The Annals of Statistics
1981, Vol. 9, No. 1, 24–44

ESTIMATION IN A MULTIVARIATE "ERRORS IN VARIABLES" REGRESSION MODEL: LARGE SAMPLE RESULTS[1]

BY LEON JAY GLESER

Purdue University

In a multivariate "errors in variables" regression model, the unknown mean vectors $\mathbf{u}_{1i} : p \times 1$, $\mathbf{u}_{2i} : r \times 1$ of the vector observations $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$, rather than the observations themselves, are assumed to follow the linear relation: $\mathbf{u}_{2i} = \alpha + B\mathbf{u}_{1i}$, $i = 1, 2, \cdots, n$. It is further assumed that the random errors $\mathbf{e}_i = \mathbf{x}_i - \mathbf{u}_i$, $\mathbf{x}_i' = (\mathbf{x}_{1i}', \mathbf{x}_{2i}')$, $\mathbf{u}_i' = (\mathbf{u}_{1i}', \mathbf{u}_{2i}')$, are i.i.d. random vectors with common covariance matrix Σ_e. Such a model is a generalization of the univariate ($r = 1$) "errors in variables" regression model which has been of interest to statisticians for over a century.

In the present paper, it is shown that when $\Sigma_e = \sigma^2 I_{p+r}$, a wide class of least squares approaches to estimation of the intercept vector α and slope matrix B all lead to identical estimators $\hat{\alpha}$ and $\hat{B}$ of these respective parameters, and that $\hat{\alpha}$ and $\hat{B}$ are also the maximum likelihood estimators (MLE's) of α and B under the assumption of normally distributed errors $\mathbf{e}_i$. Formulas for $\hat{\alpha}$, $\hat{B}$ and also the MLE's $\hat{U}_1$ and $\hat{\sigma}^2$ of the parameters $U_1 = (\mathbf{u}_{11}, \cdots, \mathbf{u}_{1n})$ and σ^2 are given. Under reasonable assumptions concerning the unknown sequence $\{\mathbf{u}_{1i}, i = 1, 2, \cdots\}$, $\hat{\alpha}$, $\hat{B}$ and $r^{-1}(r + p)\hat{\sigma}^2$ are shown to be strongly (with probability one) consistent estimators of α, B and σ^2, respectively, as $n \to \infty$, regardless of the common distribution of the errors $\mathbf{e}_i$. When this common error distribution has finite fourth moments, $\hat{\alpha}$, $\hat{B}$ and $r^{-1}(r + p)\hat{\sigma}^2$ are also shown to be asymptotically normally distributed. Finally large-sample approximate $100(1 - \nu)\%$ confidence regions for α, B and σ^2 are constructed.

1. Introduction. It is well known that the presence of errors of measurement in the independent variables in univariate (one dependent variable) linear regression makes the ordinary least squares estimators inconsistent and biased. An extensive literature, dating back to Adcock (1878), deals with estimation in models of univariate regression which incorporate "errors in variables." (For references, see Anderson (1976), Madansky (1959), Moran (1971), Sprent (1966).) Less is known concerning estimation in multivariate "errors in variables" regression models as such; although, as shown later, such models are *mathematically* equivalent to linear functional equation models and certain models of factor analysis, for which a sizeable and constantly expanding literature exists (see Anderson (1976)). Recent work relating directly to the multivariate "errors in variables" model as parameterized in the present paper has been done by Gleser and Watson (1973). Bhargava (1979), and Healy (1975).

In a multivariate "errors in variables" regression model, n random vectors $\mathbf{x}_i = (\mathbf{x}_{1i}', \mathbf{x}_{2i}')'$ are observed, where $\mathbf{x}_{1i}$ is $p \times 1$ and $\mathbf{x}_{2i}$ is $r \times 1$, $i = 1, 2, \cdots, n$. It is assumed that

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_{1i} \\ \mathbf{u}_{2i} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2i} \end{pmatrix} = \mathbf{u}_i + \mathbf{e}_i, \tag{1.1}$$

where

$$\mathbf{u}_{2i} = \alpha + B\mathbf{u}_{1i}, \qquad i = 1, 2, \cdots, n, \tag{1.2}$$

Received August 1976; revised December 1979.

[1] This research was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF under Grants AFOSR-72-2350B and AFOSR-77-3291 at Purdue University.

AMS 1970 *subject classifications.* Primary 62H10, 62E20; secondary 62F10, 62F25, 62H25, 62P15, 62P20.

Key words and phrases. Errors in variables, linear functional equations, factor analysis, multivariate linear regression, least squares, generalized least squares, orthogonally invariant norm, maximum likelihood, strong consistency, asymptotic distributions, approximate confidence regions, principal components.

and that

(1.3) the $\mathbf{e}_i$'s are i.i.d. with mean vector $\mathbf{0}$ and covariance matrix Σ_e.

The parameters $B : r \times p$ and $\mathbf{u}_{1i} : p \times 1$, $i = 1, 2, \cdots, n$ are assumed to be unknown, and are to be estimated. The parameter $\alpha : r \times 1$ is either known to be $\mathbf{0}$ (the *no-intercept model*), or else is unknown (the *intercept model*) and must be estimated. If consistent sequences of estimators for B are desired, the parametric form of Σ_e cannot be completely unspecified (see Section 5). In the present paper, it is assumed that $\Sigma_e = \sigma^2 I_{p+r}$ where the scalar $\sigma^2 > 0$ is unknown and is to be estimated. That is, the assumption (1.3) is specialized to the following:

(1.3′) The $\mathbf{e}$'s are i.i.d. with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_e = \sigma^2 I_{p+r}$.

In Section 5, it is shown that when $\Sigma_e, = \sigma^2\Sigma_0$, Σ_0 known and positive definite, then the data can be transformed linearly in such a way that (1.1), (1.2), and (1.3′) hold for the transformed data.

To condense notation, let X_1 be the $p \times n$ random matrix whose columns are $\mathbf{x}_{1i}$, $i = 1, 2, \cdots, n$, and let X_2 be the $r \times n$ random matrix whose columns are $\mathbf{x}_{2i}$, $i = 1, 2, \cdots, n$. Similarly, let U_1 be the $p \times n$ matrix whose columns are $\mathbf{u}_{1i}$, $i = 1, 2, \cdots, n$, and let U_2 be the $r \times n$ matrix whose columns are $\mathbf{u}_{2i}$, $i = 1, 2, \cdots, n$. Finally, let E be the $(p + r) \times n$ random *error matrix* whose columns are $\mathbf{e}_i$, $i = 1, 2, \cdots, n$, and let

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \qquad U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} : (p + r) \times n,$$

have columns $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ and $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$, respectively. In terms of these matrices, the model (1.1), (1.2) becomes

$$X = U + E = \begin{pmatrix} \mathbf{0} \\ \alpha \mathbf{1}'_n \end{pmatrix} + \begin{pmatrix} I_p \\ B \end{pmatrix} U_1 + E, \tag{1.4}$$

where $\mathbf{1}'_n = (1, 1, \cdots, 1) : 1 \times n$. The assumption (1.3′) can be restated as follows:

ASSUMPTION A. The columns of E are i.i.d. with common mean vector $\mathbf{0}$ and common covariance matrix Σ_e, where $\Sigma_e = \sigma^2 I_{p+r}$.

The mathematical equivalence between models of the form (1.4), linear functional equation models, and certain fixed-factors analysis models is discussed at the end of this introduction.

Two approaches to the estimation of the parameters α, B, U_1 and σ^2 have received primary attention in the literature on univariate ($r = 1$) "errors in variables" regression models. In what may be called *the* MLE *approach*, it is assumed that the common distribution of the columns of E is multivariate normal, and maximum likelihood estimators (MLE's) of the parameters are found. In contrast, the GLSE *approach* makes no assumptions about the distribution of E beyond those found in Assumption A. Since under the model (1.4), $X_2 - \alpha\mathbf{1}'_n - BX_1$ has mean $\mathbf{0} : r \times n$ and columns with common covariance matrix $\sigma^2(I_r + BB')$, values of α and B are chosen to minimize some scalar function of the normalized "error" (or "residual") matrix

$$Q(\alpha, B; X) = (I_r + BB')^{-1/2}(X_2 - \alpha\mathbf{1}'_n - BX_1), \tag{1.5}$$

where $(I_r + BB')^{1/2}$ is any square root of $I_r + BB'$. In the case $r = 1$, (1.5) is an n-dimensional vector, and minimization of the length of (1.5) yields the *generalized least squares estimators* (GLSE's) of Sprent (1966). (See also Moran, 1971; pages 246–7.) In this case, the MLE and GLSE approaches yield identical estimators. When $r > 1$, many scalar functions of (1.5) could be used. In the case $r = p$ of the no-intercept model, Gleser and Watson (1973) show that the MLE approach and the GLSE approach which minimizes the Euclidean norm $(\mathrm{tr}[QQ'])^{1/2}$ of (1.5) lead to the same estimator of B. In addition to the MLE and GLSE approaches, an approach analogous to ordinary least squares could be used, based on minimizing some scalar function of

$$R(\alpha, B, U_1; X) = X - \begin{pmatrix} \mathbf{0} \\ \alpha \mathbf{1}_n' \end{pmatrix} - \begin{pmatrix} I_p \\ B \end{pmatrix} U_1 \tag{1.6}$$

over α, B, U_1. It is well known that this method, when based on minimizing the Euclidean norm $(\text{tr}[RR'])^{1/2}$ of (1.6), and the MLE approach yield the same estimators of α, B and U_1.

In Section 2, it is shown that the GLSE approach based on minimizing any orthogonally invariant norm (see Definition 2.1) of (1.5) yields the same estimators of α and B as the MLE approach. It is also shown that the ordinary least squares approach based on minimizing any orthogonally invariant norm of (1.6) yields the same estimators of α, B and U_1 as the MLE approach. Formulas are given for the MLE's $\hat{\sigma}^2$, $\hat{B}$, $\hat{U}_1$ and $\hat{\alpha}$ of σ^2, B, U_1 and α under both the intercept and no-intercept models, and the uniqueness of these estimators as MLE's, GLSE's and ordinary least squares estimators is discussed.

In Section 3, it is shown that $\hat{B}$ and $r^{-1}(p + r)\hat{\sigma}^2$, define sequences of strongly consistent (as $n \to \infty$) estimators of B and σ^2, respectively, in both the intercept and no-intercept models; and that in the intercept model, $\hat{\alpha}$ defines a sequence of strongly consistent estimators of α. For the intercept model with $r \le p$, these results have previously been reported by Healy (1975). The strong consistency results of Section 3 are obtained without any assumptions about the distribution of E beyond those stated in Assumption A. In particular, it is not assumed that the columns of E have a multivariate normal distribution.

Strong consistency results (which assert convergence with probability one) provide deeper knowledge of the large-sample properties of estimators than do weak consistency results (which assert only convergence in probability). Results of the latter type have been obtained for $\hat{B}$ and $r^{-1}(r + p)\hat{\sigma}^2$ for the no-intercept model by Gleser and Watson (1973) and Bhargava (1979), and in the case $r = 1$ for both no-intercept and intercept models by many authors. Although weak consistency results are adequate for obtaining large-sample distributions and Fisher efficiencies of estimators, strong consistency results seem to be needed in the theory of sequential estimation (see, e.g., Gleser and Kunte (1976)) and in deriving large-deviation based measures of efficiency (Bahadur (1971)). Hence, it is worth demonstrating that a sequence $\hat{\xi}_n$ of estimators is strongly consistent for a parameter ξ.

