Panel Time-Series

Ron P. Smith
Department of Economics Mathematics and Statistics
Birkbeck University of London
e-mail: R.Smith@bbk.ac.uk

Ana-Maria Fuertes
Sir John Cass Business School,
City University, London
e-mail:A.Fuertes@city.ac.uk

August 2016

Abstract

Traditionally economic panels had large number of cross-section units and relatively few time periods and econometric methods were developed for such 'large N small T 'data. More recently panels with observations for a large numbers of time periods have become available on cross-section units like firms, industries, regions or countries. These notes explore the econometric methods developed for such 'large N large T' data. Such data allow more explicit treatment of (a) heterogeneity across units (b) dynamics, including unit roots and cointegration and (c) cross-section dependence arising from spatial interactions or unobserved common factors.

1 Introduction

1.1 Panels

Panel (or longitudinal) data provides us with observations on cross-section units (e.g. individuals, firms, industries, countries, regions), $i = 1, 2, ..., N$ over repeated time periods, $t = 1, 2, ..., T$. The cross-section units will be referred to as units or groups. We will start with the case where we want to measure the effect on a scalar dependent variable, y_{it}, of a set of regressors, a $k \times 1$ vector $\mathbf{x}_{it}$, which may include lagged y_{it} and then move onto vector dependent variables. Panel data present advantages over a single cross-section $T = 1$, or a single time series $N = 1$. We will be concerned with cases where both N and T are large. Until recently, there was no text book covering large N and T panels but Pesaran (2015) now fills that gap. Not only do such panels provide larger samples which may improve efficiency and mitigate multicollinearity they can allow for more heterogeneity than small T panels, allow for more complex dynamic models, and allow us to identify unobserved factors that influence all the units. Widely used panel data sets include the cross country Penn World Tables, regional data for the US and other countries and data on individual firms. We will emphasise balanced panels, but many of the results carry over to unbalanced panels when the number of time periods T_i differs across units.

The chapters cover:

1. Introduction
2. Static Models
3. Dynamic Linear Regression
4. Unit Roots
5. Cointegration in Single Equations
6. Cointegrating VARs
7. Cross Section Dependence: Factor Models
8. Cross Section Dependence: Estimators
9. Within and between group cointegration: the GVAR
10. Concluding comments

1.2 N and T

Econometrics has tended to become specialised into micro-econometrics where N is large and time-series econometrics where T is large. The econometric theory for panel data was largely developed for survey data where T was small (often only four or five) but N was large.[1] The asymptotic statistical theory was derived by letting $N \to \infty$ for fixed T, in contrast to the asymptotic theory for time-series analysis which was done by letting $T \to \infty$ for fixed N (often equal to one). In recent years, more large N, large T panels have become available, where there are fairly long time series for a large number of countries, industries, firms or regions.

With large T panels, the asymptotic analysis is done letting both N and T go to infinity, often known as double index asymptotics. There are various ways that this can be done. The simplest way is to let one argument, say, T go to infinity first and then let the other, say, N go to infinity second. This is called taking *sequential limits*, and one may get different answers depending on the order in which the limits are taken. A second way is to allow them to approach infinity together along a specific diagonal path of the two-dimensional array $T \times N$, say $T = T(N)$ where $T(N)$ is a specific function, as N goes to infinity. The results may not hold for other paths. A third approach looks at *joint limits*, and although it does not require a particular path, it may require restrictions on the rates at which N and T go to infinity, e.g. $N/T \to 0$.[2] When discussing the consistency of an estimator in a panel data context, always specify whether it is consistent: for $N \to \infty$; for $T \to \infty$; for either $N \to \infty$ or $T \to \infty$; or for both $N \to \infty$ and $T \to \infty$. Often this is not done, and not doing it can provoke heated arguments between micro and time-series econometricians shouting "It is consistent!" "It is not consistent!" at each other.

For example, consider the dynamic fixed effects model, with homogeneous slopes but different intercepts

$$y_{it} = \alpha_i + \beta x_{it} + \gamma y_{it-1} + u_{it};\ u_{it} \sim IN(0, \sigma^2)$$

where the independence assumption for the innovations refers to time and cross-section, $E(u_{it}u_{jt-s}) = 0$ for $i \neq j$ or $s \neq 0$, and the process is stationary $-1 < \gamma < 1$. The least squares estimates of the parameters are consistent, $T \to \infty$, N fixed; but inconsistent $N \to \infty$, T fixed: the estimate of the coefficient of the lagged dependent variable, γ, is biased downwards for small T. If both $N \to \infty$ and $T \to \infty$, then to ensure consistency of the least squares estimates T must grow sufficiently fast relative to N, such that $N/T \to \kappa$, where $0 \leq \kappa < \infty$, Alvarez and Arellano (2003). In many cases, choice of an appropriate estimator depends on the relative size of N and T.

An inevitable question is: at what values do N and T become 'large'? T should be large enough that it is sensible to estimate a different time-series model for each unit, we are considering panel time-series. If T is small different estimators than the ones considered in these notes are appropriate. The sample size T which is required for sensible estimation of the individual equations depends on the model. For a simple bivariate regression, with a strictly exogenous variable, $T = 10$ may be large enough. For more complicated models with

[1] See Hsiao (1986, 2003), Baltagi (2008), Matyas and Sevestre (1996,2008) Wooldridge (2010) and Arellano (2003) for surveys.

[2] Phillips and Moon (1999, 2000) provide an introduction.

lagged dependent variables, T would need to be sufficiently large that one could rely on the consistency properties. Similar arguments hold for N being large if averaging across units is required for consistency or for central limit theorems to be valid. In many cases, as with the dynamic fixed effect example above, the relative size of N and T also matters. Always consider what estimators N and T allow you to use. If T is large, start by estimating N separate time-series regressions and consider the distribution of the coefficients over units.

1.3 Issues

Large N, large T datasets (sometimes called data fields, Quah 1994)) raise three main issues: heterogeneity, dynamics and cross-section dependence.

1.3.1 Heterogeneity

First, since it is possible to estimate a separate regression for each unit, which is not possible in the small T case, it is natural to consider heterogeneous panel models where the parameters can differ over units. One can test for equality of the parameters, rather than assuming it, as one is forced to do in the small T case. This homogeneity hypothesis is very often rejected and the differences in the estimates between units can be large. Baltagi and Griffin (1997) discuss the dispersion in OECD country estimates for gasoline demand functions and Baltagi et al. (2002) the dispersion in regional French gasoline demand. In both cases they look at the performance of a very large range of estimators. Boyd and Smith (2002) review possible explanations for the large dispersion of estimates in models of the monetary transmission mechanism for 57 developing countries. Eberhardt and Teal (2010) argue that the treatment of heterogeneity is central to understanding of the growth process and Bond and Eberhardt (2013) suggest an estimator that allows for heterogeneity.

How one should treat *heterogeneity* depends on what the parameters of interest are, what you are trying to estimate? This may be the coefficients of the individual units, say β_i for $i = 1, 2, ..., N$, or the expected values and the variances of the coefficients over the units, $E(\beta_i)$ and $V(\beta_i)$. This will depend on the purpose of the exercise.

For illustration, consider a simple example, where for each group $y_{it} = \mu_i + \varepsilon_{it}$, $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}^2) = \sigma_\varepsilon^2$. In addition to this within group variation around a constant group-specific mean, there is also between group variation in the means $\mu_i = \mu + \eta_i$, $E(\eta_i) = 0$, $E(\eta_i^2) = \sigma_\eta^2$. Consider the various averages that we could construct.

$$\begin{array}{lll} \overline{y} = (NT)^{-1} \sum_i \sum_t y_{it}; & E(\overline{y}) = \mu; & V(\overline{y}) = (NT)^{-1}(\sigma_\varepsilon^2 + \sigma_\eta^2) \\ \overline{y}_i = (T)^{-1} \sum_t y_{it}; & E(\overline{y}_i) = \mu_i; & V(\overline{y}_i) = T^{-1}\sigma_\varepsilon^2 \\ \overline{y}_t = (N)^{-1} \sum_i y_{it}; & E(\overline{y}_t) = \mu; & V(\overline{y}_t) = N^{-1}(\sigma_\varepsilon^2 + \sigma_\eta^2) \end{array}$$

The three averages provide measures of two different things. The first and third provide unbiased estimates of the population average μ, the second provides an unbiased estimate of the group specific average μ_i. All three are consistent estimators, but they are consistent estimators of two different things. The first is consistent for μ, in that the variance goes to zero, for either large N or large T; the second is consistent for μ_i, large T; the third is consistent for μ large N.

1.3.2 Dynamics

Secondly, large T allows one to estimate less restrictive dynamic models, with lagged variables. Most economic time-series data are non-stationary, one aspect of which is their order of integration: the number of times it must be differenced to make it stationary. Many economic time-series appear to be integrated of order one, denoted I(1), and must be differenced once (after perhaps transformations, like logarithms, and removing deterministic elements like a linear trend and seasonals) to make them stationary. Such series are said to contain one unit root. However, many equilibrium or arbitrage conditions imply that specific linear combinations of these I(1) variables are stationary and this is termed cointegration.

For instance, the logarithms of the spot exchange rate for a country, s_{it}, domestic and foreign prices, p_{it} and p_{it}^* may all be I(1) but Purchasing Power Parity (PPP) predicts that the real exchange rate, $q_{it} = s_{it} - p_{it} + p_{it}^*$, should be I(0), the variables cointegrate. Thus extending the estimation and testing procedures for integrated and cointegrated series to panels is a natural development. The fact that individual time series tend to reject Purchasing Power Parity (fail to reject unit root behavior in the real exchange rate) has prompted a shift of focus toward testing PPP using panel data.[3] There are a number of issues in tests for unit roots and cointegration in panels which include problems of interpretation and the fact that the spurious regression problem usually associated with $I(1)$ variables seems to be less of a problem in panels.

Consider a standard Cobb-Douglas production function, for countries $i = 1, 2, ..., N$ in which output Y_{it} is determined by labour, L_{it}, capital, K_{it}, and technology or total factor productivity, A_{it}:

$$\begin{aligned} Y_{it} &= A_{it} L_{it}^{\beta_i} K_{it}^{\alpha_i} \\ y_{it} &= a_{it} + \beta_i l_{it} + \alpha_i k_{it} \end{aligned}$$

using lower case letter for the logarithms of the variables. It is quite possible that a_{it} is a random walk $I(1)$ variable, in which case, there is no linear combination of log output, labour and capital that is $I(0)$ and each individual production function would constitute a spurious regression.

1.3.3 Cross-section dependence

Thirdly, there is the issue of cross-section dependence, CSD, correlation between residuals in different units, which is very common in panels.[4] If the CSD

[3] See Sarno and Taylor (2002) on this and other international parity conditions. The PPP example has become a standard example for new panel time-series estimators and we will also use it as an example.

[4] If N is of the same order of magnitude or larger than T, this cannot be dealt with by using the Zellner (1962) Seemingly Unrelated Regression Estimator (SURE), the usual time series

is large and not dealt with, one may get little improvement in efficiency from panel estimators relative to a single time-series. Sources of CSD include: spatial spillovers, e.g. between house prices in different regions, Holly et al (2010) Bailey, Holly & Pesaran (2016); interaction effects through trade or other networks; or common unobserved factors that influence all groups. CSD may allow one to estimate unobserved common factors, that cannot be estimated from a single time-series. This is a distinct advantage of panels.

Suppose in the Cobb-Douglas example above we assumed that TFP was determined as

$$a_{it} = \mu_i + \gamma_i f_t + u_{it}$$

where f_t is a measure of world best-practice technology which is unobserved and may be $I(1)$, though u_{it} is $I(0)$. This will also cause the estimates of the production function parameters to be biased if l_{it} or k_{it} are correlated with f_t, which is likely. The equation is then

$$y_{it} = \mu_i + \beta_i l_{it} + \alpha_i k_{it} + \gamma_i f_t + u_{it}. \tag{1}$$

A convenient way to allow for such a common factor is the Correlated Common Effects, CCE, estimator of Pesaran (2006), which just includes the time-means of the variables, such as $\overline{y}_t = \sum_i y_{it}/N$. Average (1) over i to get

$$\begin{aligned} \overline{y}_t &= \bar{\mu} + \bar{\beta}\, \bar{l}_t + \bar{\alpha}\bar{k}_t + \bar{\gamma} f_t + \bar{u}_t + \overline{w}_t, \\ \overline{w}_t &= \left[\sum_i (\beta_i - \bar{\beta}) l_{it} + \sum_i (\alpha_i - \bar{\alpha}) k_{it} \right] /N \end{aligned}$$

where $\overline{u}_t$ and (if the parameters are independent of the regressors) $\overline{w}_t$ will be close to zero. Then, with $\bar{u}_t = \overline{w} = 0$, the factor will be a function of the means

$$f_t = \bar{\gamma}^{-1} \left(\overline{y}_t - \bar{\mu} - \bar{\beta}\bar{\bar{l}}_t - \bar{\alpha}\bar{k}_t \right)$$

and as long as $\overline{\gamma} \neq 0$, the means allow for the factor in an equation of the form

$$y_{it} = \delta_i + \beta_i l_{it} + \alpha_i k_{it} + \delta_1 \overline{y}_t + \delta_2 \bar{l}_t + \delta_3 \bar{k}_t + u_{it}.$$

This will be a cointegrating regression if u_{it} is $I(0)$.

1.4 Approach

In these notes, we examine estimation and inference methods for panels where both N and T are large with an emphasis on applied issues. There are a number of issues that we shall not cover, including a range of standard panel estimators for the small T case. A range of technical issues about proofs are largely ignored: the results are summarised and the assumptions required are sketched but not

approach, because the estimated between unit covariance matrix is rank deficient and cannot be inverted. SURE will also not be consistent if the correlation is induced by a common ommitted factor correlated with the regressors.

fully described. Bayesian estimators are largely ignored, though in many ways a Bayesian approach is more natural in this context, since it is usual to treat parameters as random. Canova (2007) has a good discussion of Bayesian estimators for panels and panel VARs in macro data. Canova also has a discussion of frequency domain methods and time varying parameters, which we do not emphasise. There is also a literature on pseudo-panels, constructed from cross-section surveys where the same individuals are not followed over time, McKenzie (2004). The focus is almost entirely on linear models, reflecting the current emphasis in this literature, though applications of dynamic panel probit models and the like are becoming more common. It should also be noted that although the literature is growing very rapidly there are many gaps in the current knowledge of the properties of panel-time-series estimators. All the standard time-series issues, such as endogeneity, non-linearity and structural breaks, carry over to panels, but our focus will be primarily on heterogeneity, dynamics particularly unit roots and cointegration[5] and cross-section dependence The starting point is usually to estimate N time-series regressions, one for each unit, dealing with all the standard problems, then examine how to exploit any commonalities across units.

1.5 PPP data

To illustrate the techniques, we will often use the example of Purchasing Power Parity (PPP), which suggests that the real exchange rate:

$$q_{it} = s_{it} - p_{it} + p_{it}^*$$

should be stationary, that is, broadly constant in the long-run. The data are time series 1973Q1-1998Q4, $T = 104$ observations for 17 countries relative to the US[6] on log prices (p_{it}), log spot exchange rates defined as the domestic price of foreign currency or local currency per US dollar (s_{it}), long term interest rates(il_{it}) and short term interest rates (is_{it}). Prices and exchange rates are scaled (before taking logs) so that their value in 1995 is 1 in order to make them comparable. There is a single foreign price, p_t^*, that of the US.

When using this sort of data, units of measurement are crucial and it is very easy to make mistakes. Most exchange rates are quoted as domestic price of foreign currency; but some, for instance, Sterling are quoted as the foreign price of domestic currency. The Sterling data have been converted to local currency per dollar. Data on interest rates are usually presented as per cent per annum, here we have converted them into proportions (divided the percentage figures by 100). If we wanted interest rates at quarterly rates, we would also need to

[5] The textbook by Maddala and Kim (1998) provides an excellent discussion of some of the issues. Banerjee (1999) provides a more technical survey. On the more general issues on cointegration and unit roots see Pesaran and Smith (1998) and Oxley and McAleer (1999).

[6] The countries are (1) Australia, (2) Austria, (3) Belgium, (4) Canada, (5) Denmark, (6) France, (7) Germany, (8) Ireland, (9) Italy, (10) Japan, (11) Netherlands, (12) New Zealand, (13) Norway, (14) South Africa, (15) Sweden, (16) Switzerland, (17) UK, (18) US.

divide them by four. Getting the time dimension of rates wrong and getting the exchange rate the wrong way up are common mistakes.

It is crucial to conduct a preliminary analysis of the data before using it in modeling and inference. The fact that the data set may be large does not remove the need to do this. This involves looking at maxima, minima, means, standard deviations, and graphing the data. In this data there is a very high maximum value for the Swedish short-term interest rate just prior to the ERM crisis in September 1992. This will create an outlier for any panel estimation using short interest rates.

2 Static Models

2.1 Cross section or Time Series?

Panel estimators build on the notion that if there are similarities in the processes generating the data in the different units, we can increase the efficiency of parameter estimation by combining the data from different units. The nature of the similarities is a central question, for instance are the parameters in different units the same or are they randomly distributed about a common mean? Panel data also allows one to answer questions that cannot be answered with time-series data alone (such as. the effect of unobserved common factors) or with cross-section data alone (such as the patterns of adjustment to change). Panels also introduce flexibility in how one defines parameter heterogeneity, e.g. over time or over units.

Suppose we have a linear model and data $(y_{it}, \mathbf{x}_{it})$ $t = 1, ..., T, i = 1, ..., N$ where $\mathbf{x}_{it}$ is a $k \times 1$ vector of exogenous variables. When T is small, the data have to be interpreted as a set of cross-sections and when N is small, as a set of time series. But when both N and T are large there is a choice.

The data could be treated as a set of T cross-section regressions, allowing the parameters to differ freely over time:

$$y_{it} = \alpha_t + \boldsymbol{\beta}_t' \mathbf{x}_{it} + u_{it};\ u_{it} \sim IN(0, \sigma_t^2). \tag{2}$$

and independence is assumed both over time and over units so that $E(u_{it}u_{js}) = 0$ for $i \neq j$ and $t \neq s$. Model (2) has $T(k+2)$ parameters: $T(k+1)$ regression coefficients and T variances. We could then test for structural stability, whether the cross-section relationship was constant over time: e.g. that $\beta_t = \beta$.

Alternatively, the data could be treated as a set of N time-series regressions, allowing the parameters to differ freely over units:

$$y_{it} = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it};\ u_{it} \sim IN(0, \sigma_i^2). \tag{3}$$

Model (3) has $N(k+2)$ parameters: $N(k+1)$ regression coefficients and N variances. We could then test for homogeneity, whether the time-series relationships were the same for each group, e.g. that $\beta_i = \beta$. It is this second interpretation (3) that we shall use as the maintained model, given that we are looking at panel time-series issues.

Whether to treat the data as a set of time-series or a set of cross-section is a crucial modelling choice. This choice has many implications and is determined by data availability, the purpose of the exercise, the relevant theory and relative fit. The Feldstein-Horioka (1980), FH, puzzle is that in regressions of investment rates on savings rates there is a large positive coefficient, often not significantly different from one. For a closed economy the coefficient would be one, with perfect capital mobility FH argue it would be zero, so a large coefficient suggests a lot less openness than one might expect. A major issue in the large literature is whether it is a time-series, solvency type relationship, as in Coakley et al. (2004) or a cross-section financial frictions type relationship as argued by Bai & Yang (2010).

Assuming equality of slopes and variances in the set of time-series, (3) , and imposing the $(N-1)(k+1)$ restrictions $\beta_i = \beta$, $\sigma_i^2 = \sigma^2$, gives the most common panel model:

$$y_{it} = \alpha_i + \boldsymbol{\beta}' \mathbf{x}_{it} + e_{it};\ e_{it} \sim IN(0, \sigma^2) \tag{4}$$

known, among other names, as the one-way Fixed Effects (FE) model. This has $N + k$ regression parameters and a single variance.

It is quite possible that the cross-section variation is measuring a different phenomenon from the time-series variation. Traditionally, it was thought that the cross-section picked up the long-run (or permanent) effects and the time-series the short-run (or transitory) effects (Baltagi and Griffin, 1984; Pesaran and Smith, 1995, Baltagi 2008, section 10.6). These effects may have different signs in cross-section and time-series. For instance, it is well known that in cross-sections the share of investment in output is positively correlated with growth, but using a panel VAR Attanasiou et al. (2000) show that investment negatively Granger causes growth. However, after taking account of a range of issues, Bond et al. (2010) find a positive relationship between investment and growth, though no relationship for the OECD countries. Brush (2007) shows that while crime is positively associated with income inequality in cross-section over US counties it is negatively associated with inequality in the time-series (first difference) regressions. Mundlak (2005) argues that differences between the various regressions can be informative. Where possible look at both cross-section and time-series relationships and where they differ try to interpret the difference in terms of theory.

2.2 Models

The one-way FE model introduced above, equation (4), has a large number of names, because it was developed independently in a number of separate areas. It is also called: (i) the analysis of covariance estimator; (ii) the Least Squares Dummy Variables (LSDV) estimator, because it can be estimated by including a dummy variable for each unit as well as by demeaning as in (6) below; and (iii) the within estimator. It is called the within estimator because the variance of y_{it} can be decomposed into two orthogonal components, *within-group* and *between-group:*

$$\sum_i \sum_t (y_{it} - \overline{y})^2 = \sum_i \sum_t (y_{it} - \overline{y}_i)^2 + T \sum_i (\overline{y}_i - \overline{y})^2 . \quad (5)$$

The one-way FE estimator, (4), just uses the within variation and can be estimated as

$$(y_{it} - \overline{y_i}) = \boldsymbol{\beta}'_W (\mathbf{x}_{it} - \overline{\mathbf{x}}_i) + u_{it}, \quad (6)$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of regressors. Notice that the degrees of freedom used in calculating variance and standard errors are $NT - K - N$, not $NT - K$ as would be used in a program running this regression.

The between (group) regression is the cross-section OLS regression on time averaged data for each group:

$$\overline{y}_i = \alpha + \boldsymbol{\beta}'_B \overline{\mathbf{x}}_i + u_i; \; u_i \sim iidN(0, \sigma^2).$$

The Pooled OLS (POLS) regression

$$y_{it} = \alpha + \boldsymbol{\beta}'_P \mathbf{x}_{it} + v_{it}; \; v_{it} \sim iidN(0, \sigma^2)$$

gives equal weight to the within and between variance. As noted above the various estimates of $\boldsymbol{\beta}$ may be measuring different things.

Equation (4) is called Fixed Effect (FE) to distinguish it from the Random Effect (RE) model discussed below. It is called 'one-way' to distinguish it from the two-way FE model

$$y_{it} = \alpha_i + \alpha_t + \boldsymbol{\beta}' \mathbf{x}_{it} + e_{it}; \; e_{it} \sim iidN(0, \sigma^2) \quad (7)$$

which allows for a completely flexible time trend, a common factor which influences each unit by the same amount. To avoid the dummy variable trap, this is usually parameterised as

$$y_{it} = \alpha + \eta_i + \eta_t + \boldsymbol{\beta}' \mathbf{x}_{it} + e_{it}; \; e_{it} \sim iidN(0, \sigma^2)$$

with the restrictions

$$\sum_{i=1}^{N} \eta_i = \sum_{t=1}^{T} \eta_t = 0.$$

It can be implemented without dummy variables by the regression

$$(y_{it} - \overline{y}_i - \overline{y}_t + \overline{y}) = \alpha + \beta(x_{it} - \overline{x}_i - \overline{x}_t + \overline{x}) + u_{it}.$$

We return to such time effects below in discussing cross section dependence.

Notice that in this case, corresponding to the two way decomposition in (5), the variance of y can be decomposed into three orthogonal components: within group and period; between group; and between period:

$$\sum_i \sum_t (y_{it} - \overline{y})^2 = \sum_i \sum_t (y_{it} - \overline{y}_i - \overline{y}_t + \overline{y})^2 + T \sum_i (\overline{y}_i - \overline{y})^2 + N \sum_t (\overline{y}_t - \overline{y})^2.$$

Similarly to the between group regression, there is also a between period regression

$$\overline{y}_t = \alpha + \boldsymbol{\beta}'\overline{\mathbf{x}}_t + u_t$$

which is just an aggregate time-series regression.

These estimators generalise naturally to more categories. Suppose that we had countries, $i = 1, 2, ..., N$ and industries, $j = 1, 2, ..., J$, then we could model the three-way data as

$$y_{ijt} = \alpha + \eta_i + \eta_j + \eta_t + \boldsymbol{\beta}'\mathbf{x}_{ijt} + e_{ijt}.$$

This involves assuming that the composite intercept for $i =$France, $j =$agriculture, $t =$1990 is the sum of a French effect (which is the same for every French industry in every year), an agriculture effect (which is the same for agriculture in every country and year) and a 1990 effect (which is the same for all countries and industries in 1990). This involves estimating $N + J + T - 1$ intercepts. A less restrictive models with more intercept heterogeneity is

$$y_{ijt} = \alpha + \eta_{ij} + \eta_t + \boldsymbol{\beta}'\mathbf{x}_{ijt} + e_{ijt}$$

which allows for NJ country-industry intercepts, a separate one for each industry in each country rather than $N + J$. Although this may look more complicated, in fact it is really only a two-way FE model

$$y_{kt} = \alpha + \eta_k + \eta_t + \boldsymbol{\beta}'\mathbf{x}_{kt} + e_{kt}$$

where $k = 1, 2, ..., NJ$. It allows for more heterogeneity in the sense that the intercept for French agriculture does not have to be the sum of a 'French' and 'agriculture' factor. Balazsi et al. (2015) discuss the estimation of multi-dimensional fixed effect panel data models in more detail.

There is also the interactive fixed effect model of Bai (2009) that will be considered in chapter 8. Here rather than being additive, $\alpha_i + \alpha_t$, the fixed effect is is multiplicative, $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$ where $\boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$ are $r \times 1$ vectors. Thus the model is

$$y_{it} = \alpha + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \boldsymbol{\beta}'\mathbf{x}_{it} + e_{it};\ e_{it} \sim iidN(0, \sigma^2).$$

2.3 Estimator Properties

Consider the properties of the estimators of the 1-way FE model, (4). Under the assumptions of homogeneous slopes, independent normal errors with constant variance and strictly exogenous x_{it} then $\widehat{\alpha}_i$ and $\widehat{\boldsymbol{\beta}}$ are best (minimum variance) linear unbiased estimators (BLUE). $\widehat{\boldsymbol{\beta}}$ is consistent for either large T or large N, whereas $\widehat{\alpha}_i$ is consistent for large T, but it is not consistent for large N. This is because every time we increase N we get another α_i to estimate, and we only have T observations to estimate each α_i on. This incidental parameter problem can occur both in the N and T dimensions. The same problem occurs with α_t, each extra time period adds an extra parameter. If we assume linear trends, with coefficients that differ over units:

$$y_{it} = \alpha_i + \delta_i t + \boldsymbol{\beta}'\mathbf{x}_{it} + e_{it};\ e_{it} \sim IN(0, \sigma^2),$$

there are said to be incidental trends. Moon, Perron and Phillips (2015) survey the issues associated with incidental parameters in dynamic panels.

