\mathbf{X}^{(l)}) \tag{4.9}$$

where $E^{(l)}(\cdots)$ denotes the expectation operator defined by the distribution (4.5) associated with $\mathbf{X}^{(l)}$. Note that when the model is rewritten, leaning on its obvious linear invariances, with

$$\mathbf{Y} \to \mathbf{Y}^* = \mathbf{YA} \qquad \text{and} \qquad \mathbf{X} \to \mathbf{X}^* = \mathbf{XB}, \tag{4.10}$$

then

$$\boldsymbol{\phi} \to \boldsymbol{\phi}^* = \mathbf{A}^{-1}\boldsymbol{\phi}\mathbf{B}^{T,-1} \tag{4.11}$$

and

$$\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^* = \mathbf{A}^T\boldsymbol{\theta}\mathbf{B}. \tag{4.12}$$

The inversion relations between (4.11) and (4.12) suggest that the parameter sets $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ should be regarded as duals.

The differential relations between the elements θ_{ij} of $\boldsymbol{\theta}$ and ϕ_{ij} of $\boldsymbol{\phi}$ may be expressed as follows. Using the fact that f in (4.7) integrates to unity, it follows that

$$d\alpha(\boldsymbol{\phi}, \mathbf{X}^{(l)}) + \sum_{u=1}^{q}\sum_{v=1}^{p} E^{(l)}(Y_u^{(l)}X_v^{(l)})\, d\phi_{uv} = 0. \tag{4.12}$$

Similarly, denoting by $\theta_{ij}^{(l)}$ the elements of the general term in (4.9), it follows that

$$\begin{aligned} d\theta_{ij}^{(l)} &= \sum_{u=1}^{q}\sum_{v=1}^{p} E^{(l)}(Y_u^{(l)}X_v^{(l)},\ Y_i^{(l)}X_j^{(l)})\, d\phi_{uv} \\ &\quad + E^{(l)}(Y_i^{(l)}X_i^{(l)})\, d\alpha(\boldsymbol{\phi}, \mathbf{X}^{(l)}) \\ &= \sum_{u=1}^{p}\sum_{v=1}^{q} C^{(l)}(Y_u^{(l)}X_v^{(l)},\ Y_i^{(l)}X_j^{(l)})\, d\phi_{uv}\,, \end{aligned} \tag{4.13}$$

where $C^{(l)}(\ldots, \ldots)$ is the covariance operator associated with $E^{(l)}$. Summing over l produces finally

$$d\theta_{ij} = \sum_{u=1}^{q}\sum_{v=1}^{p} C(u, v; i, j)\, d\phi_{uv}\,, \tag{4.14}$$

where

$$C(u, v; i, j) = \sum_{l=1}^{n} C^{(l)}(Y_u^{(l)}X_v^{(l)},\ Y_i^{(l)}X_j^{(l)}). \tag{4.15}$$

The coefficients $C(u, v; i, j)$ define a $pq \times pq$ symmetric matrix $\mathbf{C}$ relating pairs (u, v) to pairs (i, j). $\mathbf{C}$ is positive definite or semidefinite, for it defines expected products in the distribution of pairs Y_iX_j defined by choosing one of $\mathbf{X}^{(l)}$ at random with equal probabilities $1/n$ and choosing $\mathbf{Y}$ given $\mathbf{X}$ according to the distribution (4.5). It is easily shown that the $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ parameters have a one-one relationship provided that $\mathbf{C}$ is positive definite, i.e., if $\boldsymbol{\theta}$ denotes a possible matrix of moment parameters for the given $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$,..., $\mathbf{X}^{(n)}$, and if $\mathbf{C}$ is positive definite for all $\boldsymbol{\phi}$, then there is exactly one matrix of $\boldsymbol{\phi}$ parameters which yields a model with these $\boldsymbol{\theta}$ parameters.

Returning now to the log likelihood $L(\phi)$, differentiation of (4.8) gives

$$dL(\phi) = \sum_{l=1}^{n} \left\{ d\alpha(\phi, \mathbf{X}^{(l)}) + \sum_{u=1}^{q} \sum_{v=1}^{p} Y_u^{(l)} X_v^{(l)} \, d\phi_{uv} \right\}$$

$$= \sum_{u=1}^{q} \sum_{v=1}^{p} \{T_{uv} - \theta_{uv}\} \, d\phi_{uv} , \tag{4.16}$$

where

$$T_{uv} = \sum_{l=1}^{n} Y_u^{(l)} X_v^{(l)}. \tag{4.17}$$

The condition that $L(\phi)$ shall have zero first derivatives is thus seen to be that the matrix $\mathbf{T}$ of sample moments (4.17) shall equal the corresponding matrix $\boldsymbol{\theta}$ of moment parameters. From (4.16) it follows that the matrix of second derivatives of $L(\phi)$ is the negative of the covariance matrix $\mathbf{C}$ whose positive definiteness assures the convexity of $L(\phi)$. As remarked above, there is in general (under positive definiteness) at most one ϕ solving the equations $\mathbf{T} = \boldsymbol{\theta}$, and this ϕ obviously maximizes $L(\phi)$. If the equations cannot be solved, then the maximum of $L(\phi)$ occurs on a boundary of ϕ-space or is infinite and maximum likelihood estimation can make no practical sense. Particular examples of infinite $L(\phi)$ are easily constructed, as in multiple regression analysis with more predictors (p) than observations (n), or in contingency table analysis with certain margins unrepresented in the data. Theoretical aspects of uniqueness, both general and specific to particular models, are susceptible of further development.

