Abstract

This paper describes a method for estimating panels by imposing a factor structure on the residuals. The method allows SUR estimation of panel models by providing a full-rank estimator of the system covariance matrix when the usual estimate is rank-deficient. We charactersie completely the circumstances when this is possible. When the usual estimator is of full rank, our procedure provides a more parsimonious representation of the covariance matrix, which can lead to efficiency gains in finite samples. Monte Carlo analysis of convergence regressions and PPP regressions in the Heston-Summers data-set indicates that the proposed estimator has better performance in terms of RMSE and bias than standard panel or SUR estimators (where available), as well as offering unbiased inference.

Factor Residuals in SUR Regressions: Estimating Panels Allowing for Cross Sectional Correlation

Donald Robertson and James Symons

October 2000

Factor Residuals in SUR Regressions: Estimating Panels Allowing for Cross Sectional Correlation

Donald Robertson and James Symons

Acknowledgements

Donald Robertson is a lecturer in Economics at Cambridge University and Fellow of Pembroke College. James Symons is a Reader in Economics at University College London and an associate of the Centre for Economic Performance, London School of Economics.

FACTOR RESIDUALS IN SUR REGRESSIONS: ESTIMATING PANELS ALLOWING FOR CROSS SECTIONAL CORRELATION

DONALD ROBERTSON AND JAMES SYMONS

1. Introduction

The increasing availability of data structures with both time-series and cross-sectional dimensions and the possibility of overcoming the difculties associated with both pure time series and pure cross sectional work have given renewed interest to the econometric issues that arise with such structures. In this paper we propose an alternative method for the estimation of regression models on such data-sets.

The standard model characterises the relationship between the dependent variable and explanatory variables as a linear regression with random errors

$$(1) \qquad y_{it} = \circledR_i + \bar{}^0 x_{it} + u_{it} \qquad i = 1, ...n; \quad t = 1, ...T$$

Depending on the data structure available, further assumptions are generally made to identify the parameters in (??). If T is suficiently large and the errors form a vector white-noise process (with $E(uu^0) = \S$ where $u^0 = (u_{1t}, u_{2t}, ...u_{nt})$), the system can be treated as seemingly unrelated regression equations (Zellner, 1962) and $\circledR_i, \bar{}$, and the variance-covariance matrix of the errors $\S$ can be straightforwardly estimated. If, further, the errors are assumed to be jointly normally distributed, estimates are available by maximum likelihood.

If T is not suficiently large, and in particular if $n > T$, then such an approach is not feasible, because the usual estimator of the error covariance matrix is rank-de...cient. There are a variety of techniques discussed in the literature for dealing with the $n > T$ case. One method is to assume that $\S$ is diagonal. If it is further assumed that $\S$ is a scalar matrix, one has the usual panel estimator. This restricted structure is the one-way error components model. The parameters $\circledR_i$ can be treated as ...xed or random, and the model estimated by OLS, GLS, or, if one is willing to specify a joint distribution for the u_{it} and the $\circledR_i$, maximum likelihood. In these circumstances one can also allow for time ...xed e¤ects (the two-way error components model). If the explanatory variables include a lagged dependent variable, estimation is complicated in the ...xed e¤ects formulation: see Nickell (1981). In both SUR and panel estimation, it is possible to allow for some restricted serial correlation in the error process (see for example Liang and Zeger (1986) for the static case, Keane and Runkle (1992) for the dynamic case). If one relaxes the assumption that the error variances are equal, one can weight observations accordingly to obtain WLS estimators as in Fisher (1993). For $\S$ non-diagonal, one could always collapse some of the cross-sectional units into group averages to make $n < T$, as is often done for ML estimation of the factor model in ...nance. A further possibility is to replace the required inverse of $\S$ by its generalised inverse. Under certain conditions this is a BLUE estimator (see Theil, 1971), and brings the $n > T$ case within the range of SUR estimation, though we know of no applications of this

technique in a panel context. The advantage of the factor approach we propose vis-à-vis this possibility lies in its parsimony: we shall see that the factor approach is much more efficient in small samples, even in the full-rank case, at least for the cases we consider[1].

The assumptions on the correlation structure Σ are usually made for identifiability reasons, rather than descriptive accuracy. In particular the panel approach imposes zero correlation on the shocks to the cross-sectional units at each date so that estimation can proceed. The hope, presumably, is that by conditioning on sufficiently many explanatory variables x_{it}, what remains can be treated as a purely idiosyncratic shock, uncorrelated over time and the cross-sectional dimension, though with perhaps a t specific effect and a i specific effect. In practice, if the individual units are countries, firms or even individuals, this is unlikely to be the case. The econometrician rarely has sufficient explanatory variables to remove all correlated behaviour, and what is left over may not be well described by an $\alpha_i + f_t + \epsilon_{it}$ formulation[2]. Ideally one would like to allow for some sort of correlation pattern across the shocks. In a time series, it is natural to specify the correlations between disturbances to be functions of distance, measured by time, and a small set of parameters (such as in an ARMA process). In a cross-section, however, there is often no unambiguous concept of distance. One approach is to define a metric on the cross-section using some notion of physical distance and allow for a correlation structure with this measure playing the part of separation in a time series (see e.g. Conley (1999) who discusses the consistency of GMM estimation in these circumstances). Thus, for example, with a data-set consisting of many countries, one could allow for correlated shocks on adjacent countries, with zero correlation at greater distances. This approach is predicated on formulating an appropriate concept of nearness but clearly one country may be near a second for some purposes, but not for others. In the SUR framework the possibility of correlated shocks is allowed for through the covariance matrix Σ. However estimation of this matrix can often be expensive in terms of degrees of freedom, and a poor estimate of this matrix may contaminate the estimation of the parameters of interest in ??.

In this paper we propose a method that retains the flexibility of the SUR approach in allowing for correlated shocks, yet is more parsimonious than SUR, so can be expected to be more efficient in general. It can be implemented when $n > T$, and provides a more general specification than the panel approach. We impose a factor structure on the covariance matrix and estimate the factors by maximum likelihood techniques. This procedure can give an improvement in both bias and RMSE of estimators. The estimation procedure is essentially feasible GLS and so has the usual efficiency properties for large T.[3] In Section 2 we set out the details of the method. Section 3 applies the method to Barro-type convergence regressions and PPP regressions in the Heston-Summers data-set. Sections 4 and 5 assesses the performance of the factor approach by Monte Carlo and suggest some approximate methods. Section 6 concludes.

---

[1] It is to be emphasised that this is a small sample property. It is well known that, with a minor caveat, imposing restrictions on Σ brings no efficiency gains in large samples. See Greene, 1990.

[2] Frees (1995) discusses testing cross sectional correlation in panel data

[3] For fixed T, the estimators as n grows may or may not be consistent.

2. A Factor Residual Approach

We specify the model as follows

$$(2) \qquad y_t = X_t{}^- + u_t \qquad t = 1, \dots, T$$

where y_t and u_t are $n \times 1$ vectors, X_t is an $n \times v$ matrix of explanatory variables observed at t, $^-$ is a vector of unknown parameters to be estimated, u_t is a vector white noise process with $E(u_t u_t^0) = \S$, and $E(X_{ijt} u_{kt}) = 0$ all i, j, k, t.

We make the further assumption that

$$(3) \qquad \S = \text{¤}\text{¤}^0 + {}^a$$

where ¤ is a $n \times m$ matrix of so-called factor loadings and ^{a} is a diagonal $n \times n$ matrix with diagonal elements $\tilde{A}_1, \tilde{A}_2, \dots \tilde{A}_n$, where $\tilde{A}_i > 0$ reflects idiosyncratic effects. This allows for some contemporaneous correlation between shocks, expressed as a function of fewer parameters than the unconstrained $\S$ if $m < n$. Note that the factor model generalises fixed effects models directly, as the fixed effects can be entered as elements of X. A time random-effects model is equivalent to a one-factor model with ^{a} diagonal and ¤ proportional to $(1, 1, 1, \dots 1)^0$: ^{a} scalar and $m = 0$ gives the usual panel formulation.

The factor model has an attractive interpretation because it amounts to specifying that the residuals take the form $u_t = \text{¤}_1 \acute{A}_t^1 + \text{¤}_2 \acute{A}_t^2 + \dots + \text{¤}_m \acute{A}_t^m + {}^{"}_t$ where $E(\acute{A}_t^i \acute{A}_s^j) = \pm_{ij} \pm_{ts}$ and $E({}^{"}_t {}^{"}_s{}^0) = \pm_{ts}{}^a$ and ¤_k is a column vector of weights. The $\acute{A}$s can be interpreted as m common shocks and the elements of each ¤_k give the loading or impact of each of these shocks on the cross sectional units. For example, if the units are economies, the first factor $\acute{A}_t^1$ might represent a world demand shock with ¤_1 as the relative openness of the economies, the second factor $\acute{A}_t^2$ an agricultural shock etc. For $m = n$ such a decomposition can trivially be obtained; the usefulness of this decomposition arises when m may be taken to be much smaller than n. These common shocks provide cross-sectional correlation in the error structure, with the $"_t$ adding an idiosyncratic term.

Given an estimate S of $\S$, the likelihood function up to an additive constant is

$$(4) \qquad L(\S) = L(\text{¤}, {}^a, S) = {}_i \frac{T}{2} {}^i \log \det \S + \text{tr} {}^i \S^{i\,1} S^{\text{¢¢}}$$

(see Lawley and Maxwell, 1963).

