Efficiency and Robustness in a Geometrical Perspective

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

Electronic mail: **russell@ehess.cnrs-mrs.fr**

Abstract

A geometrical setting is constructed, based on Hilbert space, in which the asymptotic properties of estimators can be studied. Estimators are defined in the context of parametrised models, which are treated as submanifolds of an underlying Hilbert manifold, on which a parameter-defining mapping is defined as a submersion on to a finite-dimensional parameter space. Robustness of an estimator is defined as its root-n consistency at all points in the model, and efficiency is based on the criterion, natural in the Hilbert space setting, of the asymptotic variance. Robustness and efficiency at a given data-generating process (DGP) are given geometrical characterisations in terms of a finite-dimensional subspace, associated with asymptotically efficient estimation, of the tangent space at that DGP. Starting from an arbitrary consistent estimator, it is shown how a one-step efficient estimator can be obtained by orthogonal projection on to the efficient subspace. Examples are given, based mostly on the linear regression model.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. It is loosely based on unpublished joint work with Stanley E. Zin, to whom I am grateful for many useful discussions.

May, 1998

1. Introduction

The aim of this paper is to construct a general geometrical setting, based on Hilbert space, in which one may study various estimation techniques, in particular with respect to efficiency and robustness. Given the sort of data one wishes to study, such as continuous, discrete, *etc.*, the set of data-generating processes (DGPs) capable of generating data of that sort is given the structure of a Hilbert manifold. Statistical *models* will be treated as submanifolds of this underlying manifold.

Geometrical methods are frequently used in the study of statistical inference. One important strand of the literature is presented in Amari (1990), whose numerous earlier papers were inspired by some very abstract work of Chentsov (1972) and led to the concept of a *statistical manifold.* Other review papers and books in this tradition include Barndorff-Nielsen, Cox, and Reid (1986), Kass (1989), Murray and Rice (1993), and Barndorff-Nielsen and Cox (1994). Most of this work makes use of finite-dimensional differential manifolds, which are usually representations of models in the exponential family.

Infinite-dimensional Hilbert space methods are extensively used in another strand of literature, for which the most suitable recent reference is Small and McLeish (1994). This book contains numerous references to the original papers on which it builds. In this work, random variables are represented as elements of Hilbert space, and different probability measures (that is, different DGPs) correspond to different inner products on the Hilbert space. However, no manifold structure is imposed on the set of inner products, so that the set of DGPs, rather than the set of random variables, is not given a geometrical interpretation. Nevertheless, Small and McLeish's approach provides most of the geometrical elements used in this paper.

Davidson and MacKinnon (1987) introduced infinite-dimensional statistical manifolds, with Hilbert manifold structure, in a manner similar to that used by Dawid (1975, 1977). Infinite-dimensional differential manifolds are less frequently encountered than finite-dimensional ones, but see Lang (1972) for an excellent account. The use of infinite-dimensional manifolds avoids the need to limit attention to models in the exponential family. In this paper, that Hilbert space representation is extended, and adapted for use in the context of asymptotic theory.

In this paper, *estimators* are defined in such a way as to correspond to elements of the *tangent spaces* to the statistical manifold at DGPs belonging to the manifold. In fact, an interpretation of these tangent spaces is given as the space of random variables with zero mean and finite variance under the DGP at which the space is tangent. Since a tangent space to a Hilbert manifold is itself a Hilbert space, it can, under this interpretation, be identified with the subspace of Small and McLeish's Hilbert space corresponding to zero-mean random variables.

The principal focus in this paper is on estimators defined by the *Method of Estimating Functions*, as proposed by Godambe (1960). This method is essentially equivalent to the method known in the econometrics literature as the *Generalised Method of Moments*, introduced by Hansen (1982). With little effort, the results given in this paper can be extended to Manski's (1984) Closest Empirical Distribution class of estimators. The efficiency and/or robustness of an estimator is always treated relative to a statistical *model*, treated as a Hilbert submanifold. Since estimators estimate *parameters*, they are defined relative to a *parameter-defining mapping* defined on the model. A *parametrised model* is just the pair consisting of the model and the parameter-defining mapping.

A major result of the paper is that the tangent space to the underlying statistical Hilbert manifold at a DGP belonging to a parametrised model is the direct sum of three mutually orthogonal subspaces. The model, being a Hilbert submanifold, has its own tangent space at any DGP in it, this being a subspace of the full tangent space. The first of the three subspaces is just the orthogonal complement of the tangent space to the model. The other two are therefore complementary subspaces of the model tangent space. Of these, one is the tangent space to the subset of the model for which the model parameters do not vary, and the other, orthogonal to it, turns out to be the finite-dimensional space in which (asymptotically) *efficient* estimators are located.

Robustness of an estimator with respect to a given model is interpreted as meaning that the estimator is root-n consistent for all DGPs in the model. The property of root-n consistency is shown to have a geometrical interpretation according to which the tangents that represent the estimator are orthogonal to the second subspace described above, the one which is tangent to the space over which the parameters do not vary. Quite generally, a root-n consistent estimator can be made efficient at any given DGP by projecting it orthogonally in Hilbert space on to the finite-dimensional third subspace. Such orthogonal projections can be achieved by making use of a particular privileged basis of the third subspace, a basis that is easy to characterise in terms of Godambe's estimating functions.

In the next section, the Hilbert manifold of DGPs is constructed, and it is shown how to adapt it for use with asymptotic theory. Then in section 3, estimators are defined in a geometrical context, as also the concepts of efficiency and robustness of estimators. The main results pertaining to the three-subspace decomposition of the tangent space are proved in this section. Section 4 is an interlude of examples and illustrations, and in section 5 the results are specialised to estimators defined by estimating functions and the generalised method of moments. In section 6, the linear regression model is used as the simplest example in which the results of section 5 can be deployed, and a nontrivial application of orthogonal projection is given. Finally, concluding comments are found in section 7.

2. Data-Generating Processes in Hilbert Space

In Davidson and MacKinnon (1987) a Hilbert space representation was introduced for the set of data-generating processes (DGPs) that could have generated a given data set. The representation used here is a slight generalisation of that presented there, in that we will not restrict ourselves to samples of i.i.d. random variables.

First, it is assumed that the DGPs we are concerned with are defined on a measure space $(\Omega, \mathcal{F})$. A DGP corresponds to a probability measure, P say, defined on this space. Observed data, $\boldsymbol{y}^n \equiv \{y_t\}_{t=1}^n$, say, for a sample of size n, are interpreted as realisations of random variables on $(\Omega, \mathcal{F})$. Thus, if each observation has m components (there are m simultaneously observed dependent variables), then for each $t = 1, 2, \ldots, n$ there exists a mapping $Y_t : \Omega \to \mathbb{R}^m$, and for each sample size $n = 1, 2, \ldots$ a mapping $Y^n : \Omega \to \mathbb{R}^{nm}$, where Y_t and Y^n are respectively the random variable for observation t and the random variable for a complete sample of size n. Their stochastic properties are given by the probability measure P, or, equivalently, by the measure that P induces on $\mathbb{R}^{mn}$ by the mapping Y^n.

A *model*, for a given sample size n, will be thought of as a set of DGPs, that is, as a set of probability measures on $\mathbb{R}^{mn}$. We assume that there exists a *carrier measure* P_0^n on $\mathbb{R}^{mn}$ such that the measures associated with all DGPs in the model are absolutely continuous with respect to it. By the Radon-Nikodym theorem, this ensures for each DGP in the model the existence of a probability density for the random variable Y^n.

Consider now one single DGP in the model, and denote the density of Y^n by $L^n : \mathbb{R}^{mn} \to \mathbb{R}$. Since this is the joint density of the Y_t, $t = 1, \ldots, n$, it can be factorised as follows:

$$L^n(y_1, \ldots, y_n) = \prod_{t=1}^{n} L_t(y_t \mid y_{t-1}, \ldots, y_1), \tag{1}$$

where L_t denotes the density of the t^{th} observation, Y_t, conditional on all the observations before it in the ordering $\{1, 2, \ldots, n\}$, that is, the observations 1 through $t - 1$.

We may now make contact with the representation given in Davidson and MacKinnon (1987), by considering, not the density (1), but its square root. Analogously to (1), we write

$$\psi^n(y_1, \ldots, y_n) = \prod_{t=1}^{n} \psi_t(y_t \mid y_{t-1}, \ldots, y_1), \tag{2}$$

where $L^n(y_1, \ldots, y_n) = (\psi^n(y_1, \ldots, y_n))^2$, with a similar relation between $L_t(\cdot)$ and $\psi_t(\cdot)$, $t = 1, \ldots, n$. By construction, ψ^n belongs to the Hilbert space $L^2(\mathbb{R}^{mn}, P_0^n)$, in fact to the *unit sphere* of that space, since the integral of the square of ψ^n with respect to $d\boldsymbol{y}^n \equiv dy_1 dy_2 \ldots dy_n$ equals one. We write $\mathcal{H}^n$ for this unit sphere.

Usually we choose ψ^n and the ψ_t to be the nonnegative square roots of L^n and the L_t, but this is not necessary. Indeed, in Hilbert space, it is impossible to limit oneself to nonnegative square-root densities, since the nonnegative cone in an infinite-dimensional Hilbert space has an empty interior, and thus does not have a manifold structure. A consequence of this is that we cannot represent a given DGP *uniquely* in Hilbert space, but this does not matter for anything in this paper. Hilbert space, on the other hand, is the natural setting for mean-square convergence, and has the considerable advantage that the information matrix – to be defined later – is a smooth tensor in this representation. This would not be so if we used, for instance, the log of the density in place of the square root density.

It is clear from (2) that a convenient way to deal with arbitrary sample sizes is to consider infinite sequences of *contributions* $\{\psi_t\}_{t=1}^{\infty}$. For any given sample size n, the joint square-root density of the n observations is given by (2). For a given infinite sequence to define a DGP for each n, it is necessary and sufficient that

$$\int |\psi_t(y_t \mid y_{t-1}, \ldots, y_1)|^2 \, dy_t = 1 \tag{3}$$

for all possible values of the conditioning variables $y_1, \ldots, y_{t-1}$. We denote by $\mathbb{S}$ the set of sequences satisfying these conditions, and consider $\mathbb{S}$ as the space of DGPs for asymptotic theory, since, given any element of $\mathbb{S}$, a proper probability density can be defined for arbitrary sample size. A *model*, for the purposes of asymptotic theory, will thus be a subset of $\mathbb{S}$.

Consider first, for a given n, the *tangent space* to the unit sphere $\mathcal{H}^n$ at some DGP $\psi^n \in \mathcal{H}^n$. A tangent at ψ^n is associated with a smooth *curve* in $\mathcal{H}^n$ through ψ^n. Such a curve is a one-parameter family of DGPs that includes ψ^n. Let the curve be denoted by $\psi^n(\epsilon)$, $\epsilon \in]-1, 1[$, and $\psi^n(0) = \psi^n$. The tangent to this curve at ψ^n is then represented by the derivative of $\psi^n(\epsilon)$ at $\epsilon = 0$. The appropriate derivative in Hilbert space is a mean-square derivative, $(\psi^n)' \in L^2(\mathbb{R}^{mn}, P_0^n)$, say, that satisfies

$$\lim_{\epsilon \to 0} \left\| \frac{1}{\epsilon}\big(\psi^n(\epsilon) - \psi^n(0)\big) - (\psi^n)' \right\| = 0, \tag{4}$$

where $\|\cdot\|$ is the Hilbert space norm in $\mathcal{H}^n$.

