Jour. Ind. Soc. Ag. Statistics
43 (2), 1991 : 107—126

Missing Observations in Multivariate Analysis

B.R. Murty[1]
and
Walter T. Federer[2]
(Received : December, 1986)

Summary

An examination is made of the literature on missing observations or missing vectors of observations in multivariate situations. Missing values can occur at random, by design, or by impossibility. Most work to date has been for randomly missing values, with a concentration on computing values for the missing observations. The problems encountered in the three situations are discussed with regard to testing and estimation. Many multivariate procedures become questionable in the presence of missing values. It is concluded that a considerable amount of theoretical work is required before the problems can be resolved.

Key Words : Missing observations, Multivariate Normality, EM Algorithm, Estimation, Testing

Introduction

Multivariate problems in the real world are many as data are, in general, multivariate in nature. As pointed out recently by Gnanadesikan and Kettenring [11] most of the theoretical work is directed towards the distribution theory and inference procedure, formal mathematical proofs, extension of the univariate situations by analogy and with simplified assumptions like multivariate normality. There is a large gap between users' needs and the available statistical tools, as will be illustrated later in this paper. However, the useful applications of multivariate analysis in the recent years in biological and social sciences have given a better insight into the interrelationships between the observed variables in many situations and has helped in a meaningful interpretation of the results and planning of the strategies to be adopted. In the applications of multivariate analysis, as in biology, common problems like missing data, mixtures of distributions, and absence of information on the underlying distribution have been frustrating.

[1] Int'l. Atomic Energy Agency, United Nations, (IAEA), Vienna, Austria
Present Address : INSA Sr. Scientist, Biochemistry Div., IARI, New Delhi-12
[2] Biometrics Unit, Cornell University, Ithaca, N.Y., USA.

Hence, there is a need to assess the methods themselves in order to develop suitable modifications of the available techniques to meet the above difficult but frequent situations in the real world. On the other hand, situations are known where improper use of multivariate analysis in the hands of users with inadequate appreciation of the complexity of interpretation has led to unjustified conclusions. This is essential to avoid possible uncritical use of computer packages, unjustified conclusions, and resultant consequences. Misuse has resulted when principal component analysis, canonical analysis and factor analysis were treated as if they were interchangeable. While conceptually, the basis and computational procedures for the three are quite different as follows :

Principal Component Analysis : Let the matrix of coefficients applied to vector x be A, and Σ is the variance covariance matrix of the x's. Then $AA' = I$ and its variance covariance matrix of transformed variates which are linear combinations is $A\Sigma A'$. The eigenvalues of the variance covariance matrix are obtained by solving the differential equation $| \Sigma - \lambda I | = 0$. The linear combinations are orthogonal and maximizing the variance we can get out of the p variables.

Canonical Axioms : Of the p-dimensional sample space of the *h* universe

$$\underset{(N\times p)}{X} = \underset{(N\times h)}{Z'} \quad \underset{(h\times p)}{B} + \underset{(N\times p)}{E} \qquad \text{where } N = N_1 + N_2 + \ldots + N_h$$

and the variance covariance matrix of each of the p-variate sample errors is Σ_Ω which is estimated by the relation

$$\hat{\Sigma} = (N - h)^{-1} \left\{ \sum_{l=1}^{h} (N_l - 1)\, \hat{\Sigma}^{(l)} \right\}$$

Factor Analysis : The model is

$$\underset{(p\times 1)}{X} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{\Gamma} \quad \underset{(m\times 1)}{f} + \underset{(p\times 1)}{\omega}$$

where x is $N(\mu, \Gamma\Gamma' + \Delta)$, f is $N(O,I)$, ω is $N(O,\Delta)$ and is independent of f and Δ is a pxp diagonal matrix with non-negative elements, and the matrix $\Gamma\Gamma'$ is of rank $m < p$. Thus, principal component analysis maximizes the variance obtainable in the linear combinations which are orthogonal to each other. In discriminant analysis, the linear combinations are so chosen to maximize the ratio of treatment SS and (Treatment + Error SS). In canonical analysis, we maximize the correlation between the linear combinations of Y and linear combinations of X and the canonical vectors are mutually orthogonal. In factor analysis, it is condensation and possible deletion of factors to ensure simplifying and reducing in the dimensionality in terms of some meaningful *super-variables.*

In the discriminatory analysis, the biologist is not clear sometimes even about the null hypothesis. Sometimes the maximum discriminating ability may not be meaningful, as e.g., in intercropping experiments where a ratio of one kilogram of maize to 38 kilograms of bean was the estimate from maximum discrimination. This is unrealistic as ratios of 1:3 to 1:7 were the interpretable ones in practice. Also, the particular value of 38 depended entirely upon which cultivars were included in the experiments by Federer and Wijesinha [9].

Among the multivariate methods which have been most commonly used in eight fields of investigation (based on a survey of nearly 18,000 reports), factor analysis accounted for 55%, discriminant analysis 13%, principal component analysis 11%, and multidimensional scaling 8% while other methods are 45% in psychology, 28% in education and 6% in sociology, medicine and technology (see Gnanadesikan and Kettenring, [11]).