The first-order conditions are (Lawley and Maxwell or Jöreskog, 1967)

$$(5) \qquad \frac{@L}{@\text{¤}} = {}_i 2\S^{i\,1}(\S {}_i S)\S^{i\,1}\text{¤} = 0$$

$$\frac{@L}{@{}^a} = {}_i \text{diag}(\S^{i\,1}(\S {}_i S)\S^{i\,1}) = 0$$

Classical factor analysis assumes that S is of full rank, and thereby obtains estimates of ¤ and ^{a} by maximum likelihood: The results cannot immediately be generalised to the rank-deficient case. There are methods in the literature for dealing with this situation; for instance, Maxwell (1981) suggests estimation via norm-minimisation.[4] It is also possible to obtain factors as principal components,

---

[4] The norm in question being $\Vert B \Vert = (\text{tr} B^0 B)^{\frac{1}{2}}$, so that one would minimise
$$\Vert S {}_i \S \Vert^2 = \sum_{i,j} (S_{ij} {}_i \S_{ij})^2.$$

even in the rank-de...cient case. In the ...nance literature, Connor and Korajcyk (1986, 1988) estimate factors using the Chamberlain and Rothschild (1983) method of asymptotic principal components in the rank-de...cient case. These factors are identi...ed for asymptotically large n given certain assumptions. The advantage of maximum likelihood is that one has, in large samples at least, a natural basis for inference and model selection. The technical contribution of this paper is to show that, subject to certain restrictions on the number of ...tted factors, the likelihood function is bounded and attains its bound at an invertible §, irrespective of the rank of S. Hence maximum-likelihood estimators of ¤ and ^{a} can be obtained from direct maximisation of the likelihood function, even in the rank-de...cient case[5]. The likelihood is invariant to multiplication of ¤ by a unitary matrix so the matrices ¤ and ^{a} are identi...ed only up to such a matrix. This can be resolved by ...xing ¤ to have zeros above the diagonal, thus removing the di₵culty in estimation; though if one wished to o¤er interpretation of the factors it would, of course, be problematical.

We summarise our results concerning the likelihood function in the following. If S has full rank there is no problem in maximising the likelihood (see e.g. Jöreskog, (1967)). We show in the appendix that, as long as one does not attempt to ...t too many factors, the likelihood function for the general factor model is bounded and attains its bound, even if the initial estimate of the variance-covariance matrix is rank-de...cient. The estimator of the covariance matrix thus obtained has full rank. This result clearly has applications beyond those we explore below. To obtain estimates of the factors we can therefore operate directly on the likelihood function (??). As is well known in the case S has full rank, it can be di₵cult to ...nd the maximum of the likelihood function numerically. The situation is the same when S is rank-de...cient. The geometry of the likelihood function is explored further in the appendix.

In the situation where the likelihood is bounded, the general shape is akin to the positive orthant of a multi-dimensional volcano. The maximum is obtained on the rim or on the boundary of the admissible region (these are known as Heywood solutions). Clearly search algorithms will usually have great di₵culty with this surface. Motivated by this, we propose a mixture of search algorithms to ensure convergence. The numerical technique is discussed in detail in the appendix.

Having obtained estimates of the factors, estimates of $^-$ can then be obtained by feasible GLS, by minimising $\sum_{t=1}^{T}(y_t - X_t{}^-)'\S^{-1}(y_t - X_t{}^-)$, where $\S = \b\b' + \b$ to give $^\triangle = \left(\sum_{t=1}^{T} X_t'\S^{-1}X_t\right)^{-1}\sum_{t=1}^{T} X_t'\S^{-1}y_t$. One could in principle iterate this approach (as is usually done in a SUR framework) to obtain full maximum likelihood estimates. Standard errors can be obtained from the usual GLS formula.

Asymptotic consistency results are complicated by the incidental parameter problem (Neyman and Scott, 1946) arising particularly in this context when time or individual ...xed e¤ects are included in the model. For T-asymptotics with ...xed n, if we begin with an initial T-consistent estimate of §; the estimates of ¤ and ^{a} will be T-consistent if the factor model is true[6]: Hence our GLS estimator of $^-$ inherits the T-asymptotic properties of GLS. Thus, whereas estimates of time ...xed

---

[5]Given the factor loading matrix ¤ and ^{a} ; the factors can then be recovered by regression, if desired.

[6]By application of Slutsky's Theorem, given that, as is easily shown, the mapping S ! $\b_{Factors}$ is continuous and takes the value § at S = §:

e¤ects will not be T-consistent, other parameter estimators typically will be consistent. The gain here is a more parsimonious representation of the covariance matrix, leading, as we shall see, to potential eÇciency gains even when T is moderate.

The case of n-asymptotics is a little more murky. Clearly one could not expect to obtain n-consistent estimates of individual ...xed e¤ects, for given T: Beyond this, the Nickell result shows that, in the presence of individual ...xed-e¤ects, estimates of the structural parameters can be n-inconsistent. If $\dim(\bar{})$ is ...xed as n grows, however, one might hope to say something, but we have no results as yet. One could as well seek asymptotics for large T but larger n: It is true that, if the data are ...xed in repeated samples and S is obtained by OLS on each of the n equations, then the GLS estimate of $\hat{}_i \bar{}$ will be an odd function of the residuals, and hence median-unbiased. However, irrespective of these considerations, we shall see below that, for ...xed sample size, the proposed estimator has better performance measured by standard loss functions than both conventional SUR (where feasible) and panel methods, and in this sense can provide superior estimates. The empirical contribution of this paper is thus to suggest the use of this factor-residual model in a SUR framework, permitting estimation of panel models allowing for cross-sectional dependence. We illustrate with an empirical example.

3. An Application: Convergence Regressions

We illustrate the above techniques by applying the method to study convergence in the Heston-Summers data-set. We seek to measure convergence in GDP per capita relative to the US. This problem has been studied by many authors (Barro, 1991, Mankiw Romer Weil, 1992, Evans and Karras, 1996, Islam, 1995, Lee, Pesaran and Smith, 1995, Caselli et al.,1996, among others) mainly using the Heston-Summers database. Barro-type regressions correlate average growth rates and initial values; others attempt to control for extra right-hand-side variables such as human capital, savings rates etc. and study conditional convergence. A typical Barro result is convergence at about 2% per annum; more recently a number of panel studies have suggested a somewhat higher rate of convergence (see Evans and Karras, 1996, Islam. 1995, Lee, Pesaran and Smith, 1995, Caselli et al., 1996). We exploit the full time-series and cross-sectional dimensions of the Heston-Summers data-set, and study unconditional convergence, though our methodology could straightforwardly be applied to study conditional convergence. Imposing a factor structure means we can allow for the shocks that disturb the system to be correlated across countries – for instance oil price shocks or ...nancial crises. We specify the model as

$$(6) \qquad\qquad y_{it} = \frac{1}{2} y_{it_i 1} + {}''_{it}$$

where $y_{it} = \ln$(GDP per capita as percentage of US). We shall impose a factor structure on the $''_{it}$, allowing for correlated shocks. Note that we do not allow for individual ...xed-e¤ects. As discussed above, these would cause problems in a dynamic regression if T were small (see Nickell, 1981), though there is a developing literature addressing this problem by applying GMM techniques to the di¤erenced equation (see Islam, and Caselli et al.). In an unconditional convergence regression,

the natural formulation does not involve individual intercepts.[7] In estimation we include a simple intercept.

3.1. **Results and selection of number of factors.** We estimate this model for the subset of the OECD countries (n = 22; T = 41) and for the whole data-set (n = 103; T = 31). We also estimate (??) using a panel estimator (i.e. applying OLS to the stacked data, appropriate when § is a scalar matrix) for both the OECD and world subsets, and using SUR for the OECD countries. One key issue is the selection of the number of factors to be ...tted. Likelihood ratio tests are available, as well as information criteria such as Akaike, Schwarz-Bayes or Hannan-Quinn[8]. We discuss the performance of these information criteria later.

For the OECD subset, the estimate S of the matrix § indicates considerable cor-related structure in the errors. The average modulus of the o¤-diagonal correlations in this matrix is 0.375. With correlations of this magnitude, panel/OLS estimation will produce biased standard errors, and be less e₵cient than unrestricted SUR (see Di Liberto & Symons, 1998). The estimation results are set out in Table 1.

---

[7]Though in principle there would be no di₵culty in combining any ...xed e¤ects estimator with our factor residual model by premultiplying (??) by $\S_i^{-1=2}$; as long as an initial consistent estimate of § is available.

[8]These criterion function are of the form LogLikelihood ¡ f(T):(#params) where f(T) = 1 for Akaike, $f(T) = \frac{1}{2}\ln(T)$ for Schwarz, and we use f(T) = ln(ln(T)) for Hannan-Quinn.

Fitted Σ	intercept	$\hat{b}$	log likelihood	AIC	SBC	HQC
scalar(OLS)	0:12526 (:01314)	0:97234 (:00323)	0	¡1:00	¡1:84	¡1:31
0-factor(WLS)	0:12847 (:01321)	0:97167 (:00322)	60:01	38:01	19:43	31:29
1-factor	0:11582 (:01191)	0:97422 (:00277)	266:36	222:36	185:20	208:92
2-factor	0:11702 (:01125)	0:97463 (:00260)	305:30	240:30	185:41	220:45
3-factor	0:11445 (:01090)	0:97527 (:00252)	322:34	237:34	165:56	211:38
4-factor	0:12908 (:01191)	0:97175 (:00278)	339:37	235:37	147:55	203:62
5-factor	0:13297 (:01198)	0:97085 (:00279)	354:44	232:44	129:42	195:19
6-factor	0:13316 (:01181)	0:97071 (:00279)	370:93	231:93	114:56	189:50
7-factor	0:12902 (:01178)	0:97169 (:00280)	382:77	227:77	96:88	180:44
8-factor	0:12896 (:01109)	0:97176 (:00264)	394:44	224:44	80:89	172:54
19-factor	0:13239 (:00924)	0:97076 (:00221)	445:77	176:77	¡50:39	94:64
20-factor	0:13239 (:00924)	0:97076 (:00221)	445:77	173:77	¡55:92	90:72
n-factor(SUR)	0:13239 (:00924)	0:97076 (:00221)	445:77	192:77	¡20:88	115:52

Table 1: Factor Residual, SUR and OLS/Panel Convergence Regressions for the OECD subset of Heston-Summers

Notes: (i) Σ scalar is Panel/OLS; 0-factor is weighted OLS; n-factor is SUR. (ii) Log-likelihood is normalised to be 0 for Σ=scalar (OLS). (iii) Since ¤ is restricted to be upper diagonal, adding a further factor adds fewer than 22 parameters. (iv) Nominal standard errors in parentheses, calculated as $\hat{b}^2(X'\hat{\Sigma}^{i1}X)^{i1}$ for $\hat{b}$ in each row. These standard errors are fairly dubious a priori for a process so close to the unit circle.