Section 4 considers large-sample distributional results for $n^{1/2}(\hat{B} - B)$ and $n^{1/2}(r^{-1}(r + p)\hat{\sigma}^2 - \sigma^2)$ in both the intercept and no-intercept models, and for $n^{1/2}(\hat{\alpha} - \alpha)$ in the intercept model. If the common distribution of the columns of E has finite fourth moments, these quantities have asymptotic multivariate normal distributions. Assuming that the common distribution of the columns of E is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 I_{p+r}$ (that is, $\mathbf{e}_i \sim \text{MVN}(\mathbf{0}, \sigma^2 I_{p+r})$, all i), or that the moments of this common distribution up to, and including, the fourth-order agree with the corresponding moments of the $\text{MVN}(\mathbf{0}, \sigma^2 I_{p+r})$ distribution, approximate large-sample $100(1 - \nu)\%$ confidence regions for α, B and σ^2 are constructed. Finally, Section 5 discusses the practical relevance of the assumptions needed to carry through the large-sample theory.

Two related models. Before beginning discussion of the theoretical results mentioned above, it is worth commenting on the relationship of the model (1.4) to two mathematically equivalent models for which there is an extensive literature—the *linear functional equation model* and the *factor analysis model* with fixed factors. Both of these models start with the assumption that a $(p + r) \times n$ matrix X of observations has the structure

$$X = U + E, \tag{1.7}$$

where U is a $(p + r) \times n$ matrix of unknown constants believed to have rank p, and the columns of E are i.i.d. with common mean vector η and common covariance matrix Σ_e. The assertion that U has rank p is equivalent to asserting the existence of an $r \times (p + r)$ matrix K of full rank such that

$$KU = \mathbf{0}. \tag{1.8}$$

Even if U is known, such a matrix K is not unique. Linear functional equation models impose conditions on K and U sufficient to insure statistical identifiability of these parameters; the goal of users of such models is to estimate the matrix K of functional coefficients. Tintner (1945) and Geary (1948) assume that Σ_e is known (or estimable by an independent and consistent estimator $\hat{\Sigma}_e$), and use the MLE approach to estimate K and U. Geary (1948) proves weak consistency of the MLE $\hat{K}$ and discusses the problem of consistently estimating $n^{-1}U(I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')U'$. Using large-sample distributional results of Anderson (1948), Geary also obtains the classical Fisher efficiency of $\hat{K}$. These large-sample results (except for the efficiency) do not require the assumption that the common distribution of the columns of E is multivariate normal.

As noted by Anderson (1976), Geary's results can be made to apply to the case where it is assumed that $\Sigma_e = \sigma^2\Sigma_0$, where Σ_0 is known. Such a model can be transformed (see Section 5) into a model of the form (1.7), (1.8) in which $\Sigma_e = \sigma^2 I_{p+r}$, corresponding to Assumption A. In this case, the restriction upon K imposed by Tintner (1945) and Geary (1948) to permit statistical identifiability of the parameters is that

$$KK' = I_r. \tag{1.9}$$

Note that the "errors in variables" model (1.4) has the form of (1.7), (1.8) with $\eta' = (\mathbf{0}, \alpha')$, but instead of (1.9) requires that K has the form

$$K = (-B, I_r).$$

Another assertion equivalent to the statement that U in (1.7) has rank p is that

$$U = LF, \tag{1.10}$$

where $L : (p + r) \times p$ and $F : p \times n$ are both of rank p. Equation (1.10) is recognizable as defining a factor analysis model in which L is the matrix of factor loadings and F is the matrix of factor scores. Most of the literature on factor analysis assumes that the columns of F are i.i.d. replications of a MVN($\mathbf{0}$, I_p) distribution, and that Σ_e is an unknown diagonal matrix. However, Whittle (1952), Lindley (1953) and Anderson and Rubin (1956) discuss the case where F is assumed fixed (fixed factors), with

$$F\mathbf{1}_n = \mathbf{0}, \qquad FF' = nI_p \tag{1.11}$$

and $\Sigma_e = \sigma^2 I_{p+r}$. The restrictions (1.11) are not enough to statistically identify the model; Anderson and Rubin (1956) provide a detailed discussion of the kinds of conditions on L that can serve for this purpose. Using the restriction that $L'L$ is diagonal with distinct descending diagonal elements, Lindley (1953) uses the MLE approach to obtain estimators of L, F and σ^2. Both Lindley (1953) and Anderson and Rubin (1956) discuss weak consistency and large-sample distribution theory for the MLE of L and σ^2 (again, without assuming that the columns of E have a multivariate normal distribution). Clearly, the "errors in variables" model (1.4) has the form of (1.10) with $F = U_1$, $\eta' = (\mathbf{0}, \alpha')$ and

$$L = \begin{pmatrix} I_p \\ B \end{pmatrix}. \tag{1.12}$$

The requirement (1.12) specifies that certain elements of L have known values, corresponding to one of the methods of identifying the model (1.10) suggested by Anderson and Rubin (1956).

Since the models (1.4), (1.8), (1.10) all stem from a common linear structure (1.7), it is reasonable to expect that MLE's and large-sample results for the MLE's of the parameters of any one model can be used to obtain MLE's and large sample results for the MLE's of the other models. Provided that the parameters of the models (1.8) and (1.10) are sufficiently restricted so as to insure statistical identifiability, this expectation is (with an exception noted below) largely correct. Nevertheless, in the present paper little use is made of results appearing in the literature of linear functional equation and fixed-factor factor analysis models. In Section 2, the MLE approach is treated in the context of a broader, more distribution-free least squares approach for which few parallels appear in the literature of the related models. The large-sample results of Sections 3 and 4 are obtained without assuming that the nonzero eigenvalues of $\lim_{n\to\infty} n^{-1}UU'$ (in the no-intercept model), or of $\lim_{n\to\infty} n^{-1}U(I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')U'$ in the intercept model, are distinct, an assumption adopted (and apparently necessary) for the large-sample results obtained in the literature of the other two models relating to the case $\Sigma_e = \sigma^2 I_{p+r}$. Finally, since the parameters of the model (1.4) have important practical interpretations in may scientific contexts, and since this model has long been of interest to statisticians, it seems desirable to analyze this model strictly in terms of its own structure, particularly since such a direct approach has the advantage of clarity.

2. MLE and GLSE estimators. Because the analyses of the no-intercept and intercept forms of the model (1.4) involve similar steps, an attempt is made in the following sections to treat these two forms of the model simultaneously. To this end let

$$\begin{aligned} C_n &= I_n, && \text{for the no-intercept model,} \\ &= I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n', && \text{for the intercept model.} \end{aligned} \tag{2.1}$$

The estimators considered in this paper are based on the following reduction of the data X: Let

$$W = XC_nX', \tag{2.2}$$

and let $d_1 \geq d_2 \geq \cdots \geq d_p \geq d_{p+1} \geq \cdots \geq d_{p+r} \geq 0$ be the (ordered) eigenvalues of W. Let

$$D = \begin{pmatrix} D_{\max} & \mathbf{0} \\ \mathbf{0} & D_{\min} \end{pmatrix} = \operatorname{diag}(d_1, d_2, \cdots, d_{p+r}), \tag{2.3}$$

where $D_{\max} = \operatorname{diag}(d_1, d_2, \cdots, d_p)$, $D_{\min} = \operatorname{diag}(d_{p+1}, \cdots, d_{p+r})$. Finally, let

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} : (p + r) \times (p + r),\ G_{11} : p \times p, \tag{2.4}$$

satisfy

$$G'G = I_{p+r} = GG', \tag{2.5}$$

$$W = GDG'. \tag{2.6}$$

That is, G is an orthogonal matrix whose ith column is the eigenvector of W corresponding to d_i, $i = 1, \cdots, p + r$.

Assuming that G_{11}^{-1} and G_{22}^{-1} exist, define estimators of B, U_1 and α as follows:

$$\hat{B} = G_{21}G_{11}^{-1} = -(G_{22}')^{-1}G_{12}', \tag{2.7}$$

$$\hat{U}_1 = (G_{11}G_{11}'X_1 + G_{11}G_{21}'X_2)C_n = [(I_p - G_{12}G_{12}')X_1 - G_{12}G_{22}'X_2]C_n, \tag{2.8}$$

$$\begin{aligned} \hat{\alpha} &= \mathbf{0}, && \text{for the no-intercept model,} \\ &= (-\hat{B}, I_r)\bar{\mathbf{x}}, && \text{for the intercept model,} \end{aligned} \tag{2.9}$$

where $\bar{\mathbf{x}} = (\bar{\mathbf{x}}_1', \bar{\mathbf{x}}_2')' = n^{-1}X\mathbf{1}_n$. The equivalence of the alternative computation formulas for $\hat{B}$ and $\hat{U}_1$ given in (2.7) and (2.8) follow as a direct consequence of (2.5). Whether or not G_{11}^{-1} and G_{22}^{-1} exist, define

$$\hat{\sigma}^2 = n^{-1}(p + r)^{-1}\operatorname{tr}[D_{\min}]. \tag{2.10}$$

Since computation of $\hat{B}$ requires that G_{11}^{-1} and/or G_{22}^{-1} exist, the following two lemmas are of interest.

LEMMA 2.1. *If G is defined by* (2.4) *and* (2.5), *then G_{11} is nonsingular if and only if G_{22} is nonsingular.*

PROOF. Suppose that G_{11} is nonsingular. Let $G_{22\cdot 1} = G_{22} - G_{21}G_{11}^{-1}G_{12}$. From the determinantal equality $|G| = |G_{11}||G_{22\cdot 1}|$ and the facts that $|G| \neq 0$, $|G_{11}| \neq 0$, it follows that

$|G_{22\cdot 1}| \neq 0$. However, it is a consequence of (2.5) that $G_{21}G'_{11} + G_{22}G'_{21} = \mathbf{0}$, and hence that

$$G_{22\cdot 1} = G_{22}[I_r + (G_{11}^{-1}G_{12})'(G_{11}^{-1}G_{12})].$$

Therefore, $|G_{22}| \neq 0$ and G_{22} is nonsingular. A similar argument shows that the nonsingularity of G_{22} implies that G_{11} is nonsingular. □

LEMMA 2.2. *Let G be defined by* (2.4), (2.5) *and* (2.6). *If Assumption* A *holds, rank* $(C_n) \geq p + r$, *and the common distribution of the columns of E is absolutely continuous with respect to Lebesgue measure on* $(p + r)$*-dimensional Euclidean space, then*

$$P\{G_{11} \text{ and } G_{22} \text{ are nonsingular}\} = 1. \tag{2.11}$$

PROOF. From the given, it is straightforward to show that the elements of W on or below the diagonal have a joint distribution which is absolutely continuous with respect to $\frac{1}{2}(p + r)(p + r + 1)$-dimensional Lebesgue measure. Representing the elements of G in terms of polar coordinates, transforming from W to (G, D), and then integrating out terms not of interest, it can be shown that if $p \leq r$, the p^2 elements of G_{11} have an absolutely continuous distribution with respect to p^2-dimensional Lebesgue measure; while if $r \leq p$, the r^2 elements of G_{22} have an absolutely continuous distribution with respect to r^2-dimensional Lebesgue measure. Consider the case $p \leq r$. (The case $r \leq p$ is treated similarly.) Note that $|G_{11}|$ is a polynomial in the elements of G_{11}. By Okamoto's (1973) lemma, $P\{|G_{11}| \neq 0\} = 1$, and this, together with Lemma 2.1, implies (2.11). □

For any $s \times t$ matrix A, $s \leq t$, let $\lambda_i[A]$ be the ith largest nonzero singular value of A; that is, $\lambda_i^2[A]$ is the ith largest eigenvalue of AA', or of $A'A$, and

$$\lambda_1[A] \geq \lambda_2[A] \geq \cdots \geq \lambda_s[A] \geq 0.$$

Recall that if A is an $s \times s$ positive semidefinite matrix (A is p.s.d.), then $\lambda_1[A], \cdots, \lambda_s[A]$ are also the eigenvalues of A. If A_1, A_2 are $s \times s$ symmetric matrices, the notation $A_1 \geq A_2$ means that $A_1 - A_2$ is p.s.d. It is well known that if A_1, A_2 are $s \times s$ symmetric,

$$A_1 \geq A_2 \text{ implies that } \lambda_i[A_1] \geq \lambda_i[A_2], \qquad i = 1, 2, \cdots, s. \tag{2.12}$$

Further if $A_1 \geq A_2$ and $A_1 \neq A_2$, at least one of the differences $\lambda_i[A_1] - \lambda_i[A_2]$ must be positive.

