THE UNIVERSITY *of York*

Discussion Papers in Economics

No. 2004/07

Some Geometry for the Maximal Invariant in Linear Regression

by

Patrick Marsh

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Some Geometry for the Maximal Invariant in Linear Regression[1]

Patrick Marsh

Departments of Economics

University of York

Heslington, York

YO10 5DD

Tel. No. +44 (0)1904 433084

e-mail: pwnm1@york.ac.uk

March 1^{st} 2004

[1]AMS 1991 subject classifications. Primary- 62J05, 62F99. Secondary- 62B10

Summary

The maximal invariant forms the basis of a well established theory on hypothesis testing on the covariance structure in linear regression, see Lehman (1997). This paper examines the geometry of the maximal invariant. In particular it derives explicit expressions for both Fisher information and statistical curvature, see Efron (1975). The results apply for any sample size, for any sufficiently differentiable covariance structure and across a variety of sample densities. The results are illustrated for regressions involving autoregressive and moving average errors. Specifically, the effects of different specifications of the mean and of non-stationarity and non-invertibility may be quantified.

Some key words: Differential geometry, Efron curvature, Fisher Information, Maximal invariant.

1 Introduction

Inference upon covariance structures forms a large part of the statistical analysis of linear models. Particular cases, such as tests for serial correlation, Durbin and Watson (1950, 1951), tests for heteroskedasticity, Glejser (1969) and tests for unit roots, Sargan and Bhargava (1983) and Dufour and King (1991), have each generated sizeable literatures of their own. This paper examines the geometric properties of linear models having such covariance structures.

Geometric methods have provided many key insights into the problems of parametric inference. Rao (1945) introduced a Riemannian metric based upon Fisher information, used to assess the efficiency of maximum likelihood estimation. Development of this idea has lead to the construction of a comprehensive differential geometric framework, for example, the expected geometry of Amari (1990), see also Critchley, Marriott and Salmon (1994). Closely related is the concept of statistical curvature, which plays a crucial role in predicting the efficacy of linear methods such as linear asymptotic approximations (Efron (1975)) or locally most powerful testing procedures (Kallenberg (1981)).

Here we suppose that the data, $y = (y_1, ..., y_N)'$ is modelled according to

$$E[y] = X\beta \quad ; \quad Cov[y] = E[(y - X\beta)(y - X\beta)'] = \sigma^2\Sigma(\rho),$$

where X is a $N \times k$ matrix of covariates, β a $k \times 1$ vector of slopes, σ^2 a scalar variance and $\Sigma(\rho)$ an $N \times N$ positive definite covariance matrix depending upon a $d \times 1$ vector of parameters, ρ. Any hypothesis such that $\Sigma(\rho)$ is completely specified is invariant with respect to location-scale changes in the data. Consequently a desirable property for any test of such hypotheses, such as in the cases mentioned above, is invariance. That is the distribution of the test should not depend upon either β or σ^2 when the hypothesis is true. All invariant tests are functions of the data only through the maximal invariant, see Lehmann (1997, Chapter 6) and Lemma 1 below.

This paper derives some important differential geometric results for the maximal invariant. Under the conditions of Kariya (1980) we derive a recursive formula for obtaining the expectations of derivatives of the log-likelihood for the maximal invariant. The recursion is applied to obtain expressions for Fisher information and statistical curvature, Efron (1975). The expressions apply for any specification of the mean of the data, any sufficiently differentiable covariance structure, any sample size and over a family of sample distribution functions.

For illustrative purposes we will examine regressions involving either a linear or cyclic trend and covariances determined either by a first-order autoregression or moving average. These illustrations of the main results generalise previous results due to Ravishanker, Melnick and Tsai (1990) and van-Garderen (1999), as well as similar examples contained in Efron (1975) and Amari (1990). Specifically we may analyse the impact of the mean of the data on the geometry and crucially the impact of non-stationarity and non-invertibility, which in previous studies were assumed away.

The main results of this paper, the recursive formula and expressions for the information and curvature and some discussion of their key properties are presented in the next section. Section 3 illustrates the results in the combination of models mentioned above, specifically models with a cyclic or linear trend and with autoregressive or moving average covariance structure. For the purposes of clarity a graphical analysis is used, with the graphs provided in Appendix B.

2 Information and Curvature

Let y be a $N \times 1$ random vector and consider the following hypotheses:

$$\begin{aligned} H_0 &: Var[y] = \sigma^2 I \\ H_1 &: Var[y] = \sigma^2 \Sigma(\rho), \end{aligned} \tag{1}$$

where ρ is a scalar parameter and $\Sigma(\rho)$ an $N \times N$ positive definite symmetric matrix such that $\Sigma(0) = I_N$, we will assume the following.

Assumption 1 Let the density of y be $f(y; \beta, \sigma^2, \rho) = f(y) \in \mathcal{F}(\Sigma)$ with

$$\mathcal{F}(\Sigma) = \left\{ f : f(y; \beta, \sigma^2, \rho) = \left| \sigma^2 \Sigma(\rho) \right|^{-1/2} q\left[(y - X\beta)'(\sigma^2 \Sigma(\rho))^{-1}(y - X\beta) \right] \right\},$$

where X is a $N \times k$ matrix of covariates and β is a $k \times 1$ vector of parameters. Furthermore, we assume q is a nonincreasing convex function on $[0, \infty)$, so that $\mathcal{F}(\Sigma)$ includes, for example, contaminated Normal distributions, the multivariate t-distribution, including the multivariate Cauchy.

Under Assumption 1, defining the $N \times N - k$ matrix C by

$$CC' = M_X = I - X(X'X)^{-1}X' \quad ; \quad C'C = I_{N-k},$$

the $(N - k) \times 1$ vector $w = C'y$, and the group G with action

$$y \to ay + Xg \, , \, \beta \to a\beta + g \text{ and } \sigma^2 \to a^2\sigma^2,$$

then, summarising the results of Kariya (1980) we have the following Lemma:

Lemma 1 *(i)The maximal invariant, under G, for testing H_0 in (1) is*

$$v = w/|w| = C'y/(y'M_Xy)^{1/2}$$

and all invariant tests of H_0 are functions of y only through v.
(ii) Under H_0 v is uniformly distributed on the surface of the unit sphere in $\mathbb{R}^{N-k}$, say $v \sim \mathbb{U}(\mathbb{S}_{N-k})$, while under H_1 the density of v on $\mathbb{S}_{N-k}$ is

$$pdf(v|\rho) = c_N \det A^{-1/2} \left(v'A^{-1}v\right)^{-(N-k)/2}, \tag{2}$$

where $A = C'\Sigma(\rho)C$ and $c_N = \Gamma(\frac{N-k}{2})/2\pi^{(N-k)/2}$.