In the fixed effect models it is common to assume no heteroskedasticity, which is implausible, so one would either use GLS or robust standard errors. Feasible GLS-FE models, weight in the second stage by first stage estimates of the variance for each group estimated for instance from the residuals of a standard fixed effect estimator. Stock and Watson (2006) discuss estimation of heteroskedasticity robust standard errors for fixed effect regressions.

If there are unit fixed effects, α_i, they will pick up the effect of any time-invariant regressors, z_i, similarly the α_t will pick up the effects of any group invariant regressors, z_t. Pesaran and Zhou (2014) propose the Fixed Effects Filtered (FEF) and Fixed Effects Filtered instrumental variable (FEF-IV) estimators for estimation and inference in the case of time-invariant effects in static panel data models when N is large and T is fixed. Essentially it involves regressing the $\widehat{\alpha}_i$ on the z_i and obtaining the correct standard errors. It is shown that the FEF and FEF-IV estimators are $\sqrt{N}$-consistent, and asymptotically normally distributed. The FEF estimator is compared to other estimators including the estimator proposed by Hausman and Taylor (1981), when one of the time-invariant regressors is correlated with the fixed effects. Both FEF and FEF-IV estimators are shown to be robust to error variance heteroskedasticity and residual serial correlation.

The bias due to the incidental parameter problem can be even more severe in non-linear models and Fernandez-Val and Weidner (2015) develop bias corrections for non-linear models (e.g. logit, probit and Poisson models) with both individual and time effects when both N and T are large.

2.4 The Random Effect (Intercept) Model

The one-way FE model involves estimating N separate intercepts α_i and if N is large (e.g. in the thousands) this involves a lot of parameters and a large loss in efficiency. An alternative is the Random Effects (RE) model which treats the $\eta_i = \alpha_i - \alpha$ not as fixed parameters to be estimated, but as random variables with $E(\eta_i) = 0$ or $E(\alpha_i) = \alpha$, and $E(\eta_i^2) = V(\alpha_i) = \sigma_\eta^2$. The *random intercept* assumption implies: (i) the α_i are realizations from a probability distribution with a fixed number of parameters, (ii) they are distributed independently of u_{it}, and (iii) they are independent of $\mathbf{x}_{it}$.

With these assumptions we only have to estimate 2 parameters $\alpha \equiv E(\alpha_i)$ and $\sigma_\eta^2 \equiv V(\alpha_i)$ instead of the N intercepts α_i, $i = 1, ..., N$. The model is then:

$$y_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + (u_{it} + \eta_i)$$

where $u_{it} \sim iidN(0, \sigma^2)$. The composite error term, $v_{it} = (u_{it} + \eta_i)$ has the properties $E(v_{it}) = 0; E(v_{it}^2) = \sigma^2 + \sigma_\eta^2; E(v_{it}v_{jt-i}) = 0$, $i \neq j$ and $E(v_{it}v_{it-s}) =$

σ_η^2, $s \neq 0$. Thus this error structure introduces a very specific form of serial correlation. Estimation is by Generalised Least Squares (GLS), equivalent to OLS on the transformed equation:

$$(y_{it} - \theta \overline{y}_i) = \alpha + \beta'(x_{it} - \theta \overline{x}_i) + u_{it},$$

where

$$\theta = 1 - \sqrt{\frac{\sigma^2}{T\sigma_\eta^2 + \sigma^2}}$$

Notice that the FE estimator corresponds to the case where $\theta = 1$ or when $T\sigma_\eta^2$ is infinite (if the α_i are fixed their $E(\alpha_i)$ and $V(\alpha_i)$ are not defined which is equivalent to being infinite) whereas the POLS estimator corresponds to $\theta = 0$ or $T\sigma_\eta^2 = 0$. So the RE estimator lies between the FE and POLS estimator. Feasible GLS requires an estimate of θ; there are a number of possible consistent estimators for σ_η^2 and σ^2 based on the FE or individual OLS first-stage regressions. One can also have random time effects.

When T is large the FE and RE estimators become identical, since $\theta = 1$ as the denominator in the second term gets large.

2.5 Testing between models

The usual approach to testing between POLS and RE uses a standard LM test for heteroskedasticity, since the variances will differ between groups if the RE model is appropriate. Of course, heteroskedasticity may be caused by other things than random effects.

The usual approach to testing between RE and FE is a *Hausman (1978) test.* Let $\widehat{\beta}^{FE}$ and $V(\widehat{\beta}^{FE})$ denote the FE estimator and its covariance matrix and likewise for the RE estimator, $\widehat{\beta}^{RE}$ and $V(\widehat{\beta}^{RE})$. If the RE model is correct, $\widehat{\beta}^{RE}$ is consistent and efficient so $V(\widehat{\beta}^{FE}) > V(\widehat{\beta}^{RE})$. Let $q = \widehat{\beta}^{FE} - \widehat{\beta}^{RE}$. Under H_0 it follows that $cov(\widehat{\beta}^{FE}, \widehat{\beta}^{RE}) = 0$. This is because if $\widehat{\beta}^{RE}$ is efficient its variance cannot be reduced and, if it was correlated with $\widehat{\beta}^{FE}$, that could be used to reduce its variance. The variance of the difference is then

$$V(\hat{q}) = V(\widehat{\beta}^{F}) - V(\widehat{\beta}^{R}).$$

If the individual effects are not random but correlated with the $\mathbf{x}_{it}$ then the RE estimates are inconsistent, but the FE estimates are still consistent, since the FE model admits any degree of correlation between α_i and $\mathbf{x}_{it}$. The Hausman test statistic is:

$$\tau = q'[V(\hat{q})]^{-1} q \sim \chi^2(k) \tag{8}$$

One rejects the null hypothesis that the RE model is appropriate against the alternative that the FE model is appropriate for large τ, i.e. when the difference between the estimates is large.

A Hausman type statistic can be used in a range of settings to compare two estimators where one, the more efficient, is consistent only under H_0 and the other is inefficient under H_0 but consistent under both H_0 and H_1. If the estimates are not very different, one concludes that the H_0 is true and uses the more efficient estimator. If the estimates are very different, one rejects H_0 and uses the estimator which is consistent under H_1. Notice, that the Hausman test has an implicit null hypothesis that is different from the null hypothesis of standard parameter restrictions tested by Wald, LM and LR. This can be seen in its distribution. The degrees of freedom of the other tests is the number of restrictions, the degrees of freedom of a Hausman test is the number of parameters estimated. In some applications, the estimate of $V(\hat{q})$ may not be positive definite.

Baltagi, Bresson and Pirotte (2003) suggest pre-test estimators based on Hausman test between the FE and the RE and Hausman-Taylor (1981) model. In the latter model, there are both time-varying and time-invariant regressors some of which are independent of the random effects η_i and others are not. Subject to identifying conditions it can be estimated using the instrumental variables (IV) approach.

Pesaran and Zhou (2015) point out that even if some of the intercepts differ and are correlated with the regressors, the pooled estimator may be consistent (large N, T fixed) and more efficient than the fixed effect estimator if the differences are not too pervasive in a way that they characterise.

2.6 Random Coefficients

Whereas the RE model treats the intercepts, α_i, as random and the slope, β, as homogeneous and fixed, we could treat all the coefficients as random. The *random coefficient* assumption means that the $(\alpha_i, \boldsymbol{\beta}_i)'$ are realizations from a probability distribution with a fixed number of parameters, distributed independently of the regressors and disturbances.[7] We could also assume that some coefficients (e.g. α_i) were fixed and others (e.g. $\boldsymbol{\beta}_i$) random.[8] Notice we use the term random effect, when only the intercept is random, and random coefficient, when other parameters are also random.

To look at the heterogeneous models for simplicity we will allow $\boldsymbol{\beta}_i$ to contains both slopes and intercepts and write (3) as:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{u}_i \tag{9}$$

where $\mathbf{y}_i$ is a $T \times 1$ vector, and $\mathbf{X}_i$ is a $T \times (k+1)$ vector and $\boldsymbol{\beta}_i =$ is a $(k+1) \times 1$ vector. Then random parameters are assumed, namely, that $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i$ where $E(\boldsymbol{\eta}_i) = 0$, $E(\boldsymbol{\eta}_i \boldsymbol{\eta}_j') = \boldsymbol{\Omega}$, $i = j$, $E(\boldsymbol{\eta}_i \boldsymbol{\eta}_j') = 0$ otherwise, and that the $\boldsymbol{\eta}_i$ are independent of $\mathbf{X}_i$. The crucial assumption is the independence of the randomly varying parameters from the regressors, but here the assumption is extended

[7] Hsiao and Pesaran (2008) provide a survey.

[8] Hsiao (2003) discusses mixed fixed-random models of this sort.

from intercepts to slopes.[9]

2.6.1 Mean Group

There are a large number of estimators for $\boldsymbol{\beta} \equiv \mathbf{E}(\boldsymbol{\beta}_i)$, the expected value of the random coefficients. The simplest is to compute the OLS estimates for each group:

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{y}_i$$

and then construct the average $\overline{\boldsymbol{\beta}} = \sum_i \widehat{\boldsymbol{\beta}}_i/N$, estimating the $(k+1)\times(k+1)$ covariance matrix Ω by

$$\widehat{\boldsymbol{\Omega}} = \sum_i (\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})'/(N-1) \tag{10}$$

Pesaran and Smith (1995) call $\overline{\boldsymbol{\beta}}$ the Mean Group (MG) estimator. Its covariance matrix is

$$V(\overline{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\Omega}}/N = \sum_i (\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})'/N(N-1).$$

This provides a non-parameteric estimator of the standard errors of the elements of $\overline{\boldsymbol{\beta}}$, which does not depend on estimation of $V(\widehat{\boldsymbol{\beta}}_i)$. This is an advantage since the estimate of $V(\widehat{\boldsymbol{\beta}}_i)$ may not be robust to heteroskedasticity or autocorrelation.

As usual, the mean is sensitive to outliers and there are robust versions of the Mean Group estimator, which trim or winsorize the outliers to reduce their effect.

2.6.2 The Swamy Random Coefficient Model (RCM)

Swamy (1970) suggests a feasible GLS estimator, which is equivalent to using a weighted average of the individual OLS estimates $\widehat{\beta}_i$ instead of the MG unweighted average. Using the residuals and the unbiased estimate of the variance

$$\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i; \quad s_i^2 = \widehat{\mathbf{u}}_i'\widehat{\mathbf{u}}_i/(T-k-1),$$

respectively, the estimated covariance of $\widehat{\boldsymbol{\beta}}_i$ is

$$V\left(\widehat{\boldsymbol{\beta}}_i\right) = s_i^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}.$$

Swamy suggests estimating $\boldsymbol{\Omega}$ by

$$\widetilde{\boldsymbol{\Omega}} = \sum_i (\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})'/N - \sum_i V\left(\widehat{\boldsymbol{\beta}}_i\right)/N.$$

[9] See Pesaran, Haque and Sharma (1999) for a case where the slope independence assumption fails, producing misleading inferences in pooled regressions.

If this estimator is not positive definite (which it rarely is in small samples), the last term is set to zero and $\widehat{\mathbf{\Omega}}$ from (10) is used. Notice that as $T \to \infty$, $V\left(\widehat{\boldsymbol{\beta}}_i\right) \to 0$, so the second term vanishes for large T.

The Swamy estimator of the mean is

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \sum_i D_i \widehat{\boldsymbol{\beta}}_i, \\
D_i &= \left\{ \sum_i \left[\widetilde{\mathbf{\Omega}} + V\left(\widehat{\boldsymbol{\beta}}_i\right) \right]^{-1} \right\}^{-1} \left[\widetilde{\mathbf{\Omega}} + V\left(\widehat{\boldsymbol{\beta}}_i\right) \right]^{-1} \\
V\left(\widetilde{\boldsymbol{\beta}}\right) &= \left\{ \sum_i \left[\widetilde{\mathbf{\Omega}} + V\left(\widehat{\delta}_{\boldsymbol{\beta} i}\right) \right]^{-1} \right\}^{-1}.
\end{aligned}$$

The predictions of the individual coefficients can be improved by shrinking the OLS estimates towards the overall estimate:

$$\begin{aligned}
\mathring{\boldsymbol{\beta}}_i &= \mathbf{Q} \widehat{\boldsymbol{\beta}}_i + (I - \mathbf{Q}) \widetilde{\boldsymbol{\beta}} \qquad (11) \\
\mathbf{Q} &= \left[\widetilde{\mathbf{\Omega}}^{-1} + V\left(\widehat{\boldsymbol{\beta}}_i\right)^{-1} \right]^{-1} \widetilde{\mathbf{\Omega}}^{-1}
\end{aligned}$$

The Swamy RCM can be interpreted either as a GLS estimator or an empirical Bayes estimator, since it has the same form. Hsiao Pesaran and Tahmiscioglu (1999) review a variety of Bayes and empirical Bayes estimators of this sort and show that the MG estimator is asymptotically normal for large N and T as long as $\sqrt{N}/T \to 0$ as both N and $T \to \infty$, but is unlikely to perform well when N or T are small. In particular, the MG estimator is very sensitive to outliers which are a common feature of the group specific estimates in many applications. In principle, the weighting which the Swamy estimator applies, should reduce this problem; in practice, it may not (see Boyd and Smith, 2002). As $T \to \infty$ the difference between the Swamy and MG estimators goes to zero.

2.6.3 Relationship between the estimators

All the estimators can be regarded as different averages of the group specific estimators. Consider a simple case

$$y_{it} = \alpha_i + \beta_i x_{it} + u_{it}.$$

Define $\widetilde{y}_{it} = y_{it} - \overline{y}_i$ then the individual estimator is

$$\widehat{\beta}_i = \frac{\sum_t \widetilde{x}_{it} \widetilde{y}_{it}}{\sum_t \widetilde{x}_{it}^2}.$$

The MG and Swamy estimators are clearly averages of this, but so is the fixed effect estimator

$$\widehat{\beta}^{FE} = \frac{\sum_i \sum_t \widetilde{x}_{it} \widetilde{y}_{it}}{\sum_i \sum_t \widetilde{x}_{it}^2} = \sum_i \left(\frac{\sum_t \widetilde{x}_{it}^2}{\sum_i \sum_t \widetilde{x}_{it}^2} \right) \left(\frac{\sum_t \widetilde{x}_{it} \widetilde{y}_{it}}{\sum_t \widetilde{x}_{it}^2} \right) = \sum_i W_i \widehat{\beta}_i,$$

which weights by the proportion of the total variance in x_{it} from each group.

2.6.4 Systematic Heterogeneity

In some cases, the heterogeneity in the parameters may stem from observed individual-specific characteristics which do not vary over time. For instance, Smith and Zoega (2008) make estimates of the speed of adjustment of unemployment to shocks a function of labour market institutios of a country which influence labour market flexibility. A general model which nests a number of interesting cases is:

$$y_{it} = \alpha_i^{'}\mathbf{w}_{it} + \boldsymbol{\beta}_i^{\prime}\mathbf{x}_{it} + \varepsilon_{it},\ i = 1, ..., N, t = 1, ..., T \tag{12}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\Gamma}\mathbf{z}_i + \eta_i, \tag{13}$$

where y_{it} is a scalar dependent variable; $\boldsymbol{\alpha}_i$ and $\mathbf{w}_{it}$ are $h \times 1$ vectors of fixed unknown parameters and regressors respectively, $\boldsymbol{\beta}_i$ and $\mathbf{x}_{it}$ are $k \times 1$ vectors of random unknown parameters and regressors respectively; Γ is a $k \times s$ matrix of unknown parameters and $\mathbf{z}_i$ is a $s \times 1$ vector of observed variables that do not vary over time (e.g. the group means of the $\mathbf{x}_{it}$) with the first element of $\mathbf{z}_i$ being one to allow for an intercept. The unobserved disturbances are the scalar ε_{it} and the $k \times 1$ vector $\boldsymbol{\eta}_i$. Suppose that we are interested in α_i and $\boldsymbol{\Gamma}$. We shall assume that:

(i) The regressors $\mathbf{x}_{it}$ and $\mathbf{z}_i$ are independent of the ε_{it} and η_i;

(ii) The ε_{it} are distributed as $N(0, \sigma_i^2)$ and $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for all $i \neq j$, $t \neq s$

(iii) The $\boldsymbol{\eta}_i$ are normally distributed with $E(\boldsymbol{\eta}_i) = 0$, $E(\boldsymbol{\eta}_i\boldsymbol{\eta}_j^{\prime}) = \boldsymbol{\Omega}$, for $i = j$, and $E(\boldsymbol{\eta}_i\boldsymbol{\eta}_j^{'}) = 0$ otherwise.

If $\boldsymbol{\Gamma} = \mathbf{0}$ and $\alpha_i = 0$ we get the Swamy RCM. If $\boldsymbol{\Omega} = 0$ and $\boldsymbol{\Gamma} \neq \mathbf{0}$ we get systematic deterministic variation in the slopes. There are a variety of other special cases.[10]

2.6.5 Implications of Random parameters

We can treat the parameters as fixed or random, homogeneous or heterogeneous, and we can allow for parameter variation over individuals or over time periods or both. The panel structure thus gives us a large amount of choice; which choice will be appropriate depends on the context. Zellner (1969) shows that if the regressors are strictly exogenous and the parameters are randomly distributed independently of the regressors, all the estimators considered above will provide unbiased estimates of the expected values of the coefficients. To see why this is so, consider the heterogeneous model with a single regressor

$$y_{it} = \alpha_i + \beta_i x_{it} + u_{it}$$

where the α_i are regarded as fixed and the β_i are random parameters distributed independently of the regressor. The Swamy estimator discussed above is the efficient GLS estimator but the FE estimator gives an unbiased estimator of

[10] Hsiao (2003) discusses a variety of these mixed fixed and random models.

$E\left(\boldsymbol{\beta}_i\right) = \boldsymbol{\beta}$. As above the FE estimator of β is

$$\begin{aligned}
\widehat{\beta}^{FE} &= \frac{\sum_i \sum_t \widetilde{x}_{it}\widetilde{y}_{it}}{\sum_i \sum_t \widetilde{x}_{it}^2} \\
&= \frac{\sum_i \sum_t \widetilde{x}_{it}(\beta_i \widetilde{x}_{it} + \tilde{u}_{it})}{\sum_i \sum_t \widetilde{x}_{it}^2} \\
&= \frac{\sum_i \sum_t \widetilde{x}_{it}(\beta \widetilde{x}_{it} + \eta_i \widetilde{x}_{it} + \tilde{u}_{it})}{\sum_i \sum_t \widetilde{x}_{it}^2} \\
&= \beta + \frac{\sum_i \sum_t \widetilde{x}_{it}^2 \eta_i}{\sum_i \sum_t \widetilde{x}_{it}^2} + \frac{\sum_i \sum_t \widetilde{x}_{it}\tilde{u}_{it}}{\sum_i \sum_t \widetilde{x}_{it}^2} \\
E\left(\widehat{\beta}^{FE}\right) &= \beta
\end{aligned}$$

The expectation of the third term in the penultimate line is zero since x_{it} is strictly exogenous, $E(u_{it}\widetilde{x}_{is}) = 0 \ \forall t, s$. The expectation of the second term is zero since $\eta_i = \beta_i - \beta$ is independent of $\widetilde{x}_{it}$. Similar arguments can be applied to the other estimators, e.g. POLS if the α_i were also assumed to be random. However, this does not carry over to models with lagged dependent variables, where the regressors are not strictly exogenous, as we shall see.

2.7 Heterogeneity Testing and Model Selection

We have considered a range of models that differ in terms of the degree of heterogeneity and it is natural to consider testing between the models in determining whether to pool or not to pool. Baltagi, Bresson & Pirotte (2008) have a survey of the issues involved in the choice to pool. Testing is often not the right approach and it pays to be careful about what is being done. A testing exercise involves

- defining a null hypothesis, H_0 and an alternative H_1;
- deriving a test statistic τ, which is pivotal, i.e. it does not depend on any unknown parameters and which has a known distribution when the null hypothesis is true;
- choosing the *size* of the test α (the probability of rejecting a true H_0, Type I error);
- calculating the critical values corresponding to α so that the acceptance region is defined.
- rejecting H_0 when the test statistic lies outside the acceptance region
- evaluating the power of the test (the probability of rejecting the null when it is false, equal to one minus the probability of Type II error). This will generally depend on unkown parameters.

There are three types of classical test: (a) Lagrange Multiplier, LM, or efficient score which uses the restricted estimates (imposing the null hypothesis); (b) Wald, which use the unrestricted (not imposing the null) and (c) Likelihood Ratio, LR, which use both the restricted and unrestricted estimates. Asymptotically they are usually equivalent, but they may give different conclusions in small samples. Each is distributed χ^2 with degrees of freedom given by the number of restrictions, i.e. the number of hypotheses tested. Wald tests are not invariant to the way non-linear restrictions are written.

In the context of (3) the interesting null hypotheses are likely to be about the equality of parameters, α_i, $\boldsymbol{\beta}_i$ or σ_i^2. Given our normality assumption, and the assumed absence of cross-section dependence, OLS on each model gives the Maximum Likelihood (ML) estimator. Therefore it is straightforward to conduct Likelihood Ratio (LR) tests between the various models. H_0 is parameter homogeneity, e.g. $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, and H_1 is parameter heterogeneity $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}$. The relevant asymptotic properties of these tests are derived for N fixed as T goes to infinity. It should be noted that these LR tests are not be appropriate when N is large
razaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
to T, since the number of restrictions tested will get large as well. Pesaran and Yamagata (2007) propose a test which seems to have good properties when N is large relative to T, which we discuss below. We will also discuss a quite different sort of procedures based on Hausman (1978) tests.

2.7.1 LR Tests

Let the OLS estimators for each time series equation, (3) , be denoted by $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ and the residuals by

$$\widehat{u}_{it} = y_{it} - \widehat{\alpha}_i - \widehat{\boldsymbol{\beta}}_i' \mathbf{x}_{it}$$

The ML estimate of the variance for each unit is

$$\widehat{\sigma}_i^2 = \sum_{t=1}^{T} \widehat{u}_{it}^2 / T$$

which is biased, $E(\widehat{\sigma}_i^2) \neq \sigma_i^2$ but is consistent for large T, that is, $\widehat{\sigma}_i^2 \to \sigma_i^2$ as $T \to \infty$. The maximised log-likelihood (MLL) of model (3) for unit i is

$$MLL_i = -\frac{T}{2} (\ln 2\pi + 1) - \frac{T}{2} \ln \widehat{\sigma}_i^2.$$

and so for the whole sample we have

$$MLL_A = \sum_{i=1}^{N} MLL_i = -\frac{NT}{2} (\ln 2\pi + 1) - \frac{T}{2} \sum_{i=1}^{N} \ln \widehat{\sigma}_i^2.$$

If the variance is assumed homogeneous, $\sigma_i^2 = \sigma^2$, its ML estimate is

$$\widehat{\sigma}^2 = \sum_{i=1}^{N} \sum_{i=1}^{T} \widehat{u}_{it}^2 / NT$$

which is consistent for large T fixed N, but not for large N fixed T, since

$$E\left(\widehat{\sigma}^2\right) = \frac{NT - N(k+1)}{NT}\sigma^2 = (1 - \frac{k+1}{T})\sigma^2$$

so the bias reduces with T but not with N.

The MLL of the homogeneous variance model is

$$MLL_B = -\frac{NT}{2}(\ln 2\pi + 1) - \frac{NT}{2}\ln \widehat{\sigma}^2.$$

and so the hypothesis of common variances $H_0^1 : \sigma_i^2 = \sigma^2$ can be tested using the test statistic

$$\tau_1 = 2(MLL_A - MLL_B) \sim \chi^2(N-1).$$

Let $\overline{y}_i$ and $\overline{x}_i$ be the group means

$$\overline{y}_i = \sum_{t=1}^{T} y_{it}/T, \quad \overline{\mathbf{x}}_i = \sum_{t=1}^{T} \mathbf{x}_{it}/T.$$

Under both variance and slope equality, $\beta_i = \beta$ and $\sigma_i^2 = \sigma^2$, the ML estimator is the 1-way FE estimator[11], that is, OLS on pooled deviations from group means

$$(y_{it} - \overline{y}_i) = \boldsymbol{\beta}'(\mathbf{x}_{it} - \overline{\mathbf{x}}_i) + e_{it} \tag{14}$$

with residuals, ML estimate of the common variance and MLL of

$$\begin{aligned} \widehat{e}_{it} &= (y_{it} - \overline{y}_i) - \widehat{\boldsymbol{\beta}}'(\mathbf{x}_{it} - \overline{\mathbf{x}}_i) \\ s^2 &= \sum_i \sum_t \widehat{e}_{it}^2/NT \\ MLL_C &= -\frac{NT}{2}(\ln 2\pi + 1) - \frac{NT}{2}\ln s^2. \end{aligned}$$

The hypothesis of equality of slopes, $H_0^2 : \boldsymbol{\beta}_i = \boldsymbol{\beta}$, conditional on equality of variances ($\sigma_i^2 = \sigma^2$) can be tested using the test statistic

$$\tau_2 = 2(MLL_B - MLL_C) \sim \chi^2(N-1)k.$$

This is an asymptotic version of the familiar Chow structural stability test: a test for equality of coefficients from two samples, $N = 2$, conditional on variance equality.

The joint test for equality of both slopes and variances, $H_0^3 : \boldsymbol{\beta}_i = \boldsymbol{\beta}$ and $\sigma_i^2 = \sigma^2$, can be tested using the test statistic

$$\tau_3 = 2(MLL_A - MLL_C) \sim \chi^2(N-1)(k+1)$$

and $\tau_3 = \tau_1 + \tau_2$. Notice that although the test statistic for the joint hypothesis equals the sum of the test statistics for the two individual hypotheses; conducting the two separate individual tests is different from the joint test. The probability of not rejecting the true joint null H_0^3 is $(1-\alpha)^2$ if you conduct two individual tests and $(1-\alpha)$ if you conduct the joint test.

[11] We could also use an FE-GLS estimator which has common slopes but variances which differ across groups, the variances for each group estimated from a first step FE.

2.7.2 Wald Tests

Consider the linear regression model for each group, where X_i is a $T \times k$ matrix,

$$y_i = X_i\beta_i + u_i.$$

The homogeneity hypothesis that $H_0 : \beta_i = \beta$ can be tested by a Wald test[12] suggested by Swamy

$$S = \sum_i \left(\widehat{\beta}_i - \beta_*\right)' V\left(\widehat{\beta}_i\right)^{-1}\left(\widehat{\beta}_i - \beta_*\right) \sim \chi^2\left((k)(N-1)\right)$$

where

$$\begin{aligned} \beta_* &= \left[\sum_i V\left(\widehat{\beta}_i\right)^{-1}\right]^{-1} V\left(\widehat{\beta}_i\right)^{-1}\widehat{\beta}_i; \\ V\left(\widehat{\beta}_i\right) &= s_i^2(X_i'X_i)^{-1} \\ s_i^2 &= \widehat{u}_i'\widehat{u}_i/(T-k) \end{aligned}$$

This is equivalent to a weighted fixed effect estimator. Even if the slope coefficients are the same this test will reject if the intercepts differ. If one wanted to conduct the test allowing the intercepts to differ, $H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}$, the Swamy estimator could be applied to data in deviations from group means, i.e. treating the intercepts as fixed. This will not give the same estimates of β as the original Swamy since the covariance matrices are different.

