Econometric Theory, **25**, 2009, 1662–1681. Printed in the United States of America.
doi:10.1017/S0266466609990284

THE PROPERTIES OF KULLBACK–LEIBLER DIVERGENCE FOR THE UNIT ROOT HYPOTHESIS

PATRICK MARSH
University of York

The fundamental contributions made by Paul Newbold have highlighted how crucial it is to detect when economic time series have unit roots. This paper explores the effects that model specification has on our ability to do that. Asymptotic power, a natural choice to quantify these effects, does not accurately predict finite-sample power. Instead, here the Kullback–Leibler divergence between the unit root null and any alternative is used and its numeric and analytic properties detailed. Numerically it behaves in a similar way to finite-sample power. However, because it is analytically available we are able to prove that it is a minimizable function of the degree of trending in any included deterministic component and of the correlation of the underlying innovations. It is explicitly confirmed, therefore, that it is approximately linear trends and negative unit root moving average innovations that minimize the efficacy of unit root inferential tools. Applied to the Nelson and Plosser macroeconomic series the effect that different types of trends included in the model have on unit root inference is clearly revealed.

1. INTRODUCTION

The pathbreaking findings of Granger and Newbold (1974) have fueled econometric research for over three decades. It was demonstrated numerically, and later proved analytically in Phillips (1986), that regressions involving independent unit root processes give spurious results. The immediate implication was the necessity of building techniques to find which series had unit roots. In the original 1974 analysis only pure random walks were considered. Since then the specification of potential unit root series has broadened considerably to include the presence of deterministic components, such as trends, cycles, and breaks, and also underlying innovations with general covariance structures.

These features profoundly influence our ability to tell whether or not a series has a unit root. This paper measures this influence by detailing the numeric and

Thanks are due to Francesco Bravo, Graham Elliott, Giovanni Forchini, Bruce Hansen, Grant Hillier, Peter Phillips, Robert Taylor, two anonymous referees, and participants at seminars at the Universities of Birmingham and York and at the Conference in Honour of Paul Newbold at the University of Nottingham, Granger Centre for Time Series Econometrics. Address correspondence to Patrick Marsh, Department of Economics, University of York, Heslington, York, United Kingdom, YO10 5DD; e-mail: pwnm1@york.ac.uk.

 © 2009 Cambridge University Press 0266-4666/09 $15.00

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

analytic properties of the Kullback–Leibler divergence (henceforth *KL*). In the absence of any analytic expressions for distributions of any commonly used unit root test the literature has relied upon numerical methods to examine the impact of deterministic components and covariance structures. For details see, among many others, Durlauf and Phillips (1988), Perron (1989), Zivot and Andrews (1992), Elliott, Rothenberg, and Stock (1996), Leybourne, Mills, and Newbold (1998), and Phillips and Xiao (1998, Sect. 4). The impact of innovation covariances is most clearly seen in the numerical power comparisons of Perron and Ng (2001), that of trends in Harvey, Leybourne, and Taylor (2009), and that of trend breaks in Harris, Harvey, Leybourne, and Taylor (2009).

Here, clear evidence is presented that demonstrates that the asymptotic power envelope of Elliott et al. (1996) does not accurately predict finite-sample power. The power of a point optimal test predicted to be 0.5 by asymptotic power may in fact vary between 0.13 and 0.57, depending upon the deterministic component and correlation parameter, when the sample size is 100. Given that asymptotic power is not an accurate predictor of finite-sample power and that the latter is not analytically available, an alternative measure of the impact of model specification is required. In finite samples the numerical behavior of *KL* mirrors that of the power of a point optimal test. Both are seen to measure, in the same way, how our ability to discriminate between the unit root null and any alternative depends upon the specification of the model.

Because *KL* is available as a resolved expression it does not need to be simulated. Moreover, it is possible to detail its analytic properties as a function of both the degree of deterministic trending and innovation correlation structure, here modeled via a moving average. *KL* is proved to be differentiable and convex and thus minimizable with respect to these features. The results complement and extend the investigations of the specific impacts of sequences of polynomial trends in Phillips (1998) and linear trends in Marsh (2007). Examination of the minima helps explain why it is that a linear trend and a negative unit root moving average innovation process must inevitably imply low power.

In addition to forging an explicit link between model specification and inferential efficacy these results also have practical relevance. For each of the Nelson and Plosser (1982) macroeconomic series *KL* is estimated for three models. A general model estimates the degree of trending, and two restricted models impose a linear trend and no trend, respectively. The idea is simply to measure the effect of model specification on inferential efficacy. If, for a given series, *KL* is constant across each model, then the specification is seen not to be an influence. On the other hand if the value varies wildly across the models then this suggests that an inferential outcome may depend on the specification, rather than on whether there is, or is not, actually a unit root.

For example, for industrial production, *KL* takes the smallest observed value if no trend is included in the model, but it takes the second largest value if a trend is included; for velocity the almost exact opposite applies. For real wages, however, *KL* is close to zero if either a linear trend or no trend is included. However, if the

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

degree of trending is estimated in a more general model, *KL* is found to be far from zero. Moreover neither of the simpler models is a statistically justified restriction of the general model. These findings, not apparent from any previous study, help place into context the very different conclusions that various authors have come to regarding the presence, or not, of unit roots in these series. Justification for the consideration of noninteger trends is evidenced by the new findings that both wage series and the consumer price index (CPI) have significant nonlinear trends.

The plan for the paper is as follows. The next section derives *KL* for the unit root and details its theoretical and numerical links with power. Section 3 provides the analytic properties of *KL*, proving differentiability and convexity and evaluating its stationary points, Section 4 illustrates the results via the Nelson and Plosser (1982) data. Conclusions are followed by an Appendix containing the proofs.

2. KULLBACK–LEIBLER, POWER, AND THE SPECIFICATION OF THE MODEL

In this section it is first demonstrated that power and *KL* are both measures of the separation of sample densities. Then, when we have specified the model under consideration, *KL* is derived for the joint density of the residuals in a generalized least squares (GLS) detrending regression; its numerical behavior is compared with that of finite-sample power.

2.1. Measuring the Separation of Densities

The main idea of this paper can be summarized in the context of the following simple example. Suppose that a sample $y = (y_1, \ldots, y_T)'$ has joint density $f(y; \alpha, \phi, X)$, with α a scalar parameter, ϕ a vector of parameters, and X a matrix of covariates. Given observations on y we wish to perform the hypothesis test,

$$H_0 : \alpha = \alpha_0 \quad \text{vs.} \quad H_1 : \alpha = \alpha_1 \neq \alpha_0.$$

Even if $f(\cdot)$ is known, in general our ability to discriminate between H_0 and H_1 will depend upon both the nuisance parameters ϕ and the covariates X, neither of which are specified under either H_0 or H_1.

