ELSEVIER

Journal of Econometrics 76 (1997) 223–250

JOURNAL OF Econometrics

Inferring the rank of a matrix

John G. Cragg[a], Stephen G. Donald*[,b,c]

[a] Department of Economics, University of British Columbia, Vancouver, BC V6T 1W5, Canada
[b] Department of Economics, Boston University, Boston, MA 02215, USA
[c] Department of Econometrics, University of New South Wales, Sydney, NSW 2052, Australia

(Received June 1992; final version received August 1995)

Abstract

This paper considers methods of inference concerning the rank of a matrix $\Pi - \Xi$ based on an asymptotically normal estimate of Π and some identifiable specification for Ξ. One such specification is $\Xi = 0$, in which case one is interested in the rank of Π. We first propose, and examine the properties of, a test of the hypothesis that the rank is of a given size against the alternative that the rank is larger. We then look at the problem of estimating the rank of this matrix using model selection procedures and sequential hypothesis testing. Conditions for consistency of such procedures are obtained. Some economic applications are discussed and the various methods are compared in a Monte Carlo experiment.

Key words: Rank of matrix tests; Minimum chi-squared test; Minimum distance; Model selection; Rank estimation
JEL classification: C12; C13

1. Introduction

There are many situations in which it is useful to know the rank of a matrix for which a consistent estimate exists. An important and well-known example occurs in linear simultaneous-equation models (and equivalently instrumental-variables models) where identification relies on a certain matrix being of a

* Corresponding author.

The first author acknowledges support by the Social Sciences and Humanities Research Council of Canada under Grants 410-90-0627 and 410-94-0737. Useful comments were received on some aspects of the paper at seminars at UBC and at Nuffield College, particularly from J.A. Mirrlees and Joris Pinkse. We would like to thank two referees and Richard Blundell for comments and advice.

PII S0304-4076(95)01790-8

specific rank. Lewbel (1991) has shown that a number of results in consumer theory may depend on the rank of certain matrices that are estimable. Gill and Lewbel (1992) discuss a variety of applications. More generally, there are important examples, such as factor analysis and asset pricing, in which one is interested in determining the rank of a matrix that is the difference between an estimable matrix and another matrix that may depend on a small number of unknown parameters.

This paper investigates procedures for determining the rank in this general setup. We consider methods for testing hypotheses regarding the rank[1] and the use of model selection criteria and sequential hypothesis testing methods to estimate the rank consistently. The results are valid under the general assumption that the sample matrix has elements that are asymptotically normally distributed with a covariance matrix that can be estimated consistently.

The paper is divided into the following sections. Section 2 presents the statistic used to test the null hypothesis and examines its properties. Section 3 shows how one may use the test statistic to estimate consistently the rank of the matrix. Section 4 discusses some potential applications in factor analysis and asset pricing where the more general setup is required. In Section 5 we present a Monte Carlo experiment that compares the different procedures for estimating the rank. Section 6 offers some concluding remarks.

2. Hypothesis testing

Let Π_0 be a $K \times G$ (with $K \geqslant G$) matrix for which there exists a consistent estimator $\hat{\Pi}$. For a given specification of a matrix $\Xi(\theta)$ that is also $K \times G$ and depends on the unknown parameters[2] θ, we are interested in determining the value of L, denoted L_0, such that for any $L < L_0$ there is no $\theta \in \Theta$, where Θ is a compact set, for which $\Pi_0 - \Xi(\theta)$ has rank L. This includes the simple case where the Ξ is specified as the zero matrix, in which case we are interested in determining the rank of the matrix Π_0 itself, a problem that arises in the case of testing the rank condition for identification. However, it may be the case that $\Xi \neq 0$ as the two examples discussed in Section 4 will illustrate.

We make the following assumptions regarding the specification of $\Xi(\theta)$ and the estimator $\hat{\Pi}$ based on a sample of size N. We use the notation $\hat{\pi}$ and π_0 for

[1] Gill and Lewbel (1992) suggest a test based on a very different principle. This test is not without its problems (see Cragg and Donald, 1996) and is less general than the test discussed here.

[2] Note that we are excluding the situation where we already have estimates of θ, say $\hat{\theta}$, where the subsequent analysis would apply to $\hat{\Pi} - \Xi(\hat{\theta})$. Also note that, because Ξ is unknown, approaches based on eigenvalues or other decompositions will generally not be useful for inferring the rank of $\Pi_0 - \Xi$. However, when $\Xi = 0$ and some other restrictions to the specification are met, the procedures discussed may be reduced to examining sums of eigenvalues of appropriate matrices.

$\mathrm{vec}(\hat{\Pi})$ and $\mathrm{vec}(\Pi_0)$ respectively. Also, $\rho(.)$ will be used to denote the rank of the matrix argument.

Assumption 1. *$\Xi(\theta)$ is a continuous function on the compact set $\Theta \subset R^M$ where $M \geqslant 0$.*

In Assumption 1 we allow for the possibility that $M = 0$, in which case we interpret R^M to be the set $\{0\}$. This assumption also allows for the case where $\Xi = 0$ identically.

Assumption 2. (i) $\hat{\pi} \xrightarrow{a.s.} \pi_0$ *where* $\pi_0 \in R^{KG}$; (ii) $\sqrt{N}(\hat{\pi} - \pi_0) \xrightarrow{d} \mathrm{N}(0, W)$ *where W is finite and positive definite;*[3] (iii) *there is an estimate of W, denoted by $\hat{W}$, such that $\hat{W} \xrightarrow{a.s.} W$.*

The assumptions on the estimator are quite general and include most applications referred to in the introduction.[4]

Define the null hypothesis as involving the set

$$\Gamma(L) = \{\pi\text{: there exists } \theta \in \Theta \text{ such that } [\Pi - \Xi(\theta)]v = 0,$$
$$\text{for some } v \text{ satisfying } v'v = I_{G-L}\}.$$

Using this definition we will define L_0 as being the smallest value of L for which $\Pi_0 \in \Gamma(L)$. Defining the null hypothesis in this way makes precise the type of restrictions that are satisfied when $\Pi - \Xi(\theta)$ has rank less than or equal to L. One should note, however, that equivalently we could define

$$\Gamma(L) = \{\pi\text{: there exists } \theta \in \Theta \text{ such that } \rho(\Pi - \Xi(\theta)) \leqslant L\},$$

which implies that $\Gamma(L) \subset \Gamma(L + 1)$. Also note that the compactness assumption on Θ in Assumption 1 ensures that $\Gamma(L)$ is closed,[5] as will be demonstrated below.

We consider the test statistic $N\hat{C}(\hat{\Pi}, L)$, where

$$\hat{C}(\hat{\Pi}, L) = \min_{\pi \in \Gamma(L)} (\hat{\pi} - \pi)' \hat{W}^{-1} (\hat{\pi} - \pi). \tag{1}$$

[3] This rules out the situation where Π is symmetric, as would occur in the case of factor analysis. In Section 4 where we discuss this example in more detail, we will show how one can adapt the results of this paper to deal with that case.

[4] When one can write $W = P \otimes Q$ for positive-definite matrices P and Q, as occur in several applications, many of the results in this paper reduce to known results regarding eigenvalues.

[5] As will be shown in the proof of Lemma 1, defining the set $\Gamma(L)$ in this way ensures that it is closed. This may not be necessary as long as Π_0 is not in the closure of $\Gamma(L)$ for $L < L_0$.

To see that an approach based on $\hat{C}(\hat{\Pi}, L)$ may provide a means of inferring the value of L_0, consider the population minimization problem corresponding to (1):

$$C(\Pi_0, L) = \min_{\pi \in \Gamma(L)} (\pi_0 - \pi)' W^{-1} (\pi_0 - \pi). \tag{2}$$

An intermediate result concerns this minimization problem.

Lemma 1. For each value of L, given Assumptions 1 and 2, (i) *if* $L \geqslant L_0$ *then* $C(\Pi_0, L) = 0$, *and* (ii) *if* $L < L_0$ *then* $C(\Pi_0, L) > 0$.

The proofs of this and of all following propositions are found in the Appendix.

The next result concerns the solution to the stochastic problem (1). It will be useful for showing the power properties of the test and consistency properties of the model selection criteria considered in Section 3. Let $\hat{\Pi}^L$ denote a minimizer of the quadratic form in (1) over $\Gamma(L)$.

Lemma 2. Given Assumptions 1 and 2, (i) *if* $L \geqslant L_0$ *then* $\hat{C}(\hat{\Pi}, L) \xrightarrow{a.s.} 0$ *for* $L \geqslant L_0$ *and* $\hat{\Pi}^L \xrightarrow{a.s.} \Pi_0$, *and* (ii) *if* $L < L_0$ *then* $\hat{C}(\hat{\Pi}, L) \xrightarrow{a.s.} C(\Pi_0, L) > 0$.

This result shows that $\hat{C}(\hat{\Pi}, L)$ has the potential for making inferences about the value of L_0. In order to make inferences possible, we must develop the asymptotic distribution theory for $\hat{C}(\hat{\Pi}, L)$. To aid in this, we impose certain restrictions on $\Xi(\theta)$ beyond the assumption that L_0 is the smallest possible rank of $\Pi_0 - \Xi(\theta)$ over choices of θ.[6] Rather than making explicit restrictions, we instead assume that a full rank condition and uniqueness condition hold. In Section 4, where we examine some examples in more detail, we will discuss the conditions needed for this assumption to be satisfied.

