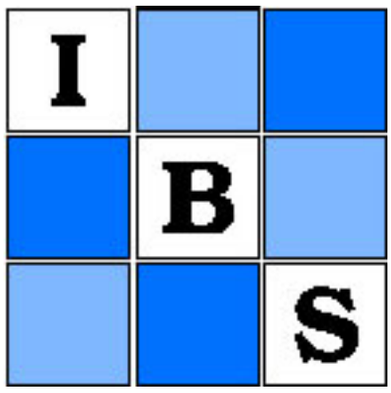

BIOMETRICS *28*, 157–75
March 1972

COVARIANCE SELECTION

A. P. DEMPSTER

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U. S. A.

SUMMARY

The covariance structure of a multivariate normal population can be simplified by setting elements of the inverse of the covariance matrix to zero. Reasons for adopting such a model and a rule for estimating its parameters are given in section 2. It is also proposed to select the zeros in the inverse from sample data. A numerical illustration of the proposed technique is given in section 3. Appendix A sketches the general theory of exponential families which underlies the special results of section 2, and Appendix B describes two approaches to computation of the proposed estimator.

1. INTRODUCTION

Two main currents of thought underlie the covariance fitting technique introduced in this paper. The first is the principle of parsimony in parametric model fitting, which suggests that parameters should be introduced sparingly and only when the data indicate they are required. The second is the exploitation of the powerful and elegant theory of exponential families of distributions, as a tool for practical data analysis. These currents come together in multivariate analysis because the complexity of even the simplest multivariate population models places a premium on the availability of both parameter reduction techniques and relatively simple general theory.

Parameter reduction involves a tradeoff between benefits and costs. If a substantial number of parameters can be set to null values, the amount of noise in a fitted model due to errors of estimation is substantially reduced. On the other hand, errors of misspecification are introduced because the null values are incorrect. Every decision to fit a model involves an implicit balance between these two kinds of errors, i.e., a decision is made not to complicate a model by adding more parameters. However, once a parametric model is adopted, the question of whether or not to thin out the parametric structure is too often settled by default, especially when optimal estimates of the complete set of parameters are easily computed. Such optimality provides no protection against the costs of introducing unnecessary parameters. For example, it is widely recognized that ordinary least squares for multiple regression analysis has many optimal properties, and yet can often be improved by selecting predictors from a full set, thus effectively reducing many regression coefficients to null values of zero.

In this paper, the covariance structure of an assumed multivariate normal

population is studied. If the number of variables p becomes moderate, the number of parameters $\frac{1}{2}p(p + 1)$ in the covariance structure becomes large. For a fixed sample size n, the number of parameters per data point increases like $\frac{1}{2}(p + 1)$ as p increases. The computational ease with which this abundance of parameters can be estimated should not be allowed to obscure the probable unwisdom of such estimation from limited data. How should the data analyst move to reduce the parameter set? One answer comes from considering the natural parameters in the representation of the hypothetical normal populations as an exponential family.

An exponential family consists of densities whose logarithms are linear in the parameters. In symbols, if $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ denotes a multivariate vector of observables, and $f(\mathbf{x}; \boldsymbol{\phi})$ denotes a meaningful probability density function of $\mathbf{x}$ given parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_r)$, then the family of densities is said to be exponential if it is expressible in the form

$$f(\mathbf{x}; \boldsymbol{\phi}) = \exp [\phi + t(\mathbf{x}) + \phi_1 t_1(\mathbf{x}) + \phi_2 t_2(\mathbf{x}) + \cdots + \phi_r t_r(\mathbf{x})], \tag{1}$$

where $t(\mathbf{x})$, $t_1(\mathbf{x})$, $t_2(\mathbf{x})$, $\cdots$, $t_r(\mathbf{x})$ are specific functions of the observable $\mathbf{x}$, and ϕ is a specific function of the parameters $\phi_1, \phi_2, \cdots, \phi_r$ satisfying

$$\int f(\mathbf{x}; \boldsymbol{\phi})\, d\mathbf{x} = 1 \tag{2}$$

or

$$e^{\phi} \times \int \exp [t(\mathbf{x}) + \phi_1 t_1(\mathbf{x}) + \cdots + \phi_r t_r(\mathbf{x})]\, d\mathbf{x} = 1. \tag{3}$$

Note that $\mathbf{x}$ can in principle consist of discrete or continuous variables, or both, and the notation $\int (\cdots)\, d\mathbf{x}$ is being used as a convenient shorthand for multiple sums or integrals, or both.

Exponential families have played a central role in mathematical statistics because optimality properties of tests and estimators are readily available. For example, if $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$ denote a random sample from an unknown member of the family (1), then it is obvious that the likelihood of the sample

$$L(\boldsymbol{\phi}; \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m) = \prod_{l=1}^{m} f(\mathbf{x}_l, \boldsymbol{\phi}) \tag{4}$$

depends only on the statistics $\sum_1^m t_1(\mathbf{x}_l)$, $\sum_1^m t_2(\mathbf{x}_l)$, $\cdots$, $\sum_1^m t_r(\mathbf{x}_l)$, which are therefore sufficient statistics.

Efforts to exploit exponential families as population models for multivariate data have been limited mainly to two special cases, the first being the family of multivariate normal densities widely used for continuous observables. The second instance appears in the recent stress on log linear models for multivariate analysis of categorical variables. In the contingency table context, *log linear* means that the log of the density follows a linear model, which is equivalent to the definition (1) of an exponential family. See Fienberg [1972] and references given there.

Undoubtedly there is scope for the use of more general exponential families in multivariate data analysis. Moreover, the concept of exponential

family is easily generalized by allowing the parameters ϕ to depend linearly on fixed variables, as indicated in Dempster [1971]. There are, however, problems to be faced in fitting more general models, some being statistical problems such as goodness-of-fit assessment or parameter reduction, and others being technical difficulties of carrying out the moment calculations which the model-fitting process requires. Moments which can be represented analytically in the case of normal distributions must be found numerically in general.

The covariance fitting technique of this paper involves the exponential family of normal distributions with unknown covariance structure, represented by the family of continuous densities

$$\left(\frac{1}{2\pi}\right)^{\frac{1}{2}p}\left(\frac{1}{\det \mathbf{\Sigma}}\right)^{\frac{1}{2}} \exp\left(-\tfrac{1}{2}\mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{x}^T\right), \tag{5}$$

where $\mathbf{\Sigma}$ and its inverse $\mathbf{\Sigma}^{-1}$ are both $p \times p$ positive definite symmetric matrices. The (i, j) element σ_{ij} of $\mathbf{\Sigma}$ is the familiar covariance of x_i and x_j, or variance of x_i when $j = i$. The (i, j) element σ^{ij} of $\mathbf{\Sigma}^{-1}$ is the less familiar *concentration* of x_i and x_j (cf. Dempster [1969]). Note that the σ^{ij} play the role of the ϕ_l in (1), and therefore beg consideration as natural parameters of the model. Specifically, (5) can be written in the form (1) where

$$\begin{aligned} r &= \tfrac{1}{2}p(p+1), \\ \phi &= (\sigma^{11}, \sigma^{12}, \cdots, \sigma^{1p}, \sigma^{22}, \sigma^{23}, \cdots, \sigma^{2p}, \sigma^{33}, \sigma^{34}, \cdots, \sigma^{3p}, \cdots, \sigma^{pp}). \\ t_1(\mathbf{x}) &= -\tfrac{1}{2}x_1^2, \; t_2(\mathbf{x}) = -x_1x_2, \cdots, t_r(\mathbf{x}) = -\tfrac{1}{2}x_p^2, \\ t(\mathbf{x}) &= 0, \quad \text{and} \\ \phi &= -\tfrac{1}{2}p \log 2\pi - \tfrac{1}{2} \log \det \mathbf{\Sigma}. \end{aligned} \tag{6}$$