An obvious way of measuring this would be to compute the power of a point optimal test and subtract the size of the test to enable valid comparisons. Letting cr_δ be any critical region of size δ in $\mathbb{R}^T$, then optimal power minus size is given by

$$\Pi = \Pi(\alpha_0, \alpha_1, \phi, X) = \sup_{cr_\delta} \int_{\mathbb{R}^T} \mathbb{I}(y \in cr_\delta)[f(y; \alpha_1, \phi, X) - f(y; \alpha_0, \phi, X)]dy, \tag{1}$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

where $\mathbb{I}(\cdot)$ is the indicator function. For the unit root problem, as for many others in econometrics, we have no resolved expressions for this quantity. Moreover, asymptotic power, as in Elliott et al. (1996), is constructed so that it does not depend on nuisance parameters. As will be demonstrated subsequently, this may not give an accurate reflection of finite-sample power.

This paper instead examines Kullback–Leibler divergence, defined by

$$KL = KL(\alpha_0, \alpha_1, \phi, X) = \int_{\mathbb{R}^T} \log \left[\frac{f(y; \alpha_0, \phi, X)}{f(y; \alpha_1, \phi, X)} \right] f(y; \alpha_0, \phi, X)\, dy.$$

The value KL is then a measure of separation of the density functions from the perspective (or preferred point, as in Critchley, Marriott, and Salmon, 1994) of the null hypothesis. Similar to the empirical version utilized by Phillips and Ploberger (2003), when KL is small statistical methods will struggle to discriminate between the values α_0 and α_1. The purpose of the paper is to detail the differential and numerical properties of KL as a function of both ϕ and X, in the unit root case detailed subsequently. Notice that for any ϕ or X such that $KL = 0$, the parameter α is not identified.

Although KL has been predominantly used in the context of model selection, here we employ it as a direct measure of the difficulty of discriminating between H_0 and H_1. It is directly related to power. Following the analysis of Würtz (1997) and Gibbs and Su (2001), optimal power is bounded above by KL through

$$\Pi \leq \frac{1}{2} \int |f(y; \alpha_1, \phi, X) - f(y; \alpha_0, \phi, X)|\, dy \leq \sqrt{\frac{KL}{2}}. \tag{2}$$

As will be demonstrated in what follows, for the unit root problem specifically, KL is minimized by particular deterministic components and correlation structures. Immediately from (2) it is evident that minimizing KL implies power must also be small.

To deliver analytic expressions, rather than numerical results obtained only through Monte Carlo methods, here it will be assumed that $f(\cdot)$ is the Gaussian density. Specifically, if $y \sim N_T(\mu, \Sigma(\alpha))$, then

$$KL = \frac{1}{2} \left(\mathrm{Tr} \left[\Sigma(\alpha_1)^{-1} \Sigma(\alpha_0) \right] + \log \frac{|\Sigma(\alpha_1)|}{|\Sigma(\alpha_0)|} - T \right). \tag{3}$$

Although strictly applying only in the Gaussian case, qualitatively this will be informative more generally, in the same way that the asymptotic Gaussian power envelope of Elliott et al. (1996) is taken as a benchmark.

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

2.2. The Model, Assumptions, and Kullback–Leibler

In common with much of the recent literature, we consider a time series $(y_t)_{t=1}^T$ generated by the following three equations:

$$y_t = \mu_t + u_t, \tag{4}$$

$$u_t = \alpha u_{t-1} + \xi_t, \tag{5}$$

$$\xi_t = \phi(L)\varepsilon_t, \tag{6}$$

where $(\varepsilon_t)_{t=1}^T$ is a sequence of independently and identically distributed random variables with $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{E}[\varepsilon_t^2] = \sigma^2 < \infty$, $\phi(L)$ is a lag polynomial, and μ_t is a deterministic component. The purpose of the paper is to analyze the impact that the specification of (4) and (6) has on inference on the interest parameter α in (5). To do so we make the following assumption on the generating mechanism for y_t.

Assumption 1.

(a) $(\xi_t)_{t=1}^T$ is a stationary Gaussian process generated according to

$$\xi_t = \sum_{j=1}^{m} \phi_j \varepsilon_{t-j} + \varepsilon_t, \qquad \varepsilon_t \sim iidN(0, \sigma^2),$$

so that the coefficients ϕ_j are such that the roots of the polynomial $\phi(l) = 1 + l\phi_1 + \cdots + l^m\phi_m$ lie outside the unit circle.

(b) The process u_t is initialized at zero, i.e., $u_0 = 0$.

(c) The deterministic component is linear in a $k \times 1$ vector of variables x_t, with

$$\mu_t = \beta' x_t = \beta_1 + \beta_2 t^p + \beta_{(3)}' d_t,$$

where $p > 0$, $\beta_{(3)} = (\beta_3, \ldots, \beta_k)'$, $x_t = \left(1, t^p, d_t'\right)'$, and d_t is any $(k-2) \times 1$ deterministic variable satisfying

(a) $d_t = o(T^p)$ and (b) $\mu_t \neq 0$ for all $\beta \neq 0$.

Assumption 1 allows us to focus on the two key features of any potential unit root model: the degree of deterministic trending and the innovation correlation structure. Although a zero startup value is assumed, a nonzero exogenous y_0 could be incorporated via the setup in Marsh (2007) by including y_0 as a regressor, in X, only for the first observation in (4). Although this allows y_0 to be regressed out, this does not solve the inferential problem because any resulting test would have critical values depending upon y_0. However, under the provisions of Assumption 1 the focus will remain upon the most strongly trending regressor, t^p.

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

Justification for consideration of nonlinear trends of the form t^p follows from numerical evidence to be given subsequently for the Nelson and Plosser (1982) data, in which both wage series and CPI are found to have trends that are significant and nonlinear. Nonlinear trends with $p < 0$ are also explored in detail in Phillips (2007).

Under Assumption 1 the model (4)–(6) defines a generalized linear regression model (GLRM). To proceed, define the following $T \times 1$ vectors, $y = (y_t)_{t=1}^T$ and $\varepsilon = (\varepsilon_t)_{t=1}^T$, and let $X = \left(x_1', x_2', \ldots, x_T'\right)'$ and let $L^{(j)}$ define a lower triangular matrix with 1's on the jth lower diagonal and 0's elsewhere. By constructing the matrices

$$\Delta_\alpha = I_T - \alpha L^{(1)} \quad \text{and} \quad K_\phi = I_T + \sum_{j=1}^{m} L^{(j)} \phi_j,$$

we have the following GLRM for y:

$$y = X\beta + \Delta_\alpha^{-1} K_\phi \varepsilon \quad \text{and} \quad y \sim N\left(X\beta, \sigma^2 \Sigma_\phi(\alpha)\right), \tag{7}$$

where $\Sigma_\phi(\alpha) = \Delta_\alpha^{-1} K_\phi K_\phi' \left(\Delta_\alpha^{-1}\right)'$.

