# The Geometry of the Wald Test

Russell Davidson *

G.R.E.Q.E. – E.H.E.S.S.
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille
France

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

* I am grateful to Mark Salmon for interesting conversations on topics related to that of this paper, and to seminar participants at Queen's University and the University of British Columbia for comments. This research was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada, gratefully acknowledged.

July, 1990

**Abstract**

The issue of the non-invariance of the Wald test under nonlinear reparametrisations of the restrictions under test is studied from a differential geometric viewpoint. Quantities that can be defined in purely geometrical terms are by construction invariant under reparametrisation, and various attempts are made to construct a Wald test out of such invariant quantities only. Despite the existence of a wide variety of possibilities, no computationally convenient invariant test statistic emerges from the analysis, since all the statistics considered need calculations equivalent in difficulty to the estimation of the restricted model, contrary to the spirit of the Wald test. On the other hand a variant of a $C(\alpha)$ test is discussed which, while not completely invariant under reparametrisation, is very nearly so, at least in the context of the model discussed by Gregory and Veall (1985), for which the conventional Wald test is particularly badly behaved. This test is easily computed from estimates of the unrestricted model only, and Monte Carlo evidence supports the conclusion that it yields as reliable inference as any other classical test even in very small samples.

**Key words and phrases:** Wald test, $C(\alpha)$ test, nonlinear restrictions, invariance under reparametrisation, differential geometry, geodesics.

# 1 Introduction.

Notions of curvature in a statistical context, with associated ideas from differential geometry, were introduced in a seminal paper of Efron (1975). Subsequently, Amari, in a series of papers, and then in a monograph (1985), developed these ideas very considerably. His work was inspired by earlier very abstract work by Chentsov (1972), in which had been established the existence of a family of *connections* on what are now called *statistical manifolds*. These manifolds corresponded to statistical models in the exponential family.

Davidson and MacKinnon (1987) introduced infinite-dimensional statistical manifolds, with a Hilbert manifold structure, in a manner similar to some early sketches of such manifolds by Dawid (1975, 1977) in two much neglected papers. The infinite-dimensional structure avoided the need to limit attention to models in the exponential family, and strongly supported an interpretation of statistical manifolds as being made up of *data-generating processes*, or DGP's, and of statistical or econometric models as being *sets* of DGP's, of either finite or infinite dimension.

A recent survey of the finite-dimensional work is provided by Barndorff-Nielsen, Cox, and Reid (1986), and a paper by Kass (1981) is useful for a rather different approach, more mathematically rather than statistically oriented.

More recently still, in some as yet largely unpublished work, Critchley, Marriott, and Salmon (1990) have tried to use the ideas of differential geometry in order to understand the reasons for the non-invariance of the Wald test under reparametrisation of the restrictions under test, and to propose new versions of the Wald test that are parametrisation-independent and that may be hoped to provide reasonably reliable inference in finite samples. The motivation for their study was the work of Gregory and Veall (1985) showing that the non-invariance of the Wald test could lead to devastatingly different inferences in certain quite simple examples. Despite high-powered follow-ups of this work by Lafontaine and White (1986) and Phillips and Park (1988), there remain a number of issues, both conceptual and practical, that remain to be resolved before any version of the Wald test can be reliably used in practice other than in linear regression models (including linear simultaneous equations models).

In this paper, I try to resolve some of these issues, in an infinite-dimensional context. Since the classical hypothesis tests, including the Wald test, are all asymptotically equivalent under local alternatives or local DGP's, the analysis is necessarily non-local. By this I mean that it is not possible to linearise everything, and treat all test statistics as though they were defined simply in the tangent space at the DGP supposed to have generated the data. I do not mean, however, that the DGP is "distant" from the model being tested. The issue is one of the *curvature* of the manifold (something that cannot be captured by an analysis restricted to a linear tangent space), in the vicinity of a given DGP.

In the next section, there is a rapid review of the essential concepts of differential geometry that will be used in an essential way in the paper. Then, in section 3, the invariance question will be studied from a geometrical standpoint. It will be shown that

a useful way to look at invariance properties is to consider as invariant such quantities as can be defined purely geometrically, without regard to any particular parametrisation chosen to characterise a model. Section 4 contains a number of suggestions for a Wald test based purely on geometric quantities, and therefore invariant under reparametrisations by construction. In one sense, the answers are not very satisfactory: the advantage of the Wald test is that it requires estimation only under the *unrestricted* model, but no invariant test exists that does not also require, at least implicitly, estimation of the model with the restrictions imposed. Invariant test statistics can indeed be defined that respect the Wald *principle* of being based on estimates of the unrestricted model, but their implementation requires computations tantamount to the estimation of the restricted model. But it seems plausible, according to the analysis presented in Section 4, that there should exist tests that, although not strictly invariant, are easy to compute once the unrestricted model has been estimated and are not so much affected by arbitrary choices as is the conventional Wald test. In Section 5, examples are discussed that make the general issues clear. The Gregory-Veall example is covered as one case of the models considered. The results of a small Monte Carlo experiment suggest that the almost invariant, easy-to-compute, tests proposed in the preceding section do indeed have desirable properties and can be recommended to use in practice. Finally, Section 6 sets forth some conclusions suggested by the analysis of the paper.

## 2 Some Background on Differential Geometry.

A *manifold* is a set endowed with a topological structure that makes it look like a linear space locally. Most statistical manifolds studied up till now have been finite-dimensional, so that they were locally like Euclidean space. The statistical manifold of Dawid (1975, 1977) and Davidson and MacKinnon (1987) is infinite dimensional, and it looks locally like Hilbert space, which is the most natural generalisation of Euclidean space to an infinite number of dimensions. A manifold is said to be *modelled* on Euclidean space or Hilbert space if it is locally isomorphic to one of these. Both Euclidean space and Hilbert space support the operation of differentiation, and so one may speak of a *smooth* manifold if the differentiable properties of these spaces, which can always be used *locally*, are inherited by the manifold *globally*. These differentiable properties may be inherited only to some extent, and so one may deal with $C^k$ manifolds, on which "smooth" functions may be continuously differentiated $k$ times, all the way to $C^\infty$ manifolds. For simplicity, we will work with a $C^\infty$ structure: functions or other objects defined on the manifold will be *smooth* if and only if they can be differentiated infinitely often.

The elements of the Hilbert manifold we shall consider are vectors of random variables, characterised by their joint density, which is assumed to exist with respect to some carrier measure defined on the space (usually just $\mathbb{R}$) in which the components of the random vector are realised. The random vectors are thought of as the observations of a sample, and so a sample of size $n$ corresponds to a random vector of $n$ components, and to a joint density of $n$ scalar random variables. A joint density can be written as follows as a function

of $n$ variables:

$$L(y^n) \equiv L(y_1, \ldots, y_n),$$

where a superscript index denotes a *sample*, with $n$ elements, and a subscript index refers to one of those elements. Normalisation implies that

$$\int dy^n L(y^n) = 1.$$

A *Hilbert* space structure is generated, not by functions, like $L$ above, that are integrable, but rather by square-integrable functions. This observation leads us to consider functions whose squares are density functions:

$$|\psi(y^n)|^2 \equiv L(y^n). \tag{1}$$

In this process we have lost the *uniqueness* of our representation of joint densities of $n$ variables, since $\psi$ is defined only up to a sign (or, more generally, a complex factor located on the unit circle of the complex plane).[2] It turns out that this does not give rise to any untoward consequences in the subsequent development of the theory. Thus, since functions $\psi$ defined as in ((1)) are square integrable to unity, we are interested in the *unit sphere* of the Hilbert space of square-integrable functions of $n$ real variables, $L^2(\mathbb{R}^n)$. We shall denote this unit sphere as $\mathbb{S}(\mathbb{R}^n)$.

It is easy to see that the unit sphere of a Hilbert space is a Hilbert manifold, but a *curved* Hilbert manifold – just as the unit sphere in 3-space is a curved two-dimensional manifold. One way to see this is to imagine the *tangent space* at any point on a sphere (of arbitrary dimension), and note that, although this space is locally isomorphic to the surface of the sphere to which it is tangent, it does not coincide with it, even locally, and, globally, it is quite different: see Figure 1.

The idea mentioned above of a *tangent space* is of cardinal importance in the analysis of manifolds. A tangent space can be fruitfully thought of as providing a *local linear approximation* to the manifold, while abstracting from all global properties, such as the topological difference between a linear space and the surface of a sphere. Formally, a tangent space at a point $m$ on a manifold $\mathbb{M}$ is defined as the set of tangents to *curves* passing through that point. Such a curve can be defined as a mapping $c : ]-1, 1[ \to \mathbb{M}$ such that $c(0) = m$. Since on smooth manifolds differentiation is permitted, we can characterise the *tangent* to the curve $c$ as the derivative $c'(0)$ of $c$ evaluated at 0. Then, by a procedure of abstraction familiar to mathematicians, a tangent at $m$ is identified with an equivalence class of curves through $m$, all with the same derivative at $m$.[3] See Figure 2.

[2] In quantum mechanics, a similar state of affairs is found. I have found no useful analogies between the physical uncertainty of Heisenberg's uncertainty principle and the statistical uncertainty that arises from the use of random variables, and so I can give no useful interpretation to such a complex factor. The fact that an identical mathematical structure appears naturally in two very different contexts is nonetheless thought-provoking.

[3] The best exposition of the general theory of differentiable manifolds that I know of is Lang (1972).