The Wald test like the LR test may not be appropriate when N is large relative to T. Pesaran and Yamagata (2008) propose a variant of this test which has good properties when N is large. The problem with the Swamy test arises because as N increases you are increasing the number of variances that you need to estimate. Rather than using $s_i^2 = \widehat{u}_i'\widehat{u}_i/(T-k)$, calculated using $\widehat{u}_i = y_i - X_i\widehat{\beta}_i$, the residuals from the individual regressions, the P&Y test uses the residuals from the fixed effect estimator, using a homogeneous coefficient, $\widetilde{u}_i = y_i - X_i\widetilde{\beta}$ in calculating an estimate of $\widetilde{\sigma}_i^2 = \widetilde{u}_i'\widetilde{u}_i/(T-1)$, which is used both in calculating the test statistic and the weighted fixed effect estimator. Although this may appear a slight adjustment, it makes a large difference when N is large relative to T, by reducing the number of parameters estimated.

Neither of these tests allow for cross-section dependence, which is usually important in economic panels. Ando and Bai (2014), review a number of tests and propose a modified version of Swamy's test that allows the error term to have a multifactor error structure of the sort discussed in chapters 7 and 8 below.

2.7.3 Issues with testing

In many cases, testing using some arbitrary nominal size or probability of Type I error, e.g. $\alpha = 5\%$, may not be appropriate and the test statistics are better

[12] Note this is a standard test of parameter restrictions unlike the Hausman test used to compare the FE and the RE.

regarded as general indications of the evidence. There are a number of issues associated with testing.

1. What is the appropriate null? E.g unit root or stationarity?

2. Decisions about which model is appropriate must depend on the purpose of the model reflected in a loss function; the size of test should reflect this. For forecasting, a size of 1% or less is often appropriate so that only very significant variables are included. Doing multiple (m) individual tests at the same significance level α can lead to a large size for the joint hypothesis, $1-(1-\alpha)^m$, as noted above and so the latter may be too often rejected. The multiple testing problem is an area of considerable current research in a variety of contexts, e.g. Bailey, Pesaran and Smith (2015).

3. Pre-testing is often used: setting a coefficient to zero if not significant, equal to the estimate if significant. This produces a biased estimator and an estimate which is a discontinuous function of the data: 1.95999 disaster, 1.96 bliss. In pre-test situations a larger size is usually appropriate, e.g. 25%.

4. It is very easy to confuse statistical significance with substantive economic significance. An effect may be large in economic terms but not statistically significant or very small in economic terms and statistically very significant. The size of the effect matters as well as its signficance.

5. In some cases one may wish to compare non-nested models, such as the set of cross-sections (2) versus the set of time-series (3) above. There are non-nested tests that generate four outcomes. Suppose the models being compared are $M1$ and $M2$, the possible outcomes are (a) reject $M1$ but do not reject $M2$ (b) reject $M2$ do not reject $M1$ (c) reject both (d) do not reject either. Whether such information is useful depends on the purpose of the exercise.

6. The nominal size, α, should fall with the number of observations, otherwise all the information is being used to reduce the Type II error, namely, the probability of not rejecting H_0 when it is false. With a fixed α and a large enough sample (typical of panels) you will always reject H_0 for any deviation from it however tiny.The test outcome will depend on the choice of null. In many cases it is not obvious what the appropriate null is, e.g. cointegration or no cointegration.

7. The p values produced by standard tests are often misinterpreted, the American Statistical Association has produced a statement on statistical significance and $p-values$ Wasserstein & Lazar (2016) and a guide to the misinterpretations.

2.7.4 Model Selection

The main issue is often one of model selection rather than testing. There are a variety of model selection criteria available. Let MLL_j be the maximised log-likelihood of model j and K_j the total number of parameters estimated in model j. The Akaike Information Criteria (AIC) chooses the model with the highest value of

$$AIC_j = MLL_j - K_j.$$

The Schwarz Bayesian Information Criterion (BIC) or Posterior Odds Criterion chooses the model with the highest value of

$$BIC_j = MLL_j - 0.5K_j \ln(NT).$$

Some econometric software packages use different versions of these formulae, based on the sum of squared errors, where you choose the model with the lowest value.

Generally the BIC is consistent in that asymptotically it will select the true model, if the true model is among the set being considered. However, there are cases in panels where it is not consistent, e.g. because of the problem of incidental trends, Moon, Perron & Phillips (2015). In some panel applications it may not be appropriate to give N and T the same weight as this formula does. It is sometimes suggested that the AIC may perform better when the true model is not in the set being considered, because being less parsimonious the extra parameters may approximate the misspecification.

These and other model selection criteria can be used both for nested or non-nested models. When comparing nested models the BIC can be interpreted as adjusting the size of the test with the number of observations. Suppose we have two models, $M1$ and $M2$, such that $M1$ has k parameters and is nested in $M2$ which has an extra variable and $k+1$ parameters. An LR test at the 5% level chooses $M2$ if $2(MLL_2 - MLL_1) > 3.84$. The BIC chooses $M2$ if $2(MLL_2 - MLL_1) > \ln NT$.

2.7.5 To pool or not to pool

The "to pool of not to pool choice" surveyed by Baltagi, Bresson and Pirotte (2008) involves the trade-off that while using a heterogeneous model avoids bias, it can cause a large loss of efficiency relative to the pooled model. Paap, Wang and Zhang (2015) suggest combining the estimators from different pooling specifications with appropriate weights to minimise mean square error.

As noted above, Pesaran and Zhou (2015) point out that even if some of the intercepts differ and are correlated with the regressors, the pooled estimator may be consistent (large N, T fixed) and more efficient than the fixed effect estimator if the differences are not too pervasive in a way that they characterise.

If one starts from heterogeneity, rather than pooling one can

- adopt a random coefficient approach;

- shrink the heterogeneous estimates to a common mean as in (11);
- have the coefficients on some variables homogeneous and others heterogeneous, as in the Pooled Mean Group estimator discussed below;
- make the heterogeneous coefficients functions of unit specific variables as in the systematic heterogeneity case in equation (13) above; or
- group the units with coefficients being homogeneous within groups and heterogeneous between groups.

If one wants to allocate the units into groups that have similar coefficients„ it is straightforward if one knows the groups, determining the groups from the data is more difficult. Bonhomme and Manresa (2012) suggest a method for doing that, which minimises a least squares criterion with respect to all possible units allowing for time-varying heterogeneity in the fixed effects. Su Shi and Phillips (2014) survey a range of methods and use penalised regression techniques, a variant of Lasso (least absolute shrinkage and selection operator) which minimises the sum of squared residuals subject to the sum of the absolute values of the coefficients being less than a constant. This leads to many coefficients being set to zero. They call their procedure for models where slope parameters are homogeneous within groups and heterogeneous across groups classifier-lasso or C-lasso. Ando and Bai (2016) also suggest a method of grouping where there is a factor error structure. Their method determines the number of groups, the number of group specific common factors and the regression coefficients. Factor error structures are discussed below under cross-section dependence.

2.8 PPP

We shall apply the various estimators to examine the Purchasing Power Parity (PPP) relationship, using the data discussed in the previous chapter. PPP suggests that the real exchange rate

$$q_{it} = s_{it} - p_{it} + p_{it}^*$$

is broadly constant in the long run. If we formulate this as $q_{it} = q_i + u_{it}$, and define the log price differential $d_{it} = p_{it} - p_{it}^*$ we have

$$\begin{aligned} s_{it} - d_{it} &= q_i + u_{it} \\ s_{it} &= q_i + d_{it} + u_{it} \\ s_{it} &= \alpha_i + \theta d_{it} + u_{it} \end{aligned}$$

so in a time-series regression of the log spot on the log price differential the slope θ should be unity and the intercept α_i should be the mean of the log real rate. Since we have rescaled the prices and exchange rates to be in the same units, the ratio $S_{it}P_{it}^*/P_{it}$ is unity in 1995 so the log real exchange rate q_{it} is zero in $t = 1995$. If they were all constant then we should have $q_i = q = 0$.

We begin by estimating the equation for each country

$$s_{it} = \alpha_i + \theta_i d_{it} + u_{it}$$

and considering the distribution of the individual country estimates $\widehat{\theta}_i$. They are distributed around unity, with a mean of 1.179 (standard error 0.077) but range from 0.679 (UK) to 1.866 (Austria), with only three countries lying between 0.9 and 1.1.

We now turn to various pooled estimators. We will estimate this relationship in five ways: pooled OLS, the 1-way FE estimator, the RE estimator, the between estimator and the Swamy RCM estimator.

The results using all $17 \times 104 = 1768$ observations are given below. The conventional standard errors reported are unlikely to be reliable because there is substantial serial correlation and heteroskedasticity, so we will focus on the size rather than the significance of the estimates. A variety of robust standard errors are available, but since this model is likely to be misspecified, because of omitted dynamics, we do not pursue this.

Table 2.1 Static PPP

Estimator	$\hat{\theta}$	SE	R^2
POLS	1.240	0.015	0.792
FE	1.107	0.018	0.866
RE	1.112	0.017	
Between	1.360	0.128	0.882
RCM	1.183	0.089	

The hypothesis that all the intercepts are equal is clearly rejected (by the LR and F statistics) and including separate intercepts results in a substantial increase in R^2 from 0.79 to 0.87. POLS is strongly rejected by an LM test for heteroskedasticity. A Hausman test just fails to reject the RE specification versus the FE model. The test statistic at 3.82 is just less than the $\chi^2(1)$ 5% critical value, ($p = 0.051$). The FE and RE estimates are very close. The RCM estimates are very similar to the MG estimates (not reported) at about 1.18. The Wald test for equality of all the coefficients strongly rejects the hypothesis, the test statistic which is 1713 is very large relative to the $\chi^2(32)$ critical value. The degrees of freedom are $(N-1)\,k$, 16 restrictions on two parameters). The between estimate is much larger than the others and this is reflected in the pooled estimate which gives equal weight to the between and within (FE) variation.

All the individual equations show substantial serial correlation. For comparison, the same estimators were applied to the first differences of the data.

$$\Delta s_{it} = \alpha_i + \theta_i \Delta d_{it} + u_{it}$$

The intercept now measures any trend in the real exchange rate. We lose one observation for each country because of differencing so the sample size is 1751. The estimates are

Table 2.2
Differenced PPP equation

Estimator	$\hat{\beta}$	SE	R^2
POLS	0.429	0.106	0.009
FE	0.243	0.118	0.018
RE	0.349	0.111	
Between	1.224	0.106	0.898
RCM	0.376	0.111	

In this case, the test results are completely different. The hypothesis that the intercepts are all equal cannot be rejected ($p^{LR} = 0.45$), the Hausman test rejects the RE specification $\left(p^H = 0.09\right)$ because the RE and FE differ significantly. The Wald homogeneity test does not reject equality of slope coefficients $\left(p^W = 0.31\right)$. The striking feature is that, except for the Between, the estimates are all much smaller than obtained with the levels regression. It is likely that the other first difference regressions are picking up the short-run effect, which is very small, while the Between, which uses the cross-section on the time averaged differences, is picking up the same long-run effect as the levels regression.

These models are all probably misspecified, in that they do not allow for enough dynamics. However, it is clear that while in cross-section there is quite strong evidence for PPP, relative inflation differentials can explain almost 90% of the variance in the percentage change in the spot rate over this period, in time-series there is rather little evidence, as suggested by the dispersion of country levels estimates and the low estimates from the first difference regressions.

2.9 Concluding questions

1. Cross-sections or Time-series?
2. Heterogeneity in coefficients and variances?
3. How to test for heterogeneity? How to group?
4. Estimators: POLS; Between; GLS RE; Within (1FE or 2FE); Mean Group; Swamy RCM?
5. Misspecification testing?
6. Standard errors for estimators if there is heteroskedasticity or autocorrelation?

3 Dynamic Linear Regressions

3.1 Single Time-series

3.1.1 ARDL

The most common time series model used to examine the relationship between a single enogeneous variable and a vector of exogenous variables is the autoregressive distributed lag (ARDL) or dynamic linear regression model. In the case of two exogenous variables (x_t, z_t) the ARDL(p, q, r) takes the form

$$y_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=0}^{q} \beta_j x_{t-j} + \sum_{j=0}^{r} \gamma_j z_{t-j} + u_t$$

where u_t is usually assumed to be white noise, though it could also be moving average. If the error is white noise, the parameters of the ARDL can be estimated by OLS, which is consistent, large T, but not unbiased in small samples. The process is stable, if all the roots (solutions), λ_i, of the characteristic equation

$$1 - \alpha_1\lambda - \alpha_2\lambda^2 - ... - \alpha_p\lambda^p = 0 \tag{15}$$

lie outside the unit circle (are greater than one in absolute value). In practice checking that $-1 < \sum_{j=1}^{p} \alpha_j < 1$ gives an indication of stability.

If it is stable the long-run solution is

$$y = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_i} + \frac{\sum_{j=0}^{q} \beta_i}{1 - \sum_{j=1}^{p} \alpha_i} x + \frac{\sum_{j=0}^{r} \gamma_i}{1 - \sum_{j=1}^{p} \alpha_i} z = \theta_0 + \theta_x x + \theta_z z$$

We will use an ARDL(1,1) for illustration, this is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t. \tag{16}$$

It is stable if $-1 < \alpha_1 < 1$, and then has a long run solution:

$$y = \frac{\alpha_0}{1 - \alpha_1} + \frac{\beta_0 + \beta_1}{1 - \alpha_1} x = \theta_0 + \theta_x x.$$

The ARDL(1,1) can be estimated by OLS and the long run parameter estimated by

$$\widehat{\theta}_x = \frac{\widehat{\beta}_0 + \widehat{\beta}_1}{1 - \widehat{\alpha}_1}.$$

This estimator will be biased for small T, both because of the bias in the estimates of the ARDL and because it is a non-linear function of the short-run coefficients. Since the expected value of a ratio is not equal to the ratio of the expected values, using bias corrected estimators of the numerator and denominator may not help and could make the bias in the long-run coefficient worse. Estimation of standard error of $\widehat{\theta}_x$ can be done using the delta method, which is programmed into most packages. How good the estimates of the confidence interval will be is unclear since the distribution of $\widehat{\theta}_x$ has no finite sample moments, because $\widehat{\alpha}_1$ can equal unity. Even, if $\widehat{\alpha}_1 = 1$ is excluded, the distribution is heavily skewed with a longer tail on the right hand side for positive θ_x, because values of $\widehat{\alpha}_1$ close to unity produce very large values of $\widehat{\theta}_x$. There is a small literature on these problems, but they are usually ignored in most applied work.

3.1.2 Error Correction models

The ARDL(1,1) can be rewritten (reparameterized) in a number of other useful ways:

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1 y_{t-1} + b_1 x_{t-1} + u_t \tag{17}$$

where

$$a_0 = \alpha_0; b_0 = \beta_0; a_1 = (\alpha_1 - 1); b_1 = \beta_0 + \beta_1;$$

or in terms of adjustment to a long-run target:

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2 (y_{t-1}^* - y_{t-1}) + u_t$$

where the long-run target or equilibrium (as calculated above) is

$$y_t^* = \theta_0 + \theta_x x_t$$

and (λ_1, λ_2) are adjustment coefficients which measure how y adjusts to changes in the target and deviations from the target, respectively. The relation between the estimated and theoretical parameters is

$$a_0 = \lambda_2 \theta_0; a_1 = -\lambda_2; b_0 = \lambda_1 \theta_x; b_1 = \lambda_2 \theta_x.$$

In this form, usually known as an error (or equilibrium) correction model (ECM), the y_t changes in response to changes in the target and the error, the deviation of the actual from the target in the previous period: $(y_{t-1}^* - y_{t-1})$.[13] An alternative parameterization, which nests the partial adjustment model ($\beta_1 = 0$) is:

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1) y_{t-1} + (\beta_0 + \beta_1) x_t - \beta_1 \Delta x_t + u_t.$$

When you **reparameterise** a model, as done above, exactly the same number of parameters are estimated (4 in this case), just written in different ways. The statistical properties of the model do not change, the estimated residuals $(\hat{u}_t)$, the standard error of the regression $(\hat{\sigma})$ and the maximised log-likelihood (MLL) are identical between the different versions. R^2 will change, because the proportion of variation explained is measured in terms of a different dependent variable, Δy_t rather than y_t. Any misspecification tests that use fitted values of the dependent variable (e.g. RESET tests) will also change. For such tests Δy_t should be used as the dependent variable, otherwise there is the danger of creating a spurious regression in the second stage regression for the diagnostic test since $\hat{y}_t$ may be $I(1)$.

3.1.3 General to specific modelling

While a reparameterisation does not change the number of parameters, when you **restrict** a model, you reduce the number of parameters estimated and such

[13] For more than one exogenous regressor, this form imposes restrictions on the estimated ECM. These restrictions are often rejected because the estimated dynamics reflects both expectations and adjustment processes, not just adjustment as written here.

restrictions are testable. Within time-series ECM modelling, the practice is to start with a very general model and then test down to a more restricted version. This is because the tests are much more likely to be valid in a very general model than in a more restricted model which is more likeley to be misspecified. A useful procedure in many circumstances is to start with a general model, and test down to specific restricted cases. This approach was adopted in section 2.7.1 on LR testing of homogeneity restrictions. Similarly one might start with long lags and test whether a model with fewer lags can be accepted. The ARDL(1,1) (16) nests a number of interesting restricted special cases, including:

(a) Static regression model: $\alpha_1 = 0; \beta_1 = 0$.

(b) First difference regression model: $\alpha_1 = 1; \beta_1 = -\beta_0$. This is the case of no long run relation and the tests are non-standard, see Pesaran, Shin and R.J. Smith (2001).

(c) Partial Adjustment Model (PAM): $\beta_1 = 0,$. generated by

$$\Delta y_t = \lambda(y_t^* - y_{t-1}).$$

(d) First order disturbance serial correlation model: $\beta_1 = -\beta_0\alpha_1$. This is got by assuming that the model is:

$$y_t = \alpha + \beta x_t + v_t;\ v_t = \rho v_{t-1} + \varepsilon_t; \varepsilon_t \sim iid(0, \sigma^2)$$

which can be written:

$$y_t = \alpha + \beta x_t + \rho v_{t-1} + \varepsilon_t$$

by noting that

$$v_t = y_t - \alpha - \beta x_t \text{ and } v_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$$

$$y_t = \alpha + \beta x_t + \rho(y_{t-1} - \alpha - \beta x_{t-1}) + \varepsilon_t$$
$$y_t = \alpha(1-\rho) + \beta x_t + \rho y_{t-1} - \beta\rho x_{t-1} + \varepsilon_t$$

which is of the same form as (16) with the restriction that the coefficient of x_{t-1} equals the negative of the product of the coefficients of x_t and y_{t-1}, i.e. $\beta_1 = -\beta_0\alpha_1$ in terms of the parameters of the unrestricted model. This is sometimes called the *common factor model*, since using the lag operator, $Ly_t = y_{t-1}$, it can be written $(1 - \rho L)y_t = (1 - \rho L)(\alpha + \beta x_t + v_t)$, that is, both sides of the static model are multiplied by the common factor $(1-\rho L)$. The restricted model, with AR(1) errors, is not linear in the parameters and is estimated by GLS or ML.

(e) Unit long-run coefficient model: $\beta_1 + \beta_0 + \alpha_1 = 1$. The restricted model can be written:

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1(y_{t-1} - x_{t-1}) + e_t.$$

and the restriction is equivalent to imposing $b_1 = -a_1$ in (17).

(f) Random Walk with drift: $\alpha_1 = 1; \beta_1 = \beta_0 = 0$.

In addition, to testing down, it is standard to carry out diagnostic or misspecification tests to check that the assumptions made about the errors are not violated by the residuals. These would normally include tests for serial correlation, heteroskedasticity, normality, a RESET test for functional form and tests for structural stability.

In panel time-series, neither general to specific modelling nor extensive diagnostic checking are as common as they should be. This is partly, because for large N it generates a lot of tests and because the null hypothesis of such panel tests, that none of the groups are misspecified, is quite strong. There is also an issue of multiple testing and how one should adjust the size of the tests in such circumstances. A review of diagnostic checking in panels is provided in Banerjee, Eberhardt and Reade (2010). Bai (2010) discusses testing for breaks in coefficients and variances in panels.

3.2 Dynamic Homogeneous Panels

Consider estimating the partial adjustment model or ARDL(1,0) equation with homogeneous slopes but different intercepts

$$y_{it} = \alpha_i + \beta x_{it} + \gamma y_{it-1} + u_{it};\ u_{it} \sim iidN(0, \sigma_i^2) \tag{18}$$

using panel data where the independence assumption for the innovations refers to time and cross-section, $E(u_{it}u_{jt-s}) = 0$ for $i \neq j$ or $s \neq 0$. This is just a dynamic version of the fixed effect model which we can estimate by OLS, to give, say, $\widehat{\beta}_{FE}$ and $\widehat{\gamma}_{FE}$. The long run solution is

$$y = \frac{\alpha_i}{1-\gamma} + \frac{\beta}{1-\gamma}x$$

and the *long run effect* of x on y is measured by $\theta_x = \beta/(1-\gamma)$. We will treat the data as being $I(0)$, though many of the results hold if the variables are $I(1)$ but cointegrated.

Let us consider the properties of these 1-way FE estimators. If $\beta_i = \beta$ and $\gamma_i = \gamma$ then $\widehat{\beta}_{FE}$ and $\widehat{\gamma}_{FE}$ are consistent as $T \to \infty$, for fixed N; though they are biased in small samples, because of the lagged dependent variable bias. However, both give inconsistent estimates of the mean effects β and γ as $N \to \infty$ for fixed T. The latter, called *Nickel (initial condition) bias* is the result of the fact that the lagged dependent variable bias arising from the initial conditions is not removed by increasing N (Nickel, 1981). Bun and Kiviet (2003) for balanced panels and Bruno (2005) for unbalanced panels discuss the size of the bias in more detail.

If both $N \to \infty$ and $T \to \infty$, then to ensure consistency of the least squares estimates T must grow sufficiently fast relative to N, Hahn & Kuersteiner (2002) and Alvarez and Arellano (2003). For GMM to work, N must be large relative to T as T grows.

3.2.1 Small T case

For a stationary autoregression

$$y_{it} = \alpha_i + \rho y_{it-1} + u_{it};$$

where u_{it} is independent across groups and $\rho < 1$, the limit of the FE estimator of ρ as $N \to \infty$ is

$$\begin{aligned} P \lim_{N\to\infty} (\widehat{\rho} - \rho) &= -\frac{1+\rho}{T-1} A_T^{-1} B_t < 0 \\ A_T &= 1 - \frac{2\rho}{(1-\rho)(T-1)} \left[1 - \frac{1-\rho^T}{T(1-\rho)} \right] \\ B_T &= 1 - \frac{1-\rho^T}{T\,(1-\rho)} \end{aligned}$$

When $\rho = 1$, the first term of the bias is of the order $-3/T$ for large N rather than $-(1+\rho)/T$, Phillips and Sul (2006). They also show that if a group specific incidental trend is included in the model

$$y_{it} = \alpha_i + b_i t + \rho y_{it-1} + u_{it};$$

the bias is approximately twice as large. When $\rho < 1$ the first term is $-2(1+\rho)/(T-2)$ and when $\rho = 1$ it is $-7.5/(T+2)$. This bias is so large that for small T the estimate of ρ can be a fairly large negative number, even when the true ρ is near unity. In certain circumstances exogenous variables reduce the bias.

A very large number of estimators have been suggested to for dynamic fixed effect models in the small T case.

The traditional approach starts by differencing the data to eliminate the α_i. This induces a correlation between the differenced error and the right hand side variables which is dealt with by IV or GMM. There is the problem that the traditional GMM estimator works very badly if the variables are I(1). For instance, if a variable is a random walk, lagged changes are not valid instruments for current changes because they are not correlated. There are various ways of dealing with this weak instrument problem and a vast literature. Roodman (2009) reviews the evidence on weak instruments. Hsiao, Pesaran and Tahmiscioglu (2002) suggest a maximum likelihood estimator and Binder, Hsiao and Pesaran (2005) suggest a maximum likelihood estimator that is consistent whether the variables are I(0) or I(1) Han and Phillips (2011) examine the properities of the maximum likelihood estimator.

Gourieux, Phillips and Yu (2010) suggest an indirect inference estimator and compare it to a number of other estimators; Han, Phillips and Sul (2014) suggest a new estimator based on a different transformation, they call X-differencing. But as yet there is little experience with the performance of these estimators.

There are a variety of bias adjusted methods, which take account that the size of the bias is a function of T and ρ_i or γ_i. Using the estimates a bias adjustment can be calculated. Phillips and Sul (2002) propose a median unbiased estimator to deal with the small sample bias in the presence of correlations between groups and heterogeneity. Hsiao & Zang (2015) examine the properties of the various estimators when either N or T or both are large.

Dhaene and Jochmans (2015) propose a split panel jackknife estimator that seems to work well in a variety of cases both for linear and non-linear models.

If the sample is of size T the bias in $\widehat{\beta}_T$ is roughly δ/T. If you estimate on a sample of $T/2$ the bias is roughly $2\delta/T$ The procedure then estimates $\widehat{\beta}_T$ from the full sample, then $\widehat{\beta}_1$ from the first $T/2$ observations, and $\widehat{\beta}_2$ from the second $T/2$ observations. The bias corrected estimate is then given by $2\widehat{\beta}-(\widehat{\beta}_1-\widehat{\beta}_2)/2$. This cancels out the bias.

There is a further problem that because the long-run coefficient, $\theta = \beta/(1-\gamma)$, is a non-linear function of the short-run coefficients, bias corrections to the short-run coefficients can worsen the bias of the long-run coefficient. There is a similar problem for the mean lag, or half life, which are non-linear functions of the short-run coefficients.

For most of these notes, we will assume that T is sufficiently large that the Nickel bias is not a problem. However, even if T is large, we also require that $N/T \to 0$, for the fixed effect estimator to be consistent, Hahn & Kuersteiner (2002).

3.3 Dynamic Heterogeneous Panels

3.3.1 Heterogeneity bias

Suppose the slopes differ, $\beta_i \neq \beta$, $\gamma_i \neq \gamma$, so that the true model is

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i y_{it-1} + u_{it};\ u_{it} \sim IN(0, \sigma_i^2) \tag{19}$$

whereas we assume the slopes are the same and estimate (18). The parameters of (19) can be efficiently estimated by OLS for each group, though there is a small T downward bias in $\widehat{\gamma}_i$. We can examine the distribution of the estimates and calculate weighted or unweighted means as discussed above. In heterogeneous panels ($\beta_i \neq \beta$ or $\gamma_i \neq \gamma$) the FE estimator, which imposes homogeneity, gives inconsistent estimates of β and γ, the expected values of β_i and γ_i even for large T. This inconsistency in heterogenous dynamic panel models was first noted by Robertson and Symons (1992) and is analysed in detail in Pesaran and Smith (1995) and Pesaran, Smith and Im (1996). It arises because under heterogeneity the composite disturbance has the form

$$w_{it} = (\beta_i - \beta)x_{it} + (\gamma_i - \gamma)y_{it-1} + u_{it}. \tag{20}$$

Even if $\gamma_i = \gamma$ this composite disturbance will be serially correlated if x_{it} is serially correlated, as it usually is, and so will not be independent of the lagged dependent variable. The heterogeneity bias, which depends on the serial correlation in the x and the variance of the random parameters, can be quite severe. Suppose x_{it} is generated by an AR1 process:

$$x_{it} = \mu_i(1-\rho) + \rho x_{it-1} + \varepsilon_{it}$$

where μ_i is the unconditional mean.

