APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 2, 99–119 (1986)

Insurance, economics and finance
Statistical validation

THE POPULATION-SAMPLE DECOMPOSITION APPROACH TO MULTIVARIATE ESTIMATION METHODS

BERNARD M. S. VAN PRAAG
Econometric Institute, Erasmus University, Rotterdam, Netherlands

JAN DE LEEUW
Department of Data Theory, Leyden University, Leyden, Netherlands

AND

TEUN KLOEK
Econometric Institute, Erasmus University, Rotterdam, Netherlands

SUMMARY

In this paper it is argued that all multivariate estimation methods, such as OLS regression, simultaneous linear equations systems and, more widely, what are known as LISREL methods, have merit as geometric approximation methods, even if the observations are not drawn from a multivariate normal parent distribution and consequently cannot be viewed as ML estimators. It is shown that for large samples the asymptotical distribution of any estimator, being a totally differentiable covariance function, may be assessed by the δ method. Finally, we stress that the design of the sample and *a priori* knowledge about the parent distribution may be incorporated to obtain more specific results.

It turns out that some fairly traditional assumptions, such as assuming some variables to be non-random, fixed over repeated samples, or the existence of a parent normal distribution, may have dramatic effects on the assessment of standard deviations and confidence bounds, if such assumptions are not realistic. The method elaborated by us does not make use of such assumptions.

KEY WORDS Multivariate estimation Large-sample theory Population-sample decomposition δ-method Curve-fitting

1. INTRODUCTION

Multivariate statistical data analysis is built around the linear model. In econometrics linear models are studied for the *dependence* of a variable on a number of other variables; in psychometrics linear models for the *interdependence* of variables get more attention. Historically the role of the linear model is even more preponderant in agricultural science and biometrics. Although the specific variants differ, it is sensible to speak about a general structure underlying linear statistical relationships. As Anderson[1] shows, there is a common structure which can be used either for linear regression models popular in econometrics of for factor-analytical approaches popular in psychology. In all cases the parameter estimators in these models are functions of the covariance matrix of the observations, so we shall call them *covariance functions.*

This then is the common structure of the techniques discussed in this paper. We are interested

8755–0024/86/030099–21$10.50

Received 29 August 1985

in statistics which are functions of the covariance matrix. These can be regression coefficients, factor loadings, eigenvalues, canonical correlations, and so on. According to our definition these are all covariance functions. In order to study their statistical properties it is necessary to specify some properties of both the sample covariances and the specific functions involved. Statistical analysis of data is, of course, usually based on a number of assumptions. In our view these assumptions are often unnecessarily restrictive. In this paper we shall try to show that it is not at all necessary to maintain all of these assumptions in order to derive statistically meaningful results.

By way of intuitive introduction let us consider a simple case first. Suppose that $\{(y_t, z_t)\}_{t=1}^{T}$ is a set of T bivariate observations in R^2. A first question in this context is how we can draw a straight line $y = b_1 z + b_0$ through the observations, represented as points in the plane, such that the line reasonably fits the observations. The standard answer to this question was proposed at the beginning of the 19th century by Lagrange, Gauss and Laplace. As a criterion of fit it uses the sum of squared residuals

$$\Delta(b_0, b_1) = \frac{1}{T} \sum_{t=1}^{T} (y_t - b_1 z_t - b_0)^2 \tag{1}$$

which is minimized with respect to the vector $\mathbf{b} = (b_0, b_1)$. The minimizing solution is denoted by $\hat{\mathbf{b}}$. The 'calculated' counterpart of y_t is $\hat{y}_t = \hat{b}_1 z_t + \hat{b}_0$, and the calculated residual is $(y_t - \hat{y}_t)$. The fitted line is traditionally called the *regression* line.

So far, we have not placed the problem, as presented above, in a statistical context. We were merely solving a fitting problem, and $\hat{\mathbf{b}}$ described the line of best fit. The vector $\hat{\mathbf{b}}$ was a *descriptive* statistic. From the beginnings of mathematical statistics, early in this century, it was understood that observed variations in the values of the statistic $\hat{\mathbf{b}}$, when they were derived from distinct samples dealing with the same phenomenon, indicated the random character of such a statistic. It was realized that these statistics were functions of the first- and second-order sample (product)-moments. If the distribution of these sample moments was known, the distribution of both intercept and slope could be computed. Initially Karl Pearson's approach to this problem, and to similar problems, was to use approximate results for large samples. But from the thirties on, notably under the influence of Fisher's work on exact sampling distributions and on the method of maximum likelihood, the emphasis shifted to assuming very specific stochastic models, and to deriving the distribution of the statistics from the *a priori* assumptions defining the model. This is also the approach used in classical econometrics.

Elementary textbooks of econometrics often start by assuming the so-called linear model[2]

$$Y_t = b_1 z_t + b_0 + \varepsilon_t$$

where the ε_t are independent normal variables with zero expectation and constant variance. The *regressor* z is supposed to be non-random. The model is supposed to be an *exact description* of reality, that is to say in our subsequent analysis we act as if all the underlying assumptions are true. The assumptions imply, for instance, that $\hat{b}_0$ and $\hat{b}_1$ are linear functions of the normal variables ε_t and are consequently normally distributed as well. The impact of this set of assumptions, which we shall call the basic postulates in this paper (see, for instance, Reference 3, pp. 80 and 81), should not be overstated. Many textbooks show that the traditional t and F test statistics can be used without change if z is a value taken by the random variable Z, provided ε_t and Z_t are independently distributed and the sample size is sufficiently large. White (Reference 4, p. 109) has extended the analysis to cover cases where the assumption of independence between a random Z_t and ε_t is not necessary, provided that certain restrictions on the fourth- and second-order moments are fulfilled. We consider that case in section 6.

In the literature a lot of space has been devoted to relaxing the basic postulates by replacing some assumptions by more realistic ones. But always covariance matrices, confidence limits, etc. are assessed on the basis of *some* set of restrictive assumptions.

The alternative approach, where we assume that $\{(Y_t, Z_t)\}_{t=1}^{T}$ is a random sample of independent observations from the same population (i.i.d.), is more natural and more general, at least in the analysis of cross-sections, but the exact distribution of $\hat{\mathbf{b}}$ cannot be derived except if the distribution of (Y_t, Z_t) is known. However, large-sample results may be obtained without such knowledge. In order to assess the large-sample distribution of $\hat{\mathbf{b}}$, we may make use of the central limit theorem, which specifies that sample moments are asymptotically normally distributed under mild conditions. We combine this with the so-called *delta method* (References 5, 6, 7 (section 28.4), 8 (section 6a.2) and 9). According to the delta method, differentiable functions of asymptotically normal variables are also asymptotically normal, with a dispersion that can be computed by using the linear terms of the Taylor series around the population value. We can expand, for instance, $\hat{b}_1$ around its population value b_1. Here $\hat{b}_1 = \hat{\sigma}_{YZ}/\hat{\sigma}_{ZZ}$, the sample covariance divided by the sample variance of the independent variable. It is a function of the sample covariance matrix. We define the population counterpart as $b_1 = \sigma_{YZ}/\sigma_{ZZ}$. The result given by the delta method is that the asymptotic distribution of $T^{1/2}(\hat{b}_1 - b_1)$ is the same as the asymptotic distribution of

$$T^{1/2}\left(\frac{\partial b_1}{\partial \sigma_{YZ}}\right)(\hat{\sigma}_{YZ} - \sigma_{YZ}) + T^{1/2}\left(\frac{\partial b_1}{\partial \sigma_{ZZ}}\right)(\hat{\sigma}_{ZZ} - \sigma_{ZZ}) \tag{2}$$

with the partials evaluated at the 'true' values. But the asymptotic distribution of this linear approximation is normal, by the central limit theorem, with variance

$$T\left\{\left(\frac{\partial b_1}{\partial \sigma_{YZ}}\right)^2 \operatorname{var}(\hat{\sigma}_{YZ}) + 2\left(\frac{\partial b_1}{\partial \sigma_{YZ}}\right)\left(\frac{\partial b_1}{\partial \sigma_{ZZ}}\right)\operatorname{cov}(\hat{\sigma}_{YZ}, \hat{\sigma}_{ZZ}) + \left(\frac{\partial b_1}{\partial \sigma_{ZZ}}\right)^2 \operatorname{var}(\hat{\sigma}_{ZZ})\right\} \tag{3}$$

This is consequently also the variance of the asymptotic normal distribution of $T^{1/2}(\hat{b}_1 - b_1)$. Observe that in this case we do not make strong assumptions about the generating model of our observations. Of course the derivation makes sense only if the central limit theorem applies, and if $\hat{\sigma}_{YZ}$ and $\hat{\sigma}_{ZZ}$ have finite variances. Thus we must assume that fourth-order moments exist. If successive observations are independent and identically distributed, then this is also the *only* assumption we need.

This non-parametric large-sample approach to the simple linear model can be extended to arbitrary covariance functions, such as the ones that occur in factor analysis or canonical correlation analysis. It has two obvious disadvantages. Firstly, it is only valid for large samples; secondly, it involves the calculation and storage of fourth-order (product)-moments. Both features made this approach unpractical until the beginning of the computer era, for there were virtually no samples which were 'large' enough, and computations involving more than a (4×4) matrix were practically impossible. At this point in time there are, however, reasons to consider the non-parametric large-sample approach, outlined above, as a promising alternative to the parametric approach based on specific stochastic models.

Logically, a method based on a smaller set of assumptions seems preferable to one based on a more restrictive set of postulates. This remark must be qualified somewhat, however. In general, weak assumptions mean a large dimensionality of the parameter space and strong assumptions mean a small dimensionality. Large dimensionality means little bias due to misspecification, but possibly a great deal of sampling variance. Small dimensionality means precision in terms of sampling variance, but this precision may be spurious because of biasedness and inconsistency. It seems to us that in many situations in econometrics and other

behavioural sciences there is not enough prior knowledge to justify the introduction of sharp models. This means that models are used mainly for data reduction and exploration, because trying to confirm something you do not really believe in seems a somewhat futile exercise. Thus most precision in the estimates must be taken with a grain of salt, because specific assumptions may be widely off the mark. We must take the 'curse of dimensionality', i.e. an increase in variance, as a consequence of our lack of prior knowledge, and we must make our models relatively weak (and our sample sizes very large).

We have already seen that the large sample approach is now technically feasible, because we often have large samples and powerful computers available. In statistics this is slowly leading to an important revolution in the choice of techniques.[10,11] The new computer-based non-parametric techniques 'replace standard assumptions about data with massive calculations' (Reference 11, p. 96). As we shall see in the following sections, the above approach will lead to a number of unexpected and convenient results which are difficult to find in the framework of established econometric or psychometric methodology. In short, we believe that the approach using the central limit theorem and the delta method is an attractive method for large samples. Finally, and this is most important, the delta method gives true asymptotic dispersions and confidence bounds under minimal assumptions. The 'basic' postulates give correct results if the model that is postulated is strictly true. In cases where these postulates are not satisfied, we do not know what confidence we should place on standard errors, t-values etc.