To summarize, under broadly applicable operating conditions, the maximum likelihood estimator for ϕ is unique and is equivalent to a method of moments in which the sample moments $\mathbf{T}$ are matched to their expectations $\boldsymbol{\theta}$ under the fitted model. From (4.15), it is seen that $\mathbf{C}$ is the covariance matrix of the estimated $\boldsymbol{\theta}$ under their conventional sampling distribution which regards the the $\mathbf{Y}^{(l)}$ as random given $\mathbf{X}^{(l)}$, and from (4.14) it follows that $\mathbf{C}^{-1}$ is a corresponding asymptotic covariance for the estimated ϕ. Alternatively, Bayesians would regard $\mathbf{C}^{-1}$ as an appropriate approximate posterior covariance for ϕ, since $\mathbf{C}$ is the negative of the matrix of second derivatives of $L(\phi)$. Either interpretation of $\mathbf{C}^{-1}$ requires that the data are sufficient to render the estimates precise in the sense that the relations between $\boldsymbol{\theta}$ and ϕ are practically linear over the region of uncertainty about the true ϕ.

The existence of unique maximum likelihood estimators $\hat{\phi}$ for ϕ is one thing, but the usefulness and relevance of fitting by maximum likelihood is more fundamental, assuming the procedure to be mathematically well-defined. Difficulties are of two kinds, numerical and inferential: Whether or not $\hat{\phi}$ can be

successfully calculated from sample data, and whether or not the resulting $\hat{\boldsymbol{\phi}}$ have any meaning. Both questions are serious owing to the potentially large number pq of parameters involved.

First consider the numerical question. Under special circumstances, an explicit formula can be given for $\hat{\boldsymbol{\phi}}$, opening the way for algorithms to compute $\hat{\boldsymbol{\phi}}$ in a finite number of steps. Examples include fitting a single multivariate normal distribution over a space S, where both means and covariances are unknown, and fitting independent but otherwise arbitrary multinomials to the simple margins of a space S in the form of a contingency table. But examples do not include such simple extensions as fitting normal distributions where inverse second moments as well as first moments depend on $X_1, X_2, \ldots, X_p$, nor contingency table fitting where all 2-way joint margins are matched to sample margins. In general, therefore, the equations for $\hat{\boldsymbol{\phi}}$ are implicit, and computing algorithms are iterative in the sense of following a sequence of $\boldsymbol{\phi}$ matrices which is known to converge mathematically to $\hat{\boldsymbol{\phi}}$. The basic iterative algorithm is the quadratic ascent procedure applied to $L(\boldsymbol{\phi})$, which is simply described because the first and second derivatives of $L(\boldsymbol{\phi})$ have the simple forms $\mathbf{T} - \boldsymbol{\theta}$ and $-\mathbf{C}$ computed above. Thus, a single iteration carries $\boldsymbol{\phi} \to \boldsymbol{\phi} + \Delta\boldsymbol{\phi}$, where

$$\Delta\boldsymbol{\phi} = (\mathbf{T} - \boldsymbol{\theta})\mathbf{C}^{-1}, \tag{4.18}$$

with the notation assuming that $\boldsymbol{\phi}$, $\mathbf{T}$ and $\boldsymbol{\theta}$ are written as $1 \times pq$ vectors meshing with the $pq \times pq$ matrix $\mathbf{C}$, and where $\boldsymbol{\theta}$ and $\mathbf{C}$ involve moments computed from the model defined by the current $\boldsymbol{\phi}$. Many other iterative procedures can be devised by following a principle of cyclic fitting, whereby a subset of the parameters $\boldsymbol{\phi}$ is adjusted to obtain matches of the corresponding elements of $\mathbf{T}$ and $\boldsymbol{\theta}$, then other subsets covering all of $\boldsymbol{\phi}$ are adjusted, and so around the subsets in order until convergence obtains. Each step increases the likelihood because it solves a restricted maximum likelihood problem, and the process converges only when $\mathbf{T} = \boldsymbol{\theta}$, so that no more adjustments are required, and the maximum of $L(\boldsymbol{\phi})$ is achieved. Cyclic fitting procedures are popular in contingency table analysis when the individual steps are themselves fitting procedures of the special kind which can be made in closed form without iterations [12, 24]. A disadvantage of cyclic fitting is that it does not, in general, provide the asymptotic covariance matrix $\mathbf{C}^{-1}$ which is a byproduct of the straightforward quadratic ascent technique.