DEFINITION 2.1. An *orthogonally invariant norm* $\|\cdot\|$ defined on $s \times t$ matrices A is a norm which also has the property that

$$\|H_1A\| = \|A\| = \|AH_2\|$$

for all $A : s \times t$, all $H_1 : s \times s$ orthogonal, all $H_2 : t \times t$ orthogonal.

Using the singular value decomposition for $s \times t$ matrices, it is easily shown that any orthogonally invariant norm $\|A\|$ is a function of A only through $\lambda_1[A], \cdots, \lambda_s[A]$. Indeed, the following is true.

LEMMA 2.3. *The function* $\|\cdot\|$ *on* $s \times t$ *matrices A is an orthogonally invariant norm if and only if*

$$\|A\| = g(\lambda_1[A], \cdots, \lambda_s[A]) \tag{2.13}$$

for some symmetric gauge function (s.g.f.) $g(\cdot)$ *defined on s-dimensional Euclidean space.*

PROOF. See von Neumann (1937).

A scalar-valued function $g(\cdot)$ defined on s-dimensional Euclidean space is an s.g.f. if $g(\cdot)$ is a norm which further satisfies

$$g(D_{\pm 1}P\boldsymbol{\lambda}) = g(\boldsymbol{\lambda})$$

for all $\boldsymbol{\lambda} : s \times 1$, all diagonal matrices $D_{\pm 1}$ having ± 1's on the diagonal, and all $s \times s$ permutation matrices P.

LEMMA 2.4. *Let $g(\cdot)$ be an* s.g.f. *on s-dimensional space.*
(i) *The function $g^*(\boldsymbol{\lambda}_1)$ defined on s_1-dimensional Euclidean space, $s_1 \leq s$, by $g^*(\boldsymbol{\lambda}_1) = g((\boldsymbol{\lambda}_1', \mathbf{0}')')$ is an* s.g.f.
(ii) *If $\boldsymbol{\lambda}^{(j)} = (\lambda_1^{(j)}, \cdots, \lambda_s^{(j)})'$, $j = 1, 2$, and $|\lambda_i^{(1)}| \leq |\lambda_i^{(2)}|$, $i = 1, 2, \cdots, s$, then $g(\boldsymbol{\lambda}^{(1)}) \leq g(\boldsymbol{\lambda}^{(2)})$.*

PROOF. Part (i) follows directly from the definition of an s.g.f. A proof of part (ii) appears in Olkin (1966; pages 55–56). □

The following definitions are implicit (or explicit) in the discussion of Section 1.

DEFINITION 2.2. The estimators $\tilde{\alpha}(X)$, $\tilde{B}(X)$, $\tilde{U}_1(X)$ are *ordinary least squares estimators* (OLSE's) of α, B, U_1, respectively, relative to an orthogonally invariant norm $\|\cdot\|$ defined on $(p + r) \times n$ matrices R, if for all α, B, U_1, X,

$$\| R(\alpha, B, U_1; X) \| \geq \| R(\tilde{\alpha}(X), \tilde{B}(X), \tilde{U}_1(X); X) \|,$$

where $R(\alpha, B, U_1; X)$ is defined by (1.6).

DEFINITION 2.3. The estimators $\alpha^*(X)$, $B^*(X)$ are GLSE's of α, B, respectively, relative to an orthogonally invariant norm $\|\cdot\|_0$ defined on $r \times n$ matrices Q, if for all α, B, X,

$$\| Q(\alpha, B; X) \|_0 \geq \| Q(\alpha^*(X), B^*(X); X) \|_0,$$

where $Q(\alpha, B; X)$ is defined by (1.5).

THEOREM 2.1. *If G_{11}^{-1} and G_{22}^{-1} exist, then $\hat{\alpha}(X)$, $\hat{B}(X)$, $\hat{U}_1(X)$ defined by* (2.9), (2.7) *and* (2.8) *are* OLSE's *of α, B, U_1 relative to any orthogonally invariant norm on $(p + r) \times n$ matrices, and further $\hat{\alpha}(X)$ and $\hat{B}(X)$ are* GLSE's *relative to any orthogonally invariant norm on $r \times n$ matrices.*

PROOF. Let $\|\cdot\|$ be any orthogonally invariant norm defined on $(p + r) \times n$ matrices. Fix α and B. Applying classical projection arguments, it can be shown that

$$R'(\alpha, B, U_1; X)R(\alpha, B, U_1; X) \geq R'(\alpha, B, \hat{U}_1(\alpha, B); X)R(\alpha, B, \hat{U}_1(\alpha, B); X), \tag{2.14}$$

where

$$\hat{U}_1(\alpha, B) = (I_p + B'B)^{-1}(I_p, B')\left[X - \begin{pmatrix} \mathbf{0} \\ \alpha \mathbf{1}_n' \end{pmatrix} \right]. \tag{2.15}$$

Indeed, equality in (2.14) holds if and only if $U_1 = \hat{U}_1(\alpha, B)$. Since the nonzero eigenvalues of $R'R$ and RR' are identical, it now follows from (2.14), (2.12), the "only if" part of Lemma 2.3, and Lemma 2.4(ii) that

$$\| R(\alpha, B, U_1; X) \| \geq \| R(\alpha, B, \hat{U}_1(\alpha, B); X) \|. \tag{2.16}$$

Note, from (1.5), (1.6) and (2.15), that

$$\begin{aligned} R(\alpha, B, \hat{U}_1(\alpha, B); X) &= \left[I_{p+r} - \begin{pmatrix} I_p \\ B \end{pmatrix} (I_p + B'B)^{-1}(I_p, B') \right] \left[X - \begin{pmatrix} \mathbf{0} \\ \alpha \mathbf{1}_n' \end{pmatrix} \right] \\ &= \left[\begin{pmatrix} -B' \\ I_r \end{pmatrix} (I_r + BB')^{-1}(-B, I_r) \right] \left[X - \begin{pmatrix} \mathbf{0} \\ \alpha \mathbf{1}_n' \end{pmatrix} \right] \qquad (2.17) \\ &= \Gamma'(B) Q(\alpha, B; X), \end{aligned}$$

where

$$\Gamma(B) = (I_r + BB')^{-1/2}(-B, I_r) \tag{2.18}$$

is an $r \times (p + r)$ row-orthogonal matrix. In consequence, p singular values of $R(\alpha, B, \hat{U}_1(\alpha, B); X)$ are zero, with the remaining r singular values being equal to the singular values of $Q(\alpha, B; X)$. From Lemma 2.3 and Lemma 2.4(i), there is an orthogonally invariant norm $\|\cdot\|_0$ defined on $r \times n$ matrices Q such that

$$\| R(\alpha, B, \hat{U}_1(\alpha, B); X) \| = \| Q(\alpha, B; X) \|_0, \tag{2.19}$$

for all α, B, X. Noting that $\hat{U}_1 = \hat{U}_1(\hat{\alpha}, \hat{B})$, it follows that if it can be shown that $\hat{\alpha}$ *and* $\hat{B}$ are GLSE's of α and B relative to any orthogonally invariant norm $\|\cdot\|_0$ on $r \times n$ matrices, then the OLSE character of $\hat{\alpha}$, $\hat{B}$, $\hat{U}_1$ realtive to $\|\cdot\|$ will follow as an immediate corollary.

Hence, let $\|\cdot\|_0$ be any orthogonally invariant norm defined on $r \times n$ matrices Q. Fix B and note that for all α, X,

$$Q(\alpha, B; X)Q'(\alpha, B; X) \geq Q(\mathbf{0}, B; XC_n)Q'(\mathbf{0}, B; XC_n) \tag{2.20}$$

with equality holding if and only if

$$\begin{aligned} \alpha = \hat{\alpha}(B) &= \mathbf{0}, && \text{in the no-intercept model,} \\ &= (-B, I_r)\bar{\mathbf{x}}, && \text{in the intercept model.} \end{aligned} \tag{2.21}$$

Consequently, (2.12) and Lemma (2.4)(ii) imply that

$$\| Q(\alpha, B; X) \|_0 \geq \| Q(\hat{\alpha}(B), B; X) \|_0 = \| Q(\mathbf{0}, B; XC_n) \|_0. \tag{2.22}$$

Now, note that

$$Q(\mathbf{0}, B; XC_n) = \Gamma(B)XC_n, \tag{2.23}$$

where $\Gamma(B)$ is defined by (2.18). It is a consequence of the Courant-Fischer min-max theorem that for all row-orthogonal matrices $\Gamma: r \times (p + r)$,

$$\lambda_i[\Gamma XC_n] \geq \lambda_{p+i}[XC_n] = \lambda_{p+i}^{1/2}[W] = d_{p+i}^{1/2}, \qquad i = 1, 2, \cdots, r, \tag{2.24}$$

with equality holding for all i in (2.24) if

$$\Gamma = \Lambda(G_{21}', G_{22}') \tag{2.25}$$

for some $r \times r$ orthogonal matrix Λ. It is easily shown that $\Gamma = \Gamma(\hat{B})$ satisfies (2.25) with $\Lambda = (G_{22}'G_{22})^{1/2}(G_{22}')^{-1}$. Thus, it follows from (2.23), (2.24), Lemma 2.3 and Lemma 2.4(ii) that for all B, X,

$$\| Q(\mathbf{0}, B; XC_n) \|_0 \geq \| Q(\mathbf{0}, \hat{B}; XC_n) \|_0 = g(d_{p+1}^{1/2}, \cdots, d_{p+r}^{1/2}), \tag{2.26}$$

where $g(\cdot)$ is the s.g.f. corresponding to $\|\cdot\|_0$. Applying (2.22) and (2.26) yields

$$\| Q(\alpha, B; X) \|_0 \geq \| Q(\hat{\alpha}, \hat{B}; X) \|_0 = g(d_{p+1}^{1/2}, \cdots, d_{p+r}^{1/2}), \tag{2.27}$$

proving that $\hat{\alpha}$ and $\hat{B}$ are GLSE's of α, B relative to $\|\cdot\|_0$. □

It is of interest to determine when $\hat{\alpha}$, $\hat{B}$, $\hat{U}_1$ are the *unique* OLSE's of α, B, U_1 relative to an orthogonally invariant norm $\|\cdot\|$, and similarly when $\hat{\alpha}$ and $\hat{B}$ are the unique GLSE's of α, B relative to an orthogonally invariant norm $\|\cdot\|_0$. Looking through the proof of Theorem 2.1 it is clear that such uniqueness is only possible if the norms $\| R \|$ or $\| Q \|_0$ are strictly monotonic in the singular values of R and Q, respectively.