Consider next a curve in $\mathbb{S}$ through the point $\psi \equiv \{\psi_t\}_{t=1}^{\infty}$. Denote the curve by $\psi(\epsilon)$, and, for each n, we have a curve in $\mathcal{H}^n$ given by

$$\psi^n(\epsilon) = \prod_{t=1}^{n} \psi_t(\epsilon).$$

In order to define the tangent to the curve $\psi(\epsilon)$, and for the purposes of asymptotic theory more generally, it is more convenient to consider, not the sequence $\{\psi^n(\epsilon)\}$ for a fixed ϵ, but rather the sequence

$$\big\{\psi^n(n^{-1/2}\epsilon)\big\}_{n=1}^{\infty}.$$

On differentiating with respect to ϵ at $\epsilon = 0$, this gives the following representation for the tangent to $\psi(\epsilon)$ at ψ:

$$\Big\{n^{-1/2}(\psi^n)'\Big\}_{n=1}^{\infty}, \tag{5}$$

where each $(\psi^n)'$ is defined as in (4).

The reason for the factor of $n^{-1/2}$ is that we can now define the norm of the sequence (5), and thus the norm of the tangent ψ' to $\psi(\epsilon)$ at ψ by the formula

$$\|\psi'\| = \lim_{n \to \infty} \big\| n^{-1/2}(\psi^n)' \big\|_{\mathcal{H}^n}, \tag{6}$$

where the norm of each $(\psi^n)'$ is calculated in the Hilbert space corresponding to sample size n. The limit in (6) will be shown shortly to exist in a wide variety of circumstances. Without the factor of $n^{-1/2}$, this would not be the case. Another way to see why the factor is useful is to note that its use converts the curve $\psi(\epsilon)$

into what Davidson and MacKinnon (1993) call a *drifting DGP*, in the sense of a Pitman drift.

Note that there is no obvious way to embed the contributions $\psi_t(\epsilon)$ in a Hilbert space or manifold, and there is therefore no direct way to compute their derivatives with respect to ϵ. An appropriate indirect way is as follows. Recall that ψ^n with a superscript refers to a product of contributions, while ψ_t with a subscript refers to a single contribution. Then define derivatives ψ'_t recursively by the relations

$$\psi'_1 = (\psi^1)', \qquad \psi^{t-1}\psi'_t = (\psi^t)' - \psi_t(\psi^{t-1})'. \tag{7}$$

For values of $(y_1, \ldots, y_{t-1})$ for which ψ^{t-1} vanishes, ψ'_t is arbitrarily set equal to zero. It should be clear that, whenever the $\psi_t(\epsilon)$ can be differentiated in any useful sense, the derivatives will satisfy (7). With that definition, it is clear that the tangent ψ' can be represented by the infinite sequence of contributions, $\{\psi'_t\}_{t=1}^{\infty}$, such that, for each n,

$$\frac{(\psi^n)'}{\psi^n} = \sum_{t=1}^{n} \frac{\psi'_t}{\psi_t}. \tag{8}$$

The construction of the tangent space at the DGP $\psi \in \mathbb{S}$ as a Hilbert space is almost complete. Tangents are represented by infinite sequences of contributions satisfying (8), with the norm (6). The final step, needed so that (6) should be positive definite, is to identify tangents of zero norm with the actual zero tangent, defined as an infinite sequence of zero contributions. In this way, the Hilbert space that we consider is the space of equivalence classes of infinite sequences of contributions satisfying (8), two sequences being equivalent if the difference between them is a sequence of zero norm using the norm (6). It will be clear shortly that the different elements of equivalence classes so defined are *asymptotically equivalent* in the usual sense of asymptotic theory. The Hilbert space thus defined, the space of tangents to $\mathbb{S}$ at the DGP ψ, will be denoted as $T_S(\psi)$.

It is now possible to give a statistical interpretation of the space $T_S(\psi)$. Consider a curve $\psi(\epsilon)$ and suppose that, for each n and for all admissible values of $\boldsymbol{y}^n \equiv (y_1, \ldots, y_n)$, $\psi^n(\epsilon; \boldsymbol{y}^n)$ is nonzero, so that $\log|\psi^n(\epsilon; \boldsymbol{y}^n)|$ exists everywhere. We remarked above that the curve corresponds to a one-parameter family of DGPs, and it is clear that $\ell^n(\epsilon, \boldsymbol{y}^n) \equiv 2\log|\psi^n(\epsilon; \boldsymbol{y}^n)|$ is the loglikelihood function corresponding to this one-parameter family. Further, $\ell_t(\epsilon; \boldsymbol{y}^t) \equiv 2\log|\psi_t(\epsilon; \boldsymbol{y}^t)|$ is just the contribution to ℓ^n from observation t, and

$$\ell^n(\epsilon; \boldsymbol{y}^n) = \sum_{t=1}^{n} \ell_t(\epsilon; \boldsymbol{y}^t). \tag{9}$$

Assuming now that ℓ^n, ℓ_t, and ψ_t can be differentiated with respect to ϵ, we see that, by (8),

$$\sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \epsilon}(0) = \frac{\partial \ell^n}{\partial \epsilon}(0) = 2\frac{(\psi^n)'}{\psi^n} = 2\sum_{t=1}^{n} \frac{\psi'_t}{\psi_t}. \tag{10}$$

The expression second from the left above is the gradient of the loglikelihood of the one-parameter family at $\epsilon = 0$, and, as such, its expectation under the DGP

ψ^n is zero. In Hilbert space, this result corresponds to a simple orthogonality property, as follows. The expectation of each expression in (10), since the square of ψ^n is the density of $\boldsymbol{y}^n$, can be written as

$$2 \int \frac{(\psi^n)'}{\psi^n} \, (\psi^n)^2 \, d\boldsymbol{y}^n = 2 \int (\psi^n)' \, \psi^n \, d\boldsymbol{y}^n, \tag{11}$$

and the right-hand side of this can be seen to be zero when the normalisation relation

$$\int (\psi^n)^2(\epsilon) \, d\boldsymbol{y}^n = 1,$$

which holds for all admissible ϵ, is differentiated with respect to ϵ and evaluated at $\epsilon = 0$ to yield

$$\int (\psi^n)'\psi^n \, d\boldsymbol{y}^n = 0, \tag{12}$$

which just says that the inner product in $L^2(\mathbb{R}^{mn}, P_0^n)$ of $(\psi^n)'$ and ψ^n is zero, so that $(\psi^n)'$ and ψ^n are orthogonal. Geometrically, this just says that a radius of the unit sphere – ψ^n – is orthogonal to a tangent to that sphere – $(\psi^n)'$.

From (3), it follows that a result like (12) holds for each contribution:

$$\int \psi_t' \, \psi_t \, dy_t = 0,$$

which, in terms of ℓ_t, becomes

$$\int \frac{\partial \ell_t}{\partial \epsilon}(0; \boldsymbol{y}^t) \, \exp\bigl(\ell_t(0; \boldsymbol{y}^t)\bigr) \, dy_t = E\left(\frac{\partial \ell_t}{\partial \epsilon}(0; \boldsymbol{y}^t) \,\middle|\, \boldsymbol{y}^{t-1} \right) = 0.$$

The second equation above implies the well-known result that the sequence

$$\left\{ \sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \epsilon}(0) \right\}_{n=1}^{\infty}$$

is a *martingale* under ψ.

Now consider the norm of the tangent ψ', as given by (6). In order to calculate it, we need the norms, in $L^2(\mathbb{R}^{mn}, P_0^n)$, of the tangents $(\psi^n)'$. These norms are given by the formula

$$\|(\psi^n)'\|^2 = \int \bigl((\psi^n)'\bigr)^2 \, d\boldsymbol{y}^n = \int \left(\frac{(\psi^n)'}{\psi^n} \right)^2 (\psi^n)^2 \, d\boldsymbol{y}^n = E_{\psi^n} \left(\left[\sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \epsilon}(0) \right]^2 \right), \tag{13}$$

where the last equality follows from (10). The martingale property allows (13) to be simplified to

$$\sum_{t=1}^{n} E_{\psi^t} \left(\left[\frac{\partial \ell_t}{\partial \epsilon}(0) \right]^2 \right). \tag{14}$$

From (6), the squared norm of ψ' is

$$\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} E_{\psi^t}\left(\left[\frac{\partial \ell_t}{\partial \epsilon}(0)\right]^2\right),$$

where the limit exists under mild regularity conditions allowing a law of large numbers to be applied. This limit can be interpreted as the limiting (asymptotic) *variance* under ψ of the sequence

$$\left\{n^{-1/2}\sum_{t=1}^{n}\frac{\partial \ell_t}{\partial \epsilon}(0)\right\}_{n=1}^{\infty}. \tag{15}$$

Although (15) is derived from a triangular martingale array rather than being a martingale, we will refer to sequences like (15) as martingales, by a slight abuse of terminology. Exactly similar considerations allow us to express the inner product of two tangents $(\psi^1)'$ and $(\psi^2)'$ in $T_S(\psi)$ as the limit of the covariance of the two random variables

$$n^{-1/2}\sum_{t=1}^{n}\frac{\partial \ell_t^1}{\partial \epsilon^1}(\epsilon^1=0) \quad \text{and} \quad n^{-1/2}\sum_{t=1}^{n}\frac{\partial \ell_t^2}{\partial \epsilon^2}(\epsilon^2=0),$$

in obvious notation.

The above considerations lead to an intuitive understanding of the Hilbert space $T_S(\psi)$ we have constructed. It is the space of equivalence classes of asymptotically equivalent sequences of random variables of the form

$$h=\left\{n^{-1/2}\sum_{t=1}^{n}h_t\right\}_{n=1}^{\infty},$$

where

$$\begin{gathered} E_\psi\big(h_t \mid h_1,\ldots,h_{t-1}\big)=0, \quad t=1,2,\ldots, \\ E_\psi\big(h_t^2\big)=\eta_t<\infty, \quad t=1,2,\ldots, \text{ and} \\ n^{-1}\sum_{t=1}^{n}\eta_t \quad \text{converges as } n\to\infty \text{ to a finite limiting variance.} \end{gathered} \tag{16}$$

The squared norm $\|h\|^2$ of such a sequence is the limiting variance, and the inner product $\langle h^1,h^2\rangle$ of two such sequences is the limiting covariance of

$$n^{-1/2}\sum_{t=1}^{n}h_t^1 \quad \text{and} \quad n^{-1/2}\sum_{t=1}^{n}h_t^2. \tag{17}$$

The construction depends heavily on the martingale property. On account of the variety of central-limit theorems applicable to martingales – see for instance

McLeish (1974) – this property also justifies considering limiting *normal* random variables to which sequences like (17) tend as $n \to \infty$.

The choice of the particular Hilbert space structure just constructed so as to define the tangent space at ψ confers a Hilbert manifold structure on the set $\mathbb{S}$ itself. It is not the aim of the present paper to conduct a full investigation of this structure, since all the remaining analysis of the paper will be local, and so just a few remarks will be made. It is clear that it would be necessary to group the elements of $\mathbb{S}$ into equivalence classes of DGPs with asymptotically equivalent properties. The regularity conditions (16) implicitly restrict the sorts of DGPs admitted into $\mathbb{S}$. These are not so strong as those imposed by Hansen (1982), who worked in a stationary ergodic framework. Methods of the sort used in White and Domowitz (1984) and White (1985) are presumably appropriate for determining just what restrictions are implicit in the present treatment.

3. Efficiency and Robustness in Hilbert Space

All procedures of estimation or inference treated here will be situated in the context of a particular model, that is, a subset of the set $\mathbb{S}$ introduced in the preceding section. If $\mathbb{M}$ denotes such a model, it is almost always interesting to define some parameters for it. A *parametrised model* will therefore be a pair, of the form $(\mathbb{M}, \boldsymbol{\theta})$, where the mapping $\boldsymbol{\theta} : \mathbb{M} \to \Theta$ is termed a *parameter-defining mapping.* The set Θ is a finite-dimensional parameter space, a subset of $\mathbb{R}^k$ for some positive integer k. A *parametrisation* would go the other way, associating a DGP to each parameter vector in Θ.