First we will define some notation. Let $\Phi_0 = \Pi_0 - \Xi(\theta_0)$, where θ_0 is a value such that $\rho(\Phi_0) = L_0$. Note that if $\rho(\Phi_0) = L_0$, then one may express $G - L_0$ columns of Φ_0 as linear combinations of the remaining L_0 columns. That is, partitioning (after any needed re-ordering of columns) Π_0 and $\Xi(\theta_0)$ into Π_0^1 and Π_0^2, Ξ_0^1 and Ξ_0^2 of $(G - L_0)$ and L_0 columns, respectively, we can write

$$\Pi_0^1 = (\Pi_0^2 - \Xi_0^2) v_0 + \Xi_0^1, \tag{3}$$

where v_0 is unique provided that $\rho(\Phi_0) = L_0$ and given the particular choice of columns which constitute Π_0^2. Thus it is possible to write Π_0^1 as a function of the

[6] Obviously, if $L_0 < G$, then there are implicitly restrictions on the nature of $\Xi(\theta)$ since, if it were completely unrestricted, then the elements could be chosen in such a way that $\hat{\Pi} - \Xi$ has any rank, implying that the 'test statistic' can be made to be zero for any specified rank. Fortunately the applications where Ξ is not zero do specify a number of restrictions on the Ξ matrix.

remaining $KL_0 + M + (G - L_0)L_0$ parameters, $\mu_0 = (\theta_0', \mathrm{vec}(v_0)', \mathrm{vec}(\Pi_0^2)')'$. In order to proceed to finding the limiting distribution of the test statistic we need a full rank condition that allows one to express $\pi \in \Gamma(L_0)$ as a function of these parameters in a neighbourhood of μ_0 so that standard expansions, as in Chamberlain (1982) and Gourieroux and Monfort (1989), may be used to find the limiting distribution of the test statistic. It should also be noted that $M = 0$ is assumed in the case where $\Xi = 0$.

Assumption 3. When $M > 0$, $\Xi(\theta)$ *is twice continuously differentiable in* θ *on the (assumed nonempty) interior of* Θ, *there is a unique value* $\theta_0 \in \mathrm{int}(\Theta)$ *such that* $\rho(\Pi_0 - \Xi(\theta_0)) = L_0$; *and in a neighbourhood of* $\mu_0 = (\theta_0', \mathrm{vec}(v_0)', \mathrm{vec}(\Pi_2^0)')'$,

$$\rho[\partial\pi_1(\mu)/\partial(\theta, \mathrm{vec}(v))] = M + (G - L_0)L_0 .$$

The following result summarizes the asymptotic properties of $N\hat{C}(\hat{\Pi}, L)$ for various values of L.

Theorem 1. Given Assumptions 1 and 2, then (i) *given Assumption 3*, $N\hat{C}(\hat{\Pi}, L_0) \rightarrow \chi^2((G - L_0)(K - L_0) - M)$, (ii) $N\hat{C}(\hat{\Pi}, L) \rightarrow \infty$ *for* $L < L_0$, *and* (iii) $N\hat{C}(\hat{\Pi}, L) \leqslant N\hat{C}(\hat{\Pi}, L_0)$ *for* $L > L_0$.

As noted in the result, Assumption 3 is only needed for part (i), Assumptions 1 and 2 being sufficient for (ii) and (iii). Furthermore, Assumption 3 is not needed at all when $\Xi = 0$, in which case Assumptions 1 and 2 suffice for the result. The proof proceeds by showing that matrices Π in a neighbourhood of Π_0 relative to $\Gamma(L_0)$ can be expressed as differentiable functions of the vector μ varying in a Euclidean neighbourhood of μ_0.

Theorem 1 shows that the statistic $N\hat{C}(\hat{\Pi}, L)$ provides a consistent test of the null hypothesis H_0: $\Pi_0 \in \Gamma(L)$ against the alternative H_1: $\Pi_0 \notin \Gamma(L)$. The implication of Theorem 1(iii) is that, if one were to hypothesize a rank L such that $L > L_0$, then all that we can say, given the current setup, is that the statistic is dominated by a statistic that is limiting chi-squared if Assumption 3 also holds. This is because an identification problem occurs since, if $L > L_0$, there is no longer a unique value v_0 such that (3) holds.[7] Although we have been unable to obtain the limiting distribution in general when testing a rank (L) larger than the actual rank (L_0), Schott (1984) and Cragg and Donald (1993) have shown

[7] Note also that in the case with a general specification of $\Xi(\theta)$ for which Assumption 3 fails the distribution of the test statistic is also unknown so that the results of Theorem 1, in particular part (i), and any conclusions regarding hypothesis tests should properly be interpreted as applying to the restricted subset of $\Gamma(L_0)$ for which Assumption 3 holds. In the case where $\Xi = 0$ this qualification is then not required since Assumption 3 is not needed for Theorem 1(i).

that, when $\Xi = 0$ and W can be factored so that $W = P \otimes Q$ for $G \times G$ and $K \times K$ positive definite matrices P and Q (with corresponding estimators satisfying Assumption 2(iii),[8] $N\hat{C}(\hat{\Pi}, L)$ is asymptotically distributed as a random variable that is stochastically dominated by a random variable that is $\chi^2((G - L)(K - L))$ when $L > L_0$. This implies that the probability of rejecting such a null is less than the nominal size of the test.

2.1. Local power properties

Here we consider the asymptotic local power properties of the test which are indicative of just how the statistic behaves, and so how the estimation procedures based on these criteria, which we investigate in the next section, perform when the hypothesized rank is less than the true rank. The sequence of local alternatives will satisfy:

Assumption 4. (i) *$\Pi_{0,N}$ converges to Π_0 such that the sequence of minimizers of $C(\Pi_{0,N}, L_0)$ denoted by $\bar{\Pi}_{0,N}$ is unique and satisfies* $\lim_{N\to\infty} \sqrt{N}(\pi_{0,N} - \bar{\pi}_{0,N}) = \delta$; (ii) *the sequence of estimators $\hat{\Pi}_N$ satisfies* $\sqrt{N}(\hat{\pi}_N - \pi_{0,N}) \to \mathrm{N}(0, W)$; *and* (iii) *there is an estimator of W, denoted by $\hat{W}$, such that $\hat{W} \xrightarrow{a.s.} W$.*

The sequence in Assumption 4 implies that $L_0 < L_{0,N}$, where $L_{0,N}$ is defined as the smallest L such that $\Pi_{0,N}$ is contained in $\Gamma(L)$ and is a slight variation on the usual Pitman-type sequences of local alternatives similar to those used in Gallant (1987). Essentially we require that the sequence get closer to the set $\Gamma(L_0)$, where the distance is measured by the metric $C(\Pi_{0,N}, L_0)$ (which is similar to the usual Euclidean distance measure).

The following theorem gives the result that the test statistic is asymptotically noncentral χ^2 under the sequence of local alternatives.

Theorem 2. Given Assumptions 1–4,

$$N\hat{C}(\hat{\Pi}, L) \xrightarrow{d} \chi^2(((G - L_0)(K - L_0) - M); \eta^2),$$

where

$$\eta^2 = \delta'(W^{-1} - W^{-1} B(\mu_0)(B(\mu_0)' W^{-1} B(\mu_0))^{-1} B(\mu_0) W^{-1})\delta$$

and

$$B(\mu) = \frac{\partial}{\partial \mu} \pi(\mu_0).$$

[8] Note that this is satisfied when $\hat{\Pi}$ is the OLS estimator for a multivariate regression model with homoskedastic and serially uncorrelated residuals.

Note that the noncentrality parameter depends on the normalised difference between the true value $\Pi_{0,N}$ and its nearest point $\bar{\Pi}_{0,N}$ in $\Gamma(L_0)$ and is not necessarily increasing in the degree to which L_0 is lower than the rank at the sequence of points $\Pi_{0,N}$. A final remark can be made regarding the behaviour of tests for $L > L_0$ under the sequence in Assumption 4. In the restricted case, discussed after Theorem 1, where $\Xi = 0$ and W can be factored as $P \otimes Q$, and under Assumption 4 with $\Pi_{0,N}$ having rank $L > L_0$ but converging to Π_0 which has rank L_0, then $N\hat{C}(\hat{\Pi}, L)$ can be shown, using a similar argument to that used in Cragg and Donald (1993) and Schott (1984), to be asymptotically distributed as a random variable that is stochastically dominated by a $\chi^2((G-L)(K-L))$ random variable. The practical implication of this is that the size of the tests may depend on how 'close' the matrix in question is to having a smaller rank than that being tested. As will be seen in Section 5 this provides some explanation for some of the Monte Carlo results.

3. Estimating the rank

We now consider the problem of estimating L_0. We shall consider two approaches that have been proposed in the literature on model selection and determining the order of ARMA processes in time-series analysis. The first approach is based on model selection criteria and the second uses a sequential hypothesis testing approach. In each case we give conditions that will result in estimates of L_0 that are consistent.

3.1. Model selection procedures

First we provide the conditions under which use of model selection criteria leads to consistent estimates of L_0. This approach has been used extensively in the areas mentioned above, specifically by Hannan and Quinn (1979), Hannan (1980, 1981), Atkinson (1981), Pötscher (1989), and Nishii (1988). The criteria have the form

$$S(L) = N\hat{C}(\hat{\Pi}, L)f(N)^{-1} - g(L), \tag{4}$$

and estimate L by letting $\hat{L}$ denote the value of L that minimizes $S(L)$.[9] The conditions for consistency of $\hat{L}$ will relate to the functions g and f and the

[9] It is possible in finite samples with some data-generating processes that the event that more than one value of L minimizes $S(L)$ is not of measure zero, in which case we can define $\hat{L}$ to be the smallest such value. The results will show, however, that under the stated conditions this possibility will disappear as N grows.

properties of $\hat{C}(\hat{\Pi}, L)$. Form (4) covers the standard criteria in model selection problems using some other minimand based on an estimation criterion for $N\hat{C}(\hat{\Pi}, L)$, for instance:

(i) Akaike Criterion (AIC) with $f(N) = 1$ and $g(L) = 2((G - L)(K - L) - M)$;

(ii) Schwarz Criterion (BIC) with $f(N) = \log(N)$ and $g(L) = ((G - L) \times (K - L) - M)$.

