Characterizing the Estimation of Parameters in Incomplete-Data Problems

DONALD B. RUBIN*

A framework is given for organizing and understanding the problems of estimating the parameters of a multivariate data set which contains blocks of missing observations. The basic technique is to decompose the original estimation problem into smaller estimation problems by factoring the likelihood of the observed data into a product of likelihoods. The result is summarized in a "factorization table," which identifies the "complete-data" factors whose parameters may be estimated using standard, well-understood complete-data techniques, and the "incomplete-data" factors whose parameters must be estimated using special missing-data methods.

1. INTRODUCTION

This article presents a framework for organizing and understanding the problem of estimating the parameters of a multivariate data set which contains blocks of missing observations. The basic technique is to decompose the original estimation problem into smaller estimation problems by factoring the likelihood of observed data into a product of likelihoods whose parameters are "distinct." The factorization is summarized in a "factorization table," which identifies the "complete-data" factors whose parameters may be estimated using standard, well-understood complete-data techniques, and the "incomplete-data" factors whose parameters must be estimated using special missing-data methods. The factorization table also identifies which rows of the data matrix are relevant for each factor and indicates the amount of data available to estimate each parameter (for example, "inestimable" parameters may be isolated). Only some particular cases in which the likelihood of observed data can be factored into a product consisting solely of complete-data factors have been previously considered.

The remainder of Section 1 introduces terminology and notation. Section 2 states and proves the factorization lemma and Section 3 describes the method for producing the factorization table. Section 4 discusses the interpretation of this table for independent and identically distributed (i.i.d.) rows of the data matrix, Section 5 extends this discussion to include conditionally i.i.d. rows, and Section 6 summarizes the advantages of producing the factorization table.

* Donald B. Rubin is research statistician, Educational Testing Service, and visiting lecturer, Department of Statistics, Princeton University, both in Princeton, N.J. The author wishes to thank W.G. Cochran, A.P. Dempster, F. Lord, T.W.F. Stroud and A.T. Yates for helpful comments on earlier drafts of this article. He especially wishes to thank Paul W. Holland for many insightful discussions.

1.1 The Likelihood of the Observed Data

Let $\mathbf{Z}$ be an $n \times p$ (n rows, p columns) data matrix representing the potential realization of p variables on a sample of n experimental units or subjects, and let $\boldsymbol{\theta}$ be the vector parameter of the density of $\mathbf{Z}$. We assume that the data analyst's primary interest is in the estimation of this parameter which lies in an open parameter space.

Let $\mathbf{M}$ be the $n \times p$ "incompleteness" matrix. If the (i, j) entry of $\mathbf{M}$ is 1, the (i, j) entry of $\mathbf{Z}$ is an observed scalar random variable and thus is a real number in the data matrix. If the (i, j) entry of $\mathbf{M}$ is 0, the (i, j) entry of $\mathbf{Z}$ is an unobserved scalar random variable. The collection of these unobserved scalar random variables is designated $\mathring{\mathbf{Z}}$.

We will assume that the observations are missing at random [14] so that the relevant likelihood is the marginal likelihood of the observed data

$$\int_{\mathring{\mathbf{Z}}} f(\mathbf{Z}; \boldsymbol{\theta}) \tag{1.1}$$

where $f(\mathbf{Z}; \boldsymbol{\theta})$ is the probability density function of $\mathbf{Z}$ evaluated at the observed portion of $\mathbf{Z}$ as a function of $\boldsymbol{\theta}$ and $\mathring{\mathbf{Z}}$, and $\int_{\mathring{\mathbf{Z}}}$ is the integral over these unobserved scalar random variables. In place of $f(\mathbf{Z}; \boldsymbol{\theta})$ we could use the more rigorous notation $P_{\boldsymbol{\theta}}^{\boldsymbol{\mu}}(\mathbf{Z})$ where $\boldsymbol{\mu}$ is the $n \times p$ random variable for which $\mathbf{Z}$ is the sample realization; however, the simpler $f(\mathbf{Z}; \boldsymbol{\theta})$ is adequate within the context of this article.

1.2 Distinct Parameters

The objective here is to factor this likelihood into a product of likelihoods whose parameters are "distinct." Such a factorization decomposes the original estimation problem into smaller estimation problems, one for each factor. As explained in detail later, this identifies which parameters can be estimated by complete-data methods, which must be estimated by special missing-data methods, and indicates the amount of data available to estimate each parameter.

To clarify what is meant by a factorization with distinct parameters, assume for the moment that $\mathbf{Z}$ is

© Journal of the American Statistical Association
June 1974, Volume 69, Number 346
Theory and Methods Section

completely observed. We then could partition the columns of $\mathbf{Z}$ into $(\mathbf{Z}_1, \cdots, \mathbf{Z}_r)$ and write the likelihood of the observed data as

$$f(\mathbf{Z}; \boldsymbol{\theta}) = \prod_{i=1}^{r} f(\mathbf{Z}_i | \mathbf{Z}_{i+}; \boldsymbol{\theta}_{i \cdot i+}), \tag{1.2}$$

where $\mathbf{Z}_{i+} = (\mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_r)$ and $f(\mathbf{Z}_i | \mathbf{Z}_{i+}; \boldsymbol{\theta}_{i \cdot i+})$ is the joint conditional density of $\mathbf{Z}_i$ given $\mathbf{Z}_{i+}$ evaluated at the observed values of the data and regarded as a function of the parameters of that density, $\boldsymbol{\theta}_{i \cdot i+} = \boldsymbol{\theta}_{i \cdot i+1, \cdots, r}$. $\mathbf{Z}_{r+}$ is empty and $\boldsymbol{\theta}_{r \cdot r+}$, the parameters of the marginal density of $\mathbf{Z}_r$, may be written as $\boldsymbol{\theta}_r$.