Models that can be estimated by maximum likelihood constitute a very straightforward class. They are special in that a parametrisation does exist for them: for each admissible parameter vector, and for each sample size, the likelihood function gives a probability density for the dependent variables, which is precisely what we mean by a DGP. The image of the parametrisation is the model, and the parameter-defining mapping is the inverse of the parametrisation. Note that the inverse will not exist if the parametrisation is not one-to-one. In such cases, the model parameters are not identified. A convenient way to impose identification of all the parameters we consider, not just in the context of maximum likelihood models, is to require the existence of a parameter-defining mapping.

In more general circumstances, a given parameter vector corresponds to an infinite number of DGPs. A simple case is that of a linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad E(u_t) = 0, \quad E(u_t^2) = \tau^2, \tag{18}$$

in which the distribution of the error terms is not specified past the first two moments. Any mean-zero error distribution with finite variance can be used in combination with a fixed parameter vector $\boldsymbol{\beta}$ and variance τ^2. Clearly, there is an infinite number of such error distributions.

In order to benefit from the Hilbert space structure introduced in the preceding section, it will be desirable to consider only models $\mathbb{M}$ that are *closed submanifolds* of $\mathbb{S}$. Locally, in a neighbourhood of a DGP $\psi \in \mathbb{M}$, this just means that, if we consider the subset of tangents at ψ generated by curves that lie entirely in the subset $\mathbb{M}$, this subset, denoted by $T_M(\psi)$, should be a closed subspace of the full tangent space $T_S(\psi)$.

If this condition is satisfied, then another regularity condition needed for the rest of the development can be imposed on the parameter-defining mappings that may be used with $\mathbb{M}$. It is that such a mapping must be a *submersion* (see, for instance Lang (1972)). Among the consequences of this technical condition is that, if $\boldsymbol{\theta}$ denotes the parameter-defining mapping, open neighbourhoods of ψ in $\mathbb{M}$ are mapped by $\boldsymbol{\theta}$ into open sets of the parameter space Θ. This avoids redundant parameters: if, for instance, one parameter, θ_1 say, was always just twice another parameter, θ_2, all points in the image of $\boldsymbol{\theta}$ would satisfy $\theta_1 = 2\theta_2$, and so the image could not be an open set. Another consequence, more important for what follows, is that $T_M(\psi)$ can be expressed as the direct sum of two orthogonal subspaces, the first, possibly infinite-dimensional, corresponding to tangents to curves along which the parameters defined by $\boldsymbol{\theta}$ are constant, and the second the orthogonal complement of the first, and necessarily of finite dimension k, where k is the number of parameters defined by $\boldsymbol{\theta}$. A maximum-likelihood model is itself of dimension k, and so the first of these orthogonal subspaces contains only the zero element of $T_M(\psi)$. In general, the first of the subspaces, that for which the parameters are constant, will be denoted as $T_M(\psi, \boldsymbol{\theta})$, and the second as $E(\psi, \boldsymbol{\theta})$. These two subspaces, together with their orthogonal complement in $T_M(\psi)$, comprise the three-subspace decomposition of $T_M(\psi)$ alluded to in the Introduction.

An *estimator* of the parameters of a given parametrised model is a sequence of random k-vectors $\hat{\boldsymbol{\theta}}^n$ which, for each n, are defined solely in terms of the random variable Y^n of which any data set of size n is a realisation. Thus $\hat{\boldsymbol{\theta}}^n$ maps from $\mathbb{R}^{mn}$ to a parameter space Θ. The estimator characterised by the sequence $\{\hat{\boldsymbol{\theta}}^n\}$ will be written as just $\hat{\boldsymbol{\theta}}$. The above definition clearly contains many useless estimators; usually we will be interested only in consistent estimators. The property of consistency can be expressed as follows. For each DGP $\psi \in \mathbb{M}$, we must have

$$\operatorname*{plim}_{n\to\infty}{}_{\psi}\, \hat{\boldsymbol{\theta}}^n = \boldsymbol{\theta}(\psi).$$

The notation means that the probability limit is calculated under the DGP ψ, and that the limit is what is given by the parameter-defining mapping $\boldsymbol{\theta}$ for that DGP.

Most root-n consistent estimators correspond to vectors of tangents at each point of the model for which they are defined. Consider a DGP ψ in a parametrised model $(\mathbb{M}, \boldsymbol{\theta})$, and let $\boldsymbol{\theta}(\psi) = \boldsymbol{\theta}_0$ be the parameter vector for ψ. Then, for a root-n consistent estimator $\hat{\boldsymbol{\theta}}$, construct the vector sequence with typical element

$$\boldsymbol{s}_t \equiv t\big(\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0\big) - (t-1)\big(\hat{\boldsymbol{\theta}}^{t-1} - \boldsymbol{\theta}_0\big). \tag{19}$$

Clearly

$$n^{1/2}\big(\hat{\boldsymbol{\theta}}^n - \boldsymbol{\theta}_0\big) = n^{-1/2}\sum_{t=1}^{n} \boldsymbol{s}_t.$$

The components of $\{\boldsymbol{s}_t\}$ may not exactly satisfy the conditions (16), but they will usually be asymptotically equivalent to sequences that do. Since such asymptotically equivalent sequences are identified in our Hilbert space structure, the estimator $\hat{\boldsymbol{\theta}}$ can be associated with the vector of tangents at ψ defined by the equivalence classes containing the components of $\{\boldsymbol{s}_t\}$. In fact, all estimators that are asymptotically equivalent to $\hat{\boldsymbol{\theta}}$ are associated with the same vector of tangents.

A simple illustration may be helpful here. The OLS estimator of the regression model (18) satisfies the relation

$$n^{1/2}\big(\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}_0\big) = \left(n^{-1}\sum_{t=1}^{n}\boldsymbol{X}_t^{\top}\boldsymbol{X}_t\right)^{-1} n^{-1/2}\sum_{t=1}^{n}\boldsymbol{X}_t^{\top}u_t \tag{20}$$

when the true parameter vector is $\boldsymbol{\beta}_0$. Under standard regularity conditions, $n^{-1}\sum_{t=1}^{n}\boldsymbol{X}_t^{\top}\boldsymbol{X}_t$ tends to a nonrandom, symmetric, positive definite limiting matrix $\boldsymbol{A}$, say. Thus the sequence with typical element (20) is asymptotically equivalent to the sequence

$$\hat{\boldsymbol{s}} \equiv \left\{n^{-1/2}\sum_{t=1}^{n}\boldsymbol{A}^{-1}\boldsymbol{X}_t^{\top}u_t\right\}_{n=1}^{\infty},$$

which clearly obeys the requirements of (16).

If the parameter space Θ is k-dimensional, we may denote the k tangents corresponding to $\hat{\boldsymbol{\theta}}$ at ψ by the vector $\hat{\boldsymbol{s}}$, with typical element $\hat{s}_i$, $i = 1,\dots,k$. It follows from the interpretation of the Hilbert space norm of a tangent as a variance that the $k \times k$ matrix with typical element $\langle \hat{s}_i, \hat{s}_j\rangle$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, that is,

$$\lim_{n\to\infty} \mathrm{Var}\left(n^{1/2}\big(\hat{\boldsymbol{\theta}}^n - \boldsymbol{\theta}_0\big)\right). \tag{21}$$

The notion of *robustness* used in this paper can be defined as follows. Suppose we have two parametrised models $(\mathbb{M}_0, \boldsymbol{\theta}_0)$ and $(\mathbb{M}_1, \boldsymbol{\theta}_1)$, where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ map into the same parameter space Θ, such that $\mathbb{M}_0 \subseteq \mathbb{M}_1$ and $\boldsymbol{\theta}_0(\psi) = \boldsymbol{\theta}_1(\psi)$ for all $\psi \in \mathbb{M}_0$. Then a consistent estimator $\hat{\boldsymbol{\theta}}$ of the parameters of the first model is said to be *robust* with respect to the second if it is also consistent for the second model. (Note that, since $\boldsymbol{\theta}_0 : \mathbb{R}^{mn} \to \Theta$, it satisfies our definition of an estimator of $(\mathbb{M}_1, \boldsymbol{\theta}_1)$.) Thus the OLS estimator of the regression model (18) restricted so as to have normal errors is robust with respect to the full model (18) with arbitrary error distribution satisfying the conditions on the first two moments.

It may happen that the "unrestricted" model $\mathbb{M}_1$ has more parameters than the "restricted" model $\mathbb{M}_0$. The above definition may still be used by limiting the parameter-defining mapping $\boldsymbol{\theta}_1$ to its projection on to those parameters that do appear in $(\mathbb{M}_0, \boldsymbol{\theta}_0)$. For instance, the unrestricted regression

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t\boldsymbol{\gamma} + u_t \tag{22}$$

contains the restricted regression

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \tag{23}$$

as a special case, but has more parameters. In order to see if an estimator for (23) is robust with respect to (22), one just forgets about the $\boldsymbol{\gamma}$ parameters for model (22). It then follows by standard arguments that the OLS estimator of (23) is robust with respect to (22) if and only if $\operatorname{plim}_{n\to\infty} n^{-1}\sum_{t=1}^{n}\boldsymbol{Z}_t^\top\boldsymbol{X}_t=\boldsymbol{0}$.

Robustness is often thought to entail a cost in terms of the *efficiency* of an estimator. One of the chief aims of this paper is to make explicit the tradeoff between these two desirable features. Before we can do so, we need a geometrical characterisation of efficiency. As with robustness, efficiency will always be defined with respect to a given parametrised model $(\mathbb{M},\boldsymbol{\theta})$. A root-$n$ consistent estimator $\hat{\boldsymbol{\theta}}$ is *efficient* for $(\mathbb{M},\boldsymbol{\theta})$ at a DGP $\psi\in\mathbb{M}$ if no other root-n consistent estimator $\check{\boldsymbol{\theta}}$ for $(\mathbb{M},\boldsymbol{\theta})$ has smaller asymptotic variance under ψ. Specifically, the difference between the asymptotic covariance matrix of $\check{\boldsymbol{\theta}}$, given by (21), and that for $\hat{\boldsymbol{\theta}}$ is a positive semi-definite matrix. The geometrical characterisation of efficiency is given by the following theorem.

Theorem 1

> Under the regularity assumed so far, the root-n consistent estimator $\hat{\boldsymbol{\theta}}$ is efficient for the parametrised model $(\mathbb{M},\boldsymbol{\theta})$ at a DGP $\psi\in\mathbb{M}$ if and only if the tangents $\hat{s}_i$, $i=1,\ldots,k$, associated with $\hat{\boldsymbol{\theta}}$ belong to the space $E(\psi,\boldsymbol{\theta})$.

In order to prove this theorem, we will develop in a series of lemmas a number of properties of root-n consistent estimators. First, note that, if the condition of the theorem is true, the $\hat{s}_i$, $i=1,\ldots,k$, span the k-dimensional space $E(\psi,\boldsymbol{\theta})$, since any linear dependence of the $\hat{s}_i$ would imply that the model parameters were not independent, contrary to the assumption that the parameter-defining mapping is a submersion.

Lemma 1

> The tangents $\check{s}_i$, $i=1,\ldots,k$, associated with a root-n consistent estimator $\check{\boldsymbol{\theta}}$ of the parametrised model $(\mathbb{M},\boldsymbol{\theta})$ at a DGP $\psi\in\mathbb{M}$ are orthogonal to the space $T_M(\psi,\boldsymbol{\theta})$.