The conceptual simplicity of the standard iterative approach should not be allowed to conceal the prodigious computational efforts which may be required to carry it out. Three troublesome elements will be sketched. First, it is necessary to make a pass through the data, or at least through summary statistics for each distinct $\mathbf{X}^{(i)}$, on each iteration. If $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}$ are all distinct and large in number, expensive data manipulation in and out of core memory may be

required. Note that $\mathbf{T}$ is sufficient for $\boldsymbol{\phi}$ in the conventional language of mathematical statistics, because $\mathbf{T}$ does include all the dependence on the random free variables $\mathbf{Y}^{(l)}$, but $L(\boldsymbol{\phi})$ also depends on the fixed variables $\mathbf{X}^{(l)}$ and the latter dependence is an important factor in the fitting procedure. Second, at each step one needs the normalizing factors $\alpha(\boldsymbol{\phi}, \mathbf{X}^{(l)})$ as in (4.8), the first moments $\boldsymbol{\theta}^{(l)}$ as in (4.9) and the second moments $\mathbf{C}^{(l)}$ as in (4.13). These integrals are obtainable analytically in the important special case of normal distributions, but hardly ever otherwise. In general, numerical integration over S is required. In the present state of the art, to require these numerical integrals for each $\mathbf{X}^{(l)}$ is prohibitive unless S is discrete and most of its probability is carried on relatively few points. Powerful numerical techniques will be required to surmount this difficulty. Some marginal succor may be obtained by not recomputing $\mathbf{C}$ at each integration, hoping that it changes sufficiently slowly to have slight effect. The third difficulty is the sheer size of the covariance matrix $\mathbf{C}$, for example, 1000×1000 if $q = 20$ and $p = 50$. Inversion of such a matrix is a feat in itself, letting alone repetitions on successive iterations. A simplifying approximation at this stage would be to replace $C(u, v; i, j)$ by

$$\tilde{C}(u, v; i, j) = \left\{ \sum_{l=1}^{n} C^{(l)}(Y_u^{(l)}, Y_i^{(l)}) \right\} \times \left\{ \sum_{l=1}^{n} X_v^{(l)} X_j^{(l)} \right\} \div n, \tag{4.19}$$

which can be inverted by separate inversions of the two smaller matrices in curly brackets. The factorization (4.19) can hold for the actual $\mathbf{C}$, but only in rare cases, as for example, when $Y_1, Y_2, \ldots, Y_q$ have a multivariate normal distribution whose mean depends linearly on $X_1, X_2, \ldots, X_p$ but whose covariance is constant.

Inferential considerations may come to the aid of numerical difficulties by advising the data analyst to cut down on the size of the parameter set to be fitted. The principle of parsimony in modeling is widely ignored when fitting is computationally easy. For example, an attempt to estimate a 50×50 correlation matrix from a sample of size 20 would produce more point estimates than original data points, and it makes little sense to regard such estimates as plausible surrogates for population values. Thus the problem of many parameters is more than a computational problem, and should be taken seriously in all situations when ambitious model-fitting is undertaken. Moreover, the problem is conceptually thorny, because it is not easily formulated within current theories of inference, and solutions to the problem are therefore tentative. Further

discussion of general aspects is deferred to Section 5.1, preceded here by some specific suggestions for exponential models.

The simplest approach to parameter reduction is selection, meaning in the present context fitting only a subset of the pq parameters ϕ while setting the rest to zero, or sometimes to another *a priori* null value. Selection of variables for multiple regression analysis is a widely used technique, with many variants, and suggests that parameter selection procedures will be useful tools for a broad class of models. In the case of parameters ϕ, one can omit rows, columns, or single elements. Serviceable techniques will need to be developed empirically from the many possibilities.

The foregoing techniques presuppose simplicity of theory in that plausible null values are available, and simplicity of data in that the data appear conformable with many of these null values. Alternative techniques may be proposed which hypothesize special structure on ϕ. For example, in [41] it is supposed that the columns of ϕ are almost identical in the case of a multinomial logit model. Another structure which could sometimes be plausible is the additive structure $\phi_{ij} = \alpha_i + \beta_j$. A more complicated approach would be to restrict ϕ to have rank $r < \min(p, q)$. The case $r = 1$ would then imply multiplicative structure for the ϕ_{ij}, and the model could be successively broadened by stepping up r. These procedures generalize canonical correlation analysis, in the sense that fitting a normal model in which the matrix of regression coefficients of Y_1, Y_2,..., Y_q on X_1, X_2,..., X_p is required to have rank r is naturally done using the canonical variables associated with the largest r canonical correlations as carriers of all the fitted correlation. The general procedure is more difficult to understand mathematically because it does not reduce to a standard eigenvalue calculation unless (4.19) holds, as it does in the canonical correlation case mentioned above. However, a back and forth type iterative fitting algorithm is easily defined, and appears to converge satisfactorily in the few cases tried by the author.

Rather than directly cutting down the parameter space as above, one can treat all of the parameters as meaningful variables but regard them as random with a distribution depending on a relatively small set of secondary parameters. Random effects models in the analysis of variance illustrate this approach. Or the approach may be regarded as an empirical Bayes approach where prior distributions are fitted from data. The application of these ideas to exponential models has been mainly restricted to simple considerations of many means or many binomial probabilities, but rapid progress and increasingly wide acceptance of these techniques appears inevitable, and indeed necessary if the model-fitting side of multivariate data analysis is to progress.