The representation (6) suggests that parameter reduction may reasonably be attempted by setting certain σ^{ij} to 0. More detailed theoretical reasons for the attractiveness of this special type of parameter reduction will be spelled out below. Having decided on a basis set of parameters, there remains a fundamental choice between setting parameters to zero on *a priori* grounds or on grounds that the data provide no evidence that individual σ_{ij} differ from 0. In this paper I follow the latter approach which is more appropriate when *a priori* knowledge is weak.

2. COVARIANCE SELECTION: THEORY

Suppose that $\mathbf{S}$ is an estimated $p \times p$ sample covariance matrix on m D.F., typically computed from a sample of $m + 1$ p-variate observation vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{m+1}$ using the formula

$$\mathbf{S} = \frac{1}{m}\sum_{l=1}^{m+1}(\mathbf{x}_l - \bar{\mathbf{x}})^T(\mathbf{x}_l - \bar{\mathbf{x}}), \tag{7}$$

where

$$\bar{\mathbf{x}} = \frac{1}{m+1} \sum_{l=1}^{m+1} \mathbf{x}_l .$$

Suppose that I denotes a subset of the index pairs (i, j) with $1 \leq i < j \leq p$ and J denotes the remaining set of pairs (i, j) with $1 \leq i \leq j \leq p$. For example, the array of pairs

$$\begin{array}{cccc} (1,1) & (1,2) & (1,3) & (1,4) \\ & (2,2) & (2,3) & (2,4) \\ & & (3,3) & (3,4) \\ & & & (4,4) \end{array}$$

corresponding to the distinct elements of $\mathbf{S}$ or $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^1$ could be partitioned into

$$I = \{(1,2), (1,3), (3,4)\} \quad \text{and}$$

$$J = \{(1,1), (2,2), (3,3), (4,4), (1,4), (2,3), (2,4)\}.$$

I will represent in general the indices of the subset of σ^{ij} parameters to be set to 0. To begin the discussion, it will be assumed that I is fixed.

The following simple recipe is proposed for defining an estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ and a corresponding estimate $\hat{\boldsymbol{\Sigma}}^{-1}$ of $\boldsymbol{\Sigma}^{-1}$:

Rule: Choose $\hat{\boldsymbol{\Sigma}}$ *to be the positive definite symmetric matrix such that* $\mathbf{S}$ *and* $\hat{\boldsymbol{\Sigma}}$ *are identical for index pairs* (i, j) *in* J *while* $\hat{\boldsymbol{\Sigma}}^{-1}$ *is identically* 0 *for index pairs* (i, j) *in* I.

The estimation rule possesses three basic properties (a), (b), (c) which enhance its attractiveness. These properties are now described, with some explanatory remarks. Since it is easier to prove theorems in the setting of a general exponential family, the derivation of the properties (a), (b), (c) is deferred to Appendix A.

(a) *Existence and uniqueness. If there is any positive definite symmetric matrix which agrees with* $\mathbf{S}$ *in the positions* (i, j) *in* J, *then there is exactly one such matrix* $\hat{\boldsymbol{\Sigma}}$ *with the additional property that* $\hat{\boldsymbol{\Sigma}}^{-1}$ *is* 0 *in positions* I. For example, $\mathbf{S}$ itself is usually a positive definite symmetric matrix which agrees with $\mathbf{S}$ in positions J, so that according to property (a) the existence and uniqueness of the estimate $\hat{\boldsymbol{\Sigma}}$ is guaranteed for such $\mathbf{S}$.

(b) *Maximum entropy model. Among all normal models* (5) *such that* $\boldsymbol{\Sigma}$ *agrees with* $\mathbf{S}$ *over the indices* J, *the special choice* $\hat{\boldsymbol{\Sigma}}$ *has maximum entropy.* In general, the entropy of a distribution specified by the density $f(\mathbf{x}; \boldsymbol{\phi})$ is defined to be

$$-\int f(\mathbf{x}; \boldsymbol{\phi}) \log f(\mathbf{x}; \boldsymbol{\phi}) \, d\mathbf{x}. \tag{9}$$

Entropy is a measure of smoothness or simplicity in a distribution. For example, among all discrete distributions over a finite number of cells, the uniform distribution has maximum entropy, or among all continuous distri-

butions with a given mean and variance, the normal distribution has maximum entropy (cf. Rao [1965]). These are good examples of smooth and simple distributions characterized by maximizing entropy. Thus (cf. Good [1963]), the principle of seeking maximum entropy is a principle of seeking maximum simplicity of explanation. In the normal covariance example, the integration (9) can be carried out analytically on the density (5) to yield

$$\tfrac{1}{2}p \log 2\pi + \tfrac{1}{2} \log \det \mathbf{\Sigma} + \tfrac{1}{2}p, \tag{10}$$

so that entropy is essentially $\log \det \mathbf{\Sigma}$. Since

$$\det \mathbf{\Sigma} = [\sigma_{11}\sigma_{22} \cdots \sigma_{pp}] \times \det \mathbf{R}, \tag{11}$$

where $\mathbf{R}$ is the correlation matrix, and since σ_{11} , σ_{22} , $\cdots$, σ_{pp} are held fixed at s_{11} , s_{22} , $\cdots$, s_{pp} in the estimation procedure, the principle of maximum entropy asserts that the choice of $\mathbf{\Sigma}$ which produces zeros in the elements I of $\mathbf{\Sigma}^{-1}$ is also the $\mathbf{\Sigma}$ which maximizes $\det \mathbf{R}$. Since $\det \mathbf{R}$ is a measure of overall correlation (cf. Dempster [1969]), the principle is also a principle of minimum overall correlation.

(c) *Maximum likelihood* (ML) *estimation. Among all normal models* (5) *such that the elements of* $\mathbf{\Sigma}^{-1}$ *in positions* I *are all* 0, *the special choice* $\hat{\mathbf{\Sigma}}$ *is the* ML *estimate of* $\mathbf{\Sigma}$. By writing down the likelihood it is trivial to see that the elements of $\mathbf{S}$ in the J positions are sufficient statistics for the restricted estimation problem. Property (c), which is less trivial, shows how to use these sufficient statistics to produce ML estimates.