While the objectives of multivariate analysis in *real world applications* are :

a) reduction in dimensionality

b) increasing the sensitivity of the analysis by analyzing the intercorrelation between variable not possible in individual univariate analyses but which can be clearly brought out in multivariate analysis.

c) exploring the underlying structure of the data for functional relationships among the variables, and

d) classification - discriminant analysis (prespecified groups as different species), clustering problems (groups obtained by data analysis),

the objectives of the work in progress on the theoretical aspects are different (see Gnanadesikan and Kettenring, [11], Kariya, Krishnaiah, and Rao, [17], as seen from the table below :

Objectives in real world application of multivariate analysis (examples in biology)	*Objectives in motivating the work on theory of multivariate analysis.*
1. Reduction in dimensionality using principal components, cluster analysis, discriminatory analyses, (Ex., evolution in biological populations, Murty [24]).	1. Evolving probabilistic models and methods by anology with univariate analysis (tests for departure from normality; missing data).
2. Increase the sensitivity of analysis by analyzing the intercorrelations among the response variables - which is not possible in separate univariate analyses but clear in multivariate analysis. (Ex. genetic diversity between varieties in plant breeding (Murty, et al [25,26,27]).	2. Development of distribution models and theory. (Mostly on multivariate normal, recently a little on elliptic distributions nonparametric approaches).

3. Explore the structure underlying the data on functional relationship among the variables. (Ex., a. Cytoplasmic differentiation in terms of protein synthesis and photosynthetic apparatus; Bhakta, [4], Thakur [40] and b. Cytochemical changes in the evolution of Lathyrus species, Narayan, [28]).

4. Classification :-

(a) Discrimination analysis between prespecified groups. (Ex., Interspecific divergence; chromosome substitution lines in wheat; Murty [24]).

(b) Clustering i.e., groups formed from the analysis of the data. (Ex. anthropometric surveys; Rao [13]; classification of world collections of crops like wheat, maize, sorghum, etc. Murty and Arunachalam [25]).

3. Development of inference procedures and of confidence regions and tests of significance and establishing optimality properties of such techniques (based on oversimplified assumptions).

Difficulties of application of multivariate analysis in the real world

Some of the problems of interpretation of results from multivariate analysis have been, to some extent, due to the uncritical use of available techniques. However, the divergence between the objectives and the order of priorities in the real-world applications and in the theoretical work in multivariate analysis has to be bridged to make a successful impact for wider use, in light of the growing number of research journals. Considering the following four desiderata proposed by Gnanadesikan and Kettenring [11],

(i) usefulness in revealing what is in the data,

(ii) ease of use,

(iii) diagnostic value - the nature of departure of the data from the key assumptions of the model, and

(iv) formal statistical properties like optimality, robustness, etc.,

the importance and priorities are in opposing directions, particularly in the case of missing data beyond the control of the experimenter. In the real world applications (i) is more important than (ii) and (iii) is more important than (iv). In published papers on the theoretical aspects, there is much more emphasis on (iii) and (iv) than on (i) and (ii). Even for (iii), when departure from key model assumptions is evident as in the case of "missing data", enough effort is not made in modifying methods for appropriately handling the data. The large output of publications on aspect (iv), statistical properties, is not linked to the need for usefulness in revealing what is in the data.

Relative utility of commonly used methods

A re-examination of the summary information by Gnanadesikan and Kettenring [11] on the relative usage of different multivariate methods and 21 criteria for comparing these methods, has revealed the inadequacy of four most commonly used multivariate methods to meet the special problem situations commonly met in the real world. The four most commonly used methods are factor analysis, discriminant analysis, principal component analysis, and cluster analysis. The criteria of concern for the most common problems in the multivariate applications are (A) Versatility (utility for several purposes), (B) Easy handling of incomplete observations, (C) Insensitivity to distributional assumptions (example, departure for normality, nonrandom missing values), and (D) robustness against outliers.

From the summary in Table 1, it is evident that the most extensively used methods are not adequate for the special situations particularly B, C, and D, and may lead to incorrect conclusions unless the method is modified. Solutions are needed for handling "missing data" that are beyond the control of the experimenter, i.e., data are not missing at random.

Missing data : The Underlying Principle

Suppose m observations are missing and our model then will be

$$\underset{(N\times 1)}{X} = \begin{bmatrix} \underset{(m\times 1)}{X_1} \\ \underset{(n-\,m\times 1)}{X_2} \end{bmatrix} = \underset{(n\times q)}{Z'}\ \underset{(q\times 1)}{\beta} + e = \begin{bmatrix} \underset{(m\times q)}{Z_1'} \\ \underset{(n-\,m\times q)}{Z_2'} \end{bmatrix} \beta + e$$

Assuming that the normal equation matrix ZZ' has a simple form and can be easily inverted, we estimate β from the N–m actual observations where

$$\hat{\beta} = (Z_2 Z_2')^{-1} Z_2 X_2 = (Z Z')^{-1} \left[I_q - Z_1 Z_1' (Z Z')^{-1} \right] Z_2 X_2$$

$$= (Z Z')^{-1} \left[I_q + Z_1 \left\{ I_m - Z_1' (Z Z')^{-1} \right\} Z_1' (Z Z')^{-1} \right] Z_2 X_2$$

Therefore $\hat{x}_1 = Z_1' \hat{\beta}$ where x_1 are the calculated m dummy observations.