A final word of caution, searching in data for suitable restrictions on ϕ or for suitable randomness hypotheses for ϕ puts even more load on the difficult

task of computational fitting. Additional difficulty is posed by missing observations (cf., Section 7), which must always be faced in real life where the use of data analytic tools is preceded by a data-cleaning operation.

5. Problems of Interpretation

5.1. *Many parameters.* Many tried and tested techniques of multivariate data analysis were invented at a time when 10 was a typical number of variables in an ambitious data collection program. Now, with automation and expanded support for scientific investigations, data sets having 100 or even 1000 variables are not uncommon. Multivariate data analysts need therefore to cultivate increasingly the habit of asking whether their data will bear the weight of their methods. The question reduces to asking whether fitted parameters are meaningful or, conversely, whether the numerical processes which produce them are not empty exercises. Sometimes evidence can be adduced after the fact by recognizing, for example, substantive meaning in clusters, factors or scales, or by successfully using a fitted model for prediction. Still, a crucial problem for theoretical statistics is to assess the evidence internally during the course of data analysis, and to alter that course where necessary so that the outputs of analysis have high signal-to-noise ratio.

Certain monotonicity relations are nearly self-evident. The potential yield of reproducible model structure from a given data set depends directly on the sharpness of differences and strength of relations in the phenomenon underlying the data. Also, the more are the variables acquired, the more is the danger that interesting differences will be undetectable. And the larger the samples of units observed, the better one is able to detect structure. For example, it is easier to detect one salient effect if 10 standard solutions are assayed once than if 100 standard solutions are assayed once; but in either case the task becomes easier if the assays are repeated.

In terms of statistical inference, the problem with many parameters is that only a relatively small subset of the parameters, or a small set of functions of the parameters, carry important messages of substantive interest, while the remaining parameters become nuisance parameters which obscure the desired information through the unknownness of their values. The problem is especially severe when the parameters of interest are not known in advance but must be identified by scanning the data.

The conceptual tools for approaching many parameters are not yet well formed, but two basic themes can be identified. The first of these is the need for measures of the confusion affecting questions of interest which is due to the uncertainty about many parameters. The decision-theoretic approach to inference evaluates

procedures in terms of operating characteristics, and there is confusion about these operating characteristics deriving from their dependence on unknown parameters. Admissibility theory generally restricts the class to Bayes rules. Openly Bayesian approaches can go further by assessing posterior probabilities for statements about parameters, but still there is confusion due to the vagueness of initial probability assessments. Approaches to inference which provide upper and lower bounds on posterior probabilities have built-in measures of confusion in the differences between the bounds. Further and more direct attempts should be made to define indices of confusion for data drawn from standard models, even if the indices are rather crude and tenuous in their relation to ideal theories of statistical inference.

The second theme is the need for guidelines in the use of confusion indices when they are available. If confusion is large, then one should consider simplifying a model, but this can be done only at the cost of making the model less realistic. At present, the tradeoff between realism and clarity of message can only be made intuitively. The author has discussed elsewhere [9] the possibility of making the tradeoff more formal.

In the case of multivariate techniques, the decision to fit large numbers of parameters is often made casually, not because adequate fit requires them, but because they arise naturally in the mathematical formulation of the model and can be fitted with relative ease. Parameter reduction techniques such as those sketched in Section 4 would then appear to offer clear prospects for more illuminating data analyses, especially when a decrease in confusion can be bought at little apparent cost in realism.

5.2. *Causality.* A major objective behind much multivariate data collection is to provide support for hypotheses of cause and to measure the strength of causal relations. In consequence, it is inadequate to avoid the persistent question: Can statistical data ever be used to demonstrate cause? The answer would appear to be yes, but with careful qualifications.

One objection to statistical data can be quickly dismissed. This is the argument which rules out evidence based on statistical data because numbers cannot specify any mechanism to explain the cause and effect relation and, in particular, can specify no deterministic mechanism. A carefully controlled and randomized experiment in which, say, 20 out of 20 drug treated animals survived, while only 2 out of 20 placebo treated animals did so, would generally carry with it a strong implication of causal connection between treatment and survival, even assuming that the biochemical action is a mystery. Further detailed knowledge of the mechanism provides a greater degree of understanding about the circumstances under which causation is operating, but there is no discrete jump to a different and more satisfying concept of cause. Statistical data and statistical

evidence do not differ in kind from empirical data and evidence generally. As argued below, *any* empirical data can have only an indirect bearing on the establishment of a hypothesized causal mechanism. Nor should one require causal explanation to be completely deterministic. Whether or not all scientific explanation can ultimately be reduced to deterministic models may be an open question for philosophical disputation. But in the real world, residual uncertainty can never be eliminated entirely and often must be faced probabilistically, especially with biological and social phenomena.