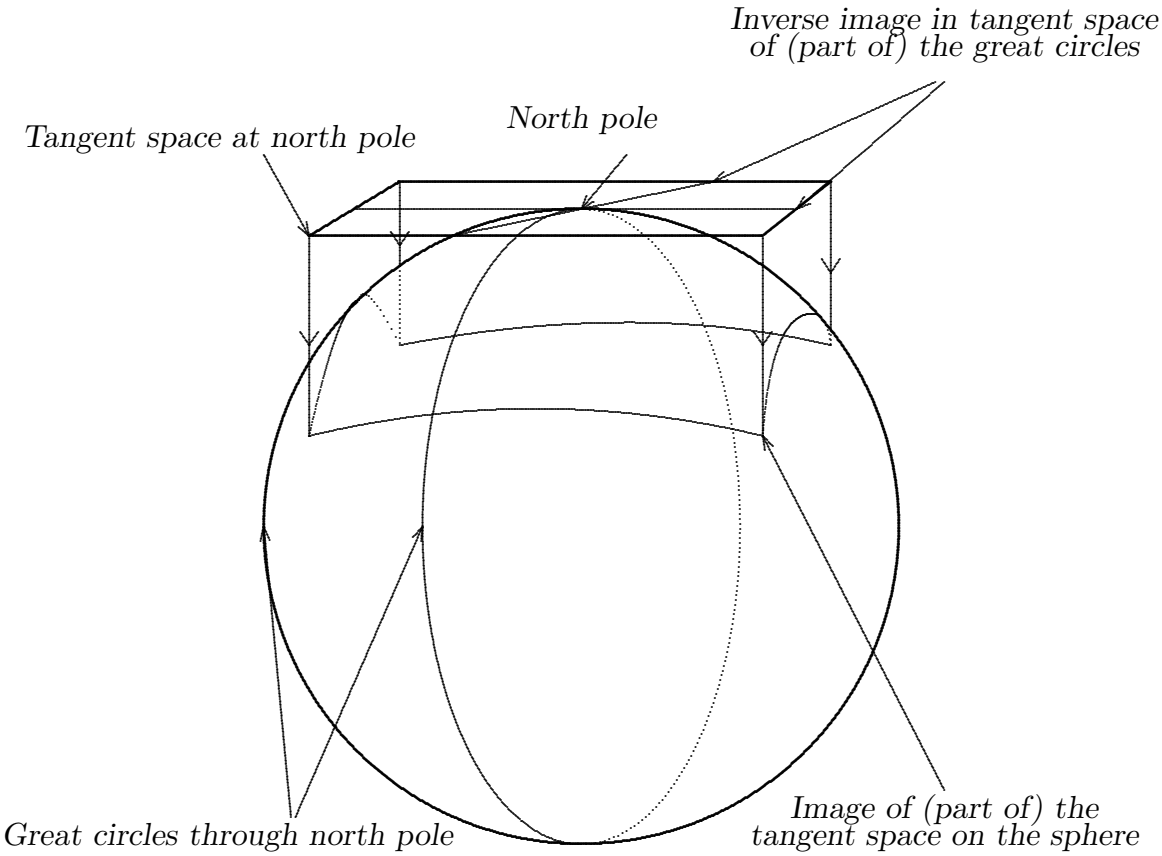


**Figure 1**

**The tangent space to a sphere.**

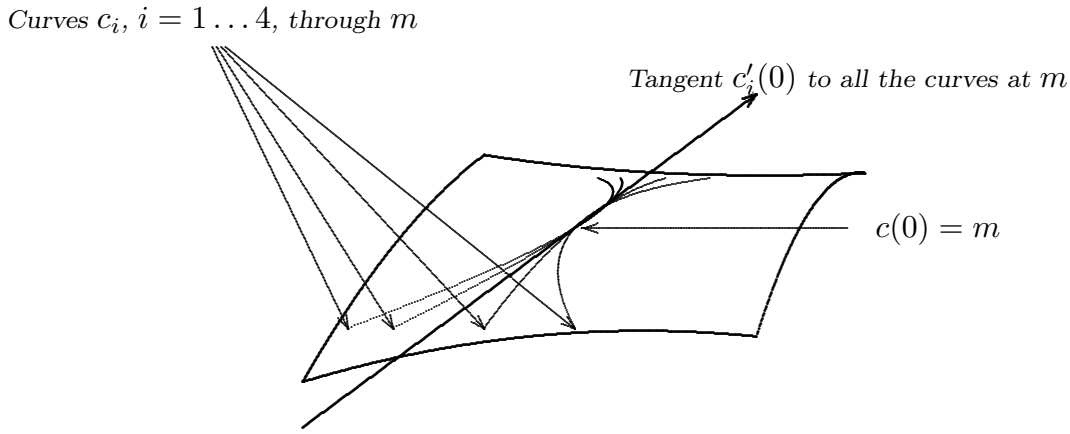


**Figure 1**

**The tangent space to a sphere.**

The set of all tangents to $\mathbb{M}$ at $m$ is the tangent space at $m$, denoted by $T\mathbb{M}_m$. This space can be seen to be locally isomorphic to the manifold in the neighbourhood of the point of tangency, $m$, and so also, therefore, to the space, Euclidean or Hilbert, on which the manifold is modelled. The tangent space thus naturally inherits a Euclidean or Hilbert space, (*i.e.*, linear) structure.

Multilinear functions defined on the set of all tangent spaces to a manifold are called *forms*. On the tangent space $T\mathbb{M}_m$ at each $m \in \mathbb{M}$, an *n-form* maps from an $n$-fold product of $T\mathbb{M}_m$ with itself into the real line $\mathbb{R}$. Of particular interest is the *metric* 2*-form*, by means of which the scalar, or inner, product of Euclidean or Hilbert space is defined. If this form is denoted as $\Omega : T\mathbb{M}_m \otimes T\mathbb{M}_m \to \mathbb{R}$, then for any two tangents $t_1, t_2 \in T\mathbb{M}_m$, the scalar product $\langle t_1, t_2 \rangle$ of $t_1$ and $t_2$ is defined as $\Omega(t_1, t_2)$. In order to be a valid inner product, $\Omega$ must be *positive definite*, in the sense that, for all nonzero $t \in T\mathbb{M}_m$, $\Omega(t, t) > 0$. Norms of tangents are defined as usual: $\|t\|^2 = \langle t, t \rangle$, and it is this use of $\Omega$ that gives it the name *metric*.

We now consider the statistical interpretation of curves and tangents to the Hilbert sphere $\mathbb{S}(\mathbb{R}^n)$. Since elements of $\mathbb{S}(\mathbb{R}^n)$ are constructed as joint densities of $n$ random variables, we shall, as in Davidson and MacKinnon (1987), interpret them as DGP's. A curve $c$ in $\mathbb{S}(\mathbb{R}^n)$ is therefore a *one-parameter family* of DGP's. If by a *model* we mean a set of DGP's, then $c$ can be thought of as simply related to the loglikelihood function for the model defined as the image of $c$: the loglikelihood for any value $\epsilon$ of the argument of $c$ is in fact $2 \log c(\epsilon)$, since $c(\epsilon)$ itself is the square root of the density associated by $c$ to the parameter $\epsilon$. The point $c(0)$ of $\mathbb{S}(\mathbb{R}^n)$ is then some DGP contained in the one-parameter model. The tangent to $c$ at $c(0)$ can be represented as the derivative of the curve at $\epsilon = 0$. It is therefore related to the derivative of the loglikelihood function at $\epsilon = 0$, that is, the score vector for the model. If we denote the loglikehood as $\ell(\epsilon)$, we have

$$c'(0) = \frac{\partial}{\partial \epsilon} c(\epsilon)\big|_{\epsilon=0} = \frac{1}{2} c(0) \frac{\partial}{\partial \epsilon} \ell(\epsilon)\big|_{\epsilon=0}. \tag{2}$$

Why is it necessary to define the tangent to $c$ as $c'(0)$ rather than as $\ell'(0)$, which is a much more familiar construct? The answer is that the Hilbert space structure necessarily had to be defined in terms of square-root densities – loglikelihoods are not necessarily square integrable – and so, if we wish to make any use of the Hilbert manifold structure, we must, to begin with at least, work with square root densities, which are square integrable. Nevertheless, the statistical interpretation of a tangent *is* that of a score vector: it is simply that the Hilbert space operations are more naturally defined on the square root density, *i.e.*, $\exp\left(\frac{1}{2}\ell(\epsilon)\right)$.

Consider next the Hilbert space operation of inner product acting on two tangents $t_1$ and $t_2$ at a DGP $m$. Since each DGP $c(\epsilon)$ in a curve can be represented by a square-root density, there is for each curve $c$ a function $\psi : \mathbb{R}^n \otimes ]-1, 1[ \to \mathbb{R}$ such that $\psi(y^n, \epsilon)$ is the square-root density associated with $\epsilon$ evaluated at $y^n$. Let $\psi_1$ and $\psi_2$ be the functions associated with the curves defining the tangents $t_1$ and $t_2$ respectively. If we use $\epsilon_1$ and $\epsilon_2$

for the parameters of the two curves, then by the usual definition of an inner product in $L^2(\mathbb{R}^n)$, we have

$$\langle t_1, t_2 \rangle = \int \frac{\partial \psi_1}{\partial \epsilon_1}(y^n, \epsilon_1 = 0) \frac{\partial \psi_2}{\partial \epsilon_2}(y^n, \epsilon_2 = 0) dy^n. \tag{3}$$

Using the relation between square-root densities and log densities and their derivatives, ((2)), and denoting by $\ell_i$, $i = 1, 2$, the loglikelihood functions of the two one-parameter models, ((3)) becomes

$$\frac{1}{4} \int \psi_1(y^n, 0) \psi_2(y^n, 0) \frac{\partial \ell_1}{\partial \epsilon_1}(y^n, 0) \frac{\partial \ell_2}{\partial \epsilon_2}(y^n, 0) dy^n. \tag{4}$$

Now both $\psi_1(0)$ and $\psi_2(0)$ are the square root of the density associated with the DGP $m$, and so the product $\psi_1(0)\psi_2(0)$ is just that density. Consequently 4 times ((4)) is the *expectation*, calculated under DGP $m$, of

$$\frac{\partial \ell_1}{\partial \epsilon_1}(y^n, 0) \frac{\partial \ell_2}{\partial \epsilon_2}(y^n, 0),$$

that is, the *covariance* under $m$ of the two zero-mean (under $m$) random variables $(\partial \ell_i / \partial \epsilon_i)(y^n, 0)$, $i = 1, 2$.

This is the interpretation of the Hilbert space inner product we were seeking. A tangent at a DGP $m \in \mathbb{M}$ is a random variable of the form $\partial \psi / \partial \epsilon$ (a random variable since it is function of the random sample $y^n$), that is related to another random variable, of the form $\partial \ell / \partial \epsilon$, that has mean zero under $m$, because it is just a score vector. Conversely it can be shown that any random variable (a function of $y^n$) with mean zero under $m$ corresponds to a tangent in $T\mathbb{M}_m$. The inner product of any two tangents in $T\mathbb{M}_m$ is, up to a factor of four, the covariance of the zero-mean random variables associated with the tangents.

There is of course no reason for which the Hilbert manifold structure could not have been defined directly on DGP's represented by their log densities rather than their square-root densities, with the (slightly) nonstandard definition of the inner product. But this procedure would give rise to a manifold with a different differentiable structure from that on $\mathbb{S}(\mathbb{R}^n)$. The manifolds themselves, considered simply as point sets, are different, since $\mathbb{S}(\mathbb{R}^n)$ can be mapped many to one into the set of densities defined on $\mathbb{R}^n$, which are however in one-one correspondence with the log-densities defined on $\mathbb{R}^n$. A more important difference is that, in the case of $\mathbb{S}(\mathbb{R}^n)$, the metric 2-form, defined as the usual Hilbert space inner product, is by construction a smooth (differentiable infinitely often) form, whereas the metric defined by the covariance operation on zero-mean random variables is not smooth. This point is obvious as soon as one notes that there are zero-mean random variables, for which the first moment necessarily exists, that have no second moment. The metric evaluated at such random variables blows up, and thus cannot be smooth. It is not hard to see that *any* representation of DGP's other than *via* their square-root densities, or a $C^\infty$ transform of their square-root densities, will suffer from this problem,

so that, in this sense, the square-root representation is the *unique* representation for which the covariance inner product is a smooth form. We shall see that the metric 2-form defined by covariance is natural in yet another way on statistical manifolds, namely as the appropriate infinite-dimensional generalisation of the information matrix, and so there are good reasons for retaining the somewhat unintuitive, but mathematically natural, square-root representation, despite the fact that DGP's are not represented uniquely in it.

Whenever a manifold is curved, there is no obvious way of relating the tangent space to the manifold at one point, $m_1$ say, to that at another point $m_2$. A *flat* manifold (*i.e.*, an open subset of a *linear* space, finite- or infinite-dimensional) can be thought of as being contained in the tangent space to it at any point. The tangent spaces are thus all just the same linear space with the origin translated to different points of the manifold, which is just a subset of the one linear space. But for a curved manifold, like $\mathbb{S}(\mathbb{R}^n)$, no such simple procedure is possible. It is nonetheless necessary for many purposes to be able to relate tangent spaces at different points of a manifold. In order for this to be possible, additional structure is needed. The appropriate structure is called a *connection*, or more formally, an *affine connection*, on the manifold.

It is not necessary for this paper to give a formal definition of an affine connection. (See Barndorff-Nielsen, Cox, and Reid (1986) for a more detailed discussion of the concept in the statistical context.) What a connection does is to permit the procedure called *parallel translation* of tangents from one tangent space to another. Thus to each tangent in $T\mathbb{M}_{m_1}$, say, there corresponds by means of the connection a tangent in $T\mathbb{M}_{m_2}$, where $m_1$ and $m_2$ are distinct points of the manifold. The parallel translation is accomplished by solving a set of differential equations, which cause a tangent to be smoothly transported in any desired direction to tangent spaces at neighbouring points on the manifold. A direction of particular interest is that defined by the tangent that is to be transported itself. Once a tangent is chosen, then, it can be followed as it "follows its nose" through different points of the manifold, thereby generating a curve. Such a curve, obtained by choosing a starting direction and following that direction and its successive parallel translations, is called a *geodesic*. The best known examples of geodesics are straight lines in linear spaces, and great circles on the surfaces of spheres.

As remarked in the introduction, at least for finite-dimensional manifolds of DGP's contained in the exponential family, there is a one-parameter family of connections, all of which seem to be natural on these manifolds. In this paper, we shall make use of all of them, and a simple way will be presented that demonstrates the sense in which they are natural.

There is one connection that is more natural than the others, namely the so-called *metric connection*. It can be shown that on *Riemannian manifolds*, that is, smooth manifolds on which a positive definite metric 2-form is defined, there exists a unique connection compatible with the metric. By "compatible", it is meant here that the geodesics generated by the connection should have the property that the *length* of a geodesic between any two points on it should be stationary with respect to perturbations of the curve passing through the two points. Usually this means simply that the shortest distance between two points on a curved manifold is the length of a geodesic in the metric connection.