If the rows of $\mathbf{Z}$ are i.i.d. and the model for each row of $\mathbf{Z}$ is the p-variate normal (Gaussian) density, the parameters $\boldsymbol{\theta}_{i \cdot i+}$ are the regression parameters (i.e., regression coefficients and residual covariance matrix) of $\mathbf{Z}_i$ given $\mathbf{Z}_{i+}$. If the rows of $\mathbf{Z}$ are i.i.d. and the model for each row of $\mathbf{Z}$ is the multinomial density for a p-way contingency table (each column in $\mathbf{Z}$ corresponding to a dimension of the table), the parameters $\boldsymbol{\theta}_{i \cdot i+}$ are the conditional probabilities of being in the cells of the table formed by the variables in $\mathbf{Z}_i$ given membership in the various cells of the marginal table formed by the variables in $\mathbf{Z}_{i+}$.

The parameters $\boldsymbol{\theta}_{i \cdot i+}$ are called "distinct" if the posterior distribution for $\boldsymbol{\theta}_{i \cdot i+}$ can be obtained using only its prior and the ith factor in (1.2), $f(\mathbf{Z}_i | \mathbf{Z}_{i+}; \boldsymbol{\theta}_{i \cdot i+})$. If all parameters of the factorization are distinct, we can maximize $f(\mathbf{Z}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ by maximizing each $f(\mathbf{Z}_i | \mathbf{Z}_{i+}; \boldsymbol{\theta}_{i \cdot i+})$ with respect to $\boldsymbol{\theta}_{i \cdot i+}$, $i = 1, \cdots, r$. For the normal and contingency table models with no constraints on the parameters and completely observed $\mathbf{Z}$, all parameters of the factorization are distinct. For other multivariate models, the parameters of a factor may or may not be distinct depending on the particular factor.

1.3 Factoring the Likelihood with Z Partially Missing

If $\mathbf{Z}$ is partially missing, the factorization of the density of $\mathbf{Z}$ in (1.2) does not necessarily imply a similar factorization of the likelihood of the observed data; that is,

$$\int_{\substack{\circ \\ \mathbf{Z}}} f(\mathbf{Z}; \boldsymbol{\theta}) = \int_{\substack{\circ \\ \mathbf{Z}}} \left[\prod_{i=1}^{r} f(\mathbf{Z}_i | \mathbf{Z}_{i+}; \boldsymbol{\theta}_{i \cdot i+})\right] \tag{1.3}$$

does not in general yield r factors with distinct parameters even for the i.i.d. normal or contingency table models. However, for special patterns of incompleteness which contain blocks of missing values, the likelihood of the observed data can be factored into a product of likelihoods with distinct parameters. The factorization of the likelihood in these cases follows from repeated application of the lemma stated and proved in Section 2.

2. THE FACTORIZATION LEMMA

The application of the following lemma requires patterns of incompleteness in which some variables are "more observed" than others and/or some variables are "never jointly observed" with others. See Exhibit A

A. Incompleteness Matrix in Which V_3 Is More Observed Than V_1, and V_1 and V_2 Are Never Jointly Observed[a]

$$\mathrm{M} = \begin{array}{cccccc} \multicolumn{2}{c}{\mathrm{V}_1} & \multicolumn{2}{c}{\mathrm{V}_2} & \multicolumn{2}{c}{\mathrm{V}_3} \\ \mathrm{X} \cdots & \mathrm{X} & 0 \cdots & 0 & 1 \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{X} \cdots & \mathrm{X} & 0 \cdots & 0 & 1 \cdots & 1 \\ 0 \cdots & 0 & \mathrm{X} \cdots & \mathrm{X} & \mathrm{X} \cdots & \mathrm{X} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 \cdots & 0 & \mathrm{X} \cdots & \mathrm{X} & \mathrm{X} \cdots & \mathrm{X} \end{array}$$

[a] 1 = observed, 0 = missing, X = arbitrary.

where the columns of a matrix $\mathbf{V}$ have been partitioned into $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, where $\mathbf{V}_3$ is more observed than $\mathbf{V}_1$, and $\mathbf{V}_1$ and $\mathbf{V}_2$ are never jointly observed. Patterns such as this are probably relatively common in practice. For example, in testing situations, alternate forms ($\mathbf{V}_1$ and $\mathbf{V}_2$) may be given to different subgroups of subjects with background variables ($\mathbf{V}_3$) recorded for all subjects; or in observational studies, data may be missing in blocks, as when a teacher forgot to hand out the pretest to her third grade classes; or in a survey, some variables may be relatively easy to record (e.g., age, race) and others far more difficult (e.g., sexual habits).

2.1 Lemma

Let $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$ be a column partition of $\mathbf{V}$, an $n \times q$ data matrix. If

a. $\mathbf{V}_3$ is more observed than $\mathbf{V}_1$
(i.e., in any row in which $\mathbf{V}_1$ is at least partially observed $\mathbf{V}_3$ is fully observed,
b. $\mathbf{V}_2$ and $\mathbf{V}_1$ are never jointly observed
(i.e., in any row in which $\mathbf{V}_1$ is at least partially observed $\mathbf{V}_2$ is completely missing), and
c. the rows of $\mathbf{V}$ given $\mathbf{V}_3$ are conditionally independent and identically distributed,

then the marginal likelihood of $\mathbf{V}$ can be factored as

$$\int_{\substack{\circ \\ \mathbf{V}}} f(\mathbf{V}; \boldsymbol{\zeta}) = \int_{\substack{\circ \\ \mathbf{V}_1}} f(\mathbf{V}_1 | \mathbf{V}_3; \boldsymbol{\zeta}_{1 \cdot 3}) \int_{\substack{\circ \quad \circ \\ \mathbf{V}_2, \mathbf{V}_3}} f(\mathbf{V}_2, \mathbf{V}_3; \boldsymbol{\zeta}_{23}) \tag{2.1}$$

where $\boldsymbol{\zeta}$ is the vector parameter of the marginal density of $\mathbf{V}$ and the other notation for writing likelihoods follows analogously from (1.1) and (1.2).