In this paper we will investigate the large-sample approach in more depth. We continue to use the classical central limit theorem and delta method, but we emphasize that in addition we could use permutational central limit theorems in combination with tools such as the bootstrap and the jackknife.[12,13] These alternative tools are even further removed from classical statistical modelling, because they substitute even more computer power for small-sample assumptions.

Although this introduction focuses on one example, namely the simple linear model and ordinary least squares regression, the same considerations hold for the whole body of classical multivariate analysis. Of course multivariate analysis is used in many other empirical sciences besides econometrics. Therefore our paper does not deal with a problem that is exclusive to econometrics. Our discussion applies equally well to the use of multivariate analysis in sociometrics, psychometrics, biometrics, and so on. Indeed in many of these fields similar reorientations on statistical techniques are going on.

Before we go on to the next section we give some references to related work. The problem of regression estimation under non-standard assumptions was studied in the econometric literature by White.[14,15] He investigated the behaviour of quasi-maximum likelihood estimators if the underlying distribution is misspecified. Van Praag[16,17] applied the approach outlined before. Chamberlain[18] proposed a related approach to simultaneous equations models. In psychometrics similar developments are going on in the analysis of covariance and correlation structures. Browne[19] studied the adaptations that were necessary if the assumption of multivariate normality was relaxed to multivariate ellipticity. Since 1982 the assumption of ellipticity has been relaxed further to merely assuming finite fourth-order product-moments.[20–26]

We try to unify aspects of all this work by using, from section 4 on, a general principle which we call the *population-sample-decomposition principle*. Applications of this general principle can also be found in References 27–29. This principle may be summarized as follows: we are interested in the consistent estimation of a population parameter vector $\mathbf{f}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the population covariance matrix of a random vector $\mathbf{X}$. The calculation of $\mathbf{f}(.)$ is a problem of calculus and optimization techniques. The value of $\mathbf{f}(\boldsymbol{\Sigma})$ depends on $\boldsymbol{\Sigma}$. It is estimated by $\hat{\boldsymbol{\Sigma}}$. If $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator and $\mathbf{f}$ is continuous, $\mathbf{f}(\hat{\boldsymbol{\Sigma}})$ is a consistent estimator of $\mathbf{f}(\boldsymbol{\Sigma})$. The

distribution of $\mathbf{f}(\hat{\Sigma})$ depends on the distribution of $\hat{\Sigma}$. Hence we obtain two aspects of any multivariate problem involving covariance functions:

(a) devising a calculation procedure to calculate $\mathbf{f}(\Sigma)$ for given Σ; this is the *population* aspect

(b) finding a consistent estimator of Σ and assessing its distribution; this is the *sample* aspect.

In this introductory section we have tried to give the reader an intuitive idea of the main idea of this paper. For illustration we have used the classical linear regression model. In section 2 we shall give a formal description of a fairly broad class of multivariate analysis techniques. In section 3 we shall derive the common statistical properties of our estimators, which all have the form of so-called *covariance functions*, under minimal or almost minimal assumptions. In section 4 we consider the modifications that result if it can be assumed that the observations are a random sample from a normal or elliptical population. In section 5 we study the case of controlled experiments and of repeatable samples. In section 6 we consider the classical situation where a linear model is assumed to be true, and where the regressors are non-random. In section 7 a numerical example is considered. Section 8 concludes.

2. THE COMMON GEOMETRIC BACKGROUND OF PROBLEMS OF BEST FIT

In this section we shall concentrate on the so-called *population* aspect. This boils down to bringing to the fore the common geometric background of most covariance functions used in multivariate analysis. Statistical properties will be considered in the following section.

Given a set of T observations of a k-vector $\mathbf{X} \in R^k$, say $\{\mathbf{X}_t\}_{t=1}^T$, natural curiosity dictates that one looks for regularity in the observations. For convenience we assume $\mathbf{E}(\mathbf{x}) = 0$. Such regularity may be described by a linear model $\mathbf{B}^T\mathbf{X} = 0$, where $\mathbf{B}^T$ is a $(p \times k)$-matrix. It describes a $(k - p)$-dimensional subspace $S(\mathbf{B})$ in R^k. Geometrically it implies that we hope that all observations are in a subspace $S(\mathbf{B})$ of R^k. In reality this will not occur except for trivial problems. In that case we may look for a space that does not fit perfectly but best. Let a distance function d on R^k be given, then we may define the point $\hat{\mathbf{x}}_{\mathbf{B}}(\mathbf{X}) \in S(\mathbf{B})$ with minimum distance to $\mathbf{X}$ (see also Reference 30, p. 54). We define the average squared minimum distance of the observations $\{\mathbf{X}_t\}_{t=1}^T$ to the subspace $S(\mathbf{B})$ as

$$\Delta_d(\mathbf{B}) = \frac{1}{T}\sum_{t=1}^{T} d^2(\mathbf{X}_t, \hat{\mathbf{x}}_{\mathbf{B}}(\mathbf{X}_t)) \tag{4}$$

The general problem of looking for a best fitting $S(\mathbf{B})$ may then be succinctly described as $\min_{\mathbf{B}} \Delta_d(\mathbf{B})$.

Let us restrict ourselves now to the traditional case, where the distance function is of Euclidean type, i.e.

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T\mathbf{Q}(\mathbf{x} - \mathbf{y}) \tag{5}$$

where $\mathbf{Q}$ is a positive-definite matrix. In that case it is well known that $\hat{\mathbf{x}}_{\mathbf{B}}(\mathbf{x}) = \mathbf{P}_{\mathbf{B}}(\mathbf{x})$ where $\mathbf{P}_{\mathbf{B}} = \mathbf{I} - \mathbf{Q}^{-1}\mathbf{B}(\mathbf{B}^T\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^T$ is a projection matrix, depending on the distance-defining matrix $\mathbf{Q}$ and the space $\mathbf{B}^T\hat{\mathbf{x}} = 0$ on which $\mathbf{x}$ is projected. The mapping $\hat{\mathbf{x}}_{\mathbf{B}}(\mathbf{x}) = \mathbf{P}_{\mathbf{B}}(\mathbf{x})$ is a linear mapping. This linearity is evidently caused by the special choice of $d^2(.,.)$ in (5).

It follows then that (4) may be written as

$$\Delta_{\mathbf{Q}}(\mathbf{B}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t^T(\mathbf{I} - \mathbf{P}_{\mathbf{B}})^T\mathbf{Q}(\mathbf{I} - \mathbf{P}_{\mathbf{B}})\mathbf{X}_t \tag{6}$$

Expression (6) is a quadratic form in the $\mathbf{x}_t$'s. Let us write $\mathbf{A} = (\mathbf{I} - \mathbf{P_B})^T\mathbf{Q}(\mathbf{I} - \mathbf{P_B})$, then (6) may be written as

$$\Delta_{\mathbf{Q}}(\mathbf{B}; \hat{\Sigma}) = \sum_{i,j}^{k} a_{ij}\hat{\sigma}_{ij} \tag{7}$$

where

$$\hat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} X_{it}X_{jt}$$

Using the explicit expression for $\mathbf{P_B}$, (7) may be written as

$$\Delta_{\mathbf{Q}}(\mathbf{B}; \hat{\Sigma}) = \text{tr } [\mathbf{B}(\mathbf{B}^T\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\hat{\Sigma}] \tag{8}$$

It follows then that minimization of (7) for given $\mathbf{Q}$ with respect to $\mathbf{B}$ yields a solution $\hat{\mathbf{B}} = \mathbf{B}(\hat{\Sigma})$, that is the solution $\hat{\mathbf{B}}$ is a function of the sample-covariance matrix. We call such a function a covariance function for short. The vector $\hat{\mathbf{x}}_{\hat{\mathbf{B}}}(\mathbf{X})$ is called the calculated counterpart of $\mathbf{X}$ and the vector $\mathbf{X} - \hat{\mathbf{x}}_{\hat{\mathbf{B}}}(\mathbf{X})$ is called the *calculated residual.*

The most familiar case is of course that where S is a hyperplane ($p = 1$) and $\mathbf{Q} = \mathbf{I}$. That case is known as orthogonal regression. If S is of dimension $p < k$, and $\mathbf{Q} = \mathbf{I}$ it is also clear that S is spanned by the first p principal components, i.e. the p eigenvectors of $\hat{\Sigma}$ with largest eigenvalues. If S is described by a set of $(k - p)$ equations $\mathbf{b}_1^T\mathbf{x} = 0$, $\mathbf{b}_2^T\mathbf{x} = 0, \ldots, \mathbf{b}_{k-p}^T\mathbf{x} = 0$, the $(k - p)$ eigenvectors of $\hat{\Sigma}$ with smallest eigenvalues may be used for $\mathbf{b}_1^T, \ldots, \mathbf{b}_{k-p}^T$. It follows also from this argument that the matrix $\mathbf{B}$ is frequently not uniquely determined, but that the geometric set it describes, the subspace S, is mainly uniquely determined. We may say then that S is geometrically identified, although not algebraically.

It is also possible to describe a subspace not by a set of equations but in its 'parametric' form by writing each point $\hat{\mathbf{x}} \in S$ as a linear combination $\hat{\mathbf{x}} = \Gamma\mathbf{f}$ of the columns of Γ. If Γ consists of p linearly independent column vectors, those vectors, constituting a basis for S, are called the 'factor loadings' and $\mathbf{f}$ the 'common' factor scores. It is then possible to reformulate the above problem from its equation form into its parametric form. Then it follows that the optimal $\hat{\Gamma}$ is a covariance function as well.

A case that frequently occurs in econometric practice is that some elements of $\mathbf{B}$ are known to be zero or one, or that functional relationships between elements of $\mathbf{B}$ are supposed to exist. In that case (6) has to be minimized with respect to $\mathbf{B}$ under the additional constraints on $\mathbf{B}$. The resulting $\hat{\mathbf{B}}$ is again a covariance function $\mathbf{B}(\hat{\Sigma})$, although not composed of eigenvectors of $\hat{\Sigma}$.[29]

Now we want to make a few additional remarks that will not be pursued explicitly in the following sections. They deal with some complications referring to the distance-defining matrix $\mathbf{Q}$.