The results are listed by decreasing degree of restriction on the covariance matrix. If Σ is restricted to be a scalar matrix, one has OLS; if Σ is diagonal (0-factors) one has weighted least squares; each factor model is nested in a higher-order model; SUR (n-factor) is the least restricted. OLS and WLS are strongly rejected against models with non-zero covariance structure by likelihood ratio tests. The restriction from seven to six factors would not be rejected at the 5% level by likelihood ratio, but further restrictions would be rejected. Using the information criteria, Akaike, Schwarz and Hannan-Quinn all select the two-factor model. Likelihood-ratio tests of constant size will not lead to consistent order selection, unlike SBC and HQC.

For the entire Heston-Summers data-set ($n = 103$; $T = 31$) we obtain the results in Table 2.

Fitted $\S$	intercept	$\hat{b}$	log likelihood	AIC	SBC	HQC
scalar(OLS)	¡ 0:01774 (:00398)	1:00617 (:00134)	0	¡ 1:00	¡ 1:70	¡ 1:22
0-factor(WLS)	¡ 0:01581 (:00321)	1:00564 (:00091)	626:91	523:91	451:75	500:83
1-factor	¡ 0:01953 (:00326)	1:00608 (:00076)	1001:15	795:15	650:83	748:98
2-factor	¡ 0:01456 (:00310)	1:00462 (:00073)	1145:27	837:27	621:49	768:24
3-factor	¡ 0:01684 (:00300)	1:00589 (:00073)	1274:88	865:88	579:34	774:21
4-factor	¡ 0:01720 (:00286)	1:00597 (:00071)	1376:93	867:93	511:32	753:85
5-factor	¡ 0:01681 (:00280)	1:00588 (:00069)	1474:56	866:56	440:60	730:29
6-factor	¡ 0:01692 (:00260)	1:00595 (:00066)	1573:85	867:85	373:23	709:62

Table 2: Factor Residual and OLS/Panel Convergence Regressions for the Heston-Summers Data-set.

Notes: As for Table 1, except that SUR is omitted since $n > T$:

Akaike is fairly ‡at between the three, four and ...ve factor models for the world regressions, Hannan-Quinn selects three factors and Schwarz points to a one-factor model, re‡ecting the heavier weight placed on extra parameters by this criterion Again a scalar or diagonal covariance matrix is strongly rejected.

The parameter estimates are in line with those obtained by other authors, and the literature on convergence clubs. We ...nd a point estimate of the rate of convergence of approximately 2.5% per annum for the OECD countries, and no convergence for the full Heston Summers data-set.

3.2. Con...dence Intervals for the Convergence Parameters. We construct con...dence intervals for the three estimates of the convergence parameter for the OECD subset by Monte Carlo. We de...ne a 95% con...dence interval as the points lying between the 2.5 percentile and 97.5 percentile points of the sampling distribution obtained from 1000 repetitions of each estimation procedure applied to arti...cial data generated by the relevant estimated parameters in Table 1 and the original estimated variance-covariance matrix S to generate the errors[9]. For the factor model, we select the estimated parameter values from the (preferred) two-factor model for the Monte Carlo, and always estimate two-factor models on the generated data. The results are set out in Table 3.

---

[9]That is, we do not generate the data using an underlying factor residual model. This will ensure that we are not biasing the Monte Carlo in favour of such models.

	true parameter	2:5 percentile	mean	97:5 percentile
OLS				
®	0:12526	0:09350	0:13105	0:17306
½	0:97234	0:96097	0:97093	0:98000
SUR				
®	0:13239	0:10678	0:13614	0:15905
½	0:97076	0:96198	0:96986	0:97686
Factors				
®	0:11702	0:09619	0:11755	0:14025
½	0:97463	0:96947	0:97449	0:97927

Table 3: Empirical Con...dence Intervals from 1000 repetitions.[10]

It will be observed that the 95% con...dence intervals for the factor model are about half the width of the OLS con...dence intervals and two-thirds of the SUR value. It is useful to compare these intervals with the apparent con...dence intervals (estimate plus or minus two estimated standard errors) from Table 1. Apparent con...dence intervals for OLS are expected to be wrong, even asymptotically, when § is non-diagonal, while SUR will be correct in large samples but wrong in small samples when an estimate of § is used in place of its true value to compute standard errors. One ...nds that the apparent con...dence interval width by OLS is about 70% of its true value, for SUR about 60%. Factors, in contrast, estimates the true con...dence interval almost exactly. For this example at least, the factor method o¤ers much superior inference. We return to this issue below.

3.3. **A Further Application: PPP.** We illustrate the technique with a further application, looking at PPP in the Heston Summers data-set. The existing literature ...nds mixed evidence for PPP using time-series, cointegration and panel methods. Using real exchange rate data from Heston Summers, we have $n = 103; T = 31$ (corresponding to the years 1960-1990). We take as our dependent variable ln(exchange rate) and normalise by subtracting from each time series its average value over 1964-69. The intention is thus to model deviations of exchange rates from some "normal" level, here taken as the average over the second half of the sixties. This is a compromise between allowing for individual country ...xed e¤ects (equivalent to subtracting time averages of the data for each country), entailing problems of bias in a dynamic panel, and normalising the exchange rates on a particular year which has the risk of choosing an anomalous reference point. We again estimate an AR(1) regression, ...tting an intercept (which should be zero if PPP holds). The average modulus of the correlations in these data is 0.275, indicating considerable cross-sectional dependence. Thus, as for the convergence regressions, conventional panel methods are likely to lead to distorted inference and loss of e¢ciency.

The results are in Table 4.

---

[10]In fact, SUR and OLS are based on 5000 repititions, factors 1000.

Fitted §	intercept	$\hat{b}$	log likelihood	AIC	SBC	HQC
scalar(OLS)	0:00205 (:00249)	0:90534 (:00911)	0	¡1:00	¡1:70	¡1:22
0-factor(WLS)	¡0:00015 (:00161)	0:95222 (:00776)	805:43	702:43	630:27	679:34
1-factor	¡0:01809 (:00197)	0:95810 (:00722)	1596:06	1390:06	1245:73	1343:89
2-factor	0:00767 (:00227)	0:86290 (:00930)	1835:08	1527:08	1311:29	1458:05
3-factor	0:00344 (:00237)	0:85584 (:00945)	2000:75	1591:75	1305:21	1500:08
4-factor	0:00328 (:00228)	0:85887 (:00929)	2054:18	1545:18	1188:58	1431:10
5-factor	0:00262 (:00224)	0:86466 (:00901)	2173:98	1565:98	1140:02	1429:71
6-factor	0:00191 (:00217)	0:84166 (:00975)	2272:95	1566:94	1072:32	1408:71

Table 4: Factor Residual and OLS/Panel PPP Regressions for the Heston Summers Dataset.

Note: Nominal standard errors in parentheses.

Both Akaike and Hannan-Quinn select a three factor model, whereas Schwarz marginally prefers two factors. Whilst panel/OLS gives an AR parameter of 0.91, allowing for the correlated error structure reveals $\hat{b}$ = 0:86: The di¤erence between the half-lives of deviations from PPP implied by these estimates is of economic signi...cance.

4. Monte Carlo analysis of the method.

4.1. Deciding on the number of factors. We take the ...tted three-factor co-variance matrix from the Heston Summers PPP regressions and use this to generate 103 £ 31 matrices of correlated normal variates. These are used to generate arti-...cial AR(1) time series of length 31, using the parameters in Table 1. An AR(1) model is then ...tted to each of the 103 time series and the residuals used to form a rank-de...cient 103 £ 103 variance-covariance matrix . AIC, SBC & HQC are then calculated to see whether they are able to identify the correct number of factors, here three. This procedure is rather time-intensive: the ...ve- and especially the six-factor models are automatically heavily overparameterised and this seems to cause convergence to be very slow. We therefore limit the number of repetitions to 100. Table 5 gives the frequency of selection of di¤erent numbers of factors by the three information criteria.

Frequency of ...tted factors:	1	2	3	4	5	6
Chosen by AIC	0	1	95	4	0	0
Chosen by SBC	0	19	81	0	0	0
Chosen by HQC	0	1	98	1	0	0

Table 5. Performance of various information criteria in identifying correct factor model.
Results from 100 repetitions of underlying 3-factor model.

The HQC does astonishingly well, getting the number of factors correct 98/100. AIC is nearly as good. SBC tends often to choose smaller models.

4.2. Factors vs SUR vs OLS - who does best? We generate Monte Carlo replications of the data using as initial values the 1950 values of log GDP for the OECD subset of Heston-Summers. We generate series for y_{it} via

$$y_{it} = \text{®} + \text{½}y_{it_i \, 1} + \text{"}_{it}$$

where ® and ½ are the estimated parameters from the SUR of this equation and the $"_{it}$ are correlated normally distributed random numbers with variance-covariance matrix as estimated from the residuals of the SUR. We use an empirical variance-covariance matrix for these Monte Carlo, rather than generating data with a given factor structure, to ensure that the experiments are not biased in favour of factors. We generate time series of di¤erent lengths T holding n = 22: On these generated data we estimate AR(1) convergence regressions by OLS/Panel, SUR (whenever n < T) and the factor approach, assuming a two-factor model, as suggested by Table 1, a one-factor model and a zero-factor model (WLS). Given that time …xed- or random-e¤ects will generate a model close to the one-factor speci…cation, it is of interest to check the performance of the factor speci…cation against a model where cross-sectional dependence is modelled as time …xed-e¤ects and this is also calcu-lated. The root-mean-squared-error and bias from 1000 repetitions are graphed in Figures 1 and 2. Note …rst that there is nothing between the one- and two-factor models. The factor-residual approach beats OLS/Panel and SUR in RMSE and bias for n = 22 and T running from 5 to about 60: One- and two-factor models outperform time …xed-e¤ects in RMSE, with similar bias performance. Eventually, as T rises, SUR starts to dominate, though of course this is holding the number of factors …xed. If the number of factors were allowed to rise as T grew, this may no longer hold. Since in these Monte Carlo experiments the variance-covariance matrix generating the data is, in fact, a genuine empirical matrix and is thus pre-sumably a 22-factor model[11], a consistent selection criterion such as SB or HQ would ultimately select m = 22 and …t the model as well as SUR. Thus the supe-rior performance of factors for T up to 60 illustrates the e¢ciency gain of parsimony, even, as here, at the expense of misspeci…cation.