Amari's (1990) expected geometry, Fisher information and the statistical curvature of Efron (1975) are derived from derivatives of the expectations of functions of derivatives of the log-likelihood. In this case we require the expectations of derivatives of the log of (2). Before proceeding, let $U = (u_1, u_2, .., u_{N-k})$ be a $N-k$ square orthogonal matrix such that $U'A^{-1}U = \Lambda = diag\{\lambda_i\}$, where the λ_i are the (ordered) eigenvalues of A^{-1} and satisfying

$$A^{-1}u_i = \lambda_i, \quad u_i'u_i = 1, \quad u_i'u_j = 0 \;\; \forall i \neq j.$$

As a consequence, the logarithm of (2) may be written as

$$L_v(\rho) = c_N^* + \frac{1}{2}\ln\det\Lambda - \frac{N-k}{2}\ln v'\Lambda v. \tag{3}$$

The eigenvalues λ_i and eigenvectors u_i are functions of ρ through A. According to Theorem 7 of Magnus & Neudecker (1999, p. 158) both are infinitely differentiable functions on a neighbourhood $\mathbb{N}(A_0^{-1}) \subset \mathbb{R}^{(N-k)\times(N-k)}$ in which,

$$\lambda_i(A_0^{-1}) = \lambda_i \quad \& \quad u_i(A_0^{-1}) = u_i.$$

The first two derivatives of each λ_i are given by

$$\begin{aligned} d_\rho^1\lambda_i &= u_i'\left(d_\rho A^{-1}\right)u_i, \\ d_\rho^2\lambda_i &= 2u_i'\left(d_\rho A^{-1}\right)\left(\lambda_i I_{N-k} - A^{-1}\right)^+\left(d_\rho A^{-1}\right)u_i \end{aligned}$$

where $d_\rho A^{-1}$ is the derivative of A^{-1} and Ξ^+ denotes the Moore-Penrose inverse of the matrix Ξ. It is clear that expectations of derivatives of (3) will involve calculation of expectations of the form

$$E_S^T = E_v \left[\prod_{s_i} \left(\frac{v'\Lambda_{(s_i)}v}{v'\Lambda v} \right)^{t_i} \right], \tag{4}$$

$\Lambda_{(s_i)}$ denotes the s_i-th derivative of Λ, S and T are indices denoting, respectively, which derivatives of Λ are involved and to what integer power (t_i) each quadratic form around $\Lambda_{(s_i)}$ should be taken. To illustrate the notation;

$$E_{(1,2,5)}^{(3,2,1)} = E_v \left[\left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v} \right)^3 \left(\frac{v'\Lambda_{(2)}v}{v'\Lambda v} \right)^2 \left(\frac{v'\Lambda_{(5)}v}{v'\Lambda v} \right) \right],$$

and so on. Defining a sequence of ratios of quadratic forms $\{r_l\}$, $l = 1, 2, ..., n = |T| = \sum t_i$, so that each of the $v'\Lambda_{(s_i)}v/\left(v'\Lambda v\right)$ appears with multiplicity t_i, then E_S^T can be evaluated via the following Lemma, proved in Appendix A.

Lemma 2 *Let $x \sim N(0, I_{N-k})$, $\bar{\Lambda}_{(l)} = \Lambda^{-1/2}\Lambda_{(l)}\Lambda^{-1/2}$ and μ^n be the n^{th} raw moment of a $\chi^2_{(N-k)}$ random variable and define*

$$Q_l = x'\bar{\Lambda}_{(l)}x,$$

then a recurrence relation for E_S^T is given by

$$E_S^T = \frac{1}{\mu^n} \left[E(Q_1)E\left(\prod_{l=2}^{n} Q_l \right) + E\left(W_n\right) \right],$$

where $n = \sum t_i$,

$$W_n = 2\sum_{l=2}^{n} \left(E\left[x'V_l x\right] \prod_{m=l+1}^{n} Q_m \right)$$

and $V_l = \Lambda^{-1/2}\Lambda_{(1)}\Lambda^{-1/2}xx'....x'\Lambda^{-1/2}\Lambda_{(j)}\Lambda^{-1/2}$.

Of primary focus in this paper are expressions for Fisher information and Efron curvature, although in principle much of Amari's (1990) expected geometry may be calculated from Lemma 2. To see this write

$$2(v'\Lambda v)\frac{dL_v\left(\rho\right)}{d\rho} + (N-k)(v'\Lambda_{(1)}v) = (v'\Lambda v)Tr[\Lambda_{(1)}\Lambda^{-1}]$$

and apply Leibnitz's formula to both sides, giving

$$
\begin{aligned}
& 2\sum_{r=0}^{n} {}^{n}C_r(v'\Lambda_{(n-r)}v)\frac{d^{r+1}L_v(\rho)}{d\rho^{r+1}} + (N-k)(v'\Lambda_{(n+1)}v) \\
= & \sum_{r=0}^{n} {}^{n}C_r(v'\Lambda_{(n-r)}v)Tr[\sum_{s=0}^{r} {}^{r}C_s\Lambda_{(r-s+1)}(-1)^s s!(\Lambda^{-1}\Lambda_{(1)})^s\Lambda_{(s+1)}],
\end{aligned}
\tag{5}
$$

where ${}^{n}C_r$ is the Binomial coefficient. Inversion of (5) then yields a recursion for the derivatives of the log-likelihood involving of products of ratios having expectations given in Lemma 2. The expected geometry then follows from the expectations of products of these derivatives and from derivatives of those expectations.