Now $\alpha_\omega^0 = X_2' X_2 - X_2' Z_2' \hat{\beta}$ with N-m-q degrees of freedom. Using these dummy observational estimates, we can show that $\alpha_\omega^0 = X' X - X' Z' \hat{\beta}$. Anderson [1] summarized the situation as follows in the case of missing observations in a simpler manner. Let $X = (Y' Z')'$ where Y has p components and Z has q components, be distributed as $N(\mu, \Sigma)$ where

Table 1. Criteria for comparing four multivariate methods

Multivariate Method	Criteria of comparison			
	Versatility A	Easy Handling of Incomplete Observations B	Insensitivity to Distributional Assumptions C	Robustness Against Outliers D
		(Missing Data)		
1. Factor Analysis	No — Does not possess the characteristic	No	Neutral ?	No — but adjunct available
2. Principal Component Analysis	Yes — Possess the characteristic	No	Neutral ?	No — "
3. Discriminant Analysis	Yes — Possess the characteristic	No	Neutral ?	No — "
4. K — means clustering	No — Does not possess the characteristic	Not at all	Possess the characteristic	No—

$$\mu = \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix}$$

and M observations are more on X and N-M additional observation of Y; the maximum likelihood estimates of μ and Σ can be obtained by expressing the likelihood function in terms of the marginal density of Y and the conditional density of Z given Y, assuming multivariate normality and the missing observations are few and are random.

Nature of missing data

The nature of missing data determines the approach to handling the same. The available methodology is mostly restricted to simple cases when missing values are few and missing at random. The utility of available methods is limited if these two assumptions are not satisfied (see Frane [10]).

Thus, missing values encountered are

(i) Missing at random.

(ii) Missing by design and probably can be estimated (see Srivastava [37] and Federer [7], Studies on intercropping) in some cases.

(iii) Missing as they are unobservable. This category includes grouping, censoring, and truncation as indicated by Dempster, Laird, and Rubin [5].

In the real world, the nonrandom missing values due to design or beyond the control of the experimenter are more a rule than an exception.

The assumption commonly made in handling missing data are :

(i) the data must be missing at random to get a good estimate of the variance covariance matrix,

(ii) each missing variable is highly correlated with one or more available variables,

(iii) the amount of missing data is not excessive, and

(iv) multivariate normality is maintained.

The problem of handling missing data in multivariate normal populations has been studied during the past three decades by the direct application of maximum likelihood for estimation. Testing of hypotheses when data are missing by design was attempted by sequentially combining covariance estimators, starting with complete observations, and adding one group at a time. Iterative methods are used more commonly to replace the missing components under simplified assumptions of multivariate normality and few values missing at random. Further improvement of procedures was made by Orchard and Woodbury [30]. Their procedure

consisted of grouping the observations into classes of identical patterns of missing and observed components, initial estimation of the mean and covariance matrix being done using the likelihood function on the complete vectors. They obtain the conditional expectation of the scores for the mean and the covariance matrix of the missing data given the observed data, then the new estimates of the parameters, and finally correcting the covariance matrix corrected for bias, where

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left[(\hat{Y}_n - \hat{\mu})(\hat{\mu}_n - \hat{\mu})^T + V_n \right] \quad \text{and} \quad \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} Y_n$$

$Y_k = Y_{k,0} + \hat{Y}_{k,m}$ where $Y_{k,0}$ is the observed portion with zero in each position corresponding to the missing component and $\hat{Y}_{k,m}$ is the estimated missing portions, with zero in the position corresponding to an observed component, and V_n is a $p \times p$ matrix for the *nth* observation.

Before discussing the procedures on estimation of missing values and testing of hypothesis and inference, the effect of missing values on multivariate normality assumption and standard test procedures that the multivariate normality assumption is not satisfied, can be examined.

Effect of missing values on multivariate normality assumptions

The present test procedures assume that multivariate normality is not violated in the presence of missing values. This is not correct. In our view, missing values of a nonrandom nature have a considerable effect on multivariate normality assumption, and consequently, on testing of hypotheses of equality of mean vectors and on the variance-covariance matrices. The nonrandom nature of missing data is bound to violate multivariate normality, as can be seen in truncated or censored cases when a large segment of values can be missing. If some variables are unobservable from the point of truncation, it is not even meaningful to estimate the missing values.

Testing for multivariate normality

The effects of anormality on the standard multivariate test procedures are not adequately examined. Testing the reasonableness of the multivariate normality assumption for a given set of data, further complicated by nonrandom missing values, is necessary to transform the data to make them approximately normally distributed and to modify the model assumed and to perform the methods of analysis. There is a need for a "variety of techniques into different sensitivities to different types of departures." Seeking a single best method would not be pragmatic. Measures of multivariate skewness and kurtosis for testing multivariate normality proposed by Malkavich and Afifi [22], are multivariate generalizations

of Fisher's univariate measures of skewness and kurtosis. These multivariate statistics are referred to as max(g) statistics g_1 and g_2, and their asymptotic distributions are derived by Machado [21], who has also provided the transformations of g_1 and g_2 to approximate standard normality. Mardia [23] defined multivariate skewness and kurtosis and formulated some omnibus tests. Thus, only recently, are tests available for multivariate normality and their utility for closeup transformation has yet to be examined.

The effect of multivariate nonnormality makes Wilks' likelihood ratio test criterion invalid. The multivariate normality assumption can be violated either due to (a) original variables being anormal or to (b) one variable being anormal (e.g., yield in maize may follow normality but yield of bean in the intercropping experiments may be anormal in distribution and a linear combination of M + bB may be anormal and no transformation suitable for both variables may be available). To transform either one of the variables, M + b logB will be absurd for interpretation.