Necessary to any assignment of cause is an *a priori* judgment that an explanatory (perhaps probabilistic) mechanism could plausibly exist. Otherwise, evidence of association in statistical data has no direct bearing on the presence or absence of causal relations. The train of logic is that a causal hypothesis is recognized to imply certain hypotheses of association or increased variability, or other observable manifestation. Then the latter type of hypotheses can be tested on data, and the failure of such tests to reject the data patterns implicit in the causal hypothesis provides negative support for the hypothesis. Negative support is unfortunately the best available, but accumulation of such support from many data sets eventually creates a sense of confidence akin to positive support. This is surely as close to proof as is logically possible with empirical phenomena.

So much for philosophy. The practical difficulties in establishing cause are even thornier than the philosophical difficulties. One must sort out competing causal hypotheses and find paths through complex patterns of multiple causes, and through hierarchical systems under which, for example, factor 1 may influence factor 2 and both may subsequently influence factor 3. The principles of experimental design can of course help with the problem of multiple causes, making some implausible by randomization and disentangling others by producing orthogonal design vectors. See [7] for some difficult aspects of design in relation to causal analysis.

The method of path analysis, due originally to Sewall Wright, has been much used in recent years by social scientists to help sort out complex patterns. The underlying idea is to create restricted linear models in which terms are inserted only where plausible causal hypotheses exist. The pattern of observed correlations is then tested, mainly by eye, for conformity with the pattern implied by the restricted linear model. Some basic references are [11, 38, 47, 48, 49], some recent references are [3, 10, 26, 40], and a bibliography [23] is available. The simultaneous linear equation methods of economics [16] are close relatives of path analysis. With a few exceptions [38, 44–46] statisticians have not entered the thickets of causal linear models. The need for supportive techniques is clear, however, and opportunities to sharpen a promising tool should not be neglected.

6. Computation

If data analysis is the substantive core of the science of statistics, then the main technological burden of doing statistics falls on the computers and computing systems which carry out data analyses. The validity of these assertions is supported by the rapid growth of the field of statistical computation. Computers are also used in statistics as tools for exploring mathematical theories, and as teaching aids to provide graphic demonstrations of theory and methods and to allow students to experience directly the practice of data analysis. The interactions among these various uses serve to enhance the richness and prospects for growth and development in the area of statistical computation.

Modern large computers operate in environments of extraordinary complexity both in terms of hardware and of time-sharing operating systems. Moreover, the technology continues to evolve rapidly, both through the trial and error process of constructing systems and through modifications of physical devices. Communication with the machine becomes ever more convenient as languages and systems of program modules are improved. Devices for input and output become more varied and faster, and memory devices become larger and less expensive. All of these factors imply that the data analyst can be in much greater contact with his data, can look at many more aspects of it quickly, and can interact with much more flexibility than was the case a few years ago.

To some, the computer analysis of data, especially data about people, has become a sinister tool, and it is indeed true that problems of privacy are real and challenging. But they are primarily moral, social, and legal problems. For data analysts, the problem is to control computers so that the results of data analysis are purposeful and informative, especially in relation to their cost.

Figure 2 is an attempt to show how an idealized data analyst user might interact with a computing system, given a system designed with a high level of interaction in mind. There is a definite separation of data structures and program structures in memory, where both can have a complex logical structure, and likewise there is a definite separation of programs into those which call, organize, and store data and those which carry through algorithmic operations of data analysis to produce numbers which carry a message. Mathematical statisticians, beginning most notably with the great generalist Gauss and his algorithm for least squares, have long been major contributors to data analysis algorithms. The practice of statistical data analysis has forced statisticians to be concerned with finding good methods for solving linear equations, finding eigenvalues and eigenvectors, maximizing functions, and so on, which in turn means keeping up with developments in numerical analysis, especially numerical linear algebra. Of central concern to data analysts is the design of building blocks which can be put together in many ways to produce higher level analyses with many possible

variations. Here it is vital that data analysts take a leading role, so that the system will have suitable levels of variety, flexibility and power relative to the desired applications. The problem of numerical accuracy of data analysis techniques calls for more cooperation between statisticians and numerical analysts. Recent papers such as [17, 42] show encouraging evidence of dialogue. Clearly, the study of accumulated round-off error, and thence of how to reduce it, merits serious ongoing treatment both empirical and theoretical by data analysts and statisticians. Likewise, increased cooperation between computer scientists and statisticians on the information processing side of statistical computing is greatly to be sought.