1. If the matrix $\mathbf{Q}$ happens to be *singular*, it may be written after diagonalization as

$$\mathbf{Q} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ & & \lambda_r & \\ 0 & & & \ddots 0 \end{bmatrix} \tag{9}$$

The set of points with equal distance ρ to the origin is given by $\mathbf{x}^T\mathbf{Q}\mathbf{x} = \rho^2$. Geometrically such a locus is not an ellipsoid but an elliptical cylinder. Consider then a hyperplane S described by the equation

$$\alpha_1 x_1 + \ldots + \alpha_r x_r + \alpha_{r+1}x_{r+1} + \ldots + \alpha_k x_k = \gamma$$

There is not a unique point in S with shortest distance to the origin. If $\mathbf{z} \in S$ is such a point, $(z + \mathbf{w})$, where $\mathbf{w} = (0, \ldots, 0, w_{r+1}, \ldots, w_k)$ and $(\alpha_{r+1} w_{r+1} + \ldots + \alpha_k w_k) = 0$ is such a point as well. It follows that $\mathbf{P_B}$ is not uniquely defined. Geometrically this is solved by projecting all observations parallel to the cylinder axis to a space of dimension r before doing any fitting. Algebraically this boils down to replacing $(\mathbf{B}^T\mathbf{Q}^{-1}\mathbf{B})^{-1}$ in $\mathbf{P_B}$ by a Moore–Penrose inverse. As before the optimizing $\mathbf{B}$ will be a covariance function $\mathbf{B}(\hat{\Sigma})$.

2. *Exogenous and endogenous variables.* The case which may be viewed in some sense as the opposite extreme of the previous case is implicit in the traditional econometric assumption, that there are *exogenous* and *endogenous* variables. This is clarified by comparing the case of orthogonal regression with OLS regression. In orthogonal regression the minimum-distance projection of $\mathbf{X}$ on the plane $\mathbf{B}^T\mathbf{X} = 0$ is found by perpendicular projection, as dictated by $\mathbf{Q} = \mathbf{I}$. In the OLS case the minimum-distance projection $\mathbf{X}$ on the plane is realized by vertical projection. In the first case all co-ordinates of $\mathbf{X}$ differ from the corresponding ones of $\hat{\mathbf{x}}(\mathbf{X})$. Let us write $\mathbf{X} = (Y, \mathbf{Z})$ where $\mathbf{Z}$ is a $(k-1)$-vector. In the second case only the co-ordinate of the endogenous component Y of $\mathbf{X}$ is changed while the $(k-1)$ exogenous components $\mathbf{Z}$ are not changed in the projection.

If $\hat{\mathbf{x}}(\mathbf{X}) = (\hat{Y}, \hat{\mathbf{Z}})$ it follows that $\hat{\mathbf{Z}} = \mathbf{Z}$. Such a projection may be seen as a limiting case where

$$\mathbf{Q} = \lim_{\lambda \to \infty} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \lambda \mathbf{I}_{k-1} \end{bmatrix}$$

(cf. also Reference 30, p. 96).

In a similar way we may consider a space of dimension $(k-p)$ defined by the system of p linear equations $\mathbf{Y} = \mathbf{B}_1^T\mathbf{Y} + \mathbf{B}_2^T\mathbf{Z}$ or $(\mathbf{I} - \mathbf{B}_1)^T\mathbf{Y} + \mathbf{B}_2^T 2\mathbf{Z} = 0$ where $\mathbf{X}$ is decomposed into a p-vector $\mathbf{Y}$ and a $(k-p)$-vector $\mathbf{Z}$. In the case of a general $\mathbf{Q}$ matrix the $\mathbf{Q}$ projection of $\mathbf{Z}$ on S will be $\hat{\mathbf{x}}(\mathbf{X})$, where all components of $\mathbf{X}$ differ from their corresponding ones in $\hat{\mathbf{x}}(\mathbf{X}) = (\hat{\mathbf{y}}, \hat{\mathbf{z}})$. Consider now the case where

$$\mathbf{Q} = \lim_{\lambda \to \infty} \begin{bmatrix} \mathbf{Q}_p & \mathbf{0}^T \\ \mathbf{0} & \lambda \mathbf{I}_{k-p} \end{bmatrix}$$

In that case there holds $\hat{\mathbf{Y}} \neq \mathbf{Y}$ and $\hat{\mathbf{Z}} = \mathbf{Z}$. This is the case of simultaneous equations, where it is assumed that $\mathbf{z}$ is exogenous and where $\mathbf{Q}_p^{-1}$ is the covariance matrix of the disturbance terms.

Frequently some additional constraints on $\mathbf{B}^T$ are added to ensure algebraic identifiability of $\mathbf{B}^T$. If $\mathbf{B}_1{}^T = 0$ we have already the reduced form or we are faced with a case of 'seemingly unrelated regressions'.[31] In all cases, however, it is obvious that a minimizing solution is a covariance function $\hat{\mathbf{B}} = \mathbf{B}(\hat{\Sigma})$.

3. *Iterative improvements.* The choice of the fitting criterion, specified by the symmetric weighting matrix $\mathbf{Q}$, implies a value judgement on the size of the residuals we like to tolerate. The result of applying a specific $\mathbf{Q}$ may be rather large residuals on one dimension and rather small on others. A special example is the case, described above, where $\mathbf{Q}$ is specified in such a way that on $(k-p)$ dimensions, corresponding to the exogenous variables, only a zero residual is tolerated.

Sometimes there may be a firm *a priori* basis in choosing a specific $\mathbf{Q}$ as fitting criterion. In most applications, however, there is no firm basis to choose a specific $\mathbf{Q}$, and one uses the identity matrix $\mathbf{I}$ to begin with. Then the result may be rather unfortunate. Let $\mathbf{X}$ be a 2-vector, and let $\text{var}(X_1 - \hat{x}_1(\mathbf{X})) = 100$ and $\text{var}(X_2 - \hat{x}_2(\mathbf{X})) = 1$. If X_1 and X_2 are roughly comparable in variation and are measured on comparable scales to begin with, a researcher may be rather unhappy with this fitting result. The natural inclination may

be to repeat the fitting after a new scaling of the variables, where the natural choice is to divide the X_1 component by 10. More generally let the covariance matrix of the residuals in the first stage be

$$\mathbf{R}^{(1)} = \frac{1}{T}\sum_t [\mathbf{X}_t - \hat{\mathbf{x}}(\mathbf{X}_t)]\,[(\mathbf{X}_t - \hat{\mathbf{x}}(\mathbf{X}_t)]^{\mathrm{T}} \tag{10}$$

then we rescale the observations as $\mathbf{X}^{(2)} = (\mathbf{R}^{(1)})^{-\frac{1}{2}}\mathbf{X}$ and apply the same fitting procedure with $\mathbf{Q} = \mathbf{I}$ as before on the rescaled observations. An alternative interpretation is that the rescaling of the observations is equivalent to a change of the weighting matrix from $\mathbf{I}$ into $\mathbf{Q}^{(2)} = (\mathbf{R}^{(1)})^{-1}$ and applying a fitting with $\mathbf{Q}^{(2)}$ on the observations measured on their *original* scale.

This fitting procedure may be repeated until $\mathbf{R}^{(n)}$, the covariance matrix of residuals resulting from the nth trial, has approached the identity matrix $\mathbf{I}$ sufficiently close. This rescaling or reweighting procedure is essentially the gist of the 'third stage' in three-stage-least-squares procedures. Except for an *a priori* fixed $\mathbf{Q}$ and a variable $\mathbf{Q}$ there are also intermediate cases where $\mathbf{Q}$ is neither completely predetermined nor completely free. In such cases $\mathbf{Q}$ may be postulated to be diagonal, or tridiagonal or block-diagonal or otherwise patterned. These special cases are treated in more detail in Reference 32. In the case of iterative improvement it is obvious that $\mathbf{R}^{(1)}$ is a function of $\hat{\boldsymbol{\Sigma}}$, as the first-stage residuals are

$$\mathbf{X} - \hat{\mathbf{x}}_1(\mathbf{X}) = \mathbf{Q}^{-1}\hat{\mathbf{B}}(\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{Q}^{-1}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}$$

with $\hat{\mathbf{B}} = \mathbf{B}^{(1)}(\hat{\boldsymbol{\Sigma}})$ the estimation result of the first fitting stage. Hence it follows that minimization of (6) with respect to $\mathbf{B}$, where $\mathbf{Q}^{(2)} = (\mathbf{R}^{(1)})^{-1}$, yields an improved estimate $\hat{\mathbf{B}}^{(2)} = \mathbf{B}^{(2)}(\hat{\boldsymbol{\Sigma}})$.

Summarizing the results of this section, we have shown that a wide variety of so-called multivariate problems, the main tools of analysis in econometrics, psychometrics and sociometrics, may be reinterpreted as fitting a linear subspace to a set of observations according to a specific criterion. The second and most important result is then that both the best-fitting $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ and the corresponding sum of squared residuals $\Delta(\mathbf{B}; \hat{\boldsymbol{\Sigma}})$ are functions of the observations' covariance matrix $\hat{\boldsymbol{\Sigma}}$, where the functional specifications of $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ and $\Delta(\mathbf{B}; \hat{\boldsymbol{\Sigma}})$ vary with the specific fitting problem.

3. STATISTICAL PROPERTIES OF MULTIVARIATE ESTIMATORS UNDER MINIMAL ASSUMPTIONS

In the previous sections we considered the regression problem and later the general multivariate best-fit problem to introduce the concept of a covariance function.

In this section we consider the situation where we have different sets of observations and accordingly different estimates $\hat{\mathbf{B}}^{(1)}$ and $\hat{\mathbf{B}}^{(2)}$. Then immediately the problem arises how to reconcile these two different findings if we feel that the two phenomena studied are not different. The answer is to assume a statistical context. The two sets are two different samples of the same population and $\hat{\mathbf{B}}^{(1)}$ and $\hat{\mathbf{B}}^{(2)}$ are two different estimates of the same population value $\mathbf{B}$. This calls for the definition of a best-fit problem in population space.[7] Consider (6). If the $\mathbf{x}_t$'s are assumed to be i.i.d. and $E(\mathbf{X}) = 0$, it follows that the population counterpart of (6) is

$$\begin{aligned}\Delta_{\mathbf{Q}}(\mathbf{B}) &= E(\mathbf{X}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{B}})^{\mathrm{T}}\mathbf{Q}(\mathbf{I} - \mathbf{P}_{\mathbf{B}})\mathbf{X}) \\ &= \mathrm{tr}[\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}]\end{aligned} \tag{11}$$

where $\boldsymbol{\Sigma} = E(\mathbf{X}\mathbf{X}^{\mathrm{T}})$.

The best-fitting linear subspace $\mathbf{B}^T\mathbf{X} = 0$ is then found by minimization of (11) with respect to $\mathbf{B}$, yielding $\mathbf{B} = \mathbf{B}(\boldsymbol{\Sigma})$. We notice that the function $\mathbf{B}(.)$ solving the *population* problem is *analytically identical* to the function $\mathbf{B}(.)$ solving the *sample* problem (8). The difference between $\hat{\mathbf{B}}^{(1)} = \mathbf{B}(\hat{\boldsymbol{\Sigma}}_1)$, $\hat{\mathbf{B}}^{(2)} = \mathbf{B}(\hat{\boldsymbol{\Sigma}}_2)$ and $\mathbf{B}(\boldsymbol{\Sigma})$ is only caused because $\hat{\boldsymbol{\Sigma}}_1 \neq \hat{\boldsymbol{\Sigma}}_2 \neq \boldsymbol{\Sigma}$. Hence, it follows that any descriptive statistic, such as the ones considered in the previous section, can be reinterpreted as an estimate of a population parameter. By definition $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ is an estimator of the corresponding population parameter $\mathbf{B}(\boldsymbol{\Sigma})$.