2.2 Distinct and Inestimable Parameters of the Factorization

The parameters $\boldsymbol{\zeta}_{1 \cdot 3}$ and $\boldsymbol{\zeta}_{23}$ are always distinct for a general normal or contingency table model because the parameters $(\zeta_{12\cdot3}, \zeta_3)$ are always distinct as are the parameters $(\zeta_{1\cdot3}, \zeta_{2\cdot3}, \zeta_3)$. The parameters of ζ that are distinct from those of $(\zeta_{1\cdot3}, \zeta_{23})$ are the distinct parameters of conditional association (e.g., partial correlation, conditional interaction) between $\mathbf{V}_1$ and $\mathbf{V}_2$ given $\mathbf{V}_3$; these are "inestimable" in the sense that their maximum likelihood estimates are restricted only by the parameter space, or equivalently, their posterior distributions equal their prior distributions. Because the parameter space is open, any function of an inestimable parameter does not possess a unique maximum likelihood estimate. It can be seen from (2.2) that these parameters of conditional association are inestimable whenever Conditions b and c in the lemma hold.

2.3 Proof of the Lemma

Let the prescript 1 indicate those rows in which $\mathbf{V}_1$ is at least partially observed and the prescript 2 indicate the remaining rows in which $\mathbf{V}_1$ is fully missing. See Exhibit A where the first block of rows corresponds to ${}_1\mathbf{V}$ and the second set corresponds to ${}_2\mathbf{V}$.

By Condition c in the lemma rewrite $\int_{\mathbf{V}}^{\circ} f(\mathbf{V}; \zeta)$ as

$$\int_{\overset{\circ}{\mathbf{V}}} [f({}_1\mathbf{V}_1, {}_1\mathbf{V}_2 | {}_1\mathbf{V}_3; \zeta_{12\cdot3}) f({}_2\mathbf{V}_1, {}_2\mathbf{V}_2, | {}_2\mathbf{V}_3; \zeta_{12\cdot3}) f(\mathbf{V}_3; \zeta_3)].$$

By definition of the prescripts 1 and 2 and by Condition b, both ${}_2\mathbf{V}_1$ and ${}_1\mathbf{V}_2$ are fully missing. Hence, taking the integral over ${}_1\mathbf{V}_2$ and ${}_2\mathbf{V}_1$ gives

$$\int_{\overset{\circ}{\mathbf{V}_3}} \left\{ \left[\int_{{}_1\overset{\circ}{\mathbf{V}_1}} f({}_1\mathbf{V}_1 | {}_1\mathbf{V}_3; \zeta_{1\cdot3}) \right] \cdot \left[\int_{{}_2\overset{\circ}{\mathbf{V}_2}} f({}_2\mathbf{V}_2 | {}_2\mathbf{V}_3; \zeta_{2\cdot3}) \right] f(\mathbf{V}_3; \zeta_3) \right\}. \quad (2.2)$$

By Condition a ${}_1\mathbf{V}_3$ is fully observed and so (2.2) becomes

$$\int_{{}_1\overset{\circ}{\mathbf{V}_1}} f({}_1\mathbf{V}_1 | {}_1\mathbf{V}_3; \zeta_{1\cdot3}) \int_{{}_2\overset{\circ}{\mathbf{V}_2}, {}_2\overset{\circ}{\mathbf{V}_3}} [f({}_2\mathbf{V}_2 | {}_2\mathbf{V}_3; \zeta_{2\cdot3}) f(\mathbf{V}_3; \zeta_3)]$$

which equals (2.1) since ${}_1\mathbf{V}_2$ and ${}_2\mathbf{V}_1$ are fully missing and ${}_1\mathbf{V}_3$ is fully observed.

3. CREATING THE FACTORIZATION TABLE

We now describe the method for factoring the likelihood of the observed data and illustrate its use with a simple example. The rows of $\mathbf{Z}$ are assumed i.i.d.; conditionally i.i.d. rows of $\mathbf{Z}$ are considered in Section 5. The resultant factorization of the likelihood is summarized in a "factorization table" which is interpreted in detail in Section 4.

Step 1. Replace all rows of $\mathbf{M}$ having the same 0–1 pattern with one row having that pattern, noting which rows of $\mathbf{Z}$ are represented by that pattern. Similarly, replace all columns of $\mathbf{M}$ having the same 0–1 pattern with one column having that pattern, noting which columns of $\mathbf{Z}$ are represented by that pattern. Call the resulting 0–1 incompleteness matrix $\tilde{\mathbf{M}}$. See Table 1 for the example of $\tilde{\mathbf{M}}$ that we will use to illustrate this method.

1. Example of Incompleteness Matrix $\tilde{M}$ for a 550×8 Data Matrix

Columns represented					Rows
1,2	*3*	*4,5*	*6,7*	*8*	
0	1	1	1	1	1–100
1	1	0	1	1	101–275
1	1	0	0	1	276–300
0	0	0	0	1	301–450
0	1	0	0	0	451–550

Step 2. Reorder and partition the columns of $\tilde{\mathbf{M}}$ into $(\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2)$ such that each column in $\tilde{\mathbf{M}}_2$ is either (a) more observed than every column in $\tilde{\mathbf{M}}_1$, or (b) never jointly observed with any column in $\tilde{\mathbf{M}}_1$. If this cannot be done, the pattern of incompleteness in $\tilde{\mathbf{M}}$ (and $\mathbf{M}$) is "irreducible," and no further progress can be made. Assuming this step can be performed, proceed to Step 3. In Table 2 see the result of Step 2 for the matrix $\tilde{\mathbf{M}}$ of Table 1. Here, every column in $\tilde{\mathbf{M}}_2$ is more observed than all columns in $\tilde{\mathbf{M}}_1$.