Most other suggested criteria[10] differ in the details of the function analogous to $\hat{C}(\hat{\Pi}, L)$ used as a criterion.

The remainder of this section gives conditions on the functions g and f that guarantee that the model selection criteria will yield consistent estimates of L_0. We consider results for both weak and strong consistency, giving respectively that $\lim_{N\to\infty} P(\hat{L} = L_0) = 1$ and that $P(\lim_{N\to\infty} \hat{L} = L_0) = 1$. Since L is discrete, the strong consistency result implies that after some finite sample size, which usually will depend on the realization of the sample and the parameters and model being investigated, N_0 say, $\hat{L} = L_0$ with probability 1. The results give some guidance regarding the choice of f and g, although they still leave quite a large range of possible choices.

First we give conditions under which these procedures are weakly consistent. In order to give necessary and sufficient conditions we need to make an assumption on the structure of $\Xi(\theta)$ that is easy to verify in practice and will allow us to provide necessary and sufficient conditions to avoid both overestimation and underestimation.

Assumption 5. *The specification $\Xi(\theta)$ is such that for any possible value of L_0 and $\varepsilon > 0$ one can find a matrix A such that $(\mathrm{tr}(A'A))^{1/2} < \varepsilon$ and a $\theta_0 \in \Theta$ for which the matrix $A + \Xi(\theta_0)$ satisfies Assumption 3 for the specified value of L_0.*

As noted, this assumption is used only to provide conditions that are necessary to avoid underestimation. In practice this assumption must be checked on a case-by-case basis. A simple situation where this assumption is easily verified is where $\Xi(\theta) = 0$, in which case it amounts to saying that there exist matrices close to the zero matrix that have any possible rank. More generally the assumption imposes a similar structure on the set $\Gamma(L)$ relative to some point in Θ.

Theorem 3. *Under Assumptions 1–3 and assuming that g is strictly decreasing in L, then* (i) *if $f(N) \to \infty$ then $\lim_{N\to\infty} P(\hat{L} > L_0) = 0$ and if $f(N)/N \to 0$ then*

[10] Note that we have not defined the g functions to be equal to the number of free parameters and instead use the number of degrees of freedom from the distribution of the test statistic. It is easy to see, however, that these only differ by a constant that does not depend on L so that one would obtain identical estimates of L from either definition of g.

$\lim_{N\to\infty} P(\hat{L} < L_0) = 0$; and (ii) $\lim_{N-\infty} P(\hat{L} > L_0) = 0$ for all possible L_0 only if $f(N) \to \infty$ and under Assumption 5 $\lim_{N\to\infty} P(\hat{L} < L_0) = 0$ for all possible L_0 only if $f(N)/N \to 0$.

The first part of Theorem 1 shows that the two conditions $f(N) \to \infty$ and $f(N)/N \to 0$ are sufficient to give rise to a weakly consistent estimator for L_0 in the class of models satisfying Assumption 3. The second part shows that these are also necessary under the additional restriction in Assumption 5. It should be noted that when $\Xi = 0$ neither of the restrictions provided by Assumptions 3 and 5 are required in order to show that the conditions on $f(N)$ are necessary and sufficient for weakly consistent model selection criteria. The result is split up to show which different features of $f(N)$ allow one to avoid overestimation and underestimation. This shows the potential tradeoffs that must be made when choosing the penalty function $f(N)$. There is, however, a wide range of possible penalty functions that give rise to weakly consistent model selection criteria.

To obtain the range of model selection criteria giving almost sure convergence of $\hat{L}$ to L, we have to first limit the almost sure behaviour of $\hat{\Pi}$ to ensure that it obeys a form of the Law of the Iterated Logarithm (LIL), as in Hannan and Quinn (1979), and second to strengthen the conditions on $f(N)$. Rather than writing down primitive conditions that guarantee that $\hat{\Pi}$ satisfies some LIL, which is difficult since different applications may give rise to different precise conditions, we follow Nishii (1988) and assume that the LIL holds. An example of conditions that guarantee that the LIL holds is provided by Wei (1985) in the case where Π is a matrix of regression coefficients. In other applications one can invoke similar results. Thus we make:

Assumption 6. $\hat{\Pi}$ satisfies the Law of the Iterated Logarithm, so that

$$\limsup \left(\frac{N}{\phi \log\log N}\right)^{1/2} W^{-1/2}(\hat{\pi} - \pi_0) \leqslant \iota_{GK} \tag{5}$$

with probability 1, where ϕ is some finite constant, $\leqslant$ is used to denote element-by-element weak inequality, and ι_{GK} is a vector of ones.

This assumption and the proof of Theorem 1 will enable demonstration of a LIL-type result for the test statistic which will provide a lower bound on the rate at which $N\hat{C}(\hat{\Pi}, L)$ converges to ∞ for $L < L_0$ to ensure strong consistency. The result characterizing the class of f functions that produce strongly consistent estimators of L follows.

Theorem 4. Under Assumptions 1–3 and 6 and assuming that g is strictly decreasing in L, then $\hat{L} \rightarrow L_0$ almost surely if both

$$\liminf \frac{f(N)}{\log\log N} = c_1 > \phi \max_{L_0} \max_{L \neq L_0} \left\{ \frac{GK}{g(L) - g(L_0)} : L > L_0 \right\}$$

(where c_1 may be infinite) and $f(N)/N \rightarrow 0$.

This result shows that f must increase at least at rate $\log\log N$ for the estimates to be strongly consistent. This is obviously more stringent than the rate needed for weak consistency, where $f(N)^{-1} = o(1)$ suffices, although it is still quite slow. Similar results for model selection in a regression context were proven by Pötscher (1989), for time-series models by Hannan and Quinn (1979) and Hannan (1981), and for model selection in a MLE context by Nishii (1988). The condition on the fastest rate of increase of f is the same as that in Theorem 3.

These are sufficient conditions, although under Assumption 5 $f(N)/N \rightarrow 0$ is necessary and sufficient to guarantee that one does not underestimate L. Although AIC satisfies condition $f(N)/N \rightarrow 0$, it is not consistent since it fails to satisfy condition $f(N) \rightarrow \infty$ as required in both Theorems 3 and 4. As a result there will generally be a positive probability of *overestimating* the rank when one uses AIC. The Schwarz criterion, known as BIC, satisfies both conditions in Theorem 4 and is therefore a strongly consistent model selection procedure. Theorem 4 suggests that there may be a range of possible criteria. Hannan (1980), in the context of determining the order of an ARMA process, has termed the formula $f(N) = c \log\log N$ for some number c satisfying the first condition in Theorem 4 as being a minimal strong consistency (MSC) approach. The lower bound on this choice is independent of L_0, thus guaranteeing consistency for any possible true rank.

3.2. Sequential hypothesis testing

The second approach to estimating the rank uses a sequential testing procedure whereby one chooses the estimate of L_0 to be the smallest value of L for which the test discussed in the previous section does not reject the null hypothesis, H_0: $\Pi_0 \in \Gamma(L)$. Although this is simple, it will not in general lead to a consistent estimate of L_0 unless some adjustment is made to the significance level, since there is always a type 1 error when one tests a true hypothesis. This type of problem occurs in a number of similar areas. Examples are estimating the order of an ARMA process and estimating the best regression model from a finite set of possible models.

The problem of overestimating L can be overcome if we allow the significance level of the test to change with the sample size as suggested by Pötscher (1983) in connection with ARMA models and by Bauer, Pötscher, and Hackl (1988) in the case of model selection. Letting α_N be the significance level, denote the estimator of L_0 by $\hat{L}$ where $\hat{L}$ is the smallest value of L such that using the statistic $N\hat{C}(\hat{\Pi}, L)$, one does not reject H_0: $\Pi_0 \in \Gamma(L)$ using significance level α_N and critical values obtained from the $\chi^2((K-L)(G-L)-M)$ distribution. The following weak consistency result can be proved using methods similar to those of Pötscher (1983).

Theorem 5. Given Assumptions 1–3 and letting α_N be such that (i) $\alpha_N \to 0$ *and* (ii) $-\log\alpha_N/N \to 0$, *then $\hat{L}$ is weakly consistent for L_0.*

4. Applications

As mentioned in the introduction, two important simple examples, where one is interested in the rank of a matrix directly (and so $\Xi = 0$ is imposed), occur in simultaneous-equation models as discussed in more detail in Cragg and Donald (1993) and in determining the rank of demand systems as considered in Lewbel (1991). In this section we discuss how the methods in this paper can be used in two further applications which illustrate nonzero specifications for Ξ. Also, we discuss in the factor analysis application how to adapt the results to deal with a singular variance–covariance matrix of the estimates.