Given a decision to fit a covariance matrix $\mathbf{\Sigma}$ using only a subset J of the elements of $\mathbf{S}$, it might at first appear more natural to propose as an estimate the matrix $\hat{\hat{\mathbf{\Sigma}}}$ which agrees with $\mathbf{S}$ in positions J and whose remaining elements are simply zero. $\hat{\hat{\mathbf{\Sigma}}}$ suffers the disadvantage that it may not be positive definite, meaning that it may not be a valid covariance matrix. Second, although $\hat{\hat{\mathbf{\Sigma}}}$ reduces certain estimated correlation coefficients to null values, it does not minimize the overall correlation measure $\det \mathbf{R}$. Finally, the elements of $\mathbf{S}$ corresponding to indices in J are not in general sufficient statistics for the model in which $\mathbf{\Sigma}$ is 0 in positions I, so that $\hat{\hat{\mathbf{\Sigma}}}$ is not an efficient estimate for the corresponding model. None of this proves that the model with 0 in positions I of $\mathbf{\Sigma}$ is not empirically more correct than the model with 0 in positions I of $\mathbf{\Sigma}^{-1}$, but in the absence of firm prior knowledge favoring the former, the theoretical advantages of the latter suggest it be given priority.

Extensive iterative computations are required to produce $\hat{\mathbf{\Sigma}}$ from $\mathbf{S}$, and several alternative approaches are available. One tack is to pass through a sequence $\mathbf{S} = \mathbf{S}^{(0)} \to \mathbf{S}^{(1)} \to \mathbf{S}^{(2)} \to \cdots$, where each $\mathbf{S}^{(k)}$ agrees with $\mathbf{S}$ and $\hat{\mathbf{\Sigma}}$ in the positions J and is such that the elements of $\mathbf{S}^{(k)-1}$ in positions I are being driven to 0 as k increases. An alternative is to pass through a sequence in which the I elements of the inverse are held constant at 0 while the J elements of the covariance matrix are driven to the corresponding values in $\mathbf{S}$. Within each of these approaches there is a choice among shifting one, several, or all variable elements in a single iteration. More detail on

computational theory, much of it applicable to exponential families generally, will be given in Appendix B.

Data-based rules for selecting the subsets I and J can be defined in various ways analogous to the various forward and backward procedures used for selecting predictor variables in multiple regression analysis (cf. Draper and Smith [1966]). A forward approach means beginning with I empty and successively adding pairs (i, j) to I until such time as a larger I appears not to improve fit significantly. A backward approach means beginning with $\hat{\mathbf{\Sigma}} = \mathbf{S}$, corresponding to I consisting of all off-diagonal elements, and then dropping pairs (i, j) from I one at a time as long as the decrease in fit is not significantly large. Exact tests of significance are not available, but several approximate tests are easily devised. For example, the change in 2 log likelihood when another parameter is added can be regarded as roughly a χ^2 variable on 1 D.F. Alternatively, the estimate $\hat{\sigma}^{ij}$ of an added exponential parameter can be divided by an estimate of its standard deviation and treated as a standard normal deviate. The two tests are asymptotically equivalent.

3. COVARIANCE SELECTION: EXAMPLE

The technique proposed in section 2 will now be illustrated numerically on the 6×6 covariance matrix S:

14.029	5.6635	1.9866	2.733	−4.867	2.0744
5.6635	14.537	0.1271	−1.347	0.206	−1.5747
1.9866	0.1271	2.068	0.294	−0.5446	−0.644
2.733	−1.347	0.294	17.11	−5.42	0.885
−4.867	0.206	−0.5446	−5.42	7.87	−1.933
2.0744	1.5747	−0.644	0.885	−1.933	3.552

used for different illustrative purposes in Cochran [1938] and Dempster [1969]. The data refer to a nocturnal insect trap. The 6th variable is the log of an insect count plus 1, while the other 5 variables measure weather conditions. S has 72 D.F. coming from successive days after removal of certain linear cycle effects.

The estimation procedures of section 2 use the sample variance of each of the p variables to estimate the corresponding population variance, whatever the choice of the subsets I and J. Moreover, the procedures produce equivalent results under linear changes of scale of the p variables. Accordingly, there is no loss of generality in working with the correlation matrix R:

1	0.396583	0.368826	0.176401	−0.463192	0.293861
0.396583	1	0.023181	−0.0854093	0.0192594	0.219141
0.368826	0.023181	1	0.049425	−0.134994	−0.237615
0.176401	−0.0854093	0.049425	1	−0.467075	0.113522
−0.463192	0.0192594	−0.134994	−0.467075	1	−0.365602
0.293861	0.219141	−0.237615	0.113522	−0.365602	1

computed from S.

A forward selection procedure was used which started with J consisting only of the diagonal elements, and successively added off-diagonal elements to J one at a time. At each stage in the selection procedure the correlation

matrix was re-estimated using the direct Newton-type algorithm described in Appendix B. The starting point in the iterations was taken to be the fitted correlation matrix from the previous stage. The decision rule for choosing the next off-diagonal element for inclusion in J was to carry out the first iteration of the fitting procedure for all of the candidates, and to compute a crude t statistic for each candidate as indicated in Appendix B. The selected parameters and corresponding approximate χ^2 from 2 log likelihood are shown in Table 1.

The first 5 selected parameters appear to be clearly significant, while the 6th is borderline. Note that at stage 6, the choice is made among 10 possible parameters, so that an overall significance level of 0.05 would require very roughly that the largest χ^2 be significant at level $0.05 \div 10$. By this standard, the 6th and later stages are introducing estimates which cannot be judged to differ from random noise.

The first 8 fitted correlation matrices are reproduced overleaf.

Note that by stage 5 all 15 of the estimated correlation coefficients are nonzero, even though only 5 parameters are in the model.

The above procedure is presented here as a speculative technique meriting further study. Simulation studies will be required to check out the improvement in estimating correlation coefficients and regression coefficients which can be achieved from covariance selection. If, as seems likely from preliminary

TABLE 1

SELECTED OFF-DIAGONAL ELEMENTS IN ORDER, WITH CORRESPONDING INCREASES IN 2 LOG LIKELIHOOD

Stage	Selected pair (i,j)	χ^2
1	(4,5)	17.72
2	(1,5)	17.39
3	(1,2)	12.32
4	(1,3)	10.53
5	(5,6)	10.33
6	(3,6)	7.10
7	(1,6)	6.40
8	(2,5)	4.63
9	(2,6)	2.88
10	(2,3)	.843
11	(2,4)	.540
12	(4,6)	.182
13	(3,5)	.116
14	(3,4)	.072
15	(1,4)	.00004

Stage 1:

1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	−0.467075	0
0	0	0	0	−0.467075	1	0
0	0	0	0	0	0	1

Stage 2:

1	0	0	0	0.216345	−0.463192	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0.216345	0	0	1	−1	−0.467075	0
−0.463192	0	0	0	−0.467075	1	0
0	0	0	0	0	0	1

Stage 3:

1	0.396583	0	0.216345	0.463192	0	
0.396583	1	0	0.0857989	−0.183694	0	
0	0	1	0	0	0	
0.216345	0.0857989	0	−1	−0.467075	0	
−0.463192	−0.183694	0	−0.467075	1	0	
0	0	0	0	0	1	

Stage 4:

1	0.396583	0.368826	0.216345	−0.463192	0	
0.396583	1	0.14627	0.0857989	−0.183694	0	
0.368826	0.14627	1	0.0797938	−0.170837	0	
0.216345	0.0857989	0.0797938	1	−0.467075	0	
−0.463192	−0.183694	−0.170837	−0.467075	1	0	
0	0	0	0	0	1	

Stage 5:

1	0.396583	0.368826	0.216345	-0.463192	0.169344
0.396583	1	0.14627	0.0857989	-0.183694	0.0671588
0.368826	0.14627	1	0.0797938	-0.170837	0.0624583
0.216345	0.0857989	0.0797938	1	-0.467075	0.170763
-0.463192	-0.183694	-0.170837	-0.467075	1	-0.365602
0.169344	0.0671588	0.0624583	0.170763	-0.365602	1

Stage 6:

1	0.396583	0.368826	0.216345	-0.463192	0.0802061
0.396583	1	0.14627	0.0857989	-0.183694	0.0318084
0.368826	0.14627	1	0.0385374	-0.0825078	-0.237615
0.216345	0.0857989	0.0385374	1	-0.467075	0.170763
-0.463192	-0.183694	-0.0825078	-0.467075	1	-0.365602
0.0802061	0.0318084	-0.237615	0.170763	-0.365602	1

Stage 7:

1	0.396583	0.368826	0.216345	-0.463192	0.293861
0.396583	1	0.14627	0.0857989	-0.183694	0.11654
0.368826	0.14627	1	0.0392016	-0.08393	-0.237615
0.216345	0.0857989	0.0392016	1	-0.467075	0.170763
-0.463192	-0.183694	-0.08393	-0.467075	1	-0.365602
0.293861	0.11654	-0.237615	0.170763	-0.365602	1

Stage 8:

1	0.396583	0.368826	0.216345	-0.463192	0.293861
0.396583	1	0.168726	-0.00899558	-0.0192594	-0.0572433
0.368826	-0.168726	1	0.0392016	-0.08393	-0.237615
0.216345	-0.00899558	0.0392016	1	-0.467075	0.170763
-0.463192	0.0192594	-0.08393	-0.467075	1	-0.365602
0.293861	0.0572433	-0.237615	0.170763	-0.365602	1

studies, the improvement can be substantial, the question may be raised: is it ever wise to simply use a sample correlation matrix or a sample covariance matrix as an estimate of the corresponding population quantity?

ACKNOWLEDGMENTS

This work was facilitated by grants GP-8774 and GP-19182 from the National Science Foundation, and by a joint study agreement between Harvard University and IBM Cambridge Scientific Center. The author wishes to thank Nanny Wermuth for carrying out the computations reported in section 3.

SELECTION DE COVARIANCE

RESUME

La structure de la covariance d'une population normale multivariate peut se simplifier en imposant à des éléments de l'inverse de la matrice de covariance d'être égaux à zéro. Des raisons pour adopter un tel modèle et une règle pour estimer ses paramètres sont données dans la section 2.

On propose aussi de sélectionner les zéros de l'inverse à partir des données de l'échantillon. A la section 3 on donne une illustration numérique de la technique proposée. L'appendice A esquisse la théorie générale des familles exponentielles sous-jacentes aux résultats particuliers de la section 2, et l'appendice B décrit deux approches pour calculer l'estimateur proposé.

REFERENCES

Cochran, W. G. [1938]. The omission or addition of an independent variate in multiple linear regression. *J. R. Statist. Soc. Suppl., 5*, 171–6.

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

Dempster, A. P. [1971]. An overview of multivariate data analysis. *J. Multivariate Analysis 1*, 316–46.

Draper, N. and Smith, H. [1966]. *Applied Regression Analysis.* Wiley, New York.

Fienberg, S. E. [1972]. The analysis of incomplete multiway contingency tables. *Biometrics 28* (to appear).

Good, I. J. [1963]. Maximum entropy for null hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist. 34*, 911–34.

Ireland, C. T. and Kullback, S. [1968]. Contingency tables with given marginals. *Biometrika 55*, 179–88.

Rao, C. R. [1965]. *Linear Statistical Inference and its Applications.* Wiley, New York.

APPENDIX A: STATISTICAL THEORY

Several basic properties of the general exponential family (1) will now be derived and the usefulness of these properties will be illustrated by application to the special family (5).

Alongside the exponential parameters $\phi = (\phi_1, \phi_2, \cdots, \phi_r)$ of the family (1), it is convenient to consider the *moment parameters* $\theta = (\theta_1, \theta_2, \cdots, \theta_r)$ defined by

$$\theta_i = \int t_i(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\phi})\, d\mathbf{x}. \tag{A1}$$

It will be assumed further that the variances of and covariances among $t_1(\mathbf{x})$, $t_2(\mathbf{x})$, $\cdots$, $t_r(\mathbf{x})$, namely the

$$\gamma_{ij} = \int [t_i(\mathbf{x}) - \theta_i][t_j(\mathbf{x}) - \theta_j] f(\mathbf{x}; \boldsymbol{\phi})\, d\mathbf{x}, \tag{A2}$$

are all finite, and that the $r \times r$ covariance matrix $\boldsymbol{\Gamma}$ with (i, j) element γ_{ij} is positive definite throughout the family. The positive definiteness assures that no linear combination of $t_1(\mathbf{x})$, $t_2(\mathbf{x})$, $\cdots$, $t_r(\mathbf{x})$ is constant and therefore that distinct $\boldsymbol{\phi}$ determine distinct members of the family. Similarly, the positive definiteness of $\boldsymbol{\Gamma}$ is sufficient to assure, as implied by Lemma A below, that distinct members of the family determine distinct $\boldsymbol{\theta}$. The expression (1) may be integrable, and therefore a valid density, only for a restricted set of vectors $\boldsymbol{\phi}$, and similarly only certain vectors $\boldsymbol{\theta}$ may result from the definition (A1). Accordingly, the positive definiteness of $\boldsymbol{\Gamma}$ implies one-to-one correspondences among the points of three mathematical spaces, namely the restricted set of valid $\boldsymbol{\phi}$, the restricted set of possible $\boldsymbol{\theta}$, and the exponential family (1) itself. It is easily checked that the set of $\boldsymbol{\phi}$ vectors and the set of $\boldsymbol{\theta}$ vectors are both convex sets.