DEFINITION 2.4. An orthogonally invariant norm $\| A \|$ on $s \times t$ matrices A, $s \leq t$ is *strictly monotonic* if $\lambda_i[A_1] \geq \lambda_i[A_2]$, $i = 1, 2, \cdots, s$, with strict inequality holding for at least one index i, implies that $\| A_1 \| > \| A_2 \|$.

Note that the classical Euclidean norm $\|A\| = (\mathrm{tr}[AA'])^{1/2}$ is orthogonally invariant and strictly monotonic, but that the orthogonally invariant norm $\|A\| = \lambda_1[A]$ is not strictly monotonic.

THEOREM 2.2 *Assume that the conditions of Lemma* 2.2 *hold. Then*
(i) $\hat{\alpha}$, $\hat{B}$ *and* $\hat{U}_1$ *are the unique* (*with probability one*) OLSE's *of* α, B *and* U_1 *relative to any strictly monotonic, orthogonally invariant norm* $\|\cdot\|$ *on* $(p + r) \times n$ *matrices* R;
(ii) $\hat{\alpha}$ *and* $\hat{B}$ *are the unique* (*with probability one*) GLSE's *of* α *and* B *relative to any strictly monotonic, orthogonally invariant norm* $\|\cdot\|_0$ *on* $r \times n$ *matrices* Q.

PROOF. Let $\mathscr{C}$ be the collection of all $X:(p+r) \times n$ for which G_{11} and G_{22} are nonsingular, and also $d_p > d_{p+1}$. It follows from Lemma 2.2 and results of Okamoto (1973) that $P(\mathscr{C}) = 1$. Fix $X \in \mathscr{C}$. Recall that $\hat{U}_1(\hat{\alpha}, \hat{B}) = \hat{U}_1$, $\hat{\alpha}(\hat{B}) = \hat{\alpha}$. Also note that since G_{22} is nonsingular, the only solution for B of the equality

$$\Gamma(B) = (I_r + BB')^{-1/2}(-B, I_r) = \Lambda(G_{12}', G_{22}') \tag{2.28}$$

is $B = \hat{B}$. However, since $d_p > d_{p+1}$, the equality (2.25) is a necessary and sufficient condition for equality to hold for all i in (2.24). Now using (2.12), the remark following (2.12), and Definition 2.4, and retracing the steps of the proof of Theorem 2.1, assertions (i) and (ii) are easily verified. □

REMARK. If the common joint distribution of the columns of E is not absolutely continuous with respect to Lebesgue measure, or if rank $(C_n) < p + r$ (so that the conditions of Lemma 2.2 fail to hold), it is conceivable that a collection of X-values could exist with positive probability for which $d_p > d_{p+1}$, but for which G_{11} and G_{22} are singular. In this case, no solution for B in (2.28) can exist and thus no B exists for which $\lambda_i[\Gamma(B)XC_n] = d_{p+i}^{1/2}$, $i = 1, 2, \cdots, r$ (see (2.24) and (2.25)). If $\|\cdot\|$, or $\|\cdot\|_0$, are strictly monotonic, then for this collection of X's, OLSE's of α, B, U_1 and GLSE's of α, B *do not exist.* This assertion follows since it is possible to show for any orthogonally invariant norm $\|\cdot\|_0$ on $r \times n$ matrices that

$$\inf_{\alpha,B} \|Q(\alpha, B; X)\|_0 = g(d_{p+1}^{1/2}, \cdots, d_{p+r}^{1/2}), \tag{2.29}$$

where $g(\cdot)$ is the s.g.f. corresponding to $\|\cdot\|_0$, and then the above argument shows that this infimum cannot be attained by any (α, B).

It remains to show that $\hat{\alpha}$, $\hat{B}$, $\hat{U}_1$ and $\hat{\sigma}^2$ are MLE's of their respective parameters.

THEOREM 2.3. *If rank* $(C_n) \geq p + r$ *and the common distribution of the columns of* E *is multivariate normal, then* $\hat{\alpha}$, $\hat{B}$, $\hat{U}_1$ *and* $\hat{\sigma}^2$ *defined by* (2.9), (2.7), (2.8) *and* (2.10), *respectively, are the unique* (*with probability one*) MLE's *of* α, B, U_1 *and* σ^2. *Further,*

$$\begin{aligned} \max_{\alpha,B,U_1,\sigma^2} L(X; \alpha, B, U_1, \sigma^2) &= L(X; \hat{\alpha}, \hat{B}, \hat{U}_1, \hat{\sigma}^2) \\ &= \left\{\frac{n(p+r)}{(2\pi e)(\mathrm{tr}[D_{\min}])}\right\}^{n(p+r)/2}, \end{aligned} \tag{2.30}$$

where $L(X; \alpha, B, U_1, \sigma^2)$ *is the likelihood of* X.

PROOF. The likelihood of the data X is

$$L(X; \alpha, B, U_1, \sigma^2) = (2\pi\sigma^2)^{-n(p+r)/2}\exp\left\{-\frac{1}{2\sigma^2}\|R(\alpha, B, U_1; X)\|^2\right\}, \tag{2.31}$$

where $R(\alpha, B, U_1; X)$ is defined by (1.6) and

$$\|R\| = (\mathrm{tr}[RR'])^{1/2} = [\textstyle\sum_{i=1}^{p+r} \lambda_i^2[R]]^{1/2}$$

is the classical Euclidean norm. Applying Theorem 2.1, together with (2.16) (2.19) and (2.27), yields the inequality

$$
\begin{aligned}
(2.32) \qquad L(X; \alpha, B, U_1, \sigma^2) &\le L(X; \hat{\alpha}, \hat{B}, \hat{U}_1, \sigma^2) \\
&= (2\pi\sigma^2)^{-n(p+r)/2} \exp\left\{ -\frac{1}{2\sigma^2} \operatorname{tr}[D_{\min}] \right\},
\end{aligned}
$$

which holds for all X for which G_{11} and G_{22} are nonsingular, all α, B, U_1 and σ^2. Since the multivariate normal distribution is absolutely continuous with respect to Lebesgue measure, it follows from Lemma 2.2 that (2.32) holds with probability one. Since the classical Euclidean norm is strictly monotonic, it follows from Theorem 2.2 that $\hat{\alpha}$, $\hat{B}$, $\hat{U}_1$ are the unique (with probability one) estimators of α, B, U_1 that achieve the upper bound in (2.32). Finally, it is well known that the upper bound in (2.32) is uniquely maximized as a function of σ^2 by $\hat{\sigma}^2$. Plugging $\hat{\sigma}^2$ into the right-hand side of (2.32) yields (2.30). □

REMARK 1. Healy (1975) obtains MLE's of parameters of a model which includes (1.4) as a special case. His proof, which is different than that given here, can be used to yield a proof of Theorem 2.3. However, Healy does not consider either the uniqueness of the MLE's or their relation to OLSE's and GLSE's. He also assumes, without proof, that G_{11} and G_{22} are nonsingular.

REMARK 2. In the intercept model with $p = r = 1$, it is well known that the MLE of the scalar B can be expressed in the form

$$
(2.33) \qquad \hat{B} = \frac{s_{22} - s_{11} + [(s_{11} - s_{22})^2 + 4s_{12}^2]^{1/2}}{2s_{12}}, \qquad s_{12} \neq 0,
$$

where $S = (n-1)^{-1} W = ((s_{ij}))$ is the sample covariance matrix. That this formula yields the same result as (2.7) can be seen by noting that the eigenvalues d_1 and d_2 of W are solutions of the quadratic equation

$$
0 = d^2 - (\operatorname{tr}[W])d + |W| = d^2 - (n-1)(s_{11} + s_{22})d + (n-1)^2(s_{11}s_{22} - s_{12}^2),
$$

so that

$$
(2.34) \qquad d_1 = (n-1)\left\{ \frac{s_{11} + s_{22} + [(s_{11} + s_{22})^2 - 4(s_{11}s_{22} - s_{12}^2)]^{1/2}}{2} \right\}.
$$

Also the scalars $G_{11} = g_{11}$ and $G_{21} = g_{21}$ satisfy

$$
d_1 \begin{pmatrix} g_{11} \\ g_{21} \end{pmatrix} = W \begin{pmatrix} g_{11} \\ g_{21} \end{pmatrix} = (n-1) S \begin{pmatrix} g_{11} \\ g_{21} \end{pmatrix}.
$$

In particular, $s_{11}g_{11} + s_{12}g_{21} = (n-1)^{-1} d_1 g_{11}$, so that

$$
g_{21}g_{11}^{-1} = \frac{(n-1)^{-1} d_1 - s_{11}}{s_{12}},
$$

and, simplifying the expression in (2.34), the equality of $g_{21}g_{11}^{-1}$ and (2.33) is established. The relationship between (2.33) and the slope $g_{21}g_{11}^{-1}$ of the major axis (first principal component) of the probability ellipse generated by S was found independently by K. Pearson and C. Gini (see Moran (1971), page 237). It should be noted that in contrast to Moran's conjecture (pages 245–246), the solutions (2.7)–(2.10) of the likelihood equations in the case $r = p = 1$ (and in general for all r, all p) do *not* define a saddlepoint of the likelihood, but actually do maximize the likelihood, as shown by Thoerem 2.3.

3. Strong consistency. In this section, strong consistency properties of $\hat{\alpha}$, $\hat{B}$ and $\hat{\sigma}^2$ are discussed. No conditions are imposed on the distribution of the columns of E beyond those appearing in Assumption A. In particular, it is *not* assumed that the common distribution of the columns of E is multivariate normal, or even that this distribution is absolutely continuous with respect to Lebesgue measure. Thus (see the remark after Theorem 2.2), it is conceivable that for fixed n, G_{11}^{-1} and G_{22}^{-1} with positive probability can fail to exist, making $\hat{\alpha}$ and $\hat{B}$ undefined. Fortunately, for every sequence $\omega = (\mathbf{x}_1, \mathbf{x}_2, \cdots)$ of observations, except perhaps for a collection of such sequences having probability zero, there exists a sample size $n(\omega)$ such that G_{11}^{-1} and G_{22}^{-1} exist for all $n \geq n(\omega)$ (see Lemma 3.3).

In this and the following sections, reference is made to the following assumptions on the sequence $\mathbf{u}_{1i}$, $i = 1, 2, \cdots$.

ASSUMPTION B. $\lim_{n\to\infty} n^{-1}U_1U_1'$ exists.

ASSUMPTION C. The matrix

$$\Delta = \lim_{n\to\infty} n^{-1}U_1C_nU_1' \tag{3.1}$$

is positive definite.