We notice at this point two things. In the first place, although the estimators we are studying are frequently identical to those appearing in traditional multivariate theory, the way in which we derive them is different from the usual procedure. We started with a criterion function based on a geometric approximation problem. This gave us descriptive statistics of the form $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$, which can be subsequently interpreted as estimates of the corresponding parameter in the population best-fit problem. In classical theory one starts usually with a parametric model, say with parameter $\mathbf{B}$. Then we find a function $\mathbf{B}(.)$, such that $\mathbf{B}(\boldsymbol{\Sigma}) = \mathbf{B}$, and it is estimated by $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$. Although the outcomes of the two approaches may be the same there is a considerable shift in emphasis. We start from the geometry and make minimal assumptions, namely that $\boldsymbol{\Sigma}$ exists. Classical theory starts from a postulated model and makes much stronger assumptions.

In the second place we note that our statistics are formally identical to ML estimators under the assumption that the $\mathbf{X}_t$s are sampled from a multivariate normal distribution. However, they continue to make sense in our frame of reference if there is no multivariate normal parent distribution. As already pointed out the statistical properties of $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ depend on those of $\hat{\boldsymbol{\Sigma}}$. Now we shall derive the asymptotical properties for $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$. Therefore we have to assume that not only second-order but also fourth-order moments exist. This will be called from now on the 'minimal' assumption. The theory presented here is thus a large-sample theory.

Let us assume in this section that $\{\mathbf{X}_t\}_{t=1}^{T}$ is a random sample, i.e. the $\mathbf{X}_t$s are drawn independently from an identical distribution (i.i.d.). Let us also assume that $E(\mathbf{X}) = 0$. Then

1. $\text{vec}(\hat{\boldsymbol{\Sigma}}) \xrightarrow{\text{a.s.}} \text{vec}(\boldsymbol{\Sigma})$, iff $\boldsymbol{\Sigma} < \infty$

2. $T^{1/2}\,\text{vec}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \xrightarrow{\text{d.}} N(0, \boldsymbol{\Pi})$, iff $\boldsymbol{\Pi} < \infty$

where

$$\boldsymbol{\Pi} = TE[\text{vec}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})]\,[\text{vec}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})]^T \tag{12}$$

The matrix $(1/T)\,\boldsymbol{\Pi}$ is the approximate covariance matrix of second-order sample moments for a sample of size T. It is also called the *kurtosis* matrix.[19,21-23,33]

The first statement is an application of the law of large numbers on second-order moments. The second statement is an application of the Lindeberg–Levy version of the central limit theorem, which states that $T^{1/2}$ times a sum of i.i.d. variables tends asymptotically to be normal if that vector has a finite covariance matrix.

The elements of $\boldsymbol{\Pi}$ are $\pi_{ij,kl} = T\,\text{cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{kl})$. There holds,

$\text{cov}(\hat{\sigma}_{ij}, \hat{\sigma}_{kl}) = E(\hat{\sigma}_{ij}, \hat{\sigma}_{kl}) - E(\hat{\sigma}_{ij})E(\hat{\sigma}_{kl})$ or, ignoring smaller-order terms in this large-sample context

$$\pi_{ij,kl} = \mu_{ijkl} - \sigma_{ij}\sigma_{kl} \tag{13}$$

with μ_{ijkl} the fourth-order central product moment of the distribution of $\mathbf{X}$. Hence, as the existence of fourth-order moments implies that second-order moments exist, we find that $T^{1/2}\,\text{vec}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ is asymptotically normal if the fourth-order central moments of $\mathbf{X}$ exist. The

matrix $\mathbf{\Pi}$ can be consistently estimated by replacing in (13) μ_{ijkl} and σ_{ij} by their sample analogues.

As $\mathbf{\Sigma}$ is symmetric, it follows that there are many identical elements in $\text{vec}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})$. To be precise there are $\frac{1}{2}(k-1)k$ equality constraints of the type $\sigma_{ij} = \sigma_{ji}$. It follows that the covariance matrix $\mathbf{\Pi}$ is singular with a defect $\frac{1}{2}(k-1)k$. However, as we shall not need to invert $\mathbf{\Pi}$ (except in section 5), this will be no problem. It is possible to replace $\text{vec}(\hat{\mathbf{\Sigma}})$ of dimension k^2 by a vector of dimension $k^2 - \frac{1}{2}(k-1)k$, being the vectorization of the lower half of the symmetric matrix $\mathbf{\Sigma}$, but the notational complication does not seem worth while.[34,35]

Consider now $\mathbf{B}(\hat{\mathbf{\Sigma}})$. If $\mathbf{B}$ is a continuous function

$$\mathbf{B}(\hat{\mathbf{\Sigma}}) \xrightarrow{\text{a.s.}} \mathbf{B}(\mathbf{\Sigma}), \text{ iff } \mathbf{\Sigma} < \infty.$$

Let us now consider the distribution of $T^{1/2}(\mathbf{B}(\hat{\mathbf{\Sigma}}) - \mathbf{B}(\mathbf{\Sigma}))$, and let us assume that $\mathbf{B}(.)$ is differentiable in a neighbourhood of $\mathbf{\Sigma}$. It is asymptotically equivalent to its Taylor expansion, i.e.

$$T^{1/2} \text{vec}(\mathbf{B}(\hat{\mathbf{\Sigma}}) - \mathbf{B}(\mathbf{\Sigma})) \approx T^{1/2}[\dot{\mathbf{B}}(\mathbf{\Sigma})]^{\mathrm{T}} \text{vec}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) \tag{14}$$

where $[\dot{\mathbf{B}}(\mathbf{\Sigma})]^{\mathrm{T}}$ stands for the matrix of first-order derivatives of $\mathbf{B}$ with respect to the elements of $\mathbf{\Sigma}$. Its dimension is $(pk \times k^2)$. From (14) it is seen that the left member equals asymptotically a linear transform of an asymptotically normal vector. Hence, it follows that

$$T^{1/2}[\mathbf{B}(\hat{\mathbf{\Sigma}}) - \mathbf{B}(\mathbf{\Sigma})] \xrightarrow{\text{d.}} \text{N}(0, \dot{\mathbf{B}}^{\mathrm{T}} \mathbf{\Pi} \dot{\mathbf{B}})$$

The method employed is called nowadays the delta method (see section 1 for some references). Thus we have derived the large-sample distribution of our best-fit statistics, but actually of course of any covariance function. Nothing else is assumed than independent and identically distributed observations with finite fourth-order moments on one side, and continuously differentiable covariance functions on the other side. These assumptions are far more general than the assumptions classically used in multivariate statistical analysis.

There are only two ingredients. In the first place we need to specify the function $\mathbf{B}(\mathbf{\Sigma})$, for instance as the function which provides the solution to a minimization problem. We also must prove that $\mathbf{B}(\mathbf{\Sigma})$ is continuously differentiable, and we have to compute its derivative. This problem can be solved in population space; it does not involve the sample. It is, in general, a problem in analysis or mathematical programming to define $\mathbf{B}(.)$ and study its properties. It is even not necessary to find an explicit expression of $\mathbf{B}$, when its derivatives with respect to $\mathbf{\Sigma}$ can be derived from a system of implicit differentiable equations.

Our second ingredient is an estimate of $\mathbf{\Sigma}$ in terms of the sample. We also need a form of the central limit theorem that applies to this estimate. But this second ingredient does not depend in any way on the particular function we are minimizing or its resulting solution $\mathbf{B}(.)$; it only depends on the statistical properties of the sample covariance matrix $\hat{\mathbf{\Sigma}}$. By carefully distinguishing between these two components, we have split up the computation of the asymptotic distribution into two subproblems. The first one, which has to do with calculating $\mathbf{B}(.)$ and its derivative, we call the *population problem*. The second one, which is the calculation of $\hat{\mathbf{\Sigma}}$ and its asymptotic distribution, is called the *sample problem*. Studying these two components separately for each problem, and combining them afterwards to a single result is called the *population-sample decomposition approach*. We shall exploit this idea in the sections to follow.

4. STATISTICAL PROPERTIES OF $\mathbf{B}(\hat{\mathbf{\Sigma}})$ IF THE DISTRIBUTION OF $\mathbf{X}$ IS KNOWN TO BE ELLIPTICAL

In the previous section we derived a general large-sample result on the distribution of $\mathbf{B}(\hat{\mathbf{\Sigma}})$, where our only assumption was that $\mathbf{X}_1, \ldots, \mathbf{X}_T$ were i.i.d. with finite fourth-order moments. The matrix $\mathbf{\Pi}$ was estimated by equation (13). In the coming sections we shall assume that we have some additional knowledge on the data-generating process.

Let us assume that we know the common distribution of $\mathbf{X}$, then it is evident that we can use that additional knowledge. The general theorem remains true (of course), but we may find an expression for $\mathbf{\Pi}$ by making use of our knowledge of the data-generating process.

Let us assume that the parent-distribution of $\mathbf{X}$ is *elliptical*, where an elliptical distribution is described by a density function $Cf(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})$ and C a normalizing constant. The normal distribution is a special case with $f(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}) = C\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})$. This distribution family, extensively studied by Muirhead,[33] among others, is very rich and nevertheless only slightly less tractable than the multivariate normal subfamily. In this special case we have

$$\mathbf{\Pi} = [\pi_{hi,jl}]$$

with

$$\pi_{hi,jl} = \varkappa(\sigma_{hi}\sigma_{jl} + \sigma_{hj}\sigma_{il} + \sigma_{hl}\sigma_{ij}) + \sigma_{hj}\sigma_{il} + \sigma_{hl}\sigma_{ij} \tag{15}$$

where $\varkappa$ is the *common kurtosis parameter*. In the case of normality $\varkappa = 0$. It can be shown that the marginal distributions of $X_1, \ldots, X_k$ have the same kurtosis defined by $3\varkappa = \varkappa_4^j/(\varkappa_2^j)^2$ $(j = 1, \ldots, k)$ where $\varkappa_2^j$ and $\varkappa_4^j$ are the second- and fourth-order cumulants of X_j.