Whilst one- and two-factor models perform about equally well in RMSE terms, rather unexpectedly, we …nd in Figure 1 that WLS does worse even than OLS for some sample sizes. In general, transforming the data so that the residuals are homoscedastic does not always lead to an increase in e¢ciency when § is non-diagonal.[12] This is a useful experiment because it shows that the e¢ciency gains from the factor approach are due to exploiting the non-diagonal nature of § rather than the inequality of the terms along the main diagonal. The fact that the one-

---

[11]Even though, in Table 1, treating this matrix as the sample estimate of an underlying m factor model, we selected m = 2.

[12]Assume ?? contains just one independent variable x_t with covariance matrix $V = E(x_t x_t^0)$. Then OLS is the best among all estimators that weight the data by a diagonal matrix (thus including WLS) if

$$\text{diag}(V\,§) = \text{¸}\,\text{diag}(V)$$

for some ¸ > 0: Examination of this condition shows that for OLS to be both best and di¤erent to WLS requires both V and § to be non-diagonal. This apart, the condition can certainly hold, which implies that there are regions of V; § space where OLS is better than WLS. These results are for the case of known §:

and two- factor models perform equally well means presumably that the efficiency gains from parsimony are off-set by losses incurred by the imposition of rejectable restrictions on the covariance matrix.

Figure 1

Figure 2

4.3. **Actual versus calculated standard errors for the various estimation procedures.** For the Monte Carlo described above we calculate the sampling standard error of the estimated parameter b from 1000 repetitions for the OLS, SUR and two-factor model. We also obtain the average calculated standard error from the $\hat{\sigma}^2(X'X)^{-1}$ formula for each estimator. The ratio of the estimated to the true standard error as the sample length (T) varies is graphed in Figure 3. We ...nd that OLS and SUR provide seriously biased inference over a wide range of sample sizes, in both cases calculated standard errors substantially understating the true sampling variation of the estimator. For instance, with a sample size of 15 (so that the panel has n=22, T=15), OLS calculated standard errors are approximately half the sampling standard error from the 1000 repetitions. The factor residual approach provides a much superior guide to the true sampling variation, though still tends to underestimate somewhat. The performance of SUR when the time and cross sectional dimensions of the data are of comparable magnitude is very poor. Of course these results depend on the correlation structure in the underlying Monte Carlo, which here is drawn from the estimates of the AR(1) model for GDP per capita on the OECD subset of the Heston and Summers database. But we have no reason to think that this is a particularly exceptional database in terms of error correlation structure, so these results must cast real doubt on inference in existing panel studies[13].

Figure 3

4.4. **Normality of the distribution of the various estimators.** For each of the Monte Carlo above we investigated the distribution of the 1000 estimates of the coe¢cient β. In general estimates derived from the factor residual technique showed less departure from normality than others. For example, for the sample $n = 22; T = 25$ we obtain the following histograms and diagnostics

---

[13]Driscoll and Kraay (1998) have recently given a method of obtaining asymptotically correct standard errors for OLS.

Figure 4 SUR Estimation

Figure 5 OLS Estimation

Figure 6 Factors Estimation

Notes to Figures 4, 5 and 6: Histograms of 1000 estimates of AR1 coe¢cient by the various estimation techniques. Sample design $n = 22; T = 25$ and underlying parameters of OLS estimate of AR1 convergence regression in Heston Summers OECD database.

SUR and OLS generate a detectably non-normal distribution of estimates, whilst for the factor residual approach it is not possible to reject the null of normality of the estimator at conventional levels.

The Monte Carlo described above is for a particular set of parameter values and a particular design, and we must be wary of generalising too freely from these results. However the conclusion of these investigations is clear. The factor residual approach is superior in RMSE and bias to both OLS and SUR (where available) over a large range of sample sizes, the estimates derived from the factor residual approach follow more closely a normal distribution when the underlying errors are normal, and their estimated standard errors provide a more reliable guide for inference.

5. Approximate Factor Techniques

Full maximum likelihood estimation of the factor model can be rather cumbersome and time-consuming due to the nature of the likelihood surface. In this section we brie‡y discuss some techniques that provide an approximation to maximum likelihood. We also set out the advantages and drawbacks of these approximations as they appear to us.

5.1. Principal Components.
One can easily approximate the rank de…cient S matrix by its m principal components and de…ning the corresponding ^{a} matrix to …t the diagonal exactly. This is very simple to implement, very quick and will usually produce an invertible estimator $\hat{§}$. We have experimented with such a principal component scheme, …tting the same number of principal components as the factor model selects, and …nd that it produces very similar results in terms of rmse etc. Thus principal components o¤ers a straightforward method to capture correlated error structures in panel data, the di¢culty being of course that, in contrast to ML, one has no obvious diagnostic for the choice of the number of principal components.

5.2. Norm Minimisation.
Principal Components can be iterated to minimise k§ ¡ Sk where § = ¤¤⁰ + ^{a}. For given ^{a} we take the eigenvectors corresponding to the largest eigenvalues of S ¡ ^{a} as our elements of ¤: We then de…ne a new ^{a} to match the diagonal elements of S exactly given this ¤ i.e. set ^{a} = max diag(S ¡ ¤¤⁰; 0): This procedure produces a non-increasing sequence of values of the norm and will converge to some $\hat{§}$. It is again simple to program and tends to converge quite quickly. The problem with norm-minimisation is that, unlike ML, it is not guaranteed to …nd an invertible $\hat{§}$: We have arti…cial examples of covariance matrices S where, if one factor is …tted, Heywood solutions of multiplicity[14] greater than unity are found by norm-minimisation. In this case the estimated $\hat{§}$ cannot be invertible.[15] This is inconvenient but not necessarily fatal as a generalised inverse could always be used for the GLS.

---

[14] The multiplicity of a Heywood solution is the number of zero components of $^{\hat{a}}$:

[15] An intuition for the singularity-hating behavior of ML can be obtained from consideration of (??): as an eigenvalue of § approaches zero, the second term on the right contributes a negative in…nity of higher order than the positive in…nity contributed by the …rst term. This tends to cause ML to choose non-singular §s. The norm, in contrast, has no such reason to shun singular matrices.

6. Conclusion

In empirical work using data structures with time and cross-sectional dimensions, it is often di¢cult to estimate correlations in the errors accurately. The solution of setting these correlations to zero may not, however, be the wisest course. This paper suggests a technique whereby estimation can exploit correlations in the error structure across cross-sectional units. By imposing a factor structure, which can have an appealing economic interpretation, we provide a full-rank estimator of the covariance matrix. Monte Carlo shows that the parameter estimator thus obtained has substantial gains in e¢ciency over the standard alternative procedures for the data-sets considered, as well as o¤ering unbiased inference. Our suggestion is that the proposed factor-residual method of estimation provides an important generalisation of standard panel techniques, and that even in situations where the conventional estimator of the covariance matrix of the residuals has full rank, it can provide a more parsimonious and e¢cient estimation technique.

Appendix A. Appendix

Properties of the Likelihood Function of the Factor Model

Our aim in this Appendix is to give a description of the geometry of the likelihood surface of the factor model, paying particular attention to the case when S is rank deficient. When S is of full rank it is possible to show by completely elementary methods that the likelihood function is bounded (Theorem 2 (i)). We do not have a similar simple demonstration that the maximum is attained (perhaps on the boundary). This is true, however, and can be shown by use of the concentration of L introduced by Lawley. When S is not of full rank, it remains true that L is bounded and attains its maximum provided that not too many factors are fitted, but here the tricky part is to show L is bounded. We achieve this by studying the behaviour of Lawley's concentrated L on the rim, the set of Σs maximising L along rays from the origin.

Definitions and Simple Results

Let matrix $S = X'X$ where X is $T \times n$. Call S admissible if no column of X consists entirely of 0s. Define the function

$$(7) \qquad L = L(\Lambda; \Psi; S) = -\left[\log \det \Sigma + \operatorname{tr} \Sigma^{-1} S \right]$$

where Λ is any $n \times m$ matrix, Ψ is $n \times n$ diagonal with entries $\psi_i > 0$ and $\Sigma(\Lambda; \Psi) = \Lambda\Lambda' + \Psi$. Note Σ is the sum of a positive definite matrix and a non-negative definite matrix and is thus itself positive definite and hence invertible. As defined, L is the likelihood function (divided by $T/2$) of the m-factor model generated by n normal variates.

Some more terminology. Denote by $\mathcal{L}$ the set of $n \times m$ matrices, by $\mathcal{P}$ the set of $n \times n$ positive diagonal matrices, and by $\mathcal{Q}$ the range of the function $\Sigma(:,:)$ so that $\Sigma : \mathcal{L} \times \mathcal{P} \to \mathcal{Q}$. This function is not one-to-one which means that $\Lambda; \Psi$ are not identified by knowledge of $\Sigma(\Lambda; \Psi)$. In particular $\Sigma(\Lambda U; \Psi) = \Sigma(\Lambda; \Psi)$ for any unitary U. If required, this problem can be solved by restrictions on $\mathcal{L}$ e.g. that it consist of Λs which are non-negative and decreasing along the diagonal, and zero above it. So restricted, the function $\Sigma(:,:)$ is bi-continuous. This is not an important issue for our purposes as we seek the Σs rather than the Λs and Ψs which generate them.

If D is a fixed $n \times n$ diagonal matrix then, up to a constant depending on D,

$$L(\Lambda; \Psi; S) = L(D'\Lambda; D'\Psi D; D'SD)$$

Thus, for the study of boundedness and the existence of maxima, no generality is lost by re-scaling the rows and columns of S. When S is admissible, the elements along the main diagonal are non-zero and S may be re-scaled to a correlation matrix. This is often convenient.

A result we shall not need, but gives some insight into the character of the likelihood surface is:

THEOREM 1. The function L has no local minima.

PROOF. If $¤_0; ^a_0$ were a local minimum then L restricted to $\pm ¤_0; \pm^a_0; \pm > 0$ has a local minimum for $\pm = 1$. Examination of $L(\pm) = L(\pm¤_0; \pm^a_0)$ shows this function has no local minimum.¤

It is well-known that L can have multiple maxima (e.g. Jöreskog, 1967). Thus likelihood surfaces can have ravines but, according to Theorem 1, no lakes.

We now set out the boundedness properties of the likelihood function.

PROPOSITION 1. If S is inadmissible, L is unbounded.