Then as x_{it} tends towards being an I(1) variable, i.e. as $\rho \to 1$, the Probability Limits of the FE estimator (taken by first letting $T \to \infty$. then letting

$N \to \infty$) are given by:

$$\underset{\rho \to 1}{P\lim}\ (\widehat{\beta}_{FE}) = 0; \quad \underset{\rho \to 1}{P\lim}\ (\widehat{\gamma}_{FE}) = 1$$

irrespective of the true values of β and γ, the expected values of β_i and γ_i. Notice that these results are not valid for $\rho = 1$, the I(1) case considered in the next section. The asymptotic bias in the estimator of the long-run coefficient $\widehat{\theta}_{FE} = \widehat{\beta}_{FE}/(1 - \widehat{\gamma}_{FE})$ is not as severe, because the biases in the numerator and denominator tend to cancel out. Notice that if ρ is positive, the usual case, the heterogeneity bias in $\widehat{\gamma}_{FE}$ is upwards, the opposite of the lagged dependent variable bias, so the two biases may offset each other at certain sample sizes. Imbs et al (2005) argue that the heterogeneity bias arising from aggregation is important for testing PPP.

In the case of the ARDL(1,1) model

$$y_{it} = \alpha_i + \beta_0 x_{it} + \beta_1 x_{i,t-1} + \gamma y_{it-1} + w_{it}$$

as $\rho \to 1$ (and x_{it} tends towards being an I(1) process) we have

$$\underset{\rho \to 1}{P\lim}\ (\widehat{\beta}_0^{FE}) = E(\beta_{0i}); \quad \underset{\rho \to 1}{P\lim}\ (\widehat{\beta}_1^{FE}) = -E(\beta_{0i}); \quad \underset{\rho \to 1}{P\lim}\ (\widehat{\gamma}^{FE}) = 1.$$

and so the estimates suggest that a first difference model, $\Delta y_{it} = \alpha_i + \beta_0 \Delta x_{it} + w_{it}$ seems appropriate, even though this is not in fact the case.

The heterogeneity bias cannot be dealt with by traditional instrumental variable estimators. As is clear from (20) anything uncorrelated with the errors will be uncorrelated with the regressors and thus not a valid instrument. However, average estimators such as the Swamy RCM or MG estimators discussed above provide consistent estimators for large N and T. For small T the average estimators will suffer small sample bias, which will not be removed however large N. For small N the average estimators are also likely to be sensitive to outliers.

3.3.2 Forms of randomness

There is an issue as to how the randomness is introduced. We could either have the randomness in the short-run coefficients

$$H_a : \beta_i = \beta + \eta_{1i}; \ \gamma_i = \gamma + \eta_{2i}$$

or in the mean lag and the long-run coefficients

$$H_b : \psi_i = \frac{\gamma_i}{1 - \gamma_i} = \psi + \xi_{1i}; \ \theta_i = \frac{\beta_i}{1 - \gamma_i} = \theta + \xi_{2i}$$

where either the η_{ki} or the ξ_{ki} $k = 1, 2$ have zero means and constant covariances.

The 'average' long-run effect of x can be estimated either as $\theta^* = \overline{\beta}/(1 - \overline{\gamma})$ or $\overline{\theta} = N^{-1} \sum \widehat{\theta}_i$, etc. As $N \to \infty$ these two estimators converge to different parameters θ^* to $E(\beta_i)/(1-E(\gamma_i))$ and $\overline{\theta}$ to $E(\beta_i/(1-\gamma_i)) = E(\theta_i)$, which will differ unless β_i and γ_i are independently distributed. In addition, $E(\theta_i)$ will not exist unless $\gamma_i \neq 1$. In most cases $E(\theta_i)$ seems to be the parameter of interest, though Phillips and Moon (1999) discussed below take a different position. However, $\overline{\theta}$ may have a large variance if some estimates of γ_i are very close to unity.

3.3.3 Testing

There are two ways to determine whether the heterogeneity bias is a problem. Suppose the model is

$$y_{it} = \alpha_i + \boldsymbol{\delta}_i' \mathbf{z}_{it} + u_{it}$$

where $\mathbf{z}_{it}$ is a $k \times 1$ vector which includes all the variables, x_{it} and y_{it-1} and the null hypothesis is $\boldsymbol{\delta}_i = \boldsymbol{\delta}$. One way is simply to test for heterogeneity using a LR test, which is χ^2 with $(N-1)\,k$ degrees of freedom. If N is small and one is interested in whether the coefficients are, in fact, the same this is the appropriate test. If N is large relative to T the Pesaran and Yamagata (2008) test described above should be used.

Alternatively one may be interested in estimating the average effects, $\delta = E(\delta_i)$. Suppose we have a model with strictly exogenous variables and a lagged dependent variable. The average effect could be estimated by either the Fixed Effect Estimator or the Mean Group or Random Coefficient Model. The FE estimator will be efficient under the null of no heterogeneity but inconsistent under the alternative of heterogeneity. The averages, MG and Swamy, will be consistent under both, but inefficient under the null. This suggests using a Hausman test. Suppose the $k \times 1$ vector of FE estimates of the slopes is $\widehat{\boldsymbol{\delta}}$ and the vector of MG or RCM estimates is $\widetilde{\boldsymbol{\delta}}$, then under the heterogeneity null

$$\left(\widetilde{\delta} - \widehat{\delta}\right)' \left[V\left(\widetilde{\delta}\right) - V\left(\widehat{\delta}\right)\right]^{-1} \left(\widetilde{\delta} - \widehat{\delta}\right) \sim \chi^2(k)$$

This statistic tests whether the two estimates of the means of the coefficient are significantly different. If the null is not rejected, they are not significantly different, one uses the fixed effect estimator, since it is efficient. If the null is rejected, they are significantly different, one uses the average estimator. If there are outliers the average estimator may have a large variance and in that case the Hausman test would have little power. Notice that the Hausman test is not a test for heterogeneity itself, but a test for the implications of heterogeneity on the consistency of different estimators, so it may have little power in certain circumstances.

While the Hausman test works for the dynamic panel regression with both exogenous variables and lagged dependent variables, it will not work for either (a) the case where all the regressors are strictly exogenous, because then both FE and MG are unbiased so $E\left(\widetilde{\delta} - \widehat{\delta}\right) = 0$ or (b) a pure autoregression, where they will both have the same asymptotic variance covariance matrix, so

$E\left[V\left(\widetilde{\delta}\right) - V\left(\widehat{\delta}\right)\right] = 0$. Pesaran and Yamagata (2008) discuss this in more detail.

3.3.4 Pooled Mean Group

The heterogeneity issue is important since when the hypothesis that the slope coefficients are in fact identical is tested, it is is often rejected. The acceptance of homogeneity for the dynamic PPP equation in our application below is quite unusual. Pesaran, Shin and R.P. Smith (1999) argue that it is more likely that long run effects are homogeneous, due to budget or solvency constraints, arbitrage conditions or common technologies. They suggest a pooled mean group estimator that constrains the long-run coefficients to be the same while allowing the short-run coefficients and variances differ over groups. Writing the ARDL(1,1) model in ECM form gives

$$\Delta y_{it} = \alpha_i + \beta_i \Delta x_{it} + \lambda_i(\theta x_{it-1} - y_{it-1}) + u_{it},\ u_{it} \sim iidN(0, \sigma_i^2)$$

where the short-run parameters $\alpha_i, \beta_i, \lambda_i$ and the variances σ_i^2 differ over groups, but the long-run parameter θ does not. In some cases, θ will be specified a priori as $+1$, or -1, as in parity conditions. This model has a non-linear cross-equation parameter restriction but is fairly easily estimated, e.g. using a Newton-Raphson method. This method can allow the dynamics to differ between groups with different lag lengths and allows some groups having $\lambda_i > 0$, cointegration and in others $\lambda_i = 0$, no cointegration. It can estimate the long-run effect even when individual θ_i cannot be identified because of exact multicollinearity between the elements of x in particular groups and can use very small samples for some cross-section units.

In the applied examples that Pesaran et al. (1999) discuss, consumption functions and energy demand functions, the hypothesis that the long-run coefficients are identical is rejected at conventional significance levels by LR tests, but the Hausman test indicates no significant differences between PMG and MG estimates of the common long-run effect.

3.4 Heterogeneity

The decision whether to pool or not will depend both on the degree of heterogeneity and on the purpose of the exercise and there are a range of alternative procedures available discussed in the section on "to pool. or not to pool above" and in Su, Shi and Phillips (2014).

When T is large enough, it is sensible to compare the estimates from the FE and MG or Swamy RCM and determine whether the differences are large, e.g. by a Hausman test or informally, and whether the differences are important for the purpose at hand. This acts as a specification test. It is possible that the estimates of the long-run parameters are not very diffferent, but that the speeds of adjustment are different. In issues like PPP and convergence the parameter of interest is the speed of adjustment, so differences in this matter. Baltagi and

Griffin (1997) use a large number of different estimators to estimate demand for gasoline in OECD countries. They find that there is a very large degree of heterogeneity in the individual country estimates but that for the purpose of medium term forecasting simple pooled estimators do well. Attanasio et al. (2000) use a variety of estimators to measure the dynamic interaction between savings, investment and growth and find little difference between FE and MG. Baltagi et al (2002) consider 23 estimators on a panel of French regions 1973-1998. For forecasting purposes, the within FE or within 2SLS do best and the heterogeneous and shrinkage estimators forecast badly, primarily because of the dispersion of the estimates. Bond and Eberhardt (2013) suggest a estimator that takes account of heterogeneity, non-stationarity and cross-section dependence.

A fundamental issue is the interpretation of the heterogeneity. It is quite common that not only is homogeneity rejected, but there is massive dispersion and many individual estimates are economically implausible, while averaged or fixed effect estimates are sensible and not very different by Hausman tests. Boyd and Smith (2002) give examples based on data for 57 developing countries where $T = 31$. In a static PPP equation of log spot on log price differential the mean was 1.13 (0.069), the RCM estimate 1.13 (0.068) the fixed effect 1.02 (0.003), but the standard deviation of the estimates was 0.5213 and the range was from -0.40 to 2.47. This does not reflect dynamic misspecification, in a first order dynamic PPP equation the dispersion is even greater. The mean long-run effect of price differential on the exchange rate is 0.986 (0.277), the RCM 1.028 (0.099) and the fixed effect 0.984(0.015); standard errors in parentheses. The standard deviation of the individual estimates was 2.095 and the range was from -9.02 to 8.20. Some of the outliers are produced because the long-run effect is measured as a ratio in which the denominator may be near zero as discussed above.

One possible explanation is omitted variable bias. Suppose we wrap all the omitted variables in w_{it} and write the model as

$$y_{it} = \boldsymbol{\beta}_i' \mathbf{x}_{it} + w_{it} + u_{it},\ i = 1, 2, ..., N; t = 1, 2, ..., T$$

and the omitted variables are correlated in particular samples with the regressors

$$w_{it} = \mathbf{b}_{iT}' \mathbf{x}_{it} + v_{it}$$

giving biased sample estimates in the regression (excluding w_{it}) of

$$E(\widehat{\boldsymbol{\beta}}_{iT}) = \boldsymbol{\beta}_i + \mathbf{b}_{iT}.$$

The specification error in the estimate $\widehat{\boldsymbol{\beta}}_{iT}$ could be large and signficant if the w_{it} have a big effect on y_{it} and are correlated with $\mathbf{x}_{it}$. If the w_{it} are structural factors operating in all time periods and countries, this would cause a systematic bias in the average estimates of $\boldsymbol{\beta}$. If they are not structural but just happen to be correlated in a particular sample then $E(\mathbf{b}_{iT}) = 0$ and these biases will cancel out when averaged across countries or over longer time-periods. Such correlated shocks will cause structural instability, because $\mathbf{b}_{iT}$ is not constant over time, and heterogeneity, because $\mathbf{b}_{iT}$ is not constant over countries. Unobserved common factors correlated with the regressors, discussed below, may also play a role in the mispecification.

In judging the importance of the heterogeneity, purpose is crucial. A parsimonious homogeneous model may be optimal for forecasting, but more heterogeneous models might be required for policy or testing theories.

3.5 PPP

The static version of the PPP equation is

$$s_{it} = q_i + \theta_i d_{it} + u_{it}$$

where q_i is the log equilibrium real exchange rate, s_{it} is the log spot, and d_{it} is the log price differential. We expect $\theta_i = 1$. Suppose that this holds only in the long-run,

$$s_{it}^* = q_i + \theta_i d_{it}$$

and there is error correction adjustment

$$\Delta s_{it} = \lambda_{1i} \Delta s_{it}^* + \lambda_{2i}(s_{it-1}^* - s_{it-1}) + u_{it}$$

then the ECM form of the dynamic version of the PPP equation is

$$\Delta s_{it} = a_{0i} + b_{0i} \Delta d_{it} + b_{1i} d_{it-1} + a_{1i} s_{it-1} + u_{it} \tag{21}$$

or in ARDL(1,1) form

$$s_{it} = \alpha_{0i} + \beta_{0i} d_{it} + \beta_{1i} d_{it-1} + \alpha_{1i} s_{it-1} + u_{it} \tag{22}$$

We have three sets of parameters, those of the estimated ARDL, estimated ECM and theoretical model. They are related by:

$$\begin{aligned}
\alpha_{0i} &= a_{0i} = \lambda_{2i} q_i; \\
\beta_{0i} &= b_{0i} = \lambda_{1i} \theta_i; \\
\theta_i &= (\beta_{0i} + \beta_{1i}) / (1 - \alpha_{1i}) = -b_{1i}/a_{1i}; \\
a_{1i} &= 1 - \alpha_{1i} = -\lambda_{2i}.
\end{aligned}$$

We first estimate 21 and 22 by POLS to see the relationship between them. We drop the first two observations.

$$\begin{array}{cccccc}
s_{it} = & 0.006 & +0.342 d_{it} & -0.312 d_{it-1} & +0.961 s_{it-1} & R^2 = 0.982 \\
 & (0.002) & (0.116) & (0.113) & (0.007) & s = 0.056
\end{array}$$

$$\begin{array}{cccccc}
\Delta s_{it} = & 0.006 & +0.342 \Delta d_{it} & +0.030 d_{it-1} & -0.039 s_{it-1} & R^2 = 0.036 \\
 & (0.002) & (0.116) & (0.010) & (0.007) & s = 0.056
\end{array}$$

We then estimate (21) by POLS, RE, FE and RCM. We cannot use the between (cross-section) estimator for dynamic models because the averages of the lagged variables are obviously endogenous. The results are reported in the Table A5.

Table 3.1. ECM estimates

	Δd_{it}	d_{it-1}	s_{it-1}	
POLS	0.342	0.030	-0.039	$R^2 = 0.036$
	(0.116)	(0.010)	(0.007)	$s = 0.056$
RE	0.351	0.033	-0.042	
	(0.117)	(0.011)	(0.007)	
FE	0.402	0.045	-0.052	$R^2 = 0.040$
	(0.123)	(0.012)	(0.009)	$s = 0.057$
RCM	0.364	0.029	-0.049	
	(0.223)	(0.021)	(0.011)	

An F test does not reject equality of intercepts, $p = 0.98$, and a Hausman test does not reject RE $p = 0.20$ and a $\chi^2(64)$ Wald test does not reject equality of all slopes and intercepts $p = 0.88$. There is very little difference between the various estimates. In the fixed effect model we cannot reject the PPP hypothesis, $b_1 + a_1 = 0$, $t = -1.169$; but we would reject the lack of a long-run relationship $b_1 = a_1 = 0$ $\chi^2(64) = 39.11$ using inappropriate standard critical values. For this dynamic PPP equation, there is very little evidence of significant heterogeneity which is unusual in panel applications. As a consequence heterogeneity bias is small and there is little difference between the RCM and FE estimates of the adjustment coefficient.

4 Unit Roots

4.1 Models for Single Time series

4.1.1 ARIMA models

Again we start with some time-series background. Suppose we have observations on some economic variable, $y_t, t = 1, 2, ..., T$, which may already have been transformed, e.g. the logarithm of GDP. It is useful to regard each y_t as a random variable with a density function, $f_t(y_t)$ and we observe one realisation from the distribution for that period. A family of random variables indexed by time is called a stochastic process, an observed time series is called a realisation of the stochastic process. A stochastic process is said to be 'strongly stationary' if its distribution, $f_t(y_t)$, is constant through time, i.e. $f(y_t)$. It is first-order stationary if it has a constant mean. It is second order, weakly or covariance stationary if it also has constant variances and constant covariances between y_t and y_{t-i}; the latter means that the autocovariances (covariances of y_t with itself in previous periods) are only a function of i (the distance apart of the observations) not t, the time they are observed.[14] These autocovariances summarise the dependence between the observations and they are often represented by the autocorrelation function (ACF) or correlogram, the vector (graph against i) of

[14] If its distribution does not have moments a series can be strongly stationary but not weakly stationary.

the autocorrelations $r_i = Cov(y_t, y_{t-i})/Var(y_t)$.[15] If the series is stationary, the correlogram converges to zero quickly. If it converges to zero slowly, it is said to be a long-memory process.

The order of integration is the number of times a series needs to be differenced in order to make it stationary. A series, y_t, is said to be:

integrated of order zero, $I(0)$, if y_t is stationary;

integrated of order one, $I(1)$, if the first difference $\Delta y_t = y_t - y_{t-1}$ is stationary;

integrated of order two, $I(2)$, if the second difference

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

is stationary.

Notice that $\Delta^2 y_t \neq \Delta_2 y_t = y_t - y_{t-2}$. The order of integration need not be an integer, but we will not consider fractional integration.

A stochastic process is said to be *white noise* if

$$E(\varepsilon_t) = 0;\ E(\varepsilon_t^2) = \sigma^2;\ E(\varepsilon_t \varepsilon_{t-i}) = 0,\ i \neq 0$$

We will use ε_t below to denote white noise processes.

A first-order (one lag) autoregressive process (AR1) takes the form:

$$y_t = \rho y_{t-1} + \varepsilon_t$$

with $E(y_t) = 0$, and if $\mid \rho \mid < 1$ it is stationary and $E(y_t) = 0$. If it is stationary, then by continuous substitution:

$$y_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + ... \tag{23}$$

the variance of y_t is $E(y_t^2) = \sigma^2/(1-\rho^2)$ and $corr(\ y_t, y_{t-k}) = \rho^k$ so it declines exponentially. A constant can be included $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$, then $E(y_t) = \alpha/(1-\rho)$.

If y_t is stationary, the parameters of the AR model can be estimated consistently by OLS, though the estimates will not be unbiased in small samples; the estimate of ρ will be biased downwards. There are various estimators that correct for this bias, but they have not been widely adopted. A central issue is the treatment of the initial conditions, y_0 in this case.

A p th order (p lags) autoregression, AR(p), takes the form:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + ... + \rho_p y_{t-p} + \varepsilon_t$$

and y_t is stationary if all the roots (solutions) of $1 - \rho_1 z - \rho_2 z^2 - ... - \rho_p z^p = 0$ lie outside the unit circle (are greater than one in absolute value).[16] If a root lies on the unit circle, some $z_i = 1$, the process is said to exhibit a unit root.

[15] The spectral density, used in frequency domain analysis of time series, is a function of the autocovariances.

[16] Some definitions use the inverse of these roots, so the roots must lie within the unit circle.

Consider the case of an AR1 process

$$y_t = \rho y_{t-1} + \varepsilon_t.$$

For stability, the solution to $(1 - \rho z) = 0$ must be greater than unity in absolute value (modulus). The solution is $z = 1/\rho$ and so the condition is $-1 < \rho < 1$. For an AR(2) we have two solutions (which may be complex) for the quadratic equation $(1 - \rho_1 z - \rho_2 z^2) = 0$ and the smallest of them must be greater than unity.

If $\rho = 1$, there is a unit root and the AR(1) becomes a random walk:

$$y_t = y_{t-1} + \varepsilon_t$$

or $\Delta y_t = \varepsilon_t$. Substituting back

$$y_t = y_0 + \varepsilon_t + \varepsilon_{t-1} + ...\varepsilon_1 = y_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j}$$

so shocks have permanent effects; in particular, $V(y_t) = t\sigma^2$. A random walk with drift is $\Delta y_t = \alpha + \varepsilon_t$ so substituting back in the same way

$$y_t = y_0 + t\alpha + \sum_{j=0}^{t-1} \varepsilon_{t-j}.$$

This has a deterministic trend, the second term, and a stochastic trend, the third term.

In both the random walk and random walk with drift cases, Δy_t is stationary, I(0), but y_t is non-stationary, I(1). If there is no drift the expected value of y_t will be constant at zero (assuming $y_0 = 0$), but the variance will increase with t. If there is a drift term we have $E(y_t) = t\alpha + y_0$ and $Var(y_t) = t\sigma^2$ and so the expected value of y_t, as well as the variance, will increase with t. Random walks appear very often in economics, e.g. the efficient market hypothesis implies that asset prices should be random walks. Random walks are a special case of $I(1)$ processes.

A first-order moving average process, MA(1), takes the form

$$y_t = \varepsilon_t + \mu \varepsilon_{t-1}$$

a qth order moving average or MA(q) proces takes the form:

$$y_t = \varepsilon_t + \mu_1 \varepsilon_{t-1} + \mu_2 \varepsilon_{t-2} + ... + \mu_q \varepsilon_{t-q}$$

An MA(q) is a q-memory process in the sense that $cov(y_t, y_{t-i}) = 0$, for $i > q$. A finite order MA process is always stationary. Any stationary process can be represented by a (possibly infinite) moving average process. The latter is known as the *Wold decomposition*.

Combining the AR and MA processes, gives the ARMA process. The first order ARMA(1,1) is

$$y_t = \rho y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

In practice, the data are differenced enough times, say d, to make them stationary and then modelled as an ARMA process of order p and q. This gives the Autoregressive Integrated Moving Average, ARIMA(p, d, q) process. For instance, the ARIMA(1,1,1) process is

$$\Delta y_t = \rho\Delta y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

ARIMA models often describe the univariate dynamics of a single economic time-series quite well and are widely used for forecasting.

4.1.2 Lag Operators

It is often convenient to use the lag operator: $Ly_t = y_{t-1}$; $L^2y_t = y_{t-2}$. $\Delta y_t = y_t - y_{t-1} = (1 - L)y_t$. Using this we can write an ARMA(p,q) process

$$y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}$$

as polynomials in the lag operator:

$$\begin{aligned} (1 - a_1L - ... - a_pL^p)y_t &= (1 + b_1L + ... + b_qL^q)\varepsilon_t \\ A(L)y_t &= B(L)\varepsilon_t. \end{aligned}$$

If an AR process, $A(L)y_t = a_0 + \varepsilon_t$, is stationary we can write it as a moving average process $y_t = A(1)^{-1}a_0 + A(L)^{-1}\varepsilon_t$ an example is (23) above, where $A(1)$ is the sum of the coefficients in $A(L)$. In the example above $A(L)^{-1} = (1 + \rho L + \rho^2L^2 + ...)$, and $A(1)^{-1} = 1/(1 - \rho)$. If a moving average process

$$y_t = B(L)\varepsilon_t,$$

is invertible we can write it as an autoregressive process,

$$B(L)^{-1}y_t = \varepsilon_t.$$

Notice that in an ARMA or ARIMA model, the MA and AR terms may cancel, because of common factors on both sides. Suppose that the true model is a random walk

$$\Delta y_t = \varepsilon_t$$

and we multiply through by $(1 - \rho L)$

$$(1 - \rho L)\Delta y_t = (1 - \rho L)\varepsilon_t.$$

This gives AR and MA terms which are individually significant in

$$\Delta y_t = \rho\Delta y_{t-1} + \varepsilon_t - \rho\varepsilon_{t-1},$$

but, because of the common factor $(1 - \rho L)$, cancel each other out.

4.1.3 Trend-Stationary and Difference-Stationary Variables

Trend-stationary processes, such as

$$y_t = a + \rho y_{t-1} + ct + \varepsilon_t; \quad -1 < \rho_i < 1 \tag{24}$$

with $-1 < \rho_i < 1$ were for long thought to be good descriptions of many economic variables such as (log) real GDP. Substituting for lagged y_t this can be written

$$y_t = \rho^t y_0 + \frac{a}{1-\rho} + \sum_{j=0}^{t-1} \rho^j \varepsilon_{i,t-j} + \frac{c}{1-\rho} t, \tag{25}$$

so the effects of shocks die out. An example of a difference-stationary (DS) process is the random walk with drift, $y_t = \alpha + y_{t-1} + \varepsilon_t$, which substituting for lagged y_t gives

$$y_t = y_0 + \alpha t + \sum_{j=0}^{t-1} \varepsilon_{t-j}$$

the sum of an initial condition, a deterministic trend and a stochastic trend (the accumulation of past errors $S_t = \sum_{j=0}^{t-1} \varepsilon_{t-j}$). Notice that shocks to a DS process are permanent (the effects persist for ever) in contrast shocks to a TS process such as (24) are transitory (the effects die out).

To distinguish between the TS and DS processes, write (24) as

$$\Delta y_t = a + (\rho - 1) y_{t-1} + (\rho - 1)\gamma t + \varepsilon_t,$$

$$\Delta y_t = a + b(y_{t-1} + \gamma t) + \varepsilon_t, \tag{26}$$

where $\gamma = c/(\rho - 1)$. Then if $\rho = 1, (b = 0)$ equation (26) collapses to $\Delta y_t = a + \varepsilon_t$.

One way of testing the unit root hypothesis, $H_0 : \rho_i = 1$, is to compute the t ratio for b in (26). Under H_0 the regression involves an I(0) variable on the left hand side and an I(1) variable on the right hand side, so this t ratio does not have a standard t distribution. The actual distribution was tabulated by Dickey and Fuller and the 5% critical value is about -2.9 without trend and -3.5 with trend. This test is known as the Dickey-Fuller test. Notice the test is written in the restricted form so that if $\rho = 1$, $b = 0$ both the lagged level and the trend drop out. If that restriction is not imposed under the null hypothesis there would be a linear trend in Δy_{it} a quadratic trend in y_{it}.

Regression (26) can be augmented by lagged changes in Δy_{it} to ensure that the error does not suffer serial correlation and this unit root test (based on the t ratio of the coefficient of y_{t-1}) is known as the Augmented Dickey Fuller (ADF) test. An alternative is to remove the serial correlation non-parametrically as in the Phillips-Perron test. For parameteric estimators one has to choose lag length, for non-parametric estimators a bandwidth parameter. There are also tests where H_0 is that the series is I(0) such as the KPSS test (Kwiatkowski et al., 1992) where H_0 is that the variance of the random walk component of the series is zero. The KPSS test is very sensitive to the choice of window or bandwidth size. There are a wide range of tests and some seem to be definitely better than the standard ADF, e.g. the quasi differencing variant of Elliott et al. (1996) and the weighted symmetric variant.[17] However, ADF tests are widely used and we will focus on them.

4.1.4 Explanations for not rejecting a unit root

There are many cases where we cannot reject the hypothesis of a unit root, but from the nature of the variable it seems unlikely that they are really $I(1)$. For instance, data on the real interest rate suggests that it has taken similar values for the last 3000 years, which is inconsistent with it being a random walk; however, it is difficult to reject a unit root in recent time series. Another instance is that after almost thirty years of testing there is still no agreement as to whether the (log) US GNP or real exchange rates have unit roots or not. Many years ago Christiano. & Eichenbaum, (1990) asked. "Unit roots in real GNP: Do we know, and do we care? The answer to both questions being no. There are a number of possible reasons for non-rejection.