When samples are large, effects of violation of multivariate normality assumption may matter little when testing hypothesis about the mean vectors but may be very serious when testing hypotheses about variance-covariance matrices. The power of the test and the significance levels are also affected (see Ito, [15], Eaton and Kovriya [6]). The question of estimation and prediction in a multivariate random effects generalized linear model under moderate departures from normality was examined by Reinsel [32]. He observed that the theoretical mean squared error expression for the random effects predictor remains valid under moderate departures from normality while the general covariance structure method is adversely affected by nonnormality. In such nonnormal cases, it may be possible to obtain predictors with somewhat smaller mean squared error by using more robust procedures. A simulated data study by him showed good agreement between the observed and theoretical results for the random effects method and individual least squares method and, in both cases, the results were not affected by a moderate nonnormality of the errors.

Beale and Little [3] have provided an alternative method of analysis with missing data which gives an estimator that does not assume a multivariate normal population while earlier work of MLE estimator assumes multivariate normality. This method was denoted by them as corrected maximum likelihood or modified Buck's Method. An approximate method of assigning standard errors to regression coefficients estimated from incomplete observations and supported by simulation studies is given by them. However, their methods also assume randomness of the missing data. Their procedure is illustrated below. Let x_{ij} represent the value of the *jth* variable in the *ith* observation and j = 1, . . . , n and i = 1, . . . , N. There are N observations and n variables. Let $A_{jk} = \Sigma_i (x_{ij} - \bar{x}_j) (x_{ik} - \bar{x}_k)$ and $\bar{x}_j = \Sigma_i \; x_{ij} /N$.

Buck's method uses the complete observations to estimate the means of all the variables and the covariance matrix. These values are then used to estimate any missing quantities x_{ij} as linear functions of the variables which are known for this observation. Substitute the estimators for the unknown variables the vector $\bar{x}_j$ and the matrix (a_{ik}) can be built. Let $\hat{x}_{ij}$ be the assumed value of the j_{th} variable in the i_{th} observation. If this value is observed, then $\hat{x}_{ij} = x_{ij}$, otherwise it is a fitted value.

Now

$$a_{jk} = \Sigma_i (\hat{x}_{ij} - \bar{x}_j)(\hat{x}_{ik} - \bar{x}_k) + c_{ijk}$$

and

$$\bar{x}_j = \Sigma_i \hat{x}_{ij} / N$$

The appropriate formula for the correction term c_{ijk} obtained by Beale and Little [3] is

$$c_{ijk} = \hat{v}_{jk} \qquad \text{if } x_{ij} \text{ and } x_{jk} \text{ are both unknown}$$

$$= 0 \text{ otherwise}$$

where v_{jk} denotes the partial covariance of x_j and x_k and u_{jk} denotes the covariance of x_j and x_k. The partial covariance is the covariance of $(x_j - \Sigma_{i \le p} \beta_{ji} x_i)$ and $(x_k - \Sigma_{i \le p} \beta_{ki} x_i)$ were β_{ji} and β_{ki} are the partial regression coefficients defining the best linear approximations to x_j and x_k respectively in terms of the variables known in the first observation. $v_{jk} = 0$ unless $j > p$ and $k > p$. v_{jk} can be estimated by pivoting on the first p diagonal elements of the matrix $(\hat{\omega}_{jk})$ where $\mu_{jk} = \frac{1}{N-2} \sum_{i=2}^{N} (x_{ij} - \tilde{x}_j)(x_{ik} - \tilde{x}_k)$ and $E\hat{v}_{jk} = \frac{N-p-2}{N-2} v_{jk}$. Taking the trial values of $\tilde{x}_j$ and $\hat{u}_{jk}$, and we use them to compute $\hat{x}_{ij}$ and c_{ijk} and hence a_{jk} and $\tilde{x}_j$ and set $\tilde{x}_j = \bar{x}_j$ and $\hat{u}_{jk} = a_{jk}/(N-1)$ and we repeat the process of iteration until no further change on any $\tilde{x}_j$ or u_{jk}.

This analysis does not assume multivariate normality but assumes that the probability of a particular variable being missing is independent of the numerical values of any of the variables for this observation. The overall covariance matrix for all variables is estimated by the corrected maximum likelihood and appropriate submatrices used for regression analysis and estimating the standard errors of the regression coefficients. They also confirm that if the missing variables are highly correlated with known variables, this method may underestimate the precision but reasonably safe. This method is a major improvement over the procedure of Orchard and Woodbury [30], and gives a correction for bias of the estimates of the covariance matrix.

Handling of missing data

In spite of the violations of some crucial underlying assumptions as multivariate normality in missing data situations, several reports on handling missing data particularly for prediction of missing values, are in the literature. The following are the most commonly used methods :

(a) elimination of subjects with any missing data,

(b) computation of "missing value" covariance matrices, and

(c) estimation of missing data.

(d) correction for covariance matrix (already discussed, Beale and Little).

The procedure (a) is possible only when very few values are missing. Even in those cases, there may be considerable loss of information as seen below. Let there be three variables X_1, X_2, and X_3.