Efron (1975) defines statistical curvature via the covariance between the first and second derivatives of the log-likelihood. Specifically, for $S_v(\rho) = dL_v(\rho)/d\rho$ and $H_v(\rho) = d^2L_v(\rho)/d\rho^2$ let

$$
M_\rho = Cov[S_v(\rho), H_v(\rho)] = \begin{pmatrix} E_v\left[S_v(\rho)^2\right] & E_v\left[S_v(\rho)H_v(\rho)\right] \\ E_v\left[H_v(\rho)S_v(\rho)\right] & Var\left[H_v(\rho)\right] \end{pmatrix},
$$

and Fisher information $I_v(\rho)$ and Efron curvature $\gamma_v(\rho)$ are then defined by

$$
\begin{aligned}
I_v(\rho) &= E_v\left[S_v(\rho)^2\right] = -E_v\left[H_v(\rho)\right] \\
\gamma_v(\rho) &= \left(\frac{\det[M_\rho]}{I_v(\rho)^3}\right)^{1/2} = \left(\frac{Var\left[H_v(\rho)\right]}{I_v(\rho)^2} - \frac{E_v\left[S_v(\rho)H_v(\rho)\right]^2}{I_v(\rho)^3}\right)^{1/2}.
\end{aligned}
\tag{6}
$$

The following Theorem, again proved in Appendix A, provides first a computationally convenient form for the information and second the relevant quantities involved in calculation of the curvature, in terms of the recursions for E_S^T.

Theorem 1 *(i) Let $I_v(\rho)$ be the (Fisher) information in v about ρ, then*

$$
I_v(\rho) = \frac{N-k}{2(N-k+2)}Tr\left[\left(A^{-1}\bar{A}\right)^2\right] - \frac{1}{2(N-k+2)}\left[Tr(A^{-1}\bar{A})\right]^2, \tag{7}
$$

where A is defined in Lemma 1, $\bar{A} = C'd_\rho\Sigma(\rho)C$, and $d_\rho\Sigma(\rho)$ is the derivative of the covariance matrix.

(ii) In terms of the expectations E_S^T the covariance between $S_v(\rho)$ and $H_v(\rho)$ and the Variance of $H_v(\rho)$ are given by

$$
\begin{aligned}
E_v\left[H_v(\rho)\,S_v(\rho)\right] &= \frac{(N-k)^2}{4}\left[E_{1,2}^{1,1} + E_1^1 E_1^2 - E_1^1 E_2^1 - E_1^3\right] \\
Var\left[H_v(\rho)\right] &= \frac{(N-k)^2}{4}\left[E_2^2 + E_1^4 + 2E_2^1 E_1^2 - \left(E_2^1\right)^2 - \left(E_1^2\right)^2 - 2E_{1,2}^{2,1}\right]. \blacksquare
\end{aligned}
$$

In order to obtain explicit representations for the curvature, Appendix A provides expressions for E_S^T with those particular values of S and T required. Before proceeding, though, it is worth highlighting some key properties of the expressions for the information and curvature.

1) Invariance: The maximal invariant has, by construction, distribution independent of the nuisance parameters β and σ^2. In fact, the quantities given in Lemma 1 and Theorem 1 obey far wider invariance principles. First, the expressions apply for any member of the family specified in Assumption 1, indeed this is exploited to compute the expectations required via expectations of Gaussian quadratic forms. Second, as noted by Efron (1975), the curvature is invariant to reparameterisations of the form $\rho \to \xi(\rho)$. On the other hand the information obeys a form of invariance related to the parameterisation of the null hypothesis in (1).

Specifically, v is invariant to transformations of the data, as in $y \to y_0 = Ty$ for non-singular T, and $f(y_0) \subset \mathcal{F}(T\Sigma T')$. Since $v = C'y/||C'y||$, then $y \to y_0$ implies a transformation on S_{N-k}, $\{S(v) : v \to v_0 = C_0'y_0/|C_0'y_0|\}$, where C_0 is the decomposition of the symmetric idempotent formed from TX. According to Lemma 4.1 of Kullback (1997, Section 1.4) the information is invariant with respect to $S(v)$, that is

$$
I_v(\rho) = \int_{v'v=1} \frac{d^2 \ln pdf(v)}{d\rho^2}(dv) = \int_{v_0'v_0=1} \frac{d^2 \ln pdf(v_0)}{d\rho^2}(dv_0) = I_{v_0}(\rho).
$$

As an immediate consequence, as far as the information in v is concerned, it is irrelevant whether we are considering hypotheses parameterised by

$$
H_0 : \Sigma(\rho = 0) = I_N \quad \text{or} \quad H_0 : T_{\rho_0}'\Sigma(\rho = \rho_0)T_{\rho_0} = I_N,
$$

that is the information in v under the alternative is independent of the null.

2) Decomposition of information (and information loss): Construction of the maximal invariant implies a transformation and decomposition of the data according to

$$y \to \begin{pmatrix} v = C'y/||C'y|| \\ \hat{\beta} = (X'X)^{-1}X'y \\ s^2 = y'M_X y \end{pmatrix},$$

which implies a factorisation of the joint sample density and hence decomposition of the information, viz

$$I_y\left(\beta, \sigma^2, \rho\right) = I_v\left(\rho\right) + I_{s^2|v}\left(\rho, \sigma^2\right) + I_{\hat{\beta}|(v,s^2)}\left(\rho, \sigma^2, \beta\right).$$

Within families satisfying Assumption 1, for the sample itself, and the statistic $w = C'y \sim \mathcal{F}(A)$, where $A = C'\Sigma(\rho)C$, the information is found to be

$$I_y\left(\beta, \sigma^2, \rho\right) = \begin{pmatrix} X'\Sigma^{-1}X/\sigma^2 & 0 & 0 \\ 0 & Tr\left[(\Sigma^{-1}d_\rho\Sigma)^2\right]/2 & Tr\left[\Sigma^{-1}d_\rho\Sigma\right]/2\sigma^2 \\ 0 & Tr\left[(\Sigma^{-1}d_\rho\Sigma)\right]/2\sigma^2 & N/2\sigma^4 \end{pmatrix}, \quad (8)$$

$$I_w\left(\sigma^2, \rho\right) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & Tr\left[(A^{-1}\bar{A})^2\right]/2 & -Tr\left[(A^{-1}\bar{A})\right]/2\sigma^2 \\ 0 & -Tr\left[(A^{-1}\bar{A})\right]/2\sigma^2 & (N-k)/2\sigma^4 \end{pmatrix} \quad (9)$$

where $\Sigma = \Sigma(\rho)$. The information loss in constructing the maximal invariant, i.e. $I_{s^2|v}\left(\rho, \sigma^2\right) + I_{\hat{\beta}|(v,s^2)}\left(\rho, \sigma^2, \beta\right)$ can be calculated from

$$I_{s^2|v}\left(\rho, \sigma^2\right) = I_w\left(\sigma^2, \rho\right) - I_v\left(\rho\right) \quad \text{and} \quad I_{\hat{\beta}|(v,s^2)}\left(\rho, \sigma^2, \beta\right) = I_y\left(\beta, \sigma^2, \rho\right) - I_w\left(\sigma^2, \rho\right).$$

Finally, noting (8), $I_y\left(\beta, \sigma^2, \rho\right)$ is seen to be block-diagonal. Moreover, the (ρ, σ^2)-block does not depend upon X. Basing a geometry upon the metric $I_y\left(\beta, \sigma^2, \rho\right)$ might therefore lead to the misleading conclusion that the properties of statistical hypotheses on the covariance do not depend upon the mean of the data.