Notice that in (7) the parameters β and ϕ are nuisance parameters. Subsequent to the analysis of Dufour and King (1991) all recent tests of $H_0 : \alpha = 1$ are constructed so as to be invariant with respect to the group of transformations defined by $G_y : y \to \sigma y + X\beta$. This is invariably accomplished by "detrending" y via a regression based on (7). If ϕ is known (in practice ϕ will need to be estimated, as will be discussed later) then under H_0 the variance of y is $\Sigma_\phi(1)$. Thus, let $z = K_\phi^{-1} \Delta_1 y$, $W = K_\phi^{-1} \Delta_1 X$, and $\upsilon = K_\phi^{-1} \Delta_1 \Delta_\alpha^{-1} K_\phi \varepsilon = \Delta_1 \Delta_\alpha^{-1} \varepsilon$, because K_ϕ and Δ_α commute, giving the regression and resulting GLS estimator for β as

$$z = W\beta + \upsilon \quad \text{and} \quad \hat{\beta}_{GLS} = (W'W)^{-1} W'z. \tag{8}$$

Consequently, $\hat{\beta}_{GLS}$ is the efficient estimator of β when the null hypothesis is true. This contrasts with the method of Elliot et al. (1996), who estimate β via the quasi-differenced regression $\Delta_{\bar{\alpha}} y = \Delta_{\bar{\alpha}} X\beta + \Delta_{\bar{\alpha}} \Delta_\alpha^{-1} K_\phi \varepsilon$, where $\bar{\alpha}$ is chosen, for each X, as the value at which the asymptotic power envelope reaches 0.5. The resulting estimator, $\hat{\beta}_{QD}$, say, is efficient under that specific chosen alternative.

Here we derive *KL* for the residuals from the GLS regression in (8), defined by

$$w = Mz \quad \text{where } M = I_T - W(W'W)^{-1} W', \tag{9}$$

so that M is symmetric, idempotent, and of rank $n = T - k$. Defining the following matrices,

$$\Omega(\alpha) = \Delta_1 \Delta_\alpha^{-1} (\Delta_\alpha^{-1})' \Delta_1' \quad \text{and} \quad A = A_\alpha = C' \Omega(\alpha) C,$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

where C is a $T \times n$ matrix satisfying

$$CC' = M \quad \text{and} \quad C'C = I_n, \tag{10}$$

yields the following lemma.

LEMMA 1. *Under Assumption 1*

(i) the distribution of w is singular-normal, with log-likelihood

$$L_w(\alpha) = -\frac{1}{2} w' C' A^{-1} C w - \frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \log|A|. \tag{11}$$

(ii) Kullback–Leibler divergence is given by

$$KL = \frac{1}{2}[\mathrm{Tr}[A^{-1}] + \log|A| - n]. \tag{12}$$

2.3. Power and Kullback–Leibler for Unit Root Inference

Although the generic statistical uses for *KL* are myriad, detailed at length in Kullback (1997), the purpose here is to examine the impact of the specification of equations (4) and (6) when testing $H_0 : \alpha = 1$ in (5). The most direct method of doing so might be to use the power envelope (the powers of the set of point optimal tests), specifically the asymptotic power envelope in the absence of detailed finite-sample results. According to Theorem 2 of Elliott et al. (1996) the asymptotic power envelope does not depend upon either ϕ or σ^2. Moreover, if $p < 0.5$, then μ_t is "slowly evolving," and the asymptotic envelope will not depend on p either.

The claim of this paper is that the asymptotic prediction that the properties of unit root inference are unaffected by ϕ and p is not borne out in practice. To fix matters, suppose that we simplify the model so that $\beta_{(3)} = 0$, $m = 1$, and $\varepsilon_t \sim N(0, 1)$, giving

$$(1 - \alpha L)\left(y_t - \beta_1 - \beta_2 t^p\right) = (1 + \phi_1 L)\,\varepsilon_t, \qquad t = 1, \ldots, T, \tag{13}$$

and test the simple hypotheses

$$H_0 : \alpha = 1 \quad \text{vs.} \quad H_1 : \alpha = \bar{\alpha}, \tag{14}$$

where $\bar{\alpha}$ is the value at which the asymptotic power envelope is 0.5 when $p < 0.5$. For given values of ϕ_1 and p the point optimal test for testing (14), based on the log-likelihood in (11), is

$$P_w = L_w(\bar{\alpha}) - L_w(1), \tag{15}$$

and H_0 is rejected if $P_w > k_\delta$, where δ is chosen so the test has size δ. For the specific comparisons to be made here we do not require invariance with respect to σ^2.

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

However, if required, the entire framework of Dufour and King (1991) follows immediately by noting that their maximal invariant is just $v = Cw/\sqrt{w'w}$.

Choosing $T = 100$ and $\delta = 0.05$, we find $\bar{\alpha} = 0.93$. The power of P_w was simulated with 50,000 replications for different values of ϕ and p. Those powers, divided by 0.5 (benchmark asymptotic power), are reported in panel A of Table 1. The asymptotic prediction, for values of $p < 0.5$, that these values should all be 1 is not accurate. In fact, if a practitioner faces a model in which $p = 0.45$ and $\phi = -0.9$ their ability to correctly identify, as measured by power in these simplified circumstances, a stationary model is in fact only one-quarter of that predicted.

To illustrate that *KL* and power have the potential to be equally informative for the unit root hypothesis specifically, *KL* was first evaluated for a benchmark model having only a constant and no innovation correlation for a sample size of $T = 100$. At $\bar{\alpha} = 0.93$ this value was 2.03. Reported in panel B of Table 1, are the relative outcomes across the same values of p and ϕ considered for power. Although *KL* is not a perfect predictor for power the qualitative relationship is clear. Moreover, although calculating power may be impractical, calculation of *KL* is routine and instantaneous. It also has additional uses beyond acting as mere

TABLE 1. Comparison of power and divergence for the unit root hypothesis

	p							
ϕ_1	0.05	0.15	0.25	0.35	0.45	*1.00*	*5.00*	*10.0*
A. The relative power of the asymptotic point optimal test (15) for testing (14) in (13) for different p and ϕ, for $T = 100$								
−0.9	0.354	0.341	0.293	0.290	0.283	*0.254*	*0.378*	*0.437*
−0.5	0.461	0.437	0.413	0.383	0.369	*0.319*	*0.587*	*0.662*
−0.1	0.763	0.559	0.430	0.428	0.374	*0.368*	*0.659*	*0.721*
0.0	0.924	0.635	0.447	0.433	0.401	*0.370*	*0.677*	*0.738*
0.1	1.05	0.926	0.454	0.440	0.407	*0.371*	*0.699*	*0.781*
0.5	1.10	1.01	0.460	0.443	0.409	*0.385*	*0.709*	*0.835*
0.9	1.13	1.04	1.02	0.952	0.424	*0.388*	*0.743*	*0.877*
B. The relative divergence, *KL* in (12), evaluated under the alternative in (14) for (13) for different p and ϕ, for $T = 100$								
−0.9	0.352	0.316	0.272	0.261	0.251	*0.231*	*0.369*	*0.438*
−0.5	0.428	0.401	0.376	0.355	0.339	*0.318*	*0.483*	*0.540*
−0.1	0.670	0.593	0.383	0.368	0.351	*0.345*	*0.530*	*0.617*
0.0	0.804	0.743	0.399	0.378	0.361	*0.348*	*0.579*	*0.655*
0.1	0.956	0.691	0.405	0.381	0.363	*0.352*	*0.592*	*0.710*
0.5	0.985	0.781	0.414	0.389	0.370	*0.362*	*0.625*	*0.741*
0.9	1.04	0.942	0.853	0.792	0.393	*0.369*	*0.662*	*0.757*

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

proxy for power. To illustrate, in both panels of Table 1, values for $p > 0.5$ are also considered. The values are italicized to emphasize that the benchmark should not be 0.5. However, numerically it seems that given a value of ϕ_1, both power and *KL* are minimized at some value of p. In the following section it will be proved that this is indeed the case for *KL*.