2. The First Partitioning of the $\tilde{M}$ of Table 1

Columns represented					Rows
1,2	*4,5*	*6,7*	*3*	*8*	
0	1	1	1	1	1–100
1	0	1	1	1	101–275
1	0	0	1	1	276–300
0	0	0	0	1	301–450
0	0	0	1	0	451–550
	$\tilde{M}_1$		$\tilde{M}_2$		

Step 3. Apply the procedure in Step 2 to both partitions created in Step 2. If both $\tilde{\mathbf{M}}_1$ and $\tilde{\mathbf{M}}_2$ are irreducible, stop; otherwise proceed trying to repartition each partition created. Continue until no partition of $\tilde{\mathbf{M}}$ can be further partitioned. See Table 3 for the final partitioning for our example. This partitioning was achieved by examining the $\tilde{\mathbf{M}}_1$ partition of Table 2 and noting that Columns (6, 7) are more observed than Columns (4, 5), and Columns (1, 2) are never jointly observed with Columns (4, 5). The $\tilde{\mathbf{M}}_2$ partition of Table 2 and all partitions in Table 3 are irreducible.

3. The Final Partitioning of the $\tilde{M}$ of Table 1

Columns represented					Rows
4,5	*1,2*	*6,7*	*3*	*8*	
1	0	1	1	1	1–100
0	1	1	1	1	101–275
0	1	0	1	1	276–300
0	0	0	0	1	301–450
0	0	0	1	0	451–550
$\tilde{M}_1$	$\tilde{M}_2$		$\tilde{M}_3$		

Step 4. Summarize the final partitions in a "factorization table" as illustrated in Table 4 for our example. Labeling the final partitions $\tilde{\mathbf{M}}_1, \cdots, \tilde{\mathbf{M}}_r$ from left to right, list for each $\tilde{\mathbf{M}}_i$:

1. the "conditioned" variables—the variables (columns of $\mathbf{Z}$) represented by the ith partition: say, $\mathbf{Z}_i$.
2. the "marginal" variables—the variables, represented in partitions to the right of the ith partition, that are more observed than variables in the ith partition: say, $\mathbf{Z}_{i*}$.
3. the "missing variables"—the variables, represented in the partitions to the right of the ith partition, that are never jointly observed with the variables in the ith partition: say, $\mathbf{Z}_{i0}$.
4. whether it is a "complete-data" partition—one column in $\tilde{\mathbf{M}}_i$, or an "incomplete-data" partition—more than one column in $\tilde{\mathbf{M}}_i$.
5. the rows of $\mathbf{Z}_i$ which are at least partially observed (i.e., the rows of $\mathbf{Z}$ represented by rows of $\tilde{\mathbf{M}}_i$ that are not all zero).

4. The Factorization Table for Table 3

Partition	Complete or incomplete	Conditioned variables	Marginal variables	Missing variables	Rows of Z
1	complete	4,5	3,6,7,8	1,2	1–100
2	incomplete	1,2,6,7	3,8	—	1–300
3	incomplete	3,8	—	—	1–550

Each partition in $\tilde{\mathbf{M}}$, and thus each row of the factorization table, corresponds to a factor of the likelihood of the observed data. The ith factor is the conditional joint distribution of the conditioned variables for that partition, $\mathbf{Z}_i$, given the marginal variables for that partition $\mathbf{Z}_{i*}$. Hence, the final factorization of the likelihood of the observed data is

$$\prod_{i=1}^{r} \int_{\mathring{\mathbf{Z}}_i} f(\mathbf{Z}_i \,|\, \mathbf{Z}_{i*}; \boldsymbol{\theta}_{i\cdot i*}) \tag{3.1}$$

where $\mathbf{Z}_{r*}$ is empty and $\boldsymbol{\theta}_{r\cdot r*}$ may be written as $\boldsymbol{\theta}_r$.

This factorization for i.i.d. rows of $\mathbf{Z}$ follows from $r - 1$ applications of the lemma of Section 2 beginning with $\mathbf{V}_1 = \mathbf{Z}_1, \mathbf{V}_2 = \mathbf{Z}_{10}, \mathbf{V}_3 = \mathbf{Z}_{1*}$; then $\mathbf{V}_1 = \mathbf{Z}_2, \mathbf{V}_2 = \mathbf{Z}_{20}, \mathbf{V}_3 = \mathbf{Z}_{2*}$; and ending with $\mathbf{V}_1 = \mathbf{Z}_{r-1}$ and $\mathbf{V}_2$ or $\mathbf{V}_3 = \mathbf{Z}_r$. The lemma can be applied because, by construction, every column in partitions to the right of the ith partition is either more observed than the columns in the ith partition or never jointly observed with the columns in the ith partition. Hence, each variable to the right of the ith partition (i.e., each column in $\mathbf{Z}_{i+}$) is either in $\mathbf{Z}_{i*}$ or $\mathbf{Z}_{i0}$.

The parameters of the factorization are distinct whenever the parameters at each step of the factorization are distinct. Hence, the parameters are always distinct for the general normal and contingency table models.

A factorization such as that in (3.1) has been previously suggested only in special cases in which all partitions are complete and apparently then only for the multivariate normal (e.g., [2, 9]).

4. INTERPRETING THE FACTORIZATION TABLE

The ith partition [the ith factor in (3.1)] represents the likelihood of the conditioned variables in the ith partition given the marginal variables in the ith partition. In our example, the three likelihoods are variables (4, 5) given variables (3, 6, 7, 8), variables (1, 2, 6, 7) given variables (3, 8), and the marginal likelihood of variables (3, 8). The vector parameter in the ith likelihood is $\boldsymbol{\theta}_{i\cdot i*}$. For the normal model, these are the regression parameters of the conditioned variables given the marginal variables. For the contingency table model these are the conditional probabilities of being in each cell of the table formed by the conditioned variables given membership in each of the cells of the table formed by the marginal variables.

Each factor involves only the rows of $\mathbf{Z}$ indicated in Column 5 of the factorization table. If the parameters of the ith factor, $\boldsymbol{\theta}_{i\cdot i*}$, are distinct, their maximum likelihood or Bayes estimates using the indicated rows and columns in the factorization table [e.g., Rows (1–100) and Columns (3–8) for the first factor] are identical to their maximum likelihood or Bayes estimates using all rows and columns of $\mathbf{Z}$. This is the definition of distinct parameters. If the parameters $\boldsymbol{\theta}_{i\cdot i*}$ are not distinct, then estimates based on the rows and columns indicated by the factorization table are simply not as efficient as they could be since other factors contain information relevant to estimating $\boldsymbol{\theta}_{i\cdot i*}$.