Variables	X_1	X_2	X_3
Individuals		0	
or	0		
treatments	0		
			0
		0	
	n_1 observations		
		n_2 observations	
			n_3 observations

Let 0 represent missing values. If we eliminate all individuals with missing data on any one variable, as above, we may be left with no individual even among the minimum group of n_1 under X_1. Thus there would be no meaningful use of the data in A, B, or C.

Examine another situation and consider the following variance-covariance matrix:

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} & \cdots & S_{1n} \\ & S_{22} & & & \\ & & & & \\ & & & S_{kk} & \\ & & & & S_{nn} \end{bmatrix}$$

n_1 observations

n_2 observations

n_3 observations

In the above case, one can have situations where S_{11} is based on all n_1 observations, but S_{12} is based on n_1-2 observations due to two missing values involving variables X_1 and X_2. Thus, in the same rows of sums of squares and sums of products, the values could be based on unequal sets of observations even if the missing values are random. Then, the structure of the variance- covariance matrix becomes very complicated, where each element of the matrix is based on different sets of observations. In such cases, any test of the covariance matrices, and equality of mean vectors is fraught with dangers. The Wilks'test criterion used in several cases of multivariate analysis is not valid. The power of this test and its significance levels are also affected; and for

$$\frac{|E|}{|E+T|} \rightarrow \text{what are the corresponding degree of freedom?}$$

Where E represents error matrix and T represents corresponding matrix for treatments.

Assume, in the data vector $X_1, \ldots, Xp$, one value missing Federer [7] suggested that we reduce l/p degree of freedom for each missing observation. This appears to be empirical but reasonable. In such cases, we change the SS and SP matrix to a uniform level for the use of suitable multipliers as follows. Assume S_{11} is based on the higher number of observations n_{11}, and S_{12} by n_{12} and all values of $n_{ij} \leq n_{11}$:

$$\begin{bmatrix} 1 \times S_{11} & S_{12}\frac{n_{11}}{n_{12}} & S_{13}\frac{n_{11}}{n_{13}} & \cdots & S_{1p}\frac{n_{11}}{n_{1p}} \\ & S_{22}\frac{n_{11}}{n_{22}} & S_{23}\frac{n_{11}}{n_{23}} & & S_{2p}\frac{n_{11}}{n_{2p}} \\ & & S_{33}\frac{n_{11}}{n_{33}} & & \\ & & & S_{ik}\frac{n_{11}}{n_{ik}} & \\ & & & & S_{pp}\frac{n_{11}}{n_{pp}} \end{bmatrix}$$

This will simplify the handling of the matrix for such tests as equality of matrices. An improvement over this method is possible and can be found as the adjusted values of the above matrix are obtained by an arbitrary multiplier.

Missing values are replaced by conditional expectation assuming that the deleted cases do not influence the maximum likelihood estimation of the regression coefficients. Automatic deletion of incomplete cases is not desirable as important

information may be lost and the deleted case can be relatively influential as reflected in appreciable changes in the fitted regression coefficients when it is removed from the data. Shih and Weisberg [35] developed a very useful procedure to detect such influential cases by deriving a one-step influence measure using the EM algorithm and demonstrated its utility with examples. Simon and Simonoff [36] used least squares estimation to provide upper and lower limits for the components of as a function of the non- randomness of the "process which causes the values to be missing". When a large proportion of degrees of freedom is lost due to missing data, higher order regression or principal component estimators may be explored (Basilevsky *et al.*, [2])

Estimation of missing data

In the standard general multivariate linear model (GLMM), the data vectors are all assumed to be complete. As is quite common for some data vectors (observations), one or more values are missing. The available procedures are, under standard assumption :

(i) data are missing at random and not excessive,

(ii) each variable with missing observations is highly correlated with one or more other variables.

Three approaches are currently being followed for estimation of missing values. These are:

(1) Regression approach — Simple linear regression; Stepwise regression cases (Frame, [10])

(a) The regression of the missing variable on all the available variables and,

(b) for each missing value on the available variables (Anderson, [17]). For (b) let data, for example, be denoted by $X=(X_1 , \hat{X}_2)$ where X_1 denotes the observed variables and $\hat{X}_2$ denotes any estimate of the missing data. Then, the Mahalanobis distance D^2 from this case to the mean is

$$D^2 = X' S^{-1} X = X_1' S_{11}^{-1} \lambda_1 + (\hat{X}_2 - \tilde{X}_2)' . (S_{22} - S_{21} S_{11}^{-1} S_{12})^{-1} (\hat{X}_2 - \tilde{X}_2)$$

where $\hat{X}_2$ is the regression estimate of the missing values from method (b). D^2 is minimized when $\hat{X}_2 = \tilde{X}_2$. In the case of (a) $\hat{X}_2$ is close to $\tilde{X}_2$ and D^2 will be *near* its minimum.

(2) *ML estimation using EM Algorithm* (Dempster, Laird and Rubin [5])

(3) *BAN estimation using MGLMM model* (Kleinbaum [18]). A comparison of (2) and (3) is made by Hosking [14] with the results of a Monte Carlo study, using complete data, pairwise deletion for some combinations of sample size, proportion

of missing values and average intercorrelations among the dependent variables. The overall superiority of ML estimation is brought out by him.

(4) *Imputation of missing values* (Greenless et al. [12]). The principle is parameter estimation in a regression model with stochastic censoring of the dependent variable. That is, the case in which the probability of nonresponse for the variable of interest depends upon the value of that variable (one should not ignore the mechanism causing the values to be missing).