Assumption B implies that Δ exists, but not that Δ is positive definite. Assumption B also implies the existence of the following quantities:

$$\boldsymbol{\mu} = \lim_{n\to\infty} n^{-1}U\mathbf{1}_n = \begin{pmatrix}\mathbf{0}\\ \alpha\end{pmatrix} + \begin{pmatrix}I_p\\ B\end{pmatrix}\boldsymbol{\mu}_1, \qquad \boldsymbol{\mu}_1 = \lim_{n\to\infty} n^{-1}U_1\mathbf{1}_n, \tag{3.2}$$

$$\tau = ((\tau_{ij})) = \lim_{n\to\infty} n^{-1}UC_nU' = \begin{pmatrix}I_p\\ B\end{pmatrix}\Delta(I_p, B'). \tag{3.3}$$

Let

$$\Theta = \sigma^2 I_{p+r} + \tau = \sigma^2 I_{p+r} + \begin{pmatrix}I_p\\ B\end{pmatrix}\Delta(I_p, B'). \tag{3.4}$$

LEMMA 3.1. *Under* (1.4), *Assumption* A *and Assumption* B,

$$\lim_{n\to\infty} n^{-1}W = \Theta, \quad \textit{with probability } 1 \text{ (w.p. 1)}. \tag{3.5}$$

PROOF. From (1.4) and (2.2)

$$n^{-1}W = n^{-1}EC_nE' + n^{-1}UC_nE' + n^{-1}EC_nU' + n^{-1}UC_nU'. \tag{3.6}$$

By Assumptions A and B, (3.3) and the SLLN,

$$\lim_{n\to\infty} n^{-1}[EC_nE' + UC_nU'] = \Theta, \qquad \text{w.p. 1}. \tag{3.7}$$

Thus, (3.5) holds if

$$\lim_{n\to\infty} n^{-1}UC_nE' = \mathbf{0}, \qquad \text{w.p. 1}. \tag{3.8}$$

Note that in the no-intercept model, $n^{-1}UC_nE' = n^{-1}UE'$, while in the intercept model

$$n^{-1}UC_nE' = n^{-1}UE' - \bar{\mathbf{u}}\bar{\mathbf{e}}',$$

where

$$\bar{\mathbf{u}} = n^{-1}U\mathbf{1}_n, \qquad \bar{\mathbf{e}} = n^{-1}E\mathbf{1}_n. \tag{3.9}$$

By Assumption A, the SLLN and (3.2),

$$\lim_{n\to\infty} \bar{\mathbf{u}} = \boldsymbol{\mu}, \quad \lim_{n\to\infty} \bar{\mathbf{e}} = \mathbf{0}, \qquad \text{w.p. 1}. \tag{3.10}$$

Thus, under either model, (3.8) holds if $\lim_{n\to\infty} n^{-1}UE' = \mathbf{0}$, w.p. 1, or equivalently if

$$\lim_{n\to\infty} n^{-1}\textstyle\sum_{k=1}^n u_{ik}e_{jk} = 0, \qquad \text{w.p. 1}, \quad \text{all} \quad i, j = 1, 2, \cdots, p + r, \tag{3.11}$$

where $U = ((u_{ik}))$, $E = ((e_{jk}))$.

From Assumption B, the limit of $n^{-1}\sum_{k=1}^n u_{ik}^2$ exists ($n \to \infty$). Using Abel's partial summation formula, it can be shown that the existence of this limit implies that

$$\sigma^2 \textstyle\sum_{k=1}^\infty k^{-2}u_{ik}^2 = \sum_{k=1}^n k^{-2}\mathrm{Var}(u_{ik}e_{jk}) < \infty.$$

Applying the corollary to Theorem 5.4.1 in Chung (1974) establishes (3.11), and thus (3.4). □

REMARK. It follows directly from (1.4), (3.9) and (3.10) that

$$\lim_{n\to\infty} \bar{\mathbf{x}} = \lim_{n\to\infty}(\bar{x}_1', \bar{x}_2')' = \boldsymbol{\mu} = \begin{pmatrix} I_p \\ B \end{pmatrix}\boldsymbol{\mu}_1 + \begin{pmatrix} \mathbf{0} \\ \alpha \end{pmatrix}, \quad \text{w.p. 1.} \tag{3.12}$$

Let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p \geq 0$ be the (ordered) eigenvalues of $\Delta(I_p + B'B)$, and let $D_\gamma = \text{diag}(\gamma_1, \gamma_2, \cdots, \gamma_p)$. Using the fact that the γ_i's are also eigenvalues of $(I_p + B'B)^{1/2}\Delta[(I_p + B'B)^{1/2}]'$ for any square root $(I_p + B'B)^{1/2}$ of $I_p + B'B$, it can be shown that there exists $\psi: p \times p$ nonsingular for which

$$\Delta(I_p + B'B)\psi = D_\gamma, \qquad \psi'(I_p + B'B)\psi = I_p. \tag{3.13}$$

Let $\Gamma(B)$ be defined by (2.18). Then

$$\Theta\begin{pmatrix} I_p \\ B \end{pmatrix}\psi = \begin{pmatrix} I_p \\ B \end{pmatrix}\psi(\sigma^2 I_p + D_\gamma), \qquad \Theta\Gamma'(B) = \Gamma'(B)\sigma^2 I_r. \tag{3.14}$$

Hence the (ordered) eigenvalues $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_{p+r} \geq 0$ of Θ are

$$\theta_i = \sigma^2 + \gamma_i, \quad i = 1, 2, \cdots, p; \qquad \theta_{p+j} = \sigma^2, \quad j = 1, 2, \cdots, r. \tag{3.15}$$

From (3.13) and (3.14), it follows that the rows of $\psi'(I_p, B')$ are orthonormal eigenvectors of Θ corresponding to the largest p eigenvalues $\theta_1, \cdots, \theta_p$ of Θ.

LEMMA 3.2. *Under the assumptions of Lemma* 4.1,

$$\lim_{n\to\infty} n^{-1}D = D_\theta = \text{diag}(\theta_1, \theta_2, \cdots, \theta_{p+r}), \quad \text{w.p. 1.} \tag{3.16}$$

PROOF. The result (3.16) follows from Lemma 3.1, and the fact that the eigenvalues d_i, $i = 1, 2, \cdots, p + r$ of the matrix W are continuous functions of the elements of W. □

COROLLARY 3.1. *Under the assumptions of Lemma* 3.1,

$$\lim_{n\to\infty} n^{-1}D_{\min} = \sigma^2 I_r, \quad \text{w.p. 1,} \tag{3.17}$$

and thus

$$\lim_{n\to\infty} r^{-1}(p + r)\hat{\sigma}^2 = \sigma^2, \quad \text{w.p. 1.} \tag{3.18}$$

PROOF. The result (3.18) is a direct consequence of (3.17), which in turn directly follows from (3.15) and (3.16). □

In the proof of Lemma 3.3, $\omega = (\mathbf{x}_1, \mathbf{x}_2, \cdots)$ is a point in the probability space of all sequences of observations, and a superscript n on a sample quantity (e.g., $W^{(n)}$) indicates that that quantity is calculated from the first n elements of ω. This notation is discontinued once Lemma 3.3 is proven.

LEMMA 3.3. *Let* (1.4) *and Assumptions* A, B, *and* C *hold. With probability one, for each* ω *there exists* $n(\omega)$, $0 < n(\omega) < \infty$, *such that* $G_{11}^{(n)}$ *and* $G_{22}^{(n)}$ *are nonsingular and* $\hat{B}^{(n)} = G_{21}^{(n)}(G_{11}^{(n)})^{-1}$ *exists, all* $n \geq n(\omega)$. *In addition,*

$$\lim_{n\to\infty} \hat{B}^{(n)} = \lim_{n\to\infty} \hat{B} = B, \quad \text{w.p. 1.} \tag{3.19}$$

PROOF. Fix ω such that (3.5) and (3.16) hold. Note that since each $G^{(n)}$ is orthogonal, the sequence $\{G^{(n)}\}$ lies in a compact subspace of $(p + r)^2$ − dimensional Euclidean space. Thus, each subsequence of $\{G^{(n)}\}$ has a convergent sub-subsequence with limit, say,

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}.$$

Since each $G^{(n)}$ is orthogonal, so is H. For all n,

$$n^{-1}W^{(n)}\begin{pmatrix} G_{11}^{(n)} \\ G_{21}^{(n)} \end{pmatrix} = \begin{pmatrix} G_{11}^{(n)} \\ G_{21}^{(n)} \end{pmatrix}(n^{-1}D_{\max}^{(n)}),$$

and taking limits over the indices of the sub-subsequence on both sides of this equality, using (3.5), (3.15), (3.16), yields:

$$\Theta\begin{pmatrix} H_{11} \\ H_{21} \end{pmatrix} = \begin{pmatrix} H_{11} \\ H_{21} \end{pmatrix}(\sigma^2 I_p + D_\gamma).$$

Thus $(H_{11}', H_{21}')'$ is in the eigensubspace corresponding to the largest p eigenvalues of Θ. Since Assumption C implies that

$$\theta_p = \sigma^2 + \gamma_p > \sigma^2 = \theta_{p+1},$$

this eigensubspace is unique. Hence, it follows from (3.14) and the orthogonality of H that there exists $\Xi : p \times p$ orthogonal such that

$$\begin{pmatrix} H_{11} \\ H_{21} \end{pmatrix} = \begin{pmatrix} I_p \\ B \end{pmatrix}\psi\Xi. \tag{3.20}$$

Note that the limit of $(G_{11}^{(n)})(G_{11}^{(n)})'$ over the indices of any convergent sub-subsequence of $\{G^{(n)}\}$ is

$$H_{11}H_{11}' = \psi\Xi\Xi'\psi' = \psi\psi' = (I_p + B'B)^{-1},$$

as can be seen from (3.20), the fact that Ξ is orthogonal, and (3.13). Since this limit is independent of the sub-subsequence, facts about limits of sequences in Euclidean space imply that

$$\lim_{n\to\infty}(G_{11}^{(n)})(G_{11}^{(n)})' = (I_p + B'B)^{-1}. \tag{3.21}$$

Using the fact that the determinant of the limit $(I_p + B'B)^{-1}$ is positive and (3.21), it is now straightforward to show that there exists $n(\omega)$, $0 \le n(\omega) < \infty$, such that $G_{11}^{(n)}$ is nonsingular for all $n \ge n(\omega)$, proving the first part of the lemma. Equation (3.19) is established by a similar argument, noting first that for all convergent sub-subsequences of $\{G^{(n)}\}$ that $(G_{11}^{(n)})^{-1}$ eventually exists, and then noting from (3.20) that the limit of $\hat{B}^{(n)} = G_{21}^{(n)}(G_{11}^{(n)})^{-1}$ is $H_{21}H_{11}^{-1} = B$ independent of the sub-subsequence. □

COROLLARY 3.2. *Under the assumptions of Lemma* 3.3,

$$\lim_{n\to\infty}\hat{\alpha} = \alpha, \qquad \text{w.p. } 1. \tag{3.22}$$

PROOF. For the no-intercept model, $\hat{\alpha} = \mathbf{0} = \alpha$ and (3.22) trivially follows. For the intercept model, (3.22) is a direct consequence of (2.9), (3.12) and (3.19). □

In Section 4, it is shown that Δ, defined by (3.1), helps to determine the covariance matrix of the asymptotic distributions of $n^{1/2}(\hat{B} - B)$ and $n^{1/2}(\hat{\alpha} - \alpha)$. For purposes of constructing large-sample approximate confidence regions for B and α, it is therefore desirable to find a consistent sequence of estimators for Δ. Let

$$\begin{aligned} \hat{\Delta} &= n^{-1}(I_p + \hat{B}'\hat{B})^{-1}\left[(I_p, \hat{B}')W\begin{pmatrix} I_p \\ \hat{B} \end{pmatrix} - r^{-1}(p + r)\hat{\sigma}^2(I_p + \hat{B}'\hat{B})\right](I_p + \hat{B}'\hat{B})^{-1} \\ &= n^{-1}[G_{11}D_{\max}G_{11}' - r^{-1}\operatorname{tr}[D_{\min}]G_{11}G_{11}']. \end{aligned} \tag{3.23}$$

The second equality in (3.23) is a consequence of (2.5), (2.6), (2.7) and (2.10).