In the special case of the family (5), the exponential parameters defined in (6) are the distinct elements σ^{ij} of $\boldsymbol{\Sigma}^{-1}$. The $t_i(\mathbf{x})$ explicitly displayed in (6) have easily computed expected values, namely

$$\boldsymbol{\theta} = (-\tfrac{1}{2}\sigma_{11}\,, -\sigma_{12}\,, \cdots, -\sigma_{1p}\,, -\tfrac{1}{2}\sigma_{22}\,, -\sigma_{23}\,, \cdots, -\sigma_{2p}\,, \cdots, -\tfrac{1}{2}\sigma_{pp}). \tag{A3}$$

Thus $\boldsymbol{\theta}$ consists of the distinct elements of the covariance matrix $\boldsymbol{\Sigma}$ except that the variances have the factors $-\frac{1}{2}$ and the covariances have the factors -1. The covariance γ_{ij} of $t_i(\mathbf{x})$ and $t_j(\mathbf{x})$ can be deduced easily from the standard formula (cf. Dempster [1969] p. 318):

$$\operatorname{cov}(x_i x_j\,, x_k x_l) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}\,, \tag{A4}$$

where again the factors $-\frac{1}{2}$ and -1 must be applied in various combinations. The family (5) is defined for all positive definite symmetric $\boldsymbol{\Sigma}$, or equivalently for all $\boldsymbol{\phi}$ such that the associated symmetric matrix with (i, j) element σ^{ij} is positive definite. The restriction on $\boldsymbol{\theta}$ is then simply the positive definiteness of $\boldsymbol{\Sigma}$. The one-to-one correspondences among $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}$, and the family (5) are well known for the case of normal distributions. Moreover, it is easily checked that $\boldsymbol{\Gamma}$ is positive definite within the family (5), for example, because no quadratic function of x_1 , x_2 , $\cdots$, x_p is constant under any distribution in the family.

An important property of the general model is that $\boldsymbol{\Gamma}$ provides the partial derivatives of the $\boldsymbol{\theta}$ parameters with respect to the $\boldsymbol{\phi}$ parameters, i.e.

$$\gamma_{ij} = \partial\theta_i/\partial\phi_j\,. \tag{A5}$$

To see this, differentiate (2) and (A1) after substituting from (1), to obtain

$$d\phi + \left[\int t_1(\mathbf{x})f(\mathbf{x};\phi)\,d\mathbf{x}\right]d\phi_1 + \cdots + \left[\int t_r(\mathbf{x})f(\mathbf{x};\phi)\,d\mathbf{x}\right]d\phi_r = 0 \qquad \text{(A6)}$$

and

$$\left[\int t_i(\mathbf{x})f(\mathbf{x};\phi)\,d\mathbf{x}\right]d\phi + \left[\int t_i(\mathbf{x})t_1(\mathbf{x})f(\mathbf{x};\phi)\,d\mathbf{x}\right]d\phi_1 + \cdots + \left[\int t_i(\mathbf{x})t_r(\mathbf{x})f(\mathbf{x};\phi)\,d\mathbf{x}\right]d\phi_r = d\theta_i\,. \qquad \text{(A7)}$$

Substituting from (A6) into (A7) yields

$$d\theta_i = \gamma_{i1}\,d\phi_1 + \gamma_{i2}\,d\phi_2 + \cdots + \gamma_{ir}\,d\phi_r\,, \qquad \text{(A8)}$$

which is equivalent to (A5). Note that, because $\boldsymbol{\Gamma}$ is symmetric,

$$\partial\theta_i/\partial\phi_j = \partial\theta_j/\partial\phi_i\,. \qquad \text{(A9)}$$

The fitting procedure described in section 2 suggests that the theory of exponential families be developed in relation to a partition of the parameter set into two classes. In general, suppose that $\phi = (\phi_1, \phi_2, \cdots, \phi_r)$ is written as

$$\phi = (\phi_1, \phi_2) \qquad \text{(A10)}$$

where $\phi_1 = (\phi_1, \phi_2, \cdots, \phi_s)$ and $\phi_2 = (\phi_{s+1}, \phi_{s+2}, \cdots, \phi_r)$, and correspondingly suppose that

$$\theta = (\theta_1, \theta_2), \qquad \text{(A11)}$$

where θ_1 and θ_2 are $1 \times s$ and $1 \times (r - s)$ vectors. The estimation rule of section 2 depends on a solution of the following type of problem:

Find a member of the family (1) *whose exponential parameters have a prespecified* ϕ_2 *and whose moment parameters simultaneously have a prespecified* θ_1.

For example, in section 2 the fitted member of the family (5) was made to agree with the moment parameters of the sample covariance in positions I and to have exponential parameters 0 in positions J. A general version of this fitting procedure, based on a sample $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$ from the density (1), would be to match θ_1 with the estimates

$$\hat{\theta}_1 = \left[\frac{1}{m}\sum_{l=1}^{m} t_1(\mathbf{x}_l), \frac{1}{m}\sum_{l=1}^{m} t_2(\mathbf{x}_l), \cdots, \frac{1}{m}\sum_{l=1}^{m} t_s(\mathbf{x}_l)\right] \qquad \text{(A12)}$$

while setting $\phi_2 = \mathbf{0}$, where $\mathbf{0}$ denotes a vector of zeros. Although this type of estimate is of greater practical importance, it is of interest to note that there exist situations of a reverse type where ϕ_2 is specified from observed data and θ_1 is fixed on *a priori* grounds. See Ireland and Kullback [1968] for a discussion of the latter type in the context of contingency tables, where exponential interaction parameters of an observed table are held constant while the observed table is modified into a fitted table with prespecified margins. Here margins are moment parameters.

The problem of simultaneously matching θ_1 and ϕ_2 gives rise to important mathematical theory which will be summarized in three lemmas related to the properties (a), (b), and (c) of section 2. The following notation will be used: $g(\mathbf{x})$ denotes a member of the family (1) which has a prespecified ϕ_2 ; $h(\mathbf{x})$ denotes a member of the family (1) which has a prespecified θ_1 and is distinct from $g(\mathbf{x})$; $k(\mathbf{x})$ denotes a member of the family (1) which possesses both ϕ_2 and θ_1 .

Lemma A. Under fairly general circumstances $k(\mathbf{x})$ exists given $g(\mathbf{x})$ and $h(\mathbf{x})$. When $k(\mathbf{x})$ exists, it is unique.

Lemma B. When $k(\mathbf{x})$ exists,

$$\int k(\mathbf{x}) \log \frac{g(\mathbf{x})}{k(\mathbf{x})}\, d\mathbf{x} > \int h(\mathbf{x}) \log \frac{g(\mathbf{x})}{h(\mathbf{x})}\, d\mathbf{x}, \tag{A13}$$

i.e., among all members $h^*(\mathbf{x})$ of the family (1) with the given θ_1 , a unique maximum of

$$\int h^*(\mathbf{x}) \log [g(\mathbf{x})/h^*(\mathbf{x})]\, d\mathbf{x} \tag{A14}$$

is attained when $h^*(\mathbf{x}) = k(\mathbf{x})$.

Lemma C. When $k(\mathbf{x})$ exists,

$$\int h(\mathbf{x}) \log \frac{k(\mathbf{x})}{h(\mathbf{x})}\, d\mathbf{x} > \int h(\mathbf{x}) \log \frac{g(\mathbf{x})}{h(\mathbf{x})}\, d\mathbf{x}, \tag{A15}$$

i.e., among all members $g^*(\mathbf{x})$ of the family (1) with the given ϕ_2 , a unique maximum of

$$\int h(\mathbf{x}) \log [g^*(\mathbf{x})/h(\mathbf{x})]\, d\mathbf{x} \tag{A16}$$

is attained when $g^*(\mathbf{x}) = k(\mathbf{x})$.