3) Extensions to multivariable covariances: Let $\rho = (\rho_1, .., \rho_d)$, $\Lambda_{(i)}$ denote the derivative of Λ with respect to ρ_i, and $\Lambda_{(i,j)}$ be the second derivative with respect to ρ_i and ρ_j, then the score vector and Hessian derived from the density (2) are

$$\begin{aligned} S_v(\rho) &= \left\{ \frac{Tr[\Lambda^{-1}\Lambda_{(i)}]}{2} - \frac{(N-k)}{2}\frac{v'\Lambda_{(i)}v}{v'\Lambda v} \right\}, \\ H_v(\rho) &= \left\{ \frac{Tr[\Lambda^{-1}(\Lambda_{(i)}\Lambda^{-1}\Lambda_{(j)} + \Lambda_{(i,j)})]}{2} - \frac{(N-k)}{2}\left[\frac{v'\Lambda_{(i,j)}v}{v'\Lambda v} - \frac{v'\Lambda_{(i)}v}{v'\Lambda v}\frac{v'\Lambda_{(j)}v}{v'\Lambda v}\right] \right\}, \\ \text{for } i,j &= 1,2,...,d. \end{aligned}$$

Neither Efron curvature, nor the curvature of Amari (1982) applies directly outside the exponential family, in the multivariable case. Although in some cases the density of the data y will be curved exponential, in general the density of the maximal invariant will not. However, a simple scalar curvature measure may be obtained directly from the recursive expectations of Lemma 1. Specifically, let $g_{ij} = \{I_v(\rho)\}_{i,j}$ and g^{ij} be its inverse, then the metric connection is defined by

$$\Gamma^m_{ij} = \frac{1}{2}g^{km}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right),$$

where $\partial_i = \partial(.)/\partial\rho_i$ and the summation convention is assumed. Since we can write,

$$g_{ij} = \frac{(N-k)Tr\left[(A^{-1}\partial_i A)\,(A^{-1}\partial_j A)\right] - Tr(A^{-1}\partial_i A)Tr(A^{-1}\partial_j A)}{2\,(N-k+2)},$$

then the connection is obtained via straightforward differentiation. Covariant differentiation of the connection yields the Riemann-Christoffel tensor, viz

$$R^m_{ijk} = \partial_j\Gamma^m_{ik} - \partial_k\Gamma^m_{ij} - \Gamma^r_{ik}\Gamma^m_{rj} - \Gamma^r_{ij}\Gamma^m_{rk},$$

and finally the Ricci curvature κ is calculated via contraction of the Reimann-Christoffel tensor, i.e.

$$\kappa = g^{uv}R^m_{umv}. \tag{10}$$

It should be borne in mind, however, that (10) is an entirely geometric measure. Without specifying the distribution of y more precisely, specifically the covariance structure itself, an embedding curvature measure such as that of Efron (1975) would prove difficult to justify, other than on a case by case basis.

In the following Section, we refer back to the idea of testing that $\Sigma(\rho)$ has a particular structure and explore the related concepts of information, distance and curvature in testing hypotheses such as (1). First, we explore the geometry of autoregressive and moving average covariances. Second, we will focus on the interplay between unit roots and linear trends, in particular upon their impact upon the available information to test the unit root hypothesis.

3 Illustration and Analysis

3.1 The Geometry of Autoregressions & Moving Averages

For the purposes of illustration we will consider two very simple regression models, viz

$$M_1 : y_i = \beta_1 + \beta_2 \sin(2\pi i/N) + \upsilon_i \quad ; \quad M_2 : y_i = \beta_1 + \beta_2(i/N) + \upsilon_i, \qquad (11)$$

and two possible models of serial correlation,

$$a_1 : \upsilon_i^{(1)} = \rho\upsilon_{i-1}^{(1)} + \varepsilon_i \quad ; \quad a_2 : \upsilon_i^{(2)} = \rho\varepsilon_{i-1} + \varepsilon_i, \qquad (12)$$

where $i = 1, 2, .., N$, the $\varepsilon_i \sim i.i.d(0, \sigma^2)$ are such that $y = (y_1, ..., y_N)$ satisfies Assumption 1, so that we consider two different covariance structures, i.e. autoregressive (a_1) and moving average dependence (a_2). We will fix $N = 25$ for all the illustrations, and will assume $\varepsilon_0 = 0$, so that all cases, including non-stationary/non-invertible processes, may be considered.

The precise quantities under consideration will be;

$$i) \sqrt{I_v(\rho)} \qquad ii)\ r_v(\rho) = \frac{I_v(\rho)}{(I_y(\beta,\sigma^2,\rho))_{\rho,\rho}} \quad iii)\ \gamma_v(\rho)^2 \ .$$

The first is the differential of the 'information distance' $\int_{\rho_0}^{\rho} \sqrt{I_v(\theta)}d\theta$, as detailed in Kass (1989), which provides a relatively simple distance measure from some null value ρ_0 and any alternative, ρ. The second measures 'information loss' and is the ratio of the information in the maximal invariant to the information in the sample y about

ρ, as given by the relevant element in (8). The third is the square of Efron curvature and is used for comparison with values reported in Efron (1975) and van-Garderen (1999).

In Appendix B, Figures 1 (autoregression) and 2 (moving average) plot $\sqrt{I_v(\rho)}$ for both M_1 and M_2 over the range of values $\rho \in (-1.1, 1.1)$. Notice that the area under the graphs, and between any two points, in Figures 1 and 2 is the 'information distance' between those points. Figures 3 (autoregression) and 4 (moving average) plot $r_v(\rho)$ for both M_1 and M_2 and additionally for $X = \mathbf{0}$ (which is equivalent to our *knowing* the value of β), over the range of values $\rho \in (-1.25, 1.25)$. Finally, Figure 5 (autoregression) and plots $\gamma_v(\rho)^2$ for both M_1 and M_2 and additionally for the case that $X = 0$ over $\rho \in (-1.1, 1.1)$, while Figure 6 (moving average) plots the same but over the range $\rho \in (-0.65, 0.65)$.