LEMMA 3.4. *Under the assumptions of Lemma* 3.3, *the estimator* $\hat{\Delta}$ *is a strongly consistent estimator of* Δ; *that is,*

$$\lim_{n\to\infty}\hat{\Delta} = \Delta, \qquad \text{w.p. 1.} \tag{3.24}$$

PROOF. The result (3.24) follows directly from (3.4), (3.5), (3.18), (3.19) and (3.23). □

REMARK 1. From (3.1), it might be conjectured that

$$n^{-1}\hat{U}_1 C_n \hat{U}_1' = n^{-1}G_{11}D_{\max}G_{11}'$$

defines a consistent sequence of estimators for Δ. However, from (3.18), (3.21), (3.23) and (3.24),

$$\begin{aligned}\lim n^{-1}\hat{U}_1 C_n \hat{U}_1' &= \lim_{n\to\infty} n^{-1}G_{11}D_{\max}G_{11}' \\ &= \Delta + \sigma^2(I_p + B'B)^{-1}, \qquad \text{w.p. 1,}\end{aligned} \tag{3.25}$$

so that this conjecture is false.

REMARK 2. Using (2.5), (2.6), and (2.7), it can be shown that when G_{22}^{-1} exists,

$$\begin{aligned}(B, -I_r)(n^{-1}W)\begin{pmatrix} B' \\ -I_r \end{pmatrix} &= (I_r + \hat{B}B')'[n^{-1}G_{22}D_{\min}G_{22}'](I_r + \hat{B}B') \\ &\quad + (\hat{B} - B)[n^{-1}G_{11}D_{\max}G_{11}'](\hat{B} - B)'.\end{aligned} \tag{3.26}$$

Taking limits as $n \to \infty$ in the equality (3.26), using (3.4), (3.5), (3.19) and (3.25), yields

$$\lim_{n\to\infty} n^{-1}G_{22}D_{\min}G_{22}' = \sigma^2(I_r + BB')^{-1}, \qquad \text{w.p. 1.} \tag{3.27}$$

4. Asymptotic distributions. From (3.6), the mean of W is

$$\mathscr{E}(W) = n\sigma^2 I_{p+r} + UC_nU', \tag{4.1}$$

where $\mathscr{E}(\cdot)$ is the expected value operator. Assume that the conditions of Lemma 3.3 hold. For $n \geq n(\omega)$, where $n(\omega)$ is defined by Lemma 3.3, G_{11}^{-1} exists, and from (2.5), (2.6), (2.7) and (4.1),

$$\begin{aligned}(I_p, B')[n^{-1/2}(W - \mathscr{E}(W))](B, -I_r)' &= [n^{1/2}(\hat{B} - B)]'(n^{-1}G_{22}D_{\min}G_{22}')(I_r + \hat{B}B') \\ &\quad - (I_p + B'\hat{B})(n^{-1}G_{11}D_{\max}G_{11}')[n^{1/2}(\hat{B} - B)]'.\end{aligned} \tag{4.2}$$

LEMMA 4.1. *If* $n^{-1/2}(W - \mathscr{E}(W))$ *has an asymptotic* (*as* $n \to \infty$) *distribution, and Assumptions* A, B, C *hold, then the asymptotic distributions of* $n^{1/2}(\hat{B} - B)$ *and*

$$F = -\Delta^{-1}(I_p + B'B)^{-1}(I_p, B')[n^{-1/2}(W - \mathscr{E}(W))](B, -I_r)' \tag{4.3}$$

are identical.

PROOF. The assertion of Lemma 4.1 is a direct consequence of (3.19), (3.25), (3.27), and (4.2). □

From (2.6), (2.7), (2.10) and (4.1),

$$\begin{aligned}rn^{1/2}[r^{-1}(p + r)\hat{\sigma}^2 - \sigma^2] &= \operatorname{tr}\{G_{22}G_{22}'(\hat{B}, -I_r)[n^{-1/2}(W - \mathscr{E}(W))](\hat{B}, -I_r)'\} \\ &\quad + \operatorname{tr}\{G_{22}G_{22}'[n^{1/2}(\hat{B} - B)][n^{-1}U_1C_nU_1'][\hat{B} - B]'\}.\end{aligned} \tag{4.4}$$

LEMMA 4.2. *Under the conditions of Lemma* 4.1, *the asymptotic distributions of* $n^{1/2}(r^{-1}(p + r)\hat{\sigma}^2 - \sigma^2)$ *and*

$$v = r^{-1}\operatorname{tr}\{(I_r + BB')^{-1}(B, -I_r)[n^{-1/2}(W - \mathscr{E}(W))](B, -I_r)'\} \tag{4.5}$$

are identical.

PROOF. From (2.5), (2.7), (3.19) and (3.21),

$$\begin{aligned}\lim_{n\to\infty} G_{22}G_{22}' &= \lim_{n\to\infty}(I_r - G_{21}G_{21}') = \lim_{n\to\infty}(I_r - \hat{B}G_{11}G_{11}'\hat{B}') \\ &= I_r - B(I_p + B'B)^{-1}B' = (I_r + BB')^{-1}, \qquad \text{w.p. } 1.\end{aligned} \tag{4.6}$$

From Lemma 4.1, $n^{1/2}(\hat{B} - B)$ has an asymptotic distribution. Thus, (3.1) and (4.6) imply that the second term on the right side of (4.4) is $o_p(1)$ as $n \to \infty$. Thus, from (3.19) and (4.6), the assertion of Lemma 4.2 follows. □

LEMMA 4.3. *In the intercept model, assume that $n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}})$ and $n^{-1/2}(W - \mathscr{E}(W))$ have an asymptotic joint distribution, and that Assumptions* A, B, C *hold. Then the asymptotic distributions of $n^{1/2}(\hat{\alpha} - \alpha)$ and*

$$\mathbf{m} = (-B, I_r)[n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}})] - F\boldsymbol{\mu}_1 \tag{4.7}$$

are identical, where F is defined by (4.3).

PROOF. Observe that from (2.9),

$$n^{1/2}(\hat{\alpha} - \alpha) = (-B, I_r)[n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}})] - [n^{1/2}(\hat{B} - B)]\bar{\mathbf{x}}_1. \tag{4.8}$$

The assertion of Lemma 4.3 is now a consequence of (3.12) and Lemma 4.1. □

Lemmas 4.1 through 4.3 motivate consideration of the asymptotic joint distribution of $n^{-1/2}(W - \mathscr{E}(W))$ and $n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}})$. For this asymptotic distribution to exist, it is sufficient that the following assumption holds.

ASSUMPTION D. Let the random vector $\mathbf{e} = (e_1, e_2, \cdots, e_{p+r})'$ have the common distribution of the columns of E. Then the elements of $\mathbf{e}$ have finite fourth moments: $\mathscr{E}(e_i) < \infty$, $i = 1, 2, \cdots, p + r$.

LEMMA 4.4. *Let $\mathbf{y}_1, \mathbf{y}_2, \cdots$ be a sequence of mutually independent s-dimensional random vectors, where $\mathbf{y}_i$ has mean vector $\mathbf{0}$ and finite covariance matrix V_i. If $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n V_i = V^*$ exists (and is finite), then*

$$n^{1/2}\bar{\mathbf{y}} = n^{-1/2}\sum_{i=1}^n \mathbf{y}_i \to_{\mathscr{L}} MVN(\mathbf{0}, V^*). \tag{4.9}$$

PROOF. Consider any linear combination

$$\mathbf{c}'[n^{1/2}\bar{\mathbf{y}}] = n^{-1/2}\sum_{i=1}^n (\mathbf{c}'\mathbf{y}_i)$$

of the elements of $n^{1/2}\mathbf{y}$. By the given

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \operatorname{Var}(\mathbf{c}'\mathbf{y}_i) = \mathbf{c}'\mathrm{V}^*\mathbf{c}.$$

From this result, the independence of the $\mathbf{c}'\mathbf{y}_i$'s, Markov's inequality, and Corollary 2 and Theorem 4 of Gnedenko and Kolmogorov (1954), pages 141, 143, it follows that for all $\mathbf{c}$,

$$\mathbf{c}'(n^{1/2}\bar{\mathbf{y}}) \to_{\mathscr{L}} N(0, \mathbf{c}'V^*\mathbf{c}).$$

From this last result, (4.9) directly follows. □

Define

$$Z_k = \mathbf{e}_k\mathbf{e}_k' - \sigma^2 I_{p+r} + (\mathbf{u}_k - \boldsymbol{\mu})\mathbf{e}_k' + \mathbf{e}_k(\mathbf{u}_k - \boldsymbol{\mu})', \tag{4.10}$$

where $\mathbf{e}_k$ is the kth column of E, $\mathbf{u}_k$ is the kth column of U, and $\boldsymbol{\mu}$ is defined by (3.2); $k = 1, 2, \cdots$. The Z_k's are independent (but not identically distributed) $(r+p)$-dimensional symmetric random matrices.

COROLLARY 4.1. *Under Assumptions* A, B, *and* D, *the elements of* $n^{-1/2}\sum_{k=1}^{n} Z_k$ *which are on or below the diagonal, together with the elements of* $n^{1/2}\,\bar{\mathbf{e}} = n^{-1/2}\sum_{k=1}^{n}\mathbf{e}_k$, *have an asymptotic* $\frac{1}{2}(p+r+3)(p+r)$*-variate normal distribution with means O and covariance matrix*

$$V^* = \begin{pmatrix} V_{11}^* & V_{12}^* \\ V_{21}^* & V_{22}^* \end{pmatrix}$$

defined as follows. Let δ_{ij} *be the Kronecker delta and* τ_{ij} *be defined by* (3.3). *Then*

(i) $V_{11}^* = ((v_{(i,j),(i',j')}^*))$ *gives the asymptotic covariance between the* (i,j)th *and* (i',j')th *elements of* $n^{-1/2}\sum_{k=1}^{n} Z_k$, $i \le j$, $i' \le j'$;

$$v_{(i,j),(i',j')}^* = \mathscr{E}(e_i e_j e_{i'} e_{j'}) - \sigma^4\delta_{ij}\delta_{i'j'} + \sigma^2(\tau_{ii'}\delta_{jj'} + \tau_{jj'}\delta_{ii'} + \tau_{ij'}\delta_{i'j} + \tau_{i'j}\delta_{ij'}); \tag{4.11}$$

(ii) $V_{12}^* = (V_{21}^*)'$ *gives the asymptotic covariance between the* (i,j)th *element of* $n^{-1/2}\sum_{k=1}^{n} Z_k$ *and the* lth *element of* $n^{1/2}\bar{\mathbf{e}}$. *If* $V_{12}^* = ((v_{(i,j),l}^*))$, *then*

$$v_{(i,j),l}^* = \mathscr{E}(e_i e_j e_l). \tag{4.12}$$

(iii) $V_{22}^* = \sigma^2 I_{p+r}$ *is the asymptotic covariance matrix of* $n^{1/2}\bar{\mathbf{e}}$.

PROOF. Let $\mathbf{y}_k$ be the $\frac{1}{2}(p+r+3)(p+r)$-dimensional random vector whose first $\frac{1}{2}(p+r+1)(p+r)$ elements are the elements on and below the diagonal of Z_k arranged (say) in lexicographic order, and whose last $(p+r)$ elements are the elements of $\mathbf{e}_k$. The assertion of Corollary 4.1 follows from Assumptions A and D, (3.2), (3.3) and Lemma 4.4. □

Note from (3.6) and (4.1) that under the no-intercept model (when $C_n = I_n$),

$$\begin{aligned} n^{-1/2}(W - \mathscr{E}(W)) &= n^{-1/2}(EE' - n\sigma^2 I_{p+r} + UE' + EU') \\ &= n^{-1/2}\textstyle\sum_{k=1}^{n}[Z_k + \boldsymbol{\mu}\mathbf{e}_k' + \mathbf{e}_k\boldsymbol{\mu}']; \end{aligned} \tag{4.13}$$

while under the intercept model ($C_n = I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$),

$$\begin{aligned} n^{-1/2}(W - \mathscr{E}(W)) &= n^{-1/2}\textstyle\sum_{k=1}^{n} Z_k + (\boldsymbol{\mu} - \bar{\mathbf{u}})(n^{1/2}\bar{\mathbf{e}})' + (n^{1/2}\bar{\mathbf{e}})(\boldsymbol{\mu} - \bar{\mathbf{u}})' \\ &= n^{-1/2}\textstyle\sum_{k=1}^{n} Z_k + o_p(1), \end{aligned} \tag{4.14}$$

since, by Corollary 4.1, $n^{1/2}\bar{\mathbf{e}} \to_L \mathrm{MVN}(\mathbf{0}, \sigma^2 I_{p+r})$ and since, by (3.2), $\lim_{n\to\infty}(\bar{\mathbf{u}} - \boldsymbol{\mu}) = \mathbf{0}$. Since also $n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}}) = n^{1/2}\bar{\mathbf{e}}$, the following theorem is a direct consequence of (4.13), (4.14) and Corollary 4.1.