Before deriving these results, the details of their application to the normal family (5) will be spelled out. Lemma A is intentionally vague, because detailed conditions on the existence of $k(\mathbf{x})$ vary from example to example. In Appendix B, however, it will be shown by a computing algorithm that $k(\mathbf{x})$ always exists for the family (5) whenever there exists any positive-definite $\mathbf{\Sigma}$ which agrees with $\mathbf{S}$ in positions J. The second part of Lemma A then assures uniqueness. It is worth noting in passing that the derivation of mathematical conditions for the existence of $k(\mathbf{x})$ is of relatively little practical importance, because in practice the standard computing algorithms are guaranteed to find $k(\mathbf{x})$ when it exists and one knows immediately when $k(\mathbf{x})$ has been found by checking out θ_1 and ϕ_2 . The failure of $k(\mathbf{x})$ to exist is generally an indication that the proposed model cannot fit the data, so that some other analysis is desirable on scientific grounds.

The expression (A14) can be written

$$-\int h^*(\mathbf{x}) \log h^*(\mathbf{x})\, d\mathbf{x} + \int h^*(\mathbf{x}) \log g(\mathbf{x})\, d\mathbf{x}. \tag{A17}$$

The second term does not vary as $h^*(\mathbf{x})$ varies provided $\phi_2 = \mathbf{0}$ and $t(\mathbf{x}) = 0$, because $\log g(\mathbf{x})$ involves terms in $t_1(\mathbf{x}), \cdots, t_s(\mathbf{x})$ but not terms in $t_{s+1}(\mathbf{x}), \cdots, t_r(\mathbf{x})$, or $t(\mathbf{x})$, and the expectations of $t_1(\mathbf{x}), \cdots, t_s(\mathbf{x})$, namely $\theta_1, \cdots, \theta_s$, are constant as $h^*(\mathbf{x})$ varies. Thus, to maximize (A14) is to maximize the first term which is entropy. This explains how Lemma B implies property (b) of section 2. Similarly, when (A16) is written

$$\int h(\mathbf{x}) \log g^*(\mathbf{x})\, d\mathbf{x} - \int h(\mathbf{x}) \log h(\mathbf{x})\, d\mathbf{x}, \tag{A18}$$

it is evident that the first term is log likelihood divided by m, while the second term does not depend on $g^*(\mathbf{x})$, so that Lemma C implies the property (c) of section 2.

The first part of Lemma A can be proved by defining a path through densities with the prespecified θ_1 beginning at $h(\mathbf{x})$ and ending at the desired $k(\mathbf{x})$. The condition that θ_1 is constant along the path implies from (A5) that the differential $d\phi = (d\phi_1, d\phi_2)$ along the path satisfies

$$d\phi_1 = -d\phi_2[\Gamma_{21}\Gamma_{11}^{-1}], \tag{A19}$$

where

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \tag{A20}$$

denotes the partition of (A2) into $s + (r - s)$ rows and columns. If a curve is defined from the initial ϕ_2 associated with $h(\mathbf{x})$ to the desired ϕ_2 associated with $k(\mathbf{x})$, and the corresponding motion in ϕ implied by (A19) remains within the space of permissible ϕ, then $k(\mathbf{x})$ has been demonstrated to exist. As already remarked, such existence cannot be proved in complete generality, but is shown for the normal family (5) in Appendix B. To prove the uniqueness of $k(\mathbf{x})$ when it exists, suppose to the contrary that distinct $k(\mathbf{x})$ and $\tilde{k}(\mathbf{x})$ possess the given θ_1 and ϕ_2 but different ϕ_1 vectors, say ϕ_1 and $\tilde{\phi}_1$. Differential motion along the line segment from (ϕ_1, ϕ_2) to $(\tilde{\phi}_1, \phi_2)$ implies a nonzero $d\phi_1$ of constant direction, which in turn implies that the elements of $d\theta_1 = d\phi_1\Gamma_{11}$ have constant signs and are not all zero which in turn implies that some of the elements of θ_1 must be different at different ends of the line segment, a contradiction.

Lemmas B and C are both simple corollaries of the familiar inequality (cf. Rao [1965])

$$-\int h(\mathbf{x}) \log h(\mathbf{x})\, d\mathbf{x} < -\int h(\mathbf{x}) \log h^*(\mathbf{x})\, d\mathbf{x} \tag{A21}$$

for any two distinct densities $h(\mathbf{x})$ and $h^*(\mathbf{x})$, so that for fixed $h(\mathbf{x})$ the expression $-\int h(\mathbf{x}) \log h^*(\mathbf{x})\, d\mathbf{x}$ achieves a unique minimum when $h^*(\mathbf{x}) = h(\mathbf{x})$. Lemma B follows from the relations

$$\int h(\mathbf{x})[\log h(\mathbf{x}) - \log g(\mathbf{x})]\, d\mathbf{x} > \int h(\mathbf{x})[\log k(\mathbf{x}) - \log g(\mathbf{x})]\, d\mathbf{x}$$

$$= \int k(\mathbf{x})[\log k(\mathbf{x}) - \log g(\mathbf{x})]\, d\mathbf{x}. \qquad \text{(A22)}$$

The inequality in (A22) is an application of (A21) while the equality follows because $\log k(\mathbf{x}) - \log g(\mathbf{x})$ has nonzero coefficients along the constant term and $t_1(\mathbf{x})$, $t_2(\mathbf{x})$, $\cdots$, $t_s(\mathbf{x})$ whose expectations $\boldsymbol{\theta}_1$ are identical under both $h(\mathbf{x})$ and $k(\mathbf{x})$. To prove Lemma C, a different application of (A21) is required, namely

$$-\int k(\mathbf{x}) \log k(\mathbf{x})\, d\mathbf{x} < -\int k(\mathbf{x}) \log g(\mathbf{x})\, d\mathbf{x}. \qquad \text{(A23)}$$

Using the equality in (A22), the inequality (A23) is expressible as

$$-\int h(\mathbf{x}) \log k(\mathbf{x})\, d\mathbf{x} < -\int h(\mathbf{x}) \log g(\mathbf{x})\, d\mathbf{x} \qquad \text{(A24)}$$

which in turn is equivalent to (A15), as required.