The estimation of $\varkappa$ and the test whether $\mathbf{X}$ is elliptically distributed can be based on the identities[7]

$$\begin{aligned}\varkappa_2^j &= \mu_2^j - (\mu_1^j)^2 \,(= \sigma_{jj})\\ \varkappa_4^j &= \mu_4^j - 4\mu_1^j\mu_3^j - 3(\mu_2^j)^2 + 12\mu_2^j(\mu_1^j)^2 - 6(\mu_1^j)^4\end{aligned} \tag{16}$$

where $\mu_i^j = E(X_j^i)$.

As in our example $\mu_1 = 0$ we find as sample estimate of $3\varkappa$

$$\hat{\mu}_4^j/(\hat{\mu}_2^j)^2 - 3 = 3\hat{\varkappa}, \quad j = 1, \ldots, k \tag{17}$$

We see that, if $\mathbf{X}$ is elliptically distributed, then

$$T^{1/2}\mathrm{vec}(\hat{\mathbf{B}} - \mathbf{B}) \to \mathrm{N}(0, \mathbf{V}), \quad \text{with } \mathbf{V} = \dot{\mathbf{B}}^T\mathbf{\Pi}\dot{\mathbf{B}}$$

and $\mathbf{\Pi}$ defined by (15). It can be shown that $\varkappa > -2/3$ (Reference, p.143), in order that the integral of the density converges. We ignore the case $\varkappa = -2/3$ which yields a two-point distribution (see Reference 36, p.88).

The advantage of the elliptical assumption is clear. If $\varkappa$ is known, we do not need to estimate any fourth-order moments of $\mathbf{X}$, as (15) may be evaluated by second-order moments only. However, for a sample not positively known to be from an elliptical population, it will not give the correct estimates of the standard deviations of $\hat{\mathbf{B}}$, for (15) will not equal (13) (see section 7 for empirical comparisons). Let us denote (15) as

$$\pi_{hi,jl} = \varkappa\pi_{hi,jl}^{(1)} + \pi_{hi,jl}^{(2)}$$

where $\mathbf{\Pi}^{(1)}$ and $\mathbf{\Pi}^{(2)}$ are defined by the first and latter two terms of (15), respectively.

Accordingly we may write

$$\Pi = \varkappa\Pi^{(1)} + \Pi^{(2)}. \tag{18}$$

The matrix Π is a positive semi-definite (p.s.d.) matrix for all $\varkappa > -2/3$. Let us define for two p.s.d. matrices $\mathbf{A}_1$, $\mathbf{A}_2$ the matrix inequality as $\mathbf{A}_1 \geq \mathbf{A}_2$ if and only if $\mathbf{x}^T\mathbf{A}_1\mathbf{x} \geq \mathbf{x}^T\mathbf{A}_2\mathbf{x}$ for all $\mathbf{x} \in R^k$. Then it is easy to see that Π is a p.s.d. matrix, increasing in $\varkappa$. This leads to an important corollary, namely that the covariance matrix of any covariance function $\mathbf{B}(\hat{\Sigma})$ is also monotonically increasing in $\varkappa$. This may be seen as follows.

Consider $\mathbf{C}^T\mathbf{A}\mathbf{C}$ with $\mathbf{A}$ positive semi-definite and $\mathbf{C}$ an arbitrary matrix, such that the vertical dimension of $\mathbf{C}$ is equal to that of $\mathbf{A}$. Then $\mathbf{C}^T\mathbf{A}\mathbf{C}$ is symmetric and it is also positive semi-definite. It follows that if $\mathbf{A}_1 \geq \mathbf{A}_2$ then $\mathbf{C}^T\mathbf{A}_1\mathbf{C} \geq \mathbf{C}^T\mathbf{A}_2\mathbf{C}$. Applying this on Π and $\dot{\mathbf{B}}^T\Pi\dot{\mathbf{B}}$ it follows that the covariance matrix

$$\mathbf{V} = \varkappa\dot{\mathbf{B}}^T\Pi^{(1)}\dot{\mathbf{B}} + \dot{\mathbf{B}}^T\Pi^{(2)}\dot{\mathbf{B}} \tag{19}$$

is increasing in $\varkappa$ as well. Notice that the second term on the right hand side of (19) is the covariance matrix when $\varkappa = 0$, that is under the assumption that $\mathbf{X}$ is drawn from a normal population.

5. CONTROLLED EXPERIMENTS AND REPEATED SAMPLES

Another bit of prior knowledge may be that some dimensions of the observations may be controlled, e.g. by a quota design, or that the same set of objects may be observed repeatedly. Up to this point we have assumed that the data-generating process could not be influenced by the researcher. In this section we shall study two situations where the researcher may influence the data-generating process.

The first situation is that of the laboratory experiment. Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ where $\mathbf{Z}$ is a vector of $(k - p)$ explanatory input variables to be determined by the researcher for each observation, and the p-vector $\mathbf{Y}$ is the resulting output to be measured and to be explained. In such a case Nature is partly replaced by the researcher; and as a consequence the marginal distribution of $\mathbf{Z}$ in the sample is determined by the researcher as well. The distribution of $\mathbf{Y}$ in the sample is conditioned by the marginal distribution of $\mathbf{Z}$. This situation is not only prevailing in laboratory or agricultural experiments but also in samples that are drawn according to a quota design; for instance, a household sample may be *designed* to include 60 per cent one-breadwinner and 40 per cent two-breadwinner families or to have a specific income distribution.

The second situation of the *repeated* sample is related to the previous one but different. Again $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ and the marginal distribution of $\mathbf{Z}$ with density $f_{\mathbf{Z}}(.)$ is assumed to be fixed over repeated samples. However, in contrast with the controlled experiment, in this case $f_{\mathbf{Z}}(.)$ is not determined by the researcher but by Nature. This situation is found if we start with a random sample $\{\mathbf{X}_t^{(1)}\}_{t=1}^T = \{(\mathbf{Y}_t^{(1)}, \mathbf{Z}_t^{(1)}\}_{t=1}^T$ and then re-observe the objects where the $\mathbf{Z}$ is kept unchanged but $\mathbf{Y}$ is changed subject to random influences. Repeated observations then yields a sample $\{\mathbf{X}_t^{(2)}\} = \{\mathbf{Y}_t^{(2)}, \mathbf{Z}_t^{(1)}\}$. This situation is frequently assumed in econometric theory, when we speak of non-random explanatory or exogenous variables $\mathbf{Z}$ or of 'fixed in repeated samples'.

Both cases have something in common. If $\{\mathbf{Z}_t\}$ is fixed, it follows that $\hat{\Sigma}_{\mathbf{ZZ}}$ is fixed, i.e. non-random. The elements $\hat{\Sigma}_{\mathbf{YY}}$ and $\hat{\Sigma}_{\mathbf{YZ}}$ of the covariance matrix $\hat{\Sigma}$ are random. This has implications for Π, the covariance matrix of the asymptotic distribution of

$$T^{\frac{1}{2}}\,\mathrm{vec}(\Delta\hat{\Sigma}) = T^{\frac{1}{2}}[\mathrm{vec}(\hat{\Sigma}) - \mathrm{vec}(\Sigma)] \tag{20}$$

For convenience we strip $\hat{\Sigma}$ of its non-distinct elements and take vec($\hat{\Sigma}$) to be a vector of length $\frac{1}{2}k(k+1)$ and Π a square matrix of the same order. Further, we partition Π as follows:

$$\Pi = \begin{bmatrix} \Pi_{YY,YY} & \Pi_{YY,YZ} & \Pi_{YY,ZZ} \\ \Pi_{YZ,YY} & \Pi_{YZ,YZ} & \Pi_{YZ,ZZ} \\ \Pi_{ZZ,YY} & \Pi_{ZZ,YZ} & \Pi_{ZZ,ZZ} \end{bmatrix}$$

where the submatrix $\Pi_{YY,ZZ}$ corresponds to the asymptotic covariance matrix of $T^{1/2}$ vec($\Delta\hat{\Sigma}_{YY}$) and $T^{1/2}$ vec($\Delta\hat{\Sigma}_{YZ}$) and the other submatrices are similarly defined. Owing to the non-randomness of $\hat{\Sigma}_{ZZ}$ it follows that $\Pi_{ZZ,ZZ} = 0$ and actually the whole third row and column of (20) vanish. It is this property that is rather attractive if we look, for instance, for the distribution of the linear regression coefficient $\hat{\mathbf{b}} = \hat{\Sigma}_{ZZ}^{-1}\hat{\Sigma}_{YZ}$ in the simple case $p = 1$, $k = 2$. If $\hat{\Sigma}_{ZZ}^{-1}$ is non-random, $\hat{\mathbf{b}}$ is just a linear function of $\hat{\Sigma}_{YZ}$.

Let us partition $\dot{\mathbf{B}}_{\Sigma}^{T} = (\dot{\mathbf{B}}_{YY}^{T}, \dot{\mathbf{B}}_{YZ}^{T}, \dot{\mathbf{B}}_{ZZ}^{T})$, then

$$\mathbf{V} = [\dot{\mathbf{B}}_{YY}^{T}, \dot{\mathbf{B}}_{YZ}^{T}] \begin{bmatrix} \Pi_{YY,YY} & \Pi_{YY,YZ} \\ \Pi_{YZ,YY} & \Pi_{YZ,YZ} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{B}}_{YY} \\ \dot{\mathbf{B}}_{YZ} \end{bmatrix} \tag{22}$$

At this point we have to deal with the two situations separately. We start with the *controlled experiment*. The question is whether $T^{1/2}\,\Delta\hat{\Sigma}_{YY}$ and $T^{1/2}\,\Delta\hat{\Sigma}_{YZ}$ tend asymptotically to normality. Consider

$$\hat{\sigma}_{YZ} = \frac{1}{T}\sum_{t=1}^{T} Y_t Z_t$$

where Y_t, Z_t are scalars and Z_t is determined by the research design. It is explicitly observed that in this context Z_t is non-random. The factors $Y_t Z_t$ in this sum may be mutually independent and drawn from *T different* distributions. In the simple and popular case that $Y_t \sim N(\beta Z_t, \sigma^2)$, it can be seen that

$$E(\hat{\sigma}_{YZ}) = \frac{1}{T}\sum_{t=1}^{T} (\beta Z_t) Z_t$$

$$\text{var}(T^{1/2}\hat{\sigma}_{YZ}) = \sigma^2 \frac{1}{T}\sum_{t=1}^{T} Z_t^2$$

However, even in that case, it is not ensured that $T^{1/2}\,\Delta\hat{\Sigma}_{YZ}$ and $T^{1/2}\,\Delta\hat{\Sigma}_{ZZ}$ tend to normality. Cases may be constructed where the Lindeberg–Feller condition does not hold, for instance let $Z_t = t (t = 1, \ldots, T)$ and let $\text{var}(Y_t) = \sigma^2$. Then $\text{var}(Z_t Y_t) = t^2\sigma^2$ and the Lindeberg–Feller condition does not hold. It follows that $T^{1/2}\,\Delta\hat{\Sigma}_{ZY}$ does not tend to normality and the δ method cannot be applied either.