4.1. Testing arbitrage pricing theory

Recent work in finance, such as McElroy and Burmeister (1988), has considered an approach to estimating and testing an arbitrage pricing model of asset returns that is based on equating factors with observed macroeconomic variables. Extending this line of work, Cragg and Donald (1992) have shown that the model may still be estimated and tested even if observed macroeconomic variables are only proxies which at best measure the factors with error. In particular, suppose that the return on asset i in time t, r_{it}, has a factor structure,

$$r_{it} = \beta_{i0t} + \phi_t'\beta_i + \varepsilon_{it},$$

where ϕ_t is a L_0-vector of unobserved common factors, with $\mathrm{E}(\phi_t) = 0$, β_i is a L_0-vector of coefficients, β_{i0t} is a constant which may take on different values over time according to some identifiable specification, and ε_{it} are random disturbances. The hypothesis of the APT is that the expected return is a linear function of the factor loadings, β_i, so that

$$\mathrm{E}(r_{it}) = \beta_{i0t} = \alpha_0 + \beta_i'\alpha.$$

Instead of applying factor analysis to determine the number of factors and test the hypothesis of APT, Cragg and Donald (1992) propose using $K \geqslant L_0$ macro variables x_t to proxy the L_0 factors via

$$\phi_t = \phi_0 + \Delta' x_t + v_t .$$

Tests of the APT are then based on the coefficients in the regression of the N asset returns on the macroeconomic variables, based on a time series of T observations:

$$r_{it} = \gamma_{i0} + x_t' \gamma_i^* + \eta_{it} .$$

One may test the APT by testing the rank restrictions

$$\rho(\Gamma^*) = \rho(\Gamma - D) = L_0 ,$$

where $\Gamma^{*\prime} = [\gamma_1^* \dots \gamma_N^*]$ is the $N \times K$ matrix of regression coefficients of the macro variables, and, letting $\gamma_i' = \{\gamma_{i0}, \gamma_i^*\}$, $\Gamma' = [\gamma_1 \dots \gamma_N]$, while $D = [d_0 \iota\ 0]$ is an $N \times (K + 1)$ matrix where d_0 is a constant (to be determined) and ι is an N vector of units. In this case, D is the Ξ matrix with $\theta_0 = d_0$ and $M = 1$. If in fact $\rho(\Gamma^*) = L_0$, then there exists a selection of L_0 columns of Γ^*, denoted Γ_2^*, that has rank L_0. It is easy to verify that the identification conditions in Assumption 3 will be satisfied provided that $\rho(\iota\ \Gamma_2^*) = L_0 + 1$ or, equivalently, if the null hypothesis is true, then γ_{i0} do not all have the same nonzero value.

Under standard assumptions on the data-generating processes, the coefficient matrices can be estimated consistently and asymptotically normally distributed at the usual $\sqrt{N}$ rate. The approach taken in Cragg and Donald (1992) is to determine the rank of Γ^* using the model selection criteria mentioned in the previous section and then to determine, either through estimation or testing, whether the rank of $\Gamma - D$ is equal to this value.[11]

4.2. Factor analysis

In the factor analysis a K-vector x_i is hypothesized to depend on $L_0 \leqslant K$ unobserved latent common factors, f_i, plus an idiosyncratic random term ε_i:

$$x_i = \varkappa_0 + \Lambda f_i + \varepsilon_i ,$$

where $\rho(\Lambda) = L_0$. Inference is based on the covariance matrix of x_i, which (assuming that the factors are uncorrelated with the residuals ε_i and that

[11] Note that the use of a consistent model selection criterion to determine the rank of Γ in conjunction with a test that $\Gamma - D$ has the same rank provides an asymptotically valid procedure for testing the model; see Pötscher (1991). It should be noted, however, that if one is interested in the parameters themselves, then the use of a consistent model selection procedure in the first stage followed by estimation may result in estimators with undesirable properties; see Pötscher (1991, Section 4, Remark (iii)).

$\mathrm{E}(f_i f_i') = I$) has population value

$$\Sigma = \Lambda\Lambda' + \Psi .$$

Here $\Sigma = \mathrm{E}((x_i - \mathrm{E}(x_i))(x_i - \mathrm{E}(x_i))')$ and $\Psi = \mathrm{diag}\{\mathrm{var}(\varepsilon_i)\}$. Thus the number of factors G can be determined by determining the rank of the matrix $\Lambda\Lambda'$ or, equivalently, the rank of the matrix $\Sigma - \Psi$.[12] This fits into the framework of Section 2, since the variance–covariance matrix of x_i can be estimated from data by

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})' ,$$

and under standard conditions,

$$\sqrt{N}(\mathrm{vech}\,\hat{\Sigma} - \mathrm{vech}\,\Sigma) \to \mathrm{N}(0, V) ,$$

where vech is a vector containing the unique elements of the matrix argument. Under reasonable conditions a consistent estimator of V can be obtained by using appropriate sample moments of the x_i. For example, a natural estimate of the covariance between the (j, k) and (l, m) elements of $\hat{\Sigma}$ is given by

$$\sum_{i=1}^{N} (\hat{a}_j \hat{a}_k \hat{a}_l \hat{a}_m - \hat{\Sigma}_{jk} \hat{\Sigma}_{lm})/N ,$$

where $\hat{a}_j = x_{ij} - \bar{x}_j$.

In this case the Ψ matrix corresponds to the Ξ matrix of Section 2 and $M = K$. The number of distinct elements of Σ is $K(K+1)/2$, so some adjustments to the results of Sections 2 and 3 are required. In particular one should modify the objective function to solve[13]

$$\min_{\Gamma(L)} \{\mathrm{vech}(\hat{\Sigma} - \Sigma)' \hat{V}^{-1} \mathrm{vech}(\hat{\Sigma} - \Sigma)\} ,$$

[12] Note that this differs from the Factor Analysis method suggested in Gill and Lewbel (1992) who consider the rank of the covariance matrix between nonoverlapping subvectors of x_i. As noted, this provides a lower bound for the number of factors and all possible partitions are needed to determine the number of factors. Gill and Lewbel (1992) also suggest this as means for testing the CAPM restriction on the APT model which differs from the method suggested in Section 4.1, where imperfect proxies for the factors are used.

[13] Cragg and Donald (1995) have recently shown that this type of objective function can be used to estimate factor analysis models efficiently under a wide class of distributional assumptions on the variables. The same approach can be taken in cases where some elements of Π are known *a priori* for example when a submatrix is known to be zero, or when they and their estimates obey some other constraints.

over the set

$$\Gamma(L) = \{\Sigma\colon [\Sigma - \Psi(\psi)]v = 0, \text{ for some } v \text{ such that } v'v = I_{G-L},$$
$$\text{where } \psi \text{ belongs to a compact set and } \Sigma = \Sigma'\}.$$

It is easy to show that imposing symmetry on the solution in addition to the rank restriction results in a statistic which is χ^2 with degrees of freedom equal to

$$[(K - L)(K - L + 1)/2] - K.$$

Thus, in order for the rank restriction to be testable we need that

$$[(K - L)(K - L + 1)/2]/ > K.$$

This condition corresponds to the condition for there to exist overidentifying restrictions in the model.

5. Monte Carlo comparison

The procedures outlined in Sections 2 and 3 rely for their justification on asymptotic properties together with whatever more or less heuristic considerations led to their advancement. The questions arise as to how well any of them perform in reasonably sized samples and which of them appear to be superior.

We investigate these questions with a Monte Carlo study involving two models of some possible interest. The models are based on the results presented in Cragg and Donald (1992), involving the regression of the returns of the stocks of 60 individual companies on a large number of macro variables. Two multivariate regression models were estimated in which the population matrix of regression coefficients was of rank 5. One model had 45 dependent variables and 17 independent variables. The other model had 10 dependent variables and 6 independent variables. Thus in one model $G = 17$, $K = 45$, and in the other $G = 6$, $K = 10$. The model is such that $\Xi = 0$ and $W = P \otimes Q$. The test statistic here consists of the sum of the $G - L$ smallest roots of $N\hat{\Pi}'\hat{Q}^{-1}\hat{\Pi}$ in the metric of $\hat{P}$.

All the independent variables and the error terms were generated as independently normally distributed pseudo-random numbers.[14] Letting Y be the $N \times K$ matrix of dependent variables and X the $N \times G$ matrix of explanatory variables, the $\hat{\Pi}$ matrix is given by $\hat{\Pi} = Y'X(X'X)^{-1}$, while $\hat{P} = (Y'Y - Y'X \times (X'X)^{-1}X'Y)/(N - K)$ and $\hat{Q} = (X'X/N)^{-1}$. The vector of ordered eigenvalues

[14] All pseudo-random numbers were generated using a linear congruent generator, cf. Kennedy and Gentile (1980). Programming was done in Microsoft FORTRAN and code is available from the authors on request. The experiments were run on a 486 IBM clone.

of $N\Pi_0'Q^{-1}\Pi_0$ in the metric of P, λ, was chosen, on the basis of results of Cragg and Donald (1992), to be

$$\lambda^0 = \{0, \dots, 0, 0.21, 0.24, 0.32, 0.41, 1.81\}. \qquad (6)$$

Thus there are four nonzero roots of quite similar magnitude and one substantially larger root. These values mean that the analysis is based on regressions for which the nonzero population values of R^2 vary between 0.17 and 0.64.

Each experiment consisted of 2526 independent replications (which gives a standard error on estimated frequencies of less than 0.01, whatever the true probability, and of 0.001 when the true probability is 0.05). The replications in different experiments were also independent. The initial set of experiments used values[15] of N of 128, 256, and 1024. In using the testing criterion (TC), α was chosen to be $\kappa/\log(N)$ with κ chosen to give a significance level of 0.05 when $N = 256$.

The results are summarized[16] in Table 1. Prominent features are the very weak performance of all methods in the large model with only 128 observations and how the performance tended to improve with more observations or when the smaller model was investigated. No method pointed frequently to the correct rank (which is 5) when the large model was estimated with 128 observations. The BIC and MSC tended to point to too small a rank; the others to too large a value. The results for the small model with 128 observations are somewhat different. The AIC and TC pointed to the correct rank more that 75% of the time and the BIC and MSC usually indicated too small a rank.

Serious underestimation of the rank occurred in the large model with 256 observations when BIC and MSC were used. By contrast, the modal estimates of AIC and TC were correct, though especially in the latter case there was a tendency to overestimate the rank. The BIC was correct in more than half the cases, and the AIC and TC each produced the right answer more often than nine times out of ten.

With 1024 observations, the BIC and MSC were too small always in the large model, never in the small model. The AIC and the TC both pointed to the right rank in over 90% of the cases, and otherwise pointed towards a higher rank, usually rank 6.

We examined the extent to which the size of the population roots affects the reliability of the procedures by running the experiments with 256 observations, and roots four times larger and one-fourth the size of those used in the base

[15] The number of observations in Cragg and Donald (1992) was 240. One experiment was tried with $N = 64$. The results for the large model were wretched with no criterion ever pointing to the right rank. The likely reason is not surprising: $\hat{P}$ would be singular if it were estimated with only three fewer observations.