PROOF. If S is inadmissible then the jth (say) row and column consist of 0s. Choose $¤ = 0$, so that $L = ¡ (\sum_{i=1}^{n} \log \tilde{A}_i + \sum_{i=1}^{n} s_{ii}=\tilde{A}_i)$ and let $\tilde{A}_j ! 0$. There is a term in L; $¡ \log \tilde{A}_j$, which is unmatched by a term $¡ 1=\tilde{A}_j$ so that $L ! 1$. ¤

PROPOSITION 2. If rank(S) = n; then L is bounded.

PROOF . Choose unitary U to diagonalise $§(¤;^a)$ so that

$$L = ¡ \sum_{i=1} (\log ¾_i + x_i^0 S x_i=¾_i)$$

where $¾_1 ¸ ¾_2 ::: ¸ ¾_n > 0$ are the eigenvalues of § with corresponding eigenvectors x_i. Since $x^0 S x$ takes values in $[s_n; s_1]$ (where s_i denotes the eigenvalues of S and $s_1 ¸ s_2 ::: ¸ s_n > 0$) for vectors x with $k x k = 1$,

$$L \cdot ¡ \sum_{i=1} (\log ¾_i + s_n=¾_i) \cdot ¡ n (\log s_n + 1)$$

because $¡ (\log ¾_i + s_n=¾_i)$ is maximised at $¾_i = s_n:$¤

PROPOSITION 3. If rank(S) = r < n and m = r then L is unbounded.

PROOF. Consider, for $¹ > 0$,

$$§_¹ = ¹ I + x_1 x_1^0 + ::: + x_r x_r^0$$

where $r = \text{rank} S$ and $x_i; i = 1; :::; r$ are the normalised eigenvectors of S of non-zero eigenvalues s_i. It follows that $§_¹ 2 Q$. Now the eigenvectors of § are precisely those of S with eigenvalues $¹ + 1$ for $i = 1; :::r$, and $¹$ otherwise. It follows that we may simultaneously diagonalise $§_¹$ and S in (??) to obtain

$$L = ¡ \left(r \log (1 + ¹) + (n ¡ r) \log ¹ + \sum_{i=1} s_i=(1 + ¹) \right)$$

$$! 1 \text{ as } ¹ ! 0:¤$$

If we were to ...t m factors in Proposition 3 where $m \cdot r < n$ then the likelihood function takes the form

$$L = ¡ \left(m \log (1 + ¹) + (n ¡ m) \log ¹ + \sum_{i=1} \frac{s_i}{1 + ¹} + \sum_{i=m+1} \frac{s_i}{¹} \right)$$

and it is easy to see that this function is bounded as a function of $¹$ as long as m < r: This might lead one to conjecture that L is bounded provided one ...ts fewer than r factors. However this is incorrect: certain null space structures of the matrix S further reduce the number of possible factors. To see this suppose that for all

z 2 N(S) (the null space of S), $z_j = 0$ for some index j; $1 \cdot j \cdot n$, i.e. the null space has a row of zeros at the jth position.[16] This implies that e_j ?N(S) where e_j is the jth element of the canonical basis of R^n. It follows that e_j is a linear combination of the eigenvectors of S of non-zero eigenvalue whence

$$\mathbf{X}_{i=1} x_i x_i^0 + {}^1 I = \mathbf{X}^1_{i=1} y_i y_i^0 + {}^a$$

where the y_i are linear combinations of the x_i; and ^{a} is diagonal. This enables the likelihood for the r-factor model to be written as an r ¡ 1 factor model. Since for rank(S) = r, the r-factor model gives an unbounded likelihood, each row of zeros in the null space reduces by one the number of factors that can be ...tted. Thus the structure of the null space plays a key part in determining the number of factors that can be ...tted. It turns out that the appropriate condition is a generalisation of the number of rows of zero in the null space which we now develop. For convenience we state the main result.

THEOREM 2. If S is inadmissible then L is unbounded. If S is admissible:
(i) If rank(S) = n then L is bounded
(ii) If rank(S) = r < n and m ¸ r ¡ d(S) then L is unbounded
(iii) If rank(S) = r < n and m < r ¡ d(S) then L is bounded
where d(S) is the defect of matrix S, de...ned below.

Propositions 1-3 establish the ...rst part. We now proceed to the proof of parts (ii) and (iii). This builds on the concentration of the likelihood introduced by Lawley.

Lawley's Machinery

The following results are essentially due to Lawley (1940, 1942, 1943); Lawley and Maxwell (1963) give a convenient condensed version; Jöreskog (1967) also provides a useful account:

Assume § = §(¤; ^{a}). Then, by routine calculation,

(8)
$$@L=@¤ = ¡ 2§^{¡1}(§ ¡ S)§^{¡1}¤$$

(9)
$$@L=@^a = ¡ \text{diag}^i §^{¡1} (§ ¡ S) §^{¡1} {}^{\text{¢}}$$

(10)
$$§^{¡1} = a^{¡1} ¡ a^{¡1}¤^i I_m + ¤^0 a^{¡1} ¤ {}^{\text{¢} ¡1} ¤^0 a^{¡1}$$

If in addition, $@L=@¤ = 0$, then a little algebra shows

(11)
$$S a^{¡1}¤^i I_m + ¤^0 a^{¡1}¤ {}^{\text{¢}¡1} = ¤$$

(12)
$$§^{¡1}(§ ¡ S)§^{¡1} = a^{¡1}(§ ¡ S)a^{¡1}$$

(13)
$$@L=@^a = ¡ a^{¡1}\text{diag}(§ ¡ S) a^{¡1}$$

We make the normalisations $S^¤ = a^{¡\frac{1}{2}} S a^{¡\frac{1}{2}}$; $¤^¤ = a^{¡\frac{1}{2}}¤$: Then (??) is transformed to

(14)
$$S^¤¤^¤ (I_m + ¤^¤ ¤^{¤0})^{¡1} = ¤^¤$$

---

[16]An abuse of terminology that we will often ...nd, as here, too convenient to resist is to identify a linear space with its basis vectors written as a matrix.

Equation (??) is essentially a collection of eigenvector equations from which optimal Λ can be obtained for each value of a: Lawley-Jöreskog base an ML procedure on (??), searching over a: This is for the full-rank case wherein the existence of an ML solution is guaranteed according to Proposition 3. In the rank-de...cient case, with no such assurance in general, we need to proceed with a little more care. Nevertheless, we shall see below that, for given a, optimal Λ are indeed determined by (??).

Concentrating out Λ

Substitution from (??) into (??) and a little manipulation yields

(15)
$$L = \left[\log\det a + \log\det(I_m + \Lambda^{\Lambda_0}\Lambda^\Lambda) + \operatorname{tr} S^\Lambda \operatorname{tr} \Lambda^{\Lambda_0}S^\Lambda\Lambda^\Lambda(I_m + \Lambda^{\Lambda_0}\Lambda^\Lambda)^{-1} \right]$$

Since $L(a; \Lambda U) = L(a; \Lambda)$ for any conformable orthogonal matrix U it follows that we may replace Λ^Λ by $\Lambda^\Lambda U$ in (??), choosing U so that the columns of $\Lambda^\Lambda U$ are orthogonal vectors or 0s. Assume this has been done and extract the terms in Λ^Λ from (??):

$$L_0 = \sum_{i=1}^{n} [\log(1 + x_i^0 x_i) - x_i^0 S^\Lambda x_i / (1 + x_i^0 x_i)]$$

where the x_i are the columns of Λ^Λ. We wish to maximise L_0 over all systems of vectors $x_i; i = 1; ::::; n$ where the x_i are zero or pairwise orthogonal. Hold each $x_i^0 x_i$ ...xed and regard L_0 as a function of the normalised x_i. Since the supremum of $\sum_{i=1}^{m} x_i^0 S^\Lambda x_i$ is attained at eigenvectors of S^Λ or zero vectors (Rao, page 63) it follows that L_0 is a sum of terms of the form: $s_i^\Lambda x_i^0 x_i = (1 + x_i^0 x_i) - \log(1 + x_i^0 x_i)$. These terms contribute non-negatively if and only if $s_i^\Lambda \geq 1$. If $s_i^\Lambda < 1$ then L_0 is maximised by choosing $x_i = 0$. The optimal system of $x_i; i = 1; ::::; m$ then consists of the set of eigenvectors of S^Λ with modulus determined by

(16)
$$1 + x_i^0 x_i = \max(1; s_i^\Lambda)$$

Let $\frac{1}{4}(S^\Lambda)$ be the number of eigenvalues of S^Λ greater than 1 and de...ne $m_0 = \min(\frac{1}{4}(S^\Lambda); m)$. The columns of the optimal $\Lambda^\Lambda = \Lambda^\Lambda(a)$ in (??) thus consist of the ...rst m_0 eigenvectors of S^Λ, with modulus determined by (??), 0s elsewhere. De...ne $\Lambda(a) = a^{\frac{1}{2}}\Lambda^\Lambda(a)$. The concentrated likelihood function is now $L^c(a) = L(a; \Lambda(a))$. This function is well–de...ned for all S, be it full-rank or rank-de...cient, admissible or inadmissible.

Substituting in (??), one ...nds, up to a constant,

(17)
$$L^c(a) = \left[\log\det a + \sum_{i=1}^{m_0} (\log s_i^\Lambda - s_i^\Lambda) + \operatorname{tr} S^\Lambda \right]$$

This is the function we need to bound.

De...ne $F = f a \; 2 \; P; \tilde{A}_i \cdot s_{ii} g$. Then since $@L=@\Lambda = 0$ at $(a; \Lambda(a))$ and $@L^c=@a = @L=@a$ (the envelope theorem) it follows from (??) that $@L^c=@\tilde{A}_i < 0$ for $\tilde{A}_i > s_{ii}$ with the implication that, for each $a \; \not\in \; F$, there is an element of F for which the concentrated L takes a higher value. We de...ne the rim R as the set of a which maximise the concentrated likelihood function along rays from the origin. Although

outside F all directional derivatives are strictly negative, the rim does not necessarily lie within the bounded set F; however, given a point on the rim not within F, one can ...nd a dominating point within F by moving along a directional derivative towards F, subsequently passing back to the rim along a ray from the origin. It is easy to see that iterating this procedure leads to a dominating value in $R \setminus F$. It follows that the maximum is attained on $R \setminus F$ or on the boundary of F. Thus we have:

THEOREM 3. If L is bounded, it attains its maximum in F or $\overline{F} \cap F$ in the sense that

$$\sup L = \lim_{p \to 1} L^c(a^p)$$

where $\lim_{p \to 1} a^p = a^0 \in \overline{F} \cap F$ and $a^p \in R \setminus F$ for all p.