Let us suppose, however, that asymptotic normality of $T^{1/2}\,\Delta\hat{\Sigma}_{YY}$ and $T^{1/2}\,\Delta\Sigma_{YZ}$ may be assumed. Then the elements of the middle factor in (22) are assessed by (13). It is obvious that the choice of our sample design $\{\mathbf{Z}_t\}_{t=1}^{T}$ is basic. It determines Σ_{ZZ} and the distribution of $\hat{\Sigma}_{YZ}$ and $\hat{\Sigma}_{YY}$ will vary with the choice of the sample design (except if we hypothesize the existence of a linear model, describing the relationship between $\mathbf{Y}$ and $\mathbf{Z}$; this case will be considered in Section 6).

Let us now consider the case of the repeated sample. To obtain more insight into this case we define a *primary* sample $\{\mathbf{X}_t^{(1)}\}_{t=1}^{T} = \{\mathbf{Y}_t^{(1)}, \mathbf{Z}_t^{(1)}\}_{t=1}^{T}$ and a repetition of it in some sense, namely $\{\mathbf{X}_t^{(2)}\}_{t=1}^{T} = \{\mathbf{Y}_t^{(2)}, \mathbf{Z}_t^{(1)}\}_{t=1}^{T}$. The $\mathbf{Z}_t$s are kept equal in the repeated sample to their counterparts in the primary sample. If $\mathbf{X}^{(1)}$ is a sample of i.i.d. drawings, the corresponding

sample covariance matrix $\hat{\Sigma}^{(1)}$ is a consistent estimator. In a repeated sample framework the question is now what will be the distribution of $\mathbf{B}(\hat{\Sigma})$ over repeated samples, that is samples where $\mathbf{Z}^{(1)} = \mathbf{Z}^{(2)}$ and consequently $\hat{\Sigma}_{ZZ}^{(1)} = \hat{\Sigma}_{ZZ}^{(2)}$. The answer is found by considering the distribution of $\text{vec}(\hat{\Sigma})$ given that $\hat{\Sigma}_{ZZ}$ is constant and equal to $\hat{\Sigma}_{ZZ}^{(1)}$. First we notice that if $\text{vec}(\hat{\Sigma})$ tends to be normal if T is large, then, under some continuity conditions, the conditional distribution of $\text{vec}(\hat{\Sigma})$ given that $\hat{\Sigma}_{ZZ} = \hat{\Sigma}_{ZZ}^{(1)}$ will be asymptotically normal as well.[37] Let us now partition the covariance matrix (20) of the asymptotic distribution of $T^{\frac{1}{2}}\,\text{vec}(\Delta\Sigma)$ as

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}$$

where Π_{22} corresponds to $\hat{\Sigma}_{ZZ}$. Then the conditional distribution of $T^{\frac{1}{2}}\,\text{vec}(\Delta\Sigma)$ is asymptotically normal, and if the sample size is sufficiently large the conditional expectation of $\text{vec}(\hat{\Sigma})$ is approximately given by

$$E\left[\begin{matrix}\text{vec}(\hat{\Sigma}_{YY}) \\ \text{vec}(\hat{\Sigma}_{YZ})\end{matrix} \,\middle|\, \hat{\Sigma}_{ZZ} = \hat{\Sigma}_{ZZ}^{(1)}\right] = \begin{bmatrix}\text{vec}(\Sigma_{YY}) \\ \text{vec}(\Sigma_{YZ})\end{bmatrix} + \Pi_{12}\Pi_{22}^{-1}(\text{vec}(\hat{\Sigma}_{ZZ} - \Sigma_{ZZ})) \tag{24}$$

However, asymptotically $\hat{\Sigma}_{ZZ}$ is unbiased, so the second term on the RHS vanishes. It follows that Σ_{YY} and $\hat{\Sigma}_{YZ}$ are asymptotically unbiased in repeated samples.

In a similar way we find for the conditional covariance matrix that it equals

$$\Pi_{11}^{(2)} = \Pi_{11} - \Pi_{12}\Pi_{22}^{-1}\Pi_{21} \tag{25}$$

where we notice that Π_{22} is non-singular, as we use in this section $\text{vec}(\hat{\Sigma})$ in its reduced version, stripped of its non-distinct elements. Consider now the distribution of $\mathbf{B}(\hat{\Sigma})$ under the condition that Σ_{ZZ} is fixed, that is $\mathbf{B}(\hat{\Sigma}^{(2)})$. It is approximately normal in large samples with expectation $\mathbf{B}(\Sigma)$ and covariance matrix $(1/T)$ times as defined in (22) where the middle matrix is $\Pi_{11}^{(2)}$ as given in (25). Partitioning $\dot{\mathbf{B}}_\Sigma^{\mathsf{T}} = (\dot{\mathbf{B}}_1^{\mathsf{T}}, \dot{\mathbf{B}}_2^{\mathsf{T}})$ like we partitioned Π in (23) we find

$$\mathbf{V}^{(2)} = \dot{\mathbf{B}}_1^{\mathsf{T}}\Pi_{11}^{(2)}\dot{\mathbf{B}}_1 \tag{26}$$

for the conditional analogue of $\mathbf{V}$, i.e. if $\Pi_{11}^{(2)}$ is used rather than Π_{11}.

Now we may make a comparison between the conditional covariance matrix of the approximate large-sample distribution of $\hat{\mathbf{B}}$ and the unconditional one by rewriting (26) as

$$\mathbf{V}^{(2)} = \begin{bmatrix}\dot{\mathbf{B}}_1^{\mathsf{T}}, \dot{\mathbf{B}}_2^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\Pi_{11}^{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\dot{\mathbf{B}}_1 \\ \dot{\mathbf{B}}_2\end{bmatrix} = \dot{\mathbf{B}}^{\mathsf{T}}\Pi^{(2)}\dot{\mathbf{B}} \tag{27}$$

where $\Pi^{(2)}$ stands for the middle factor in (27). The unconditional covariance matrix is evaluated as before by

$$\mathbf{V} = \begin{bmatrix}\dot{\mathbf{B}}_1^{\mathsf{T}}, \dot{\mathbf{B}}_2^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22}\end{bmatrix}\begin{bmatrix}\dot{\mathbf{B}}_1 \\ \dot{\mathbf{B}}_2\end{bmatrix} = \dot{\mathbf{B}}^{\mathsf{T}}\Pi\dot{\mathbf{B}} \tag{28}$$

It can be shown that $\mathbf{V}^{(2)} \leq \mathbf{V}$, which implies for the diagonal elements, that all standard deviations of $\hat{\mathbf{B}}^{(2)}$ are estimated smaller than their unconditional counterparts. That $\mathbf{V}^{(2)} \leq \mathbf{V}$ follows by comparison of $\Pi^{(2)}$ with Π. Consider the difference $(\Pi - \Pi^{(2)})$; it can be shown to be positive semi-definite. We have

$$\Pi - \Pi^{(2)} = \begin{bmatrix}\Pi_{12}\Pi_{22}^{-1}\Pi_{21} & \Pi_{12} \\ \Pi_{21} & \Pi_{22}\end{bmatrix} = \begin{bmatrix}\Pi_{12} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\Pi_{22}^{-1} & \mathbf{I} \\ \mathbf{I} & \Pi_{22}\end{bmatrix}\begin{bmatrix}\Pi_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix} \tag{29}$$

It follows that $(\Pi - \Pi^{(2)}) \geq 0$, if the middle factor in the last matrix product is positive semi-definite. Π_{22} is positive definite as it is a covariance matrix of $\hat{\Sigma}_{ZZ}$. Let $m = \frac{1}{2}(k-p)(k-p+1)$

so that the matrix $\mathbf{\Pi}_{22}$ is of order $m \times m$. Then

$$\begin{bmatrix} \mathbf{\Pi}_{22}^{-1} & \mathbf{I} \\ \mathbf{I} & \mathbf{\Pi}_{22} \end{bmatrix} \tag{30}$$

is a $2m \times 2m$ matrix with m positive eigenvalues and m zero eigenvalues. Its eigenvalues and eigenvectors depend in a simple way on the eigenvalues and eigenvectors of $\mathbf{\Pi}_{22}$. Let λ be a (positive) eigenvalue of $\mathbf{\Pi}_{22}$, with corresponding eigen(row)vector $\boldsymbol{\xi}^{\mathrm{T}}$, then $(\boldsymbol{\xi}^{\mathrm{T}}, \lambda\boldsymbol{\xi}^{\mathrm{T}})$ is an eigenvector of (30) with eigenvalue $\lambda + \lambda^{-1} > 0$ and $(\boldsymbol{\xi}^{\mathrm{T}}, -(1/\lambda)\boldsymbol{\xi}^{\mathrm{T}})$ is an eigenvector of (30) with eigenvalue zero. If follows that $\mathbf{\Pi} - \mathbf{\Pi}^{(2)} \geq 0$ and hence that $\mathbf{\Pi}^{(2)} \leq \mathbf{\Pi}$. The difference is zero if and only if $[\hat{\mathbf{B}}_1^{\mathrm{T}}\mathbf{\Pi}_{12}\hat{\mathbf{B}}_2^{\mathrm{T}}]$ is in the null space of (30). This is a serious restriction, an example of which will be given in the next section. In all other cases we have to end with the rather disturbing but intuitively plausible result that assuming variables to be constant over samples, when they actually vary over samples, yields too small, that is too optimistic, standard deviation estimates for multivariate estimators of the covariance function type. From the numerical example in section 7 it will be seen that this underestimation may be quite serious.

6. THE MODEL ASSUMPTION

In this section we shall look for the difference between the best-fitting approach outlined above and the classical approach. We consider at first the most simple case of OLS regression.

We have a sample of i.d.d. two-dimensional observations, measured as deviations from their means, $\{\mathbf{X}_t\} = \{Y_t, Z_t\}_{t=1}^{T}$ and we look for a linear regression line through the observations that gives the best fit to the sample in the sense of the OLS distance definition given in section 2, i.e. $(1/T)\sum_{t=1}^{T}(Y_t - \beta Z_t)^2$ is minimized with respect to β, yielding an estimator $b = \sum Y_t Z_t / \sum Z_t^2$. Its large-sample distribution has been derived in the introductory section.

Let us write $\hat{\mathbf{X}}_t = (\beta Z_t, Z_t)$, then the residual vector is $\mathbf{X}_t - \hat{\mathbf{X}}_t$ and we may write $Y_t = \hat{b}Z_t + (Y_t - \hat{Y}_t) = \hat{b}Z_t + e_t$. So by the definition of the residual e_t any Y_t may be written as the sum of a structural part and a residual. Minimization of the sum of squared residuals entails the normal equation $(1/T)\sum_{t=1}^{T}(Y_t - \hat{Y}_t)Z_t = 0$, or in words the residual is uncorrelated with the explanatory variable Z_t. A similar argument may be pursued for the population space.