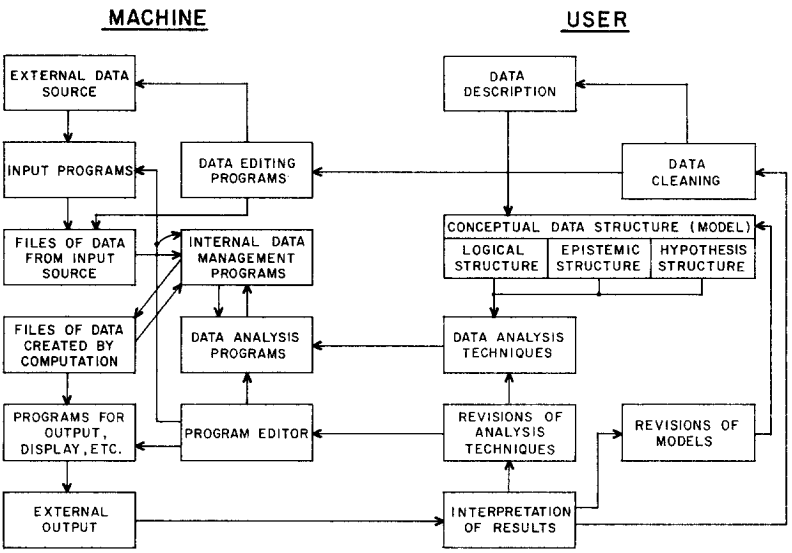

FIG. 2. Schematic diagram of the interaction between the computing machine and the data analyst user. Note that system functions of the machine (i.e., control of flow, error diagnostics, wrappers for programs and data structures, etc.) are not explicitly represented.

For more insight into the breadth and complexity of statistical computation, see [32, 34].

7. Data Cleaning and Missing Values

After a data set is stored in a machine readable form, such as punched cards or magnetic tape, and before any main body of analysis is performed, it is

generally necessary to perform a screening analysis to detect impossible or implausible values and to decide what device or combination of devices is to be used to cope with the complications due to missing values. These processes are often more time-consuming and agonizing than the main body of data analysis. They are closely related, because a detected wild value may become a missing value, and because both processes can involve assessing a plausible range of values for a potential wild value in one case and for a missing value in the other case.

Errors enter data during the measuring, recording and transmitting phases of data collection. For the most part, the detection of errors is beyond the power of the data analyst, excepting of course data from experiments specifically designed to assess biases and error distributions in measurement processes. It is necessary to believe at some point that the data are of adequate quality to make analysis meaningful, and to proceed while admitting that all variables are not exactly what they purport to be. After all, none of the tools in the data analyst's arsenal is completely firm, whether they be models, or inference procedures, or numerical computations, and one must rely on intuition and experience to judge whether or not the whole ship is tight enough to carry the analysis.

The term data cleaning refers mainly to the search for outlying values or values otherwise known to be impossible. The standard approach is to look at sample distributions of variables, from which extreme values can be picked up directly, or to look at simple moments such as fourth moments about the mean which can be highly inflated relative to the square of the sample variance. In the presence of correlated response variables, it is possible to predict the values of a variable from its correlates and thence to pick up extreme deviates not only from single distributions but also from distributions of prediction errors. For example, if Y is highly correlated with X_1, X_2, X_3, then a value of Y which is far from its predicted value, given X_1, X_2, X_3, may be suspicious even though that value of Y is not at all extreme relative to the marginal distribution of Y.

Extreme observations pose questions of competing causes which are beyond the scope of the data itself to answer. Sometimes it is possible to retrace steps through the collection process and to obtain corroborating evidence of error which may in fact be overwhelming. In other cases, searching may be impossible or no evidence of irregularity may be found. Then the degree of extremity must be balanced against the possibility of a genuine measurement which happens to be very unusual. *In extremis*, the value may be expunged and treated as missing, but a safer course may be to create an extra parameter for the extreme value, and to admit that the interpretation may be an erroneous wild value or may be a genuine drawing from a long-tailed distribution. Either way, there is likely to be much vagueness about the parameter associated with the outlier, and the

main emphasis should be on modeling and analysis which prevent this vagueness from contaminating the answers to questions on which the data have much to offer. Univariate examples are the trimmed and Winsorized means methods popularized by Tukey [39]. In principle, any analysis technique based on a probabilistic model could have an attachment to fit parameters to extreme observations of any value which can be predicted by the model, assuming that the model appears to fit apart from a modest number of such suspicious values. The development of such attachments for multivariate methods is a worthy objective, but virtually unattempted so far.

Missing values cause difficulties whose severity depends on many factors such as the number of missing values, their pattern, including whether they appear in fixed or response variables, and the degree of correlation between the indicator of missingness on a variable and other variables. If only a few values are missing, it may be possible to show that any set of plausible inserts will produce almost the same final analysis, so that any *ad hoc* procedure of filling in missing values may be used. With increasing numbers of missing values, it becomes possible in principle to assess the correlation patterns between missingness indicators and other variables, and even to include missingness in a model to be fitted. The potential here is great, as are the difficulties posed by the need to relate a large number of indicator variables to a data structure which may be rich in detail quite apart from missing values.

Some of the more common practical expedients are as follows. Most of these implicitly accept the randomness of missing values. In the case of response variables, maximum-likelihood model-fitting can be carried out whether or not complete data vectors are available; but it is computationally difficult to do so. Still, there is considerable scope for further development of maximum likelihood techniques, or at least of reasonable approximations to them. Another general approach is to fill in missing values with good estimates and then proceed to analyze as though the estimates were actual observations. In the case of missing response variables, this may often be reasonably close to the maximum likelihood approach, but in the case of missing fixed variables it usually means filling in the mean value of the variable, which may be a poor value and which in any case trades heavily on the randomness of the missing values. Still another approach applicable to missing fixed values is to attempt to estimate various summary statistics which would have been computed from the full data set. For example, in multiple regression analysis, each of the covariances entering the normal equations can be estimated by summing over those units where both variables are recorded, using different sets of units for different covariances. Both of the techniques just described for fixed values attempt to recoup the advantages of analysis techniques which are well-defined when data sets are not complicated by missingness. In particular, the former method of estimating each missing

value is able to take advantage of simplifications of analysis from balance which would have characterized the complete data set (cf., Section 2.1). See [1] for a review and study of certain missing value techniques depending on the assumption of randomness.