From the graphs, generally, moving averages would seem to be much less sensitive to the different specifications of the mean. Autoregression is more sensitive, particularly so for values of ρ close to the non-stationary boundary, $\rho = 1$, but not $\rho = -1$. Most interesting is that with a linear trend (M_1) the information in an autoregression vanishes at $\rho = 1$. Figure 3 reveals that when there is a linear trend essentially none of the information in the sample is contained in the maximal invariant near $\rho = 1$, while with a cyclical trend more than 80% is available. Perhaps this provides some useful evidence as to why there are difficulties in testing for the presence of unit roots in the presence of trending regressors, see for example Bhargava (1996). At the origin the information is the same for both covariances, as one would expect.

More fundamental is the lack of symmetry in these geometric measures, which has not been revealed in the related examples of Efron (1975) and van Garderen (1999). Monte Carlo evidence due to King and Giles (1977) has pointed toward this, though. As expected, given the values for the information near $\rho = 1$, the curvature of autoregression becomes enormous near this point, Figure 5 is truncated, for that reason. Efron (1975) pointed to a value of $\gamma_v(\rho)^2$ of around 1/8, such that linear approximations become poorer for curvature larger than this value, see also Kallenberg (1981). For autoregression and since $\gamma_v(\rho) \to 0$ as $N \to \infty$, then for the cases here a sample size of $N = 100$ turns out to be sufficient for curvature to fall below this value, except at $\rho = 1$. The curvature for the moving average, Figure 6, is explained very simply by the following argument: Efron (1975) defines curvature to be zero for a linear exponential model. Autoregressions are curved exponential models, and moving averages may be approximated via

$$v_i = \rho\varepsilon_{i-1} + \varepsilon_i \quad \sim \quad v_i = \sum(-\rho)^j v_{i-j} + \varepsilon_i,$$

with the sum converging only if $|\rho| < 1$. The larger $|\rho|$ is the more terms are required for the approximation and hence the more 'curved' that approximating autoregression is.

Acknowledgments

Thanks are due for comments on this and related work, to Francesco Bravo, Giovanni Forchini, Peter Phillips and Robert Taylor, as well as an anonymous referee and participants at seminars at the Universities of Birmingham and York.

APPENDIX A (Proofs)

1) Proof of Lemma 2

Let

$$r_{s_i}^{t_i}(v) = \left(\frac{v'\Lambda_{(s_i)}v}{v'\Lambda v}\right)^{t_i},$$

so that $E_S^T(\rho) = E\left[\prod_{s_i} r_{s_i}^{t_i}(v)\right]$ and $z \sim N(0, \Lambda)$ so that if we define $z/|z| = v^*$, then

$$pdf(v^*; \rho) = c_N \det \Lambda^{1/2} \left((v^*)'\Lambda v^*\right)^{(N-k)/2},$$

and so

$$E\left[\prod_{s_i} r_{s_i}^{t_i}(v^*)\right] = E\left[\prod_{s_i} r_{s_i}^{t_i}(v)\right] = E\left[\prod_{s_i} r_{s_i}^{t_i}(z)\right],$$

where

$$r_{s_i}^{t_i}(z) = \left(\frac{z'\Lambda_{(s_i)}z}{z'\Lambda z}\right)^{t_i}.$$

Since both $\Lambda_{(s_i)}$ and Λ are diagonal then for $x = \Lambda^{-1/2}z \sim N(0, I_{N-k})$ and

$$r_{s_i}^{t_i}(x) = \left(\frac{x'\Lambda^{-1/2}\Lambda_{(s_i)}\Lambda^{-1/2}x}{x'x}\right)^{t_i},$$

we have

$$E\left[\prod_{s_i}\right] = E\left[\prod_{s_i} r_{s_i}^{t_i}(x)\right] = E\left[\frac{\prod_{s_i} x'\Lambda^{-1/2}\Lambda_{(s_i)}\Lambda^{-1/2}x}{(x'x)^n}\right],$$

where $n = \sum t_i$. Furthermore since $r_{s_i}^{t_i}(x)$ is independent of the length of x, $|x|$,(see Jones (1987)) then

$$E\left[|x|^{2n}\prod_{s_i} r_{s_i}^{t_i}(x)\right] = E[|x|^{2n}]E\left[\prod_{s_i} r_{s_i}^{t_i}(x)\right],$$

and so

$$\begin{aligned} E\left[\prod_{s_i} r_{s_i}^{t_i}(x)\right] &= E\left[\prod_{s_i} x'\Lambda^{-1/2}\Lambda_{(s_i)}\Lambda^{-1/2}x\right]/E[(x'x)^n] \\ &= \frac{1}{\mu^n}E\left[\prod_{s_i} x'\Lambda^{-1/2}\Lambda_{(s_i)}\Lambda^{-1/2}x\right]. \end{aligned}$$

Finally, write $\prod_{s_i} x'\Lambda^{-1/2}\Lambda_{(s_i)}\Lambda^{-1/2}x = \prod_{l=1}^{n} Q_l$, with the Q_l defined in the statement of the lemma, so that

$$E\left[\prod_{s_i} r_{s_i}^{t_i}(x)\right] = \frac{1}{\mu^n}E\left[\prod_{l=1}^{n} Q_l\right],$$

and the recurrence relation itself follows from Theorem 1 of Ghazal (1996).