It follows that for any random vector $\mathbf{X} = (Y, Z)$ with zero mean we can maintain that there is a β such that in the population space

1. $Y = \beta Z + \varepsilon$
2. $E(\varepsilon^2)$ is minimized (31)
3. $E(\varepsilon) = 0, E(\varepsilon Z) = 0$

It follows also that for any random sample $\{\mathbf{X}_t\} = \{Y_t, Z_t)\}_{t=1}^{T}$ we may find a $\hat{b}$ such that

1. $Y_t = \hat{b}Z_t + e_t$
2. $\frac{1}{T}\sum e_t^2$ is minimized (32)
3. $\frac{1}{T}\sum (e_t Z_t) = 0, \frac{1}{T}\sum e_t = 0$

The properties 1, 2 and 3 are the traditional assumptions of OLS regression. Logically they appear not to be assumptions or postulates for they hold always. What then is the difference between classical OLS regression and the approach outlined above? That there is a difference

is clearly demonstrated by the fact that we found in section 1 an expression for $\text{var}(\hat{b})$ different from the classical expression. The difference with the classical model has to do with the nature of the disturbance term ε.

Consider formula (3). We may write $\hat{\sigma}_{YZ} = \hat{\sigma}_{(\beta Z+\varepsilon)Z} = \beta\hat{\sigma}_{ZZ} + \hat{\sigma}_{\varepsilon Z}$. Using this expression we obtain for the variance of the normal approximation of the large sample distribution

$$T\,\text{var}(\hat{b}) = T\Bigg\{\frac{1}{(\sigma_{ZZ})^2}\,[\beta^2\,\text{var}(\hat{\sigma}_{ZZ}) + 2\beta\,\text{cov}(\hat{\sigma}_{ZZ}, \hat{\sigma}_{Z\varepsilon}) + \text{var}(\hat{\sigma}_{Z\varepsilon})]$$

$$-\frac{2\beta}{\sigma_{ZZ}^2}\,[\beta\,\text{var}(\hat{\sigma}_{ZZ}) + \text{cov}(\hat{\sigma}_{Z\varepsilon}, \hat{\sigma}_{ZZ})] + \frac{\beta^2}{\sigma_{ZZ}^2}\,\text{var}(\hat{\sigma}_{ZZ})\Bigg\} \tag{33}$$

After reordering this expression it simplifies to

$$\text{var}(\hat{b}) = \frac{1}{\sigma_{ZZ}^2}\,\text{var}(\hat{\sigma}_{Z\varepsilon}) \tag{34}$$

The classical formula is

$$\text{var}(\hat{b}) = \frac{1}{T\sigma_{ZZ}^2}\,\sigma_{ZZ}\sigma_{\varepsilon\varepsilon} = \frac{\sigma_{\varepsilon\varepsilon}}{T\sigma_{ZZ}} \tag{35}$$

The question is now when $\text{var}(\hat{\sigma}_{Z\varepsilon})$ may be written as $\sigma_{ZZ}\sigma_{\varepsilon\varepsilon}/T$. This is possible if A and ε are mutually independent. To be sure, cases may be constructed of 'higher-than-fourth-order' dependence where $\text{var}(\hat{\sigma}_{Z\varepsilon}) = \sigma_{ZZ}\sigma_{\varepsilon\varepsilon}$ still holds, but for all practical purposes the distinction between those cases and independence is irrelevant.

In the present context where we give special attention to the covariance matrix of the sample moments it is interesting to realize that equality of (34) and (35) implies a restriction on $\mathbf{\Pi}$. Writing $\Delta\beta = b - \beta$, we have

$$\Delta\beta = \frac{1}{\sigma_{ZZ}}\,(\Delta\sigma_{YZ} - \beta\Delta\sigma_{ZZ}) \tag{36}$$

The variance of the expression between brackets equals $\text{var}(\hat{\sigma}_{Z\varepsilon})$. Equality of (34) and (35) implies

$$\mathbf{\Pi}_{YZ,YZ} - 2\beta\mathbf{\Pi}_{YZ,ZZ} + \beta^2\mathbf{\Pi}_{ZZ,ZZ} = \sigma_{\varepsilon\varepsilon}\sigma_{ZZ}/T \tag{37}$$

with

$$\beta = \sigma_{YZ}/\sigma_{ZZ} \quad \text{and} \quad \sigma_{\varepsilon\varepsilon} = \sigma_{YY} - \sigma_{YZ}^2/\sigma_{ZZ}$$

The implication of (37) is a rather strong restriction on the second and higher moments of (Y, Z).

For intuitive reasons we would prefer to speak of a true model to exist if Z and ε are *mutually independent*. A straightforward generalization in terms of our general approach is then the following model definition:

Definition of a model

Let $\{\mathbf{X}_t\}_{t=1}^{T}$ stand for a random sample in R^k with existing fourth-order moments, then we say that $\{\mathbf{X}_t\}$ obeys a model if there exist a proper linear subspace S and a metric $\mathbf{Q}$ on R^k, such that the random observation $\mathbf{X}$ may be written as

$$\mathbf{X} = \hat{\mathbf{X}} + \varepsilon$$

where $\hat{\mathbf{X}} = \hat{\mathbf{x}}_S(\mathbf{X})$ is the projection of $\mathbf{X}$ on S according to the $\mathbf{Q}$-metric, and where $\hat{\mathbf{X}}$ and ε are *mutually independent*. This is clearly the case if $\mathbf{X}$ is drawn from a multivariate normal distribution $N(0, \Sigma)$ where $\mathbf{Q} = \Sigma^{-1}$.

7. A NUMERICAL EXAMPLE

Our database consists of a cross-section of 2206 observations of households. For each household we know family size fs, its household income after-tax y_c and an estimate by the head of the household of the minimum income $y_{\min}$ 'he would need to make ends meet for his household'. Obviously this amount $y_{\min}$ depends on fs, while it is 'anchored' to current income y_c. So a relation

$$\ln y_{\min} \approx \beta_0 + \beta_1 \ln \text{fs} + \beta_2 \ln y_c \tag{38}$$

lies at hand. For the theoretical background, which is evidently psychologically flavoured, we refer to References 38–40. Our regression estimate is

$$\ln y_{\min} = 4{\cdot}563 + 0{\cdot}157 \ln \text{fs} + 0{\cdot}508 \ln y_c, \qquad R^2 = 0{\cdot}527, N = 2206$$

The standard deviations of the regression coefficients are evaluated under various assumptions. We compare in Table I the following assumptions:

1. $y_{\min}$, fs, y_c are i.i.d. random observations (section 3).
2. $y_{\min}$, fs, y_c are drawn from an elliptical parent distribution with $3\varkappa = -1, 0, 2, 4, 6$ (this includes the case of a normal parent distribution ($\varkappa = 0$), section 4).
3. The sample is a *controlled* experiment with respect to fs, y_c.
4. $y_{\min}$, fs, y_c are drawn from a sample, *repeated* with respect to fs, y_c (section 5).
5. The sample is drawn from a population for which the approximating model is the true model, that is there holds

$$\ln Y_{\min} = \beta_0 + \beta_1 \ln \text{fs} + \beta_2 \ln y_c + \varepsilon$$

   with ε a normal error, not depending on the random variables fs and y_c, but fs and y_c are random. The joint distribution of fs, y_c is unknown.
6. The sample satisfies the traditional OLS assumption, i.e. a controllable experiment (fs, y_c non-random) and the model holds with $\varepsilon \sim N(0, \sigma^2)$.
7. Finally we repeated the minimal approach in (1) and investigated its sensitivity to the sample size. We partitioned the sample into 11 subsamples of 200 observations each (ignoring the last 6) and calculated the s.d.s for these samples. In line 7 we report the average s.d together with their sample deviation over the 11 estimators in parentheses. As hopefully expected, those s.d.s are about $\sqrt{11}$ times as large, and the small-sample deviations indicate that even for relatively small samples the deviation about the mean is quite acceptable.

The first rather striking result emanating from Table I is that the reliability assessment of estimates, as reflected by their standard deviations (s.d.), varies a great deal with the prior assumptions made. For instance, when we consider the elliptical assumption with a varying value of $\varkappa$ we find that normality is just one parameter choice among many and that the reliability of the same estimator on the same sample decreases rapidly if we opt for greater values of $\varkappa$. Not by coincidence do the classical linear model assumptions yield standard deviations identical to that under the normal assumption ($\varkappa = 0$).[2] From the table it is also evident that cases 3 and 4 are different, but that 3 and 5 are equivalent. It follows that the reliability estimates of the classical linear model (or of the normal parent population) cannot have any

Table I. Standard deviations under various assumptions

	β_0	β_1	β_2
1. Random sample	0·330	0·017	0·034
2. Elliptical sample			
$\varkappa = -1$	0·075	0·007	0·008
$\varkappa = 0$	0·129	0·012	0·013
$\varkappa = 2$	0·167	0·015	0·017
$\varkappa = 4$	0·197	0·018	0·020
$\varkappa = 6$	0·223	0·020	0·023
3. Controlled experiment	0·262	0·017	0·026
4. Repeatable sample	0·227	0·013	0·023
5. Linear model with normal error	0·262	0·017	0·026
6. Classical linear model	0·129	0·012	0·013
7. 11 subsamples	0·820	0·048	0·084
	(1·923)	(0·043)	(0·095)

special claim as being more valid than other estimates presented in the table. This is especially relevant as most s.d. estimates presented in Table I are much larger, which implies that the OLS standard deviations may give a rather optimistic outlook on the reliability. The values of the t-statistics used for significance tests may be too high; similar considerations hold for other frequently used test statistics. The main conclusion is that all assumptions 2 to 6, also that of OLS, are arbitrary choices and consequently that all reliability estimates have an arbitrary basis as well.

Obviously, there is only one reliability estimate which does not suffer from arbitrary assumptions. That is the estimator under minimal assumptions investigated in section 3, of which the corresponding standard deviations are given in line 1. It is seen that for this example the standard deviations under minimal assumptions are larger than for most other assumptions. However, this is not always true (cf. $\varkappa = 6$). The only fact that always holds is that the assumption of repeatability diminishes standard deviations compared to the minimal assumption.

In this paper we have only considered the stochastic properties of large-sample estimators of covariance functions. In classical statistics parameter estimation is the counterpart.

In this paper we do not touch on the extension to the testing of hypotheses, primarily for lack of space. However, it is fairly easy to test hypotheses for their credibility. Let us assume that our parameter of interest is $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ and let our hypothesis be that the true $\mathbf{B} \in A$ where A is a set in the parameter space. As the asymptotic distribution of $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ is known, also the chance $P(\mathbf{B}(\hat{\boldsymbol{\Sigma}}) \in A)$ may be assessed. If it is small, we have to reject the hypothesis. For instance, we may construct 95 per cent confidence ellipsoids as

$$\{\mathbf{B} \mid (\mathbf{B} - \hat{\mathbf{B}})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{B} - \hat{\mathbf{B}}) < \chi^2_{0\cdot 95}\}$$

using the fact that RHS in the inequality is χ^2-distributed. It does not seem relevant to consider traditional F-tests, as they are based on the linear model assumptions with respect to residuals, which we do not employ in our approach.