Proof: If $T_M(\psi,\boldsymbol{\theta})$ consists only of the zero tangent, the lemma is trivial. Otherwise, consider a curve in $\mathbb{M}$ through ψ such that, for all points $\psi(\epsilon)$ on the curve, $\boldsymbol{\theta}(\psi(\epsilon))=\boldsymbol{\theta}(\psi)$. The tangent ψ' to this curve belongs to $T_M(\psi,\boldsymbol{\theta})$ by definition, and any element of $T_M(\psi,\boldsymbol{\theta})$ can be generated by such a curve. Then, for all admissible ϵ, the expectation of $\check{\boldsymbol{\theta}}^n$ under $\psi(\epsilon)$ tends to $\boldsymbol{\theta}_0\equiv\boldsymbol{\theta}(\psi)$ as $n\to\infty$.

Suppose that the curve is expressed in terms of contributions $\ell_t(\epsilon)$, as in (9), and that the tangents $\check{s}_i$ correspond to components $\check{s}_{ti}$ satisfying (16). Then we have

$$0=E_{\psi(\epsilon)}\big(\check{s}_{ti}(\boldsymbol{y}^t)\mid\boldsymbol{y}^{t-1}\big)=\int\exp\big(\ell_t(\epsilon;\boldsymbol{y}^t)\big)\,\check{s}_{ti}(\boldsymbol{y}^t)\,dy_t.$$

From (19), it is clear that, since $\boldsymbol{\theta}(\psi(\epsilon))$ is independent of ϵ, so too is $\check{s}_{ti}$ along the curve $\psi(\epsilon)$. Thus differentiating with respect to ϵ and evaluating at $\epsilon=0$ gives

$$\int\exp\big(\ell_t(0;\boldsymbol{y}^t)\big)\,\frac{\partial\ell_t}{\partial\epsilon}(0;\boldsymbol{y}^t)\,\check{s}_{ti}(\boldsymbol{y}^t)\,dy_t=E_\psi\left(\frac{\partial\ell_t}{\partial\epsilon}(0;\boldsymbol{y}^t)\,\check{s}_{ti}(\boldsymbol{y}^t)\;\middle|\;\boldsymbol{y}^{t-1}\right)=0.$$

Thus the random variables $\partial\ell_t/\partial\epsilon(0)$ and $\check{s}_{ti}$ have zero covariance at ψ conditional on $\boldsymbol{y}^{t-1}$. By the martingale property (compare (14)), this implies that the unconditional covariance of $n^{-1/2}\sum_{t=1}^n \partial\ell_t/\partial\epsilon(0)$ and $n^{-1/2}\sum_{t=1}^n \check{s}_{ti}$ is zero, and so, letting $n\to\infty$ gives

$$\lim_{n\to\infty} E_\psi\left(n^{-1/2}\sum_{t=1}^n \frac{\partial\ell_t}{\partial\epsilon}(0,\boldsymbol{y}^n)\; n^{1/2}\left(\check{\boldsymbol{\theta}}^n(\boldsymbol{y}^n)-\boldsymbol{\theta}_0\right)\right)=\boldsymbol{0}. \tag{24}$$

Since the left-hand side of this is the limiting covariance of $n^{1/2}(\check{\boldsymbol{\theta}}^n-\boldsymbol{\theta}_0)$ and $n^{-1/2}\sum_{t=1}^n(\partial\ell_t/\partial\epsilon)$ under ψ, the typical element of (24) becomes

$$\langle\psi',\check{s}_i\rangle=0.$$

Since ψ' is an arbitrary element of $T_M(\psi,\boldsymbol{\theta})$, this completes the proof. ▌

Lemma 1 shows that any $\check{s}_i$ can be expressed as the sum of a component in $E(\psi,\boldsymbol{\theta})$ and a component in the orthogonal complement of $T_M(\psi)$ in $T_S(\psi)$. The two terms of this sum are themselves orthogonal. According to Theorem 1, the second term must vanish for an efficient estimator. In fact, the efficient estimator will turn out to be asymptotically unique.

For the next lemma, for each $j=1,\ldots,k$, consider any curve $\psi_j(\epsilon)$ in $\mathbb{M}$ that satisfies the relation

$$\boldsymbol{\theta}\big(\psi_j(\epsilon)\big)=\boldsymbol{\theta}_0+\epsilon\boldsymbol{e}_j, \tag{25}$$

where $\boldsymbol{e}_j$ is a k-vector all the components of which are zero except for component j, which equals one. The existence of such curves is once more guaranteed by the requirement that $\boldsymbol{\theta}$ be a submersion.

Lemma 2

> For any root-n consistent estimator $\check{\boldsymbol{\theta}}$ characterised at the DGP ψ in the parametrised model $(\mathbb{M},\boldsymbol{\theta})$ by the tangents $\check{s}_i$, $i=1,\ldots,k$, and for any curve $\psi_j(\epsilon)$ satisfying (25), the inner product $\langle\psi_j',\check{s}_i\rangle=\delta_{ij}$.

Proof: Suppose as in the proof of Lemma 1 that the $\check{s}_i$ correspond to components $\check{s}_{ti}$ satisfying (16). From (19) and (25) it follows that $\partial\check{s}_{ti}/\partial\epsilon=\delta_{ij}$ along $\psi_j(\epsilon)$. Letting $\psi_j(\epsilon)$ be expressed in terms of contributions $(\ell_j)_t(\epsilon)$, then, by exactly the same arguments as in the proof of Lemma 1, for $i=1,\ldots,k$, we see that

$$E_\psi\left(\frac{\partial(\ell_j)_t}{\partial\epsilon}(0;\boldsymbol{y}^t)\;\check{s}_{ti}(\boldsymbol{y}^t)\;\middle|\;\boldsymbol{y}^{t-1}\right)=\delta_{ij}.$$

This implies that $\langle\psi_j',\check{s}_i\rangle=\delta_{ij}$, as required. ▌

Lemma 3

> At each DGP ψ in the parametrised model $(\mathbb{M},\boldsymbol{\theta})$, there exist unique tangents $\hat{s}_i$, $i=1,\ldots,k$ in the space $E(\psi,\boldsymbol{\theta})$ such that for any root-n consistent estimator $\check{\boldsymbol{\theta}}$ characterised at ψ by the tangents $\check{s}_i$, $i=1,\ldots,k$, $\check{s}_i=\hat{s}_i+v_i$, where v_i belongs to the orthogonal complement of $T_M(\psi)$

in $T_S(\psi)$. Similarly, for all $j = 1, \ldots, k$ and for any curve $\psi_j(\epsilon)$ satisfying (25), there exist unique tangents σ_j in $E(\psi, \boldsymbol{\theta})$ such that $\psi'_j = \sigma_j + w_j$, where w_j belongs to $T_M(\psi, \boldsymbol{\theta})$.

Proof: For any $\check{\boldsymbol{\theta}}$, we know from Lemma 1 that $\check{s}_i$ can be expressed as a sum of a tangent in $E(\psi, \boldsymbol{\theta})$ and some v_i in the orthogonal complement of $T_M(\psi)$ in $T_S(\psi)$. This decomposition is unique, because it is orthogonal. Thus we may choose an arbitrary estimator $\check{\boldsymbol{\theta}}$ and *define* the $\hat{s}_i$ by $\check{s}_i = \hat{s}_i + v_i$. Similarly, for any given set of curves $\psi_j(\epsilon)$ satisfying (25), since the ψ'_j lie in $T_M(\psi)$, we may *define* the tangents σ_j by $\psi'_j = \sigma_j + w_j$, $\sigma_j \in E(\psi, \boldsymbol{\theta})$, $w_j \in T_M(\psi, \boldsymbol{\theta})$. Clearly the σ_j span $E(\psi, \boldsymbol{\theta})$.

By Lemma 2, we have

$$\delta_{ij} = \langle \psi'_j, \check{s}_i \rangle = \langle \sigma_j + w_j, \hat{s}_i + v_i \rangle = \langle \sigma_j, \hat{s}_i \rangle, \tag{26}$$

since v_i, being orthogonal to $T_M(\psi)$, is orthogonal to both σ_j and w_j, and w_j, being orthogonal to $E(\psi, \boldsymbol{\theta})$, is orthogonal to $\hat{s}_i$.

Consider any other root-n consistent estimator characterised by tangents $\tilde{s}_i$ such that $\tilde{s}_i = t_i + u_i$, $t_i \in E(\psi, \boldsymbol{\theta})$, u_i orthogonal to t_i. Then (26) applies to the $\tilde{s}_i$, and so

$$\langle \sigma_j, t_i \rangle = \delta_{ij}.$$

Since the σ_j span $E(\psi, \boldsymbol{\theta})$, and the t_i and the $\hat{s}_i$ belong to $E(\psi, \boldsymbol{\theta})$ and have the same inner products with the basis vectors σ_j, we have $t_i = \hat{s}_i$, and so the $\hat{s}_i$ are unique, as claimed. The uniqueness of the σ_j follows by an exactly similar argument starting from any other set of curves satisfying (25). ∎

Since all the tangents in the above Lemma can be represented by martingales, the results of the Lemma can be expressed in terms of contributions, as follows:

$$\begin{aligned} E_\psi\big(\hat{s}_{ti}(\boldsymbol{y}^t)\, v_{ti}(\boldsymbol{y}^t) \bigm| \boldsymbol{y}^{t-1}\big) &= 0, \text{ and} \\ E_\psi\big(\sigma_{tj}(\boldsymbol{y}^t)\, w_{tj}(\boldsymbol{y}^t) \bigm| \boldsymbol{y}^{t-1}\big) &= 0. \end{aligned}$$

The relations

$$\langle \sigma_j, \hat{s}_i \rangle = \delta_{ij} \tag{27}$$

can be expressed by saying that the σ_j and the $\hat{s}_i$ constitute a pair of *dual bases* for $E(\psi, \boldsymbol{\theta})$. This property also implies that the $k \times k$ matrix with typical element $\langle \hat{s}_i, \hat{s}_j \rangle$ is the inverse of the matrix with typical element $\langle \sigma_i, \sigma_j \rangle$. Since the former matrix is the asymptotic covariance matrix of the estimator $\boldsymbol{\theta}$, the latter can be thought of as performing the role of the asymptotic *information matrix* – in a maximum likelihood model, it would be the asymptotic information matrix in the usual sense. Since the scalar product is a smooth tensor on Hilbert space or a Hilbert manifold, it is seen that the information matrix is smooth in our Hilbert space construction.

The proof of Theorem 1 can now be finished easily. Any estimator $\hat{\boldsymbol{\theta}}$ satisfying the condition of the theorem is characterised by tangents lying in $E(\psi, \boldsymbol{\theta})$, which, by the uniqueness given by Lemma 3, must be the $\hat{s}_i$ of that lemma. Any other estimator $\check{\boldsymbol{\theta}}$ has associated tangents of the form $\hat{s}_i + v_i$. Since all the $\hat{s}_i$ are orthogonal to all the v_i, the asymptotic covariance matrix of $\check{\boldsymbol{\theta}}$ equals the matrix of inner products of the $\hat{s}_i$ plus the matrix of inner products of the v_i. Since all of these matrices are covariance matrices, they are all positive semi-definite, and so the difference between the asymptotic covariance matrix of $\check{\boldsymbol{\theta}}$ and that of $\hat{\boldsymbol{\theta}}$ is positive semi-definite, as required. ∎

4. Examples and Illustrations

As a textbook example, consider the linear regression model (18) with normal errors. Since asymptotic theory is hardly necessary to treat this model, we can consider a finite sample size n. The model can be written in matrix notation as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \tag{18}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $n \times 1$, $\boldsymbol{X}$ is $n \times k$, and $\boldsymbol{\beta}$ is $k \times 1$. We also consider the model (22):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{u}. \tag{22}$$

As we saw, the OLS estimator for (18) is not robust with respect to (22) if $\boldsymbol{Z}^\top\boldsymbol{X}$ is nonzero. However the OLS estimator for (22), restricted to the parameters $\boldsymbol{\beta}$, is consistent, but not efficient, for (18). The OLS estimator from (18) is

$$\hat{\boldsymbol{\beta}} \equiv \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{y},$$

and the estimator of $\boldsymbol{\beta}$ from (22) is

$$\check{\boldsymbol{\beta}} \equiv \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{y}, \tag{28}$$

where $\boldsymbol{M}_Z \equiv \mathbf{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ is the orthogonal projection on to the orthogonal complement of the span of the extra regressors $\boldsymbol{Z}$. It is easy to show that (see, for instance, Davidson and MacKinnon (1993) Chapter 11)

$$\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \big(\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{M}_X\boldsymbol{y}, \tag{29}$$

with $\boldsymbol{M}_X$ defined similarly to $\boldsymbol{M}_Z$.