It is probably not surprising that the metric connection on statistical manifolds turns out to the "natural" connection for the square-root manifold $\mathbb{S}(\mathbb{R}^n)$. Denote by $\mathbb{P}^\delta(\mathbb{R}^n)$ the statistical manifold defined by means of the fractional power $\delta$ of the density, where $\delta$ is not in general equal to one-half. Of course $\mathbb{P}^\delta(\mathbb{R}^n)$ is not a *Hilbert* manifold for $\delta \neq \frac{1}{2}$. For the limiting case of $\delta = 0$, the manifold will be defined by the log of the density. The manifolds $\mathbb{P}^\delta(\mathbb{R}^n)$ have natural connections, as we shall see later, but they are different from the metric connection. In fact varying the fractional power to which a density is raised generates representations for which the natural connections are precisely the members of the one-parameter family of connections alluded to above.

To conclude this preliminary section, we give the definition of the length of a curve on a manifold. This is necessary if we wish to verify whether any given connection is the metric connection. Consider a curve $c : [0, 1] \to \mathbb{M}$. A tangent can be defined for each $\epsilon \in [0, 1]$, call it $c'(\epsilon)$, although these tangents belong to different spaces. Each tangent has a norm, however, that we can denote $\|c'(\epsilon)\|$, which is a positive real number. The *length* of the curve $c$ is then, by definition:

$$\int_0^1 \|c'(\epsilon)\| d\epsilon. \tag{5}$$

It is important to note that the length of a curve is a property of that curve as a point set in the manifold, that is, it is independent of the particular parametrisation used to generate the curve. If $f$ denotes a monotone transformation mapping the interval $[1, 0]$ to the interval $[f(0), f(1)]$, the curve $\tilde{c} \equiv c \circ f^{-1}$ which maps $[f(0), f(1)]$ to the same image in $\mathbb{M}$ as does $c$, can be seen, by simple calculation, to have the same length as $c$. A parametrisation of any curve considered as a point set which has the property that $c'(\epsilon)$ is constant for all $\epsilon$ in its range of definition defines what is called an *arc-length* parameter for the curve. Such parametrisations will be of use to us later.

## 3 Parametrisation Invariance as a Geometrical Property.

In the previous section, a loglikelihood function was always a random variable, a function of $y^n$. It was a member of a statistical manifold, $\mathbb{S}(\mathbb{R}^n)$ for preference. If the random variable $y^n$ is realised as a draw from some DGP $m$, then we have observed a random sample *generated* by $m$. To avoid possible confusion, the realised sample will be denoted as $Y^n$. This sample now defines a function on the statistical manifold $\mathbb{S}(\mathbb{R}^n)$, and in fact on the whole class of statistical manifolds $\mathbb{P}^\delta(\mathbb{R}^n)$, by evaluation. Denote the function by $Y : \mathbb{P}^\delta(\mathbb{R}^n) \to \mathbb{R}$. Then, if $\ell$ is the log density associated to the DGP $m$, $Y$ is defined by

$$Y(m) = \ell(Y^n). \tag{6}$$

By a slight abuse of notation, we use one symbol $Y$ for the mappings from any of the manifolds $\mathbb{P}^\delta(\mathbb{R}^n)$. Observe that in general $Y$ is not *bounded* on any $\mathbb{P}^\delta(\mathbb{R}^n)$. The mapping(s) $Y$ exists independently of any econometric *model*. But if we now choose to define a model $\mathbb{M}$

as a subset of $\mathbb{S}(\mathbb{R}^n)$ of finite dimension $k$, and further to parametrise the model by means of a one-to-one parametrisation $\lambda$ that maps a parameter space $\Theta \subset \mathbb{R}^k$ to $\mathbb{M}$, such that for all $\theta \in \Theta$, there is one and only one $m(\theta) \in \mathbb{M}$, then the mapping $l_Y \equiv Y \circ \lambda : \Theta \to \mathbb{R}$ is precisely the loglikelihood function for this parametrised model.

In this section, we try to leave the last step of the above analysis out to the extent that we can. That is, we consider models as subsets of $\mathbb{S}(\mathbb{R}^n)$ *without* any particular parametrisation. In this way, we will readily see what aspects of what we are now about to study, namely hypothesis testing, are parametrisation-independent, and what are not. The term *intrinsic* is often used to describe quantities or procedures that can be defined on a manifold in a way that is independent of any particular representation or parametrisation of the manifold.

For simplicity, we remain in the classical scheme of hypothesis testing. We begin from a maintained or *alternative* hypothesis, or model, denoted as $\mathbb{M}_1 \in \mathbb{S}(\mathbb{R}^n)$. The null hypothesis, or model, is a subset of $\mathbb{M}_1$, denoted as $\mathbb{M}_0$. It is of strictly lower dimension than $\mathbb{M}_1$. In the classical approach, only DGP's in $\mathbb{M}_1$ are ever considered, although this restriction is entirely unnecessary – see Davidson and MacKinnon (1987).

Maximum likelihood estimation of either of the models $\mathbb{M}_0$ or $\mathbb{M}_1$ can be defined in a parameter-free way, as follows. The mapping $Y$ restricted to a model $\mathbb{M}_i, i = 0, 1$ maps from that (finite-dimensional) model to $\mathbb{R}$. Let us denote the restricted mappings by $Y_i, i = 0, 1$. The DGP $\hat{m}_i$ which maximises the value of $Y_i$ over $\mathbb{M}_i$ will be defined as the *maximum likelihood estimate* for model $i$ and for the realised sample $Y^n$. Thus we consider estimation as a procedure for estimating a DGP rather than a vector of parameters. In the maximum likelihood context, this clearly does no violence to our usual intuition, and we have clearly found a parameter-free way of treating the matter. The maximised loglikelihood is now defined straightforwardly as $Y(\hat{m}_i)$ for model $i$, using the mapping $Y$ defined in ((6)).

We may proceed directly to define the Likelihood Ratio (LR) test of $\mathbb{M}_0$ against $\mathbb{M}_1$. For the sample $Y^n$ the value of the test statistic is

$$LR = 2\left(Y(\hat{m}_1) - Y(\hat{m}_0)\right).$$

The above definition is parametrisation independent, but it clearly coincides with conventional definitions of the test statistic based on parametrised models.

The Lagrange Multiplier (LM) test statistic can also be defined in an intrinsic fashion, with a little more work. Associated to any smooth real-valued function $f$ defined on a manifold $\mathbb{M}$ there is a 1-form denoted by *df* and called the *exterior derivative* of $f$. At any point $m \in \mathbb{M}$, the form *df* is defined by specifying its value for any tangent at $m$, $t \in T\mathbb{M}_m$. The definition is made in terms of a curve $c$ through $m$ such that its tangent at $m$ is $t$, as follows:

$$df(t) = \frac{d}{d\epsilon} f\left(c(\epsilon)\right)|_{\epsilon=0}. \tag{7}$$

It can be shown that the 1-form is well defined by ((7)), that is, the definition is independent of the choice of the curve $c$ provided its tangent at $m$ is $t$. Manifolds modelled on Euclidean

or Hilbert space have the property that any 1-form, that is, a set of linear functions defined on all of the tangent spaces to the manifold, is in one-one correspondence with a field of tangents defined on the manifold. By this is meant that, just as the form defines a linear function on each tangent space, the field of tangents defines a particular element of each tangent space. This property is expressed by saying that Euclidean and Hilbert spaces are *self-dual*. It means for present purposes that the action of the 1-form $df$ at $m$ on a tangent $t$ at $m$ is given by the operation of taking the inner product of $t$ with the tangent, $\nabla_m f$ say, that corresponds to $df$ at $m$:

$$df(t) = \langle \nabla_m f, t \rangle.$$

As the notation suggests, we call $\nabla_m f$ the *gradient* of the function $f$ at $m$.

We will need the gradient of the mapping $Y$ in order to construct the LM test. However it turns out to be impossible to define this gradient on the entire statistical manifold $\mathbb{S}(\mathbb{R}^n)$, because not only is $Y$ not in general bounded on $\mathbb{S}(\mathbb{R}^n)$, it is not smooth either: it is not even continuous.[4] This is not a difficulty if we consider instead the restricted mappings $Y_i$. These, being defined on finite-dimensional manifolds, are smooth under very general assumptions about the models $\mathbb{M}_i, i = 0, 1$.

The LM test statistic is defined as follows:

$$LM = \langle \nabla_{\hat{m}_0} Y_1, \nabla_{\hat{m}_0} Y_1 \rangle. \tag{8}$$

That is, the statistic is the squared norm of the gradient of the evaluation mapping $Y$ restricted to the alternative model, evaluated at the maximum likelihood estimate of the null model. This definition is probably rather unfamiliar! On the other hand, it is simple and also intrinsic. To see that it is equivalent to conventional definitions, it is perhaps easiest to have recourse to the methods of Davidson and MacKinnon (1987).

There is no simple means of defining what could be recognised as a Wald test statistic in an intrinsic manner. *If* we are prepared to accept a particular representation of the restrictions that specify $\mathbb{M}_0$ as a subset of $\mathbb{M}_1$, then we can at least write down what we may call the Wald test associated with that representation of the restrictions in an intrinsic fashion, and thereby see more clearly why the test statistic depends on the representation chosen. Let $R : \mathbb{M}_1 \to \mathbb{R}^r$ be a smooth mapping with the property that $R(m) = 0 \forall m \in \mathbb{M}_0$, but $R(m) \neq 0 \quad \forall m \in \mathbb{M}_1, m \notin \mathbb{M}_0$. Clearly $\mathbb{M}_0$ can be defined as

$$\mathbb{M}_0 = \{ m \in \mathbb{M}_1 | R(m) = 0 \in \mathbb{R}^r \}. \tag{9}$$

The Wald test statistic associated with the representation $R$ of the restrictions is then constructed as follows. Define an $r \times r$ matrix $V$ with typical element

$$V_{ij} = \langle \nabla_{\hat{m}_1} R_i, \nabla_{\hat{m}_1} R_j \rangle,$$

[4] No evaluation mapping is smooth on an $L^2$ Hilbert space of functions. To see this, observe that one can find two functions which are arbitrarily close in $L^2$ norm and which have fixed different values at some chosen point.

where $R_i$, $R_j$ are respectively the $i^{\text{th}}$ and $j^{\text{th}}$ component functions of $R$. Then the statistic is

$$W = R^\top(\hat{m}_1)V^{-1}R(\hat{m}_1). \tag{10}$$

The definition ((10)) is indeed intrinsic, but it is particular to the representation $R$. Other representations of the restrictions may be obtained by considering any mapping $Q : \mathbb{R}^r \to \mathbb{R}^r$ that maps the origin and only the origin to the origin.[5] Then $R' \equiv Q \circ R : \mathbb{M}_1 \to \mathbb{R}^r$ provides an equally valid way of characterising the null model: indeed, ((9)) is true with $R$ replaced by $R'$. If we denote by $\hat{J}$ the Jacobian of the mapping $Q$ evaluated at $R(\hat{m}_1) \in \mathbb{R}^r$, the matrix $V$ will be replaced by

$$V' \equiv \hat{J}V\hat{J}^\top.$$

If, but only if, $Q$ is a *linear* mapping, so that $J$ is a constant matrix such that for any $x \in \mathbb{R}^r$, $Q(x) = Jx$, will $W$ be the same for both $R$ and $R'$.

The definition ((10)) and the above reasoning show clearly that the Wald statistic is unaffected by the parametrisation of either of the *models* $\mathbb{M}_1$ or $\mathbb{M}_0$, but that it *is* affected by the "parametrisation" of the *restrictions*, that is, the choice of $R$. There seems to be no obvious way of escaping the dependence on the parametrisation of the restrictions as long as we stick with the conventional idea of the Wald test as a test of whether the restrictions are significantly violated at the unrestricted maximum likelihood estimate.