3. DIFFERENTIAL PROPERTIES OF *KL*

Given Assumption 1 we interpret *KL* as a function of both p and ϕ (and also the autocorrelation coefficient α) as $KL(\alpha, p, \phi)$, i.e., $KL : \mathbb{R}^{m+2} \to \mathbb{R}$. Here we establish the differential properties of that euclidean mapping. *KL* is a function of α, p, and ϕ through $A = C'\Omega(\alpha)C$ and thus depends on p and ϕ only through C, itself a function of $W = K_\phi^{-1}\Delta_1 X$. Letting the derivatives of any matrix Ψ with respect to p and the elements ϕ_j of ϕ be denoted $\partial_p\Psi$ and $\partial_{\phi_j}\Psi$, then the following theorem establishes differentiability of C and gives the required derivatives.

THEOREM 1. *Let C be the singular value decomposition of the symmetric idempotent $M = CC' = I_T - W(W'W)^{-1}W'$ and let C_0 and W_0 define points in $\mathbb{R}^{T\times n}$ and $\mathbb{R}^{T\times k}$. Then*

(i) if W is differentiable in a neighborhood of W_0, C is differentiable in a neighborhood of C_0, and

(ii) the derivatives of C are

$$\partial_p C = -W(W'W)^{-1}(\partial_p W)'C \quad \text{and} \quad \partial_{\phi_j} C = -W(W'W)^{-1}(\partial_{\phi_j} W)'C. \tag{16}$$

Theorem 1 establishes that C is differentiable in both the trend parameter p and moving average parameters ϕ. We are thus in a position to state the main theorem of this paper and demonstrate that *KL* is a continuous, differential, convex function of both p and ϕ and thus minimizable with respect to those arguments.

THEOREM 2. *Let KL be defined as in Lemma 1. Then*

(i) KL is differentiable, and therefore continuous, with respect to p, with derivative given by

$$\frac{\partial KL}{\partial p} = \operatorname{Tr}\left[A^{-1}\left(\partial_p W\right)'(W'W)^{-1}W'\Omega(\alpha)C\left(I_n - A^{-1}\right)\right], \tag{17}$$

where

$$\partial_p W = K_\phi^{-1}\Delta_1(\partial_p X) = K_\phi^{-1}\Delta_1\left(0, t^p\log(t), 0, \ldots, 0\right)_{t=1}^{T};$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

(ii) KL is differentiable, and therefore continuous, with respect to $\phi = \{\phi_j\}_{j=1}^m$, *with partial derivatives given by*

$$\frac{\partial KL}{\partial \phi_j} = \mathrm{Tr}\left[A^{-1}\left(\partial_{\phi_j} W\right)'(W'W)^{-1}W'\Omega(\alpha)C\left(I_n - A^{-1}\right)\right], \tag{18}$$

where

$$\partial_{\phi_j} W = -K_\phi^{-1} L^{(j)} W; \quad \textit{and}$$

(iii) $KL(\alpha, p, \phi)$ *is quasi-convex over both* p *and* ϕ, *with*

$$\frac{\partial^2 KL}{\partial p^2} \geq 0 \quad \textit{and} \quad \frac{\partial^2 KL}{\partial \phi_j^2} \geq 0, \qquad j = 1, \ldots, m,$$

and therefore any solutions in p *to (17) and in each* ϕ_j *to (18) are at minima.*

(iv) Let $\hat{\alpha}$, $\hat{p}$, *and* $\hat{\phi}$ *be consistent estimators for* α, p, *and* ϕ. *Defining* $\overline{KL}(\alpha, p, \phi) = T^{-2}KL(\alpha, p, \phi)$, *then*

$$\overline{KL}(\hat{\alpha}, \hat{p}, \hat{\phi}) = \overline{KL}(\alpha, p, \phi) + o_p(1).$$

Theorem 2 proves that *KL* is differentiable and convex in the model parameters p and ϕ. As a consequence minima in these model features may be explicitly characterized. Moreover, continuity ensures that our analysis can be made feasible in that *KL* is consistently estimable, provided we have consistent estimators for α, p, and ϕ.

First we illustrate the minimum of *KL* in p via a simplified model with no error autocorrelation:

$$(1 - \alpha L)(y_t - \beta_1 - \beta_2 t^p) = \varepsilon_t; \qquad \varepsilon_t \sim iidN(0, \sigma^2), \qquad t = 1, \ldots, T. \tag{19}$$

We may solve

$$\left.\frac{\partial KL}{\partial p}\right|_{p=p^*} = 0$$

and plot the solutions p^* for different sample sizes ($T = 25, 50$, and 100), giving Figure 1. Notice that for moderate sample sizes and for alternatives "close" to the null the stationary points do not correspond exactly to a linear time trend. In a neighborhood of the unit root *KL* is, to the second order, proportional to Fisher information; see Kullback (1997, p. 28). Marsh (2007, Table 1) has shown that when $p = 1$ Fisher information in the statistic $C'w$ is $O(T)$ at $\alpha = 1$, rather than $O(T^2)$ when either $p \neq 1$ or $\alpha = (1 - c/T)$ for any $c > 0$. In finite samples, this

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

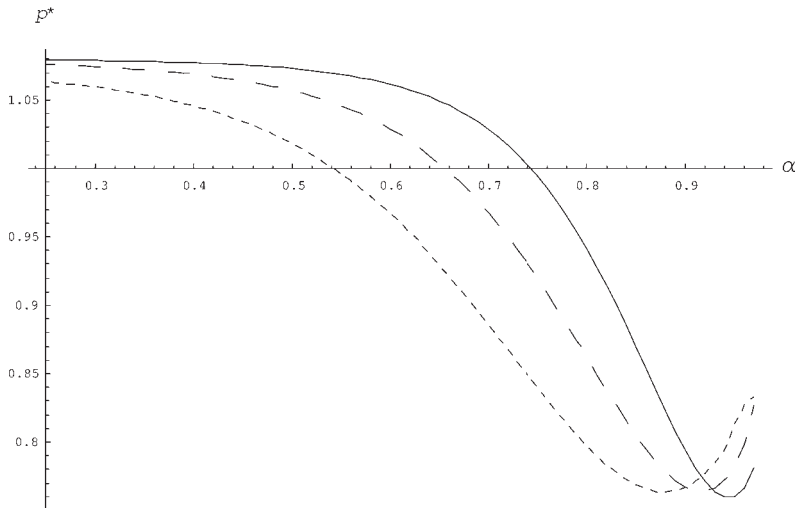


FIGURE 1. The minima of *KL*, p^*, as a function of α for model (19) for $T = 100$ (—), $T = 50$ (- - -), and $T = 25$ ($\cdots$).

phenomenon clearly has the effect, shown in Figure 1, of "pulling" p^* toward 1, for values of α very close to 1.