THEOREM 4.1. *Under Assumptions* A, B, *and* D, *the following results hold*:

(i) *In the no-intercept model, the elements of* $n^{-1/2}(W - \mathscr{E}(W))$ *on or below the diagonal have an asymptotic* $\frac{1}{2}(p+r+1)(p+r)$*-variate normal distribution with zero means and with covariance between the* (i,j)th *and* (i',j')th *elements given by*

$$\begin{aligned} &v_{(i,j),(i',j')}^* + \mu_i\mathscr{E}(e_j e_{i'} e_{j'}) + \mu_j\mathscr{E}(e_i e_{i'} e_{j'}) \\ &\qquad + \mu_{i'}\mathscr{E}(e_{j'} e_i e_j) + \mu_{j'}\mathscr{E}(e_i e_j e_{i'}) \end{aligned} \tag{4.15}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_{p+r})'$.

(ii) *In the intercept model, the elements of* $n^{-1/2}(W - \mathscr{E}(W))$ *on or below the diagonal, together with the elements of* $n^{1/2}(\bar{\mathbf{x}} - \bar{\mathbf{u}})$, *have an asymptotic* $\frac{1}{2}(p+r+3)(p+r)$*-variate normal distribution with means zero and with covariance matrix* V^*. *Thus, for example,* $V_{11}^* = ((v_{(i,j),(i',j')}^*))$ *gives the asymptotic covariances between the* (i,j)th *and* (i',j')th *elements of* $n^{-1/2}(W - \mathscr{E}(W))$.

From Theorem 4.1 and Lemmas 4.1 through 4.3, the asymptotic distributions of $n^{1/2}(\hat{B} - B)$ and $n^{1/2}(r^{-1}(p + r)\hat{\sigma}^2 - \sigma^2)$ in the no-intercept model, and the asymptotic distributions of $n^{1/2}(\hat{\alpha} - \alpha)$, $n^{1/2}(\hat{B} - B)$ and $n^{1/2}(r^{-1}(p + r)\hat{\sigma}^2 - \sigma^2)$ in the intercept model, can be straightforwardly obtained. These asymptotic distributions are all (multivariate) normal with zero means, but the covariance matrices of these distributions involve third-order and fourth-order cross moments of **e**, and thus are both complicated in form and hard to estimate from data. The following assumption is therefore imposed to simplify results, and permit construction of large-sample approximate confidence regions for the parameters of the models.

ASSUMPTION E. The cross-moments of the common distribution of the columns of E are identical, up to and including moments of order four, to the corresponding moments of the MVN($\mathbf{0}$, $\sigma^2 I_{p+r}$) distribution.

THEOREM 4.2. *Assume that Assumptions* A, B, C, *and* E *hold. Under both the no-intercept and intercept models*

(i) $n^{1/2}(r^{-1}(p+ r)\hat{\sigma}^2 - \sigma^2) \to_L N(0, 2r^{-1}\sigma^4)$;

(ii) *the elements of* $n^{1/2}(\hat{B} - B)$ *have an asymptotic rp-variate normal distribution with zero means and covariance between the* (i, j)th *and* (i', j')th *elements given by* $\sigma^2[\sigma^2\Delta^{-1}(I_p + B'B)^{-1}\Delta^{-1} + \Delta^{-1}]_{jj'}[I_r + BB']_{ii'}$;

(iii) *in the intercept model*

$$n^{1/2}(\hat{\alpha} - \alpha) \to_L \mathrm{MVN}(\mathbf{0}, \rho(I_r + BB')),$$

where

$$\rho = \sigma^2\{1 + \mu_1'[\sigma^2\Delta^{-1}(I_p + B'B)^{-1}\Delta^{-1} + \Delta^{-1}]\mu_1\},$$

and further the elements of $n^{1/2}(\hat{B} - B)$ *and* $n^{1/2}(\hat{\alpha} - \alpha)$ *have an asymptotic* $r(p + 1)$*-variate normal distribution where the asymptotic covariance between the* (i, j)th *element of* $n^{1/2}(\hat{B} - B)$ *and the* lth *element of* $n^{1/2}(\hat{\alpha} - \alpha)$ *is*

$$\sigma^2(I_r + B'B)_{il}\{[\sigma^2\Delta^{-1}(I_p + B'B)^{-1}\Delta^{-1} + \Delta^{-1}]\mu_1\}_j.$$

In using the above results, it should be remembered that Δ is defined by (3.1) in a different fashion in the no-intercept and intercept models.

PROOF. Assumption E implies Assumption D. The assertions of the theorem now follow from Lemmas 4.1 through 4.3, Theorem 4.1, and straightforward calculation of the covariances. □

Let $\chi_t^2[1 - \nu]$ be the $100(1 - \nu)$th percentile of the chi-squared distribution with t degrees of freedom. Note that

$$\sigma^2\Delta^{-1}(I_p + B'B)^{-1}\Delta^{-1}+\Delta^{-1} = \Delta^{-1}[\Delta + \sigma^2(I_p + B'B)^{-}\ ]\Delta^{-1} \tag{4.16}$$

is consistently estimated by $\hat{\Delta}^{-1}[n^{-1}G_{11}D_{\max}G_{11}']\hat{\Delta}^{-1}$, as can be seen from (3.24) and (3.25). The following results now follow from Theorem 4.2, (3.12), (3.18) and (3.19).

COROLLARY 4.2. *Under the assumptions of Theorem* 4.2, *and under both the no-intercept and intercept models, a large-sample approximate* $100(1 - \nu)\%$ *confidence interval for* σ^2 *is*

$$\{\sigma^2: |\sigma^2 - (nr)^{-1}\mathrm{tr}[D_{\min}]| \le (n^{3/2}r)^{-1}\mathrm{tr}[D_{\min}](2\chi_1^2[1 - \nu])^{1/2}\}.$$

COROLLARY 4.3. *Under the assumptions of Theorem* 4.2, *and under both the no-intercept and intercept models, a large-sample approximate* $100(1 - \nu)\%$ *confidence region for the elements of B is*

$$\{B:\mathrm{tr}[n(I_r + \hat{B}\hat{B}')^{-1}(\hat{B} - B)\hat{\Delta}(n^{-1}G_{11}D_{\max}G_{11}')^{-1}\hat{\Delta}(\hat{B} - B)'] \leq r^{-1}(p + r)\hat{\sigma}^{2}\chi^2_{rp}[1 - \nu]\},$$

which is equivalent to the following computationally more convenient region:

$$\{B:\mathrm{tr}[(I_r + \hat{B}\hat{B}')^{-1}(\hat{B} - B)G_{11}D_{\max}^{-1}(D_{\max} - \mathrm{tr}[r^{-1}D_{\min}]I_p)^2G_{11}'((\hat{B} - B)'] \leq (rn)^{-1}\mathrm{tr}[D_{\min}]\chi^2_{rp}[1 - \nu]\}.$$

COROLLARY 4.4. *In the intercept model, under the assumptions of Theorem 4.2 a $100(1 - \nu)\%$ confidence region for α is*

$$\{\alpha: n(\hat{\alpha} - \alpha)'(I_r + \hat{B}\hat{B}')^{-1}(\hat{\alpha} - \alpha) \leq \hat{\rho}\chi^2_p[1 - \nu]\},$$

where

$$\hat{\rho} = (nr)^{-1}\mathrm{tr}[D_{\min}]\{1 + \bar{\mathbf{x}}_1'G_{11}D_{\max}^{-1}(D_{\max} - \mathrm{tr}[r^{-1}D_{\min}]I_p)^2G_{11}'\bar{\mathbf{x}}_1\}.$$

Theorem 4.2 can also be used to construct a large-sample joint confidence region for the elements of α and B in the intercept model; however, for reasons of space, and since in practice α and B are usually studied separately, the formula for such a confidence region is omitted.

Because of their relatively simple form, the large-sample confidence regions of Corollaries 4.2 through 4.4 have practical appeal. However, potential users of these regions should be aware that errors-in-variables models frequently behave counter to the maxims of large-sample theory developed for parametrically more regular models. For example, in the case $r = p = 1$ with Assumption E strengthened to require multivariate normality of the columns of the error matrix E, Theorem 4.2 shows that $\hat{B}$ (a scalar in this case) has finite asymptotic variance. Nevertheless, *for any sample size n* (no matter how large), *the exact variance of $\hat{B}$ is infinite;* indeed, $\mathscr{E}(\hat{B})$ is not well-defined (Anderson (1976)). This fact should not dissuade practitioners from considering the use of the confidence region for B in Corollary 4.3, but it would be wise to be cautious in assigning an exact confidence value to such a region, at least until analytic and/or simulation studies in fixed-sample situations give more insight into the exact properties of that region.

5. Comments on assumptions. The major results of this paper require two assumptions for their validity;

(i) that Δ exists and is positive definite (Assumption C).

(ii) that the covariance matrix, Σ_e of the errors equals $\sigma^2 I_{p+r}$.

It is, of course, open to argument whether either of these assumptions can be even approximately valid in practical contexts.

Of the above two assumptions, assumption (i) is the easiest to justify. First, in econometric and psychometric applications of "errors in variables" models, it is often assumed that the sequence $\{\mathbf{u}_{1i}, i = 1, 2, \cdots\}$ is obtained from independent observations of a random vector $\mathbf{u}_1$ having a distribution of known form (usually normal) with a finite covariance matrix Σ_u. As Moran (1971, page 246) remarks, this assumption converts the model from a functional equation model into a *structural* equation model, and thus changes the nature of the statistical problem. Nevertheless, it is worth noting that this assumption implies that assumption (i) holds with probability one through application of the SLLN.

Loosely speaking, assumption (i) requires that the $\mathbf{u}_{1i}$ observations sample all directions in p-dimensional space with relative frequencies (over the sequence $\{\mathbf{u}_{1i}, i = 1, 2, \cdots\}$) of commensurate magnitude. Certainly, any well-designed classical response surface experiment would require that the observations on the vector of independent variables have this property, and in the present model the $\mathbf{u}_{1\iota}$'s play the role of the vector of independent variables. Using simulations in the case $p = r = 3$, Gleser and Watson (1973) under-sampled one dimension in 3-dimensional space. The resulting estimates of B bore little resemblance to the true value. On the other hand, when all dimensions were nearly equally sampled, the estimates of B were reasonably accurate even in moderately small samples.