[16] Ranks for which entries are not reported all had only zero entries.

Table 1
Estimates of rank $-L_0 = 5$, frequencies (%) of different rank estimates

	Large model				Small model			
Rank	AIC	BIC	MSC	TC	AIC	BIC	MSC	TC
$n = 128$								
0	0.0	0.0	100.0	0.0	0.0	0.0	5.4	0.0
1	0.0	63.1	0.0	0.0	0.0	5.5	94.4	0.0
2	0.0	34.9	0.0	0.0	0.0	33.1	0.2	0.0
3	0.0	1.9	0.0	0.0	0.4	38.9	0.0	0.5
4	1.1	0.0	0.0	0.0	12.3	18.8	0.0	24.1
5	14.3	0.0	0.0	2.4	83.4	3.7	0.0	72.3
6	43.6	0.0	0.0	24.9	3.9	0.0	0.0	3.0
7	33.5	0.0	0.0	43.9				
8	6.9	0.0	0.0	24.5				
9	0.4	0.0	0.0	4.0				
10	0.0	0.0	0.0	0.2				
$n = 256$								
0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
1	0.0	98.6	0.0	0.0	0.0	0.0	92.4	0.0
2	0.0	1.4	0.0	0.0	0.0	0.3	7.6	0.0
3	0.3	0.0	0.0	0.0	0.0	4.3	0.1	0.0
4	13.7	0.0	0.0	6.0	0.0	26.0	0.0	0.2
5	59.3	0.0	0.0	46.8	93.4	69.4	0.0	95.8
6	24.8	0.0	0.0	39.5	6.6	0.0	0.0	4.0
7	1.9	0.0	0.0	7.2				
8	0.0	0.0	0.0	0.5				
$n = 1024$								
0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	38.4	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	58.8	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0
5	91.3	0.0	0.0	93.2	92.7	100.0	100.0	97.3
6	8.7	0.0	0.0	6.5	7.3	0.0	0.0	2.7
7	0.0	0.0	0.0	0.4				
17	0.0	0.0	0.0	0.0				

experiments – with the multiplicative factor indicated by the parameter v when needed in tables. The results are summarized in Table 2. Having smaller roots very strongly pushed the methods towards underestimating the rank just as the remarks after Theorem 2 suggested might occur. In the large model, both the BIC and the MSC almost always pointed to rank 0. The others usually pointed to too low a rank. Much the same also occurred in the small model. Somewhat

Table 2
Estimates of rank $-L_0 = 5$, effects of different size roots, frequencies (%) of different rank estimates

	Large model				Small model			
Rank	AIC	BIC	MSC	TC	AIC	BIC	MSC	TC
$n = 256$:	Roots 0.25 times base roots ($v = 0.25$)							
0	0.0	98.5	100.0	0.0	0.0	0.8	98.1	0.0
1	1.1	1.5	0.0	0.0	0.0	93.0	1.9	0.0
2	17.6	0.0	0.0	3.3	4.0	6.1	0.0	3.0
3	47.0	0.0	0.0	27.2	24.9	0.1	0.0	33.4
4	30.0	0.0	0.0	46.0	45.5	0.0	0.0	12.7
5	4.3	0.0	0.0	20.3	24.5	0.0	0.0	12.7
6	0.1	0.0	0.0	3.1	1.1	0.0	0.0	0.4
7	0.0	0.0	0.0	0.0				
$n = 256$:	Roots 4.0 times base roots ($v = 4.0$)							
0	0 0	0.0	64.5	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	35.5	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	7.8	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	56.8	0.0	0.0	0.0	0.0	0.0	0.0
5	33.3	35.4	0.0	18.9	92.3	100.0	100.0	94.9
6	54.9	0.0	0.0	51.0	7.7	0.0	0.0	5.1
7	11.2	0.0	0.0	26.9				
8	0.7	0.0	0.0	3.0				
9	0.0	0.0	0.0	0.2				

opposite tendencies emerged with the larger roots. Here the AIC and TC in the large model usually opted for ranks greater than 5, and the other two, while making too small estimates, were not as poor as with small roots. The small model results were very similar to those obtained in the base case.

These results suggests that the patterns are dependent not only on the number of observations and the number of parameters but also on the sizes of the nonzero roots. This is further investigated in Table 3 where we tabulate the average values across experiments in the cumulative sums of the roots,[17] that is the averages of the $N\hat{C}(\hat{\Pi}, L)$, $L = G - 1, \dots, 0$. For the large model, $N\hat{C}(\hat{\Pi}, L_0)$ is the sum of 12 roots. Asymptotic theory would lead us to expect a value of 480 or less, based on Theorem 1 and the remarks after Theorem 2, the same as the number of degrees of freedom. Instead, all entries, except the one for the experiment using the smallest roots with 256 observations, are greater than

[17] Standard errors are not reported because they are very small, for example ranging between 0.04 and 1.44 for the entries in the second column for the large model. Others are of similar magnitude.

Table 3
Sums of roots

Number of roots	D.F.	$n = 128$ $v = 1.0$	$n = 256$ $v = 0.25$	$n = 256$ $v = 1.0$	$n = 256$ $v = 4.0$	$n = 1024$ $v = 1.0$
Large model						
1	29	11.84	10.32	10.92	11.37	10.64
2	60	28.70	24.73	26.32	27.39	25.53
3	93	50.66	43.33	46.19	48.05	44.64
4	128	78.18	66.28	70.94	73.78	68.14
5	165	111.96	93.91	100.86	105.07	96.30
6	204	152.70	126.75	136.43	142.43	129.84
7	245	201.34	165.15	178.55	186.86	169.34
8	288	259.15	209.92	228.04	239.20	215.41
9	333	327.61	261.72	285.97	300.85	268.91
10	380	408.84	321.53	353.95	373.62	331.14
11	429	505.26	390.34	434.12	460.63	404.35
12	480	620.01	469.60	530.32	569.10	493.10
13	533	757.98	560.99	650.65	849.38	746.78
14	588	926.45	667.39	800.51	1216.26	1064.07
15	645	1136.47	792.78	988.36	1698.47	1472.53
16	704	1414.76	946.14	1231.75	2346.41	1998.10
17	765	1957.69	1184.39	1893.01	4742.45	4006.47
Small model						
1	5	4.41	3.60	4.80	5.05	5.00
2	12	26.10	15.21	49.92	197.10	203.08
3	21	63.63	36.45	119.21	456.32	460.49
4	32	121.04	70.02	217.86	811.14	805.57
5	45	209.48	122.59	359.10	1299.26	1263.09
6	60	486.45	262.47	863.66	3265.37	3153.25

this value.[18] By contrast, with the smallest roots, there is a significant underestimate. In addition, for all the sums involving less than thirteen terms, the averages for given population root sizes decline as the number of observations increases and they decline with increasing size of roots when $N = 256$, again in all cases significantly so at the 0.01 level. Furthermore, the suggestion that the sum of roots when nonzero population roots are also involved would follow the noncentral χ^2 distribution received very little support in the sense that all values differ significantly from the value arising from the corresponding noncentral χ^2 distribution.

[18] In all cases the entries are significantly greater at the 0.01 level. Since the asymptotic results would be exact in this model if W were used rather than $\hat{W}$, the difficulties here arise from needing to estimate W.

The results in the small model are closer to what asymptotic theory might lead one to expect. Here, with 1024 observations, the mean of the smallest root happened to equal the degrees of freedom to beyond two decimals. There was no significant difference when the large roots were used with 256 observations. In the other cases, the sum was significantly smaller than 5.

A number of conclusions emerge from this part of the Monte Carlo experiments:

- The performance in estimating the rank is affected by the values of the nonzero population eigenvalues and by the size of the model.
- While it appears that the BIC and MSC criteria are far too severe in their premia for parsimony, making no adjustment of the premium at all with increasing numbers of observations does lead to the AIC's exhibiting some of the weakness predicted by the way in which it is inconsistent. By contrast, the decline in α values used in the particular version of TC implemented here might appear not to be adequately rapid.

6. Conclusion

We have been exploring using a particular test statistic for testing the hypothesis that a matrix has a particular rank and for estimating the rank of the matrix in cases where that is unknown. The situation being contemplated is more general than has usually been investigated and the statistic is asymptotically equivalent to the one commonly used, including the usual likelihood based one. The form of the problem presents some nonstandard features in defining what quantities the test statistic is investigating when the hypothesized rank is too small and also when it is larger than the actual rank. We demonstrate consistency of the test when too small a rank is hypothesized and suggest that the test is conservative when the hypothesized rank is larger than the actual rank in the sense that the asymptotic size of the test is smaller than that given by the asymptotic distribution of the statistic when the hypothesized rank is correct. Local power considerations suggest that this may also be true when the rank is almost smaller than a true hypothesized value.

Armed with these results, we then consider using model selection and sequential testing procedures based on the statistic to estimate the rank and establish conditions for the estimates to be strongly and weakly consistent. We next consider how the procedures work in finite samples in two examples through a Monte Carlo study. Two things stand out. Not surprisingly, the test statistic does not behave entirely as suggested by the asymptotic theory, probably mainly because of the need to use an estimated covariance matrix; but in models which are not too large relative to the number of observations and in which the population matrix is not too near to being of smaller rank, the performance indicated by the asymptotic theory seems to be quite useful. Second, it emerges that the parsimony penalties suggested by standard consistent model selection criteria are too stringent in moderate sized samples, leading to underestimation of the rank, though the usual problem of the Akaike criterion overestimating is evident. Judicious choice of significance level dependent on sample size does lead to sequential testing being a useful estimation procedure. This is encouraging, since as the substantive examples discussed indicate, there are serious problems which fit the framework used in this paper.