From an applied perspective the deterministics μ_t are a choice made by the modeler to attempt to capture the trending behavior of the data, specifically to ensure invariance with respect to those trends. On the other hand, the correlation structure of the innovations is a property of the underlying statistical process. That does not mean, however, that understanding the effects that particular autocorrelation structures have is not important.

For the minima of *KL* in ϕ, consider

$$(1-\alpha L)(y_t - \beta_1 - \beta_2 t) = (1+\phi_1 L)\varepsilon_t; \qquad \varepsilon_t \sim iidN(0,\sigma^2), \tag{20}$$

so that y_t follows an $ARIMA(0,1,1)$ process under the null. As α varies we can calculate the minimum arguments of *KL*, ϕ_1^*, for sample sizes of $T = 25, 50$, and 100 for model (20). These values are plotted in Figure 2. As we should expect, it is large negative values of ϕ_1 that make the distance small. Again, the result is that it is *not quite* an MA(1) with a negative unit root that minimizes the distance, although, as in the case with a linear trend, there is some uniformity in that the value of ϕ_1^* is not particularly sensitive with respect to α.

In the following section the consistent estimation of *KL* established in part (iv) of Theorem 2 is exploited to examine the effect of model specification in the context of the Nelson and Plosser (1982) data.

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

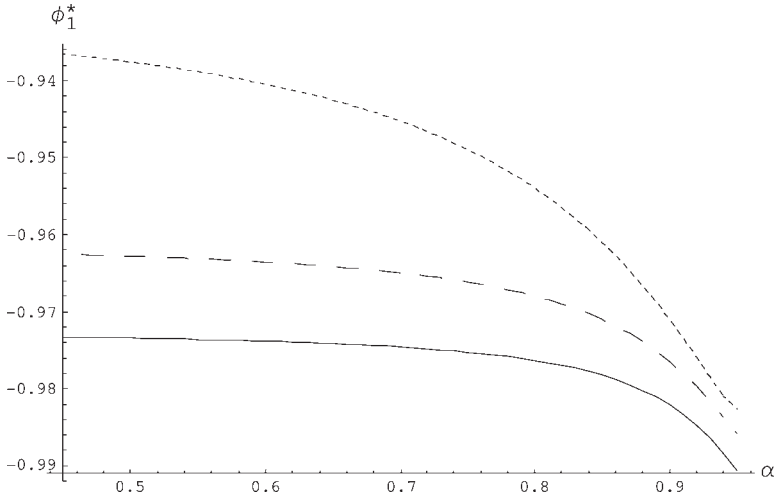


FIGURE 2. The minima of KL, ϕ_1^*, as a function of α for model (20) and for $T = 100$ (—), $T = 50$ (- - -), and $T = 25$ ($\cdots$).

4. ILLUSTRATION FOR THE NELSON AND PLOSSER DATA

The purpose here is to measure the impact of different model specifications on our ability to discriminate between unit root and (trend) stationary processes, as measured by KL. These data have been much analyzed in the literature, with authors coming to different conclusions about the existence of unit roots within some of the series. See, e.g., the differing perspectives of Phillips (1991) and Dejong and Whiteman (1991), particularly with regard to the trending characteristics of the deterministic component.

Here we fit the following general model to each of the series:

$$M_0: \quad (1-\alpha L)\left(y_t - \beta_1 - \beta_2 t^p\right) = \left(1+\phi_1 L+\phi_2 L^2\right)\varepsilon_t, \qquad t = 1,\ldots,T, \tag{21}$$

under the conditions given in Assumption 1. We also fit the two most commonly employed deterministic components, obtained as restrictions of M_0. These restrictions are

$$M_1: \beta_2 = 0 \quad \text{and} \quad M_2: p = 1, \qquad \beta_2 \neq 0, \tag{22}$$

so that M_2 employes a linear trend whereas M_1 has no trend. Following the analysis given in Section 2, for each model KL is a function of the autoregressive parameter, α, the moving average parameters, $\phi = (\phi_1, \phi_2)'$, and, for M_0, the trend parameter, p. By construction it depends on neither β nor σ^2.

To proceed let $\theta_0 = (\alpha, p, \phi_1, \phi_2)'$ and $\theta = (\alpha, \phi_1, \phi_2)'$ and let $\hat{\theta}_0$ denote the estimator (of θ_0) obtained via Gaussian maximum likelihood applied to M_0 and

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

$\hat{\theta}_m$, the estimator (of θ) obtained from model M_m, $m = 1, 2$. These yield estimated standardized divergences, as in

$$\widehat{KL}_0 = \overline{KL}_0\left(\hat{\theta}_0\right) \quad \text{and} \quad \widehat{KL}_m = \overline{KL}_m\left(\hat{\theta}_m\right), \qquad m = 1, 2,$$

that are consistent for their population counterparts via Theorem 2(iv).

For each series in the Nelson and Plosser (1982) data, the maximum likelihood estimator (MLE) for all unknown parameters was computed and the estimated measures of entropy calculated for each of the three models. For M_0 (similar notation applies also to M_1 and M_2) define $\gamma = \left(\alpha, \beta_1, \beta_2, p, \phi_1, \phi_2, \sigma^2\right)'$; then the MLE $\hat{\gamma}$ is found via numerical optimization of the Gaussian likelihood, $L_y(\alpha)$, implied by (7). Proceeding under the maintained assumption that $|\alpha| < 1$ (note if $\alpha = 1$ then $KL = 0$, by definition) then $\hat{\gamma}$ is consistent and asymptotically normal, with

$$\left(H_n\left(\hat{\gamma}\right)\right)^{1/2}\left(\hat{\gamma} - \gamma\right) \to_d N(0, I_7),$$

where

$$H_n\left(\hat{\gamma}\right) = -\left.\frac{\partial^2 L_y(\alpha)}{\partial\gamma\,\partial\gamma'}\right|_{\gamma=\hat{\gamma}}. \tag{23}$$

To justify, to some extent, the assumption of an MA(2) innovation process, for each model we calculate the first-order residual correlation,

$$r_m = \sqrt{T - k_m}\,\frac{\sum_{k_m+1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-j}}{\sum_{k_m+1}^{T}\hat{\varepsilon}_t^2},$$

where k_m is the total number of parameters in M_m, $m = 0, 1, 2$. The maximum value of $|r_m|$ obtained in any estimated model was 1.815, which is not significant according to standard asymptotic critical values at the 5% significance level. The outcomes of all regressions are recorded in Table 2. For M_0, the estimated autoregressive parameter, $\hat{\alpha}$, and also the estimated trend degree $\hat{p}$ and its associated slope coefficient $\hat{\beta}_2$, are given along with the estimated divergence measure, $T^2\widehat{KL}_0$. For models M_1 and M_2 only $\hat{\alpha}$ and $T^2\widehat{KL}_m$, $m = 1, 2$, are recorded.