With respect to assumption (ii), it should be noted that if the requirement in Assumption A that $\Sigma_e = \sigma^2 I_{p+r}$ is replaced by the more general requirement

$$\Sigma_e = \sigma^2 \Sigma_0, \qquad \Sigma_0 \text{ known}, \tag{5.1}$$

then transforming the data $\mathbf{x}_\iota$, $i = 1, 2, \cdots, n$ to new data $\Sigma_0^{-1/2}\, \mathbf{x}_\iota$, $i = 1, 2, \cdots, n$, where $\Sigma_0^{-1/2}$ has the form

$$\Sigma_0^{-1/2} = \begin{pmatrix} T_{11} & \mathbf{0} \\ T_{21} & T_{22} \end{pmatrix}; \qquad T_{11}\colon p \times p, \tag{5.2}$$

yields a model for the transformed data of the form (1.4) with covariance matrix $\Sigma_e = \sigma^2 I_{p+r}$. If α^* and B^* are the intercept vector and slope matrix of the new model, then the identities

$$\alpha^* = T_{22}\alpha, \qquad B^* = T_{21}T_{11}^{-1} + T_{22}BT_{11}^{-1}, \tag{5.3}$$

allow estimators for α^*, B^* based on the transformed data to be converted to estimators for α and B. Further, the estimator $\hat{\sigma}^2$ of σ^2 based on the original data is

$$\hat{\sigma}^2 = n^{-1}(p + r)^{-1} \textstyle\sum_{i=p+1}^{p+r} \lambda_\iota[\Sigma_0^{-1}\, W], \tag{5.4}$$

where $\lambda_\iota[\Sigma_0^{-1}\, W]$ is the ith largest singular value (in this case, eigenvalue) of $\Sigma_0^{-1}\, W$.

Thus, the basic question that must be asked about the practicality of using assumption (ii) is whether or not the form of the error covariance matrix Σ_e can be known up to a constant scalar multiple.

In the geophysical surveying problem mentioned in Gleser and Watson (1973), there is some justification for this assumption. First, prior experience with the errors involved in measuring the longitude, latitude, and altitude of a stake placed on a glacier can yield the matrix of correlations $P_0 = \Sigma_0$ among the three kinds of measurement error. By symmetry, one can argue that the error variances in measuring the three dimensions are approximately equal (to σ^2). Thus, $\mathscr{E}(\mathbf{e}_{1\iota}\mathbf{e}_{1\iota}') = \mathscr{E}(\mathbf{e}_{2\iota}\mathbf{e}_{2\iota}') = \sigma^2 P_0$. Further, since the first and second surveys of any stake are widely separated in time, the errors made in these two surveys should be independent of one another. Putting these arguments together,

$$\Sigma_e = \sigma^2 \begin{pmatrix} P_0 & 0 \\ 0 & P_0 \end{pmatrix}.$$

In some problems, such as in psychometric testing (Lord (1973)), it may be possible to obtain m_ι independent measurements on each $\mathbf{x}_\iota$, $i = 1, 2, \cdots, n$. In this case, Σ_e in its entirety can be estimated by the pooled sample covariance matrix S with $\Sigma_{i=1}^{n}\, (m_i - 1)$ degrees of freedom. However, rather than use the methods of this paper to analyze the data in such cases, it would be advisable to consult the basic paper by Anderson (1951), where the relevant maximum likelihood theory is worked out in full.

Turning back to the original problem (where $m_\iota = 1$, all i), the following theorem shows that unless some fairly strong assumptions relating $\Sigma_{11} = \mathscr{E}(\mathbf{e}_{1i}\mathbf{e}_{1i}')$ to $\Sigma_{22} = \mathscr{E}(\mathbf{e}_{2i}\mathbf{e}_{2\iota}')$ are required, the MLE approach cannot be applied to estimate α, B, U_1 and Σ_e.

THEOREM 5.1. *For the model* (1.4), *relax Assumption* A *to allow* Σ_e *to be an unknown* $(p + r) \times (p + r)$ *positive definite matrix in a class* $\mathscr{S}$ *containing matrices of the form*

$$\Sigma_e = \sigma^2 \begin{pmatrix} I_p & \mathbf{0} \\ \mathbf{0} & \kappa I_r \end{pmatrix} = \sigma^2\, \Sigma_0(\kappa) \tag{5.5}$$

where σ^2, κ *are unrestricted. Then if the common distribution of the columns of* E *is multivariate normal, the likelihood of the data has an infinite supremum.*

PROOF. This theorem generalizes a result of Anderson and Rubin (1956), pages 129–130. Note first that the supremum of the likelihood over α, B, U_1, and Σ_e in $\mathscr{S}$ is no less than the supremum of the likelihood over α, B, U_1 and Σ_e of the form (5.5). Assuming Σ_e has the form (5.5), fix κ, and use the transformation of the data mentioned at the start of the discussion of assumption (ii) along with Theorem 2.3 to maximize the likelihood $L(X;\, \alpha, B, U_1, \sigma^2, \kappa)$ over α, B, U_1 and σ^2. This yields $L(X;\, \hat{\alpha}, \hat{B}, \hat{U}_1, \hat{\sigma}^2, \kappa)$ proportional to

$$\left(\sum_{i=p+1}^{p+r} \lambda_i[(\Sigma_0(\kappa))^{-1} W]\right)^{-n(p+r)/2}$$

and it is easy to see that this last expression tends to infinity when $\kappa \to \infty$. □

REMARK 1. The arguments in the proof of Theorem 5.1 easily extend to cover cases in which Σ_e has the form

$$\Sigma_e = \begin{pmatrix} \kappa_1^{1/2} I_p & \mathbf{0} \\ \mathbf{0} & \kappa_2^{1/2} I_r \end{pmatrix} \Sigma_0 \begin{pmatrix} \kappa_1^{1/2} I_p & \mathbf{0} \\ \mathbf{0} & \kappa_2^{1/2} I_r \end{pmatrix}$$

where Σ_0 can either be fixed or allowed to vary over any class of positive definite matrices, and where $\kappa_1, \kappa_2 > 0$ are unrestricted.

REMARK 2. In the intercept model with $r = p = 1$ and the columns of E multivariate normally distributed with unknown covariance matrix Σ_e, one can partially differentiate the likelihood with respect to the parameters and arrive at solvable likelihood equations. From Theorem 5.1, it is seen that these equations do not yield MLE's, despite assertions to the contrary in the literature. Instead, as shown by Solari (1969), the solutions to these equations provide saddlepoints for the likelihood.

An even stronger argument than Theorem 5.1 against trying to fit "errors in variables" models in which Σ_e is completely unrestricted is provided by Nussbaum (1977). For the intercept model with $r = p = 1$ and the columns of E multivariate normally distributed, Nussbaum shows that *if Σ_e is unrestricted, no strongly consistent estimator for B can exist.* Nussbaum's arguments can be straightforwardly extended to the multivariate case ($r, p \geq 1$), both in the intercept and no-intercept models.

The above arguments should not be interpreted as stating that no models of the form (1.4) in which Σ_e is of a more general form than (5.1) can be analyzed. Bhargava (1977) discusses a model in the case $r = p$ where Σ_e is an unknown block diagonal matrix, and proves the existence of MLE's of the parameters of this model. The work of Anderson and Rubin (1956), and a considerable recent literature, suggests that allowing Σ_e to be an unknown diagonal matrix subject to certain restrictions adopted in factor analysis, and regarding the model (1.4) as a factor analysis model where certain elements of the matrix L of factor loadings have specified values (see Section 1), can lead to consistent methods for estimating α and B.

Finally, it should be noted that the methods of analysis of Sections 3 and 4 can be used to study robustness of the large-sample properties of $\hat{\alpha}$, $\hat{B}$ and $\hat{\sigma}^2$ when $\Sigma_e \neq \sigma^2 I_{p+r}$. In particular, it should be possible to identify a class of matrices Σ_e under which $\hat{B}$ remains a strongly consistent estimator of B. Some results concerning this problem are planned for a forthcoming paper.

Acknowledgment. I am indebted to three anonymous referees of this paper for their careful reading of this paper, and for valuable comments which helped me to improve its organization and broaden its content.

REFERENCES

[1] ADCOCK, R. J. (1878). A problem in least squares. *The Analyst* **5**, 53–54.

[2] ANDERSON, T. W. (1948). The asymptotic distribution of the roots of certain determinantal equations. *J. Roy. Statist. Soc., Ser. B* **10** 132–139.

[3] ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22** 327–351.

[4] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148.

[5] ANDERSON, T. W. (1976). Estimation of linear functional relationships: approximate distributions and connection with simultaneous equations in econometrics (with discussion). *J. Roy. Statist. Soc., Ser. B* **38** 1–36.

[6] ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **5** 111–150. Univ. California Press.

[7] BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

[8] BHARGAVA, A. K. (1977). Maximum likelihood estimation in a multivariate 'errors in variables' regression model with unknown error covariance matrix. *Comm. Statist. Theor. Meth.* **A6** 587–601.

[9] BHARGAVA, A. K. (1979). Estimation of a linear transformation and an associated distributional problem. *J. Statist. Planning and Inference* **3** 19–26.

[10] CHUNG, KAI LAI (1974). *A Course in Probability Theory.* Academic Press, New York.

[11] GEARY, R. (1948). Studies in relations between economic time series. *J. Roy. Statist. Soc., Ser. B* **10** 140–158.

[12] GLESER, L. J. and KUNTE, S. (1976). On asymptotically optimal sequential Bayes interval estimation procedures. *Ann. Statist.* **4** 685–711.

[13] GLESER, L. J. and WATSON, G. S. (1973). Estimation of a linear transformation. *Biometrika* **60** 525–534.

[14] GNEDENKO, V. and KOLMOGOROV, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* (K. L. Chung, translator). Addison-Wesley, Cambridge, Mass.

[15] HEALY, JOHN DOUGLAS (1975). Estimation and tests for unknown restrictions in multivariate linear models. Purdue Univ., Depart. Statist., Mimeo Series 471.

[16] LAWLEY, D. N. (1953). A modified method of estimation in factor analysis and some large sample results. In *Upsalla Symp. Psychological Factor Analysis.* 17–19. 35–42. Almqvist and Wicksell, Uppsala.

[17] LORD, FREDERIC M. (1973). Testing if two measuring procedures measure the same dimension. *Psych. Bull.* **79** 71–72.

[18] MADANSKY, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Assoc.* **54** 173–205.

[19] MORAN, P. A. P. (1971). Estimating structural and functional relationships. *J. Multivariate Anal.* **1** 232–255.

[20] NUSSBAUM, N. (1977). Asymptotic optimality of estimators of a linear functional relation if the ratio of the error variances is known. *Math. Op. Statist. Ser. Statist.* **8** 173–198.

[21] OKAMOTO, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.* **1** 763–765.

[22] OLKIN, I. (1966). Special topics in matrix theory. Lecture notes at Stanford Univ.

[23] SOLARI, M. E. (1969). The 'maximum likelihood solution' of the problem of estimating a linear functional relationship. *J. Roy. Statist. Soc., Ser. B.* **31** 372–375.

[24] SPRENT, P. (1966). A generalized least squares approach to linear functional relationships. *J. Roy. Statist. Soc., Ser. B* **28** 278–297.

[25] TINTNER, G. (1945). A note on rank, multicollinearity, and multiple regression. *Ann. Math. Statist.* **16** 304–308.

[26] VON NEUMANN, J. (1937). Some matrix inequalities and metrization of matric-space. *Tomsk. Univ. Rev.* **1** 286–300. (In. *Collected Works,* Vol. IV, Pergamon Press, 1962).

[27] WHITTLE, P. (1952). On principal components and least square methods of factor analysis. *Skand. Aktuaretidskr.* **35** 223–239.

[28] WILLIAMS, E. J. (1955). Significance tests for discriminant functions and linear functional relationships. *Biometrika* **42** 308–381.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, IN 47907