1. All these tests have very low power in the near unit root case i.e. a low probability of rejecting the unit root null when in fact the process is stationary but with a coefficient close to unity.

2. Inference is sensitive to treatment of possible serial correlation in the disturbance (e.g. too few augmentation lags in the ADF equation result in size distortions whereas too many lags produce the correct size at the expense of power) and to treatment of means and trends, see Phillips (2004). KPSS tests are very sensitive to the window size chosen.

3. The presence of shifts in mean or trend will tend to lead to non-rejection of the unit root null, even if the process is stationary around the shifting mean or trend. Distinguishing structural changes, large infrequent permanent innovations, from unit roots, small frequent permanent innovations is difficult. Perron (2006) discusses the difficulties.

4. The power depends on the data span (number of years) not on the number of observations: long span time-series (over a century) tend to look I(0), short ones (over decades) I(1).

5. There are problems of aggregation, both over components and over time. Even if all components adjust rapidly, the total may show considerable persistence. If y_t is an aggregate $y_t = \sum_{i=1}^{N} y_{it}$ where $\Delta y_{it} = \phi_i y_{it-1} + \varepsilon_{it}$, the estimate from the aggregate data $\widehat{\phi}$, i.e. $\Delta y_t = \phi y_{t-1} + \varepsilon_t$ can be much closer to zero than the average of the $\widehat{\phi}_i$ (see Taylor, 2001).

6. There may be non-linearities. The series may look I(1) because the underlying process is a random walk within a band, but returns to the band

[17] See Phillips and Xiao (1998) for an introduction.

very rapidly if it strays beyond the band, Kapetanios et al. (2003) provide a simple test for cases such as these assuming the process has non-linear mean reversion of the form

$$\Delta y_t = \phi y_{t-1} + \gamma y_{t-1}\left[1 - \exp(-\theta y_{t-1}^2)\right]$$

for a mean zero process. When $y_{t-1} = 0$ the second term equals zero and ϕ is small, but as y_{t-1} becomes larger in absolute value it produces stronger mean reversion ($\phi + \gamma$ is the speed of adjustment). For 10 real exchange rates they find that θ is around 0.02. The LM test for such non-linearity is very easy to implement, it just involves including y_{t-1}^3, where y_t is mean zero, in the regression.

4.2 Panel Unit Root Tests

Unit root tests differ in

- whether the null is a unit root, like ADF type tests, or stationarity, like KPSS type tests; and
- whether serial correlation is removed parametrically, like ADF type tests, or non-parametrically like Phillips-Perron tests.

There are also adjustments that improve the power of the tests, e.g. GLS or weighted symmetric versions of the ADF. All these issues will arise in panels, but we focus on ADF type tests. In panels, further issues arise:

- how much heterogeneity across units is allowed under the alternative (it is homogeneous under the unit root null);
- how do you interpret the null and alternatives (e.g. rejecting the hypothesis that they are all I(1) does not imply that they are all I(0));
- how do you combine the tests for the individual units; and
- how do you control for cross-section dependence.

First generation tests assumed independence across groups, but the tests were very sensitive to the failure of this assumption. Second generation tests allow for cross-section dependence. It is crucial to allow for cross-section dependence when doing panel unit root tests, though we will ignore it for this section to introduce the other issues. Breitung and Pesaran (2008) provide a survey of unit roots and cointegration in panels.

Consider the Dickey Fuller regression:

$$\Delta y_{it} = a_i + (\rho_i - 1)y_{i,t-1} + (\rho_i - 1)\gamma_i t + \varepsilon_{it},$$

or

$$\Delta y_{it} = a_i + b_i y_{i,t-1} + c_i t + \varepsilon_{it} \tag{27}$$

where lagged values of Δy_{it} may also be included to ensure that the disturbances are not serially correlated. The unit root hypothesis, $\rho_i = 1$, implies that $b_i = 0$, for all i. Given T sufficiently large, this can be tested on each group using the t ratio for b_i and the non-standard critical values. But as noted above such tests lack power. The hope of panel unit-root tests is that power can be increased by increasing the sample through use of the cross-section dimension. Breitung and Meyer (1994) and Quah (1994) are early examples.

Levin Lin & Chu (2002), LLC, (based on 1992 work by Levin and Lin) consider a two way fixed effect version of this model, which allows for a flexible common trend, which may pick up some cross-section dependence:

$$\Delta y_{it} = a_i + \alpha_t + by_{i,t-1} + \varepsilon_{it} \tag{28}$$

and devise a test for the null $H_0 : b = 0$, against the alternative $b < 0$. They also allow for serial correlation by augmenting with lagged changes. The assumption of homogeneity under the alternative, that $b_i = b$, is clearly restrictive and subject to the possible heterogeneity bias of the fixed effect estimator discussed above.

Im Pesaran and Shin (2003), IPS, based on 1996 work, allow the b_i to differ under the alternative; under the null they are homgeneous, $b_i = 0$. They use the estimates of (27) directly (augmenting with lagged changes if necessary); calculate the average ADF statistics (t ratios for b_i) and provide simulated test statistics for the mean and variance of the average t ratio, which allows testing of the hypothesis $H_0 : b_i = 0$, for all i. The alternative is that $b_1 < 0, ..., b_K < 0$, $K \leq N$, some subset are stationary, with $K/N \to k$, as $N \to \infty$. Harris et al. (2010) discuss the problem of sensitivity to initial conditions in the IPS test.

The second generation variant, that deals with cross-section dependence uses the Common Correlated Effect, CCE approach Pesaran (2006, 2007) adds means, for instance estimating

$$\Delta y_{it} = a_i + b_i y_{i,t-1} + c_i t + \delta_1 \overline{\Delta y}_t + \overline{y}_{t-1} + \varepsilon_{it}. \tag{29}$$

This assumes a single factor, Pesaran, Smith and Yamagata (2013) allow for a multifactor structure by adding time means of additional variables. Westerlund (2013) shows that adding covariates, additional variables, increases power substantially and outweighs the power loss associated with the incidental trends problem discussed in Moon, Perron and Phillips (2007).

There is a further issue that is apparent in (28). Suppose $b < 0$, and the test correctly rejects the hypothesis $b = 0$. However, suppose α_t is $I(1)$, then the test will have rejected a unit root, whereas every series does in fact have a unit root arising from the common stochastic trend in the mean of the series, $\overline{y}_t$. Thus there is a joint hypothesis: $b < 0$ can arise either if the series are $I(0)$ and α_t is $I(0)$ or if α_t is $I(1)$ and the y_{it} cointegrate with α_t. A similar problem arises with, (29), when time means are included. $\overline{y}_{t-1}$. Kapetanios and Pesaran (2007) and Kapetanios, Pesaran and Yamagata (2008) have a more general discussion of non-stationary models with a multi-factor error structure. De Silva, Hadri and Tremayne (2009) review the finite sample performance of a number of the unit root tests in the presence of cross-section dependence.

Maddala and Wu (1999) agree that a heterogenous alternative is better, but argue that averaging the ADF statistics is not the most efficient way to use the information. They propose a test statistic, based on a suggestion of R.A. Fisher, which is $-2\sum_{i=1}^{N}(\ln p_i)$ where p_i is the p value for the ith test. Under the null of a unit root, this is distributed $\chi^2(2N)$. Their simulations suggest that in a variety of situations the Fisher test is more powerful than the IPS test which is more powerful than the Levin and Lin test, though this depends whether the coefficients are homogeneous or heterogeneous under the alternative.

Before one can apply the individual ADF tests one has to determine the degree of augmentation (the number of lags) and whether a trend should be included. This raises questions about the criterion that should be used (AIC, BIC etc) and the possibility of pre-test biases. In a panel setting, one must determine whether to choose augmentation and trend on a group specific basis or use a common specification for each group. Similar issues arise in choosing window size in non-parameteric tests. None of these tests allow for covariances between groups, though the use of year effect dummies or data in deviations from year means allows for a simple covariance structure.

Shin and Snell (2006) develop a panel KPSS type test, which differs from the standard KPSS test in using a parametric rather than a non-parameteric estimator of the long-run variance. They show that this performs much better and avoids the bandwidth choice problem. They also suggest a procedure which uses both their test of stationarity and the IPS test of non-stationarity to provide evidence on the proportion of cases that are stationary or non-stationary.

One of the reasons researchers were concerned about the presence of unit roots was to avoid the problem of spurious regressions. But in panels, where the spurious regression problem is reduced by averaging, as discussed below, this motivation is less pressing. A similar point applies to testing for cointegration to which we turn in the next chapter. There is an issue about the interpretation of such tests, e.g. Karlsson and Lothgren (2000). The null hypothesis is that $b_i = 0$, for all i. This can be rejected (in sufficiently large samples) if any one of the N coefficients b_i is non-zero. Rejection of the null certainly does not indicate that all the series are stationary, though the way the alternative is written in the Levin-Lin variant may give that impression. If the hypothesis of interest is that all the series are stationary, then the appropriate test would be a panel variant of the KPSS test. But again rejection could reflect the fact that a single series was non-stationary, which may not be interesting. As always with pooling, some judgement on the commonality or degree of similarity between groups is needed. Pesaran, (2012) argues that rejection of the panel unit root hypothesis should be interpreted as evidence that a statistically significant proportion of the units are stationary. Accordingly, in the event of a rejection, and in applications where the time dimension of the panel is relatively large, it recommends the test outcome to be augmented with an estimate of the proportion of the cross-section units for which the individual unit root tests are rejected. The economic importance of the rejection can be measured by the magnitude of this proportion. Westerlund & Breitung (2013) review a number of facts about the IPS and LLC tests and warn against approaching the testing problem from a narrow time-series perspective. Moon & Perron (2012) suggest using the false discovery rate in evaluating the classification of individual series into I(0)/I(1). Procedures which classify units into groups with similar coefficients, like that of Su, Shi and Phillips (2014) can also be used to separate into stationary and unit root groups.

4.3 PPP

Tests for unit roots in the real exchange rate use ADF regressions of the form:

$$\Delta q_{it} = a_{i0} + b_i q_{it-1} + \gamma_i \sum_{j=1}^{k} \Delta q_{it-j} + u_{it}, \quad u_{it} \sim iidN(0, \sigma^2)$$

where the appropriate augmentation order k can be selected by a testing down procedure or using the AIC and BIC.

The focus can be either on estimating b_i, the speed of adjustment or some transformation of it such as the half life, or testing whether $H_0 : b_i = 0$. This distinction is important since the distribution is different under the null and the alternative. One may estimate b_i under the alternative, $H_1 : b_i < 0$ and its confidence interval does not cover zero. But because the distribution is different under the null, one cannot reject the hypothesis $b_i = 0$. Imbs et al. (2005) show that aggregation biases b_i towards zero and that once one allows for heterogeneity and between group dependence one gets much faster adjustment.

Pesaran, Smith, Yamagata and Hvozdyk (2009) approach the question rather differently. Given data on $N + 1$ countries, 18 in our case, the usual procedure, which we have folowed in these notes, is to look at the N real exchange rates against a base country, the US in our case. But it is possible that the real exchange rate between the UK and France is stationary, but both their real rates against the US dollar are non-stationary. PSYH consider all N(N+1)/2 pairs of real exchange rates among the N+1 countries and try to estimate the proportion of pairs of real exchange rates that are stationary using a variety of unit root tests. This proportion can be estimated consistently for large N even if there is between group dependence. They find that the probability of rejection depends on the size of the shocks: where there are large shocks rejection of unit roots is almost universal. Large deviations from equilibrium are necessary to be able to estimate the speed of adjustment back to equilibrium. They also find evidence for non-linear adjustment.

5 Cointegration in Single Equations

5.1 Spurious Regression and cointegration

One reason the order of integration is interesting is that there is a danger of spurious regression if the variables are $I(1)$. Suppose the random walks $\Delta y_{it} =$

ε_{1it} and $\Delta x_{it} = \varepsilon_{2it}$ are unrelated, that is ε_{1it} and ε_{2it} are independent. In a regression of y_{it} on x_{it}, say

$$y_{it} = \theta_{0i} + \theta_i x_{it} + u_{it}, \tag{30}$$

as $T \to \infty$ the OLS estimate of the regression coefficient, $\widehat{\theta}_i$, and its t ratio will not go to zero as they should, but to non-zero random variables and the R^2 of the regression will go to unity and the Durbin-Watson statistic to zero. This problem of spurious regression was pointed out by Granger and Newbold (1974) and Phillips (1986) provides a theoretical treatment. In general, if y_{it} and x_{it} are I(1), any linear combination, say $u_{it} = y_{it} - \theta_i x_{it}$ will also be I(1). Define $\widetilde{y}_{it} = y_{it} - \overline{y}_i$ where $\overline{y}_i = \sum_{i=1}^{T} y_{it}/T$ is the mean for each group, and similarly for $\widetilde{x}_{it}$; the individual OLS estimate is:

$$\widehat{\theta}_i = \frac{\sum_{t=1}^{T} \widetilde{x}_{it}\widetilde{y}_{it}}{\sum_{t=1}^{T} \widetilde{x}_{it}^2} = \theta_i + \frac{\sum_{t=1}^{T} \widetilde{x}_{it} u_{it}}{\sum_{t=1}^{T} \widetilde{x}_{it}^2}$$

The spurious regression problem arises because the noise in the OLS estimate $\widehat{\theta}_i$ caused by the covariance of two I(1) random variables, x_{it} and u_{it} (which does not go to zero as $T \to \infty$, even when suitably scaled), swamps the signal, the true value of θ_i, which in the case of a spurious regression is zero. This problem is much less severe in ECM such as

$$\Delta y_{it} = a_{0i} + b_{0i}\Delta x_{it} + \lambda_i(\theta_i x_{i,t-1} - y_{i,t-1}) + \varepsilon_{1it} \tag{31}$$

or

$$\Delta y_{it} = a_{0i} + b_{0i}\Delta x_{it} + a_{1i} y_{i,t-1} + b_{1i} x_{i,t-1} + \varepsilon_{1it} \tag{32}$$

because this regression (32) nests the true DGP, $\Delta y_{it} = \varepsilon_{1it}$, whereas the levels regression (30) does not.

If there exists a linear combination of the I(1) variables,

$$z_{it} = y_{it} - \theta_i x_{it},$$

which is I(0), then the two variables are said to be cointegrated, Engle and Granger (1987). Cointegration is the condition required for the regression of y_{it} on x_{it}, equation (30), not to be spurious, i.e. for $\hat{\theta}_i$ to be consistent for the true θ_i. The error u_{it} is then I(0) and the noise does not swamp the signal. If they cointegrate, x_{it} and y_{it} have a common stochastic trend which is cancelled out by the linear combination. The parameter vector $(1, -\theta_i)$, is called the cointegrating vector, and the cointegrating relationship, z_{it}, is often interpreted as a disequilibrium measure. Notice we are free to normalise the cointegrating vector, since if z_{it} is I(0), $a_{1i} y_{it-1} + b_{1i} x_{it-1}$ is also I(0). Thus we could also have called the cointegrating vector (a_{1i}, b_{1i}) from the ECM.

The ECM (31) balances if the variables are both I(1), the left hand side Δy_{it} is I(0) and the two I(1) terms y_{it-1} and x_{it-1} on the right hand side form a linear combination, $a_{1i} y_{it-1} + b_{1i} x_{it-1}$, which is I(0); that is, if y_{it} and x_{it}

cointegrate. Notice that if they cointegrate λ_i must be non-zero and positive, and θ_i non zero. This can be tested using the null hypothesis that there is no long-run relationship, i.e. that $a_{1i} = b_{1i} = 0$, which can be tested with an F test. The critical values are non standard, but are provided by Pesaran, Shin and R.J. Smith, (2001). They propose a bounds test that can be used when it is not known whether the variables are I(0), I(1) and cointegrated or purely I(1). Thus if the calculated F statistic is below the I(0) critical value, there is definitely no long-run relationship; if it is above the I(1) critical value, there is definitely a long run-relationship; and it is lies in between it depends on the order or integration of the variables.

If y_{it} and x_{it} are $I(0)$, OLS produces $\sqrt{T}$ consistent estimates of all the parameters. If y_{it} and x_{it} are $I(1)$ and cointegrated, OLS on the ECM produces T consistent (superconsistent) estimates of the long-run parameter θ and $\sqrt{T}$ consistent estimates of the remaining (short-run) parameters. Notice that if we estimate (32), $\hat{a}_{1i}$ and $\hat{b}_{1i}$ will be $\sqrt{T}$ consistent but $\hat{b}_{1i}/\hat{a}_{1i}$ will be T-consistent. The ECM can also allow for different orders of integration. By making $a_{1i} = 0$, for instance, (32) can allow y_{it} to be I(1) and x_{it} to be $I(0)$.

With only two I(1) variables there can only be a single cointegrating vector, but with more than two variables there can be more than one cointegrating vector and any linear combination of these cointegrating vectors will also be a cointegrating vector. The cointegrating vector may also be related to I(0) variables. We deal with multiple cointegrating vectors in the next chapter.

5.2 Approaches to testing for cointegration

If the cointegrating vector is known a priori, as with PPP, we can form the hypothesised I(0) linear combination, the real exchange rate, and use an ADF test to determine whether it is in fact I(0).

If there is a single unknown cointegrating vector, there are a number of routes we can take.

We can use the tests for multiple unknown cointegrating vectors discussed in the next chapter.

If we know which is the dependent variable and know that the regressors are weakly exogenous, we can estimate an ECM model and test for the existence of a long-run relationship, i.e. test the null hypothesis that the levels x_{it-1} and y_{it-1} should not appear in the equation, or equivalently that $\lambda_i = 0$ in (31) above, using the appropriate (non-standard) critical values from Pesaran Shin and Smith (2001). Westerlund (2007) discusses an extension of this approach to panels, though using a different null from PSS (2001).

We can estimate the levels equation (30) by OLS and test whether the residuals are I(1), using an ADF test and the appropriate critical values, which are different from those of the ADF test applied to an ordinary variable. This is the original Engle-Granger (EG) procedure. Although the estimates are 'superconsistent' (converge to their true values at rate T rather than $\sqrt{T}$), equation (30) is clearly misspecified because it omits the dynamics, thus the estimate of $\hat{\theta}$ can be badly biased in small samples, unless the R^2 is close to unity. It can also be sensitive to normalisation: which variable is chosen as the dependent variable in the regression. In using the residuals, it also imposes implicit restrictions on the dynamics

$$\begin{aligned}\Delta\hat{v}_{it} &= b_i\hat{v}_{it-1} + w_{it} \\ \Delta y_{it} - \hat{\theta}\Delta x_{it} &= b_i(y_{it-1} - \hat{\theta}x_{it-1}) + w_{it}\end{aligned}$$

whereas the ECM has more flexible dynamics. Thus the original EG procedure is not recommended in most cases.

Similarly in panel, where it is hypothesised that there is a single cointegrating vector, with known coefficients, panel unit root tests can be applied. Where there is a single cointegrating vector with unknown coefficients but known dependent variable ECM or EG type procedures can be used and one can estimate by OLS either the heterogeneous individual levels equation or the homogeneous pooled levels equation. Panel unit root tests, with either heterogeneous or homogeneous autoregressive parameters, can be applied to the residuals from the EG regressions (but not the ECM regressions); with appropriate adjustments to the critical values. This is essentially the procedure suggested by Pedroni (1999, 2004). Westerlund (2005) suggests an alternative residual based test which does not require an adjustment for serial correlation. Westerlund and Basher (2008) show the results of the cointegration tests are sensitive to whether parameteric and non-parametric adjustment is used. These test the null hypothesis of no cointegration. KPSS like variants, where the null is cointegration, rather than lack of cointegration, are given in McCoskey and Kao (1998) and Hadri (2000). Breitung and Pesaran (2008) provide a detailed survey. We return to testing for cointegration in the next chapter. Wagner and Hlouskova (2010) provide a review and simulation study of a large number of panel cointegration tests and estimators.

5.3 Single Equation Estimation with I(1) Variables

Consider the simple model with $I(1)$ x_{it}

$$\begin{aligned}x_{it} &= \mu + x_{it-1} + v_{it} \\ y_{it} &= \alpha_i + \beta_i x_{it} + u_{it} \\ u_{it} &= \rho_i u_{it-1} + \varepsilon_{it}\end{aligned} \tag{33}$$

There are a number of cases. If $\rho_i = 1$, $I(1)$ errors, there is no cointegration; if $\rho_i < 1$ there is cointegration. If $\rho_i < 1$ and $\beta_i = \beta$ then there is homogeneous cointegration. If $\rho_i < 1$ and $\beta_i \neq \beta$, there is heterogeneous cointegration. If there is heterogeneous cointegration and homogeneity is imposed then what is estimated is

$$y_{it} = \alpha_i + \beta x_{it} + \{(\beta_i - \beta)x_{it} + u_{it}\}.$$

If x_{it} is $I(1)$ the composite error term, in $\{..\}$, will be $I(1)$, there will be no cointegration, despite the fact that every group cointegrates. A possible exception is where each of the x_{it} are driven by a single I(1) common factor. Trapani and Urga (2010) discuss the relation between micro and macro cointegration.

The homogeneous cointegration case corresponds to the common long-run coefficients assumed in the pooled mean group estimator discussed above, with the short-run coefficients differing. The heterogeneous cointegration raises no new issues since a cointegrating model can be estimated for each group and the estimates averaged. Of course if it is known that there is no cointegration (the error u_{it} is I(1), $\rho_i = 1$) the sensible procedure would be to use the first differenced data which will produce $\sqrt{T}$ consistent (individual OLS) estimates of β_i and $\sqrt{NT}$ consistent (pooled OLS) estimates of a common β or of the mean of the β_i (mean group).

There may be good economic explanations for the error term u_{it} being I(1). In PPP the equilibrium real exchange rate is determined by relative productivities through Harrod-Balassa-Samuelson effects and these may be I(1), so exchange rates and prices do not cointegrate. In production functions technical progress may be I(1) so ouput does not cointegrate with labour and capital. In money demand functions the effect of financial innovation may be I(1), so real money demand does not cointegrate with income and interest rates.

Because one may not know whether $\rho_i = 1$ there is some interest in the properties of the level regression where there is no cointegration, either because the errors are $I(1)$ or because homogeneity is wrongly imposed. For a single time-series this would be the case of a spurious regression, where $\widehat{\beta}_i$ converges to a random variable. Pesaran and Smith (1995) noted that the problem of spurious regression does not arise in a cross-section regression of the form

$$\overline{y}_i = \theta_0 + \theta\overline{x}_i + \overline{u}_i$$

even if x_{it} y_{it} and u_{it} all contain unit roots. Under the strong assumptions of random parameters and strictly exogenous x_{it}, they show that $\widehat{\theta}$ consistently (large T large N) estimates the long-run effect of x_{it} on y_{it}. Note that strict exogeneity is needed because by averaging (33) over time we have an error term $\overline{u}_i$ that contains u_{it-1}, u_{it}, u_{it+1} and so forth and likewise for $\overline{x}_i$ so in order to have $cov(\overline{x}_i, \overline{u}_i) = 0$ we need $cov(x_{it}, u_{is}) = 0$ for all t, s. In this case the between regression provides an estimate of the long-run coefficient even if there is no cointegration. Madsen (2005) extends this analysis. Phillips and Moon (1999, 2000) and Kao (1999) show that a similar result holds for the pooled estimator, under weaker assumptions and that the pooled estimator will be more efficient than the cross-section estimator.

5.3.1 Four cases

No relationship, No cointegration Consider a number of cases. First, suppose v_{it} is I(1) and y_{it} and x_{it} are independent random walks, so that the true value of θ is zero. In a single time-series this is the spurious regression case and $\widehat{\theta}_i$ will have a non-zero probability limit as $T \to \infty$, because the noise, $\sum_{t=1}^{T} \widetilde{x}_{it}v_{it}$, swamps the signal. However, in the pooled estimator we are also averaging over i, and this attenuates the noise and allows us to obtain a consistent estimate of the parameter. Thus in the case of a spurious regression, we get a consistent

($T \to \infty$, $N \to \infty$) estimate of the true value of the parameter, zero (Kao, 1999). By using pooled data we can avoid the problem of spurious regression. The conventional standard errors of $\widehat{\theta}$ will be wrong, and the standard t statistics will diverge from zero, but correct standard errors can be calculated.

Some relationship, no cointegration Secondly, suppose y_{it} and x_{it} are not cointegrated, but they are related. This could be either through a homogenous θ or a heterogenous θ_i :

$$y_{it} = \theta_{0i} + \theta_i x_{it} + u_{it},$$

where u_{it} is $I(1)$. Again the averaging over i will attenuate the noise in the relationship and $\widehat{\theta}$, the pooled FE estimator, will consistently estimate that relationship. It is important to be clear what is being estimated here. $\widehat{\theta}^{FE}$ provides a consistent estimate of what Phillips and Moon (1999) call θ^{LRA} the *long-run average* (LRA) regression coefficient the ratio of the average across groups of the long-run covariance between x and y to the average across groups of the long-run variance of the x,:

$$\theta^{LRA} = E(\Omega^i_{yx})(E(\Omega^i_{xx}))^{-1}.$$

Notice that in general, θ^{LRA} will be different from the *average long-run* regression (ALR) coefficient

$$\theta^{ALR} = E(\theta_i) = E(\Omega^i_{yx}(\Omega^i_{xx})^{-1})$$

estimated by the average across groups of the individual regression coefficients $\widehat{\theta}^{MG} = \sum_i \widehat{\theta}_i/N$, since the expected value of a ratio is not equal to the ratio of the expected values. A similar issue arose in section 3.3 with respect to the form of randomness.

The two estimators use the cross-section within-group variation differently, and measure different parameters of interest. The average long-run coefficient $\theta^{ALR} = E(\theta_i)$ represents the average behaviour of the agents. There are some cases where the 'long-run average' and 'average long-run' coefficients coincide, e.g. if the variance of x is the same for each group. The relation between the two estimators is discussed further below. Coakley, Fuertes and Smith (2006) present Monte Carlo evidence of the performance of various panel estimators when the error term is I(1), i.e. there is no cointegration, and show that the variance of the estimators falls with $\sqrt{N}$. Coakley Flood Fuertes and Taylor (2005) use various measures to estimate the panel elasticity of exchange rates to price differences in a PPP context, which is robust to a lack of cointegration.