4.1 Inestimable Parameters

The scalar parameters of $\boldsymbol{\theta}$ that are distinct from those of $(\boldsymbol{\theta}_{1\cdot 1*}, \boldsymbol{\theta}_{2\cdot 2*}, \cdots, \boldsymbol{\theta}_r)$ are inestimable from the data in the previous sense (Section 2) of prior equals posterior. For a contingency-table or normal model, these are the parameters of conditional association (partial correlation) between the conditioned variables and missing variables given the marginal variables in each partition. In our example, the parameters of conditional association between variables (4, 5) and (1, 2) given variables (3, 6, 7, 8) would be inestimable for these densities.

Other parameters may also be inestimable. For instance, by the result in Section 2, whenever two variables are never jointly observed, the distinct parameters of conditional association between them are inestimable even if the variables are in the same partition. Our example does not contain such a pattern of incompleteness.

4.2 A Complete-Data Factor

If the ith partition is "complete," the ith factor is a "complete-data" factor in the sense that it involves a completely observed data matrix. This follows because each row of $\mathbf{Z}_i$ is either fully observed or fully missing (since $\mathbf{Z}_i$ is represented by only one column in $\tilde{\mathbf{M}}_i$), and taking the integral over $\mathring{\mathbf{Z}}_i$ (the fully missing rows of $\mathbf{Z}_i$) leaves only rows in which $\mathbf{Z}_i$ and $\mathbf{Z}_{i*}$ are fully observed. In our example, the first partition is complete and involves Rows 1–100 as indicated in Column 5 of Table 4. Thus, the first factor, variables (4, 5) given variables (3, 6, 7, 8), is a complete-data factor and involves only Rows (1–100) and Columns (3–8) of $\mathbf{Z}$ which form a completely observed data matrix.

Given a complete-data factor, we are on familiar ground in that

1. Its parameters, $\boldsymbol{\theta}_{i \cdot i*}$, can be estimated using any of the usual computational methods (e.g., least squares, ridge regression, iterative-fitting, cross tabulation) designed for a completely observed data matrix, and
2. The properties of the resultant estimates are the same as in any complete-data problem (e.g., least squares estimates of regression coefficients are maximum likelihood under the normal model and are unbiased under the more general linear regression model).

In our example, if $\mathbf{Z}$ is normal, one would use complete-data regression techniques to estimate the regression parameters of variables (4, 5) on variables (3, 6, 7, 8) using the fully observed 100×6 data matrix consisting of Rows (1–100) and Columns (3–8) of $\mathbf{Z}$. Similarly, if the model for Z is a contingency table, one would use complete-data contingency table techniques to estimate the conditional probabilities of being in each cell of the two-way table formed by variables (4, 5) given membership in each cell of the four-way table formed by variables (3, 6, 7, 8), again using the fully observed 100×6 data matrix.

As with any complete-data problem, if the number of observation vectors (i.e., number of rows in which $\mathbf{Z}_i$ is observed) is small compared to the number of parameters being estimated, the resulting estimates may not be very good even if the model is correct. For example, in the normal case, the number of degrees of freedom for the estimation of the residual covariance matrix must be positive for unique maximum likelihood estimates to exist, and if the degrees of freedom are small the estimates may be quite poor (e.g., with respect to mean-squared error). Hence, if the number of observed rows in $\mathbf{Z}_i$ is small, one might want to consider fitting a reduced model with restrictions on the parameters (see [7] for the normal model and [4] for the contingency table model). In any case, knowing the number of observation vectors available for estimating the parameters of a complete-data factor is as important as knowing the number of observation vectors available in a problem with no missing data, and this is not easily ascertainable without factoring the likelihood of the observed data.

4.3 An Incomplete-Data Factor

If the ith partition is "incomplete," the ith factor is an "incomplete-data" factor in the sense that it involves an incompletely observed data matrix which cannot be factored. This follows because taking the integral over the fully missing rows in $\mathbf{Z}_i$ still leaves those rows in which only some variables in $\mathbf{Z}_i$ are observed and all variables in $\mathbf{Z}_{i*}$ are observed. In our example, the second and third partitions are incomplete and, as indicated in Table 1, involve Columns (1–3, 6–8) and Rows (1–300), and Columns (3, 8) and Rows (1–500), respectively.

When estimating the parameters $\boldsymbol{\theta}_{i \cdot i*}$ in an incomplete-data factor we are on unfamiliar ground even assuming that the parameters of the factorization are distinct. In principle, we can always find posterior densities and, if they exist, unique maximum likelihood estimates. Iterative maximum likelihood algorithms are available for the normal (see [11]) and the multinomial (see [12]). However, we cannot rely on complete-data results to imply that these estimates and/or their estimated variances are either (1) reasonable in small or moderate samples if the model is correct, or (2) robust if the model is not correct.

The same two problems exist with estimation methods based on "pseudosufficient" statistics. The most common example of this approach is to use the marginal means and variances and the "missing-observation" correlation matrix in the normal case (see [10] for a Monte Carlo study of this method). Similarly, marginal and conditional "pseudosufficient" frequencies can be used in the contingency table case. Although the estimates found by these methods are consistent, they can lead to strange results in small samples (e.g., negative roots for correlation matrix, probabilities over all cells do not sum to unity), and deciding when the sample is small is not as simple as in a complete-data problem.