Appendix of proofs

In the proofs we use $\|.\|$ to denote the Euclidean norm in the appropriate Euclidean space. $B(., \delta)$ and $\bar{B}(., \delta)$ are used to denote the open and closed balls around the first argument with radius δ. Δ will denote a generic large finite constant. Finally, let $V(L) = \{v\colon\ v'v = I_{G-L}\}$ and define the function, for $v \in V(L)$,

$$F_L(\Pi, \theta, v) = (\Pi - \Xi(\theta))v\,.$$

Proof of Lemma 1

Let $d_W(\pi_0, \pi) = \{(\pi_0 - \pi)' W^{-1}(\pi_0 - \pi)\}^{1/2}$. Note that d generates the same topology as the Euclidean metric in R^{KG}. Therefore, $d_W(\pi_0, \pi) \geqslant 0$ and $d_W(\pi_0, \pi) = 0$ if and only if $\pi = \pi_0$. Next we show that $\Gamma(L)$ is a closed set. To do so let π_i be a sequence in $\Gamma(L)$ such that $\pi_i \to \bar{\pi}$. Then we must show that $\bar{\pi} \in \Gamma(L)$. Since, $\pi_i \in \Gamma(L)$, then for each i there is a $\theta_i \in \Theta$ and $v_i \in V(L)$ for which $F_L(\Pi_i, \theta_i, v_i) = 0$. Since Θ and $V(L)$ are compact in Euclidean space, there is a subsequence (θ_{i_j}, v_{i_j}) that converges to a point, $(\bar{\theta}, \bar{v})$ say, in $\Theta \times V(L)$. But π_{i_j} must also converge to $\bar{\pi}$. Then by construction and the continuity of the function F in all its arguments,

$$\lim_{i_j \to \infty} F_L(\Pi_{i_j}, \theta_{i_j}, v_{i_j}) = F_L(\bar{\Pi}, \bar{\theta}, \bar{v}) = 0\,,$$

so that $\bar{\pi} \in \Gamma(L)$ and so $\Gamma(L)$ is closed. Since, $\pi_0 \in \Gamma(L)$ for $L \geqslant L_0$, (i) follows. (ii) follows from the facts that $\Gamma(L)$ is closed and π_0 is not in $\Gamma(L)$ for $L < L_0$. □

Proof of Lemma 2

Using Assumptions 1 and 2 we first show that for large enough N any solution to either (1) or (2) is almost surely contained inside a compact set. In the notation of the proof of Lemma 1, $\hat{C}(\hat{\Pi}, L) = \min_{\pi \in \Gamma(L)} d_{\hat{W}}(\hat{\pi}, \pi)^2$. Now, suppose that $\tilde{\pi}$ is a minimizer (over $\Gamma(L)$). Then,

$$d_{\hat{W}}(0, \tilde{\pi}) \leqslant d_{\hat{W}}(0, \hat{\pi}) + d_{\hat{W}}(\hat{\pi}, \tilde{\pi}) \leqslant d_{\hat{W}}(0, \hat{\pi}) + d_{\hat{W}}(\hat{\pi}, \pi^*),$$

where π^* is an arbitrary fixed element of $\Gamma(L)$. The right-hand side converges almost surely to $d_W(0, \pi_0) + d_W(\pi_0, \pi^*)$. Let $\Delta > d_W(0, \pi_0) + d_W(\pi_0, \pi^*)$, we have that $d_{\hat{W}}(0, \tilde{\pi}) \leqslant \Delta$ for large N almost surely. Since $\| \tilde{\pi} \| \leqslant d_{\hat{W}}(0, \tilde{\pi}) / \lambda_{\min}^{1/2}(\hat{W}^{-1})$ and since $\lambda_{\min}^{1/2}(\hat{W}^{-1}) \xrightarrow{a.s.} \lambda_{\min}^{1/2}(W^{-1})$, then for a constant $\Delta' > \Delta / \lambda_{\min}^{1/2}(W^{-1})$ we have that $\| \tilde{\pi} \| \leqslant \Delta'$ for large N almost surely, so that the minimum of (1) occurs on the compact ball $\bar{B}(0, \Delta')$ for large N almost surely. A similar argument shows that the minimum of (2) occurs on another compact ball around the origin. Thus for large enough N the minima of (1) and (2) occur in some compact ball, B say, almost surely.

Next we show that there is uniform convergence on the compact set $B \cap \Gamma(L) = \bar{\Gamma}(L)$:

$$\sup_{\pi \in \bar{\Gamma}(L)} |(\hat{\pi} - \pi)' \hat{W}^{-1}(\hat{\pi} - \pi) - (\pi_0 - \pi)' W^{-1}(\pi_0 - \pi)| \xrightarrow{a.s.} 0.$$

Note that

$$\begin{aligned}
&\sup_{\pi \in \bar{\Gamma}(L)} |(\hat{\pi} - \pi)' \hat{W}^{-1}(\hat{\pi} - \pi) - (\pi_0 - \pi)' W^{-1}(\pi_0 - \pi)| \\
&\leqslant \sup_{\pi \in \bar{\Gamma}(L)} |(\hat{\pi} - \pi_0)' \hat{W}^{-1}(\hat{\pi} - \pi)| + \sup_{\pi \in \bar{\Gamma}(L)} |(\pi_0 - \pi)' \hat{W}^{-1}(\hat{\pi} - \pi_0)| \\
&\quad + \sup_{\pi \in \bar{\Gamma}(L)} |(\pi - \pi_0)'(\hat{W}^{-1} - W^{-1})(\pi - \pi_0)|.
\end{aligned}$$

For each of the three terms on the right-hand side we have, respectively,

$$\begin{aligned}
\sup_{\pi \in \bar{\Gamma}(L)} |(\hat{\pi} - \pi_0)' \hat{W}^{-1}(\hat{\pi} - \pi)| &\leqslant \| \hat{\pi} - \pi_0 \| \lambda_{\max}(\hat{W}^{-1}) \\
&\quad \times \left(\| \hat{\pi} \| + \sup_{\pi \in \bar{\Gamma}(L)} \| \pi \| \right),
\end{aligned}$$

$$\begin{aligned}
\sup_{\pi \in \bar{\Gamma}(L)} |(\pi_0 - \pi)' \hat{W}^{-1}(\hat{\pi} - \pi_0)| &\leqslant \| \hat{\pi} - \pi_0 \| \lambda_{\max}(\hat{W}^{-1}) \\
&\quad \times \left(\| \pi_0 \| + \sup_{\pi \in \bar{\Gamma}(L)} \| \pi \| \right),
\end{aligned}$$

$$\begin{aligned}
&\sup_{\pi \in \bar{\Gamma}(L)} |(\pi - \pi_0)'(\hat{W}^{-1} - W^{-1})(\pi - \pi_0)| \\
&\leqslant \max\{ |\lambda_{\min}(\hat{W}^{-1} - W^{-1})|, |\lambda_{\max}(\hat{W}^{-1} - W^{-1})| \} \\
&\quad \times \left(\| \pi_0 \|^2 + 2 \| \pi_0 \| \sup_{\pi \in \bar{\Gamma}(L)} \| \pi \| + \sup_{\pi \in \bar{\Gamma}(L)} \| \pi \|^2 \right).
\end{aligned}$$

By Assumption 2 and the compactness of $\bar{\Gamma}(L)$, there exists a large enough N for which each of these terms can be made arbitrarily small so that uniform convergence almost surely follows. Using standard consistency arguments (see Gallant, 1987, p.180, for example) the results follow by Lemma 1. □

Proof of Theorem 1

Part (ii) is obvious from Lemmas 1 and 2.

Part (*i*): By Lemma 2, for any open (relative to $\Gamma(L_0)$) ball around π_0, say $U(\delta) = B(\pi_0, \delta) \cap \Gamma(L_0)$ for any $\delta > 0$, for large enough N the minimum of $(\hat{\pi} - \pi)' \hat{W}^{-1} (\hat{\pi} - \pi)$ over this neighbourhood equals $\hat{C}(\hat{\Pi}, L_0)$ with probability 1. As noted in the text Π_0^1 can be written as a function of μ_0. First, we show that, for any π in the neighbourhood $U(\delta)$ for sufficiently small δ, there is a neighbourhood of μ_0 for which we can write π as a function of μ in this neighbourhood and, moreover, that as δ becomes small, the neighbourhood of μ_0 becomes small. To show this let, for any $\pi \in U(\delta)$, $\theta(\pi)$ be the set of $\theta \in \Theta$ such that $F_{L_0}(\Pi, \theta, v) = 0$ for some $v \in V(L_0)$. Next, using the same argument used to show that $\Gamma(L)$ is closed, we can show that $\theta(\pi)$ is a closed set for any $\pi \in U(\delta)$. Since $\theta(\pi) \subset \Theta$, we have that $\theta(\pi)$ is compact.

Next, it is the case that for any sequence $\pi_i \in \Gamma(L_0)$ converging to π_0,

$$\sup_{\theta \in \theta(\pi_i)} \|\theta - \theta_0\| = \max_{\theta \in \theta(\pi_i)} \|\theta - \theta_0\| \to 0.$$

To see this, suppose there is a sequence π_i converging to π_0 for which this does not hold. Then there must exist some subsequence (θ_{i_j}, v_{i_j}) which converges to some point $(\bar{\theta}, v) \in \Theta \times V(L_0)$ such that $\bar{\theta} \neq \theta_0$. But by continuity of F_{L_0} we would have

$$0 = \lim_{i_j \to \infty} F_{L_0}(\Pi_{i_j}, \theta_{i_j}, v_{i_j}) = F_{L_0}(\Pi_0, \bar{\theta}, \bar{v}),$$

contradicting the uniqueness condition in Assumption 3. Thus, letting

$$\varepsilon(\delta) = \sup_{\pi \in U(\delta)} \sup_{\theta \in \theta(\pi)} \|\theta - \theta_0\|,$$

$\varepsilon(\delta) \to 0$ as $\delta \to 0$.