Heterogeneous cointegration Thirdly, suppose the model is again

$$y_{it} = \theta_{0i} + \theta_i x_{it} + u_{it},$$

but now u_{it} is I(0), there is heterogeneous cointegration within each group. The pooled estimator $\widehat{\theta}$ again provides a consistent estimator of the 'long-run average' regression coefficient' θ^{LRA}. Notice that the error in the pooled regression, which imposes $\theta_i = \theta$ is $v_{it} = u_{it} + (\theta_i - \theta)x_{it}$ which will be I(1) if $\theta_i \neq \theta$ and x_{it} is I(1), but as in the spurious regression case averaging the variance and covariance across groups attenuates the noise. Again, in general, the 'long-run average' regression coefficient is different from the 'average long run coefficient', the expected value of the θ_i the estimated cointegrating coefficient. In contrast to the previous case, $\widehat{\theta}_i$ is a consistent estimator of θ_i as $T \to \infty$. To see the difference between the ALR and LRA, note that the 'long-run average' regression coefficient can be written as a weighted average of the $\widehat{\theta}_i$:

$$\begin{aligned} \widehat{\theta}^{LRA} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}\widetilde{y}_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2} = \sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\widetilde{x}_{it}^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}\left(\frac{\sum_{t=1}^{T}\widetilde{x}_{it}\widetilde{y}_{it}}{\sum_{t=1}^{T}\widetilde{x}_{it}^2}\right) \\ &= \sum_{i=1}^{N}W_i\widehat{\theta}_i = \sum_{i=1}^{N}W_i\theta_i + \sum_{i=1}^{N}W_i\left\{\frac{\sum_{t=1}^{T}\widetilde{x}_{it}u_{it}}{\sum_{t=1}^{T}\widetilde{x}_{it}^2}\right\} \end{aligned}$$

In the case where there is a cointegrating relationship, the component in curly brackets goes to zero as $T \to \infty$. Let $\theta = E(\theta_i)$, then since $\sum_{i=1}^{N} W_i = 1$

$$E(\widehat{\theta}^{LRA}) = \theta^{LRA} = \theta + E(\sum_{i=1}^{N} W_i(\hat{\theta}_i - \theta)).$$

Thus the expected value of $\widehat{\theta}^{LRA}$ will only be the same as the expected value of $\hat{\theta}_i$ where the weights, the share of a groups variance in the total variance of x_{it}, $(\Omega_{xx}^i / \sum_i \Omega_{xx}^i)$ are uncorrelated with the θ_i. Informally, one can think of this as independence of the θ_i from the x_{it}, though the conditions Phillips and Moon use to establish that the relevant limits exist involve different assumptions about which parameters are random. Clearly if there is heterogeneous cointegration it is much better to estimate the individual cointegrating regressions and average, since this estimate is $T\sqrt{N}$ consistent, rather than pooling to get a $\sqrt{N}$ consistent estimator of a different parameter.

Homogeneous cointegration Fourthly, suppose that there is a common cointegrating parameter θ, then the FE estimator of $\widehat{\theta}$ will consistently estimate it for large T and N. Phillips and Moon (1999) propose a Fully Modified (FM) pooled estimator of the long-run parameter for this case to control for the biases induced by endogeneity and serial correlation. This case is the same as the PMG considered in Pesaran, Shin and R.P. Smith (1999) for cointegrating I(1) variables with common long-run parameters, but heterogeneous short-run parameters. As discussed above in ECM form the PMG model is

$$\begin{aligned} \Delta y_{it} &= \alpha_i + \beta_i \Delta x_{it} + \lambda_i(\theta x_{it-1} - y_{it-1}) + u_{it}; \\ u_{it} &\sim IN(0, \sigma_i^2) \end{aligned}$$

where the short-run parameters $\alpha_i, \beta_i, \lambda_i$ and the variances σ_i^2 differ over groups, but the long-run parameter θ does not. This method can allow the dynamics to differ between groups (i.e. different lag lengths); can estimate the long-run effects even when individual θ_i cannot be identified because of exact multicollinearity between the elements of x in particular groups and can use very small samples for some groups. The difference from the FM estimator is that the pooled mean group estimator has parametric short-run dynamics, rather than relying on non-parametric methods to remove the effects of the dynamics (and any endogeneity) in v_{it} on the long-run coefficients as Phillips and Moon do. In many cases the parameters of interest are those of the short run dynamic processes, e.g. speeds of adjustment or rates of convergence, and FM estimators do not provide estimates of those short-run parameters. Because the short-run parameters are heterogeneous, the PMG estimator allows a mix of cointegration $\lambda_i > 0$, and non-cointegration $\lambda_i = 0$. Westerlund & Hess (2011) suggest a poolability test for cointegrated panels, which uses a Hausman type test.

6 Cointegrating VARs

6.1 Vector Autoregressions

The most common framework used to examine multiple cointegrating vectors is the Vector Autoregression, VAR. Canova & Ciccarelli (2013) provide a survey of panel VARs. But as before we start with reviewing the issues for single time-sseries.

If we have two I(1) variables, there can only be a single cointegrating vector. But with more than two variables, there may be more than one cointegrating vector. In the case of PPP, where the variables are logs of prices, p_{it} and spot exchange rates, s_{it} and short, is_{it}, and long il_{it} interest rates, , $d_{it} = p_{it} - p_t^*$, , is_t^*, il_{it}, il_t^*, $\Delta p_{it}, \Delta p_t^*$, there are a large number of possibly stationary linear combinations given by the various parity conditions. PPP says real exchange rates: $s_{it} - p_{it} + p_{it}^*$, are I(0) and price levels might be I(2), but differentials $p_{it} - p_t^*$ I(1) which cointegrate with I(1) spot exchange rates. Real interest rates $is_{it} - \Delta p_{it}$, or their differentials might be I(0), etc.

For exposition, consider a VAR2, (of order two, with two lags) :

$$\mathbf{y}_t = \mathbf{a} + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\Sigma})$$

where $\mathbf{y}_t$ is now a $m \times 1$ vector, $\mathbf{a}$ a $m \times 1$ vector, $\mathbf{A}_1$ and $\mathbf{A}_2$ are $m \times m$ matrices and $\boldsymbol{\Sigma}$ is a $m \times m$ matrix with elements σ_{ij}.

For $m = 2$, $\mathbf{y}_t = (y_{1t}, y_{2t})'$ the VAR is:

$$\begin{aligned} y_{1t} &= a_1^0 + a_{11}^1 y_{1t-1} + a_{12}^1 y_{2t-1} + a_{11}^2 y_{1t-2} + a_{12}^2 y_{2t-2} + \varepsilon_{1t}, \\ y_{2t} &= a_2^0 + a_{21}^1 y_{1t-1} + a_{22}^1 y_{2t-1} + a_{21}^2 y_{1t-2} + a_{22}^2 y_{2t-2} + \varepsilon_{2t}. \end{aligned}$$

Each equation of the VAR can be estimated consistently by OLS[18] and the Maximum Likelihood estimate of the covariance matrix $\boldsymbol{\Sigma}$ can be obtained from

[18] This is because each equation of the VAR has the same right hand side variables. If the regressors differ there are more efficient systems estimators like SURE.

the OLS residuals,

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t^{'}$$

with elements

$$\widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_{it}\widehat{\varepsilon}_{jt}$$

where $\widehat{\sigma}_{11}$ is the estimated variance of ε_{1t}, $\widehat{\sigma}_{12}$ the estimated covariance of ε_{1t} and ε_{2t}. The Maximised log-likelihood for the system is

$$MLL = -\frac{mT}{2}\ln 2\pi - \frac{mT}{2}\ln \mid \widehat{\mathbf{\Sigma}} \mid$$

and this can be used for LR tests of restrictions on the system, e.g. choosing lag lengths.

A variable y_{2t} is said to Granger cause y_{1t} if knowing current values of y_2 helps you to predict future values of y_1; equivalently, current y_1 is explained by past y_2. In this case, y_2 is Granger causal with respect to y_1 if either a_{12}^1 or a_{12}^2 are non zero. You can test that they are both zero with a standard F or χ^2 tests of linear restrictions if the variables are $I(0)$. The restricted model just excludes y_{2t-1} and y_{2t-2} from the equation for y_{1t}. If the variables are I(1), the usual tests are no longer valid, but Toda and Yamamoto (1995) suggest that the problem can be dealt with by adding extra lags, beyond the optimal number, which you do not use in the tests.[19]

Granger causality is rarely the same as economic causality, particularly because expectations allow consequences to precede their cause: weather forecasts Granger Cause the weather.

The appropriate lag length can be determined by Likelihood Ratio tests or model selection criteria like the AIC or BIC[20], though these can be inconsistent in the presence of incidental parameters in dynamic panels, Moon, Perron and Phillips (2015). VARs can easily use up degrees of freedom. If the lag length is p, each equation of a VAR with m variables and intercept has $1+mp$ parameters. This can get large, 4 lags in a 4 variable VAR gives 17 parameters including a constant in each equation.

A pth order VARp, p lags,

$$\mathbf{y}_t = \mathbf{a} + \sum_{i=1}^{p}\mathbf{A}_i\mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t$$

is stable if all the roots of the determinantal equation $\mid \mathbf{I} - \mathbf{A}_1\mathbf{z} - ... - \mathbf{A}_p\mathbf{z}^p \mid = 0$ lie outside the unit circle. If there are unit roots, stochastic trends, some roots will lie on the unit circle.

[19] Dave Giles blog has a good discussion of this and many other topics: http://davegiles.blogspot.ca/2011/04/testing-for-granger-causality.html

[20] In testing for lag length, use the same sample for the restricted and unrestricted model; i.e. do not use the extra observation that becomes available when you shorten the lag length.

6.1.1 The VAR as reduced form of SEM and solution to a RE model

VARs arise naturally from standard economic models. Consider the structural form of a standard simultaneous equation model, SEM, with no exogenous variables, but predetermined variables playing the same role and supressing the intercept

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{u}_t \tag{34}$$

with

$$E(\mathbf{u}_t\mathbf{u}_t^{'}) = \mathbf{\Omega} \tag{35}$$

the reduced form is the VAR

$$\begin{aligned} \mathbf{y}_t &= \mathbf{B}_0^{-1}\mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_0^{-1}\mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{A}_1\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \\ E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') &= \mathbf{B}_0^{-1}\mathbf{\Omega}\mathbf{B}_0^{-1'} = \mathbf{\Sigma} \end{aligned}$$

The reduced form has $m+m^2$ regression parameters plus $m(m+1)/2$ independent covariance matrix parameters. These can be estimated consistently. However the structural form has $m+2m^2$ regression parameters plus $m(m+1)/2$ independent covariance matrix parameters. To obtain the structural form parameters from the VAR requires m^2 just identifying restrictions which specify the parameter values on the basis of economic theory, of which m will be provided by normalisation, dependent variables are given coefficients of unity.

Alternatively note that

$$\mathbf{A}_1 = \mathbf{B}_0^{-1}\mathbf{B}_1 = \mathbf{B}_0^{-1}\mathbf{P}^{-1}\mathbf{P}\mathbf{B}_1 = \widetilde{\mathbf{B}}_0^{-1}\widetilde{\mathbf{B}}_1,$$

for any non-singular $m \times m$ matrix $\mathbf{P}$ so $\mathbf{B}_i$ and $\widetilde{\mathbf{B}}_i$ are observationally equivalent, both consistent with the estimated $\mathbf{A}_1$. Identification can be thought of as the choice of the m^2 elements of $\mathbf{P}$.

The identifying restrictions can be either on the coefficients $\mathbf{B}_0, \mathbf{b}, \mathbf{B}_1$ or the covariance matrix $\mathbf{\Omega}$. If there are enough restrictions (e.g. that certain elements of $\mathbf{B}_0, \mathbf{B}_1$ are zero; exclusion restrictions) the system can be estimated, by for instance, two-stage least squares. Just identifying restrictions are not testable. However, if there are more restrictions than needed to identify the system, these over-identifying restrictions can be tested.

A special case of an identified system is a recursive structure. Suppose the variables can be ordered so that $\mathbf{B}_0$ is lower triangular, the elements above the diagonal are zero. This provides $m(m-1)$ restrictions. Suppose also that $\mathbf{\Omega}$ is diagonal, all the off-diagonal elements are zero, this provides another $m(m-1)$ restrictions. Finally there are m normalisation restrictions, the coefficients of the dependent variables are unity. Such recursive systems are not only identified (since there are m^2 restrictions) each equation can be estimated by OLS because the errors are not correlated with the right hand side endogenous variables. These are the assumptions (imposed by specifying the order of the variables in the Choleski decomposition) used in calculating orthogonalised impulse response functions for VARS. Generalised impulse response functions do not require identifying restrictions, but cannot be given a structural interpretation.

A similar analysis can be applied to Rational Expectations, RE, models. Suppose we have

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2E_t(\mathbf{y}_{t+1}) + \mathbf{u}_t$$

where $E_t(\mathbf{y}_{t+1})$ is the rational expectation of $\mathbf{y}_{t+1}$, given information at t. The solution is given by

$$\begin{aligned}\mathbf{y}_t &= \mathbf{C}\mathbf{y}_{t-1} + \mathbf{B}_0^{-1}\mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{A}_1\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t\end{aligned}$$

where $\mathbf{C}$ solves $\mathbf{B}_2\mathbf{C}^2 - \mathbf{B}_0C + \mathbf{B}_1 = 0$. The solution is unique and stationary if all the eigenvalues of $\mathbf{C}$ and $(\mathbf{I} - \mathbf{B}_2\mathbf{C})^{-1}\mathbf{B}_2$ lie strictly inside the unit circle. Notice the solution is the same as the same reduced form of the SEM. In this case the order condition is that we require $2m^2$ restrictions, rather than m^2 in the case of the SEM.

6.1.2 Panel VARs

Canova & Ciccarelli (2013) provide a survey of panel VARs. Suppose the model is an unrestricted VAR (which is the unrestricted form of the cointegrating VECM discussed below) and we assume homogeneity of slopes across cross section units. This is very easy to estimate, since each equation of the VAR has standard FE form

$$y_{it} = \alpha_i + \beta'\mathbf{x}_{it} + u_{it}$$

$\mathbf{x}_{it}$ is a mp vector of lags on the m endogenous variables. It can be estimated in the usual FE way by OLS on the NT observations, $y_{it} - \overline{y}_i$, as long as T is large relative to N. Dhaene and Jochmans (2015) suggest a bias correction for cases where T is large but of the same order as N. As with the dynamic fixed effect as well as the Nickel downward bias there is also a possible upward heterogeneity bias.

Aksoy et al (2015) use a panel VARX, VAR with exogenous variables, to examine the effect of demography on macroeconomic trends. They have $N = 20$, OECD countries, $T = 1970 - 2007$. The endogenous vector Y_{it} includes the growth rate of the real GDP, $g_{i,t}$, the share of investment in GDP, $I_{i,t}$, the share of personal savings in GDP, $S_{i,t}$, the logarithms of hours worked $H_{i,t}$, the real short interest rate, $R_{i,t}$ and the rate of inflation $\pi_{i,t}$. $Y_{i,t} = (g_{i,t}, I_{i,t}, S_{i,t}, H_{i,t}, rr_{i,t}, \pi_{i,t})'$. for countries $i = 1, 2, ..., N$.

The exogenous variables are $W_{i,t}$ a vector of shares of the population in 7 age groups, allowing for the fact that the shares add up to one.

Assuming slope homogeneity across countries but allow for intercept heterogeneity through a_i they estimate a one way fixed effect panel VARX(1) of the form:

$$Y_{it} = a_i + AY_{i,t-1} + DW_{i,t} + u_{it},$$

plus two other controls: lagged oil price and population growth, assumed exogenous.

A key assumption is that the demographic structure is exogenous. To check this they run a VAR with both vectors Y_{it} and W_{it} treated as endogenous and estimate

$$\begin{bmatrix} Y_{it} \\ W_{it} \end{bmatrix} = a_i + \begin{bmatrix} \tilde{A} & \tilde{D} \\ B_1 & B_2 \end{bmatrix} \begin{bmatrix} Y_{i,t-1} \\ W_{i,t-1} \end{bmatrix} + u_{it}$$

and verify whether matrix B_1 is equal to zero. Although some parameters in B_1 are found to be singificantly different then zero they are all small.

Having estimated a_i, A and D, the long-run moving equilibrium for the system (ignoring population growth and oil prices) is then given by

$$Y_{it}^* = (I - A)^{-1} a_i + (I - A)^{-1} DW_{it},$$

where the effect of the demographic variables is given by $D_{LR} = (I - A)^{-1} D$, which reflects both the direct effect of demographics on each variable and the feedback between the endogenous variables.

Like ordinary VARs, panel VARs can be used to estimate impulse response functions (IRF) that show how the variables respond to shocks, using the moving average representation.

6.2 Cointegration

We can reparameterise the VAR(2):

$$\mathbf{y}_t = \mathbf{a} + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t$$

as the Vector Error Correction Model, VECM:

$$\begin{aligned} \mathbf{y}_t - \mathbf{y}_{t-1} &= \mathbf{a} + (\mathbf{A}_1 + \mathbf{A}_2 - \mathbf{I})\mathbf{y}_{t-1} - \mathbf{A}_2(\mathbf{y}_{t-1} - \mathbf{y}_{t-2}) + \boldsymbol{\varepsilon}_t \\ \Delta\mathbf{y}_t &= \mathbf{a} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\Gamma}\Delta\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \end{aligned}$$

and the VAR(p) as :

$$\Delta\mathbf{y}_t = \mathbf{a} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta\mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t.$$

Notice that this is the vector equivalent of the ADF regression that we used above to test for unit roots and it may include trends, which may be restricted in the same way.

If all the variables, the m elements of $\mathbf{y}_t$, are I(0), $\boldsymbol{\Pi}$ is a full rank matrix. If all the variables are I(1) and not cointegrated, $\boldsymbol{\Pi} = \mathbf{0}$, and a VAR in first differences is appropriate. If the variables are I(1) and cointegrated, with r cointegrating vectors, then $\boldsymbol{\Pi}$ has rank r. In this case, there are r cointegrating relations, combinations of $\mathbf{y}_t$ that are I(0),

$$\mathbf{z}_t = \boldsymbol{\beta}'\mathbf{y}_t$$

where $\mathbf{z}_t$ is a $r \times 1$ vector and $\boldsymbol{\beta}'$ is a $r \times m$ matrix. Then we can write the model as:

$$\begin{aligned}
\Delta \mathbf{y}_t &= \mathbf{a} + \mathbf{\Pi} \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \\
\Delta \mathbf{y}_t &= \mathbf{a} + \boldsymbol{\alpha} \mathbf{z}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \\
\Delta \mathbf{y}_t &= \mathbf{a} + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t.
\end{aligned}$$

The I(0) dependent variable is explained by the r I(0) variables $\mathbf{z}_{t-1}$ and the I(0) lagged changes; $\boldsymbol{\alpha}$ is a $m \times r$ matrix of 'adjustment coefficients' which measure how the deviations from equilibrium, $\mathbf{z}_{t-1}$, feed back on the changes; and $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$ has rank $r < m$ if there are r cointegrating vectors.

If there are $r < m$ cointegrating vectors, then $\mathbf{y}_t$ will also be determined by $m - r$ stochastic trends. If there is cointegration, some of the $\boldsymbol{\alpha}$ must be non-zero, there must be some feedback on the $\mathbf{y}_t$ to keep them from diverging, i.e. there must be some Granger causality in the system.

If there are r cointegrating vectors and $\mathbf{\Pi}$ has rank r, it will have r non-zero eigenvalues and Johansen (1988) provided a way of estimating the eigenvalues and two tests for determining cointegrating rank, the trace and the maximal eigenvalue tests. The trace tests the null of say $r \leq 1$ against the alternative of $r \geq 2$, the maximum eigenvalue tests the null of $r \leq 1$ against the alternative of $r = 2$. These allow us to determine r, though the two tests may give different answers. The trace test is probably better. The Johansen estimates of the cointegrating vectors β are the associated eigenvectors.

There is an identification problem with the cointegrating vectors, similar to the problem in the SEM discussed above. Since

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}' = (\boldsymbol{\alpha}\mathbf{P}^{-1})(\mathbf{P}\boldsymbol{\beta}) = \widetilde{\alpha}\widetilde{\beta}'$$

the α and β are not uniquely determined. We can always choose a non-singular $r \times r$ matrix P such that $\boldsymbol{\alpha}\boldsymbol{\beta}'$ and the new estimates $\widetilde{\alpha}\widetilde{\beta}'$ are observationally equivalent (consistent with the same estimated $\mathbf{\Pi}$), though they might have very different economic interpretations. Put differently, if $\mathbf{z}_{t-1} = \boldsymbol{\beta}'\mathbf{y}_{t-1}$ are I(0) so are $\mathbf{z}^*_{t-1} = \mathbf{P}\boldsymbol{\beta}'\mathbf{y}_{t-1}$, since any linear combination of I(0) variables is I(0). We need to choose the appropriate $\mathbf{P}$ matrix to allow us to interpret the estimates. This requires r^2 restrictions, r on each cointegrating vector. One of these is provided by normalisation, we set the coefficient of the 'dependent variable' to unity, so if $r = 1$ this is straightforward (though it requires the coefficient set to unity to be non-zero). Unlike The Engle-Granger estimates the Johansen estimates are invariant to the normalisation chosen. If there is more than one cointegrating vector it requires prior economic assumptions. The Johansen identification assumption, that the β are eigenvectors with unit length and orthogonal, do not allow an economic interpretation. However, they can be used for forecasting or counterfactual exercises which are invariant to the just-identifying restrictions used.

As we saw above with the Dickey Fuller regression, there is also a problem with the treatment of the deterministic elements. If we have a linear trend in the VAR, and do not restrict the trends, the variables will be determined by $m-r$ quadratic trends. To avoid this (economic variables tend to show linear not quadratic trends), we enter the trends in the cointegrating vectors. Most programs give you a choice of how you enter trends and intercepts; unrestricted intercepts and restricted trends, option 4 in many programs, is a good choice for trended economic data. We return to the treatment of the deterministic elements below.

6.3 Relation between VECM and ECM

Suppose that we have the bivariate VAR1

$$\begin{aligned} y_t &= a_{10} + a_{11} y_{t-1} + a_{12} x_{t-1} + \varepsilon_{1t} \\ x_t &= a_{20} + a_{21} y_{t-1} + a_{22} x_{t-1} + \varepsilon_{2t} \end{aligned}$$

with $E(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij}$, $i, j = 1, 2$. If y_t and x_t cointegrate, then we can write it

$$\begin{aligned} \Delta y_t &= a_{10} + \alpha_1 (y_{t-1} - \beta x_{t-1}) + \varepsilon_{1t} \\ \Delta x_t &= a_{20} + \alpha_2 (y_{t-1} - \beta x_{t-1}) + \varepsilon_{2t} \end{aligned}$$

with

$$\mathbf{\Pi} = \begin{bmatrix} \alpha_1 & -\alpha_1\beta \\ \alpha_2 & -\alpha_2\beta \end{bmatrix}$$

which clearly has rank one. If $\alpha_2 = 0$, then x_t is weakly exogenous for β because there is no information in the marginal distribution of x_t (the equation determining Δx_t) which helps us determine β. However, if $\sigma_{12} \neq 0$, there is information in Δx_t which helps explain Δy_t, since then

$$\begin{aligned} E(\varepsilon_{1t}|\varepsilon_{2t}) &= \sigma_{12}\sigma_{22}^{-1}\varepsilon_{2t} \\ &= \sigma_{12}\sigma_{22}^{-1}(\Delta x_t - a_{20}) \end{aligned}$$

so

$$\Delta y_t = \left(a_{10} - \sigma_{12}\sigma_{22}^{-1}a_{20}\right) + \sigma_{12}\sigma_{22}^{-1}\Delta x_t + a_1 (y_{t-1} - \beta x_{t-1}) + u_t$$

$$\Delta y_t = a_0 + b_0 \Delta x_t + \alpha(y_{t-1} - \beta x_{i,t-1}) + u_t \tag{36}$$

an ECM. If $\sigma_{12} \neq 0$, this will fit better than the VAR equation for Δy_t since $E(u_t^2) = \sigma_{11} - \left(\sigma_{12}\sigma_{22}^{-1}\right)^2 \sigma_{22} = \sigma_{11} - \sigma_{12}^2\sigma_{22}^{-1} < \sigma_{11}$.

6.4 Panel cointegration

The Johansen procedure for estimating cointegrating VARS has become standard in time-series, though there are a number of issues in its application in panels. Suppose $\mathbf{y}_{it}$ is a $m_i \times 1$ vector of $I(1)$ variables, with typical element y_{kit}; $k = 1, ..., K$. Unlike the $N = 1$ case there may be cointegration both between different variables within the same group, between the same variable in different groups; and between different variables in different groups, y_{kit} and y_{sjt}, Banerjee et al. (2004). Log income in the UK may be cointegrated with both log consumption in the UK, log income in the US and real interest rates in the US.We will return to this issue below, but let us begin by assuming that between group cointegration does not arise and consider other issues.

Suppose that the variables are $I(1)$ and cointegrated, with r_i cointegrating vectors, combinations of $\mathbf{y}_{it}$ that are I(0),

$$\mathbf{z}_{it} = \boldsymbol{\beta}_i' \mathbf{y}_{it}$$

where $\mathbf{z}_{it}$ is a $r_i \times 1$ vector and β_i' is a $r_i \times m_i$ matrix. Starting from the structural form of the SEM:

$$B_{0i} \Delta \mathbf{y}_{it} = \boldsymbol{\mu}_{i0} + \boldsymbol{\mu}_{i1} t + \Lambda_i \mathbf{z}_{i,t-1} + \sum_{l=1}^{p_i-1} \Xi_l \Delta \mathbf{y}_{i,t-l} + \mathbf{u}_{it}$$

with $E(u_{it}) = 0$, $E(u_{it} u_{it}') = \Sigma$. The reduced form is

$$\Delta \mathbf{y}_{it} = \kappa_{i0} + \kappa_{i1} t + \alpha_i \mathbf{z}_{i,t-1} + \sum_{l=1}^{p_i-1} \Gamma_l \Delta \mathbf{y}_{i,t-l} + \boldsymbol{\varepsilon}_{it},$$

Where $K_{i0} = B_{0i}^{-1} \mu_{i0}$, $\boldsymbol{\alpha}_i = B_{0i}^{-1} \Lambda_i$, $\boldsymbol{\varepsilon}_{it} = B_{0i}^{-1} \mathbf{u}_{it}$, $\Gamma_l = B_{0i}^{-1} \Gamma_l$, $E(\boldsymbol{\varepsilon}_{it} \boldsymbol{\varepsilon}_{it}') = \Omega_i = B_{0i}^{-1} \Sigma B_{0i}^{-1\prime}$. Notice that even if Λ_i has a simple structure, $\alpha_i = B_{0i}^{-1} \Lambda_i$ will not. The SEM identification problem is to provide the m_i^2 just-identifying restrictions (m_i of which will be provided by normalisation restrictions) in order to determine B_{0i}.

The reduced form can also be written:

$$\Delta \mathbf{y}_{it} = \boldsymbol{\kappa}_{i0} + \boldsymbol{\kappa}_{i1} t + \alpha_i \boldsymbol{\beta}_i' \mathbf{y}_{i,t-1} + \sum_{l=1}^{p_i-1} \Gamma_l \Delta \mathbf{y}_{i,t-l} + \boldsymbol{\varepsilon}_{it}.$$

The cointegration identification problem is to provide the r_i^2 just identifying restrictions (r_i of which are provided by normalisations) to identify $\boldsymbol{\beta}_i$. The intercept and trend have been left unrestricted in this model, this will produce $m_i - r_i$ quadratic trends in the data. To avoid this the trends (and if required the intercepts) can be restricted to lie within the cointegration space by specifying the model

$$\Delta \mathbf{y}_{it} = \boldsymbol{\kappa}_{i0} + \alpha_i (\boldsymbol{\beta}_i' \mathbf{y}_{i,t-1} + \kappa_{i1}^* t) + \sum_{l=1}^{p_i-1} \Gamma_l \Delta \mathbf{y}_{i,t-l} + \boldsymbol{\varepsilon}_{it}.$$

6.4.1 Within country cointegration

Assuming no between country cointegration, cointegration analysis requires determining for each country:

- The m_i endogenous variables to be included in the analysis and any exogenous variables;
- The order of integration of the variables;
- The lag order p_i;
- The treatment of deterministic elements, κ_{ij}, restrictions on intercept and trend;
- The number of cointegrating vectors r_i;
- The treatment of any cross-section dependence;
- The r_i^2 long-run just-identifying restrictions;
- The m_i^2 short-run just-identifying restrictions if one intends to do structural analysis.