2) Proof of Theorem 1

(i) The Score and the Hessian are immediately available from (3), yielding

$$\begin{aligned} S_v(\rho) &= \frac{dL_v(\rho)}{d\rho} = \frac{Tr[\Lambda_{(1)}\Lambda^{-1}]}{2} - \frac{(N-k)}{2}\frac{v'\Lambda_{(1)}v}{v'\Lambda v}, && (13) \\ H_v(\rho) &= \frac{d^2L_v(\rho)}{d\rho^2} \\ &= \frac{Tr[\Lambda_{(2)}\Lambda^{-1} - \left(\Lambda_{(1)}\Lambda^{-1}\right)^2]}{2} - \frac{(N-k)}{2}\left[\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right]. && (14) \end{aligned}$$

Before proceeding note that since $pdf(v|\rho)$ is a density the first Bartlett identity applies, i.e. $E[S_v(\rho)] = 0$, and hence rearranging (13) and taking expectations, we have

$$E_1^1 = E_v\left[\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right] = \frac{Tr[\Lambda_{(1)}\Lambda^{-1}]}{N-k}. \tag{15}$$

The information is given by

$$I_v(\rho) = \frac{1}{2}Tr[\Lambda_{(2)}\Lambda^{-1} - \left(\Lambda_{(1)}\Lambda^{-1}\right)^2] - \frac{N-k}{2}\left(E_2^1(\rho) - E_1^2(\rho)\right), \tag{16}$$

and substitution of the expressions for E_2^1 and E_1^2 from the Appendix, yields

$$I_v(\rho) = \frac{N-k}{2(N-k+2)}Tr\left[\left(\Lambda^{-1}\Lambda_{(1)}\right)^2\right] - \frac{1}{2(N-k+2)}\left[Tr[\Lambda^{-1}\Lambda_{(1)}]\right]^2.$$

In terms of the eigenvalues of A^{-1} and their derivatives, we have

$$I_v(\rho) = \frac{N-k}{2}\sum_{i=1}^{N-k}\left(\frac{d_\rho^1\lambda_i}{\lambda_i}\right)^2 - \frac{1}{2}\left[\sum_{i=1}^{N-k}\left(\frac{d_\rho^1\lambda_i}{\lambda_i}\right)\right]^2. \tag{17}$$

The derivative of A^{-1} is given by

$$d_\rho A^{-1} = -A^{-1}C'd_\rho\Sigma(\rho)CA^{-1} = -A^{-1}\bar{A}A^{-1},$$

and noting $A^{-1}u_i = \lambda_i u_i$ so that (17) may be rewritten as

$$I_v(\rho) = \frac{1}{2}\sum_{i=1}^{N-k}\left(u_i'A^{-1}\bar{A}u_i\right)^2 - \frac{1}{2(N-k)}\left[\sum_{i=1}^{N-k}u_i'A^{-1}\bar{A}u_i\right]^2. \tag{18}$$

We may write

$$\sum_{i=1}^{N-k}u_i'A^{-1}\bar{A}u_i = Tr\left[\sum_{i=1}^{N-k}u_i'A^{-1}\bar{A}u_i\right] = Tr\left[A^{-1}\bar{A}\sum_{i=1}^{N-k}U_i\right],$$

where the $U_i = u_iu_i'$ are symmetric idempotent matrices satisfying $\sum_{i=1}^{N-k}U_i = I_{N-k}$, thus

$$\left[\sum_{i=1}^{N-k}u_i'A^{-1}\bar{A}u_i\right]^2 = \left[Tr\left(A^{-1}\bar{A}\right)\right]^2. \tag{19}$$

Similarly,

$$\begin{aligned}\sum_{i=1}^{N-k}\left(u_i'A^{-1}\bar{A}u_i\right)^2 &= \sum_{i=1}^{N-k}Tr\left(u_i'A^{-1}\bar{A}u_iu_i'A^{-1}\bar{A}u_i\right) = Tr\left[\sum_{i=1}^{N-k}\left(A^{-1}\bar{A}U_i\right)^2\right] \\ &= Tr\left[\sum_{i=1}^{N-k}\left(A^{-1}\bar{A}\right)^2U_i\right],\end{aligned}$$

since U_i is idempotent and noting again that $\sum_{i=1}^{N-k} U_i = I_{N-k}$ we have

$$\sum_{i=1}^{N-k} \left(u_i' A^{-1} \bar{A} u_i\right)^2 = \left[Tr\left(A^{-1}\bar{A}\right)^2\right]. \tag{20}$$

Finally substitution of (19) and (20) into (18) proves the theorem. ■

(ii) If we note the first Bartlett identity (15) which implies

$$Tr\left[\Lambda_{(1)}\Lambda^{-1}\right] = (N-k)\, E_1^1,$$

and also since $E_v[S_v(\rho)] = 0$, then

$$\begin{aligned} E_v[S_v(\rho)\, H_v(\rho)] &= \frac{(N-k)^2}{4} E_v\left[\left(E_1^1 - \frac{v^1\Lambda_{(1)}v}{v'\Lambda v}\right)\left(\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right)\right] \\ &= \frac{(N-k)^2}{4}\left\{E_1^1\left(E_v\left[\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right]\right)\right. \\ &\quad \left. -E_v\left[\frac{v^1\Lambda_{(1)}v}{v'\Lambda v}\left(\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right)\right]\right\} \\ &= \frac{(N-k)^2}{4}\left[E_1^1\left(E_2^1 - E_1^2\right) - \left(E_{1,2}^{1,1} - E_1^3\right)\right], \end{aligned}$$

which upon rearranging gives the result.

For the variance of the Hessian, note that

$$\begin{aligned} Var[H_v(\rho)] &= Var\left[\frac{N-k}{2}\left(\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right)\right] \\ &= \frac{(N-k)^2}{4}\left\{E_v\left[\left(\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right)^2\right]\right. \\ &\quad \left. -E_v\left[\frac{v'\Lambda_{(2)}v}{v'\Lambda v} - \left(\frac{v'\Lambda_{(1)}v}{v'\Lambda v}\right)^2\right]^2\right\} \\ &= \frac{(N-k)^2}{4}\left(E_2^2 - 2E_{1,2}^{2,1} + E_1^4 - E_2^1 - E_1^2\right), \end{aligned}$$

which again immediately yields the result. ■

3) Resolved Recursions

Lemma 1 gives a recurrence relation for the expectation of powers ratios of quadratic forms in v. In order to derive the information and Efron curvature we require only the following few expressions, let $\mu_n = (N-k)(N-k+2)..(N-k+(2n-1))$, then for $i = 1, 2$,

$$E_i^1 = Tr[\Lambda_{(i)}\Lambda^{-1}]/\mu_1$$

$$E_i^2 = \left(Tr[\Lambda_{(i)}\Lambda^{-1}]^2 + 2Tr[(\Lambda_{(i)}\Lambda^{-1})^2]\right)/\mu_2,$$

$$E_i^3 = \left(Tr[\Lambda_{(i)}\Lambda^{-1}]^3 + 6Tr[(\Lambda_{(i)}\Lambda^{-1})^2]Tr[\Lambda_{(i)}\Lambda^{-1}] + 8Tr[(\Lambda_{(i)}\Lambda^{-1})^3]\right)/\mu_3$$