There exists a fourth classical test statistic, not nearly so well known as the trinity of LR, LM, and Wald, but asymptotically equivalent to them under DGP's close to the null hypothesis. It is the so-called $C(\alpha)$ test of Neyman. The original reference is Neyman (1959), and the test has been reconsidered by Smith(1987), Dagenais and Dufour (1989), and Davidson and MacKinnon (1991). Like the LM and Wald tests, the $C(\alpha)$ test is constructed about a particular DGP, but unlike the better-known tests, which are constructed around either the unrestricted or the restricted MLE, the $C(\alpha)$ test can be computed at any DGP in $\mathbb{M}_0$. For data generated by DGP's close to the DGP used to compute the test, the statistic is asymptotically equivalent to the other classical tests. The DGP used for the $C(\alpha)$ test will typically be the result of a consistent but inefficient estimation procedure applied to the null hypothesis. If this DGP is denoted by $\tilde{m}_0$, the test statistic can be expressed intrinsically as follows:

$$C(\alpha) = \langle \nabla_{\tilde{m}_0} Y_1, \nabla_{\tilde{m}_0} Y_1 \rangle - \langle \nabla_{\tilde{m}_0} Y_0, \nabla_{\tilde{m}_0} Y_0 \rangle . \tag{11}$$

It is readily seen that the LM test is in fact a special case of the $C(\alpha)$ test, since, at the restricted MLE, we have

$$\nabla_{\hat{m}_0} Y_0 = 0$$

by the first-order conditions for a restricted maximum of the loglikelihood function. Consequently ((11)) reduces to ((8)) when $\tilde{m}_0 = \hat{m}_0$.

[5] And still other representations exist, which respect only the set on which $R(m) = 0$. But we need not consider these.

In the next section we shall consider a variety of test statistics which, although not Wald tests *stricto sensu*, are based on the Wald principle, in the sense that they are calculated in terms of the unrestricted MLE. Among these will be some tests that are actually $C(\alpha)$ tests. However the DGP at which they are computed is based on the unrestricted MLE. Although most of the procedures considered turn out to be unsatisfactory, either because they are not invariant to reparametrisations of the alternative model or of the restrictions, or else are invariant but not simple to compute, there is one $C(\alpha)$-style procedure which, though not invariant, is much less badly behaved in that regard than other non-invariant procedures, and is easily computed.

## 4 Suggestions for an Invariant Wald Test.

The classical hypothesis tests use a variety of ways of measuring how far the unrestricted MLE, $\hat{m}_1$, is from the null model $\mathbb{M}_0$. The LR test does so directly in terms of the difference of the maximised loglikelihood functions; the LM test does so indirectly by testing the gradient of the loglikelihood function at the restricted MLE $\hat{m}_0$; and the Wald test does so at the unrestricted MLE $\hat{m}_1$ in terms of a highly arbitrary metric defined in terms of some representation $R$ of the restrictions that define the null model $\mathbb{M}_0$.

The differential-geometric approach sketched in section 2 suggests naturally a variety of ways of measuring statistical distance on statistical manifolds. These ways fall into two broad categories – local and global. The LR test yields a global measure, since for different realised samples $Y^n$ the unrestricted and restricted MLE's may be anywhere on the manifolds $\mathbb{M}_1$ and $\mathbb{M}_0$. The LM test uses a local measure, since it is defined exclusively in terms of quantities defined *at* the restricted MLE $\hat{m}_0$; similarly the $C(\alpha)$ test uses a local measure at the DGP $\tilde{m}_0$ belonging to $\mathbb{M}_0$. The definition ((10)) of the Wald test is also a local one: everything that appears in the definition is evaluated at the unrestricted MLE $\hat{m}_1$.

In the search for an invariant Wald-style test, we shall consider both local and global possibilities. It will turn out that there is no obvious test that is both local and invariant. The reason, roughly speaking is this: the evaluation mapping $Y_1$ gives its global maximum over $\mathbb{M}_1$ at the MLE $\hat{m}_1$. Consequently its gradient $\nabla_{m_1} Y_1$ is zero, just as $\nabla_{m_0} Y_0$ is zero at the restricted MLE. There is therefore no material from which to construct a test statistic available from the gradient of $Y_1$ at $m_1$. The value $Y(m_1)$ is used in the LR test, and the Hessian, essentially the gradient of the gradient of $Y$, may be used for the purposes of estimating the information matrix, but there is no first-derivative information available. In order to create some, the usual Wald procedure makes use of the restrictions under test, but then the problem arises that the way in which the restrictions are formulated affects the test statistic. If a *natural* way of formulating restrictions existed, then a Wald test based on this formulation would be free of the arbitrariness that afflicts the Wald test in general, but no such natural formulation has been found.

For this reason, global possibilities are worth considering. The LR test exemplifies such possibilities, but it requires knowledge of the restricted estimate $\hat{m}_0$ as well as of

$\hat{m}_1$. Can other global (and invariant) procedures be defined in terms solely of $\hat{m}_1$? If not, what is the minimum of extra information needed to construct a test? Do tests so constructed have reasonable properties in finite samples? It appears that the answers to these questions are negative or discouraging: by definition a *global* procedure cannot be defined in terms of a single point; the extra information needed for a test is more or less that provided by estimating the restricted model; and there is no reason to believe that finite-sample properties will be any better for newly invented tests than for existing ones like LR.

The dilemma is then that simplicity of test construction requires a local procedure, but local procedures necessarily involve an unwelcome degree of arbitrariness. Global procedures, at least if they are defined intrinsically, are not arbitrary, but need at least as much effort to compute as the LR test. By considering various suggestions that have been made in order to find an invariant Wald test, we shall see the force of this dilemma, but it will also turn out that in some circumstances approximations exist that make the computation of reasonably well-behaved tests possible.

Critchley, Marriott, and Salmon (1990) (CMS) propose what they call a *geodesic* version of the Wald test. Their idea is to treat $\mathbb{M}_1$ as a statistical manifold, just as we have done here, and to trace out, within $\mathbb{M}_1$, a geodesic from $\hat{m}_1$ to $\mathbb{M}_0$. That is, they seek a curve, in $\mathbb{M}_1$, starting from $\hat{m}_1$ and ending somewhere on $\mathbb{M}_0$, of minimum length, as measured by the arc length formula ((5)). The geodesic test statistic is then the square of the minimised arc length, and it will have, asymptotically, a central chi-squared distribution with $r$ degrees of freedom under the null hypothesis.

There are various problems and ambiguities associated with this procedure, quite apart from the sometimes nearly insuperable computational difficulties of finding and measuring the length of the geodesic. First, why should the geodesic be restricted to lie in $\mathbb{M}_1$ rather than simply being calculated in $\mathbb{S}(\mathbb{R}^n)$, the big manifold in which $\mathbb{M}_1$ is embedded? Even if there were no other considerations, it would be vastly simpler to locate and measure the unrestricted geodesic in $\mathbb{S}(\mathbb{R}^n)$ than the restricted one in $\mathbb{M}_1$. Further, the unrestricted geodesic would necessarily be no longer than the restricted one. If geodesics do indeed measure something that can usefully be thought of as *statistical distance*, then use of a geodesic artificially restricted to pass through $\mathbb{M}_1$ will presumably overstate the distance from $\hat{m}_1$ to $\mathbb{M}_0$, the more so the more curved is $\mathbb{M}_1$. Figure 3 illustrates this point. Preliminary Monte Carlo work by CMS suggests that this is indeed the case: their test rejects true nulls much too often if asymptotic nominal critical values are used.

Another question arises from the existence of a whole one-parameter family of natural connections on $\mathbb{S}(\mathbb{R}^n)$. It is natural to ask which one should be used in order to determine the geodesic curve from $\hat{m}_1$ to $\mathbb{M}_0$. CMS use the metric connection, arguing that it is more natural than the others. Again, use of any other connection will necessarily lead to a longer geodesic than that given by the metric connection, which, by construction, minimises the arc length of the geodesic.

Even if one can find reasons to prefer one connection over another, why should the geodesic used for calculating the test statistic be the one that minimises arc length? The

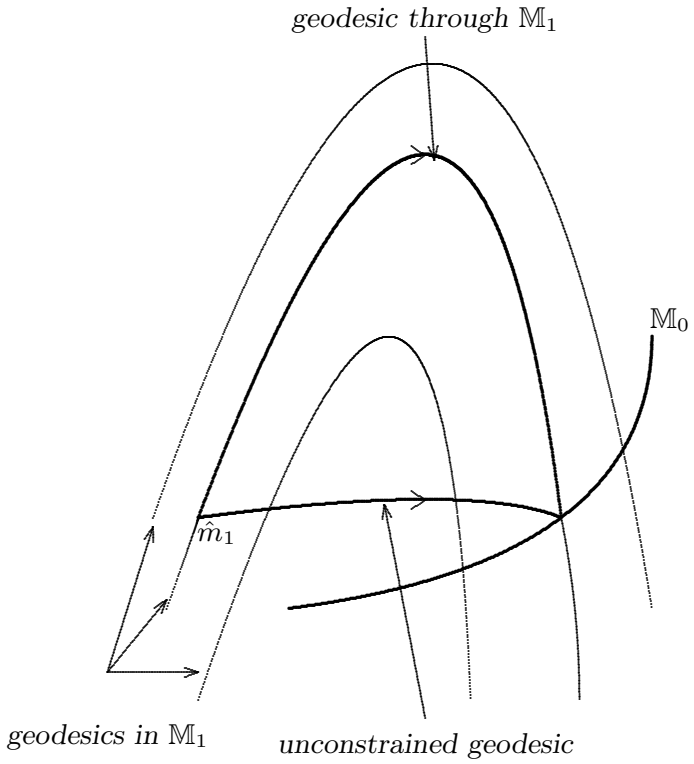


**Figure 3**

**Constrained and unconstrained geodesics**

point of this question is that, if a null model is false, then the maximum likelihood estimate of that model converges for large sample sizes to what we may call, by analogy with the terminology used in connection with quasi-maximum likelihood estimation (see White (1980)), the *pseudo-true* DGP. This pseudo-true DGP, call it $m_Q$, is defined as the DGP in $\mathbb{M}_0$ which minimises the Kullback-Leibler information criterion (KLIC) for distance from the true DGP to $\mathbb{M}_0$. The KLIC is an *asymmetric* distance measure, so that it is necessary to specify the *direction* in which it is used: here one starts from the true DGP and measures from it to the DGP $m_Q$ contained in the null model. Its asymmetry ensures that the KLIC is not equivalent to any distance measure of the sort we have considered here. No connection can ever lead to geodesics through two given points that look different in one direction and the other. But there is a strong intuitive case for using a geodesic, in whatever connection, that links the unrestricted MLE $\hat{m}_1$ to the point $m_Q$ of $\mathbb{M}_0$ that minimises the KLIC from $\hat{m}_1$ to $\mathbb{M}_0$ rather than any other measure of distance. In fact, as will be clear in the next section, in some cases $m_Q$ is just the restricted MLE $\hat{m}_0$.

The KLIC is asymmetric, but it is intrinsic. In fact, if two DGP's $m_1$ and $m_2$ in $\mathbb{S}(\mathbb{R}^n)$ are characterised by log densities $\ell_1$ and $\ell_2$, then the KLIC from $m_1$ to $m_2$ is

$$E_{m_1}(\ell_1 - \ell_2). \tag{12}$$

The notation means that the expectation is calculated using the probability density associated with DGP $m_1$, namely $\exp(\ell_1)$. In the other direction, from $m_2$ to $m_1$, the KLIC is

$$E_{m_2}(\ell_2 - \ell_1).$$

The usual interpretation of the KLIC, and the reason for which its messy asymmetric character is tolerated, is that it measures the power that statistical tests have, used on samples that were actually generated by the source DGP $m_1$, to discriminate between $m_1$ and $m_2$. Since $m_1$ and $m_2$ are in general of differing degrees of "noisiness", there is no reason to suppose that this discriminatory power will be same for data generated by each of them.

If it is possible to choose a DGP in $\mathbb{M}_0$ in a non-arbitrary way, and somehow measure the distance to it from $\hat{m}_1$, then it is again possible to consider a local test. The reason is that a *direction* is involved, namely that from $\hat{m}_1$ to the chosen DGP in $\mathbb{M}_0$, and a direction can be expressed as a tangent at $\hat{m}_1$, that is, in a local fashion. Some such local tests will be examined later.