In an incomplete-data factor with distinct parameters, different scalar parameters of $\boldsymbol{\theta}_{i \cdot i*}$ can be estimated with various efficiencies but none with the efficiency associated with all the partially observed rows in $\mathbf{Z}_i$. Hence, the number of rows indicated in Column 5 of the factorization table is an overestimate of the effective sample size. A possible definition for the effective sample size for the estimation of a scalar parameter is the minimum number of row deletions such that the parameter becomes inestimable. This issue is interesting but beyond the scope of this article; the point, however, is clear: an awareness of the number of observations used to estimate a parameter is essential to statistically intelligent conclusions based on the estimate and its estimated (asymptotic) variance.

The study of the effects of deviations from assumptions on maximum likelihood (or Bayes) estimates based on incomplete data is an essentially unexplored field (by the previous discussion, the estimates from complete-data factors are, of course, excluded).

5. DEPENDENT AND INDEPENDENT VARIABLES

Often the data analyst's primary interest is not in the parameters of the joint distribution of all variables but rather in the parameters of the conditional distribution of some of the variables, the "dependent variables," given the other variables, the "independent variables." Partitioning the columns of $\mathbf{Z}$ into $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$ where $\mathbf{X}$ represents the independent variables and $\mathbf{Y}$ the dependent variables, the likelihood of the observed data can be written as

$$\int_{\overset{\circ}{\mathbf{Z}}} f(\mathbf{Z}; \boldsymbol{\theta}) = \int_{\overset{\circ}{\mathbf{X}}} \int_{\overset{\circ}{\mathbf{Y}}} f(\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\phi}) f(\mathbf{X}; \boldsymbol{\psi}). \quad (5.1)$$

where $\boldsymbol{\phi}$ is the vector parameter of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, and $\boldsymbol{\psi}$ is the vector parameter of the marginal distribution of $\mathbf{X}$. Generally, the marginal distribution of the independent variables is not specified, thus implying $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are distinct parameters. Also, the rows of $\mathbf{Z}$ are usually assumed conditionally i.i.d. given $\mathbf{X}$.

5.1 X More Observed Than Y

If $\mathbf{X}$ is more observed than $\mathbf{Y}$, we can apply the lemma of Section 1 with $\mathbf{V}_1 = \mathbf{Y}$ and $\mathbf{V}_2 = \mathbf{X}$ to obtain the following factorization of (5.1):

$$\int_{\overset{\circ}{\mathbf{Y}}} f(\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\phi}) \int_{\overset{\circ}{\mathbf{X}}} f(\mathbf{X}; \boldsymbol{\psi}). \quad (5.2)$$

Hence, in this case, the likelihood of $\mathbf{Y}$ given $\mathbf{X}$ is a factor of the likelihood of the observed data with distinct parameters, and thus the marginal likelihood of $\mathbf{X}$ can be ignored as usual in conditional analyses. Notice that $\mathbf{X}$ more observed than $\mathbf{Y}$ is very common in analysis of variance models in which $\mathbf{X}$ is the design matrix and the missing $\mathbf{Y}$ observations are those that destroy "balance" in the design (see [15]).

When forming the factorization table with $\mathbf{X}$ more observed than $\mathbf{Y}$, one need consider only the columns of $\mathbf{Z}$ that correspond to the $\mathbf{Y}$ variables. The ith factor in the resultant factorization table then gives the conditional density of the conditioned variables, $\mathbf{Y}_i$, given the marginal variables, $\mathbf{Y}_{i*}$, and implicitly the independent variables, $\mathbf{X}$. Hence, assuming r partitions of $\mathbf{Y}$, the factorization of the likelihood of the observed data implied by the factorization table is

$$\left[\prod_{i=1}^{r} \int_{\overset{\circ}{\mathbf{Y}_i}} f(\mathbf{Y}_i \mid \mathbf{Y}_{i*}, \mathbf{X}; \boldsymbol{\phi}_{i\cdot i*})\right] \int_{\overset{\circ}{\mathbf{X}}} f(\mathbf{X}; \boldsymbol{\psi}),$$

where $\mathbf{Y}_{r*}$ is empty, and $\boldsymbol{\phi}_{r\cdot r*}$ may be written as $\boldsymbol{\phi}_r$, the parameters of the marginal density of $\mathbf{Y}_r$ given $\mathbf{X}$. Assuming conditionally i.i.d. rows of $\mathbf{Z}$ given $\mathbf{X}$, this result follows from r applications of the lemma starting with $\mathbf{V}_1 = \mathbf{Y}_1$, $\mathbf{V}_2 = \mathbf{Y}_{10}$, $\mathbf{V}_3 = (\mathbf{Y}_{1*}, \mathbf{X})$ and ending with $\mathbf{V}_1 = \mathbf{Y}_r$ and $\mathbf{V}_3 = \mathbf{X}$. The parameters of the factorization are distinct whenever the parameters at each step of the factorization are distinct; hence, the parameters are always distinct for the general linear (i.e., conditional normal) and loglinear (i.e., conditional contingency table) models. Thus, for $\mathbf{X}$ more observed than $\mathbf{Y}$, the method of Section 3 and the discussion of Section 4 are completely appropriate with $\mathbf{Y}$ replacing $\mathbf{Z}$, the parameter $\boldsymbol{\phi}$ replacing $\boldsymbol{\theta}$, and each factor additionally being conditional on $\mathbf{X}$.

5.2 X Not More Observed Than Y

If some $\mathbf{X}$ variables are not more observed than all of the $\mathbf{Y}$ variables, a joint density must be specified for some $\mathbf{X}$ variables in order to have a well-defined marginal likelihood for $\mathbf{Y}$ given $\mathbf{X}$. More specifically, reorder and partition the columns of $\mathbf{X}$ as $(\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_2$ is more observed than $\mathbf{X}_1$ and $\mathbf{Y}$; then, from (5.1) and the factorization lemma, the likelihood of the observed data can be written as

$$\int_{\overset{\circ}{\mathbf{X}_1}} \int_{\overset{\circ}{\mathbf{Y}}} [f(\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\phi}) f(\mathbf{X}_1 \mid \mathbf{X}_2; \boldsymbol{\psi}_{1\cdot 2})] \int_{\overset{\circ}{\mathbf{X}_2}} f(\mathbf{X}_2; \boldsymbol{\psi}_2). \quad (5.3)$$

If $\mathbf{X}_1$ is not more observed than $\mathbf{Y}$, (5.3) does not reduce to (5.2) and thus the density $f(\mathbf{X}_1 | \mathbf{X}_2; \boldsymbol{\psi}_{1\cdot 2})$ must be specified to obtain a well-defined marginal likelihood for $\mathbf{Y}$ given $\mathbf{X}$. The specification of $f(\mathbf{X}_1 | \mathbf{X}_2; \boldsymbol{\psi}_{1\cdot 2})$ seems reasonable because in order for a variable to be incompletely observed, some kind of "sampling" took place even if it was merely the process of recording a "fixed" value.