The maximizations in Lemmas B and C can be corroborated by computing derivatives of the quantities (A14) and (A16). Suppose that (A14) is denoted B^* and the parameters associated with $h^*(\mathbf{x})$ are denoted $\boldsymbol{\phi}^*$, $\boldsymbol{\theta}^*$, and $\boldsymbol{\Gamma}^*$. If $h^*(\mathbf{x})$ undergoes differential motion obeying $d\boldsymbol{\phi}_1^* = -d\boldsymbol{\phi}_2^* \boldsymbol{\Gamma}_{21}^* \boldsymbol{\Gamma}_{11}^{*-1}$ as in (A19) to keep $\boldsymbol{\theta}_1^*$ constant, it is easily checked that

$$dB^* = d\boldsymbol{\phi}_2^* [\boldsymbol{\Gamma}_{22}^* - \boldsymbol{\Gamma}_{21}^* \boldsymbol{\Gamma}_{11}^{*-1} \boldsymbol{\Gamma}_{12}^*](\boldsymbol{\phi}_2 - \boldsymbol{\phi}_2^*)^T, \qquad \text{(A25)}$$

where $\boldsymbol{\phi}_2$ is the fixed parameter set associated with $g(\mathbf{x})$. From (A25) it is seen that B^* has zero derivatives when $\boldsymbol{\phi}_2 = \boldsymbol{\phi}_2^*$, i.e. at the maximum indicated by Lemma B. Similarly, if (A16) is denoted C^* and the parameters associated with $g^*(\mathbf{x})$ in Lemma C are denoted $\boldsymbol{\phi}^*$, $\boldsymbol{\theta}^*$, and $\boldsymbol{\Gamma}^*$, then differential changes in $g^*(\mathbf{x})$ keeping $\boldsymbol{\phi}_2^*$ constant at the prespecified $\boldsymbol{\phi}_2$ yield

$$dC^* = d\boldsymbol{\phi}_1^* (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^*)^T \qquad \text{(A26)}$$

which shows that $dC^* = 0$ when $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*$, i.e. at the maximum indicated by Lemma C.

Finally, the first derivative relation (A25) and (A26) can be differentiated trivially to show that the matrix of second partial derivatives of B^* with respect to the elements of $\boldsymbol{\phi}_2^*$ is

$$-(\boldsymbol{\Gamma}_{22}^* - \boldsymbol{\Gamma}_{21}^* \boldsymbol{\Gamma}_{11}^{*-1} \boldsymbol{\Gamma}_{12}^*), \qquad \text{(A27)}$$

and the matrix of second partial derivatives of C^* with respect to the elements of $\boldsymbol{\phi}_1^*$ is

$$-\boldsymbol{\Gamma}_{11}^* . \qquad \text{(A28)}$$

It is an interesting property of exponential families that $\boldsymbol{\Gamma}_{11}^*$ provides both first derivatives of $\boldsymbol{\theta}_1^*$ with respect to $\boldsymbol{\phi}_1^*$ as shown by (A5) and second de-

rivatives C^* as shown by (A28). One application of this coincidence is given in Appendix B. Another important application of (A28) is to ML estimation. As remarked above, the estimation rule of section 2 illustrates ML estimation within the subfamily of (1) with $\phi_2 = \mathbf{0}$ where θ_1 is a vector of sufficient estimators. Since mC^* is log likelihood, it follows that $-m\Gamma_{11}^*$ is the matrix of second derivatives of log likelihood with respect to the parameters ϕ_1^*. Thus $(1/m)\Gamma_{11}^{-1}$ calculated at the ML estimates provides an estimate of the asymptotic covariance matrix of the estimated ϕ_1 , which in turn implies from (A5) that $(1/m)\Gamma_{11}$ is an approximate asymptotic covariance matrix for the estimated θ_1 . The latter result can also be verified directly, for if Γ_{11} could be calculated at the true parameter values, it would give exactly the sampling covariance matrix of $t_1(\mathbf{x})$, $t_2(\mathbf{x})$, $\cdots$, $t_s(\mathbf{x})$, so that $(1/m)\Gamma_{11}$ defines the covariance matrix of the sufficient estimates $(1/m)\sum_1^m t_1(\mathbf{x}_l)$, $\cdots$, $(1/m)\sum_1^m t_s(\mathbf{x}_l)$ of θ_1 .

After carrying out the fitting procedure matching θ_1 with the sufficient estimates and matching ϕ_2 with $\mathbf{0}$, any element of the fitted ϕ_1 can be tested approximately for significant difference from zero by dividing by the square root of the corresponding diagonal element of $(1/m)\Gamma_{11}^{-1}$ and comparing to a standard normal deviate. An asymptotically equivalent test is to treat twice the reduction in log likelihood when adding the parameter to the model as a χ^2 on 1 D.F. Either of these tests can be used as a basis for parameter selection.

APPENDIX B: COMPUTATIONAL THEORY

Two specific iterative procedures of the form $\phi^{(0)} \to \phi^{(1)} \to \phi^{(2)} \to \cdots$ will be described which converge to the ϕ defining a member of the family (1) specified by given θ_1 and ϕ_2 as in (A10) and (A11). In the first procedure, the partitions $\phi^{(i)} = (\phi_1^{(i)}, \phi_2^{(i)})$ are characterized by holding $\phi_2^{(i)}$ at the desired ϕ_2 , so that only $\phi_1^{(i)}$ changes with i in such a way that the corresponding $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)})$ has $\theta_1^{(i)}$ converging to the desired θ_1 . In the second procedure $\theta_1^{(i)}$ is held fixed at the desired θ_1 which means that both $\phi_1^{(i)}$ and $\phi_2^{(i)}$ change with i as $\phi_2^{(i)}$ converges to the desired ϕ_2 .

The first procedure is most simply regarded as a straightforward application of Newton's method for solving implicit equations. With ϕ_2 fixed, θ_1 can be regarded as a function $\theta_1(\phi_1)$ and the problem is to solve the equations $\theta_1 = \theta_1(\phi_1)$ for ϕ_1 given θ_1 . Having $\theta_1^{(i)} = \theta_1(\phi_1^{(i)})$ at stage i, one expands θ_1 in a Taylor series about $\phi_1^{(i)}$ to obtain

$$\theta_1 = \theta_1^{(i)} + (\phi_1 - \phi_1^{(i)})\Gamma_{11}^{(i)} + \cdots , \tag{B1}$$

where $\Gamma_{11}^{(i)}$ denotes the Γ_{11} part of (A20) calculated at $\phi^{(i)}$. Newton's method says one should define $\phi_1^{(i+1)}$ by solving the first term equation (B1) to obtain

$$\phi_1^{(i+1)} = \phi_1^{(i)} + (\theta_1 - \theta_1^{(i)})\Gamma_{11}^{(i)-1}. \tag{B2}$$

The process (B2) may also be regarded as an ascent procedure based on a quadratic approximation to the log likelihood. Denoting log likelihood at

ϕ_1 and $\phi_1^{(i)}$ by mC and $mC^{(i)}$, respectively, it follows from (A26) and (A28) that a two-term Taylor series expansion of mC about $\phi_1^{(i)}$ is given by

$$mC = m[C^{(i)} + (\phi_1 - \phi_1^{(i)})(\theta_1 - \theta_1^{(i)})^T - \tfrac{1}{2}(\phi_1 - \phi_1^{(i)})\Gamma_{11}^{(i)}(\phi_1 - \phi_1^{(i)})^T + \cdots]. \tag{B3}$$

It is easily checked that the maximum of the quadratic expression (B3) occurs when $\phi_1 = \phi_1^{(i+1)}$ as defined in (B2). The process (B2) is not guaranteed to converge as it stands, but from (B3) it follows that the change vector $\phi^{(i+1)} - \phi^{(i)}$ does set out in a direction such that the likelihood is increasing in a neighborhood of $\phi^{(i)}$. Consequently, even if the full step (B2) should reduce likelihood, a shortened step in the same direction will increase likelihood. Thus, a minor modification of the process (B2) yields a monotone sequence of increasing likelihoods which must converge to the unique maximum when it exists.