Before discussion of the impact of model specification on estimated divergence consider the results for M_0. On the basis of the diagonal elements of $H_n\left(\hat{\gamma}\right)^{-1}$ obtained from (23) approximate 95% confidence intervals for both β_2 and p can be constructed, e.g., for β_2 : $\hat{\beta}_2 \pm 1.96\sqrt{H_n(\hat{\gamma})^{-1}_{\beta_2,\beta_2}}$. In Table 2 if the interval for β_2 includes 0 (implying no trend) this is indicated on the estimated value. Likewise for p: if the interval includes either 0 (i.e., β_2 is unidentified) or 1 (a linear trend), or both, this is appropriately indicated. Of prime relevance to the current study are the findings that the fitted models for both real and nominal wages and also for CPI have significant, nonlinear trends. This implies that models having nonlinear trends of the type specified in Assumption 1 have empirical relevance. The intervals can also be used as an indicator for which of the restricted

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

TABLE 2. Estimated values for α and KL in M_0 of (21) and M_1 and M_2 of (22) and also β_2 and p for M_0, for the Nelson and Plosser data sets

	M_0				M_1		M_2	
	$\hat{\alpha}_0$	$\hat{\beta}_2$	$\hat{p}$	$T^2\widehat{KL}_0$	$\hat{\alpha}$	$T^2\widehat{KL}_1$	$\hat{\alpha}$	$T^2\widehat{KL}_2$
Real GNP	1.005	0.172*	−0.011*	0.087	1.001	0.003	0.991	0.002
Nominal GNP	0.982	0.340	0.825#	0.011	1.008	0.104	0.990	0.002
GNP per.capita	1.003	−0.053*	0.518*	0.001	1.002	0.009	0.993	0.001
Bond yield	0.970	0.013*	1.470*,#	0.072	1.000	0.000	0.949	0.302
Nominal wage	0.980	0.226	0.813	0.002	1.006	0.081	0.989	0.003
Real wage	0.880	0.909	0.305	3.267	1.000	0.000	0.998	0.000
Unemployment	0.750	1.118	−0.161*	53.71	0.755	52.38	0.751	15.80
Employment	1.001	0.010*	0.810#	0.000	1.001	0.001	0.998	0.000
Money	0.979	0.072	0.999#	0.036	1.003	0.023	0.975	0.055
S&P500	0.946	0.019*	1.170#	1.205	1.009	0.305	0.930	1.519
Velocity	0.962	0.084*	−0.566*	4.016	0.959	4.862	1.011	0.008
Industrial production	0.884	0.024	1.104#	6.042	1.000	0.000	0.854	9.622
CPI	0.975	0.563	0.571	0.130	1.021	1.942	0.983	0.035

Note: *denotes that 95% asymptotic confidence interval contains 0.
#denotes that 95% asymptotic confidence interval contains 1.

models may be appropriate for each series. A significant linear trend, i.e., model M_2, is indicated for nominal gross national product (GNP), money supply and industrial production.

Because the majority of studies of these data only consider models M_1 and M_2, these were estimated for all series, regardless of the statistical significance of the trend parameter or its slope. The interest here is the effect that model specification has on the estimated value of KL, which for most series is actually minor. Heuristically, we would therefore not expect the decision as to whether or not those series had unit roots to be influenced by the specification of any deterministic trend included. However, there are some clear exceptions to focus our analysis on.

For nominal GNP imposing no trend implies an estimated model having significantly higher divergence. However, that finding is given context via the results for M_0 indicating that such a restriction is not justified. For nominal wages, although neither restriction is statistically justified, the impact of wrongly doing so is minor. The results for real wages are most intriguing. Neither restriction is justified statistically, and imposing either would give a misleading estimated divergence of approximately zero. Fitting the nonlinear trend instead yields a divergence very far from zero. This novel finding may only be illuminated by following the approach taken in this paper.

For both velocity and industrial production M_1 and M_2 give very different values for the divergence but in completely opposite directions. For each it is possible

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

to "choose" a specification to give a fitted model very close to a unit root, as measured by *KL*. Only consideration of the more general model gives these results a proper context, because it is from applying the statistically justified restriction that appropriate weight can be given to these outcomes in assessing the nature of these series.

5. CONCLUSIONS

This paper has explored the influences of model specification on the properties of unit root inference. Kullback–Leibler divergence is used for this purpose. It is shown that, like the power envelope, it is expressible as an integral of a simple function of the joint sample densities, under the null and alternative. Moreover, for the unit root problem specifically, the numerical properties of the divergence are almost identical to those of power. However, unlike power the divergence is available in closed form, and thus its analytic and numeric properties may be detailed. It is shown to be continuous, differentiable, and convex (i.e., minimizable) in the degree of trending of a deterministic component and also the correlation structure of the underlying innovation process. Confirming the expansive experimental evidence on the question, it is trends that are linear and innovations with a large negative first-order autocorrelation that minimize the divergence and thus, through the bounds given previously, reduce power.

The results also have practical importance. In the context of the Nelson and Plosser (1982) data set, the entropic measures can be estimated via the fitted model and the impact of trend specification explicitly measured. In the case of real wages is it is revealed that imposition of either a linear trend or no trend gives similar, but misleading, results, because neither restriction is statistically justified. It is the approach of this paper that is able to uncover such subtleties as we expand the allowed specification beyond that of the original work of Granger and Newbold (1974).

REFERENCES

Critchley, F., P. Marriott, & M. Salmon (1994) Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *Annals of Statistics* 22, 1587–1602.

Dejong, D.N. & C.H. Whiteman (1991) The case for trend-stationarity is stronger than we thought. *Journal of Applied Econometrics* 6, 413–421.

Dufour, J.-M. & M.L. King (1991) Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR(1) errors. *Journal of Econometrics* 47, 115–143.

Durlauf, S.N. & P.C.B. Phillips (1988) Trends versus random walks in time series analysis. *Econometrica* 56, 1333–1354.

Elliott, G., T.J. Rothenberg, & J.H. Stock (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.

Gibbs, A.L. & F.E. Su (2002) Choosing and bounding probability metrics. *International Statistical Review* 70, 419–435.

Granger, C.W.J. & P. Newbold (1974) Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–120.