If the rows of $\mathbf{Z}$ are conditionally i.i.d. given $\mathbf{X}_2$, we have returned to the case considered previously in this section in which the "depleted independent" variables, $\mathbf{X}_2$, are more observed than the "augmented dependent" variables $(\mathbf{Y}, \mathbf{X}_1)$; hence, we form the factorization table examining the pattern of incompleteness for $(\mathbf{Y}, \mathbf{X}_1)$, where each factor is conditional on $\mathbf{X}_2$. This approach seems extremely appropriate for problems such as missing covariates in analysis of variance designs, since these covariates are often essentially the same as dependent variables except that the investigator assumes they are not affected by the treatments (e.g., pretest scores).

5.3 Other Methods for Missing Independent Variables

Hartley and Hocking [11] have described a different "likelihood" method for handling missing independent variables which essentially puts a flat prior on the conditional parameters, $\boldsymbol{\phi}$, and on the unobserved scalar random variables, $\overset{\circ}{\mathbf{X}}$, and then estimates $\boldsymbol{\phi}$ by maximizing the posterior with respect to both $\boldsymbol{\phi}$ and $\overset{\circ}{\mathbf{X}}$. Actually they consider only univariate $\mathbf{Y}$ and show that with continuous independent variables [on $(-\infty, \infty)$] this approach for estimating $\boldsymbol{\phi}$ is equivalent to discarding all units with missing $\mathbf{X}$'s and for a discrete $\mathbf{X}$ is equivalent to a nonlinear integer programming problem. Even with univariate $\mathbf{Y}$, their approach is clearly not equivalent to the one proposed here in which a missing $\mathbf{X}$ value is considered to be just like a missing $\mathbf{Y}$ value in that it is associated with the observed values of the row and column in which it appears.

A special case in which these two methods for estimating $\boldsymbol{\phi}$ are equivalent is $\mathbf{X}$ more observed than $\mathbf{Y}$, for then both utilize only the rows of $\mathbf{Z}$ for which all the

$\mathbf{X}$ variables are observed. If, in addition, each row of $\mathbf{Y}$ is fully observed or fully missing (e.g., univariate $\mathbf{Y}$) and the model for $\mathbf{Y}$ given $\mathbf{X}$ is the usual conditional normal, the "least-squares" approach to missing values (which minimizes a norm of the residual cross-products matrix over the parameters $\boldsymbol{\phi}$ and unobserved random variables $\mathring{\mathbf{X}}$ and $\mathring{\mathbf{Y}}$) agrees with the preceding two methods; this follows because all three methods discard all rows with any missing observations and estimate $\boldsymbol{\phi}$ by least squares (except for the divisor of the residual covariance matrix).

The fact that both least squares and maximum likelihood discard incompletely observed rows for the special case of simple regression with complete $\mathbf{X}$ (i.e., univariate complete $\mathbf{X}$ and conditionally normal univariate $\mathbf{Y}$) has been mentioned by several writers, e.g., Afifi and Elashoff [1] and Edgett [8]. Note, however, from (5.2) that maximum likelihood or Bayes estimation of $\boldsymbol{\phi}$ always ignoring rows with any missing observations is not tied to univariate $\mathbf{Y}$, complete $\mathbf{X}$, and normally distributed residuals but rather to (a) rows of $\mathbf{Z}$ that are either fully observed or are missing all dependent variables, and (b) conditionally i.i.d. rows of $\mathbf{Z}$ given $\mathbf{X}$. For an example of this fact with a bivariate $\mathbf{Y}$ having a nonlinear regression on $\mathbf{X}$, see Box, *et al.* [5, p. 619].

6. ADVANTAGES OF THE FACTORIZATION

A method has been presented for factoring the likelihood of the observed data into a product of complete-data and incomplete-data likelihoods with distinct parameters. This factorization can be summarized in a factorization table which identifies which parameters can be estimated by usual complete-data methods, which parameters must be estimated using special incomplete-data methods, which rows of the data matrix are relevant for estimating each parameter, and indicates how much data is available to estimate each parameter.

The primary advantage of the factorization is this characterization of the problems of estimating parameters. However, there are other advantages especially if an efficient algorithm is developed for the production of the factorization table. For example, an examination of $\tilde{\mathbf{M}}$ highlights particular patterns of incompleteness, and thus may allow the investigator to recover missing data (e.g., one teacher forgot to turn in the second questionnaire), or to discard inappropriate groups of variables or units (e.g., almost no one was willing to answer a group of questions, or one group of subjects refused to answer almost all questions), or perhaps simply to understand better the process that "caused" the missing data.

6.1 Computing Maximum Likelihood Estimates

The factorization may also be useful in computing maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ (or $\boldsymbol{\phi}$ if there are independent variables). Each individual factor is a reduced problem in the sense that it contains fewer rows, fewer variables that may have missing values, and fewer parameters than the original problem. Hence, the storage and time requirements for the estimation of the parameters of the factorization may be less than for the parameters of the original problem, especially if some factors are complete. If a unique maximum exists, we can find maximum likelihood estimates of the original parameters $\boldsymbol{\theta}$ (or $\boldsymbol{\phi}$) from maximum likelihood estimates of the parameters of the factorization by the usual complete-data operators: "sweep" (see [3] or [6]) for the (conditional) normal model and "direct-product" for the (conditional) contingency table model. When unique maximum likelihood estimates do not exist for all parameters, the factorization can isolate inestimable parameters and/or parameters of factors with too few rows to yield unique estimates. The isolation of these parameters is important not only for statistically intelligent conclusions, but also because algorithms may not converge when attacking problems with nonunique maximums.