Next note that in a neighbourhood of π_0 (relative to $\Gamma(L_0)$) and in a neighbourhood of θ_0 $\rho(\Phi^2) = \rho(\Pi^2 - \Xi^2(\theta)) = L_0$ so that, since $\pi \in \Gamma(L_0)$, $\Phi^1 = \Phi^2 v$, where $v = (\Phi^{2\prime} \Phi^2)^{-1} \Phi^{2\prime} \Phi^1$. Since this is a continuous function of Π and θ, then there exist neighbourhoods of π_0 and θ_0 such that v is in a neighbourhood of $v_0 = (\Phi_0^{2\prime} \Phi_0^2)^{-1} \Phi_0^{2\prime} \Phi_0^1$. Thus we can find a $\delta > 0$, such that if $\pi \in U(\delta)$ then we can express $\Pi^1 = (\Pi^2 - \Xi^2(\theta))v + \Xi^1(\theta)$ for $\pi^2 \in B(\pi_0^2, \delta)$, $\theta \in B(\theta_0, \varepsilon(\delta))$, and $\mathrm{vec}(v) \in B(\mathrm{vec} v_0, \eta(\delta))$ for $\varepsilon(\delta) > 0$ and $\eta(\delta) > 0$. Letting $\mu = (\theta', \mathrm{vec}(v)', \pi^{2\prime})'$

denote by $\mathcal{N}(\mu_0)$ the neighbourhood $B(\theta_0, \varepsilon(\delta)) \times B(\mathrm{vec}(v_0), \eta(\delta)) \times B(\pi_0^2, \delta)$. Assumption 3 implies that any $\pi \in U(\delta)$ can be expressed as a unique function of $\mu \in \mathcal{N}(\mu_0)$, $\pi(\mu)$, where

$$(\Pi^1, \Pi^2) = [(\Pi^2 - \Xi^2(\theta))v + \Xi^1(\theta), \Pi^2],$$

and moreover that $\pi(\mu)$ is twice continuously differentiable. Since $\hat{\Pi}^L$ is contained in the appropriate neighbourhood of Π_0 for large enough N almost surely, then for enough N with probability 1,

$$\hat{C}(\hat{\Pi}, L_0) = \min_{\mu \in \mathcal{N}(\mu_0)} (\hat{\pi} - \pi(\mu))' \hat{W}^{-1} (\hat{\pi} - \pi(\mu)), \tag{A.1}$$

where

$$\pi(\mu) = \begin{pmatrix} (v' \otimes I)(\pi^2 - \mathrm{vec}(\Xi_2(\theta))) + \mathrm{vec}(\Xi_1(\theta)) \\ \pi^2 \end{pmatrix}.$$

Also, using the above arguments and Lemmas 1 and 2, the minimizer $\hat{\mu} \xrightarrow{a.s.} \mu_0$.

Standard expansions such as those in Chamberlain (1982, Proposition 8) and Gourieroux and Monfort (1989) may then be used. Defining

$$B(\mu) = \frac{\partial}{\partial \mu} \pi(\mu) = \begin{pmatrix} \dfrac{\partial}{\partial(\theta, \mathrm{vec}(v))} \pi^1(\mu) & (v' \otimes I) \\ 0 & I \end{pmatrix},$$

we have by Assumption 3(ii) that $\rho(B(\mu_0)) = KL_0 + L_0(G - L_0) + M$. First note that expanding $\sqrt{N}(\hat{\pi} - \pi(\hat{\mu}))$ about μ_0,

$$\sqrt{N}(\hat{\pi} - \pi(\hat{\mu})) = \sqrt{N}(\hat{\pi} - \pi(\mu_0)) + \sqrt{N}\, B(\mu_0)(\hat{\mu} - \mu_0) + o_p(1),$$

and substituting into $N\hat{C}(\hat{\Pi}, L_0)$, we get

$$\begin{aligned} N\hat{C}(\hat{\Pi}, L_0) &= N(\hat{\pi} - \pi(\hat{\mu}))' \hat{W}^{-1} (\hat{\pi} - \pi(\hat{\mu})) \\ &= N(\hat{\pi} - \pi(\mu_0))' \hat{W}^{-1} (\hat{\pi} - \pi(\mu_0)) \\ &\quad - 2N(\hat{\pi} - \pi(\mu_0))' \hat{W}^{-1} B(\mu_0)(\hat{\mu} - \mu_0) \\ &\quad + N(\hat{\mu} - \mu_0)' B(\mu_0)' \hat{W}^{-1} B(\mu_0)(\hat{\mu} - \mu_0) + o_p(1). \end{aligned}$$

Then expanding the first-order conditions for minimizing (A.1) yields

$$\begin{aligned} 0 &= \sqrt{N} B(\hat{\mu})' \hat{W}^{-1} (\hat{\pi} - \pi(\hat{\mu})) \\ &= \sqrt{N} B(\mu_0)' \hat{W}^{-1} (\hat{\pi} - \pi(\mu_0)) - \sqrt{N}\, B(\mu_0)' \hat{W}^{-1} B(\mu_0)(\hat{\mu} - \mu_0) + o_p(1), \end{aligned}$$

so that

$$\sqrt{N}(\hat{\mu} - \mu_0) = (B(\mu_0)' \hat{W}^{-1} B(\mu_0))^{-1} B(\mu_0)' \hat{W}^{-1} \sqrt{N}(\hat{\pi} - \pi(\mu_0)) + o_p(1).$$

Substituting for $\sqrt{N}(\hat{\mu}-\mu_0)$ in the expansion for $N\hat{C}(\hat{\Pi}, L_0)$, we get

$$\begin{aligned} N\hat{C}(\hat{\Pi}, L_0) &= N(\hat{\pi}-\pi_0)'\,\hat{W}^{-1/2}(I-\hat{A}(\hat{A}'\hat{A})^{-1}\hat{A}')\,\hat{W}^{-1/2}(\hat{\pi}-\pi_0)+o_p(1) \\ &= N(\hat{\pi}-\pi_0)'\,W^{-1/2}(I-A(A'A)^{-1}A')\,W^{-1/2}(\hat{\pi}-\pi_0)+o_p(1) \\ &\xrightarrow{d} \chi^2(GK-\rho(B(\mu_0))), \end{aligned}$$

where $\hat{A}=\hat{W}^{-1/2}B(\mu_0)$ and $A=W^{-1/2}B(\mu_0)$ and the result follows.

When $\Xi=0$, the proof is similar except that Assumption 3 is not needed, the arguments showing (A.1) are simpler (not requiring the neighbourhood of θ_0), and in the expansion,

$$B(\mu)=\begin{pmatrix}(I\otimes\Pi^2) & (v'\otimes I)\\ 0 & I\end{pmatrix}$$

and the degrees of freedom are $(G-L)(K-L)$.

Part (*ii*): Obvious from Lemmas 1 and 2.

Part (*iii*): This follows from the fact that $\Gamma(L_0)\subset\Gamma(L)$ for $L>L_0$. □

Proof of Theorem 2

The proof is essentially the same as that of Theorem 1 (i), except that one expands around the point $\bar{\pi}_{0,N}$. □

Proof of Theorem 3

(i) Note that $\hat{L}>L_0$ (remembering that since there is the possibility of ties, this is interpreted as meaning that one of the minimizers of $S(L)$ is larger than L_0) implies that, for some $L>L_0$, $S(L)\leqslant S(L_0)$. Thus, $P(\hat{L}>L_0)\leqslant\sum_{L=L_0+1}^{G}P(S(L)\leqslant S(L_0))$. But, since

$$S(L_0)-S(L)=f(N)^{-1}N(\hat{C}(\hat{\Pi},L_0)-\hat{C}(\hat{\Pi},L))+g(L)-g(L_0),$$

we have for $L>L_0$ that

$$P\{S(L_0)\geqslant S(L)\}\leqslant P\{N\hat{C}(\hat{\Pi},L_0)\geqslant f(N)(g(L_0)-g(L))\}\to 0,$$

because $f(N)\to\infty$, $g(L)<g(L_0)$, and $\hat{C}(\hat{\Pi},L_0)$ is $\chi^2((G-L_0)(K-L_0))$ asymptotically by Theorem 1. Thus, $f(N)\to\infty$ implies that $P(\hat{L}>L_0)\to 0$. Similarly, for $L<L_0$,

$$\begin{aligned} P\{S(L)\leqslant S(L_0)\} &= P\{f(N)^{-1}N(\hat{C}(\hat{\Pi},L)-\hat{C}(\hat{\Pi},L_0)) \\ &\qquad + g(L)-g(L_0)\leqslant 0\}\to 0, \end{aligned}$$

by Lemma 1 so that $f(N)/N\to 0$ implies that $P(\hat{L}<L_0)\to 0$.