8. CONCLUDING REMARKS

In this paper we consider multivariate estimators as estimators of approximating models in the population space. We explicitly do not assume that those approximating models are the *true*

structure; we even do not take a standpoint as to whether the observations are generated by a structural model of any kind at all.

Under our minimal assumption multivariate estimators are just covariance functions $\hat{\mathbf{B}} = \mathbf{f}(\hat{\boldsymbol{\Sigma}})$, where $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$ and hencce $\mathbf{f}(\hat{\boldsymbol{\Sigma}})$ of the population parameter $\mathbf{B} = \mathbf{f}(\boldsymbol{\Sigma})$. If T is large, it implies that $\text{vec}(\Delta\mathbf{B}) \approx \dot{\mathbf{B}}_{\Sigma}^{\mathrm{T}} \text{vec}(\Delta\boldsymbol{\Sigma})$, where $\text{vec}\Delta\boldsymbol{\Sigma}$ is asymptotically normal and $\dot{\mathbf{B}}_{\Sigma}^{\mathrm{T}}$ a gradient vector. It follows then that $\Delta\mathbf{B}$ may be seen as a linear combination of a normal vector $\text{vec}\Delta\boldsymbol{\Sigma}$, and that

$$T^{1/2}\,\text{vec}(\Delta\hat{\mathbf{B}}) \to \mathrm{N}(0, \mathbf{V}) \quad \text{with} \quad \mathbf{V} = \dot{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\Pi}\dot{\mathbf{B}} \tag{39}$$

where $\boldsymbol{\Pi}$ is the asymptotic covariance matrix of $T^{1/2}\,\text{vec}\Delta\hat{\boldsymbol{\Sigma}}$. The specification of the vector $\hat{\mathbf{B}}$ depends on the specific multivariate estimator used. The distribution of $\Delta\mathbf{B}$ depends on the stochastic properties of $\Delta\boldsymbol{\Sigma}$. It follows that different multivariate estimators derived from *one* sample have basically the same stochastic properties, determined by $\boldsymbol{\Pi}$. The functional specification $\dot{\mathbf{B}}$ does not depend on the sample, but only on the population. Its evaluation $\mathbf{B}(\hat{\boldsymbol{\Sigma}})$ as a consistent estimate of $\mathbf{B}(\boldsymbol{\Sigma})$ requires only that $\hat{\boldsymbol{\Sigma}}$ is consistent.

If specific additional assumptions are made, either with respect to the population distribution of the observed vector $\mathbf{X}$, or on the sampling procedure, or on the existence of a true model generating $\mathbf{X}$, they affect only $\boldsymbol{\Pi}$ and not $\mathbf{B}$ and $\dot{\mathbf{B}}$. It follows that the stochastic properties of all multivariate estimators will be influenced in a similar manner. Only the middle factor in (39) changes.

When looking at multivariate estimators it follows that there are two dimensions, the choice of the function $\mathbf{B} = \mathbf{f}(\boldsymbol{\Sigma})$ and the stochastic properties of $\hat{\boldsymbol{\Sigma}}$. We call this decomposition the *population-sample decomposition* (PSD). Compared to the established theory there seem to be two major methodological advantages inherent in the method we advocate.

1. Usually one derives the distribution of $\hat{\mathbf{B}}$ under a model hypothesis H_0. The choice of H_0 is mostly loaded with arbitrary elements, one of the heaviest being the requirement of simple calculations. It follows, however, that the reliability assessment of an estimate depends on the arbitrary choice of H_0.
2. The distribution of $\hat{\mathbf{B}}$ is calculated under H_0. However, if one or all of the assumptions of H_0 are violated, we do not know the distribution of $\hat{\mathbf{B}}$ (cf. however References 4 and 14 under the alternative). In our approach we do not make use of a maintained hypothesis H_0 and an alternative H_1. This implies that we do not meet such difficulties. The result under minimal assumptions derived with respect to the distribution of $\hat{\mathbf{B}}$ holds always. If there is an additional structure, its impact will show up in the general result automatically, without making any analytical provisions.

ACKNOWLEDGEMENTS

B. M. S. van Praag is at present affiliated with the Econometric Institute of Erasmus University, Rotterdam, but was with the Center for Research in Public Economics, Leyden University during most of the preparation time. Moreover he benefitted from being a fellow at the Netherlands Institute of Advanced Studies (N.I.A.S.) during the year 1983–1984. The paper has been produced with the support of the Nederlands Foundation for the Advancement of Pure Science. We are very grateful to A. Bertram Wesselman and Jan Koster who helped us in various stages with their critical but constructive remarks and programming support for section 7.

REFERENCES

1. T. W. Anderson, 'Estimating linear statistical relationships', *The Annals of Statistics*, **12**, 1–45 (1984).
2. A. Goldberger, *Econometric Theory*, Wiley, New York, 1964.
3. E. Malinvaud, *Statistical Methods of Econometrics*, North Holland Publishing Company, Amsterdam, 1970.
4. H. White, *Asymptotic Theory for Econometricians*, Academic Press Inc., Orlando, Florida, U.S.A., 1984.
5. J. L. Doob, 'The limiting distribution of certain statistics', *Annals of Mathematical Statistics*, **6**, 160–170 (1935).
6. H. B. Mann and A. Wald, 'On stochastic limit and order relationships', *Annals of Mathematical Statistics*, **14**, 217–226 (1943).
7. H. Cramér, *Mathematical Methods of Statistics*, University Press, Princeton, 1946.
8. C. R. Rao, *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.
9. J. Tiago de Oliveira, 'The δ-method for obtention of asymptotic distributions; applications', *Publications de l'Institut de Statistique de l'Université de Paris*, **27**, (1), 27–48 (1982).
10. B. Efron, 'Bootstrap methods: another look at the jackknife', *The Annals of Statistics*, **7**, 1–26 (1979).
11. P. Diaconis and B. Efron, 'Computer-intensive methods in statistics', *Scientific American*, May, 96–108 (1983).
12. D. A. Freedman and S. C. Peters, 'Bootstrapping an econometric model: some empirical results', *Journal of Business and Economic Statistics*, **2**, 150–158 (1984).
13. D. A. Freedman and S. C. Peters, 'Bootstrapping a regression equation: some empirical results', *Journal of the American Statistical Association*, **79**, 97–106 (1984).
14. H. White, 'Using least squares to approximate unknown regression functions', *International Economic Review*, **21**, 149–170 (1980).
15. H. White 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1982).
16. B. M. S. Van Praag, 'The multivariate approach in linear regression theory', in L. C. A. Corsten and J. Hermans (eds), *Proceedings in Computational Statistics, COMPSTAT 1978*, Physika Verlag, Vienna, 1978.
17. B. M. S. Van Praag, 'Model-free regression', *Economics Letters*, **7**, 139–144 (1980).
18. G. Chamberlain, 'Multivariate regression models for panel data', *Annals of Applied Econometrics*, **1**, 5–46 (1982).
19. M. W. Browne, 'Covariance structures', in D. M. Hawkins (ed.), *Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge, 1982.
20. J. H. Steiger and A. R. Hakstian, 'The asymptotic distribution of elements of a correlation matrix', *British Journal of Mathematical and Statistical Psychology*, **37**, 62–83 (1982).
21. P. M. Bentler, 'Simultaneous equation systems as moment structure models with an introduction to latent variable models', *Journal of Econometrics*, **22**, 13–42 (1983).
22. P. M. Bentler, 'Some contributions to efficient statistics in structural models: specification and estimation of moment structures', *Psychometrika*, **48**, 493–517 (1983).
23. J. De Leeuw, 'Models and methods for the analysis of correlation coefficients', *Journal of Econometrics*, **22**, 113–137 (1983).
24. M. W. Browne, 'Asymptotically distribution-free methods for the analysis of covariance structures', *British Journal of Mathematical and Statistical Psychology*, **37**, 62–83 (1984).
25. P. M. Bentler and T. Dijkstra, 'Efficient estimation via linearization in structural models', in P. R. Krishnaiah (ed.), *Multivariate Analysis VI*, North Holland Publishing Company, Amsterdam, 1984.
26. A. Mooijaart, 'Factor analysis for nonnormal variables', *Psychometrika*, **49**, (1984).
27. B. M. S. Van Praag, T. K. Dijkstra and J. van Velzen, 'Least-squares theory based on general distribution assumptions with an application to the incomplete observations problem', *Psychometrika*, **50**, 25–35 (1985).
28. B. M. S. Van Praag and A. M. Wesselman, 'The hot-deck method. An analytical and empirical evaluation', *Computational Statistics Quarterly*, **1**, (3), 205–231 (1984).
29. B. M. S. Van Praag and J. T. A. Koster, 'Specification in simultaneous linear equations models: the relation between *a priori* specifications and resulting estimators' in T. K. Dijkstra (ed.), *Misspecifications Analysis*, Springer-Verlag, Berlin, 1984.
30. D. S. G. Pollock, *The Algebra of Econometrics*, Wiley, New York, 1979.
31. A. Zellner, 'An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias', *Journal of the American Statistical Association*, **57**, 348–368 (1962).
32. J. De Leeuw, A. Mooijaart and R. Van der Leeden, 'Fixed factor score models with linear restrictions', submitted for publication, 1985.
33. R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
34. J. Magnus and H. Neudecker, 'The elimination matrix: some theorems and applications', *SIAM Journal on Algebraic and Discrete Methods*, **1**, 422–449 (1980).
35. Ch. E. McCulloch, 'Symmetric matrix derivatives with applications', *Journal of the American Statistical Association*, **77**, 679–682 (1982).
36. M. C. Kendall and A. Stuart, *Advanced Theory of Statistics*, Vol. 1, Griffin, London, 1967.
37. G. P. Steck, 'Limit theorems for conditional distributions', University of California Publications in Statistics, **2**, 237–284, 1957.
38. H. Helson, *Adaptation-Level Theory; An Experimental and Systematic Approach to Behavior*, Harper, New York, 1964.

39. Th. Goedhart, V. Halberstadt, A. Kapteyn and B. M. S. Van Praag, 'The poverty line: concept and measurement', *The Journal of Human Resources*, **12**, 503–520 (1977).
40. B. M. S. Van Praag, 'The welfare function of income in Belgium; an empirical investigation', *European Economic Review*, **2**, 337–369, (1971).
41. B. M. S. Van Praag, 'Household cost functions and equivalence scales', Report 8527/A, Econometric Institute, Erasmus University, Rotterdam (1985).