Such boundary solutions where some $\tilde{A}_i = 0$ can occur and are called Heywood solutions. At a Heywood solution one need not have $\partial L = \partial a = 0$: Since the likelihood function is evidently continuous where de...ned, it can be unbounded only at such boundary solutions. The study of the boundedness of the likelihood function thus consists in large part of the study of Heywood solutions. Examination of (??) reveals we need an apparatus to analyse the simultaneous behaviour of the $\tilde{A}_i$ and the s^π_j as some of the $\tilde{A}_i \to 0$: We proceed to this.

The box diagram

To complete the proof of Theorem 2 we consider the spectrum of S^π. Assume S is admissible. Consider a sequence $a^p \in F$ with $\lim_{p \to 1} a^p = a^0$ where $\tilde{A}^0_i = 0$ for some i: By passing to a subsequence and renaming indices if necessary, we can assume

$$\tilde{A}^p_1 \geq \tilde{A}^p_2 \geq ::: \geq \tilde{A}^p_n$$

for each p. We assume $\tilde{A}^p_j \to 0$ if and only if $j < N$. De...ne $S^{p\pi}$ corresponding to a^p and assume it has eigenvalues $s^{p\pi}_1, s^{p\pi}_2, :::, s^{p\pi}_n, \geq 0$ with corresponding unit eigenvectors x^p_i. If S is rank-de...cient; $r = \text{rank} S < n$; then the last $n_i r$ of the $s^{p\pi}_i$ are zero. De...ne $a^p_i = s^{p\pi}_i a^p$. Then the eigenvalue equation for $S^{p\pi}$ takes the form

(18) $$S(a^p_i)^{\frac{1}{2}} x^p_i = (a^p_i)^{\frac{1}{2}} x^p_i \qquad 1 = 1; :::; r$$

For each p, the system $x^p_i; i = 1; :::; r$, constitutes an orthonormal set in R^n and the compactness of the unit ball implies there exists a limit orthonormal system $x^0_i; i = 1; :::; r$ where $x^0_i = \lim_{p \to 1} x^p_i$ (passing to a subsequence if necessary). Treating (??) as an equation of the form $S x = y$, one deduces

$$(a^p_i)^{\frac{1}{2}} x^p_i = S^i (a^p_i)^{\frac{1}{2}} x^p_i + z$$

where $z \in N(S)$, and S^i is any generalised inverse of S, which we may take to be positive de...nite. Since $N(S^{p\pi}) = (a^p_i)^{\frac{1}{2}} N(S)$, it follows that

(19) $$x^p_i = (a^p_i)^{\frac{1}{2}} S^i (a^p_i)^{\frac{1}{2}} x^p_i + z^\pi$$

where $z^\pi \in N(S^{p\pi})$. Thus, since $x^p_i ? z^\pi$;

(20) $$1 = x^{p0}_i (a^p_i)^{\frac{1}{2}} S^i (a^p_i)^{\frac{1}{2}} x^p_i \qquad i = 1; :::; r$$

It follows from (??) that

$$(21) \qquad 1 = s_{max}^i \cdot \overset{\circ}{\underset{\circ}{\circ}} (a_i^p)^{\frac{1}{2}} x_i^{p \overset{\circ}{\underset{\circ}{\circ}} 2} \cdot 1 = s_{min}^i$$

where s_{max}^i and s_{min}^i denote the largest and smallest eigenvalues of S^i, respectively. Thus, passing if necessary to a subsequence of a^p, we deduce that $\tilde{A}_j^p s^{p\pi}$ approaches a ...nite limit on the support of x_i^0 (the indices r for which $x_{ir}^0 \ne 0$) and a non-zero limit for at least one r in the support. The limiting behaviour of $\tilde{A}_j^p s_i^\pi$ is indicated in the diagram:

Figure A1 The Box Diagram

The shaded areas indicate $(i; j)$ pairs for which $\tilde{A}_j s_i^\pi$ approaches a non-zero limit. The diagram incorporates the following properties:

1. The pairs $(i; j)$ are arranged as non-overlapping boxes. Each $i \cdot r$ corresponds to one box. If $(i; j)$ and $(i^0; j^0)$ belong to the same box then $\tilde{A}_j s_i^\pi = \tilde{A}_{j^0} s_{i^0}^\pi$; $\tilde{A}_j = \tilde{A}_{j^0}$ and $s_i^\pi = s_{i^0}^\pi$ all approach non-zero limits. To see this, observe that the boxes could be constructed as follows. For each i, the j for which $\tilde{A}_j s_i^\pi$ approaches a non-zero limit form a vertical line-segment by the monotonicity of $\tilde{A}_j$ with respect to j. Next note that if two such line-segments were to overlap (share a common j), they would lie in a box with the stated properties.

2. Above the boxes, $\tilde{A}_j s_i^\pi \,!\, 0$: This follows by construction.

3. Beneath the boxes, $x_{ik}^0 = 0$. This follows by construction since $\tilde{A}_j s_i^\pi$ approaches a ...nite limit on the support of x_i^0:

4. The boxes start in the top left hand corner. If not we would have $\tilde{A}_1 s_1^{\alpha} \to 0$ and hence $\tilde{A}_1 s_i^{\alpha} \to 0$ for all i. But $\mathrm{tr} S^{\alpha} = \sum_{i=1}^{r} s_i^{\alpha} = \sum_{i=1}^{n} s_{ii} = \tilde{A}_i$ so then $0 = \lim \tilde{A}_1 \sum_{i=1}^{r} s_i^{\alpha} \geq s_{11} > 0:$

5. The boxes are tall: their height is not exceeded by their breadth. To see this, ...rst de...ne

$$G(x_i^0) = \lim_{p \to 1} {}^{a} p^{\frac{1}{2}} x_i^0$$

and extend to the span by linearity; de...ne also

$$H(x) = (S^i)^{\frac{1}{2}} G(x):$$

If $u; v$ on the horizontal axis belong to the same box then

$$H(x_u^0)^0 H(x_v^0) = \lim_{p \to 1} x_u^{0 \, 0a} p^{\frac{1}{2}} S^i {}^{a} p^{\frac{1}{2}} x_v^0$$
$$\neq 0$$

since the limits of ${}^{a}{}_{u}$ and ${}^{a}{}_{v}$ di¤er by a non-zero multiplicative constant in the same box. It follows that H and hence G preserve dimension on the span of the x_i^0 corresponding to each box. But the support of each $G(x_i^0)$ lies within such a box so the result follows.

The Null Space

So far we have not de...ned limit eigenvectors in the null space $N(S^{p\alpha})$: Let $z_{r+1}; ::::; z_n$ be a basis for $N(S)$ in reduced column-echelon form and, for $j = 1; ::::; n;$ de...ne $V(j)$ as the largest index $r + k$ for which the leading 1 has row-number greater than or equal to $j:$ In the event that the basis has rows of zeros at the bottom, we de...ne $V(j) = r$ for these rows. The step-shaped path $(j; V(j))$ thus de...nes the frontier between the zero and non-zero regions of $N(S):$ Let s be the minimum number of non-zero entries among all non-zero vectors of $N(S)$ and de...ne the defect of $S; d(S)$ by

$$s = r_i d(S) + 1$$

Reduction to echelon form will produce $r + 1$ or fewer non-zero entries in z_n and $d(S) > 0$ indicates extra structure in $N(S):$ This can be so only if an algebraic identity holds among the variates used to generate the covariance matrix S, typically a zero-probability event. Note that each row of zeros in the null space contributes one to the defect. It is possible, perhaps having ...rst permuted the indices $i;$ to write a basis for $N(S)$ in column-echelon form so that $V(r_i d(S) + 1) = n.$ In this case the right-most column in the null space has non-zero entries for $j \cdot r_i d(S) + 1,$ zeros elsewhere. In general (without necessarily permuting the indices), column-echelon form delivers $V(r_i d + 1) = n$ for some $d \geq d(S)$ where $r_i d + 1$ is the number of terms lying above (and including) the leading 1 in $z_n.$[17]

Column-echelon form ensures the following properties of the function V:

1. $V(j) = n$ for $j \cdot r_i d + 1:$
2. $0 \geq \mathbb{C}V(j) \geq i 1;$ where $\mathbb{C}V(j) = V(j)_i V(j_i 1):$
3. $\mathbb{C}V(r_i d + 1) = i 1:$

Now de...ne $F(j)$ by

$$V(j)_i F(j) = n_i j \text{ for } j = 1; ::::; n$$

---

[17] As de...ned, d is invariant to the particular column-echelon form.

Then F inherits the properties:

1. $F(j) = j$ for $j \cdot r_i d + 1$:
2. $1 \ge \mathfrak{C}F(j) \ge 0$:
3. $\mathfrak{C}F(r_i d + 1) = 0$:
4. $i_i F(j) \cdot 1$ for $j > r_i d + 1$;

where property 4 follows from 2 and 3 immediately above.

Limit eigenvectors in $N(S^{p\alpha})$ are constructed as follows. In the event that $N \cdot r_i d + 1$ then $x^0_{r+i} = {}^a 0^{\frac{1}{2}}z_{r+i}$, $i = 1; ::::; n_i r$, are linearly independent, each a limit of the sequence ${}^a p^{\frac{1}{2}}z_{r+i}$ and orthogonal to x^0_i; $i = 1; ::::; r$: When $N > r_i d + 1$, the last $n_i V(N)$ vectors as de...ned above vanish, but these can be replaced by $\lim_{p! \ 1} ({}^a p_{=}\tilde{A}_k)^{\frac{1}{2}}z_{r+i}$, where k is the row number of the leading 1 in z_{r+i}: These are all limits of sequences in $N(S^{p\alpha})$ and have some non-zero elements for indices j less than N: In both cases we are led to limit null-space vectors $x^0_{r+1}; ::::; x^0_n$ for which: (a) If $N < r_i d + 1$ then the limit vectors have zeros for indices less than N: (b) If $N \ge r_i d + 1$ then the limit vectors can be divided into two groups, one with zero elements at indices above N, one with zero elements at indices below N, the dividing vertex being $(N; V(N))$: This structure is summarised in Figure A2 below.