When (18) is specified with normal errors, the model is finite-dimensional, $k+1$-dimensional in fact, if τ^2 is allowed to vary. Since the loglikelihood of the model is

$$\ell(\boldsymbol{\beta}, \tau^2) = -\frac{n}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2,$$

the tangents to the curves along which just one component of $\boldsymbol{\beta}$ varies are represented by the k-vector of zero-mean random variables

$$\boldsymbol{\sigma} \equiv n^{-1/2}\frac{1}{\tau^2}\boldsymbol{X}^\top\boldsymbol{u}. \tag{30}$$

The only way to vary the DGP without changing the parameter vector $\boldsymbol{\beta}$ is to vary the error variance. Thus the space $T_M(\psi, \boldsymbol{\beta})$ is one-dimensional in this case, and is generated by the tangent represented by

$$n^{-1/2}\frac{\partial \ell}{\partial \tau^2} = n^{-1/2}\frac{1}{2\tau^2}\sum_{t=1}^{n}\left(\frac{u^2}{\tau^2} - 1\right), \tag{31}$$

which has zero covariance with all the components of (30). This means that these components lie in $E(\psi, \boldsymbol{\beta})$, thereby justifying the notation $\boldsymbol{\sigma}$.

The OLS estimator $\hat{\boldsymbol{\beta}}$ is associated with the tangents

$$\hat{\boldsymbol{s}} \equiv \left(n^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1} n^{-1/2}\boldsymbol{X}^\top\boldsymbol{u}, \tag{32}$$

which are seen immediately to be linear combinations of the components of $\boldsymbol{\sigma}$ in (30). The tangents $\hat{\boldsymbol{s}}$ therefore also lie in $E(\psi, \boldsymbol{\beta})$ and so $\hat{\boldsymbol{\beta}}$ is seen to be asymptotically efficient. Note also that the matrix of inner products of the components of $\boldsymbol{\sigma}$ and $\hat{\boldsymbol{s}}$ is the expectation of

$$n^{-1/2}\frac{1}{\tau^2}\boldsymbol{X}^\top\boldsymbol{u}\; n^{-1/2}\boldsymbol{u}^\top\boldsymbol{X}\left(n^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1} = \mathbf{I},$$

confirming the dual basis property (27).

The tangents corresponding to the estimator (28) are seen, from (29), to be

$$\check{\boldsymbol{s}} = \hat{\boldsymbol{s}} + \left(n^{-1}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{X}\right)^{-1} n^{-1/2}\boldsymbol{X}^\top\boldsymbol{M}_Z\boldsymbol{M}_X\boldsymbol{u}.$$

It is simple to check that the covariances of the second term of this with the components of (32) are zero, as also with (31), which represents the tangent that generates $T_M(\psi, \boldsymbol{\beta})$. Thus this second term represents a tangent orthogonal to all of $T_M(\psi)$, as required by the theory of the preceding section.

As a slightly less trivial example, consider again the regression model (18), without imposing the normality of the error terms. The OLS estimator is of course *robust* for any model at all that satisfies the regression equation with zero-mean errors, but it is interesting to enquire under what conditions it is also efficient.

Consider a parametrised model $(\mathbb{M}, \boldsymbol{\beta})$ the DGPs of which satisfy (18), but do not necessarily have normal errors. The OLS estimator is still characterised by the tangents (32), and its robustness implies, by Lemma 1, that these tangents are orthogonal to $T_M(\psi, \boldsymbol{\beta})$ for all $\psi \in \mathbb{M}$. Consequently, the estimator is efficient at a given ψ if $\mathbb{M}$ is large enough to contain the tangents (32) in its tangent space $T_M(\psi)$ at ψ, since then, being orthogonal to $T_M(\psi, \boldsymbol{\beta})$, they must belong to $E(\psi, \boldsymbol{\beta})$. Although it is difficult to state a precise condition that will guarantee this property, intuitively it can be seen that the model must include the case of normal errors.

It can be checked that, if the model specifies the error distribution, up to a scale factor, then *only* normal errors are compatible with the efficiency of the OLS estimator. Suppose that the error density, scaled to have unit variance, is denoted by f. Then the log-density of observation t of the model (18) is

$$\log\left(\frac{1}{\tau} f\left(\frac{y_t - \boldsymbol{X}_t\boldsymbol{\beta}}{\tau}\right)\right).$$

The tangent corresponding to a variation of β_i is then represented by

$$-\frac{1}{\tau} n^{-1/2} \sum_{t=1}^{n} X_{ti} \frac{f'(e_t)}{f(e_t)},$$

where $e_t \equiv (y_t - \boldsymbol{X}_t\boldsymbol{\beta})/\tau$. If the tangents $\hat{\boldsymbol{s}}$ given by (32) are linear combinations of those above, for $i = 1, \ldots, k$, then it is necessary that

$$\frac{f'(e_t)}{f(e_t)} = ce_t, \tag{33}$$

for some constant c independent of t. The general solution to the differential equation (33) is

$$f(e) = C \exp\left(ce^2/2\right),$$

(C another constant) and, since this density must have mean zero and unit variance, it must be the standard normal density, with $C = (2\pi)^{-1/2}$, and $c = -1$.

In general, if one wishes to improve the precision of a parameter estimate, more information of some sort is necessary. Such information may take the form of the true value of some other parameter, or the true nature of the error density, or the like. When models are considered as sets of DGPs, this sort of information corresponds to a *reduction* of the model size, since only DGPs satisfying the constraints imposed by the new information can belong to the model. Then efficiency gains are possible because estimators that would not have been robust with respect to the original model may be so for the reduced model. In some circumstances, though, this is not so, in which case the extra information is uninformative concerning the parameters.

These general considerations can be illustrated geometrically. Consider Figure 1, which represents the space $E(\psi, \boldsymbol{\theta})$ for some parametrised model as a two-dimensional space, with $\boldsymbol{\theta} = [\beta \,\vdots\, \gamma]$.

The origin corresponds to the DGP ψ, at which it is supposed that $\beta = \beta_0$, $\gamma = 0$. The dual bases $\{\hat{s}_\beta, \hat{s}_\gamma\}$ and $\{\sigma_\beta, \sigma_\gamma\}$ are drawn, and it can be seen that $\hat{s}_\beta$ is orthogonal to σ_γ, and $\hat{s}_\gamma$ to σ_β. The tangent σ_γ gives the direction in which only γ varies, and so it is labelled $\beta = \beta_0$. Similarly, σ_β is labelled $\gamma = 0$.

Now suppose that we are provided with the information that $\gamma = 0$. The model must now be restricted to DGPs that satisfy that property. The two-dimensional $E(\psi, \boldsymbol{\theta})$ depicted in the Figure is reduced to the one-dimensional line in the direction of σ_β, that being the direction in which γ remains constant at zero. But $\hat{s}_\beta$ does not belong to the one-dimensional $E(\psi, \beta)$, and so is no longer efficient for the constrained model. The efficient estimator for that model is obtained by projecting $\hat{s}_\beta$ orthogonally on to the direction of σ_β, using the projection denoted

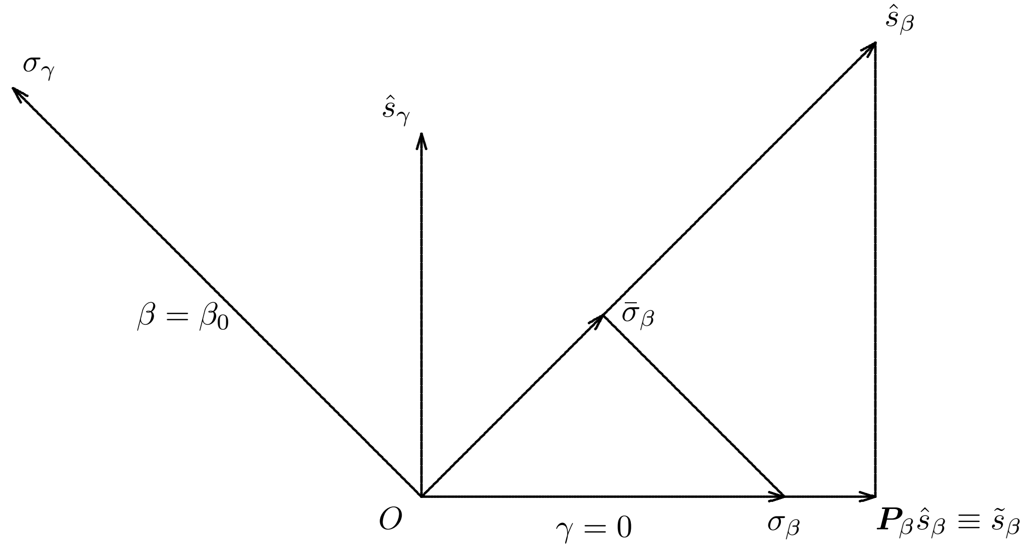

Figure 1

Efficiency gain from added information

by $\boldsymbol{P}_\beta$ in the figure. This gives rise to a new consistent estimator associated with $\tilde{s}_\beta \equiv \boldsymbol{P}_\beta \hat{s}_\beta$. Since $\tilde{s}_\beta$ is obtained from $\hat{s}_\beta$ by an orthogonal projection, it is of smaller norm, or in statistical terms, of smaller asymptotic variance. In addition, the orthogonal projection means that $\tilde{s}_\beta$ has the same inner product with σ_β as does $\hat{s}_\beta$, and so it satisfies the condition of Lemma 2 for a consistent estimator. The result of Lemma 1 is also seen to be satisfied: the inefficient estimator $\hat{s}_\beta$ for the constrained model equals the efficient estimator $\tilde{s}_\beta$ plus something orthogonal to the constrained model.

If γ is a "nuisance" parameter, the value of which is used only to improve the precision of the estimate of β, then it could have been left out of the parameter-defining mapping of the original, unreduced, model. If so, the omission of γ once more leads to a one-dimensional $E(\psi, \beta)$, but this time in the direction of $\hat{s}_\beta$. This is because $E(\psi, \beta)$ must be orthogonal to all directions in which β does not vary, now *including* the direction of σ_γ. This time, it is σ_β which is projected orthogonally on to the direction of $\hat{s}_\beta$ to yield $\bar{\sigma}_\beta$, which replaces σ_β for the model with γ dropped. This orthogonal projection, which means that $\bar{\sigma}_\beta$ has smaller norm than σ_β, corresponds to a reduction in information about β. Notice that the *estimator* $\hat{s}_\beta$ is unchanged whether or not γ is dropped. The information gain moving from $\bar{\sigma}_\beta$ to σ_β is not realised so long as β and γ are estimated jointly, and is realised only when information about γ is available.

If σ_β and σ_γ were orthogonal, so would be $\hat{s}_\beta$ and $\hat{s}_\gamma$, and the directions of σ_β and $\hat{s}_\beta$ would coincide, as would those of σ_γ and $\hat{s}_\gamma$. Redrawing Figure 1 to reflect this state of affairs shows that information about γ no longer leads to any gain in the precision of the estimate of β. This is perfectly intuitive, since the orthogonality means that the asymptotic covariance matrix of the parameter estimates is diagonal.