ML estimation using EM algorithm

There are two steps for each iteration:

(a) expectation step followed by

(b) maximization step.

The main computation *is to find the parameters of the conditional multivariate normal distribution of the missing values given the observed values in that row.* That is, given a partially observed X, we replace the missing parts of sums, sums of squares, and sums of products by their conditional expectations given the observed data and current fitted population parameters.

- This method is advantageous because of its simplicity and generality. This method also assumes that data are missing at random.

- The EM procedure has also been given for nonrandom missing values like grouped or censored data commonly encountered. But the multivariate normality assumption is still maintained (variables are jointly normally distributed) particularly in factory analysis.

- When there are several missing values, EM algorithm is rather slow.

- EM algorithm does not provide estimates of standard error since calculation and inversion of the information matrix is avoided. However, its advantage is that it provides fitted values for the missing data. The results depend on the pattern of the missing data.

- When the number of missing cells is large in the contingency tables, the iterative fitting procedure is more efficient than the EM algorithm.

Therefore, we should examine the relevance of estimating variances by the EM method for the unobservable values, as pointed out by Searle [34]. EM is still a good procedure among those available for estimating missing values.

EM algorithm provides correct maximum likelihood estimation for many missing data problems assuming multivariate normality, by maximizing the likelihood function $L_1(\theta, Z_m / \tilde{Z}_p) = f(Z_m, \tilde{Z}_p / \theta)$ with respect to (θ, Z_m). Recently, Little and Rubin [20] have suggested a procedure for handling incomplete data when the data are *not* missing at random by maximizing the actual likelihood function $L_2(\theta / \tilde{Z}_p) = \int f(Z_m, \tilde{Z}_p / \theta)\, dZ_m$ which means that the complete-data likelihood is integrated over the missing data Z_m. This procedure is more appropriate than joint maximizing of the complete data likelihood function $L_1(\theta, Z_m / \tilde{Z}_p) = f(Z_m, \tilde{Z}_p / \theta)$ with respect to Z_m and θ which is useful only when few values are missing. Thus even the most recent work using EM algorithm does not meet the needs when the assumptions of multivariate normality are not valid and the missing data are large but sample size remains fixed.

A method of handling non-randomly missing data in arrayed contingency tables using Turner's syndrome data is described by Nordheim [29] where the incompleteness of the data is dependent on the category identity of the observations. Sensitivity analysis incorporating parameters related to the missing data mechanism are recommended for estimates and testing. By introducing a parameter R, which is the ratio of probabilities of uncertain classification, some information on the missing data mechanism can be obtained independently of the data. Such R values are provided by a rough estimate from knowledgeable workers in the particular area.

Testing of hypotheses and inference with missing data

There is very little work on this aspect. The work of Srivastava [37] is an attempt on prediction and is not useful for several situations. The work of Sarkar, Sinha, and Krishnaiah [33] and Kariya, Krishnaiah, and Rao [17] on some aspects of missing data in hierarchical classification is the only available information. The paper by Kariya, Krishnaiah, and Rao [17] is closer to some common situations but is still based on multivariate normality and random missing data. This is an extension of testing unbalanced data from a bivariate normal distribution (Sarkar, Sinha, and Krishnaiah, [33]. Let us consider n_1 paired observations, n_2 additional observations on X only, and n_3 additional observations on Y only.

$$n_1 \leq n_2 \leq n_3$$

Define a new variable $$X_i^* = \lambda_1 X_i - (1 - \lambda_1) \sum_{j=1}^{n_2} c_{ij}^{(1)} X_{n_1 + j}$$

where the matrices $C_1 = \left(< c_{ij}^{(1)} \right)$ and $C_2 = \left(< c_{ik}^{(2)} \right)$ satisfy $C_1 \underset{\sim}{1}_{n_2} = \underset{\sim}{1}_{n_1}$

and

$$C_1 C_1' = (n_1 / n_2) I_{n_1} \; ; \; C_2 \underset{\sim}{I}_{n_3} = \underset{\sim}{I}_{n_1} ; C_2 C_2' = (n_1 / n_3) I_{n_1}$$

then

$$\left(X_i^*, Y_i^*\right) \qquad i = 1, 2, \ldots, n \text{ are iid.}$$

Therefore, tests on hypotheses, $H_{01} : \mu_2 = 0$, $H_{02} : \mu_1 = \mu_2$ and $H_{03} : \rho = 0$ against the alternatives remain invariant under the group of transformations.

$$X_i^* = a + bX_i \,, \; X_{n_1+j} \rightarrow a + bX_{n_1+j}$$

$$Y_i^* = c + dY_i \,, \; Y^*_{n_1+n_2+k} \rightarrow c + dY_{n_1+n_2+k}$$

and

$$-\infty < a, b, c, d < \infty \text{ and } b, d \neq 0$$

Thus, Kariya, Krishnaiah, and Rao [17] considered with Kariya's earlier works, show that by using conditional distributions, it is possible to provide tests for equality of mean vectors of correlated multivariate normal populations. Gupta and Rohatgi [13] considered only a bivariate normal case and estimation of covariance with missing data in that special case. Srivastava's [37] paper on multi-response experiments is also inadequate except for prediction in special cases. In all the above three papers, assumptions like multivariate normality and data missing on several variables (probably at random), are made which are not useful in many situations arising in practice.