Of course, estimates other than maximum likelihood can be used in this procedure. For example, in the normal model one could use estimates based on the missing observation correlation matrix made positive by a ridge procedure.

6.2 Computing Estimates by Filling in and/or Deleting Values

One class of methods for computing estimates is rather obvious within the context of factoring the likelihood. Such methods involve deleting and/or filling in observations until the pattern of incompleteness in each complete-data factor is modified so that a full factorization to complete-data factors is possible.

Any process that deletes observations in order to obtain a special pattern of incompleteness is a special case of missing at random and thus the resulting estimates suffer only from reduced efficiency. In some cases, deleting observations and proceeding from the reduced data matrix may be more reasonable than calculating estimates and their (asymptotic) variances by maximum likelihood. For example, if in a partition there are two variables and they have only one row in common, it seems more reasonable to delete one value and factor the likelihood into two marginal likelihoods than to obtain estimates of the parameters of conditional association (or functions of these parameters) and their "asymptotic" variances based on a sample size of one. In our example, we could delete the 100 observations in Rows (451–550) on variable 3 and the 50 observations in rows (276–300) on variables (1, 2) and then partition both $\tilde{\mathbf{M}}_2$ and $\tilde{\mathbf{M}}_3$ of Table 3. There are of course other ways to delete observations and fully reduce our example to complete-data factors. A "common" choice is to delete all rows of $\mathbf{Z}$ in which there are any missing observations—in our example this would discard all the data.

If instead of deleting observations in the example, we filled in Rows (276–300) for variables (6, 7) and Rows (451–550) for variable 8, we could partition both $\tilde{\mathbf{M}}_2$ and $\tilde{\mathbf{M}}_3$ into complete partitions. There are other choices for which missing observations to fill in. A common choice is to fill in all values—in our example, a majority of the values in the resultant data matrix would then have been filled in. Filling in observations is preferably done by recovering the original values; this might be practical if there are only a few missing observations. Otherwise, "expert" opinion could be used to estimate the values. Methods based on prediction from the data at hand (e.g., the mean of the variable for the normal, the most frequent category of the variable for the multinomial, a regression predicted value for the conditional normal) have been used and, to some extent, studied in simple cases (e.g., [9]). Typically, the resultant data produce consistent estimates of some parameters, but not of others (e.g., filling in a regression predicted value does not bias the estimate of the regression line but biases the estimate of the residual variance towards zero). Also, filling in observations by prediction from the data effectively inflates the sample size and thus can lead to faulty conclusions about the accuracy of estimates.

A combination of filling in and deleting observations is also possible. In our example one could delete Rows (276–300) for variables (1, 2), fill in Rows (451–550) for variable 8, and then partition both $\tilde{\mathbf{M}}_2$ and $\tilde{\mathbf{M}}_3$ in Table 3. How to fill in and delete observations so that all partitions are complete with "minimal disturbance" to $\mathbf{Z}$ is an interesting practical problem and related to the theoretical one mentioned earlier concerning the number of observations available to estimate each parameter.

[Received May 1971. Revised October 1972.]

REFERENCES

[1] Afifi, A.A., and Elashoff, R.M., "Missing Observations in Multivariate Statistics, I: Review of the Literature," *Journal of the American Statistical Association*, 61 (September 1966), 595–604.

[2] Anderson, T.W., "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations Are Missing," *Journal of the American Statistical Association*, 52 (June 1957), 200–3.

[3] Beaton, A.E., "The Use of Special Matrix Operations in Statistical Calculus," Educational Testing Service Research Bulletin, RB-64-51.

[4] Bishop, Y., "Full Contingency Tables, Logits, and Split Contingency Tables," *Biometrics*, 25 (June 1969), 383–400.

[5] Box, M.J., Draper, N.R. and Hunter, W.G., "Missing Values in Multiresponse Nonlinear Model Fitting," *Technometrics*, 12 (August 1970), 613–20.

[6] Dempster, A.P., *Elements of Continuous Multivariate Analysis*, Reading, Mass.: Addison-Wesley Publishing Co., 1969.

[7] ———, "Covariance Selection," *Biometrics*, 28 (March 1972), 157–76.

[8] Edgett, G.L., "Multiple Regression with Missing Observations among the Independent Variables," *Journal of the American Statistical Association*, 51 (March 1956), 122–31.

[9] Elashoff, R.M. and Afifi, A.A., "Missing Observations in Multivariate Statistics, III, IV," *Journal of the American* Statistical Association, 64 (March 1969), 337–65.

[10] Haitovsky, Y., "Missing Data in Regression Analysis," *Journal of the American Statistical Association*, 63 (March 1968), 67–81.

[11] Hartley, H.O. and Hocking, R.R., "The Analysis of Incomplete Data," *Biometrics*, 27 (December 1971), 783–823.

[12] Hocking, R.R. and Oxspring, H.H., "Maximum Likelihood Estimation with Incomplete Multinomial Data," *Journal of the American Statistical Association*, 66 (March 1971), 65–70.

[13] ——— and Smith, W.B., "Estimation of Parameters in the Multivariate Normal Distributions with Missing Observations," *Journal of the American Statistical Association*, 63 (March 1968), 159–73.

[14] Rubin, D.B., "Missing at Random—What Does It Mean?" Educational Testing Service Research Bulletin, RB-73-2.

[15] ———, "A Noniterative Algorithm for Least Squares Estimation of Missing Values in Any Analysis of Variance Design," *Applied Statistics*, 21 (February 1972), 136–41.

[16] Wilks, S.S., "Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples," *Annals of Mathematical Statistics*, 2 (August 1932), 163–95.