$$\begin{aligned} E_i^4 = {} & \left(Tr[\Lambda_{(i)}\Lambda^{-1}]^4 + 32Tr[(\Lambda_{(i)}\Lambda^{-1})^3]Tr[\Lambda_{(i)}\Lambda^{-1}] + 12Tr[\Lambda_{(i)}\Lambda^{-1}]^2 Tr[(\Lambda_{(i)}\Lambda^{-1})^2]\right. \\ & \left. +12Tr[(\Lambda_{(i)}\Lambda^{-1})^2]^2 + 48Tr[(\Lambda_{(i)}\Lambda^{-1})^4]\right)/\mu_4. \end{aligned}$$

while for the cross product terms required we have

$$E_{(1,2)}^{(1,1)} = \left(Tr[\Lambda_{(1)}\Lambda^{-1}]Tr[\Lambda_{(2)}\Lambda^{-1}] + 2Tr[(\Lambda_{(1)}\Lambda^{-1})(\Lambda_{(2)}\Lambda^{-1})]\right)/\mu_2,$$

$$\begin{aligned} E_{(1,2)}^{(2,1)} = {} & \left(Tr[\Lambda_{(1)}\Lambda^{-1}]^2 Tr[\Lambda_{(2)}\Lambda^{-1}] + 4Tr[(\Lambda_{(1)}\Lambda^{-1})]Tr[(\Lambda_{(1)}\Lambda^{-1})(\Lambda_{(2)}\Lambda^{-1})]\right. \\ & \left. +2Tr[(\Lambda_{(1)}\Lambda^{-1})^2]Tr[\Lambda_{(2)}\Lambda^{-1}] + 8Tr[(\Lambda_{(1)}\Lambda^{-1})^2(\Lambda_{(2)}\Lambda^{-1})]\right)/\mu_3\,. \end{aligned}$$

APPENDIX B (Figures)

Figure 1: $\sqrt{I_v(\rho)}$ plotted for M_1- dashed and M_2- dotted

for $\upsilon_t \sim AR(1)$

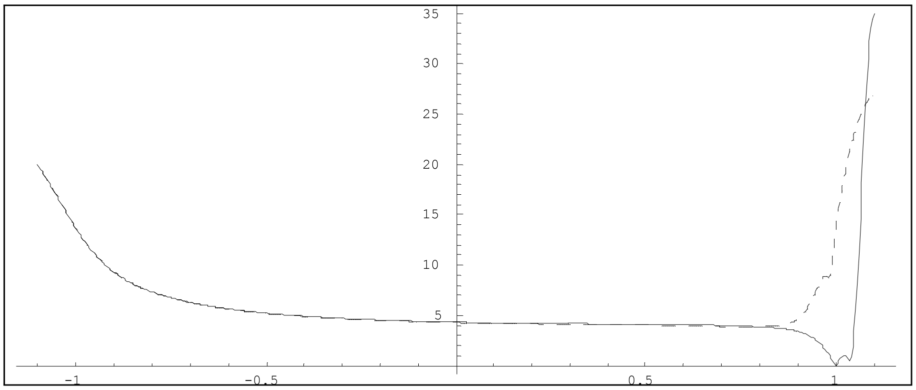


Figure 2: $\sqrt{I_v(\rho)}$ plotted for M_1- dashed and M_2- dotted

for $\upsilon_t \sim MA(1)$

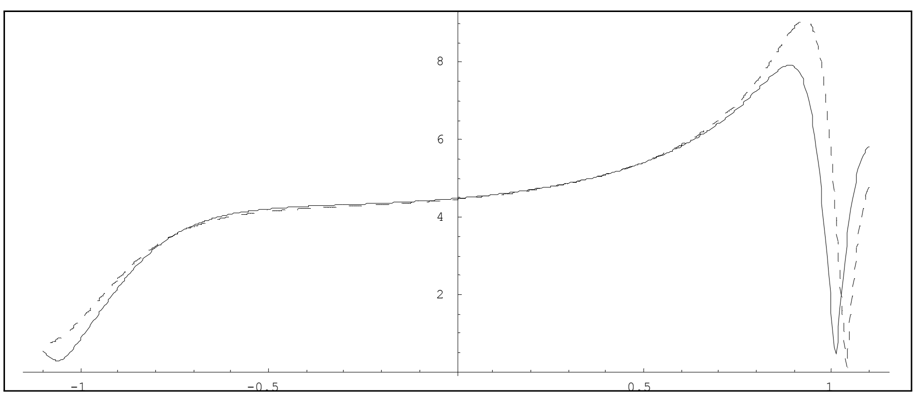


Figure 3: $r_v(\rho)$ plotted for M_1- dashed, M_2- dotted and $X = 0$- solid for $\upsilon_t \sim AR(1)$

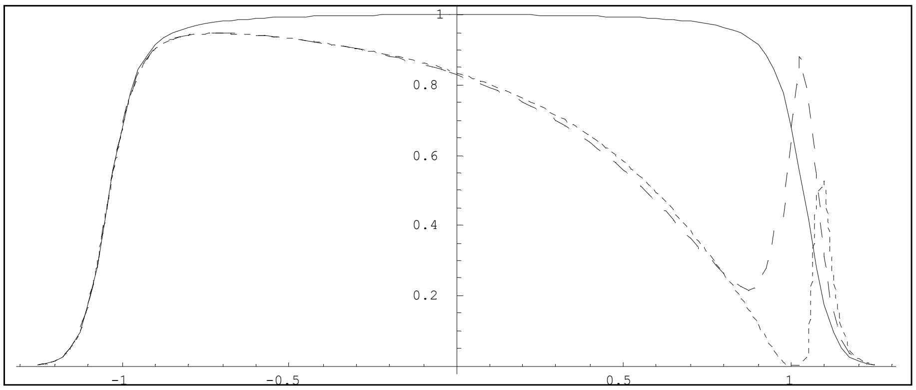


Figure 4: $r_v(\rho)$ plotted for M_1- dashed, M_2- dotted and $X = 0$- solid for $\upsilon_t \sim MA(1)$