These choices interact, tests for r_i are sensitive to choice of p_i for instance. It is not obvious in what order the choices should be made, e.g. whether it is worth pre-testing for the order of integration of each variable, or let it be revealed by the Johansen tests. Some would argue that it is better just to use the unrestricted VAR in levels, which although less efficient does not risk the biases resulting from imposing inappropriate restrictions. These choices are difficult enough even when considering a single country, they multiply in panel.

Suppose for simplicity $m_i = m$, and the same variables are used for each country, a restricted model which may be of interest in certain circumstances is $r_i = r$, $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, with the same just identifying restrictions for each unit. If β is known a priori this is straighforward to implement, since z_{it} can be constructed and its lag included in the VECM. For a known r and $\boldsymbol{\beta}$, this can be easily tested on each country and panel version of the tests applied, e.g. the Fisher type test $-2\sum_{i=1}^{N}(\ln\, p_i)$ where p_i is the p value for the ith test. Johansen estimation of a common $\boldsymbol{\beta}$ is also straightforward. Suppose we assume unrestricted intercepts and trend. For each country we calculate the vectors of residuals r_{0it} from regressing $\Delta\mathbf{y}_{it}$ on $(1, t, \Delta\mathbf{y}_{it-1}, .., \Delta\mathbf{y}_{it-p})$ and r_{1it} from regressing $\mathbf{y}_{it-1}$ on $(1, t, \Delta\mathbf{y}_{it-1}, .., \Delta\mathbf{y}_{it-p})$. We then stack the NT residuals to form:

$$S_{KH} = (NT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} r_{Kit}r_{Hit}, \quad K, H = 0, 1$$

and use the standard Johansen estimator of β e.g. as in Maddala and Kim (1998). This is based on the eigenvalues and eigenvectors of

$$S_{00}^{-1}S_{01}S_{11}^{-1}S_{10}.$$

This allows all the short-run parameters to be different but the long-run cointegrating vectors to be the same. Thus it is an extension to systems of the single equation Pooled Mean Group procedure discussed above. However, heterogeneous cointegration seems a more promising route and will be pursued below.

6.5 Testing for cointegration in panels

As noted above Unit root (cointegration) tests differ in

- whether to test the null of a unit root (no cointegration) against a stationary alternative (cointegration) as with ADF (Johansen) tests do or test a null of stationarity (cointegration) against an alternative of a unit root (no cointegration), as with KPSS type tests?
- whether serial correlation is removed parametrically, as with ADF (Johansen) procedures by using lagged changes, or non-parametrically like Phillips-Perron tests and fully modified estimators?

In panels, further issues arise:

- how much heterogeneity to allow across units, including in number of cointegrating vectors etc?
- how to combine the statistics for different units?
- how to interpret the null and alternatives? e.g. rejecting the hypothesis that they are all do not cointegrate does not imply that they all cointegrate;
- how to control for cross-section dependence, in second generation [procedures?.

Most of the tests discussed below assume that T is large. Binder, Hsiao and Pesaran (2005) examine the small T case. They provide tests for unit roots and cointegrating rank for a dynamic FE model (i.e. with slope homogeneity) using a maximum likelihood estimator which is consistent for large N whether the underlying series are I(0) or I(1).

Rather than using unit root tests on the residuals as discussed in the single equation case, one could use Johansen tests for each group. Larsson, Lyhagen and Lothgren (2001) suggest using the average of the Johansen trace statistics. Rather than use IPS like averages one could also use Maddala and Wu type Fisher combinations. Gengenbach, Urbain & Westerlund (2009) provide a second generation panel test for a single cointegrating vector in an error correction regression for each unit with a global factor approximated by time means.

Heterogeneity in panel cointegration tests raise even more difficulties of interpretation than in panel unit root tests, where there was a single parameter than could be heterogeneous. For a vector of data y_{it}, with r_i I(0) cointegrating relationships $z_{it} = \beta_i' y_{it}$, there are a variety of possible hypotheses. One could test for the same number of cointegrating vectors, $r_i = r$, in each group; that there are at least $r_{\min}$ cointegrating vectors or that there are identical cointegrating vectors, β in each group. As with panel unit root tests it is not clear how rejection should be interpreted. Pre-testing problems in cointegrating models are more severe than in the unit root tests, where the issues are just lag lengths and treatment of the deterministic elements.

In panels, cointegration between cross-section units becomes a possibility, Normally, in time series one considers within unit cointegration, i.e. investigates whether there is a linear combination of the data for a particular unit $z_{it} = \beta_i' y_{it}$ which is $I(0)$. But in many economic examples it is equally or more plausible that there are linear combinations across units, e.g. linear combinations of y_{it} and y_{jt} which are $I(0)$. Between unit cointegration can distort the results of within unit cointegration tests. When there is both between and within cointegration we need to impose more structure on the problem an issue to which we return when discussing GVARs.

Gutierrez (2003) reviews the power of panel cointegration tests. Which test performs best depends on T. This is not surprising since the tests differ in the amount of heterogeneity they allow and the optimal amount of heterogeneity will depend on the length of the time-series available to estimate the parameters. Wagner and Hlouskova (2010) consider the performance of a variety of panel cointegration tests and estimators using a large simulation study.

7 Cross-section Dependence and Factor Models

7.1 Introduction

Cross-section dependence, CSD, seems pervasive in panels, it seems rare that the covariance of the errors is zero: $E(e_{it}e_{jt}) = 0$. In recent years there has been much progress in characterising and modelling CSD. Phillips and Sul (2003) note the consequences of ignoring CSD can be serious: pooling may provide little gain in efficiency over single equation estimation; estimates may be badly biased and tests for unit roots and cointegration may be misleading. CSD has always been central in spatial econometrics (discussed by Pesaran, 2015) where there is a natural way to characterise dependence in terms of distance, but for most economic problems there is no obvious distance measure. For instance, trade between countries reflects not just geographical distance, but transport costs (transport by sea may be cheaper than by land), common language, policy and historical factors, such as colonial links. For large T, it is straightforward to test for cross-section dependence either using the squared correlations between the residuals, the Breusch-Pagan variant, or the correlations themselves. Pesaran, Ullah and Yamagata (2007) survey the various tests and propose new ones. Sarafidis, Yamagata and Robertson (2009) suggest a test for the case where N is large relative to T. Hsiao, Pesaran and Pick, (2012) consider the problem of testing for cross section independence in limited dependent variable panel data models.

Chudik and Pesaran (2014) provide an overview of the recent literature on estimation and inference in large panel data models with cross-sectional dependence. It reviews panel data models with strictly exogenous regressors as well as dynamic models with weakly exogenous regressors. The paper begins with a review of the concepts of weak and strong cross-sectional dependence, and discusses the exponent of cross-sectional dependence that characterizes the different degrees of cross-sectional dependence. It considers a number of alternative estimators for static and dynamic panel data models, distinguishing between factor and spatial models of cross-sectional dependence. The paper also provides an overview of tests of independence and weak cross-sectional dependence.

Pesaran (2015a) considers testing the hypothesis that errors in a panel data model are weakly cross sectionally dependent, using the exponent of cross-sectional dependence, introduced in Bailey, Kapetanios and Pesaran (2012).

7.1.1 CSD from spatial effects

There are various sources of CSD (neighborhood or network effects, the influence of a dominant unit or the influence of common unobserved factors) and various representations of CSD.Pesaran and Tosetti (2009) discuss the links between the various forms of CSD and give more precise definitions than used here.

Suppose that we are considering $N \times 1$ units, such as US states. Spatial models use a weight matrix $\mathbf{W}$, a known $N \times N$ positive matrix, with its diagonal elements equal to zero. For instance, $\mathbf{W}$ could reflect whether the units share a common border. Then the elements $w_{ij} = 1$ if two units share a common border, zero otherwise. Alternatively $w_{ij} = d_{ij}$ the distance between i and j. Sometimes it is parameterised for instance $w_{ij} = 1/d_{ij}^{\gamma}$ and γ is a distance decay factor.

Consider $N \times 1$ units, with a spatial model of the form

$$\begin{aligned}
\mathbf{y} &= \rho \mathbf{W}\mathbf{y} + \alpha \boldsymbol{\iota}_N + \mathbf{X}\beta_1 + \mathbf{W}\mathbf{X}\beta_2 + \mathbf{u} \\
\mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}_t
\end{aligned}$$

where $\boldsymbol{\varepsilon}_t$ is cross-sectionally independent and $\mathbf{X}$ is $N \times k$. The parameter vectors β_1 and β_2 are $k \times 1, \rho, \alpha$ and λ are scalars. Vega and Elhorst (2013) refer to this as the general nesting model and one can get other commonly used spatial models as special cases of this. Unlike the temporal lag terms the spatial lag term $\mathbf{W}\mathbf{y}$ is endogenous, this has to be allowed for, but there are maximum likelihood and GMM estimators for the cases when $\rho \neq 0, \lambda \neq 0$. This general model is rarely estimated since it is only weakly identified because the spatial autoregression and spatial error specifications are difficult to distinguish. There is considerable controversy about which of the spatial special cases of this general model is most useful. In the spatial lag of X, SLX, model, which has no endogeneity problem,

$$\mathbf{y} = \alpha \boldsymbol{\iota}_N + \mathbf{X}\beta_1 + \mathbf{W}\mathbf{X}\beta_2 + \mathbf{u}$$

the direct effect of the exogenous variables on the locality itself is given by $\mathbf{X}\beta_1$ the spillovers from neighboring localities by $\mathbf{WX}\beta_2$. These spillovers are local only to units connected through $\mathbf{W}$. In the spatial autoregression, SAR, model

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\boldsymbol{\iota}_N + \mathbf{X}\beta_1 + \mathbf{u}$$

the reduced form is

$$\mathbf{y} = (I - \rho\mathbf{W})^{-1}\alpha\boldsymbol{\iota}_N + (I - \rho\mathbf{W})^{-1}\mathbf{X}\beta_1 + (I - \rho\mathbf{W})^{-1}\mathbf{u}$$

so the spillovers are global transmitted through the $(I - \rho\mathbf{W})^{-1}$ matrix.

Vega and Elhorst (2013) discuss the advantages of parameterising the $\mathbf{W}$ matrix e.g. having the elements decay with distance, e.g $w_{ij} = d_{ij}^{-\gamma}$, where γ is estimated.

7.1.2 CSD from unobserved factors

Factor models interpret the cross-section dependence as reflecting the fact that the errors are determined by a vector of unobserved common factors

$$y_{it} = \mathbf{z}_t'\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i'\mathbf{x}_{it} + \boldsymbol{\gamma}_i'\mathbf{f}_t + \varepsilon_{it} \tag{37}$$

where y_{it} is a scalar dependent variable; z_t is a $k_z\times$ vector of variables that do not differ over units, e.g. intercept and trend; $\mathbf{x}_{it}$ is a $k_x \times 1$ vector of observed regressors which differ over units; $\mathbf{f}_t$ is a $r \times 1$ vector of unobserved factors, which may influence each unit differently and which may be correlated with the $\mathbf{x}_{it}$; and ε_{it} is an unobserved disturbance with $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}^2) = \sigma_i^2$ which is independently distributed across i and (possibly) t. The covariance between the errors $e_{it} = \boldsymbol{\gamma}_i'\mathbf{f}_t + \varepsilon_{it}$ is determined by the factor loadings $\boldsymbol{\gamma}_i$. Notice that if $\mathbf{f}_t$ is correlated with $\mathbf{x}_{it}$, as is likely in many economic applications such as. global cycles, then not allowing for CSD by ommitting $\mathbf{f}_t$ causes the estimates of $\boldsymbol{\beta}_i$ to be biased and inconsistent.

Infinite VARS treat the CSD as reflecting completely flexible interdependence between the $\mathbf{N} \times 1$ vector of observations on each unit (variable)

$$\mathbf{y}_t = \boldsymbol{\Phi}\mathbf{y}_{t-1} + \mathbf{u}_t$$

and consider approximations to this structure as $N \to \infty$, Chudik and Pesaran (2009). Cubadda et al. (2008) consider a related representation.

7.1.3 Weak and strong CSD

One can distinguish weak and strong CSD. With weak CSD, the dependences are local and decline with N. This could be the case with spatial models discussed above, where each cross-section unit is correlated with near neighbors but not others. With strong CSD the dependences influence all units as in the case of the factor models above, with $\boldsymbol{\gamma}_i' \neq 0$. The distinction can be expressed in various ways. Suppose the elements of $\mathbf{y}_t$ are stationary, e.g. growth rates, and the weighted average of the elements $\overline{y}_t = \sum_i y_{it}/N$, where the weights are 'granular', go to zero as $N \to \infty$. Then with weak CSD the variance of $\overline{y}_t$ goes to zero as $N \to \infty$. If there is strong CSD it does not, for instance there may be a global cycle in $\overline{y}_t$. If there is weak CSD the influence of the factors, $\sum_i \boldsymbol{\gamma}_i^2$ is bounded as $N \to \infty$, if there is strong dependence it goes to infinity with N. If there is weak dependence, all the eigenvalues of the covariance matrix of the errors $E(e_{it}e_{jt})$ are bounded as $N \to \infty$. If there is strong dependence, the largest eigenvalue goes to infinity with N. See Chudik, Pesaran & Tosetti (2010) and Bailey, Kapetanios & Pesaran (2012) who characterise the strength of the dependence in terms of the exponent of CSD, defined as $\alpha = \ln(n)/\ln(N)$ where n is the number of units (out of the total N) with non zero factor loadings. In the case of a strong factor $\alpha = 1$. Bailey et al. find that their estimates for a variety of cases suggest $\alpha < 1$. This way of characterising the strength of dependence allows for intermediate forms between weak and strong, such as semi strong.

The spatial and factor models treat CSD as an exogenous feature of the data. Kapetanios, Mitchell and Shin (2014) have an interesting non-linear model, which can endogenously generate both weak and strong CSD. The model relies on unit specific aggregates or summaries of past values of variables relating to other units. The model can capture aspects of herding and learning.

CSD is central to all the issues discussed in these notes. For instance, there is a growing literature on testing for structural change in panels, However, the apparent structural change may result from having left out an unobserved global variable, the f_t above. If f_t is omitted and the correlation between the f_t and the x_{it} changes, this will change the estimate of β_i giving the appearance of structural change. Similarly, an ommitted factor may give the impression of non-linearity. Cerrato, de Peretti and Sarantis (2007) extend the Kapetanios, Shin and Snell (2003) test for a unit root against a non-linear ESTAR alternative to allow for cross-section dependence. Since unobserved factors play a major role in the treatment of CSD, we begin by discussing the estimation of such factors and postpone issues of estimation of models with CSD to the next chapter. The implications for estimation are different depending on whether the $\mathbf{f}_t$ are merely regarded as nuisance parameters that we wish to control for in order to get better estimates of β or whether they are the parameters of interest: one wishes to estimate the $\mathbf{f}_t$ as variables of economic interest in their own right.

7.2 Factor models

The meaning of the term 'factor' depends very much on context and it has a variety of different meanings in different areas. Here it means that some observed variables x_{it}, $i = 1, 2, ..., N$ are determined by some unobserved factors, f_{jt}, $j = 1, 2, ..., r$ e.g.

$$x_{it} = \lambda_{io} + \sum_{j=1}^{r} \lambda_{ij} f_{jt} + e_{it} \tag{38}$$

the λ_{ij} are often called factor loadings, the e_{it} are often called idiosyncratic effects. Usually r is much smaller than N so the variation in a large number of observed variables can be reduced to a few unobserved factors which determine them. Although the notation suggests a panel structure, this need not be the case. T often is but need not be time periods, i may index cross-section units, variables, or other things. The case where the x_{it} are errors in a regression equation as in (37) above, will be considered in detail in the next chapter.

7.2.1 Uses

Factor models are used in various applications:

- In economics the oldest is probably the decomposition of time series into unobserved factors labelled trend, cycle, seasonals etc. This remains a common quest.
- More generally, it may be believed that the observed series are generated by some underlying unobserved factors and the objective is to measure them. This was developed most extensively in psychometrics, where the x_{it} are answers to a variety of questions by a sample of people. The underlying factors are aspects of personality, e.g. neuroticism, openness, conscientiousness, agreeableness and extroversion. It has also been used in economics for unobserved variables like: development, natural rates, permanent components, core inflation, etc.
- Factor models can be used to measure the dimension of the independent variation in a set of data, e.g. how many factors are needed to account for most of the variation in x_{it}. For I(1) series these dimensions may be the stochastic trends.
- Factor models can be used to reduce the dimensionality of a set of possible explanatory variables in regression or forecasting models, i.e. replace the large number of possible x_{it} by a few f_{jt} which contain most of the information in the x_{it}. This may reduce omitted variable problems.
- Factor models are used to model residual cross-section dependence in panel data models.
- Factor models have been used to choose instruments for IV or GMM estimators when there is a large number of potential instruments.

7.2.2 Estimation Methods

There are various ways to estimate factors:

- Univariate ($N = 1$) filters (e.g. the Hodrick-Prescott filter for trends).
- Multivariate ($N > 1$) filters such as the Kalman filter used to estimate unobserved-component models, Canova (2007) discusses this approach.

- Multivariate judgemental approaches, e.g. NBER cycle dating based on many series.
- Using a priori weighted averages of the variables.
- Deriving estimates from a model, e.g. Beveridge Nelson decompositions which treat the unobserved variable as the long-horizon forecast from a model.
- Principal component, PC, based methods.

The relative attractiveness of these methods depends on the number of observed series, N, and the number of unobserved factors, r. The method emphasised here is Principal Components, PC. This can be appropriate for large N small r. Unobserved component models for small N tend to put more parameteric structure on the factors, which PCs do not. As always size of N and T are crucial. It may be the case that there are some methods that work for small N, other methods that work for large N, but no obvious methods for the medium sized N that we have in practice.

Factor models have a long history. In the early days of econometrics, it was not clear whether the errors in variables model (observed data generated by unobserved factors) or the errors in equation model was appropriate and both models were used, for instance Stone (1947) used PCs to show that most of the variation in a large number of national accounts series could be accounted for by three factors, which could be interpreted as trend, cycle and rate of change of the cycle. Factor analysis was extensively developed in psychometrics and as the errors in equation model came to dominate played relatively little role in the development of econometrics. However, in the last few years there has been an explosion of papers on factor models in economics. The original statistical theory was developed for cases where one dimension, say N was fixed and the other say T went to infinity. It is only recently that the theory for large panels, where where both dimensions can go to infinity, has been developed.[21]

Below we briefly review the construction of PCs, discuss some other issues with their use and look at a few applications.

7.3 Calculating Principal Components

7.3.1 Static Models

Suppose that we have a $T \times N$ data matrix, X with typical element x_{it}, observations on a variable for units $i = 1, 2, ..., N$ and periods $t = 1, 2, ..., T$. We will interpret i as denoting units such as countries and t as time periods, but PCs can be applied to lots of other types of data. The direction in which you take the factors could also be reversed, i.e. treat X as an $N \times T$ matrix. We assume that the $T \times N$ data matrix X is generated by a smaller set of r unobserved

[21] See Bai (2003) and Bai and Ng (2002).

factors stacked in the $T \times r$ matrix F. In matrix notation, (38) can be written:

$$X = F\Lambda + E \tag{39}$$

where Λ is a $r \times N$ matrix of factor loadings and E is a $T \times N$ matrix of idiosyncratic components. Units can differ in the weight that is given to each of the factors. Strictly factor analysis involves making some distributional assumptions about e_{it} and applying maximum likelihood to estimate the factor loadings, but we will estimate the factors as the PCs of the data matrix.

The Principal Components of X are the linear combinations of X that have maximal variance and are orthogonal to (uncorrelated with) each other. Often the X matrix is first standardised (subtracting the mean and dividing by the standard deviation), to remove the effect of units of measurement on the variance, $X'X$ is then the correlation matrix. To obtain the first PC we construct a $T \times 1$ vector $f_1 = Xa_1$, such that $f_1'f_1 = a_1'X'Xa_1$ is maximised. We need some normalisation, for this to make sense, so use $a_1'a_1 = 1$. The problem is then to choose a_1 to maximise the variance of f_1 subject to this constraint. The Langrangian is

$$\pounds = a_1'X'Xa_1 - \phi_1(a_1'a_1 - 1)$$
$$\frac{\partial \pounds}{\partial a_1} = 2X'Xa_1 - 2\phi_1 a_1 = 0$$
$$X'Xa_1 = \phi_1 a_1$$

so a_1 is the first eigenvector of $X'X$, (the one corresponding to the largest eigenvalue, ϕ_1); or the first eigenvector of the correlation matrix of X if the data are standardised. This gives us the weights we need for the first PC. The second PC $f_2 = Xa_2$ is the linear combination which has the second largest variance, subject to being uncorrelated with a_1 i.e. $a_2'a_1 = 0$, so a_2 is the second eigenvector. If X is of full rank, there are N distinct eigenvalues and associated eigenvectors and the number of PCs is N. Note $AA' = I_N$ so $A' = A^{-1}$.

We can stack the results

$$X'XA = A\Phi.$$

where A is the matrix of eigenvectors and Φ is the diagonal matrix of eigenvalues. We can also write this

$$\begin{aligned} A'X'XA &= \Phi \\ F'F &= \Phi \end{aligned}$$

The eigenvalues can thus be used to calculate the proportion of the variation in X that each principal component explains: $\phi_i / \sum \phi_i$. If the data are standardised $\sum \phi_i = N$, the total variance. Forming the PCs is a purely mathematical operation replacing the $T \times N$ matrix X, by the $T \times N$ matrix F.

We define the PCs as $F = XA$, but we can also write $X = FA'$ defining X in terms of the PCs. Usually we want to reduce the number of PCs that we use, to reduce the dimensionality of the problem so we can write it

$$X = F_1 A_1' + F_2 A_2'$$

where the $T \times r$ matrix F_1 contains the $r < N$ largest PCs, the $r \times N$ matrix A_1' contains the first r eigenvectors corresponding to the largest eigenvalues. We treat F_1 as the common factors corresponding to the f_{jt}, and $F_2 A_2$ as the idiosyncratic factors corresponding to the e_{it} in (38). While it is an abuse of this notation, we will usually write F_1 as F and $F_2 A_2'$ as E.

7.3.2 Dynamic Models

Above t did not have to represent time, the arguments applied to any two dimensional data matrix, but suppose that t does represent time and we write the factor model in time series form

$$\mathbf{x}_t = \mathbf{\Lambda f}_t + \mathbf{e}_t$$

where x_t is an $N \times 1$ vector, Λ an $N \times r$ matrix of loadings, f_t a $r \times 1$ vector of factors and e_t a $N \times 1$ vector of errors. In using PCs to calculate the factors we have ignored all the information in the lagged values of the x_t. It may be that some lagged elements of x_{it-j} contain information that help predict x_{it}, e.g. factors influence the variables at different times. Standard PCs, which just extract the information from the covariance matrix, are often called static factor models, because they ignore the dynamic information in the autocorrelations and the idiosyncratic component, e_t, may be serially correlated. There are also dynamic factor models which extract the PCs of the long-run covariance matrix or spectral density matrix, Forni et al. (2000, 2003, 2005). Journal of Econometrics, 119, (2004) has a set of papers on dynamic factor models. The spectral density matrix is estimated using some weight function, like the Bartlett or Parzen windows, with some truncation lag.

The dynamic factor model gives us different factors, say

$$\mathbf{x}_t = \mathbf{\Lambda}^* \mathbf{f}_t^* + \mathbf{e}_t^* \tag{40}$$

where f_t is a $r^* \times 1$ vector. In practice we can approximate the dynamic factors in many applications by using lagged values of the static factors,

$$\mathbf{x}_t = \mathbf{\Lambda}(L)\mathbf{f}_t + \mathbf{e}_t^s. \tag{41}$$

where $\mathbf{\Lambda}(L)$ is a pth order lag polynomial. This may be less efficient in the sense that the $r^* < rp$: one can get the same degree of fit with fewer parameters using the dynamic factors than using current and lagged static factors. Determining whether the dynamics in $\mathbf{x}_t$ comes from an autoregression in $\mathbf{x}_t$, dynamics in $\mathbf{f}_t$ or serial correlation in $\mathbf{e}_t$ raises quite difficult issues of identification.

Dynamic PCs are two sided filters, taking account of future as well as past information, thus are less suitable for forecasting purposes. This problem does not arise with using current and lagged static factors. Forni et al. (2003) discuss one sided dynamic PCs which can be used for forecasting. Forecasting also includes 'nowcasting', where one has a series, say quarterly GDP, produced with a lag but various monthly series produced very quickly, such as industrial production and retail sales. PCs of the rapidly produced series are then used to provide a 'flash' estimate of current GDP. Mitchell et al. (2005) find that PCs based on a number of series forecast current GDP worse than one just based on industrial production. Stock and Watson (2010, 2016) contain surveys.

7.4 Issues in using PCs

7.4.1 How to choose r

Choosing how many factors to use, i.e. determining r, is a difficult issue. There are a variety of statistical criteria suggested but the choice will also depend on the purpose of the exercise and the context (e.g. the relevant economic theory). Traditional rules of thumb for determining r included choosing the PCs that correspond to eigenvalues that are above average value or equivalently greater than unity for standardised data or graphing the eigenvalues and seeing where they drop off sharply, if they do. There are also various tests and information criteria for N fixed T going to infinity. Write the relationship

$$x_{it} = \lambda_i' f_t + e_{it}$$

where f_t are the observations on the $r \times 1$ PCs corresponding to the largest eigenvalues. By construction, these PCs minimise the unexplained variance

$$V(r) = (NT)^{-1} \sum_i \sum_t (x_{it} - \lambda_i' f_t)^2.$$

Bai and Ng (2002) review a number of criteria and show that the number of factors r can be estimated consistently as $\min(N,T) \to \infty$ by minimising one of the following information criteria

$$\begin{aligned} IC_1(r) &= \log(V(r)) + r\left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right) \\ IC_2(r) &= \log(V(r)) + r\left(\frac{N+T}{NT}\right)\log\left(\min(N,T)\right) \end{aligned}$$

Although the Bai and Ng criteria have been widely used, they may not work well when N or T are small, leading to too many factors being estimated, e.g. always choosing the maximum number allowed. De Silva et al. (2009) discuss the performance of the information criteria in choosing r in the context of panel unit root testing.

Having chosen r denote the estimates

$$\begin{aligned} \widetilde{e}_{it} &= x_{it} - \widetilde{\lambda}_i' \widetilde{f}_t \\ \widetilde{E} &= X - \widetilde{F}\widetilde{\Lambda} \end{aligned}$$

where $\widetilde{\Lambda}$ is $N \times r$ and $\widetilde{F}$ is $r \times T$. If they are constructed from standardised data: $\widetilde{F}'\widetilde{F}/T = I$. Bai (2003) provides formulae for estimating covariance matrices of the estimated factors and loadings.