It should be emphasised that everything suggested up to now is asymptotically equivalent to everything else. All of the natural connections on $\mathbb{S}(\mathbb{R}^n)$ are locally the same, and all the geodesics that they generate linking $\hat{m}_1$ to $\mathbb{M}_0$ have the same length asymptotically under DGP's that drift at a rate proportional to $n^{-\frac{1}{2}}$ towards a DGP contained in $\mathbb{M}_0$.[6] Thus what is at stake in all this discussion, besides ease of computation, is the finite-sample properties of the proposed test statistics. After all, the conventional Wald statistic, ((10)), *is* asymptotically independent of the representation $R$ of the restrictions used to compute the test. But rather than plunge directly into Monte Carlo investigations or stochastic expansions designed to elucidate finite-sample behaviour, it is prudent to eliminate from consideration tests that are clearly dominated by existing tests. It will appear that not much is left after this elimination process.

It may be useful to present at this point a catalogue of suggestions for invariant test statistics, and discuss how they may be computed. For none of them will it be necessary to represent the restrictions explicitly. The requirement imposed by CMS that their geodesic pass through $\mathbb{M}_1$ seems arbitrary and potentially dangerous for the size of the test. None of the following suggestions uses it. First, for global tests, it is necessary to select a DGP in $\mathbb{M}_0$ with a view to computing the distance between it and the MLE $\hat{m}_1$. This may be the restricted MLE, $\hat{m}_0$, or the DGP $m_Q$ that minimises the KLIC from $\hat{m}_1$ to $\mathbb{M}_0$ (these two may be the same), or some other DGP obtained, perhaps, by a consistent estimation procedure other than maximum likelihood, as for a $C(\alpha)$ test, or, more geometrically, the DGP that minimises the arc length of the geodesic from $\hat{m}_1$ to $\mathbb{M}_0$ in one or other of the natural connections. Once this DGP, call it $\check{m}_0$, has been selected, a geodesic is drawn from $\hat{m}_1$ to $\check{m}_0$, defined once more by use of one of the natural connections, although even if one was used to select $\check{m}_0$ it need not be the same one here. Then the arc-length of this geodesic is computed, and its square serves as the test statistic. Local tests can be defined

[6] The concept of a *drifting DGP* is defined in Davidson and MacKinnon (1987), in which it is called by the less satisfactory name of *local DGP*. The concept is very useful in analyses of test power.

if it is possible to determine a direction at $\hat{m}_1$. This could be for instance the tangent to a geodesic, constructed as for a global test, at the beginning point of the geodesic, that is, $\hat{m}_1$. The test statistic in this case would be the squared norm of the tangent.

It is now time to make explicit how to find geodesic curves from one DGP $m_1 \in \mathbb{S}(\mathbb{R}^n)$ to another, $m_2$, for the different natural connections defined on $\mathbb{S}(\mathbb{R}^n)$. It turns out that such curves can be written down simply, by means of convex combinations, for different ways of representing the elements of $\mathbb{S}(\mathbb{R}^n)$. In the natural representation, namely that which uses the square-root density, we consider the set of convex combinations of the square-root density, $\psi_1$, say, associated with DGP $m_1$ and $\psi_2$, associated with $m_2$. We obtain

$$\psi_\lambda \equiv (1-\lambda)\psi_1 + \lambda\psi_2. \tag{13}$$

Of course $\psi_\lambda$ is not a square-root density for $\lambda \neq 1, 2$, since it does not satisfy the normalisation condition for a density. This is easily rectified: we simply renormalise, as follows:

$$\psi(\lambda) = \frac{\psi_\lambda}{\|\psi_\lambda\|}. \tag{14}$$

Here

$$\|\psi_\lambda\| = (1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\cos\phi$$

where $\cos\phi \equiv \langle\psi_1, \psi_2\rangle$ is the cosine of the angle $\phi$ between the two DGP's. We claim (without proof) that the curve $c(\lambda)$ given by ((14)) is the geodesic in the metric connection joining $m_1$ and $m_2$. In order to compute its length, we must construct the tangent to $c(\lambda)$ for each $\lambda \in [0, 1]$. This tangent is represented simply by the derivative of $c(\lambda)$, which is

$$c'(\lambda) = \left( \frac{1}{\|\psi_\lambda\|^2} \{(\psi_1 - \psi_0)\|\psi_\lambda\| + 2(1-2\lambda)(1-\cos\phi)\psi_\lambda\} \right) \tag{15}$$

Next one calculates the norm of the above expression as a function of $\lambda$, and integrates the result from 0 to 1, according to formula ((5)). The calculation is tedious, but the answer is simple: it is just $\phi$, the angle between the two DGP's. This is of course in accord with the usual distance between two points on the surface of any sphere of unit radius, measured as the length of the arc of the great circle through the two points.

The reason for which the calculation of the length is complicated is that $\lambda$ is not an *arc-length* parameter along $c(\lambda)$, since the norm of $c'(\lambda)$ depends on $\lambda$. The reason for this should be clear from Figure 4. $\lambda$ would be an arc-length parameter along the *secant* joining $m_1$ to $m_2$, but not along the *arc*. For small arcs, the difference is of the second order of smalls, and so asymptotically it is something else that can be ignored.

If a geodesic is defined using a genuine arc-length parameter, $\epsilon$ say, the calculation of the length of the geodesic is greatly simplified. For an arc-length parameter, $\|c'(\epsilon)\|$ is constant by definition. Thus if the geodesic is defined from $\epsilon = \epsilon_1$ to $\epsilon = \epsilon_2$, its length is

$$\int_{\epsilon_1}^{\epsilon_2} \|c'(\epsilon)\| d\epsilon,$$

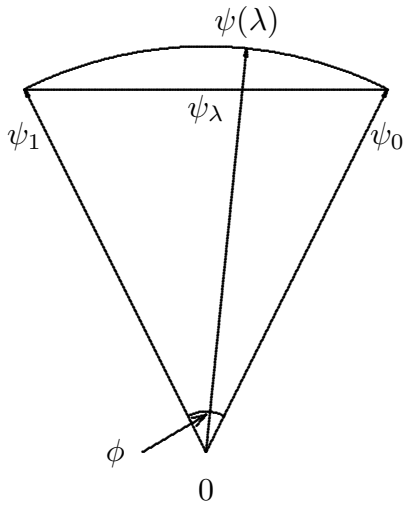


**Figure 3**

**Constrained and unconstrained geodesics**

which is just

$$(\epsilon_2 - \epsilon_1)\|c'(0)\|.$$

Thus it is sufficient to calculate the norm of $c'(\epsilon)$ for just one value of $\epsilon$, at the beginning of the curve, say. This suggests a way of implementing local tests based on geodesics. We saw above that the "natural" parameter $\lambda$ was almost an arc-length parameter. It will almost always be much easier to use than a genuine arc-length parameter. In particular, the length of the (metric) geodesic from $\psi_1$ to $\psi_2$ can be approximated by $\|c'(0)\|$, which, from ((15)), is $(2(1-\cos\phi))^{\frac{1}{2}}$. The leading term in a Taylor expansion of this for small $\phi$ is indeed just $\phi$. This approximation clearly merits the name *local*, since it uses information only about the start of the geodesic, which, for all the proposed statistics, is $\hat{m}_1$.

The approximation is not of much use for the metric connection, for which the length is simply computed as just the angle between two square-root densities. But it is of great use for the other connections that we consider. Let $\delta$ be a parameter between zero and unity, and suppose that we are interested in a connection, indexed by $\delta$, between two DGP's with log-densities of $\ell_1$ and $\ell_2$. The connection turns out to be natural in the manifold $\mathbb{P}^\delta(\mathbb{R}^n)$, as follows. The DGP's are represented in $\mathbb{P}^\delta(\mathbb{R}^n)$ by the $\delta^{\text{th}}$ power of their densities, which can be written as $\exp(\delta\ell_i)$, $i = 1, 2$. Just as with square-root densities, in ((13)), consider a curve defined by convex combinations of these powers, as follows:

$$\exp\left(\delta\ell(\lambda)\right) = \left[(1-\lambda)\exp\left(\delta\ell_1\right)) + \lambda\exp\left(\delta\ell_2\right))\right]/N_\delta(\lambda) \tag{16}$$

where the normalisation factor $N_\delta(\lambda)$ is defined by the formula

$$N_\delta(\lambda) = \left[\int \left\{(1-\lambda)e^{\delta\ell_1} + \lambda e^{\delta\ell_2}\right\}^{\frac{1}{\delta}} dy^n\right]^\delta$$

Although $\lambda$ is still not an arc-length parameter, the curve $c_\delta(\lambda)$ defined by ((16)) is (we state this without proof) the geodesic from $m_1$ to $m_2$ in the connection indexed by $\delta$. Two special cases are of particular interest. If $\lambda = 1$, we obtain the so-called *mixture* connection, given by simple convex combinations of densities. Note that in this case, $N_1(\lambda) = 1$. The other case is the limit as $\delta \to 0$. We shall see in a moment that this corresponds to the so-called *exponential* connection, in which geodesics are defined by means of convex combinations of log-densities.

Let us now compute the approximate length of these geodesics, as $\|c'_\delta(0)\|$. We find without too much trouble from ((16)) that

$$\ell'(0) = \frac{1}{\delta}\left(\exp(\delta(\ell_2 - \ell_1)) - C(\delta)\right),$$

where the important function $C$ is defined as follows:

$$C(\delta) = \int \exp\left((1-\delta)\ell_2 + \delta\ell_1\right) dy^n. \tag{17}$$

The function $C$ can in fact be interpreted in several ways. It is the expectation under $m_1$ of $\exp(\delta(\ell_2 - \ell_1))$, and thus the moment-generating function under $m_1$ of the random variable $\ell_2 - \ell_1$. It is also the expectation under $m_2$ of $\exp(\delta(\ell_1 - \ell_2))$, with a corresponding interpretation as a moment-generating function. We shall see that most of our proposed statistics can be expressed in terms of $C$ and its derivatives.

The norm of the tangent $c'_\delta(0)$ can be computed using $\ell'(0)$ and the expectation definition of the norm in $\mathbb{S}(\mathbb{R}^n)$. We obtain

$$\begin{aligned} E_0(\ell'(0))^2 &= \frac{1}{\delta^2}\int e^{\ell_0}\left(e^{\delta(\ell_2-\ell_1)} - C(\delta)\right)^2 \\ &= \frac{1}{\delta^2}E_0\left(e^{2\delta(\ell_2-\ell_1)} - 2C(\delta)e^{\delta(\ell_2-\ell_1)} + C^2(\delta)\right). \\ &= \frac{1}{\delta^2}\left(C(2\delta) - C^2(\delta)\right) \end{aligned} \tag{18}$$

Notice that the limit of this last expression as $\delta \to 0$ is

$$C''(0) - (C'(0))^2. \tag{19}$$

The formula ((18)) permits the implementation of a series of local possibilities for a Wald test. The DGP $m_1$ is replaced by the unrestricted ML estimate $\hat{m}_1$, and $m_2$ is replaced by the "target" DGP $\check{m}_0$.