ML estimates and likelihood ratio statistics and their asymptotic null and non-null distributions are derived easily for the k- population testing and estimation problem with patterned means and covariance matrices in the presence of missing data. The standard delta method is used for deriving the asymptotic non- null distributions. Iterative algorithms for finding MLE and asymptotic distributions of the MLE and likelihood ratio statistics (LRS) are presented using the EM algorithm.

Srivastava [38] presented a general approach for obtaining ML estimates when the missing values are few in number compared to the sample available even when data are missing by design. The asymptotic distribution of the statistics on which the likelihood ratio test is based is derived. However, the work of Szatrowski [39] and Srivastava [38] does not meet the realistic need where sample size is not very large and missing data are not few. Giving asymptotic results for the missing data problem eliminates the problem of missing data but does not solve the problem of estimation and testing for finite sample size. As stated by Little and Rubin [20] the asymptotic approach to the missing data problem involves a "trivial assumption in which the proportion of missing data goes to zero as the sample size increases."

Distribution-free procedures in multivariate data with missing observations

As multivariate normality assumptions are violated in many cases of missing data, the use of tests where the dispersion matrix is not necessarily p-variate normal is of interest. These are useful in the case of discrete variables. Klotz [19] proposed a distribution-free procedure for missing observations in ordered categorical data, and proposed a modified Cochran-Friedman test. This procedure can be used for testing the equality of treatment means and also to construct a linear combination of treatment subgroups.

The diversity coefficient (DIVC) discussed by Rao [31] to measure diversity between and within populations for several variables including discrete ones, does not require a normality assumption. This measure could be related to Mahalanobis D^2 statistic commonly used to quantify the divergence between populations, but does not take into consideration all situations with missing data.

The paper by Klotz [19] is only a beginning in the desired direction with *assumption of two or three randomly missing values. There is also loss of information in his procedure as part of the data with equal ranks for the treatments is discarded from analysis.* This is more a problem of the limited range of the scale of the variable (in this case a score of 2-7 was used). If censoring or truncation is responsible for missing values, we have the same problem with discrete data as with multivariate continuous distributions. More work is needed in the area of nonparametric analysis for missing data.

The robustness of multivariate tests

Even if multivariate normality is not satisfied as in some special distributions like θ (n)-invariant distributions (including elliptically symmetrical distributions), the usual MANOVA tests like the likelihood ratio test, Roy's test, Lawley-Hotelling's test, and Pillai's test, which are uniformly most powerful invariant (UMPI) under multivariate normality, are still UMPI in these above two anormal distributions. Tests for equality of covariance matrices for non normal data can be done under some conditions (Kariya, [16]). It remains to be seen if these situations are similar with missing data.

Conclusions:

1) There is need to compare the estimation/prediction, testing for equality of mean vectors or means, and testing for equality of covariance matrices using simulated data with a) complete data, b) data missing at random, c) data missing due to design, and d) data missing due to truncation or censoring.

2) The effect of missing data, particularly in type (c) and (d) on multivariate normality assumptions and standard test procedures is unknown.

3) Estimation of means and tests of equality of mean vectors are available now. When we need discriminating analysis, tests of equality of variance-covariance matrices are yet to be developed for missing data.

4) A closer look at the maximum likelihood estimation from incomplete data with the EM algorithm, particularly for the situation where the elements of a variance-covariance matrix are based on unequal observations, will be of interest.

5) There is a need to assess the methods themselves and provide modifications of the methods in the case of missing values of a nonrandom nature. Work on asymptotic distributions of the maximum likelihood estimates and likelihood ratio statistics, assume missing values are few, does not meet the realistic needs as sample sizes are finite and missing data are not few.

6) Continuous interaction of the theoretical and applied users is needed to eliminate the gaps between theory and actual practice.

REFERENCES

[1] Anderson, T.W., 1984. *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc., New York, Ed. 2, 1-675.

[2] Basilevsky, A., Sabourun, D., Huns, D., and Anderson, A., 1985. Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications In Statistics- Simulation Computation* **14**, 371-394.

[3] Beale, E.M.L., and Little, R.J.A., 1975. Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, B **37**, 129-145.

[4] Bhakta, S.T., 1980. Population dynamics in some elite populations of pearl millet and inter-racial crosses of Brassica. Ph. D. Thesis in Genetics, Indian Agricultural Research Institute, New Delhi.

[5] Dempster, A.P., Laird, N.M. and Rubin D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy, Statist. Soc.* B **39**, 1-38.

[6] Eaton, M., and Kovriya, T., 1983. Multivariate tests with incomplete data. *Annals of Statistics* **11**, 654-665.

[7] Federer, W.T., 1984. *Statistical Design and Analysis of Intercropping Experiments.*

[8] Federer, W.T., and Murty, B.R., 1986. Use, limitations and requirements of multivariate analyses for intercropping experiments. Proc. Symposium in Statistics and Festchrift in honour of Joshi, V.M., D. Reidel Publishing Company, Boston, Massachusetts. USA, 269–283.

[9] Federer, W.T. and Wijesinha, A., 1981. Statistical design and analysis for intercropping experiments. BU-735-M in the Biometrics Unit series, Cornell University.