LEMMA 1. In the box-diagram, the right-most box lies beneath $j = N$ if and only if $F(N) < r + 1$: In this case, the North-West vertex of the box lies on the path $(N; F(N))$:

PROOF. We prove the second part ...rst. Assume therefore that the right-most box lies beneath $j = N$.: Taking limits in (??), we deduce that the subvectors of each x^0_i; $i = 1; ::::; r$ lying above $j = N$ belong to the limit null-space, analogously truncated. Exploiting the orthogonality structure implicit in Figure A2, and the mutual orthogonality between $x^0_1; ::::; x^0_r$ and $x^0_{r+1}; ::::; x^0_n$, one can deduce that these subvectors are identically zero. It follows that the limit eigenvectors in the right-most box, together with the limit null space vectors up to index $V(N)$, all have zeros above $j = N$, whereas all other eigenvectors have zeros below $j = N$: Figure A3 illustrates the argument (note we have drawn the V and F functions as straight lines for simplicity). Considerations of dimensionality now show that the submatrix of the right-most box lying beneath $j = N$ shown shaded in Figure 3, together with the correspondingly truncated vectors in the adjacent region of the limit null space, forms a square. The result now follows from the de...nition of the function F:

Figure A2 Location of zeros in the limit null space

Figure A3 Interaction between the null space and the right-most box

If there is no box beneath $j = N$ then the above argument shows that the ﬁrst block of truncated vectors in the null space themselves form a square, and are linearly independent by construction. It follows that

$$V(j) = j + r \quad \text{for } j = 1, \ldots, N - r$$

whence $F(j) = 1 + r.$ Clearly this argument is reversible, so the Lemma is proved. ¤

Proof of Theorem 2

We ﬁrst prove boundedness for $m < r - d(S).$ Assume, by way of contradiction, L^c is unbounded for some sequence $a^p \to a^0.$ In choosing the elements $\tilde{A}_j^p$ to be monotonic in j, we have implicitly re-ordered the indices $j.$ We have however $m < r - d(S) \cdot r - d$ where $V(r - d + 1) = n.$ Clearly we may take all a^p to lie in $R \setminus F$ so that $a^0 \in \overline{F} \cap F$. We ﬁrst prove the following.

LEMMA 2. On the rim,

$$(22) \qquad \operatorname{tr} S^\pi - \sum_{i=1}^{N_0} s_i^\pi = n - m_0$$

and, up to an additive constant,

$$(23) \qquad L^c = \left(\sum_{i=1}^{\tilde{A}} \log \tilde{A}_i + \sum_{i=1}^{N_0} \log s_i^\pi \right)$$

Moreover, $m = m_0$; i.e. $s_i^\pi \geq 1$ for $i \leq m.$

PROOF. For $\pm > 0$, one has from (??)

$$L^c(\pm^a) = \left(\sum \log \det^a + \sum_{i=1}^{N_0} \log s_i^\pi + (n - m_0)\log \pm + (\operatorname{tr} S^\pi - \sum_{i=1}^{N_0} s_i^\pi) = \pm \right)$$

Now $\operatorname{tr} S^\pi = \sum_{i=1}^{n} s_i^\pi$ so $\operatorname{tr} S^\pi > \sum_{i=1}^{m_0} s_i^\pi$ in virtue of the assumption that $m <$ rank $S = $ rank $S^\pi.$ The integer m_0 can depend on $\pm$ but, irregardless, $L^c(\pm^a)$ can be made arbitrarily negative by choice of $\pm$ suﬃciently small or large. Continuity of $L^c(\pm^a)$ then guarantees that a maximum, at $\pm_m$ say, is attained. Fixing m_0 to be the value at this maximum, one deduces that $\pm_m = (\operatorname{tr} S^\pi - \sum_{i=1}^{P} s_i^\pi) = (n - m_0).$ Since $\pm_m = 1$ if a already belongs to the rim, the ﬁrst result follows and the second is obtained by direct substitution in ??. Finally, m_0 can be smaller than m only if $s_{m_0+1} < 1$; however ?? implies that the average of the numbers $s_i^\pi; i = m_0 + 1, \ldots, n$, some of which are zero, is unity, inconsistent with this. ¤

The implication of (??) is that, for a^p on the rim, $0 \leq s_i^\pi \leq 1$ for $i = m + 1, \ldots, n.$ We may thus choose a subsequence of a^p for which each s_i^π converges, which implies that $\tilde{A}_j s_i^\pi$ converges to a ﬁnite limit for $j \geq N - r; i > m.$ (??) further implies that at least one of these limits is non-zero. It follows that the right-most box lies beneath $j = N$ in the box diagram and hence that the North-West vertex of this box is given by $(N, F(N))$ where $m + 1 \geq F(N).$ If $N > r - d + 1$ then, since $m < r - d$; we have

$$m + 1 < r_i \; d + 1 \cdot \; F(N)$$

by properties (i) and (ii) of F above. This is contrary to $m + 1 \; \geq \; F(N)$ and we conclude that $N \cdot \; r_i \; d + 1$:[18] It follows that $F(N) = N$ by property (i) of F above i.e. the NW vertex of the right-most box lies on the diagonal. But since boxes are "tall", all boxes must then lie along the diagonal with the implication that $\tilde{A}_i s_i^{\alpha}$ converges to a non-zero limit for all $i \cdot \; r$: If (??) is recast in the form

$$L^c = \; i \; \left(\sum_{i=1}^{\tilde{A}_{x^n}} \log \tilde{A}_i s_i^{\alpha} + \sum_{i=m+1}^{x^i} \log \tilde{A}_i \right)!$$

then boundedness now follows since $N = F(N) \cdot \; m + 1$: This completes the proof of Theorem 2(iii).

Now assume that $m \; \geq \; r_i \; d(S) \; , r < n$: We assume $d > 0$; for otherwise unboundedness would follow from Proposition 3. Thus $m \cdot \; r_i \; 2$: We re-order the indices j so that a reduced column-echelon form for $N(S)$ has leading 1 in place $r_i \; d(S) + 1$ for z_n. We choose a convergent sequence $^{a\,p}$ with $N = m + 2$ so that $n \; \geq \; N > r_i \; d(S) + 1$: No harm is done by assuming m_0 is ...xed as p changes. If $m_0 < m$, we must have $s_{m_0+1}^{p\alpha} \cdot \; 1$, so it follows that $m_0 + 1 \; \geq \; F(N)$: Write (??) in the form

$$L^c = \; i \; \left(\sum_{i=1}^{\tilde{A}_{x^{m_0}}} \log \tilde{A}_i s_i^{\alpha} + \sum_{i=m_0+1}^{x^i} s_i^{\alpha} + \sum_{i=m_0+1}^{x^i} \log \tilde{A}_i \right)!$$

Considering these three terms in turn, note that the ...rst approaches a ...nite limit or $+1$ since the diagonal in the box diagram intersects the boxes or lies above them; the second is bounded because since $m_0 + 1 \; \geq \; F(N)$; while the third contains a term in $i \; \log \tilde{A}_{m_0+1}$ which approaches $+1$ since $m_0 + 1 < N$: This proves Theorem 2(ii) so the proof of Theorem 2 is now complete.

Figure 4 gives the box-diagram for a convergent ML estimation of the factor model.

---

[18]This argument has shown, in fact, that $N > r_i \; d + 1$ implies the sequence $^{a\,p}$ is not on the rim.

Figure A4 The box diagram for convergent maximum likelihood

In the shaded boxes, $\tilde{A}_j\, s_i^\sharp$ approaches a non-zero limit, above them, zero; beneath the boxes the limit eigenvectors x_i^0 have zero support. The ...gure also gives the zero-structure for the limit null space. We have also established that $m \geq N_i\; 1$: De...ne the multiplicity of a Heywood solution as the number of indices for which $\tilde{A}_j\;!\;0$, $N_i\; 1$ in our terminology. Thus our proof of Theorem 2 has established:

COROLLARY. The multiplicity of a Heywood solution is less than or equal to the number of ...tted factors.

THEOREM 4. Assume $m < r_i\; d$. Then there exists an invertible $\S^0$ in $R \setminus F$ or its boundary to maximise L:

PROOF. By Theorems 2 and 3, if $m < r_i\; d$, there exists a sequence $^{a\,p}$ in $R \setminus F$ with limit $^{a\,0}$ such that $\sup L = \lim_{p!\; 1} L^c(^{a\,p})$: It was shown in the proof of 2(iii) that $F(N) = N$: If $\S^0$ corresponds to $^{a\,0}$ then

$$\S^0 = \lim_{p!\; 1} (^{a\,p} + ¤^p ¤ p^0)$$

(24)
$$= \lim_{p!\; 1} {}^{a\,p\frac{1}{2}} (I + ¤^\sharp ¤ ^{¤^0})^{a\,p\frac{1}{2}}$$

where $¤^\sharp$ consists of the ...rst m eigenvectors of $S^{p\sharp}$ with modulus determined by (??). Let x_i^p, $i = 1; :::n$, be an orthonormal basis of eigenvectors of $S^{\sharp p}$, ranked by eigenvalue, and de...ne $q_i^p = s_i^{\sharp p}$ for $i \cdot m$, 1 for $i > m$: Then $I = \mathbf{P}_{i=1}^{n} x_i^p x_i^{p^0}$ so

it follows that

$$\S^0 = \lim_{p \to 1} {}^{a} {}^{p\frac{1}{2}} \left(\sum_{i=1}^{\mathbf{X}} q_i^p x_i^p x_i^{p^0} \right) {}^{a} {}^{p\frac{1}{2}}$$

$$= \sum_{i=1}^{N^{-1}} G(x_i^0) G(x_i^0)^0 + \sum_{i=N}^{\mathbf{X}} q_i^{0 a\, 0\frac{1}{2}} x_i^0 x_i^{0^0} {}^{a}{}^{0\frac{1}{2}}$$

where G is as de...ned in the proof of property (v) of the box-diagram, $\tilde{A}_j^0 > 0$ for $j \, \mathbf{,} \, N$; and $q_i^p \, \mathbf{,} \, 1$. The $G(x_i^0)$ are linearly independent in the same box and orthogonal between boxes, as well as orthogonal to each ${}^{a}{}^{0\frac{1}{2}} x_i^0$, $i \, \mathbf{,} \, N$, in virtue of the structure of the box-diagram. Exploiting the fact that the ${}^{a}{}^{0\frac{1}{2}} x_i^0$ vanish above $j = N$ for $i \, \mathbf{,} \, N$, one deduces that the ${}^{a}{}^{0\frac{1}{2}} x_i^0$ are linearly independent and the result follows. ¤