Consider now a geodesic from $m_1$ to $m_2$ in the exponential connection, which corresponds to the limiting case $\delta \to 0$. The geodesic itself is given by the set of normalised convex combinations

$$\ell(\lambda) = (1-\lambda)\ell_1 + \lambda\ell_2 - \log C(\lambda) \tag{20}$$

in which the normalising term[7] turns out to be nothing other than the function $C$ of ((17)), evaluated at $\lambda$. The derivative of $\ell(\lambda)$is

$$\ell'(\lambda) = \ell_2 - \ell_1 - \frac{C'(\lambda)}{C(\lambda)}$$

and so the squared norm of the tangent to the geodesic at $\lambda$ is

$$E_{m_\lambda}\left(\ell_2 - \ell_1 - \frac{C'(\lambda)}{C(\lambda)}\right)^2. \tag{21}$$

Now notice that

$$\begin{aligned} E_{m_\lambda}(\ell_2 - \ell_1) &= \frac{1}{C(\lambda)}\int (\ell_2 - \ell_1)\exp\left((1-\lambda)\ell_1 + \lambda\ell_2\right) dy^n \\ &= \frac{1}{C(\lambda)}\frac{\partial}{\partial\lambda}\int \exp\left((1-\lambda)\ell_1 + \lambda\ell_2\right) dy^n \\ &= \frac{C'(\lambda)}{C(\lambda)}, \end{aligned} \tag{22}$$

and, similarly,

$$E_{m_\lambda}(\ell_2 - \ell_1)^2 = \frac{C''(\lambda)}{C(\lambda)}.$$

Consequently expression ((21)) becomes

$$\frac{C''(\lambda)}{C(\lambda)} - \left(\frac{C'(\lambda)}{C(\lambda)}\right)^2,$$

(compare ((19))), which can be expressed more simply as

$$\frac{\partial^2}{\partial\lambda^2}\log C(\lambda),$$

or, if we define $\gamma(\lambda) \equiv \log C(\lambda)$, as just $\gamma''(\lambda)$. The length of the exponential geodesic from $m_1$ to $m_2$ is therefore

$$\int_0^1 \sqrt{\gamma''(\lambda)}d\lambda. \tag{23}$$

Notice that, as we mentioned earlier, this quantity is defined solely in terms of the function $C$ of ((17)).

The exponential connection was chosen over other possibilities in the above calculation of an exact geodesic length for two reasons. One is simply that the plain log density is mathematically more tractable than fractional powers of the density. The other reason is

[7] A term rather than a factor because we are working with a log density.

deeper, and provides a theoretical justification for preferring the exponential connection to the other connections in the context of hypothesis testing. It is that the function $C$ and its logarithm $\gamma$ inherit from the log-densities $\ell_1$ and $\ell_2$ a structure by which the information in successive independent observations is accumulated *additively* in both the approximate and exact lengths of the exponential geodesic connecting them, ((19)) and ((23)) respectively.

Although the reasoning can be extended without major difficulty to the case of dependent observations, the present point can be made most clearly if we assume that the DGP's $m_1$ and $m_2$ generate $n$ independent, but not necessarily identically distributed, observations. In this case, each of the log-densities $\ell_i, i = 1, 2$ can be additively decomposed as follows:

$$\ell(y^n) = \sum_{t=1}^{n} \ell_t(y_t), \tag{24}$$

where $\ell_t$ is the log-density of observation $t$. Similarly $C(\lambda)$ can be written as

$$\begin{aligned} C(\lambda) &= \int \exp\left(\sum_{t=0}^{n} \left((1-\lambda)(\ell_1)_t + \lambda(\ell_2)_t\right)\right) dy^n \\ &= \int \prod_{t=1}^{n} \exp\left((1-\lambda)(\ell_1)_t + \lambda(\ell_2)_t\right) dy^n \\ &= \prod_{t=1}^{n} \int \exp\left((1-\lambda)(\ell_1)_t + \lambda(\ell_2)_t\right) dy_t, \end{aligned}$$

so that

$$\gamma(\lambda) = \sum_{t=1}^{n} \gamma_t(\lambda), \tag{25}$$

where the *contribution* to $\gamma(\lambda)$ from observation $t$, $\gamma_t(\lambda)$, is just

$$\gamma_t(\lambda) = \log \int \exp\left((1-\lambda)(\ell_1)_t + \lambda(\ell_2)_t\right) dy_t.$$

Now consider the DGP's on the exponential geodesic for values of $\lambda$ different from 1 or 0. From ((20)) along with ((24)) and ((25)) we have

$$\ell(\lambda) = \sum_{t=1}^{n} \left\{(1-\lambda)(\ell_1)_t + \lambda(\ell_2)_t - \gamma_t(\lambda)\right\}.$$

We see that the additive structure of the log densities across observations is preserved, or, in other words, if the DGP's $m_1$ and $m_2$ generate independent observations, so do the intermediate DGP's on the exponential geodesic connecting them. It is important to observe that this is *not* the case with any other connection, including the metric connection.

One way of expressing the particular property of the exponential connection we have just described is as follows. If we restrict attention to DGP's for which the successive

observations are independent, this corresponds to restricting attention to a submanifold of $\mathbb{S}(\mathbb{R}^n)$, call it $\mathbb{P}(\mathbb{R}^n)$, which is a *product manifold* of $n$ unit spheres in $\mathbb{S}(\mathbb{R})$:

$$\mathbb{P}(\mathbb{R}^n) = \bigotimes_{t=1}^{n} \mathbb{S}(\mathbb{R}).$$

It is clear that $\mathbb{P}(\mathbb{R}^n)$ is the subset of $\mathbb{S}(\mathbb{R}^n)$ containing square-root densities corresponding to independent observations, that is, densities that factorise into $n$ marginal densities. On this product manifold, it can be checked that the inner product is the product of the $\mathbb{P}(\mathbb{R}^n)$ inner products of the factors. Only the exponential connection preserves the structure of $\mathbb{P}(\mathbb{R}^n)$, in the sense that geodesics between points of $\mathbb{P}(\mathbb{R}^n)$ lie exclusively in $\mathbb{P}(\mathbb{R}^n)$ only for the exponential connection.

We may now write down both the exact and approximate distances in the exponential connection between two DGP's that generate independent observations. From ((23)) and ((25)) we conclude, first, that the exact distance is

$$\int_0^1 \sqrt{\sum_{t=1}^{n} \gamma_t''(\lambda) d\lambda}; \tag{26}$$

while the approximate distance is

$$\sqrt{\sum_{t=1}^{n} \gamma_t''(0)}. \tag{27}$$

If connections other than the exponential one are used, there is no additive structure for the exact distance. However since the approximate distance, given by ((18)), is expressed solely in terms of the function $C$, its multiplicative structure can be used in order to permit a calculation in which each observation can be treated separately. The exact lengths are in any event a good deal harder to compute, even for a single observation, and we shall spend no time on them here.

## 5 Some Examples.

The examples in this section will be restricted to regression models, not necessarily linear, with normal errors. The first, and simplest, example is of a straightforward location model, which can conveniently be written as

$$y_t = \mu + \epsilon_t; \qquad t = 1, \ldots, n \qquad \boldsymbol{\epsilon} \sim N(0, I). \tag{28}$$

This is, as it stands, a one-parameter model of the type considered earlier: it defines a *curve* in $\mathbb{S}(\mathbb{R})$. The curve is parametrised by $\mu$. If as usual $\ell_t$ denotes the log-density of observation $t$, we have

$$\ell_t(y_t, \mu) = -\frac{1}{2}\log 2\pi - \frac{1}{2}(y_t - \mu)^2.$$

Consider the determination of the distance between two DGP's in the family ((28)) characterised by the parameters $\mu_1$ and $\mu_2$. We first calculate the function $C$ of ((17)) for these two DGP's. We obtain for each observation $t$:

$$C_t(\lambda) = \exp\left\{-\frac{1}{2}(\mu_1 - \mu_2)^2\lambda(1-\lambda)\right\},$$

so that

$$\gamma_t(\lambda) = -\frac{1}{2}(\mu_1 - \mu_2)^2\lambda(1-\lambda). \tag{29}$$

The derivatives of $\gamma_t$ are

$$\gamma_t'(\lambda) = -\frac{1}{2}(\mu_1 - \mu_2)^2(1-2\lambda),$$

and

$$\gamma_t''(\lambda) = (\mu_1 - \mu_2)^2.$$

The distance between the DGP's characterised by $\mu_1$ and $\mu_2$ in the exponential connection is therefore, by ((23)), for each observation:

$$\int_0^1 \sqrt{\gamma_t''(\lambda)}d\lambda = |\mu_2 - \mu_1|, \tag{30}$$

where we have chosen the positive square root in order to obtain a positive length. For a sample of $n$ observations, the combination rule ((26)) can be easily applied: the distance is

$$\sqrt{n}|\mu_2 - \mu_1|. \tag{31}$$

Notice that ((31)) is both the exact and the approximate distance for the exponential connection, or, in other words, $\mu$ is an arc length parameter on the exponential geodesic. For the connection indexed by $\delta$, the approximate squared distance for each observation is given by ((18)):

$$\frac{1}{\delta^2}\left(C(2\delta) - C^2(\delta)\right) =$$

$$\frac{1}{\delta^2}\left[\exp\left(-(\mu_1 - \mu_2)^2\delta(1-2\delta)\right) - \exp\left(-(\mu_1 - \mu_2)^2\delta(1-\delta)\right)\right].$$

For small values of $|\mu_1 - \mu_2|$ this expression can be expanded in a Taylor series. The result is

$$(\mu_1 - \mu_2)^2\left(1 - \frac{\delta(2-3\delta)}{2}(\mu_1 - \mu_2)^2 + \ldots\right), \tag{32}$$

from which we can see that the leading term is independent of the choice of $\delta$, and coincides with the result ((30)) for the exponential connection. The approximation is plainly far from perfect, however, since we know that by construction the smallest *exact* squared distance is given by the metric connection, for which $\delta = \frac{1}{2}$. The approximation to the approximate distance given by ((32)) is however minimised at $\delta = \frac{1}{3}$.

Things are a little less simple if we allow the variance to be variable. Consider next the following model:

$$y_t = \mu_t + \epsilon_t, \qquad t = 1, \ldots, n \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I). \tag{33}$$

Although we have allowed for another element of generality in ((33)), namely the possibility that the mean of each observation be different, this does not in fact lead to any further complexity. Suppose that we consider two DGP's, $m_1$ and $m_2$, each described by ((33)), with different parameters $(\boldsymbol{\mu}_1, \sigma_1)$ and $(\boldsymbol{\mu}_2, \sigma_2)$. Tedious but not very difficult calculations demonstrate that the exponential geodesic through $m_1$ and $m_2$ passes through DGP's which are all of the form ((33)), so that the exponential connection preserves the structure of the normal model.[8] The contributions $\gamma_t(\lambda)$ can be computed, and they are, if $\Delta\mu_t$ denotes $(\mu_2)_t - (\mu_2)_t$ and $\Delta\tau$ denotes $\log(\sigma_2/\sigma_1)$:

$$\gamma_t(\lambda) = -\lambda\Delta\tau - \frac{1}{2}\log\left(1 - \lambda + \lambda e^{-2\Delta\tau}\right) - \frac{1}{2\sigma_1^2}\frac{\lambda(1-\lambda)(\Delta\mu_t)^2 e^{-2\Delta\tau}}{(1 - \lambda + \lambda e^{-2\Delta\tau})}. \tag{34}$$

This expression, like ((29)), depends on $\mu_1$ and $\mu_2$ only through $\Delta\mu_t$, and it is easy to see that in fact $\gamma(\lambda)$ itself depends only on $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. The second derivative of ((34)) is rather messy:

$$\begin{aligned}\gamma_t''(\lambda) = &\frac{1}{2}\frac{\left(1 - e^{-2\Delta\tau}\right)^2}{\left(1 - \lambda + \lambda e^{-2\Delta\tau}\right)^2} + \frac{(\Delta\mu_t)^2 e^{-2\tau_2}}{(1 - \lambda + \lambda e^{-2\Delta\tau})} \\ &- \frac{(1-2\lambda)(\Delta\mu_t)^2 e^{-2\tau_2}\left(1 - e^{-2\Delta\tau}\right)}{\left(1 - \lambda + \lambda e^{-2\Delta\tau}\right)^2} \\ &- \frac{\lambda(1-\lambda)(\Delta\mu_t)^2 e^{-2\tau_2}\left(1 - e^{-2\Delta\tau}\right)^2}{\left(1 - \lambda + \lambda e^{-2\Delta\tau}\right)^3}.\end{aligned} \tag{35}$$

The above expression simplifies considerably when it is evaluated at $\lambda = 0$. We obtain

$$\gamma_t''(0) = \frac{1}{2}\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^2 + \frac{\sigma_1^2}{\sigma_2^4}\left(\Delta\mu_t\right)^2. \tag{36}$$

It is certainly theoretically possible to compute the integral of the square root of ((35)) with respect to $\lambda$ from 0 to 1, since the integrand is the square root of a rational function of $\lambda$. However the complexity of the calculation is such that the effort hardly seems justified. We make do therefore with ((36)), which permits us to use an approximate length for the geodesic from $m_1$ to $m_2$. The square of this approximate length is, for a sample size of $n$:

$$\frac{n}{2}\left(1 - \frac{\sigma_1^2}{\sigma_2^2}\right)^2 + \frac{\sigma_1^2}{\sigma_2^4}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2\|^2. \tag{37}$$

[8] It is well known, see Amari (1985) for example, that it preserves the structure of the exponential family. The normal family is a subfamily of the exponential family, the more special structure of which is also preserved.