A simple example of such orthogonality is provided by the linear regression model (18) with normal errors, for which the tangents (30) corresponding to variation of the parameters $\boldsymbol{\beta}$ of the regression function are orthogonal to the tangent (31), which corresponds to variation of τ^2. As is well known, knowledge of the value of the error variance is uninformative about $\boldsymbol{\beta}$. In much the same way, it was seen above that if normal errors are not assumed in (18), then, if it were learnt that the errors were in fact normal, this new information would not lead to any gain as regards estimation of $\boldsymbol{\beta}$, since the OLS estimator remains efficient for normal errors.

5. Estimating Functions and GMM

The generalised method of moments (GMM) was proposed by Hansen (1982) apparently without knowledge of a very similar method proposed by Godambe (1960); see also Godambe and Thompson (1989) and Godambe (1991). It is convenient to refer to Godambe's method as the *method of estimating functions*. Both approaches start from what in the estimating function context are called *elementary zero functions*, which are functions, at least one for each observation of the sample, of both data and parameters. When these functions are evaluated at the correct values for any given DGP, their expectation under that DGP is zero. The simplest example is, as usual, the linear regression model (18), for which the elementary zero functions are the $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$, one for each observation t.

Specifying a set of elementary zero functions is very similar to specifying a model and a parameter-defining mapping. Suppose that, for each observation t, the elementary zero functions are written as $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$, where, as before, the argument $\boldsymbol{y}^t \in \mathbb{R}^{mt}$ corresponds to the observed data in observations 1 through t, and $\boldsymbol{\theta}$ is a k-vector of parameters. The p-vector-valued function $\boldsymbol{f}_t$ will usually depend on explanatory variables (covariates), hence the index t. The natural way to proceed is to specify the model as the set of those DGPs ψ for which there exists a unique parameter vector $\boldsymbol{\theta}$ such that $E_\psi(\boldsymbol{f}_t(Y^t, \boldsymbol{\theta})) = 0$. (Recall that Y^t is the random variable of which observations 1 through t are a realisation.) The parameter-defining mapping then maps ψ to this unique $\boldsymbol{\theta}$.

The above way of defining a parametrised model needs to be qualified somewhat, for a number of reasons. The first is that, in order to perform inference, it is necessary to be able to estimate not only the parameter vector $\boldsymbol{\theta}$, but also the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$, and, for this, one needs the existence of higher moments. It would therefore be preferable to limit the model to those DGPs for which those higher moments exist.

The second reason reveals a difficulty that arises whenever a model, parameter-defining mapping, or estimation method makes use of *moments*, that is expectations. It is that, in any commonly used stochastic topology, including the one used here based on Hilbert space norms, (see Billingsley (1968, 1979)) expectations of unbounded random variables are not continuous functions of the DGP under which they are calculated. For instance, even the smallest admixture of the Cauchy distribution with the normal is enough to destroy the existence of the first moment. The unfortunate consequence is that, if a model is defined by moments, it will not be a smooth submanifold of the overall set of DGPs, $\mathbb{S}$.

The lack of continuity of moments is a problem for establishing appropriate regularity conditions in many contexts, not just the geometrical one. For present purposes, the easiest solution is just to require that the elementary zero functions $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ should be bounded functions of $\boldsymbol{y}^t$. Of course, this assumption excludes most interesting models, even the linear regression model, but, since the emphasis of this paper is geometrical, it does not seem worthwhile to look further for more suitable regularity conditions. In particular, imposing the existence of moments on a model is not informative about that model's parameters. This can be seen by considering a very simple problem, namely that of estimating the mean of a set of scalar i.i.d. observations. If these observations may take values anywhere in an unbounded set, then the set of DGPs defined by requiring that the observations be i.i.d. drawings from a distribution for which the mean exists is not a smooth submanifold of $\mathbb{S}$. However, the set is *dense* in such a submanifold. To see why, consider the Hilbert space $L^2(\mathbb{R})$ in which the unit sphere represents all univariate densities defined on $\mathbb{R}$. Then, for ψ in this unit sphere, the mean of the density to which ψ corresponds, if it exists, is

$$\int_{-\infty}^{\infty} |\psi(y)|^2 y \, dy. \tag{34}$$

The integral above defines an unbounded quadratic operator on $L^2(\mathbb{R})$, the domain of which is dense in $L^2(\mathbb{R})$. In other words, the densities for which the mean exists are dense in the unit sphere of $L^2(\mathbb{R})$. It is straightforward to extend this univariate result to the asymptotic Hilbert space $\mathbb{S}$.

Clearly the model implicitly defined by the problem of estimating the mean is just the set of all i.i.d. sequences, and this set does constitute a smooth submanifold because the requirement that all the observations be i.i.d. can be expressed by the relations $\psi_t = \psi$, for some ψ independent of $t = 1, \ldots$, and these relations are trivially continuously differentiable in the Hilbert space norm of $L^2(\mathbb{R})$. The set of DGPs in this model for which the mean actually exists is a dense set, so that its closure is the full submanifold. However, any information gain that could lead to increased precision of the estimate of the mean must involve, as we saw above, a reduction in the dimension of the model. Since we have seen that imposing a finite mean does not reduce the dimension, no information gain is possible from the knowledge that the mean exists.

Any expectation can be expressed as an integral similar to (34), and can therefore be used to define an unbounded operator on the Hilbert spaces for finite samples. Thus the argument above generalises to all models defined using the expectations of unbounded random variables, and so, for the purposes of geometrical discussions of efficiency and robustness, we must limit ourselves to models defined in terms of the expectations of bounded random variables, for instance, variables obtained by censoring unbounded variables above some suitably high threshold.

Suppose then that we define a parametrised model $(\mathbb{M}, \boldsymbol{\theta})$ by a set of elementary zero functions given by the components of the p-vector $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$, as above. Suppose further that the parameter-defining mapping $\boldsymbol{\theta}$ thus implicitly defined is in fact defined for all $\psi \in \mathbb{M}$, so that the identification condition is satisfied, and that $\boldsymbol{\theta}$ is a submersion, as we required earlier. Consider any $\psi \in \mathbb{M}$ and suppose that

$\boldsymbol{\theta}(\psi) = \boldsymbol{\theta}_0$. Then, for each component f_{ti}, $i = 1, \ldots, p$, of $\boldsymbol{f}_t$, $E_\psi(f_{ti}(Y^t, \boldsymbol{\theta}_0)) = 0$, and so in some circumstances it may be that the sequence

$$n^{-1/2} \sum_{t=1}^{n} \sum_{i=1}^{p} a_{ti}(\boldsymbol{y}^{t-1}) f_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}_0) \tag{35}$$

represents a vector of tangents at ψ, where the $a_{ti}(\boldsymbol{y}^{t-1})$ are predetermined at t, and such that $\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} \sum_{i=1}^{p} E_\psi(a_{ti}^2)$ is finite. This may equally well not be so, because, since only the *unconditional* expectations of the zero functions must vanish, the sequence may not be a martingale, as required by the first equation of (16). For the moment, we suppose that the martingale property is satisfied. Then we have

Lemma 4

> For a parametrised model $(\mathbb{M}, \boldsymbol{\theta})$ defined by a set of elementary zero functions $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ obeying the above regularity conditions and such that, for all $\psi \in \mathbb{M}$, the sequence (35) is a martingale, the tangents represented by the components of (35) are orthogonal to $T_M(\psi, \boldsymbol{\theta})$, the space of tangents at ψ that correspond to curves within the model along which the parameters are constant at $\boldsymbol{\theta}_0$.

Proof: As in the proof of Lemma 1, consider a curve $\psi(\epsilon)$ in $T_M(\psi, \boldsymbol{\theta})$, represented by the log-density contributions $\ell_t(\epsilon)$. Then, as in Lemma 1,

$$\int f_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}_0) \, \frac{\partial \ell_t}{\partial \epsilon}(\epsilon; \boldsymbol{y}^t) \; \exp\bigl(\ell_t(\epsilon; \boldsymbol{y}^t)\bigr) \, dy_t = 0. \tag{36}$$

On multiplying by a_t, the martingale property implies the result. ∎

Lemma 4 is the geometrical expression of the fact that, under suitable regularity conditions, the parameters $\boldsymbol{\theta}$ can be consistently estimated by solving any k linearly and functionally independent equations of the form

$$\sum_{t=1}^{n} \sum_{i=1}^{p} a_{ti}(Y^{t-1}) f_{ti}\bigl(Y^t, \hat{\boldsymbol{\theta}}\bigr) = 0, \tag{37}$$

where the a_{ti}, $t = 1, \ldots, n$, $i = 1, \ldots, p$, are predetermined at t. Standard arguments based on a short Taylor expansion (see for instance Davidson and MacKinnon (1993), Chapter 17) show that, asymptotically, the components of the sequence $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ are linear combinations of the tangents represented by

$$n^{-1/2} \sum_{t=1}^{n} \sum_{i=1}^{p} a_{ti}(\boldsymbol{y}^{t-1}) f_{ti}\bigl(\boldsymbol{y}^t, \boldsymbol{\theta}_0\bigr), \tag{38}$$

provided that a law of large numbers can be applied to the sequences

$$\left\{ \frac{\partial f_{ti}}{\partial \theta_j}(Y^t, \boldsymbol{\theta}_0) \right\}_{t=1}^{\infty}.$$

Although this point will not be developed here, regularity conditions like these are the algebraic counterparts of the geometrical regularity conditions, involving identifiability and the submersion property, discussed above. For further discussion, see Newey and McFadden (1994).

Just as with the tangents $\check{s}_i$ used in Lemma 1, the tangents (38) can be expressed as the sum of two orthogonal components, one in the k-dimensional space $E(\psi, \boldsymbol{\theta})$, and the other orthogonal to the model $\mathbb{M}$. The first component corresponds to the asymptotically efficient estimator, and so, in order to find an efficient estimator, we wish to *project* the tangents (38) orthogonally on to $E(\psi, \boldsymbol{\theta})$. Intuitively, this orthogonal projection is on to the model $\mathbb{M}$ itself, since the tangents are already orthogonal to $T_M(\psi, \boldsymbol{\theta})$, the orthogonal complement of $E(\psi, \boldsymbol{\theta})$ in the tangent space to the model $\mathbb{M}$.

We can perform the orthogonal projection by expressing the unique tangents σ_j, $j = 1, \ldots, k$, defined in Lemma 3, in the form (38). As seen in the proof of that lemma, we can compute the inner product of any tangent of the form (38) with σ_j by considering a curve satisfying (25), since the tangent to such a curve equals σ_j plus a component orthogonal to everything like (38). These inner products are given by the following lemma.

Lemma 5

> For a parametrised model $(\mathbb{M}, \boldsymbol{\theta})$ defined by a set of elementary zero functions $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ obeying the regularity conditions of Lemma 4, the tangent σ_j at DGP $\psi \in \mathbb{M}$ corresponding to component $j = 1, \ldots, k$ of $\boldsymbol{\theta}$ can be represented by the sequence of contributions
>
> $$\sigma_{tj} \equiv \sum_{i=1}^{p} \sigma_{tij} f_{ti}(\boldsymbol{\theta}_0). \tag{39}$$
>
> Further, for all $i = 1, \ldots, p$, $j = 1, \ldots, k$,
>
> $$E_\psi\big(\sigma_{tj} f_{ti}(\boldsymbol{\theta}_0) \mid \boldsymbol{y}^{t-1}\big) = -E_\psi\left(\frac{\partial f_{ti}}{\partial \theta_j}(\boldsymbol{\theta}_0) \mid \boldsymbol{y}^{t-1}\right). \tag{40}$$
>
> If the covariance matrix of $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta}_0)$ under ψ, conditional on $\boldsymbol{y}^{t-1}$, is $\boldsymbol{\varOmega}_t(\boldsymbol{y}^{t-1})$, then
>
> $$\sigma_{tij}(\boldsymbol{y}^{t-1}) = -\sum_{l=1}^{p} \big(\boldsymbol{\varOmega}_t^{-1}(\boldsymbol{y}^{t-1})\big)_{il} E_\psi\left(\frac{\partial f_{tl}}{\partial \theta_j}(\boldsymbol{y}^t) \;\middle|\; \boldsymbol{y}^{t-1}\right). \tag{41}$$

Proof: The first statement of the lemma, (39), simply requires the equations that define the asymptotically efficient estimator to take the general form (37). For this to be false, there would need to exist consistent estimators defined by equations that did not take this form. But the model is defined by the expectations of the elementary zero functions, and expectations of nonlinear functions of these will not in general be zero. Thus a consistent estimator cannot be defined using nonlinear functions of the elementary zero functions, and so (39) is true.