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

Harris, D., D.I. Harvey, S.J. Leybourne, & A.M.R. Taylor (2009) Testing for a unit root in the presence of a possible break in trend. *Econometric Theory* 25, 1545–1588 (this issue).

Harvey D.I., S.J. Leybourne, & A.M.R. Taylor (2009) Unit root testing in practice: Dealing with uncertainty over the trend and initial condition. *Econometric Theory* 25, 587–636.

Kullback, S. (1997) *Information Theory and Statistics*. Dover.

Leybourne, S.J., T.C. Mills, & P. Newbold (1998) Spurious rejections by Dickey–Fuller tests in the presence of a break under the null. *Journal of Econometrics* 87, 191–203.

Magnus, J.R. & H. Neudecker (1999) *Matrix Differential Calculus, with Applications in Statistics and Econometrics*, (rev. ed.) Wiley.

Marsh, P. (2007) The available information for invariant tests of a unit root. *Econometric Theory* 23, 686–710.

Nelson, C.R. & C.I. Plosser (1982) Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10, 139–162.

Perron, P. (1989) The Great Crash, the oil price shock and the unit root hypothesis. *Econometrica* 57, 1361–1401 (Erratum, 61, 248–249).

Perron, P. & S. Ng (2001) Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69, 1519–1554.

Phillips, P.C.B. (1986) Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–340.

Phillips, P.C.B. (1991) To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6, 333–364.

Phillips, P.C.B. (1998) New tools for understanding spurious regressions. *Econometrica* 66, 1299–1325.

Phillips, P.C.B. (2007) Regression with slowly varying regressors and nonlinear trends. *Econometric Theory* 23, 557–614.

Phillips, P.C.B. & W. Ploberger (2003) Empirical limits for time series econometric models. *Econometrica* 71, 627–673.

Phillips, P.C.B. & Z. Xiao (1998) A primer on unit root testing. *Journal of Economic Surveys* 12, 423–470.

Würtz, A. (1997) A universal upper bound on power functions. University of New South Wales, Discussion paper 97/17.

Zivot, E. & D.W.K. Andrews (1992) Further evidence on the Great Crash, the Oil Price Shock, and the unit root hypothesis. *Journal of Business & Economic Statistics* 10, 251–270.

APPENDIX

Proof of Lemma 1. Part (i): M is square, symmetric, and idempotent of rank $n = T - k$, and hence $w = Mz$ has singular normal distribution, with mean $\mathrm{E}[w] = \mu = MW\,\beta = 0$ and variance $V[w] = \sigma^2\Sigma = \sigma^2 M\Omega(\alpha)M$. The density function of w is therefore given by

$$f(w;\alpha) = \frac{\exp\left\{-\frac{1}{2\sigma^2}w'\Sigma^{-}w\right\}}{\left(2\pi\sigma^2\right)^{n/2}\sqrt{\prod_{i=1}^{n}\lambda_i}}, \tag{A.1}$$

where the λ_i, $i = 1, 2, \ldots, n$, are the ordered nonzero eigenvalues of Σ and Σ^{-} denotes its generalized inverse.

To find a suitable Σ^{-}, write

$$M\Omega(\alpha)M = CC'\Omega(\alpha)CC' = CAC',$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

where the $T \times n$ matrix C is the singular value decomposition of M defined in (10). Consequently, note that

$$\Sigma C A^{-1} C' \Sigma = C A C' C A^{-1} C' C A C' = C A C' = \Sigma,$$
$$C A^{-1} C' \Sigma C A^{-1} C' = C A^{-1} C' C A C' C A^{-1} C' = C A^{-1} C',$$
$$\left(C A^{-1} C' \Sigma \right)' = \left(C C' \right)' = M' = M = C C' = \Sigma C A^{-1} C',$$
$$\left(\Sigma C A^{-1} C' \right)' = \left(C C' \right)' = M' = M = C C' = C A^{-1} C' \Sigma,$$

so that $C A^{-1} C'$ satisfies the four conditions that define the unique Moore–Penrose inverse of Σ; i.e., we use $\Sigma^- = C A^{-1} C'$.

Because the singular value decomposition is unique only up to orthogonal transformation, and because the matrix A is symmetric, it has a spectral decomposition,

$$A = R \Lambda R', \qquad \Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\},$$

where $R'R = RR' = I_n$ and $R = (r_1, r_2, \ldots, r_n)$ is the matrix of eigenvectors, which then satisfy the equations

$$A r_i = \lambda_i r_i, \qquad i = 1, 2, \ldots, n.$$

That is, the λ_i are the eigenvalues of A. Moreover, we can also assume, without loss of generality, that in addition to (10), C also satisfies

$$C' A C = C' \Lambda C \quad \text{and} \quad C' A^{-1} C = C' \Lambda^{-1} C.$$

Taking logs of (A.1), noting $|A| = \prod_{i=1}^{n} \lambda_i$, we find

$$L_w(\alpha) = -\frac{w' C A^{-1} C' w}{2\sigma^2} - \frac{n \ln\left(2\pi\sigma^2\right)}{2} - \frac{\ln |A|}{2},$$

as required.

Part (ii): This follows from straightforward application of the definition given in (3) to an $n \times 1$ Gaussian random vector $C'w \sim N(0, \sigma^2 A)$. ■

Proof of Theorem 1. Because $W = K_\phi^{-1} \Delta_1 X$, then immediately W is differentiable with respect to ϕ_j. Now under Assumption 1 the rank of X is constant, and so X is differentiable with respect to p. Consequently, the rank of W is constant, and therefore W is also differentiable with respect to p, with generic derivative $\partial W = K_\phi^{-1} \Delta_1 (\partial X)$. In fact W is an analytic (matrix) function of both p and ϕ.

To establish differentiability of C (with respect to either parameter) we note that C is defined as the singular value decomposition of $M = I_T - W(W'W)^{-1} W'$, and is therefore the unique solution (up to orthogonal transformation), in $\mathbb{R}^{T \times n}$, to the equations

$$M = C C' \quad \text{and} \quad C'C = I_n. \tag{A.2}$$

We first show that (A.2) implies and is implied by

$$W'C = 0 \quad \text{and} \quad C'C = I_n. \tag{A.3}$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

To do this note that

$$M = CC' \Longleftrightarrow (I_T - M)C = 0 \tag{A.4}$$

and define

$$P = I_T - M = W(W'W)^{-1}W' = WW^{+} = \left(W^{+}\right)'W',$$

where W^{+} denotes the Moore–Penrose inverse of W that exists and is unique because the rank of W is also constant. Writing (A.4) as $(W^{+})'W'C = 0$, and because

$$W^{+} = (W'W)^{+}W' \quad \text{and} \quad W = W(W'W)^{+}(W'W),$$

we have

$$(W^{+})'W'C = 0 \Longleftrightarrow (W'W)^{+}W'C = 0,$$

which leads to

$$(W'W)^{+}W'C = 0 \Longleftrightarrow (W'W)(W'W)^{+}W'C = 0 \Longleftrightarrow W'C = 0,$$

as required.