We have now developed enough tools to be able to consider the case considered by Gregory and Veall (1985) (GV), in which a linear regression model is subject to nonlinear restrictions on the regression parameters. The estimation of the unrestricted model is simple, and can be done by ordinary least squares, although presumably the maximum likelihood estimate of the error variance will be preferred over the OLS one in this context. Suppose that this estimation yields an estimate $\hat{\sigma}^2$ for the error variance and a vector of fitted values $\hat{\boldsymbol{\mu}}$ with components $\hat{\mu}_t$. The estimate $\hat{\sigma}^2$ and the fitted values $\hat{\boldsymbol{\mu}}$ define the DGP from which our geodesic is to start – it corresponds to the $m_1$ of the general theory above. Theory has also suggested that for $m_2$ we take the DGP in the null model that minimises the KLIC from $m_1$. Alternatives would include the DGP that minimises the exact exponential geodesic distance from $m_1$ to the null model, but that would be computationally awkward. It is of course simple to minimise the approximate distance ((37)), and that provides another possibility.

First let us consider minimising the KLIC. From ((12)) and ((22)) it can be seen that the KLIC from $m_1$ to $m_2$ is just $-\gamma'(0)$. From ((34)) it can be computed that

$$\gamma'(0) = \frac{n}{2} \log \left( \frac{\sigma_1}{\sigma_2} \right)^2 + \frac{n}{2} \left( 1 - \frac{\sigma_1^2}{\sigma_2^2} \right) - \frac{1}{2\sigma_2^2} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2.$$

It is immediate that this is minimised with respect to $\boldsymbol{\mu}_2$ and $\sigma_2$ by minimising the sum-of-squares $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ so as to obtain $\tilde{\boldsymbol{\mu}}$ say, and then choosing $\sigma_2^2$ as

$$\tilde{\sigma}_2^2 = \sigma_1^2 + \frac{1}{n} \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_2\|^2. \tag{38}$$

The result ((38)) is familiar from the theory of pseudo-true values for misspecified regression models.

Minimisation of the approximate length of the exponential geodesic, given by ((37)), does not yield the same result. In fact, the sum-of-squares is again minimised, so that the minimising $\boldsymbol{\mu}_2$ is still $\tilde{\boldsymbol{\mu}}_2$, but the value for the variance parameter is different. It is readily seen to be:

$$\hat{\sigma}_2^2 = \sigma_1^2 + \frac{2}{n} \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_2\|^2. \tag{39}$$

We now have two relatively simple candidates for an invariant Wald test. We can calculate the approximate length of the exponential geodesic, as given by ((27)), from the unrestricted estimate of the DGP, characterised by $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$, to the point of the null model characterised by $\tilde{\boldsymbol{\mu}}$ and either the $\tilde{\sigma}_2^2$ of ((38)) or the $\hat{\sigma}_2^2$ of ((39)). The results are, if it is the KLIC that is minimised,

$$\frac{1}{\hat{\sigma}^2} \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2 \left( 1 - n^{-1} \frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} \right)^{-2} \left( 1 + \frac{1}{2} n^{-1} \frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} \right), \tag{40}$$

or, if the approximate length of the geodesic is minimised,

$$\frac{1}{\hat{\sigma}^2}\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2 \left(1 - 2n^{-1}\frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2}\right)^{-2} \left(1 + n^{-1}\frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2}\right). \tag{41}$$

Asymptotically, both of these are just

$$\frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2},$$

which, apart from some numerical constants related to numbers of degrees of freedom, is just the standard $F$ test, since, as is clear from its construction as a minimiser of a sum of squares, $\tilde{\boldsymbol{\mu}}$ is the vector of fitted values from the (in general nonlinear) least squares estimation of the restricted, null, model. Further, both of them can be expressed as simple functions of this $F$-statistic and the sample size. For the special case in which the null model is linear, their exact distributions could be computed, and they would have all the usual optimality properties of the $F$ test. Notice however that they are both numerically larger than (the appropriate multiple of) the $F$-statistic. That is of course a well-known feature of conventional Wald tests.

It is perhaps worth a word that the minimised KLIC, times 2, is also a suitable asymptotic test statistic. In this case, the minimised KLIC can be computed as

$$\frac{n}{2}\log\left(1 + n^{-1}\frac{\|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2}\right), \tag{42}$$

which is in fact precisely one half of the LR statistic, and also a numerical function of the $F$-statistic.

It is evident from the above discussion that all the test statistics considered, ((40)), ((41)), and ((42)) require knowledge of $\tilde{\boldsymbol{\mu}}$, which is just the vector of parameter estimates obtained from estimating the restricted model. This violates the basic idea of the Wald principle, and in the present context requires nonlinear estimation where the unrestricted model needs only linear OLS. It is worthwhile therefore to consider tests that do not have such strong requirements, and here $C(\alpha)$ tests seem natural candidates, since they can be computed at any DGP in $\mathbb{M}_0$. We will see how this can be done for the specific example considered by GV; the method is easily seen to be quite general.

Gregory and Veall consider a model of the form

$$y_t = x_{1t}\beta_1 + x_{2t}\beta_2 + \epsilon_t, \qquad t = 1, \ldots, n \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I), \tag{43}$$

with the nonlinear restriction

$$\beta_1\beta_2 - 1 = 0. \tag{44}$$

Model ((43)) can evidently be estimated simply by OLS. Whether or not the restriction ((44)) is true, the OLS estimates are consistent for $\beta_1$ and $\beta_2$. Therefore at least two possible DGP's in the model $\mathbb{M}_0$ defined by ((43)) with ((44)) suggest themselves as DGP's

at which a $C(\alpha)$ test may be computed, namely those characterised by the parameters $(\hat{\beta}_1, 1/\hat{\beta}_1)$ or $(1/\hat{\beta}_2, \beta_2)$.

As is clear from the general theory, the $C(\alpha)$ tests are defined invariantly, but it is equally clear that the tests are not unique. It is true that, *in the chosen parametrisation of model* ((43)), there are only two obvious points at which to consider computing a $C(\alpha)$ test, but even then many more can be found with little expenditure of imagination. For instance, a geometric mean of the form $(\beta_1/\beta_2)^{\frac{1}{2}}$ and its reciprocal could well seem particularly well suited to the model in hand. The invariance of the $C(\alpha)$ test thus seems to be offset by the fact that it could be evaluated at a wide variety of points, each of which will presumably yield a different value for the test statistic. (In general that is: with linear restrictions that would not be the case.) We are still in effect confronted by the problem that there is no natural direction towards $\mathbb{M}_0$ if we are restricted to working with information yielded by the unrestricted (OLS) estimates.

Although this difficulty seems to be at least as serious as the non-invariance of the conventional Wald test, it has a different origin, as can be made clear by geometric reasoning. Whereas the Wald test is affected by the representation of the *restrictions*, a $C(\alpha)$ test clearly is not, precisely because it *is* invariant. Suppose that we choose to parametrise ((43)) in such a way that the restriction ((44)) appears as a zero restriction on one of the parameters, for instance as

$$y_t = x_{1t}\gamma_1 + x_{2t}(\gamma_2 + \frac{1}{\gamma_1}) + \epsilon_t.$$

Here $\gamma_1 = \beta_1$, and $\gamma_2 = \beta_2 - (1/\beta_1)$. The restriction is just that $\gamma_2 = 0$. The unrestricted estimates are therefore $\hat{\gamma}_1 = \hat{\beta}_1$ and $\hat{\gamma}_2 = \hat{\beta}_2 - (1/\hat{\beta}_1)$. It seems to make sense to do a $C(\alpha)$ test at the DGP with parameters $(\hat{\beta}_1, 0)$. But this conclusion is completely dependent on our having chosen $\beta_1$ as the *other* parameter of the model along with $\gamma_2$. If instead we had defined $\gamma_1$ as $\beta_2$, we would have had a different result. In Figure 5, we see two possibilities out of an infinite number. In both, the coordinate curves corresponding to $\gamma_2 =$ constant are the same: one of them corresponds to the restricted model $\mathbb{M}_0$. But the other coordinate curves correspond to different choices of the other parameter, $\gamma_1$. It is clear that the "natural" point of $\mathbb{M}_0$ given by the unrestricted estimate can be virtually anything on $\mathbb{M}_0$. This point should be borne in mind when reading the paper of Dagenais and Dufour (1989).

A possible *approximate* way out of this difficulty is suggested by the theory of artificial regressions – see Davidson and MacKinnon (1990). Use of artificial regressions (*linear* regressions) makes it possible to perform *one-step efficient estimation* in a simple way. The idea is that, just as with $C(\alpha)$ tests themselves, it is possible to start from any DGP in $\mathbb{M}_0$, and, if it is close enough to the true DGP, also supposed to belong to $\mathbb{M}_0$, obtain by performing one linear regression an estimator asymptotically equivalent to maximum likelihood. Without giving details – they are in Davidson and MacKinnon (1990) – what one needs to do in the present context is to choose, in what we have seen must be an essentially arbitrary fashion, some value of $\beta_1$ from which to start – call it $\check{\beta}_1$. Then one

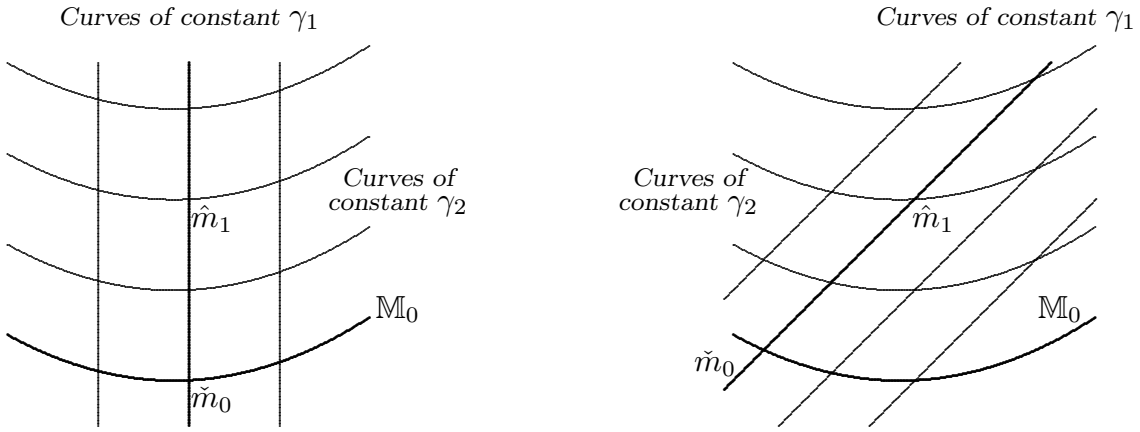


**Figure 3**

**Constrained and unconstrained geodesics**

runs the (univariate) regression

$$r_t(\check{\beta}_1) = R_t(\check{\beta}_1)b + \text{residual}, \tag{45}$$

where the regressand is

$$r_t(\check{\beta}_1) = y_t - x_{1t}\check{\beta}_1 - x_{2t}\frac{1}{\check{\beta}_1}$$

and the regressor is

$$R_t(\check{\beta}_1) = x_{1t} - \frac{x_{2t}}{\check{\beta}_2^2}.$$

The one-step efficient estimator of $\beta_1$ is then

$$\tilde{\beta}_1 \equiv \check{\beta}_1 + \hat{b}, \tag{46}$$

where $\hat{b}$ is the OLS estimate from ((45)).

Although the estimator ((46)) is not the exact restricted MLE, provided that the true DGP is close to $\mathbb{M}_0$, it should be reasonably close in many circumstances. The DGP defined by ((46)) could well be a good place at which to compute a $C(\alpha)$ test. Such a test would of course still depend on the original point of departure, $\check{\beta}_1$, but this dependence can reasonably be expected to be less important than it would be if the $C(\alpha)$ test were computed at $\check{\beta}_1$ itself.