[10] Frane, J.W., 1976. Some simple procedures for handling missing data in multivariate analysis. *Psychometrika* **41**(3), 409-415.

[11] Gnanadesikan, R. and Kettenring, J.R., 1984. A pragmatic review of multivariate methods in applications. *In Statistics: An Appraisal. Proc. 50th Anniversary Conference, Iowa State Statistical Laboratory*, Iowa State University Press, Ames, Iowa, 309-337.

[12] Greenless, J.S., Reece W.S., and Zieschang, K.D., 1982. Imputation of missing values when the probability of response depends on the variable being imputed. *JASA* **77**, 251–261.

[13] Gupta, A.K. and Rohatgi, V.K., 1982. Estimation of covariance from unbalanced data. *Sankhya, Ser. B* **44**(2), 143-153.

[14] Hosking, J.D., 1984. A comparison of several procedures for estimation in incomplete multivariate linear models. In *Program of the Joint Statistical Meetings of the ASA and the Biometric Society*, Aug. 13-16, 245 (abstract).

[15] Ito, K., 1966. On the heteroscedasticity in the linear normal regression model. *Res. Papers Statist. Festchr. Neyman,* (F. David, ed.) 147-155.

[16] Kariya, T., 1981. Robustness of multivariate tests. *Ann. Statist.* **9**(6), 1267-1275.

[17] Kariya, T., Krishnaiah, P.R. and Rao, C.R., 1983. Inference on parameters of multivariate normal population when data are missing. In *Developments in Statistics,* Vol. 4 (Krishnaiah, P.R. ed.) Academic Press (Harcourt Brace Jovanovich, Publishers), New York, London. 137-184.

[18] Kleinbaum, D.G., 1973. A generalization of the growth curve model which allows missing data. *J. Multivariate Analysis* **3**, 117-124.

[19] Klotz, J., 1980. A modified Cochran-Friedman test with missing observations and ordered categorical data. *Biometrics* **36**(4), 665- 670.

[20] Little, R.J.A., and Rubin, D.B., 1983. On jointly estimating parameters and missing data maximizing the complete data likelihood. *The American Statistician* **37**, 218-220.

[21] Machado, S.G., 1983. Two statistics for testing multivariate normality. *Biometrika* **70**(3), 713-718.

[22] Malkovich, J.F. and Afifi, A.A., 1983. On tests for multivariate normality. *J. Amer. Statist. Assoc.* **68**, 176-179.

[23] Mardia, K.V., 1984. Mardia's test of multi-normality. Unpublished technical report.

[24] Murty, B.R., 1973. Biometrical classification of the genus Sorghum. *In "Sorghum in Seventies Int. Symp. 1971*, House, L.R. and Rao, N.G.P., eds. Oxford University Press, New Delhi, 14-38.

[25] Murty, B.R. and Arunachalam. V., 1966. The nature of divergence in relation to breeding system in crop plants. *Ind. J. Genetics* **26A**, 188-198.

[26] Murty, B.R., Arunachalam, V. and Jain, O.P., 1970. Factor analysis in relation to breeding system. *Genetica* **41**, 179-189.

[27] Murty, B.R., Mathuj, J.B.L., and Arunachalam, V., 1965. Self- incompatibility and genetic divergence in *Brassica campestris var brown sarson. Sankhya, Ser. B*, **27**, 271-278.

[28] Narayan, R.K.J., 1982. Discontinuous DNA variation in the evolution of plant species: the genus *Lathyrus. Evolution* **36**(5), 877-891.

[29] Nordheim, E.V., 1984. Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's Syndrome. *Journal of the American Statistical Association* **79**, 772-780.

[30] Orchard, T., and Woodhury, M.A., 1976. A missing information principle: Theory and applications. In Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. I. 697-713.

[31] Rao, C.R., 1982. Diversity and dissimilarity coefficients: A unified approach. *Theor. Population Biology* **21**(1), 24-42.

[32] Reinsel, G., 1984. Estimation and prediction in a multivariate random effect generalized linear model. *JASA* **79**, 406-414.

[33] Sarkar, S.K., Sinha, B.K. and Krishnaiah, P.R., 1983. Some tests with unbalanced data from a bivariate normal population, *Ann. Inst. Statist. Math., A,* **35**, 63-75.

[34] Searle, S.R., 1973. *Linear Models,* John Wiley & Sons, New York.

[35] Shih, W.J. and Weisberg, S., 1986. Assessing influence in multiple linear regression with incomplete data. *Technometrics* **28**, 231-239.

[36] Simon, G.A., and Simonoff, J.S., 1986. Diagnostic plots for missing data in least squares regression. *Journal of the American Statistical Association* **81**, 501-505.

[37] Srivastava, J.N., 1968. On a general class of designs for multiresponse experiments. *Ann. Math. Statist.* **39**, 1825-1843.

[38] Srivastava, M.S., 1985. Multivariate data with missing observations, *Communications In Statistics-Theor. Math.* **14**, 775- 792.

[39] Szatrowski, T.H., 1985. Missing data in the k-population multivariate normal patterned mean and covariance matrix testing and estimation problem. *Communications in Statistics-Simula. Computa.* **14**, 357-370.

[40] Thakur, S.R., 1980. Mutation processes for cytoplasmic male steriles with reference to disease in pearl millet. Ph.D. Thesis in genetics, Indian Agricultural Research Institute, New Delhi.