An alternative tack to missing values is to drop from the analysis any unit for which a complete set of variables is not measured, and to proceed with the reduced data set as though it were complete. This approach is close in spirit to regarding missing values as kin to outliers, and presumably offers some protection against the possibility that missingness correlates with extreme values of the missing variables. That is, relations are studied in the more central regions of the distributions. In practice, reduction to completely observed units is feasible only if the number of such units is adequate. A variant is to select important variables and to require units to be complete on the selected variables.

Finally, a different type of missingness deserves mention. Instead of identified units with only partial information, there may be no units having particular interesting values on some variable. The absence of these units may result from missing values on the variables concerned, but not necessarily. In the case of response variables which are continuous or nearly so, missing levels or sparsely observed levels are the norm, and the solution is typically to fit a smooth density to repeated observations.

In the case of fixed variables, missing levels may be a matter of design, as when a certain drug may not have been tried, or perhaps a certain dose level of a drug may not have been selected. The effects of missing levels on a fixed variable must be assessed either by extrapolating a smooth relation, as by fitting a dose response curve, or by making an assumption of randomness, as might hold if it was assumed that the effect of an untried drug would be as though the drug was exchangeable with the drugs tried in the experiment. The former involves curve-fitting; the latter involves random effects models. These devices are part of the bread and butter of data analysis.

8. Reprise

Rapid progress is both probable and vital in a pair of directions.

First, exploratory and especially graphical techniques should be pushed and made computationally more convenient, so that the data analyst may see the the main outlines of his data and be able to explore for relations in a broad catalog of known types. More approaches in the spirit of [13–15] should be tried, especially in the traditional area of searching for relations which may be used for prediction. The empirical development of structure-seeking techniques such as clustering and scaling has proceeded rapidly, as if filling a vacuum, and

standards for empirically judging these techniques may be expected to rise. The standards for relation-seeking analyses are already quite high, because of a long history of development of least squares, analysis of variance, multiple discriminant analysis, and contingency table methods, so that the would-be provider of new insights must grapple with a large body of available methods and exercise nontrivial ingenuity.

Second, theoretical statisticians need to develop models and associated supportive reasoning techniques which will help the data analyst to separate the meaningful from chance artifacts, and which will make it possible to quantitatively assess meaningful features of data. The task is especially difficult in the area of structure-seeking techniques, where progress has been slow, due to the intricacies of having to search through complicated parameter spaces. As indicated above, however, even in the more traditional relation–seeking analyses, parameter spaces quickly become large. Detailed studies of the effect of prior information will greatly clarify the study of causal effects. Finally, the science of statistical inference itself should grow and shift in emphasis as it addresses itself to the tasks of data analysis.

In short, the twin supports of data analysis, in computational and manipulative technique, on the one hand, and in statistical reasoning to secure reliable knowledge on the other, have clear prospects for growth into a healthy balanced science.

References

[1] Affifi, A. A. and Elashoff, R. M. (1966, 1967, 1969). Missing observations in multivariate statistics. I. Review of the literature. II. Point estimation in simple linear regression, III. Large sample analysis of simple linear regression, IV. A note on simple linear regression. *J. Amer. Statist. Assoc.* **61** 595–605; **62** 10–29; **64** 337–358; 359–365.

[2] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[3] Blalock, H. M. (1967). Path coefficients versus regression coefficients. *Amer. J. Sociol.* **72** 675–676.

[4] Bock, R. D. (1969). Estimating multinomial response relation. In *Contributions to Statistics and Probability: Essays in Memory of S. N. Roy* (R. C. Bose, Ed.). University of North Carolina Press, Chapel Hill.

[5] Carroll, J. D. and Chang, Jih-Jie (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika* **35** 283–319.

[6] Cattell, R. B. (1965). Factor analysis: An introduction to essentials. I. The purpose and underlying models. II. The role of factor analysis in research. *Biometrics* **21** 190–215, 405–435.

[7] Cochran, W. G. (1965). The planning of observational studies in human populations. *J. Roy. Statist. Soc. Ser. A* **128** 234–65.

[8] DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

[9] DEMPSTER, A. P. (1970). Model searching and estimation in the logic of inference. In *Proceedings of the International Symposium on Statistical Inference* (held in Waterloo, Ontario, April 1970). Holt, Rinehart and Winston, to appear.

[10] DUNCAN, O. D. (1966). Path analysis: sociological examples. *Amer. J. Sociology* **72** 1–16.

[11] ENGELHARDT, MAX D. (1936). The technique of path coefficients. *Psychometrika* **1** 287–293.