Maximising the Likelihood

There is a voluminous literature on ML estimation of the factor model (see e.g. Lee and Jennrich, 1979, Rubin and Thayer, 1982). Analytic maximisation function is intractable, so numerical techniques are used. Our prime requirement for numerous Monte Carlo experiments was guaranteed convergence, at the expense, if necessary, of speed. Our reading of the literature on factor ML algorithms is that choice of algorithm can a¤ect convergence speed by a factor of about 3, though not by factors of 10 or more. Given the speed of modern machines this would indicate that the particular algorithm chosen may be of secondary importance. After some experimentation, we have chosen a mixture of steepest ascent and Fletcher-Powell algorithms. Use of steepest ascent alone can be very time consuming, the geometry of the likelihood function being such that a large number of very small steps can be required (hem-stitching). Fletcher-Powell, on the other hand, can get stuck where, in the search direction, there is no appreciable likelihood value improvement, but the gradient is non-zero in a di¤erent direction i.e. the likelihood function is not well-approximated by an ellipsoid in certain regions. Mixing the two schemes ensures that the speed of Fletcher-Powell is used where it improves the likelihood function, but, if it runs into problems, we switch to the more robust steepest ascent to move the search along. We also implement a scheme whereby the step length increases following a successful step (i.e. one that improves the likelihood). It seems likely that the appropriate mix of steepest ascent and Fletcher-Powell and the step expansion, is problem dependent; we have not attempted to optimise in this direction. We have found that this mixture provides a very robust maximisation procedure. For the regressions in this paper we used the following scheme. The starting values are derived by assuming that ${}^{a}{}_0$ is scalar with magnitude the average of the smallest $n \, \mathbf{i} \, m + 1$ eigenvalues of S; and as ¤$_0$ the implied square root of $S \, \mathbf{i} \, {}^{a}{}_0$. The search method uses steepest ascent for the ...rst 100 steps, then switches to Fletcher-Powell. Following a successful step, the step-length is expanded by 20%. Convergence is checked by the change in the likelihood function being below some tolerance. At this point the gradient of the likelihood function is also checked. If the gradient exceeds a critical value we switch back to steepest ascent for a further 40 steps and then revert to Fletcher-Powell. This process continues until both the change in the likelihood and the slope of the likelihood fall beneath required thresholds. Convergence characteristics with this algorithm seem good; in several thousand Monte Carlo we have never failed to ...nd a maximum.

References

[1] Barro, R. (1991) "Economic Growth in a Cross Section of Countries", Journal of Political Economy, 100(2) pp 223-51.

[2] Caselli, F. Esquivel, G and F. Lefort, (1996) "Reopening the Convergence Debate: A New Look at Cross Country Growth Empirics" Journal of Economic Growth, September, pp 363-389.

[3] Chamberlain, G and M.Rothschild (1983) "Arbitrage and Mean Variance Analysis on Large Asset Markets", Econometrica, 51, pp1281-1304.

[4] Conley, T.G. (1999) "GMM estimation with cross sectional dependence", Journal of Econometrics, 92, pp1-45.

[5] Connor, G and R.Korajczyk (1986) "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis" Journal of Financial Economics, 15, pp373-394

[6] Connor, G and R.Korajczyk (1988) "Risk and Return in an Equilibrium APT: Application of a New Test Methodology" Journal of Financial Economics, 21, pp255-289.

[7] Di Liberto, A. and J.Symons. (1998) "Econometric Issues in Convergence Regressions" (mimeo).

[8] Driscoll, J.C. and A.C. Kraay (1998) "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data", The Review of Economics and Statistics, 80, pp 549-560.

[9] Evans, P. and G.Karras (1996) "Convergence Revisited", Journal of Monetary Economics, 37, pp 249-265.

[10] Fisher, S. (1993) "The role of macroeconomic factors in growth", Journal of Monetary Economics pp 485-512.

[11] Frees, E.W. (1995) "Assessing cross-sectional correlation in panel data", Journal of Econometrics, 69, pp393-414

[12] Greene, W (1990) Econometric Analysis, Macmillan.

[13] Islam, N (1995) "Growth Empirics: A Panel Data Approach" Quarterly Journal of Economics, November.

[14] Lee, S.Y. and R.Jennrich (1979) "A Study of Algorithms for Covariance Structure. Analysis with Speci...c Comparisons Using Factor Analysis", Psychometrica, 44(1), pp 99-113.

[15] Jöreskog, K.G. (1967) "Some Contributions to Maximum Likelihood Factor Analysis", Psychometrica, 32(4), pp443-482.

[16] Keane, M. and D.Runkle (1992) "On the Estimation of Panel Data with Serial Correlation when Instruments are not Strictly Exogenous", Journal of Business and Economic Statistics, 10(1) pp 1-9.

[17] Lawley, D.N. (1940) "The Estimation of Factor Loadings by the Method of Maximum Likelihood", Proceedings of the Royal Society of Edinburgh Section A, 60, pp 64-82

[18] Lawley, D N. (1942) "Further Investigations in Factor Estimation" Proceedings of the Royal Society of Edinburgh Section A, 61, pp176-189

[19] Lawley, D.N. (1943) "The application of the maximum likelihood method to factor analysis" British Journal of Psychology, 33, pp172-175

[20] Lawley, D.N. and A.E.Maxwell, (1963) Factor Analysis as a Statistical Method, London, Butterworths

[21] Liang, K-Y. and S.L.Zeger (1986) "Longitudinal Data Analysis using Generalised Linear Models" Biometrika 73,1, pp13-22.

[22] Lee, K., Pesaran, H. and R.Smith (1995) "Growth and Convergence; A Multicountry Empirical Analysis of the Solow Growth Model" (mimeo)

[23] Mankiw, N. Romer, D. and D.Weil (1992) "A Contribution to the Empirics of Economics Growth", Quarterly Journal of Economics, 107(2), pp 407-37.

[24] Maxwell,A.E. (1981) "Factor Analysis" in Encyclopedia of Statistical Sciences Vol 3 ed. S.Kotz and N.L.Johnson, John Wiley & Sons.

[25] Neyman, J and E.Scott, (1948) "Consistent Estimates Based on Partially Consistent Observations" Econometrica, 16(1) pp1-32.

[26] Nickell, S.J. (1981) "Biases in Dynamic Models with Fixed E¤ects", Econometrica, 47(5), pp 1249-66.

[27] Rao, C. R. (1973) Linear Statistical Inference and its Applications, Wiley, New York.

[28] Rubin D and D.Thayer (1982) "EM Algorithms for ML Factor Analysis", Psychometrica, 47(1), pp 69-76.

[29] Sargan, J. (1988) Lectures on Advanced Econometric Theory, Blackwell.

[30] Theil, H. (1971) Principles of Econometrics, Wiley, New York

[31] Zellner, A. (1962) "An E¢cient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias" Journal of the American Statistical Association, vol 57, pp 348-368.

University of Cambridge

University College London

472	Brian Bell Stephen Nickell Glenda Quintini	Wage Equations, Wage Curves and All That
471	Marek Dabrowski Stanislaw Gomulka Jacek Rostowski	Whence Reform? A Critique of the Stiglitz Perspective
470	B. Petrongolo C. A. Pissarides	Looking Into the Black Box: A Survey of the Matching Function
469	W. H. Buiter	Monetary Misconceptions
468	A. S. Litwin	Trade Unions and Industrial Injury in Great Britain
467	P. B. Kenen	Currency Areas, Policy Domains and the Institutionalization of Fixed Exchange Rates
466	S. Gomulka J. Lane	A Simple Model of the Transformational Recession Under a Limited Mobility Constraint
465	F. Green S. McIntosh	Working on the Chain Gang? An Examination of Rising Effort Levels in Europe in the 1990s
464	J. P. Neary	R&D in Developing Countries: What Should Governments Do?
463	M. Güell	Employment Protection and Unemployment in an Efficiency Wage Model
462	W. H. Buiter	Optimal Currency Areas: Why Does the Exchange Rate Regime Matter?
461	M. Güell	Fixed-Term Contracts and Unemployment: An Efficiency Wage Analysis
460	P. Ramezzana	Per Capita Income, Demand for Variety, and International Trade: Linder Reconsidered
459	H. Lehmann J. Wadsworth	Tenures that Shook the World: Worker Turnover in Russia, Poland and Britain
458	R. Griffith S. Redding J. Van Reenen	Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries

457	J. Swaffield	Gender, Motivation, Experience and Wages
456	C. Dougherty	Impact of Work Experience and Training in the Current and Previous Occupations on Earnings: Micro Evidence from the National Longitudinal Survey of Youth
455	S. Machin	Union Decline in Britain
454	D. Marsden	Teachers Before the 'Threshold'
453	H. Gospel G. Lockwood P. Willman	The Right to Know: Disclosure of Information for Collective Bargaining and Joint Consultation
452	D. Metcalf K. Hansen A. Charlwood	Unions and the Sword of Justice: Unions and Pay Systems, Pay Inequality, Pay Discrimination and Low Pay
451	P. Martin H. Rey	Financial Integration and Asset Returns
450	P. Martin H. Rey	Financial Super-Markets: Size Matters for Asset Trade
449	A. Manning	Labour Supply, Search and Taxes
448	S. Nickell L. Nunziata	Employment Patterns in OECD Countries
447	W. Buiter	The Fallacy of the Fiscal Theory of the Price Level
446	R. Portes H. Rey	The Determinants of Cross-Border Equity Flows
445	G. Duranton D. Puga	Nursery Cities: Urban Diversity, Process Innovation and the Life-Cycle of Products
444	D. Quah	UK Phillips Curves and Monetary Policy
443	L. Feinstein	The Relative Economic Importance of Academic, Psychological and Behavioural Attributes Developed in Childhood
442	D. Quah	Cross-Country Growth Comparison: Theory to Empirics