For (40), consider, as in Lemma 2, a curve $\psi_j(\epsilon)$ in $\mathbb{M}$ satisfying (25). Then we have, for $i = 1, \ldots, p$, $j = 1, \ldots, k$, $t = 1, \ldots$, and all admissible ϵ,

$$\int \exp\big((\ell_j)_t(\boldsymbol{y}^t;\epsilon)\big)\, f_{ti}(\boldsymbol{y}^t, \boldsymbol{\theta}_0 + \epsilon \boldsymbol{e}_j)\; dy_t = 0,$$

and so, on differentiating with respect to ϵ, and setting $\epsilon = 0$, we get

$$E_\psi\left(\frac{\partial(\ell_j)_t(0)}{\partial\epsilon} f_{ti}(\boldsymbol{\theta}_0) \;\middle|\; \boldsymbol{y}^{t-1}\right) = -E_\psi\left(\frac{\partial f_{ti}}{\partial\theta_j}(\boldsymbol{\theta}_0) \;\middle|\; \boldsymbol{y}^{t-1}\right).$$

By Lemma 3 and the remark following it, we may replace $\partial(\ell_j)_t(0)/\partial\epsilon$ in the left-hand side above by the contribution σ_{tj}, thus yielding (40).

Substituting the expression for σ_{tj} in (39) into (40) gives

$$\sum_{l=1}^{p} \sigma_{tlj} E_\psi\big(f_{tl}(\boldsymbol{\theta}_0)\, f_{ti}(\boldsymbol{\theta}_0) \mid \boldsymbol{y}^{t-1}\big) = -E_\psi\left(\frac{\partial f_{ti}}{\partial\theta_j}(\boldsymbol{\theta}_0) \;\middle|\; \boldsymbol{y}^{t-1}\right),$$

from which (41) follows, since by definition $(\boldsymbol{\Omega}_t)_{li} = E_\psi(f_{tl}(\boldsymbol{\theta}_0)\, f_{tl}(\boldsymbol{\theta}_0) \mid \boldsymbol{y}^{t-1})$. ▌

The following theorem now follows immediately from Lemma 5.

Theorem 2

Let the parametrised model $(\mathbb{M}, \boldsymbol{\theta})$ be defined by means of the set of bounded elementary zero functions $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ with the restriction that for all $\psi \in \mathbb{M}$, the sequences $f_{ti}(Y^t, \boldsymbol{\theta}_0)$, $i = 1, \ldots, p$, satisfy the conditions of (16) under ψ. Define the sequences of random variables σ_{ti} by (39), with the coefficients $\sigma_{tij}(\boldsymbol{y}^{t-1})$, predetermined at t, given by (41). Then the estimator $\hat{\boldsymbol{\theta}}$ obtained by solving the equations

$$\sum_{t=1}^{n}\sum_{i=1}^{p} \sigma_{tij}(Y^{t-1}) f_{ti}\big(Y^t, \hat{\boldsymbol{\theta}}\big) = 0, \tag{42}$$

for $j = 1, \ldots, k$, is asymptotically efficient for $(\mathbb{M}, \boldsymbol{\theta})$.

Proof: As mentioned above, $\hat{\boldsymbol{\theta}}$ can be expressed asymptotically as a linear combination of the tangents (38) with the a_{ti} replaced by σ_{tij}. By Lemma 5, these tangents belong to $E(\psi, \boldsymbol{\theta})$ for all $\psi \in \mathbb{M}$. By Theorem 1, $\hat{\boldsymbol{\theta}}$ is asymptotically efficient. ▌

The conditions (16) are quite essential for Theorem 2, in particular the martingale condition. However, if the elementary zero functions do not satisfy that condition, it is often possible to find linear combinations, $\boldsymbol{g}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$ of the $\boldsymbol{f}_s(\boldsymbol{y}^s, \boldsymbol{\theta})$, $s = 1, \ldots, t$, that do. The transformation from the $\boldsymbol{f}_t$ to the $\boldsymbol{g}_t$ is analogous to the transformation used to estimate models by GLS rather than OLS.

For instance, suppose that there is just one elementary zero function $f_t(\boldsymbol{\theta})$ per observation (that is, $p = 1$), and denote the covariance matrix of the f_t, for sample size n, by the $n \times n$ matrix $\boldsymbol{V}$. $\boldsymbol{V}$ may depend on $\boldsymbol{\theta}$, and possibly on other parameters as well, such as autocorrelation coefficients. Let $\boldsymbol{\phi}$ denote the complete set of parameters, and let $\boldsymbol{\phi}_0$ be the parameter values for the DGP ψ. In addition, since we are interested in the conditional covariance structure, $\boldsymbol{V}_{ts}$, if $t \geq s$, may depend on Y^{s-1}.

Then if the lower-triangular matrix $\boldsymbol{P}(\boldsymbol{\phi})$ is such that $\boldsymbol{P}^{\top}(\boldsymbol{\phi})\boldsymbol{P}(\boldsymbol{\phi}) = \boldsymbol{V}^{-1}(\boldsymbol{\phi})$, we may form the vector of zero functions $\boldsymbol{g}(\boldsymbol{\phi}) = \boldsymbol{P}(\boldsymbol{\phi})\boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{f}$ is $n \times 1$, with typical element f_t. Note that $\boldsymbol{P}_{ts}(\boldsymbol{\phi})$ is nonzero only if $t \geq s$, and in that case it may depend on $\boldsymbol{y}^{s-1}$. The covariance matrix of $\boldsymbol{g}(\boldsymbol{\phi})$ is then just the identity matrix, and the martingale condition is satisfied.

In order to obtain the optimal estimating equations, we use the relation

$$E_{\psi}\left(\frac{\partial g_t}{\partial \phi_j}(\boldsymbol{y}^t, \boldsymbol{\phi}_0) \,\middle|\, \boldsymbol{y}^{t-1}\right) = \sum_{s=1}^{t} \boldsymbol{P}_{ts}(\boldsymbol{y}^{s-1}, \boldsymbol{\phi}) E_{\psi}\left(\frac{\partial f_s}{\partial \phi_j}(\boldsymbol{y}^s, \boldsymbol{\theta}_0) \,\middle|\, \boldsymbol{y}^{s-1}\right),$$

which holds since

$$\begin{aligned} &E_{\psi}\left(\frac{\partial \boldsymbol{P}_{ts}}{\partial \phi_j}(\boldsymbol{y}^{s-1}, \boldsymbol{\phi}) f_s(\boldsymbol{y}^s, \boldsymbol{\theta}_0) \,\middle|\, \boldsymbol{y}^{s-1}\right) \\ =&\frac{\partial \boldsymbol{P}_{ts}}{\partial \phi_j}(\boldsymbol{y}^{s-1}, \boldsymbol{\phi}) E_{\psi}\big(f_s(\boldsymbol{y}^s, \boldsymbol{\theta}_0) \mid \boldsymbol{y}^{s-1}\big) = 0. \end{aligned}$$

Thus the equations that define the asymptotically efficient estimator are obtained from (42) with g_t in place of f_{ti} (the index i is omitted since we have assumed that $p = 1$), and σ_{t1j} is defined by (41) with $\boldsymbol{g}$ in place of $\boldsymbol{f}$. Since $p = 1$, $\boldsymbol{\Omega}_t$ is just a scalar, equal to one, since the covariance matrix of $\boldsymbol{g}$ is the identity matrix. Putting all this together gives as the estimating equation

$$\sum_{t=1}^{n}\sum_{s=1}^{n} E_{\hat{\psi}}\left(\frac{\partial f_t}{\partial \phi_j}(\hat{\boldsymbol{\phi}}) \,\middle|\, Y^{t-1}\right) \boldsymbol{V}^{-1}(\hat{\boldsymbol{\phi}}) f_s(Y^s, \hat{\boldsymbol{\phi}}) = 0. \tag{43}$$

The notation is intended to indicate that the conditional expectation of $\partial f_t/\partial \phi_j$ must be estimated in some manner – we need not specify details, since many procedures exist. The result (43) is standard in the estimating functions literature, and can be found, for instance, in Godambe (1960) and Godambe and Thompson (1989).

All of the results of this section can be extended quite simply to the sort of model usually found in the context of the generalised method of moments. Such models are still defined in terms of elementary zero functions $\boldsymbol{f}_t(\boldsymbol{y}^t, \boldsymbol{\theta})$, but the requirement that these have zero mean is strengthened so as to require that their means conditional on some set of random variables be zero.

More formally, let $\mathcal{F}_t$, $t = 1, \ldots$, be a nested set of sigma-algebras, with $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$, and such that $Y^t \in \mathcal{F}_t$. Then the condition on the zero functions becomes

$$E_{\psi}\big(f_{ti}(Y^t, \boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}\big) = 0,$$

where as usual $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\psi)$, $t = 1, \ldots$, and $i = 1, \ldots, p$. An equivalent way of expressing the condition is to require that, for all random variables $h_{t-1} \in \mathcal{F}_{t-1}$, the unconditional expectation of $h_{t-1} f_{ti}(\boldsymbol{\theta}_0)$ be zero. Lemma 5 can be applied to all of these new zero functions, and it follows that (40) and (41) are now true with the expectations conditional on $\mathcal{F}_{t-1}$ rather than just on Y^{t-1}.

In addition, Theorem 2 continues to hold with the σ_{tij} defined by the modified (41). This result is most often expressed in terms of the *optimal instruments* for GMM estimation – see for instance Davidson and MacKinnon (1993), Chapter 17.

6. The Linear Regression Model

Despite its simplicity, the linear regression model (18) can be used to illustrate most of the theory of the preceding section. The elementary zero functions, one per observation, are given by

$$f_t(y_t, \boldsymbol{\beta}) = y_t - \boldsymbol{X}_t\boldsymbol{\beta}. \tag{44}$$

In order to use (41), we compute $\partial f_t/\partial\beta_j = -X_{tj}$ Thus, if the f_t are homoskedastic, and provided that $X_{tj} \in \mathcal{F}_{t-1}$, it follows from Theorem 2 that solving the estimating equations

$$\sum_{t=1}^{n} X_{tj}\big(y_t - \boldsymbol{X}_t\boldsymbol{\beta}\big) = 0, \quad j = 1, \ldots, k,$$

yields an asymptotically efficient estimator, namely the OLS estimator, as required by the Gauss-Markov theorem. In case of heteroskedasticity, if $E\big((y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2\big) = \tau_t^2$, the estimating equations are

$$\sum_{t=1}^{n} \frac{1}{\tau_t^2} X_{tj}\big(y_t - \boldsymbol{X}_t\boldsymbol{\beta}\big) = 0,$$

and they yield the Aitken GLS estimator. If the explanatory variables $\boldsymbol{X}_t$ are endogenous and do not belong to $\mathcal{F}_{t-1}$, then the estimating equations are, assuming homoskedasticity,

$$\sum_{t=1}^{n} E\big(X_{tj} \mid \mathcal{F}_{t-1}\big)\,\big(y_t - \boldsymbol{X}_t\boldsymbol{\beta}\big) = 0.$$

In simultaneous equations models, the expectations of endogenous explanatory variables can be expressed in terms of exogenous instrumental variables: the equation above then defines an instrumental variables estimator with optimal instruments.