[12] FIENBERG, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41** 907–917.

[13] GENTLEMAN, W. MORVEN, GILBERT, JOHN P. AND TUKEY, J. W. (1969). The smear-and-sweep analysis. In *The National Halothane Study* (John P. Bunker *et al.*, Eds.), pp. 287–315. U. S. Government Printing Office, Washington, D.C.

[14] GNANADESIKAN, R. AND WILK, M. B. (1969). Data analytic methods in multivariate statistical analysis. In [28] pp. 593–638.

[15] GNANADESIKAN, R. AND LEE, E. T. (1970). Graphical techniques for internal comparisons amongst equal degree of freedom grouping in multiresponse experiments. *Biometrika* **57** 229–238.

[16] GOLDBERGER, A. S. (1966). *Econometric Theory.* Wiley, New York.

[17] GOLUB, G. H. (1969). Matrix decompositions and statistical calculations. In [32], pp. 365–398.

[18] GOODMAN, LEO A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Assoc.* **65** 226–256.

[19] GOWER, J. C. (1967). Comparison of some methods of cluster analysis. *Biometrics* **23** 623–637.

[20] GOWER, J. C. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.* **18** 54–64.

[21] HARMAN, H. H. (1967). *Modern Factor Analysis*, 2nd. ed. University of Chicago Press, Chicago, Ill.

[22] HARTIGAN, J. A. (1967). Representation of similarity matrices by trees. *J. Amer. Statist. Assoc.* **62** 1140–1158.

[23] HAUCK, W. (1970). A bibliography on causal inference. Research Report CP-1, Department of Statistics, Harvard University, Cambridge, Mass.

[24] IRELAND, C. T. AND KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

[25] JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **32** 241–254.

[26] JÖRESKOG, K. G. (1970). A general method for analysis of covariance structures. *Biometrika* **57** 239–252.

[27] KRISHNAIAH, P. R. (Ed.) (1966). *Multivariate Analysis. Proceedings of an International Symposium held in Dayton, Ohio, June* 14–19, 1965. Academic Press, New York.

[28] KRISHNAIAH, P. R. (Ed.) (1969). *Multivariate Analysis-II.* Academic Press, New York.

[29] KRUSKAL, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika* **29** 1–27.

[30] KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29** 115–129.

[31] KRUSKAL, J. B. AND CARROLL, J. D. (1969). Geometrical models and badness-of-fit functions. In [28], pp. 639–671.

[32] MILTON, R. C. AND NELDER, J. A. (Ed.) (1969). *Statistical Computation.* Academic Press, New York.
[33] MORRISON, D. F. (1967). *Multivariate Statistical Methods.* McGraw-Hill, New York.
[34] MULLER, M. E. (1970). Computers as an instrument for data analysis. *Technometrics* **12** 259–294.
[35] SHEPARD, R. N. (1962). The analysis of proximities: multinomial scaling with an unknown distance function I, II. *Psychometrika* **27** 125–140, 219–246.
[36] SHEPARD, R. N. AND CARROLL, J. D. (1966). Parametric representation of nonlinear data structures. In [27], pp. 561–592.
[37] SOKAL, R. R. AND SNEATH, P. H. A. (1963). *Principles of Numerical Taxonomy.* Freeman, San Francisco.
[38] TUKEY, J. W. (1954). Causation, regression and path analysis. *Statistics and Mathematics in Biology* (O. Kempthorne *et al.*, Ed.), pp. 35–66. Iowa State College Press, Cedar Falls, Iowa.
[39] TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **44** 1–67.
[40] TURNER, M. E. AND STEVENS, C. D. (1969). The regression analysis of causal paths. *Biometrics* **15** 236–258.
[41] WALKER, S. H. AND DUNCAN, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54** 167–180.
[42] WAMPLER, R. H. (1970). On the accuracy of least squares computer programs. *J. Amer. Statist. Assoc.* **65** 549–565.
[43] WISHART, D. M. G. (1969). An algorithm for hierarchical classification. *Biometrics* **25** 165–170.
[44] WOLD, H. (1956). Causal inference for observational data: A review of ends and means. *J. Roy. Statist. Soc. Ser. A* **119** 28–61.
[45] WOLD, H. (1959). Ends and means in econometric model-building. In *Probability and Statistics* (U. Grenander, Ed.). Wiley, New York.
[46] WOLD, H. (1964). Forecasting by the chain principle. In *Econometric Model Building: Essays on the Causal Chain Approach.* Contributions to Econometric Analysis, No. 36, pp. 5–36, North-Holland, Amsterdam.
[47] WRIGHT, SEWALL (1918). On the nature of size factors. *Genetics* **3** 367–374.
[48] WRIGHT, SEWALL (1954). The interpretation of multivariate systems. In *Statistics and Mathematics in Biology* (O. Kempthorne *et al.*, Ed.), pp. 11–34. Iowa State College Press, Cedar Falls, Iowa.
[49] WRIGHT, SEWALL (1960). Path coefficients or path regressions: alternative or complementary concepts. *Biometrics* **16** 189–202.

Printed in Belgium