The computations reported in section 2 were developed from (B2). The computation of $\phi^{(i+1)} - \phi^{(i)}$ given $\theta_1 - \theta_1^{(i)}$ and $\Gamma_{11}^{(i)}$ is formally identical to the computation of a vector of regression coefficients from normal equations in least squares. For the latter, there are many variants on detailed calculations, and the one used in section 2 was Beaton's SWP operator as discussed in Dempster [1969]. This process is not one of inverting $\Gamma_{11}^{(i)}$ and then multiplying by $\theta_1 - \theta_1^{(i)}$. Rather, it carries out both processes simultaneously, and gradually modifies $\theta_1 - \theta_1^{(i)}$ into $\phi^{(i+1)} - \phi^{(i)}$.

Under the second computing procedure, the stage of passing from $\phi^{(i)}$ to $\phi^{(i+1)}$ consists of a finite sequence of operations O_j for $j = s + 1, s + 2, \cdots, p$, where the operation O_j is the modification of any current ϕ^* which leaves all of the elements of θ^* constant except θ_j^* and simultaneously adjusts ϕ_j^* to the desired ϕ_j. In other words, each O_j is an example of the general computing problem, but with the partition $p = s + (p - s)$ replaced by $p = (p - 1) + 1$. After applying O_j for $j = s + 1, s + 2, \cdots, p$ it is clear that $\theta_1 = (\theta_1, \theta_2, \cdots, \theta_s)$ is unchanged, but unfortunately the ϕ_2^* are not in general all at their desired values because the adjustment O_j generally alters all ϕ_2^* and, while matching ϕ_j^* to its desired value ϕ_j, it destroys any matches on the remaining ϕ_i. Nevertheless, each O_j does produce an increase in the expression (A14), as follows from applying Lemma B to the partition $p = (p - 1) + 1$. Thus the sequence of operations $\phi^{(1)} \to \phi^{(2)} \to \cdots$ does produce an increasing sequence of values of (A14) which only stops increasing in the limit when ϕ_2 attains its desired values. Thus, provided there is any solution, the process as defined will converge to it.

The second computing process is of interest only if the individual operations O_j are simple, as happens in the case of the covariance fitting example of section 2. It is also easy to see in this case that each operation O_j produces a resulting Σ which remains within the class of positive-definite matrices, and that the criterion (A14) which is essentially log det Σ is bounded above because

$$\log \det \Sigma \leq \log (\sigma_{11}\, \sigma_{22}, \cdots \sigma_{pp}). \tag{B4}$$

Thus the iterative procedure produces a monotone increasing sequence of log det Σ values which are bounded above and therefore converge to a finite limit. This limit must occur when the subset of σ^{ij} for pairs (i, j) in I match the desired values 0, because otherwise the next cycle would produce another finite increase in log det Σ. In this way the existence of the estimator $\hat{\Sigma}$ defined in section 2 is proved.

To understand why the O_i in the covariance fitting example are simple, one needs some facility with various different parametrizations of positive-definite symmetric matrices. Specifically, the information in Σ is equivalent to the information in SWP $[i_1, i_2, \cdots, i_t]\Sigma$, where $\{i_1, i_2, \cdots, i_t\}$ is a subset of $\{1, 2, \cdots, p\}$ and SWP denotes the Beaton sweep discussed in Dempster [1969]. In particular, consider the equivalent representations given by

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{B5}$$

$$\text{SWP } [1, 2, \cdots, p-2] \quad \downarrow \quad \uparrow \quad \text{RSW } [1, 2, \cdots, p-2]$$

$$\begin{bmatrix} -\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{B6}$$

$$\text{SWP } [p-1, p] \quad \downarrow \quad \uparrow \quad \text{RSW } [p-1, p]$$

$$\begin{bmatrix} -\Sigma^{11} & -\Sigma^{12} \\ -\Sigma^{21} & -\Sigma^{22} \end{bmatrix}, \tag{B7}$$

where the partitions refer to $p = (p-2) + 2$, (B5) defines the partition of Σ, (B7) defines the partition of $-\Sigma^{-1}$, and (B6) is an intermediate representation, where $-\tilde{\Sigma}_{11} = -\Sigma_{11}^{-1}$, $\tilde{\Sigma}_{12} = \Sigma_{11}^{-1}\Sigma_{12}$, and $\tilde{\Sigma}_{22} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. With this background consider the O_i-type operation which leaves all of the elements σ_{ij} of Σ intact except $\sigma_{p-1,p}$ and which changes the $(p-1, p)$ element $\sigma^{p-1,p}$ of Σ^{-1} to 0. A simple prescription for carrying out this operation given Σ is (i) pass from (B5) to (B6), (ii) set the $(p-1, p)$ element in (B6) to 0, and (iii) pass from the new (B6) back to a new (B5). Alternatively, one can accept the Σ^{-1} parametrization as basic and (i) pass from (B7) to (B6), (ii) alter (B6) as above, (iii) return to (B7). The latter is computationally more desirable since sweeping on only 2 indices rather than $p-2$ is required. It is easy to check from the definitions of SWP and RSW in Dempster [1969] that the prescriptions satisfy the requirements. In addition, stage (ii) modifies $\tilde{\Sigma}_{22}$ in such a way as to leave it positive-definite, so that the new Σ and Σ^{-1} are also positive-definite. Finally, it can be checked that det Σ is increased by the factor $(1 - r^2)^{-1}$, where r^2 is the squared correlation coefficient computed from $\tilde{\Sigma}_{22}$.

Besides its mathematical use as a tool to prove the existence of $\hat{\Sigma}$, the second process can be developed into a practical computing tool for finding

$\hat{\boldsymbol{\Sigma}}$ from S. The steps (i), (ii), (iii) above can be streamlined for this purpose. It is planned to report elsewhere on precise algorithms.

Experience has shown, however, that the Newton-type algorithm is generally much faster than the second algorithm. In addition, the Newton-type algorithm produces $\boldsymbol{\Gamma}_{11}^{-1}$ as a by-product, thus allowing approximate tests of significance to be carried out on estimated $\hat{\phi}_1$ parameters, here non-zero elements of $\hat{\boldsymbol{\Sigma}}^{-1}$. For the purposes of selecting a new σ^{ij} parameter to put into the fitted model, it is likely to be impractical to fit all possible choices by an iterative calculation.

Received August 1971

Key Words: Exponential families; Covariance estimation; Parameter selection; Maximum likelihood computation; Maximum entropy.