To continue, define the matrix-valued function $h : \mathbb{R}^{T\times k} \times \mathbb{R}^{T\times n} \to \mathbb{R}^{T\times n}$ of C and W by

$$h(C, W) = \begin{pmatrix} W'C \\ C'C - I_n \end{pmatrix};$$

then following a similar argument to Magnus and Neudecker (1999, Thm. 8.7), h is differentiable on $\mathbb{R}^{T\times k} \times \mathbb{R}^{T\times n}$. Letting the point C_0, W_0 in $\mathbb{R}^{T\times k} \times \mathbb{R}^{T\times n}$ satisfy

$$h(C_0, W_0) = 0,$$

and further

$$\det[J_0] = \det\left[\left.\frac{h(C, W)}{\partial C}\right|_{C_0, W_0}\right] = \det\begin{bmatrix} W_0' \\ 2C_0' \end{bmatrix} \neq 0,$$

because by definition $W'C = 0$, then the conditions for the implicit function theorem are met (see Magnus and Neudecker, 1999, Sect. 7, Thm. A.3). Consequently, there exists a neighborhood in $\mathbb{R}^{T\times k}$, $V(W_0)$ and a unique (up to orthogonal transformation) matrix-valued function $C : V(W_0) \to \mathbb{R}^{T\times n}$ for which the following statements hold:

(a) C is differentiable on $V(W_0)$,
(b) $C(W_0) = C_0$, and
(c) $W'C = 0$ and $C'C = I_n$ for all $W \in V(W_0)$,

which concludes the proof of part (i).

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

For part (ii) we require an explicit relationship between the differential of C and that of W. From (A.3) we have

$$W'C = 0,$$

so that denoting the differentials of W and C by ∂W and ∂C, respectively (suppressing for the moment which variable we are differentiating with respect to), we have

$$(\partial W)'C + W'(\partial C) = 0,$$

giving

$$(W')^{+}W'(\partial C) = -(W')^{+}(\partial W)'C. \tag{A.5}$$

Consider the matrix defined by

$$H = (W')^{+}W' + CC' = P + M = I_T,$$

and so

$$\partial C = H(\partial C) = \left((W')^{+}W' + CC'\right)(\partial C) = (W')^{+}W'(\partial C),$$

because $C'(\partial C) = 0$. Using (A.5) the generic derivative of C is

$$\partial C = -(W')^{+}(\partial W)'C = -W(W'W)^{-1}(\partial W)'C, \tag{A.6}$$

which then gives the expressions in (16). ■

Proof of Theorem 2. For parts (i) and (ii), to establish differentiability of KL, as given in Lemma 1, we utilize *Cauchy's rule of invariance* for (possibly) matrix-valued functions of matrix arguments. If F is differentiable at D and G is differentiable at $E = F(D)$, then the composite function, defined by

$$H(D, U) = G \circ F,$$

is differentiable for all $n \times m$ matrices U and

$$\partial H(D, U) = \partial G(E; \partial F(D; U)).$$

From Theorem 1, C is differentiable with respect to both p and ϕ_j, and so differentiability of A and thence KL follows immediately.

To proceed denote the generic derivative of KL, A, and C with respect to either p or any ϕ_j, for $j = 1, \ldots, m$, by ∂KL, ∂A, and ∂C and similarly for any scalar or matrix function of A by $\partial h(A)$; then because $A = C'\Omega(\alpha)C$, we have

$$\partial A = \left[(\partial C)'\Omega(\alpha)C + C'\Omega(\alpha)\partial C\right]. \tag{A.7}$$

Substitution of (A.6) into (A.7) yields

$$\partial A = -C'[D + D']C, \tag{A.8}$$

where $D = (\partial W)(W'W)^{-1}W'\Omega(\alpha)$. Finally, noting the following standard differentials,

$$\partial \ln|A| = \mathrm{Tr}\left[A^{-1}(\partial A)\right] \quad \text{and} \quad \partial A^{-1} = -A^{-1}(\partial A)A^{-1},$$

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press

then we obtain

$$\partial KL = \frac{1}{2}\left(\mathrm{Tr}\left[(\partial A)\,A^{-1}\left(I_n - A^{-1}\right)\right]\right). \tag{A.9}$$

Substituting (A.8) into (A.9) and then rearranging proves both part (i) and part (ii) for the derivatives with respect to p and ϕ_j, in turn.

For part (iii), notice that A and hence A^{-1} are $n \times n$ positive definite, symmetric matrices of full rank. Consequently, following Magnus and Neudecker (1999, p. 152), the kth derivatives (guaranteed to exist by Theorem 1) of the matrix functions A^{-1} and $\ln|A|$, with respect to p, are

$$\partial^k\left(A^{-1}\right) = (-1)^k k!\left(A^{-1}(\partial A)\right)^k A^{-1} \quad \text{and}$$

$$\partial^k(\ln|A|) = (-1)^{k-1}(k-1)!\mathrm{Tr}\left[\left(A^{-1}(\partial A)\right)^k\right],$$

where $\partial^k(\cdot)$ denotes a generic kth derivative. As a consequence the second derivatives of entropy are given by

$$\partial^2 KL = \frac{1}{2}\mathrm{Tr}\left[\left(A^{-1}\partial A\right)^2\left(2A^{-1} - I_n\right)\right].$$

Note that, by construction $\mathrm{Tr}[A^{-1}] = \mathrm{Tr}\left[(A^{-1/2})'(A^{-1/2})\right]$ is a norm on the space of positive definite symmetric matrices of full rank and is thus minimized at $A = A^{-1/2} = I_n$, i.e., when $\alpha = 1$. Hence $\mathrm{Tr}[A^{-1}] \geq n$, and so $\mathrm{Tr}\left[2A^{-1} - I\right] > 0$, and therefore

$$\mathrm{Tr}\left[\left(A^{-1}\partial A\right)^2\left(2A^{-1} - I\right)\right] \geq 0 \rightarrow \partial^2(KL) \geq 0.$$

Consequently, $KL(\alpha, p, \phi)$ is (quasi) convex in p and ϕ_j, and so any solution to either

$$\frac{\partial KL}{\partial p} = 0 \quad \text{or} \quad \frac{\partial KL}{\partial \phi_j} = 0$$

is at a minimum.

For part (iv), because KL is formed from taking expectations under the null then, as in Elliott et al. (1996, App. B), $T^{-2}L_w(\alpha) = O_p(1)$, for all α. Consequently, $\overline{KL}(\alpha, p, \phi) = \mathrm{E}\left[T^{-2}\left(L_w(1) - L_w(\alpha)\right)\right] = O(1)$, and a continuous mapping theorem follows given continuity and differentiability (which is trivially extended for α), as established in parts (i) and (ii). ■

https://doi.org/10.1017/S0266466609990284 Published online by Cambridge University Press