(ii) Suppose that $L_0 < G$. Note that $S(G) < S(L_0)$ implies that $\hat{L} \neq L_0$ so that $P(S(G) < S(L)) \leqslant P(\hat{L} \neq L_0)$. Suppose $f(N)$ has a subsequence that is bounded by Δ. Since $\hat{C}(\hat{\Pi}, G) = 0$, then along this subsequence,

$$\lim P(S(G) < S(L_0)) \geqslant \lim P(N\hat{C}(\hat{\Pi}, L_0) > \Delta(g(L_0) - g(L)))$$

$$= P\{\chi^2((G - L_0)(K - L_0) - M) > \Delta(g(L) - g(L_0))\} > 0.$$

This subsequence satisfies $f(N)/N \to 0$ so that $P(\hat{L} < L_0) \to 0$ by (i), so using the fact that

$$P(\hat{L} > L_0) = P(\hat{L} \neq L_0) - P(\hat{L} < L_0) \geqslant P(S(G) < S(L_0)) - P(\hat{L} < L_0),$$

we have lim sup $P(\hat{L} > L_0) > 0$ so that $P(\hat{L} = L_0) \not\to 0$. Finally, suppose that $f(N)/N \not\to 0$, then these is a subsequence along which $f(N)/N \geqslant c > 0$. Since, along this subsequence $f(N) \to \infty$, we have by (i) that lim $P(\hat{L} > L_0) = 0$. Note that $P(S(L) < S(L_0)) \leqslant P(\hat{L} \neq L_0)$, and for $L < L_0$ we have

$$P\{S(L) < S(L_0)\} = P\left\{\frac{f(N)}{N} - \frac{1}{g(L) - g(L_0)}(\hat{C}(\hat{\Pi}, L) - \hat{C}(\hat{\Pi}, L_0)) > 0\right\},$$

and along the appropriate subsequence,

$$\operatorname{plim}\frac{f(N)}{N} - \operatorname{plim}\frac{1}{g(L) - g(L_0)}(\hat{C}(\hat{\Pi}, L) - \hat{C}(\hat{\Pi}, L_0))$$

$$\geqslant c - \frac{1}{g(L) - g(L_0)}C(\Pi_0, L),$$

where

$$C(\Pi_0, L) = \min_{\Gamma(L)}(\pi_0 - \pi)' W^{-1}(\pi_0 - \pi),$$

and it will be shown that this can be made arbitrarily small by appropriate choice of π_0. By Assumption 5 we can let $\Pi_0 = A + \Xi(\theta_0)$, where trace $(A'A)^{1/2} \leqslant \varepsilon$, and such that Assumption 3 is satisfied. Thus, $\pi_0 \in \Gamma(L_0)$, but $\pi_0 \notin \Gamma(L)$ for $L < L_0$. On the other hand, let $\pi = \mathrm{vec}(\Xi(\theta_0))$ and note that $\pi \in \Gamma(L)$ for all L, so that using these values we have that

$$C(\Pi_0, L) \leqslant (\mathrm{vec}(A))' W^{-1}(\mathrm{vec}(A)) \leqslant \lambda_{\min}^{-1}(W)\varepsilon^2,$$

which can be made small by appropriate choice of ε. Therefore, lim sup $P(\hat{L} < L_0) = 1$ using similar arguments to those used previously so that $P(\hat{L} < L_0) \not\to 0$. □

Proofs of Theorem 4

Note that (b) can be shown to be sufficient to rule out the limit of $\hat{L}$ being $L < L_0$ in a similar way as in the proof of Theorem 3, by changing convergence in probability to almost sure convergence. To show that the limit of $\hat{L}$ will almost surely be no greater than L_0, note that for any $L > L_0$,

$$P(\limsup (S(L_0) - S(L)) < 0)$$

$$\geqslant P\{\limsup N\hat{C}(\hat{\Pi}, L_0) < \liminf f(N) g(L) - g(L_0))\}$$

$$\geqslant P\left\{\limsup \frac{N}{\phi \log\log N} \hat{C}(\hat{\Pi}, L) < \liminf \frac{f(N)}{\phi \log\log N}(g(L) - g(L_0))\right\} = 1$$

by condition (a), since under Assumptions 1, 2, 3, and 6,

$$\limsup \frac{N\hat{C}(\hat{\Pi}, L_0)}{\phi \log\log N} \leqslant \iota_{GK}'(I - W^{-1/2} B_0(B_0' W^{-1} B_0)^{-1} B_0 W^{-1/2})\iota_{GK} \leqslant GK,$$

with probability 1, where $B_0 = B(\mu_0)$. □

Proof of Theorem 5

Let E_L denote the event that one can reject the hypothesis H_0: $\Pi \in \Gamma(L)$ for the specified value of L using the test statistic $N\hat{C}(\hat{\Pi}, L)$ and a critical value based on a sequence α_N satisfying the conditions of the theorem. (Note that E_L also depends on N.) Also let $\bar{E}_L$ denote the event that one cannot reject the hypothesis. Let $c(\alpha_N, L)$ denote the critical value corresponding to the significance level α_N for testing the hypothesis corresponding to L. By Pötscher (1983, Thm. 5.8), if $\alpha_N \to 0$, then $c(\alpha_N, L) \to \infty$, and if $-\log \alpha_N/N \to 0$, then $c(\alpha_N, L)/N \to 0$. In general, letting $\hat{L}$ denote the estimator of L_0, then for any value of L,

$$P(\hat{L} = L) = P(E_0 \cap E_1 \cap \ldots \cap \bar{E}_L).$$

Then for any $L < L_0$,

$$P(\hat{L} = L) \leqslant P(\bar{E}_L) = 1 - P(E_L).$$

But

$$P(E_L) = P(\hat{C}(\hat{\Pi}, L) > c(\alpha_N, L)/N) \to 1,$$

since $\hat{C}(\hat{\Pi}, L) \xrightarrow{p} C(\Pi_0, L) > 0$ for $L < L_0$ and $c(\alpha_N, L)/N \to 0$. Thus, $\lim_{N\to\infty} P(\hat{L} < L_0) = 0$. Now, for $L > L_0$ we have that

$$P(\hat{L} > L_0) \leqslant P(E_{L_0}) = P(N\hat{C}(\hat{\Pi}, L_0) > c(\alpha_N, L_0)).$$

Fix any $\varepsilon > 0$ and let Δ be such that $P(\chi^2((K - L_0)(G - L_0) - M) > \Delta) = \varepsilon$. Since $c(\alpha_n, L_0) \to \infty$, there exists a $\bar{N}$ such that, for all $N > \bar{N}$, $c(\alpha_N, L_0) > \Delta$. Also since by Theorem 1 we have that

$$N\hat{C}(\hat{\Pi}, L_0) \xrightarrow{d} \chi^2((K - L_0)(G - L_0) - M),$$

then there exists a $\tilde{N}$ such that for all $N > \tilde{N}$ we have

$$|P(N\hat{C}(\hat{\Pi}, L_0) > \Delta) - P(\chi^2((K - L_0)(G - L_0) - M) > \Delta)| \leqslant \varepsilon.$$

Then, for $N > \max\{\bar{N}, \tilde{N}\}$, we have by construction

$$P(E_{L_0}) \leqslant P(N\hat{C}(\hat{\Pi}, L_0) > \Delta) \leqslant 2\varepsilon,$$

and since ε is arbitrary, we have that lim $P(\hat{L} > L_0) = 0$. Thus we have that lim $P(\hat{L} = L_0) = 1$. □

References

Akaike, Hirotugu, 1987, Factor analysis and AIC, Psychometrika 52, 317–332.

Anderson, T.W., 1958, An introduction to multivariate statistical analysis (Wiley, New York, NY).

Anderson, T.W. and H. Rubin, 1956, Statistical inference in factor analysis, in: J. Neyman, ed., Proceedings of the third Berkeley symposium on mathematical statistics and probability, Vol. 5 (University of California Press, Berkeley, CA) 111–150.

Andrews, D.W.K., 1989, Generic uniform convergence, Cowles Foundation discussion paper 940 (Yale University, New Haven, CT).

Atkinson, A.C., 1981, Likelihood ratios, posterior odds and information criteria, Journal of Econometrics 16, 15–20.

Bauer, P., B.M. Pötscher, and P. Hackl, 1988, Model selection by multiple tests, Statistics 19, 39–44.

Bozdogan, Hamparsum, 1987, Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, Psychometrika 52, 345–370.

Chamberlain, G., 1982, Multivariate regression models for panel data, Journal of Econometrics 18, 5–46.

Cragg, J.G. and S.G. Donald, 1992, Testing and determining arbitrage pricing structure from regressions on macro variables, Discussion paper 92-14 (University of British Columbia, Vancouver).

Cragg, J.G. and S.G. Donald, 1993, Testing identifiability and specification in instrumental variables models, Econometric Theory 9, 222–240.

Cragg, J.G. and S.G. Donald, 1995, Factor analysis under more general conditions with reference to heteroskedasticity of unknown form, in: G.S. Maddala, Peter C.B. Phillips, and T.N. Srinivasan, Advances in econometrics and quantitative economics: Essays in honor of Professor C.R. Rao (Blackwell, Cambridge, MA) 291–310.

Cragg, J.G. and S.G. Donald, 1996, On the asymptotic properties of LDU based tests for the rank of a matrix, Journal of the American Statistical Association, forthcoming.

Gallant, A.R., 1987, Nonlinear statistical models (Wiley, New York, NY).

Gill, L. and A. Lewbel, 1992, Testing the rank and definiteness of estimated matrices with applications to factor, state space and ARMA models, Journal of the American Statistical Association 87, 766–776.

Gourieroux, C. and A. Monfort, 1989, A general framework for testing a null hypothesis in a mixed form, Econometric Theory 5, 63–82.

Hannan, E.J., 1980, The estimation of the order of an ARMA process, Annals of Statistics 8, 1071–1081.

Hannan, E.J., 1981, Estimating the dimension of a linear system, Journal of Multivariate Analysis 11, 459–473.

Hannan, E.J. and B.G. Quinn, 1979, The determination of the order of an autoregression, Journal of the Royal Statistical Society B 41, 190–195.

Lewbel, A., 1991, The rank of demand systems, Econometrica 59, 711–730.

Nishii, R., 1988, Maximum likelihood principle and model selection when the true model is unspecified, Journal of Multivariate Analysis 27, 392–403.

Phillips, P.C.B., 1989, Partially identified econometric models, Econometric Theory 5, 181–240.

Pötscher, B.M., 1983, Order estimation in ARMA models by Lagrange multiplier tests, Annals of Statistics 11, 872–885.

Pötscher, B.M., 1989, Model selection under nonstationarity: Autoregressive models and stochastic linear regression models, Annals of Statistics 17, 1257–1274.

Pötscher, B.M., 1991, Effects of model selection on inference, Econometric Theory 7, 163–185.

Schott, James R., 1984, Optimal bounds for the distribution of some test criteria for dimensionality, Biometrika 71, 561–567.

Wei, C.Z., 1985, Asymptotic properties of least squares estimates in stochastic regression models, Annals of Statistics 13, 1498–1508.