Table 1 shows the results of a small Monte Carlo experiment run to test these ideas. Some of these results duplicate the work of GV and show that there is a fair amount of experimental error both in their work and in that presented here. Variance reduction methods are certainly available to rectify this state of affairs, but for present purposes their use is not really necessary.

# Table 1

Percentage of Rejections in 10,000 Trials from Tests of Nominal Size 5 per cent Null Hypothesis True.

| DGP | Test | $n = 20$ | $n = 30$ | $n = 40$ | $n = 50$ | $n = 100$ |
|---|---|---|---|---|---|---|
| $\beta_1 = 10, \quad \beta_2 = 0.1$ | WA | 31.63 | 26.13 | 22.91 | 20.38 | 15.48 |
| | WB | 7.08 | 5.91 | 5.94 | 5.20 | 5.14 |
| | CA1 | 8.27 | 6.73 | 6.51 | 5.63 | 5.46 |
| | CA2 | 32.71 | 26.98 | 23.59 | 20.82 | 15.72 |
| | OS1 | 8.27 | 6.73 | 6.51 | 5.63 | 5.46 |
| | OS2 | 8.28 | 6.73 | 6.51 | 5.63 | 5.46 |
| | LM1 | 8.27 | 6.73 | 6.51 | 5.63 | 5.46 |
| | LM2 | 28.90 | 24.51 | 21.48 | 19.52 | 14.81 |
| $\beta_1 = 5, \quad \beta_2 = 0.2$ | WA | 18.91 | 16.36 | 14.40 | 12.83 | 10.47 |
| | WB | 6.94 | 6.03 | 6.10 | 5.66 | 5.26 |
| | CA1 | 8.03 | 6.93 | 6.83 | 6.10 | 5.40 |
| | CA2 | 20.02 | 16.96 | 14.81 | 13.22 | 10.67 |
| | OS1 | 8.00 | 6.92 | 6.83 | 6.11 | 5.39 |
| | OS2 | 8.04 | 6.92 | 6.83 | 6.10 | 5.39 |
| | LM1 | 8.00 | 6.92 | 6.83 | 6.11 | 5.39 |
| | LM2 | 18.37 | 14.98 | 13.10 | 11.23 | 8.77 |
| $\beta_1 = 2, \quad \beta_2 = 0.5$ | WA | 10.19 | 8.83 | 8.07 | 7.34 | 6.00 |
| | WB | 6.46 | 5.54 | 5.71 | 5.61 | 4.88 |
| | CA1 | 7.72 | 6.49 | 6.34 | 6.16 | 5.11 |
| | CA2 | 11.14 | 9.39 | 8.43 | 7.61 | 6.20 |
| | OS1 | 7.69 | 6.45 | 6.39 | 6.12 | 5.09 |
| | OS2 | 7.87 | 6.54 | 6.38 | 6.13 | 5.10 |
| | LM1 | 7.69 | 6.46 | 6.39 | 6.12 | 5.10 |
| | LM2 | 10.27 | 7.87 | 7.30 | 6.74 | 5.27 |
| $\beta_1 = 1, \quad \beta_2 = 1$ | WA | 4.63 | 4.70 | 4.72 | 4.57 | 4.87 |
| | WB | 6.17 | 6.05 | 5.89 | 5.51 | 5.22 |
| | CA1 | 5.99 | 5.54 | 5.16 | 5.02 | 5.23 |
| | CA2 | 5.76 | 5.41 | 5.27 | 4.93 | 5.10 |
| | OS1 | 7.20 | 6.69 | 6.16 | 5.73 | 5.49 |
| | OS2 | 6.74 | 6.44 | 6.04 | 5.68 | 5.49 |
| | LM1 | 8.33 | 7.15 | 6.58 | 6.04 | 5.60 |
| | LM2 | 8.90 | 7.65 | 6.99 | 6.45 | 5.70 |

## Table 2

Percentage of Rejections in 10,000 Trials from Tests of Nominal Size 5 per cent Null Hypothesis False.

| DGP | Test | $n = 20$ | $n = 30$ | $n = 40$ | $n = 50$ | $n = 100$ |
|---|---|---|---|---|---|---|
| $\beta_1 = 1.5, \quad \beta_2 = 1$ | WA | 69.49 | 83.75 | 91.61 | 96.18 | 99.87 |
| | WB | 56.21 | 73.92 | 85.47 | 92.64 | 99.76 |
| | CA1 | 60.41 | 76.05 | 86.83 | 93.35 | 99.78 |
| | CA2 | 71.35 | 84.72 | 92.12 | 96.36 | 99.87 |
| | OS1 | 58.70 | 75.23 | 86.45 | 93.19 | 99.77 |
| | OS2 | 60.59 | 76.14 | 86.87 | 93.41 | 99.77 |
| | LM1 | 59.07 | 75.34 | 86.52 | 93.21 | 99.77 |
| | LM2 | 64.14 | 78.59 | 88.15 | 94.14 | 99.79 |
| $\beta_1 = 1, \quad \beta_2 = 0.5$ | WA | 25.16 | 46.00 | 68.30 | 84.84 | 99.94 |
| | WB | 85.56 | 94.71 | 97.90 | 99.25 | 99.99 |
| | CA1 | 78.92 | 92.71 | 97.18 | 99.05 | 99.99 |
| | CA2 | 28.94 | 50.06 | 71.25 | 86.56 | 99.94 |
| | OS1 | 83.85 | 93.91 | 97.45 | 99.09 | 99.99 |
| | OS2 | 67.77 | 85.30 | 92.87 | 96.12 | 99.79 |
| | LM1 | 85.76 | 95.43 | 97.72 | 99.19 | 99.99 |
| | LM2 | 88.14 | 95.41 | 98.09 | 99.34 | 99.99 |

In addition to the two Wald tests computed by GV, called WA and WB in the Table, three sets of easily computed statistics are shown, each set containing two statistics. The first set, called CA1 and CA2, contains two "naïve" $C(\alpha)$ tests, computed at the DGP's characterised by the parameters $(\hat{\beta}_1, 1/\hat{\beta}_1)$ and $(1/\hat{\beta}_2, \hat{\beta}_2)$. Then in the second set are the $C(\alpha)$ tests computed at the points obtained from the two used in the first set by a one-step efficient estimation. These statistics are called OS1 and OS2 in the Table. Finally, since the points used in the second set are asymptotically equivalent to maximum likelihood estimates, the third set contains two LM test statistics computed just as if they actually were ML estimates. The statistics in the third set are therefore always numerically greater than those in the second set, by construction. In Table 2 the tests are tried out on two DGP's that do not satisfy the null hypothesis under test.

The results confirm the intuition about the $C(\alpha)$ tests computed at the one-step efficient estimates. In Table 1, one sees by comparing the entries in the various lines OS1 and OS2 that the two statistics yield very similar inferences indeed. In Table 2, as one might expect, their performances are not so similar, especially in very small samples. The first two lines in each battery of tests correspond to the two tests considered by GV. The extreme dependence of the test results on the way in which the restrictions are formulated

is again clearly visible. The next two lines, with the naïve $C(\alpha)$ statistics, confirm the point made above that, although these tests are invariant to model parametrisation and to the formulation of the restrictions, they are affected by the choice of the point at which they are evaluated to at least the same extent as is the Wald test by the way in which the restrictions are formulated. The last two lines, containing the results yielded by the pseudo LM tests, show that these, too, must be treated with caution. Although one or the other of the one-step estimates is often close enough to the true restricted MLE for there to be no appreciable difference between the $C(\alpha)$ test and the pseudo LM test, both virtually never are at the same time, and sometimes both give bad pseudo LM tests.

In contrast, the OS tests not only agree rather well between themselves, but yield very acceptable inference by any standards. Remember that no degrees-of-freedom correction has been applied even in samples of size no more than 20. The slight over-rejection observed is therefore in no way objectionable. It would appear, then, that at very least for testing nonlinear restrictions on linear models, a satisfactory alternative to the Wald test is an OS test, performed by using the unrestricted MLE to obtain some, rather arbitrary, DGP in the null model, then taking one step towards an asymptotically efficient estimate by means of an artificial linear regression, and finally computing a $C(\alpha)$ test at the efficient estimate. The procedure involves no nonlinear procedures at all, and even if the unrestricted model needed to be estimated nonlinearly, that one estimation would be the only nonlinear estimation needed in order to perform a test. The proposed OS test does not absolutely respect the Wald principle of using *only* information about the unrestricted MLE – we saw that that was incompatible with invariance – and it is not absolutely invariant either, but its computation requires little more than knowledge of the unrestricted MLE, and its lack of invariance is not very troubling, at least in the example considered, one which, up till now, has been thought of as exemplifying the worst aspects of the lack of invariance of the conventional Wald test.

## 6 Conclusion.

A survey of differential-geometric methods as applied to econometrics has been given and applied to the long-standing problem of the lack of invariance of the Wald test under reparametrisations of the restrictions under test. A very large variety of intrinsic, and hence invariant, possible test statistics has been unearthed, some simple to compute, and others distinctly awkward. The geometrical analysis shows clearly what the reasons are for the lack of invariance of the Wald test. It turns out that the less well-known $C(\alpha)$ test is also subject to a lack of invariance, but for different reasons. At the root of all the difficulties however is the fact that any *projection* (defined in whatever intrinsic fashion we please) of an unrestricted MLE on to a null model is in general hard to compute, in the sense that it involves as much work as estimating the null model.

The "solution" proposed here is to use what I provisionally dub OS tests, as described in the last section. These tests have an intuitive justification, and, in the small Monte Carlo study, show themselves to behave remarkably well in circumstances that lead the

conventional Wald test to misbehave quite remarkably. Even if OS tests are not exclusively based on information provided by the unrestricted estimates, and if they are not genuinely invariant, they are still simple to compute once the unrestricted MLE is known, and do not suffer very severely from lack of invariance. They appear, in short, a viable alternative to the Wald test in circumstances in which the latter is suspected of yielding faulty inference in reasonably sized samples.

## References.


Amari, S.-I. (1985) *Differential Geometrical Methods in Statistics*, New York, Springer-Verlag.

Barndorff-Nielsen, O.E., D.R. Cox, N. Reid (1986) "The role of differential geometry in statistical theory," *International Statistical Review*, **54**, 83–96.

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

Critchley, F., P. Marriott, M. Salmon (1990) "On the differential geometry of the Wald test with nonlinear restrictions", Working Paper.

Dagenais, M.G., J.-M. Dufour, (1989) "Invariance, nonlinear models and asymptotic tests," C.R.D.E working paper, Université de Montréal.

Davidson, R., J.G. MacKinnon (1987) "Implicit alternatives and the local power of test statistics," *Econometrica*, **55**, 1305–1329.

Davidson, R., J.G. MacKinnon (1990) "Specification tests based on artificial regressions," *Journal of the American Statistical Association*, **85**, 220–227.

Davidson, R., J.G. MacKinnon (1991) "Artificial regressions and $C(\alpha)$ tests," *Economics Letters*, forthcoming.

Dawid, A.-P., (1975) Discussion of Efron (1975).

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

Efron, B. (1975) "Defining the curvature of a statistical problem," *Annals of Statistics*, **3**, 1189-1242.

Gregory, A.W., M.R. Veall (1985) "On formulating Wald tests for nonlinear restrictions," *Econometrica*, **53**, 1465–1468.

Kass, R.E. (1981) "The Geometry of Asymptotic Inference," Carnegie-Mellon University Technical Report No 215.

Lafontaine, F., K.J. White (1986) "Obtaining any Wald test you want," *Economics Letters*, **21**, 35–40.

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

Neyman, J. (1959) "Optimal asymptotic tests of composite statistical hypotheses," in U. Grenander, ed., *Probability and Statistics*, Wiley, New York.

Phillips, P.C.B., J.Y. Park (1988) "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, **56**, 1065–1083.

Smith, R.J., (1987) "Alternative asymptotically optimal tests and their application to dynamic specification," *Review of Economic Studies*, **54**, 665–680.