It is clear that, if the model $\mathbb{M}$ includes heteroskedastic as well as homoskedastic DGPs, then there will be no estimator that is at the same time robust with respect to the whole model and efficient at every DGP in the model, unless there is a way of consistently estimating the variances τ_t^2. This will be the case whenever feasible GLS can be used, but not more generally.

In section 4, it was seen that, in regression models in which the error density is specified, the OLS estimator is efficient only if that density is normal. It is of interest to see if an efficiency gain with respect to OLS can be realised when the density is not normal, but is nonetheless of unknown form. We will now derive an estimator that can be used with homoskedastic errors, is robust against nonnormal error densities, and is more efficient than OLS in some cases of nonnormality. This estimator, which was recently proposed by Im (1996) using arguments somewhat different from those here, can be derived directly using the theory of the preceding section.

We rename the zero function (44) as $u_t(y_t, \boldsymbol{\beta})$, and introduce a new zero function and a new parameter by the relation

$$v_t(y_t, \boldsymbol{\beta}, \tau^2) = u_t^2(y_t, \boldsymbol{\beta}) - \tau^2,$$

where the homoskedasticity assumption is made explicit in terms of the error variance τ^2. It will also be assumed that the v_t are homoskedastic, and that the expectation of $u_t v_t$ does not depend on t. If this last assumption does not hold, the estimator we are about to derive is still robust, but is no longer efficient.

Analogously to (39), tangents that span the efficient space for the present model can be defined by the contributions

$$\begin{aligned} \sigma_{ti} &= a_{ti}u_t + b_{ti}v_t, \quad i = 1, \ldots, k, \text{ and} \\ \sigma_{t\tau} &= a_{t\tau}u_t + b_{t\tau}v_t, \end{aligned} \tag{45}$$

where a_{ti}, b_{ti}, $a_{t\tau}$, and $b_{t\tau}$ are exogenous or predetermined at t. Now by (40) we have

$$\begin{aligned} E(\sigma_{ti}u_t) &= -E\left(\frac{\partial u_t}{\partial \beta_i}\right) = X_{ti}, \\ E(\sigma_{ti}v_t) &= -E\left(\frac{\partial v_t}{\partial \beta_i}\right) = 2E\big(X_{ti}u_t\big) = 0, \\ E(\sigma_{t\tau}u_t) &= -E\left(\frac{\partial u_t}{\partial \tau^2}\right) = 0, \text{ and} \\ E(\sigma_{t\tau}v_t) &= -E\left(\frac{\partial v_t}{\partial \tau^2}\right) = 1. \end{aligned}$$

Thus, on substituting the definitions (45) into the above, we obtain the following equations for the a_{ti}, *etc.*:

$$\begin{aligned} a_{ti}\tau^2 + b_{ti}\gamma &= X_{ti}; \\ a_{ti}\gamma + b_{ti}\kappa &= 0; \\ a_{t\tau}\tau^2 + b_{t\tau}\gamma &= 0; \\ a_{t\tau}\gamma + b_{t\tau}\kappa &= 1, \end{aligned}$$

where $\gamma \equiv E(u_t^3)$ and $\kappa \equiv E(u_t^4) - \tau^4$ are independent of t by assumption, and equal 0 and $2\tau^4$ under normality. Letting $\delta = \tau^2 - \gamma^2/\kappa$, we find that

$$\delta a_{ti} = X_{ti}; \quad \delta b_{ti} = -(\gamma/\kappa)X_{ti}, \quad \delta a_{t\tau} = -\gamma/\kappa, \quad \delta b_{t\tau} = \tau^2/\kappa.$$

Thus, according to Theorem 2, the estimating equations for $\boldsymbol{\beta}$ are

$$\sum_{t=1}^{n} X_{ti}\left(u_t - \frac{\gamma}{\kappa}\big(u_t^2 - \tau^2\big)\right) = 0, \tag{46}$$

If γ is known to be zero, that is, if the error density is not skewed, these equations give the OLS estimator. Otherwise, γ can be consistently estimated by $n^{-1}\sum_{t=1}^{n}\hat{u}_t^3$, and κ by $n^{-1}\sum_{t=1}^{n}\hat{u}_t^4 - \hat{\tau}^4$, where $\hat{u}_t$ and $\hat{\tau}^2$ can be obtained by, for example, OLS.

The procedure suggested by Im (1996) makes use of the artificial regression

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + (\hat{u}_t^2 - \hat{\tau}^2)\theta + \text{residual},$$

where θ is an auxiliary parameter. A little algebra shows that the OLS estimate of $\boldsymbol{\beta}$ from this is the solution of (46). Im shows, both theoretically, and by Monte Carlo simulation, that his estimator, which he calls RALS, for Residuals Augmented Least Squares, is more efficient than OLS when the error terms are skewed. In fact, he goes further, and, by introducing a third zero function

$$w_t(y_t, \boldsymbol{\beta}, \tau^2, \gamma) = u_t^3(y_t, \boldsymbol{\beta}) - 3\tau^2 u_t - \gamma,$$

in which γ, as defined above, becomes an explicit parameter, shows that further efficiency gains relative to OLS are available if the errors have nonnormal kurtosis. The approach of the preceding paragraph can again be used to derive the explicit form of this estimator. As Im points out, estimators of this sort are constructed in the spirit of adaptive estimation – see for instance Newey (1988).

7. Concluding Remarks

In this paper, geometrical characterisations have been given of efficiency and robustness for estimators of model parameters, with special reference to estimators defined by the method of estimating functions and/or the generalised method of moments. It has been shown that, when a parametrised model is considered as a Hilbert manifold in an underlying space of DGPs, the tangent space at any DGP of the model can be expressed as the direct sum of three mutually orthogonal subspaces. Consistent estimators of the model parameters all have the same component in one of these subspaces, which is finite-dimensional, with dimension equal to the number of parameters. This space contains the asymptotically efficient estimator. Inefficient estimators also have a non-vanishing component in the orthogonal complement of the tangent space to the model, and they thus lose efficiency by including random variation in directions excluded by the specification of the model.

Information about parameters is represented geometrically by tangents that lie within the tangent space to the model. There is a unique tangent in the finite-dimensional efficient subspace that corresponds to the variation of each parameter, and the tangent to any curve along which that parameter alone varies is the sum of this unique tangent and a component in the tangent subspace in which the model parameters do not vary. These information tangents form a basis of the efficient subspace that is dual to that provided by the efficient estimators.

Efficient estimating equations for model parameters can be obtained by projecting arbitrary root-n consistent estimators on to the efficient subspace.

Lemma 5 provides a simple method of performing this sort of projection. As seen in the example of the RALS estimator, the projection can often be implemented by an artificial regression. More generally, as shown in Davidson and MacKinnon (1990) in the context of fully parametrised models, artificial regressions can be used in many one-step efficient estimation procedures that are equivalent to projection on to the efficient subspace. The theory of this paper suggests that artificial regressions can be developed to perform such projections in greater generality.

One-step estimators of a seemingly different sort have been proposed recently by Imbens (1997), and it is claimed that their finite-sample properties are substantially better than those of conventional, asymptotically efficient, GMM estimators. Although these estimators are not implemented by artificial regression, they are of course the result of implicit projection on to the efficient subspace. Another asymptotically efficient estimation method with finite-sample properties different from those of GMM has been proposed by Kitamura and Stutzer (1997), based on minimisation of the Kullback-Leibler information criterion. It seems probable that this minimisation is another asymptotically equivalent way of projecting on to the efficient subspace.

It is hoped that the geometrical construction laid out in this paper will serve as a unified framework for the discussion of asymptotic efficiency and robustness.

References

Amari, S. (1990). *Differential-Geometrical Methods in Statistics*, second edition, Lecture Notes in Statistics No. 28, Springer-Verlag, Berlin.

Barndorff-Nielsen, O. E., and D. R. Cox (1994). *Inference and Asymptotics*, Monographs on Statistics and Applied Probability No. 52, Chapman and Hall, London.

Barndorff-Nielsen, O. E., D. R. Cox, and N. Reid (1986). "The Role of Differential Geometry in Statistical Theory," *International Statistical Review*, 54, 83–96.

Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley & Sons, New York.

Billingsley, P. (1979). *Probability and Measure*, John Wiley & Sons, New York.

Chentsov, N. N. (1972). *Statistical Decision Rules and Optimal Inference* (in Russian), Nauka, Moscow. English translation (1982), *Translations of Mathematical Monographs*, Vol. 53, American Mathematical Society, Providence, Rhode Island.

Davidson, R. and J. G. MacKinnon (1987). "Implicit alternatives and the local power of test statistics," *Econometrica*, **55**, 1305–29.

Davidson, R. and J. G. MacKinnon (1990). “Specification tests based on artificial regressions,” *Journal of the American Statistical Association*, 85, 220–27.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, Oxford, New York.

Dawid, A. P. (1975). Discussion of Efron, B. (1975), “Defining the curvature of a statistical problem,” *Annals of Statistics*, 3, 1189–1242.

Dawid, A. P. (1977). “Further comments on a paper by Bradley Efron,” *Annals of Statistics*, 5, 1249.

Godambe, V. P. (1960). “An optimum property of regular maximum likelihood estimation,” *Annals of Mathematical Statistics*, 31, 1208-1212.

Godambe, V. P. (1991). *Estimating Functions*, ed. V. P. Godambe, Clarendon Press, Oxford.

Godambe, V. P., and M. E. Thompson (1989). “An extension of quasi-likelihood estimation,” (with discussion), *Journal of Statistical Planning and Inference*, 22, 137–72.

Hansen, L. P. (1982). “Large sample properties of generalized method of moments estimators,” *Econometrica*, 50, 1029–54.

Im, Kyung So, (1996). “Least Square Approach to Non-Normal Disturbances,” DAE working paper No. 9603, University of Cambridge.

Imbens, G. W. (1997). “One-Step Estimators for Over-Identified Generalized Method of Moments Models,” *Review of Economic Studies*, 64, 359–83.

Kass, R. E. (1989). “The Geometry of Asymptotic Inference (with discussion),” *Statistical Science*, 4, 187–263.

Kitamura, Y., and M. Stutzer (1997). “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica*, 65, 861–74.

Lang, S. (1972). *Differential Manifolds*, Addison-Wesley, Reading, Mass. Reprinted 1985.

Manski, C. F. (1983). “Closest Empirical Distribution Estimation,” *Econometrica*, 51, 305–20.

McLeish, D. L. (1974). “Dependent central limit theorems and invariance principles,” *Annals of Probability*, 2, 620–28.

Murray, M. K., and J. W. Rice (1993). *Differential Geometry and Statistics*, Monographs on Statistics and Applied Probability No. 48, Chapman and Hall, London.

Newey, W. K. (1988). “Adaptive Estimation of Non-Linear Models,” *Journal of Econometrics*, 38, 301–39.

Newey, W. K., and D. McFadden (1994). “Estimation in Large Samples,” in *Handbook of Econometrics*, Vol. 4, eds. D. McFadden and R. F. Engle, North Holland, Amsterdam.

Small, C. G., and D. L. McLeish, (1994). *Hilbert Space Methods in Probability and Statistical Inference*, Wiley-Interscience, New York.

White, H. (1985). *Asymptotic Theory for Econometricians*, Academic Press, New York.

White, H., and I. Domowitz (1984). “Nonlinear regression with dependent observations,” *Econometrica*, 52, 143–61.