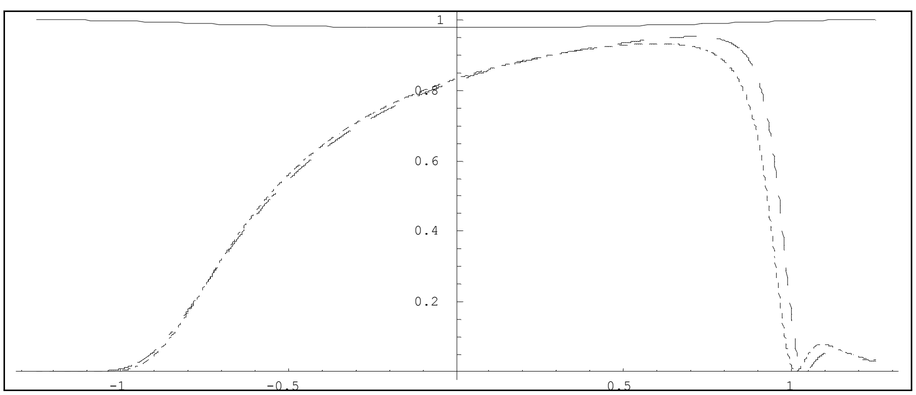


Figure 5: $\gamma_v(\rho)^2$ plotted for M_1- dashed, M_2- dotted and $X=0$- solid for $\upsilon_t \sim AR(1)$

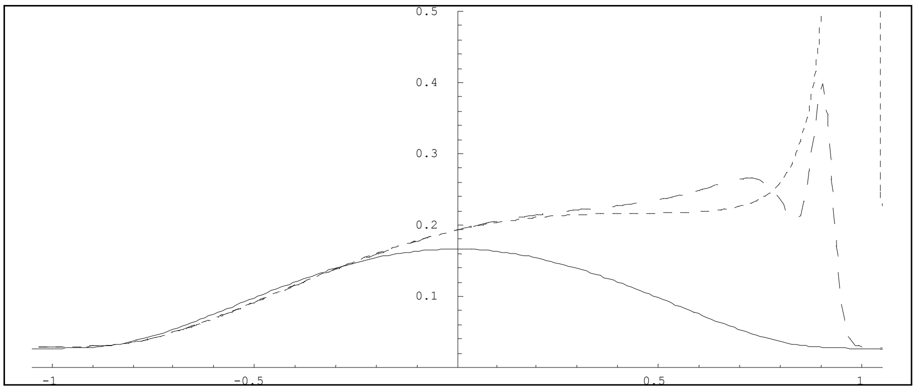
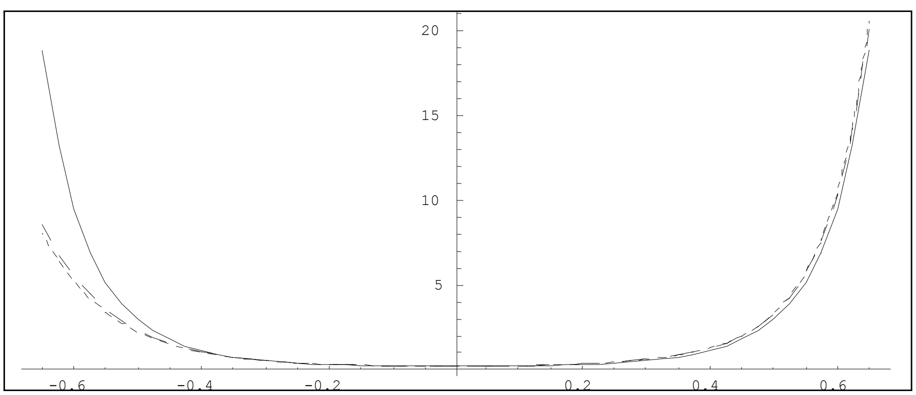


Figure 6: $\gamma_v(\rho)^2$ plotted for M_1- dashed, M_2- dotted and $X=0$- solid for $\upsilon_t \sim MA(1)$



References

AMARI, S.-I. (1982). Differential geometry of curved exponential families - curvature and information loss. *Annals of Statistics*, **10**, 357-385.

AMARI, S.-I. (1990). *Differential-Geometric Methods in Statistics,* 2nd ed. *Lecture Notes in Statistics* **28**. Springer, Berlin.

BHARGAVA, A. (1996). *Some Properties of Exact Tests for Unit Roots*, Biometrika, **83**, 944-949.

CRITCHLEY, F., MARRIOTT, P. and SALMON, M. (1994). Preferred Point Geometry and the Local Differential Geometry of the Kullback-Leibler Divergence. *Annals of Statistics*, **22**, 1587-1602.

DURBIN, J. and WATSON, G.B. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409-428.

DURBIN, J. and WATSON, G.B. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159-178.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, **3**, 1189-1242.

GHAZAL, G.A. (1996). Recurrence formula for expectations of products of quadratic forms. *Statistics and Probability Letters*, **27**, 101-109.

GLEJSER, H. (1969). A new test of heteroskedasticity. *Journal of the American Statistical Association,* **64**, 316-323.

JONES, M.C. (1987). On moments of ratios of quadratic forms in normal variables. *Statistics and Probability Letters*, **6**, 129-136.

KARIYA, T. (1980). Locally Robust Tests for Serial Correlation in Least Squares Regression. *Annals of Statistics*, **8**, 1065-1070.

KALLENBERG, W. C. M. (1981). The shortcoming of locally most powerful tests in curved exponential families. *Annals of Statistics*, **9**, 673–677.

KASS, R.E. (1989). The geometry of asymptotic inference. *Statistical Science*, **4**, 188-234.

KING, M. L. and GILES, D. E. A. (1977). A Note on Wallis' Bounds Test and Negative Autocorrelation, *Econometrica*, **45**, 1023-1026.

KULLBACK, S. (1997). *Information Theory and Statistics.* Dover, New York.

LEHMANN, E.L. (1997): *Testing Statistical Hypotheses* (2nd ed.), Springer texts in statistics.

MAGNUS, J.R. and NEUDECKER, H. (1999). *Matrix Differential Calculus, with Applications in Statistics and Econometrics* (revised edition), Chichester, Wiley.

RAO, C.R. (1945). Information and accuracy available in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81-89.

RAVISHANKER, N., MELNICK, E.L. and TSAI, C.-L. (1990). Differential geometry of ARMA models. *Journal of Time Series Analysis*, **11**, 259-274.

SARGAN, J.D. and BHARGAVA, A. (1983). Testing residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica*, **51**, 153-174.

van GARDEREN, K-J. (1999). Exact geometry of autoregressive models. *Journal of Time Series Analysis*, **20**, 1-21.