Onatski (2009) suggests a function of the largest eigenvalues of the spectral density matrix at a specified frequency to choose r.

7.4.2 How to choose N and T

One may have very large amounts of potential data available (e.g. thousands of time series on different economic, social, and demographic variables for different countries) and an issue is how many of them you should use in constructing the principal components. It may seem that more information is better so one should include as many as possible, but this may not be the case. Adding variables that are weakly dependent on the common factors will add very little information. To calculate the PCs the weights on the series have to be estimated and adding more series adds more parameters to be estimated. This increases the noise due to parameter estimation error. If the series have little information on the factors of interest, they just add noise, worsening the estimation problem. The series may be determined by different factors, increasing the number of factors needed to explain the variance. They may also have outliers or idiosyncratic jumps and this will introduce a lot of variance which may be picked up by the estimated factors. Many of the disputes in the literature about the relevant number of factors reflect the range of series used to construct the PCs. If the series are mainly different sort of price and output measures, two factors may be adequate; but if one adds financial series such as stock prices and interest rates, or international series such as foreign variables, more factors may be needed. One may be able to look at the factor loading matrix and see whether it has a block structure, certain factors loading on certain sets of variables. If this is the case one may want to split the data using different data-sets to estimate different factors. But in practice it may be difficult to determine the structure of the factor loading matrix.

N may be larger than the number of variables, if transformations of the variables (e.g. logarithms, powers, first differences, etc.) are also included. This trade-off between the size of the information set and the errors introduced by estimation may be a particular issue in forecasting, where parsimony tends to produce better forecasting models. Then using more data may not improve forecasts, e.g. Mitchell et al. (2005) and Elliott and Timmerman (2008).

In forecasting we would need to update our estimates of the F_t, and perhaps r the number of factors, as T our sample size changes. This raises similar issues to recursive estimation of parameters in a regression: whether we should change the model (the variables included in the regression) or only change the estimates for a constant model and whether we should drop some of the early part of the sample because of structural change.

7.4.3 How to Identify and interpret factors

To interpret the factors requires just identifying restrictions. Suppose that we have obtained estimates:

$$X = F\Lambda + E.$$

Then for any non-singular matrix P, the new factors and loadings $(FP)(P^{-1}\Lambda)$ are observationally equivalent to $F\Lambda$. The new loadings are $\widetilde{\Lambda} = (P^{-1}\Lambda)$ and factors are $\widetilde{F} = (FP)$. Note the similarity to the issue of choosing restrictions in

SEM and cointegration identification above. The just identifying restrictions, the P matrix, used to calculate PCs are the unit length and orthogonality normalisations which come from treating it as an eigenvalue problem. Thus the factors are only defined up to a non-singular transformation. In many cases a major problem in applications is to interpret the estimated PCs. Often in time-series the first PC has roughly equal weights and corresponds to the mean of the series. Looking at the factor loadings and the graphs of the PCs may help interpret them. The choice of P, just identifying restrictions, called 'rotations' in psychometrics, is an important part of traditional factor analysis where they help interpretation. Rotations in psychometrics are as controversial as just-identifying restrictions in economics, so while many psychologists agree that there are five dimensions to personality, $r = 5$, how they are described differs widely and the list given above, neuroticism etc., is only one candidate.

The same identification issue arises in simple regression. For

$$y = X\beta + u$$

is observationally equivalent to the reparameterisation

$$\begin{aligned} y &= (XP^{-1})(P\beta) + u \\ y &= Z\delta + u. \end{aligned}$$

For instance, Z could be the PCs, which have the advantage that they are orthogonal and so the estimates of the factor coefficients are invariant to which other factors are included. But there could be other P matrices. To interpret the regression coefficients we need to choose a parameterisation, k^2 restrictions that specify P. We tend to take the parameterisation for granted in economics, so this is not usually called an identification problem.

For some purposes, e.g. forecasting, one may not need to identify the factors, but for other purposes their interpretation is crucial. It is quite often the case that one estimates the PCs and has no idea what they mean or measure.

7.4.4 Estimated or imposed weights?

Factors are estimated as linear combinations of observed data series. Above it has been assumed that the weights in the linear combination should be estimated to optimise some criterion function, e.g. to maximise variance explained in the case of PCs. However, in many cases there are possible *a priori* weights that one might use instead, imposing the weights rather than estimating them. Examples are equal weights as in the mean or trade weights as in effective exchange rates. There is a bias-variance trade-off as with imposing coefficients in regression. The imposed weights are almost certainly biased, but have zero variance. The estimated weights may be unbiased but may have large variance because of estimation error. The imposed weights may be better than the estimated weights in the sense of having smaller mean square error (bias squared plus variance) Forecast evaluation of regression models indicates that simple models with imposed coefficients tend to do very well. Measures constructed with imposed weights are usually also much easier to interpret.

7.4.5 Explanation using PCs

Suppose the model of interest is

$$y = X\beta + u$$

where β is a $N \times 1$ vector and you wish to reduce the dimension of the $T \times N$ matrix X. This could be because there are a very large number of candidate variables or because there is multicollinearity. As seen above replacing X by all the PCs $F = XA$ is just a reparameterisation which does not change the statistical relation

$$\begin{aligned} y &= XAA'\beta + u \\ y &= F\delta + u. \end{aligned}$$

However, we could reduce the number of PCs by writing it

$$\begin{aligned} y &= F_1\delta_1 + F_2\delta_2 + u \\ y &= XA_1A_1'\beta + XA_2A_2'\beta + u \end{aligned}$$

where F_1 are the $r < N$ largest PCs (corresponding to the largest eigenvalues) then setting $\delta_2 = A_2'\beta = 0$ to give

$$y = F_1\delta_1 + v.$$

If required the original coefficients could be recovered as $\beta = A_1\delta_1$. The hypothesis $\delta_2 = 0$ can be tested as long as $N < T$, which it may not be. There is the difficulty that a PC which has a small eigenvalue and explains a very small part of the total variation of X may explain a large part of the variation of y. The PCs are chosen on the basis of their ability to explain X not y, but the regression is designed to explain y. There is now a large literature on choosing regressors when there are more potential explanatory variables than observations and it is believed that the true model is sparce, with only a few explanatory variables. Partial Least Squares and LASSO are possible estimators in this case.

Secondly unless the F_1 can be given an interpretation, e.g. as an unobserved variable, it is not clear whether the hypothesis $\delta_2 = A_2'\beta = 0$ has prior plausibility or what the interpretation of the estimated regression is. Thirdly, estimation error is being introduced by using F_1 and these are generated regressors with implications for the estimation of the standard errors of δ_1. As a result, until recently with the Factor augmented VARS and ECMs discussed below, economists have tended not to use PCs as explanatory variables in regressions. Instead multicollinearity tended to be dealt with through the use of theoretical information, either explicitly through Bayesian estimators or implicitly by *a priori* weights e.g. through the construction of aggregates. Notice that we could include certain elements of X directly and have others summarised in factors.

7.4.6 Example: interest rates

Cochrane (2001, p379) proviedes an example where the decomposition has a natural interpretation and correspond to the *a priori* weights that one would expect. This gives the eigenvalue decomposition of the covariance matrix of zero coupon bond yields of different maturities. We follow his notation. The table below summarises the results.

TABLE 7.1
BOND YIELD FACTOR ANALYSIS

σ	1	2	3	4	5
6.36	0.45	0.45	0.45	0.44	0.44
0.61	-0.75	-0.21	0.12	0.36	0.50
0.10	0.47	-0.62	-0.41	0.11	0.46
0.08	0.10	-0.49	0.39	0.55	-0.55
0.07	0.07	-0.36	0.68	-0.60	0.21

The first column, σ, gives the square root of the eigenvalues. The columns marked 1-5 give the eigenvectors corresponding to 1-5 year zero-coupon bond yields. The covariance matrix is decomposed as $\Sigma = Q\Lambda Q'$; σ^2 gives the diagonal entries in Λ and the rest of the table gives the entries of Q. With this decomposition, we can say that the bond yields are generated by $y = Q\Lambda^{1/2}\varepsilon$, $E(\varepsilon\varepsilon') = I$. Thus Q give loadings on the shocks ε. The first eigenvector, which accounts for the bulk of the variation, has roughly equal weights and measures the level of interest rates. The second, measures the slope of the yield curve. The third the curvature in the yield curve. Thus they capture the main features that we would look at in a yield curve.

Moon and Perron (2007) study non-stationarity in a panel of interest rates using a linear dynamic factor model, estimating principal components. However, whether it is better to estimate the factors by principal components or use a priori weights which capture the level, slope and curvature of the yield curve directly is an open question, see e.g. Diebold et al. (2005) who argue for imposed weights.

7.5 FAVARs

In recent work, dynamic models which mix observables and estimated factors have become popular under the name factor-augmented VARS, discussed below, or factor augmented error correction models, see Bannerjee and Marcellino (2009). The standard practice is to first difference the data to make it stationary and then extract the stationary factors from the stationary data. This loses all long-run level information, for instance first differencing interest rates loses the information in the yield curve. In addition if there is a level relationship, first differencing the error introduces a non-invertible moving average component, which means that the first difference version will not have an autoregressive representation, which is required for the VAR. When you take factors of non-stationary variables you need to be concerned about orders of integration, cointegration and spurious regressions between the observables and the factors. Determining whether the dynamics comes from the factor or the variable can raise difficulties, though the factor structure may impose cross-equation restrictions on the dynamics.

The analysis of monetary policy often involves estimating a small VAR in some focus variables, e.g. output, inflation and interest rates. Then the VAR is used to examine the effect of a monetary shock to interest rates on the time paths of the variables, the impulse response functions. To identify the monetary shock involves making some short-run just identifying assumptions, e.g. a Choleski decomposition imposes a recursive causal ordering, in which some variables, e.g. output and inflation, are assumed to respond slowly, and others, e.g. interest rates, to respond fast, i.e. within the same period. VARs plus identifying assumptions are often called structural VARS. Generalised Impulse Response Functions do not require any just identifying assumptions but cannot be given a structural interpretation.

Small VARs can give implausible impulse response functions, e.g. the "price puzzle", that a contractionary monetary shock was followed by a price increase rather than a price decrease as economic theory would predict. This was interpreted as reflecting misspecification errors, the exclusion of relevant conditioning information. One response was to add variables and use larger VARs, but this route rapidly runs out of degrees of freedom, since Central Bankers monitor hundreds of variables. The results are also sensitive to the choice of which variables to add. Another response was Factor Augmented VARs, FAVARS. These are used to measure US monetary policy in Bernanke Boivin and Eliasz (2005), BBE; UK monetary policy in Lagana and Mountford (2005), LM; US and Eurozone monetary policy in Favero Marcellini and Neglia (2005), FMN. The technical econometric issues are discussed in more detail by Stock and Watson (2005), SW.

Consider a $M \times 1$ vector of observed focus variables $\mathbf{Y}_t$, a $K \times 1$ vector of unobserved factors $\mathbf{F}_t$ with a VAR structure

$$\begin{pmatrix} \mathbf{F}_t \\ \mathbf{Y}_t \end{pmatrix} = \mathbf{A}(L) \begin{pmatrix} \mathbf{F}_{t-1} \\ \mathbf{Y}_{t-1} \end{pmatrix} + \mathbf{v}_t.$$

where $\mathbf{A}(\mathbf{L})$ is a polynomial in the lag operator. The unobserved factors $\mathbf{F}_t$ are related to a $N \times 1$ vector $\mathbf{X}_t$, which contains a large number (BBE use $N = 120$, LM $N = 105$) of potentially relevant observed variables by

$$\mathbf{X}_t = \mathbf{\Lambda F}_t + \mathbf{e}_t,$$

where the $\mathbf{F}_t$ are estimated as the principal components of the $\mathbf{X}_t$, which may include the $\mathbf{Y}_t$.

The argument is that (a) a small number of factors can account for a large proportion of the variance of the $\mathbf{X}_t$ and thus parsimoniously reduce omitted variable bias in the VAR ; (b) the factor structure for $\mathbf{X}_t$ allows one to calculate impulse response functions for all the elements of $\mathbf{X}_t$ in response to a (structural) shock in $\mathbf{Y}_t$ transmitted through $\mathbf{F}_t$; (c) the factors may be better measures of underlying theoretical variables such as economic activity than the observed proxies such as GDP or industrial production; (d) FAVARs may forecast better than standard VARs (e) factor models can approximate infinite dimensional VARs, Chudik and Pesaran (2010, 2011).

BBE conclude: "the results provide some support for the view that the "price puzzle" results from the exclusion of conditioning information. The conditioning information also leads to reasonable responses of monetary aggregates".

The simplest approach (called the two step method) is to (1) estimate K PCs from the $\mathbf{X}$, (2) estimate the VAR treating the PCs as measures of $\mathbf{F}_t$ variables along with the M observed focus variables $\mathbf{Y}_t$. The standard errors produced by the two-step estimates of the FAVAR are subject to the generated regressor problem and thus can potentially lead to misleading inference. In large samples $\mathbf{F}_t$ can be treated as known, thus there is no generated regressor problem, but it is not clear how good this approximation is in practice.

Choosing M and K, the number of focus variables and the number of factors, raises difficult issues. SW for the US and LM for the UK argue for 7 factors, BBE argue for smaller numbers e.g. $M = 3$, $K = 1$, or $M = 1$, $K = 3$. They use monthly data with either output, inflation and the interest rate as focus variables and one factor or the interest rate as the only observed focus variable and 3 unobserved factors, their preferred specification. If a large number of factors are needed, it reduces the attraction of the procedure and may make interpretation of the factors more difficult. The procedure is sensitive to the choice of $\mathbf{X}_t$. Just making the set of variables large does not solve the problem, because there may be factors that are very important in explaining $\mathbf{X}_t$, but do not help in explaining $\mathbf{Y}_t$ and vice versa. BBE motivate the exercise with the standard 3 equation macro model with the unobserved factors being the natural level of output and supply shocks. However, they do not use this interpretation in the empirical work, just note the need to interpret the estimated factors more explicitly. In subsequent work Boivin and Giannoni (2006) do use the theory putting the factor model in the context of a DSGE with imperfect measurement of the theoretical variables. The dynamic factor structure leads to testable over-identifying restrictions, which SW find are rejected for BBE type models, but the economic effect of rejection is small.

LM largely follow BBE in selecting the set of variables $\mathbf{X}_t$, but the relevant variables are likely to be very different for a small open economy like the UK than a large almost closed economy like the US, a fact LM note. In particular, they do not consider any foreign variables apart from some trade variables. They note that UK monetary authorities would take account of US interest rates, but do not include them in $\mathbf{X}_t$. The estimated factors explain a much smaller proportion of the variance of the $\mathbf{X}_t$ in the UK than in the US application.

7.5.1 Panel FAVARs, example

It is noticeable that movements in the unemployment rate in different OECD countries are quite highly correlated, therefore the PCs of OECD unemployment rates may act as a proxy for general world influences on the natural rate and one can then examine the country specific movements independent of the general global influences.

Smith and Zoega (2008) SZ find that using data on unemployment for 21 countries 1960-2002, the first PC accounts for 69% of the variance, and the first four PCs account for 93% of the variance. For investment, the proportions were somewhat lower, 58% and 83%. In both cases the first PC was close to the mean across countries of the series for each year. The correlation between the first PCs was 0.93, suggesting a single common factor drove both investment and unemployment. For none of the countries can we reject a unit root in unemployment. But this could be because the natural rate had a unit root and the ADF statistic for the average unemployment rate over the 18 countries was -1.6, so an apparently I(1) global shock could be the explanation. When we allow for global shocks that are very persistent, there is more evidence for adjustment to a long-run equilibrium. Thus there appears to be cointegration between national unemployment rates and the first global factor in most countries, justifying the extraction of factors from $I(1)$ variables. If national unemployment rates cointegrate with the first factor extracted from global unemployment rates implies the other factors extracted from unemployment must all be I(0).

They then estimate a panel FAVAR explaining investment and unemployment in each country by their own lags and the lags of the factors using the Swamy RCM estimator and find that the global investment factor drives unemployment.

Table 7.2
Country loadings for unemployment

	PC_1	PC_2	PC_3	PC_4
Australia	0.2643	0.0743	0.0564	0.0836
Austria	0.2410	-0.2519	-0.0647	-0.1342
Belgium	0.2622	0.0852	-0.1648	0.0572
Canada	0.2397	0.2324	0.1406	0.2301
Denmark	0.2500	0.1935	0.1625	-0.0233
Finland	0.2203	-0.2693	0.2888	0.3204
France	0.2663	-0.0704	-0.1409	-0.0595
Germany	0.2596	-0.1149	-0.2041	0.0534
Ireland	0.2210	0.2724	0.0462	-0.4313
Italy	0.2517	-0.1905	-0.2027	-0.1485
Japan	0.2038	-0.2337	-0.5051	0.2833
Netherlands	0.2288	0.2988	-0.1991	-0.0989
New Zealand	0.2466	-0.1192	0.2519	-0.2462
Norway	0.2251	-0.1356	0.5144	-0.2423
Spain	0.2654	0.0204	-0.0883	-0.1387
Sweden	0.1991	-0.3310	0.2478	0.3975
UK	0.2444	0.2393	-0.1318	0.0140
US	0.0959	0.5361	0.1478	0.4629

7.6 Infinite VARs

Chudik and Pesaran (2011a) discuss the 'curse of dimensionality' in the case of infinite dimensional vector autoregressive (IVAR) models. It is assumed that each unit or variable in the IVAR is related to a small number of neighbors and a large number of non-neighbors. The neighborhood effects are fixed and do not change with the number of units (N), but the coefficients of non-neighboring units are restricted to vanish in the limit as N tends to infinity. Problems of estimation and inference in a stationary IVAR model with an unknown number of unobserved common factors are investigated. A cross section augmented least squares (CALS) estimator is proposed and its asymptotic distribution is derived. An empirical illustration examines dynamic spill-over effects in modelling of U.S. house prices.

Pesaran and Chudik, (2013) extends the analysis of infinite dimensional vector autoregressive models (IVAR) to the case where one of the variables or the cross section units in the IVAR model is dominant or pervasive. The dominant unit influences the rest of the variables in the IVAR model both directly and indirectly, and its effects do not vanish even as the dimension of the model (N) tends to infinity. The dominant unit acts as a dynamic factor in the regressions of the non-dominant units and yields an infinite order distributed lag relationship between the two types of units. Despite this it is shown that the effects of the dominant unit as well as those of the neighborhood units can be consistently estimated by running augmented least squares regressions that include distributed lag functions of the dominant unit.

The curse of dimensionality applies also to estimating the covariance matrix and Bailey, Pesaran and Smith (2015) propose a regularisation method for the estimation of large covariance matrices. The method tests the statistical significance of individual pair-wise correlations and sets to zero those elements that are not statistically significant, taking account of the multiple testing, MT, nature of the problem. The procedure is straightforward to implement and is readily adapted to deal with non-Gaussian observations.The MT estimator performs well and tends to outperform the other estimators, particularly when the cross-sectional dimension, N, is larger than the time series dimension, T. If the inverse covariance matrix is also of interest, then a shrinkage version of the MT estimator is proposed that ensures positive definiteness.

Chudik, Mohaddes, Pesaran and Raissi, (2016) develop a cross-sectionally augmented distributed lag (CS-DL) approach to the estimation of long-run effects in large dynamic heterogeneous panel data models with cross- sectionally dependent errors. The asymptotic distribution of the CS-DL estimator is derived under coefficient heterogeneity in the case where the time dimension (T) and the cross-section dimension (N) are both large. The CS-DL approach is compared with more standard panel data estimators that are based on autoregressive distributed lag (ARDL) specifications. It is shown that unlike the ARDL type estimator, the CS-DL estimator is robust to misspecification of dynamics and error serial correlation. The theoretical results are illustrated with small sample evidence obtained by means of Monte Carlo simulations, which suggest that the performance of the CS-DL approach is often superior to the alternative panel ARDL estimates particularly when T is not too large and lies in the range of $30<T<100$.

8 Cross Section Dependence: estimation

CSD has attracted considerable attention in recent years and a large number of estimators have been suggested to deal with it. This chapter reviews some of them. Currently, the market leader, according to Monte Carlo studies, for large T cases appears to be CCE type estimators. However, there are a number of issues of interpretation. Kuersteiner and Pruch (2015) discuss flexible GMM estimators for the small T case, where there are observed distance measures. An early comparison of the estimators is Coakley, Fuertes and Smith (2006).

As usual issues interact. Banerjee and Carrion-i-Silvestre (2015) consider testing the null hypothesis of no cointegration when there are both structural breaks and cross-section dependence.

8.1 Estimators

8.1.1 Including Means, the CCE estimator

If one just wishes to treat the factors as nuisance parameters and remove the effect of CSD, a simple and effective procedure, for large N and T, is the correlated common effect, CCE, estimator of Pesaran (2006). This involves adding the means of the dependent and independent variables to the regression for each unit

$$y_{it} = \mathbf{z}_t'\boldsymbol{\alpha}_i + \mathbf{x}_{it}'\boldsymbol{\beta}_i + \delta_{0i}\overline{y}_t + \boldsymbol{\delta}_i'\overline{\mathbf{x}}_t + u_{it}. \tag{42}$$

A similar procedure can be used for unit root tests, etc. To see the motivation for this procedure, assume a single factor and average (37) across units to give

$$\begin{aligned} \overline{y}_t &= \mathbf{z}_t'\overline{\boldsymbol{\alpha}} + \overline{\beta}'\overline{\mathbf{x}}_t + \overline{\gamma} f_t + \overline{\varepsilon}_t + N^{-1}\sum(\beta_i - \overline{\beta})'\mathbf{x}_{it} \\ f_t &= \overline{\gamma}^{-1}\{\overline{y}_t - \mathbf{z}_t'\overline{\boldsymbol{\alpha}} + \overline{\beta}'\overline{\mathbf{x}} + \overline{\varepsilon}_t + N^{-1}\sum(\beta_i - \overline{\beta})'\mathbf{x}_{it}\} \end{aligned}$$

so the $\overline{y}_t$ and $\overline{\mathbf{x}}_t$ provide a proxy for the unobserved factor. Notice that the covariance between $\overline{y}_t$ and ε_{it} goes to zero with N, so for large N there is no endogeneity problem. The CCE generalises to many factors and lagged dependent variables, but requires that $\overline{\gamma}$, or the vector equivalent, is non-zero. This formulation assumes heterogeneous coefficients, there are homogeneous versions. There are sometimes economic reasons for adding averages, as in (42) , but in other cases the economic interpretation is not straightforward. In a variety of circumstances estimating the factors by the means, as the CCE does, seems to work better than estimating them directly by the principal component estimator discussed below and Westerlund & Urbain (2014) examine why this should be the case. Whereas the principal component procedures only work for strong factors the CCE estimators will work more generally.

The CCE procedure is simple to apply; can handle multiple factors which are $I(0)$ or $I(1)$ Kapetanios, Pesaran and Yamagata (2011); which can be correlated with the regressors; and handles serial correlation in the errors. The consistency proof holds for any linear combination of the dependent variable and the regressors, not just the arithmetic mean, subject to the assumptions that the weights w_i satisfy

$$(i): \; w_i = O\left(\frac{1}{N}\right), \; (ii): \; \sum_{i=1}^{N} | \, w_i \, | < K, \text{ and } (iii): \; \sum_{i=1}^{N} w_i \gamma_i \neq 0.$$

These clearly hold for the mean:

$$w_i = \frac{1}{N}, \; \sum_{i=1}^{N} | \, w_i \, | = 1, \text{ and } \sum_{i=1}^{N} w_i \gamma_i = N^{-1} \sum_{i=1}^{N} \gamma_i = \overline{\gamma}_N$$

as long as the mean effect of the factor on the dependent variable is non-zero. Notice that this procedure determines the weights a priori rather than estimating them by PCs. Not estimating the weights seems to improve the performance of the procedure, show that this procedure is robust to a wide variety of data generation processes including unit roots. See also Kapetanios and Pesaran (2007). Chudik and Pesaran (2015a) extend the procedure to dynamic models. Bailey, Holly, and Pesaran (2016) use cross unit averages to extract common factors (viewed as a source of strong cross-sectional dependence) and then apply multiple testing procedures to the de-factored observations to determine significant bilateral correlations. They apply this to US metropolitan area house prices.

8.1.2 Residual Principal components

Coakley, Fuertes and Smith (2002) suggested estimating a first stage model

$$y_{it} = b_i' \mathbf{x}_{it} + e_{it}$$

and then estimating the factors as the principal components of $\widehat{e}_{it}$, using some test or information criteria to choose the number of factors, $\widehat{\mathbf{f}}_t$. These factors are then included in a second stage regression

$$y_{it} = \boldsymbol{\beta}_i' \mathbf{x}_{it} + c_i' \widehat{\mathbf{f}}_t + v_{it}.$$

Assume that the x_{it} are generated by

$$x_{it} = \phi_{1i} f_t + \phi_{2i} z_t + \phi_{3i} \chi_t + v_{it} \tag{43}$$

where χ_t are common factors that influence x_{it} but not y_{it}. So

$$\overline{x}_t = \overline{\phi}_1 f_t + \overline{\phi}_2 z_t + \overline{\phi}_3 \chi_t + \overline{v}_t$$

This estimate using the residual principal component (RPC) will be a consistent estimator of β_i (for large N and T) if either $\phi_{1i} = 0$ or $\phi_{3i} = 0$ in (43). Otherwise using inconsistent estimates of $\widehat{e}_{it}$ causes it to be biased. The Coakley et al demonstration that the estimator was consistent had assumed that $\phi_{3i} = 0$. If $\phi_{1i} = 0$ this is a computationally convenient way to implement an approximation to SURE. Pesaran (2006) shows that under the case of a single regressor and a single factor the asymptotic the difference between the RPC estimator and true value will be zero only if either the factor is uncorrelated with $\overline{x}_t$; or if the factor is perfectly correlated with $\overline{x}_t$. If as N gets large the factor is perfectly correlated with $\overline{x}_t$, then it is obviously sensible to use $\overline{x}_t$.

8.1.3 Interactive fixed effects

Bai (2009) considers the model

$$\begin{aligned} Y_{it} &= X_{it}^{'}\beta + \lambda_i^{'}F_t + \varepsilon_{it}; \\ Y &= X\beta + F\Lambda' + \varepsilon, \end{aligned}$$

$i = 1, 2, ..., N$; $t = 1, 2, ..., T$, $X_{it}^{'}$ is a $p \times 1$ vector of observable regressors and F_t is a $r \times 1$ vector of unobserved factors. This model, is similar to the RPC of the previous subsection, the difference being it assumes homogeneous β and it requires interation. Bai interprets it as a generalisation of the usual addititve two way fixed effect model, e.g. when $\lambda_i^{'} = \lambda^{'}$, then $\alpha_t = \lambda^{'}F_t$. The just identifying restrictions for the factors are $F'F/T = I_r$ and $\Lambda'\Lambda$.diagonal. Bai suggests a least squares estimator which, rather than just using two steps as the RPC estimator above does, iterates between estimation of F and Λ by principal components and estimation of β, avoiding the inconsistency of the RPC estimator. Bai also considers various extensions, including bias corrections. The issue of choosing r remains, but Moon and Weidner (2015) show that it does not matter as long as you over-estimate, rather than underestimate, the number of factors. See also Bai (2013) and Westerlund and Norkute (2014) for a factor analytic method to estimate the model.

Hayakawa et al. (2014) discuss an ML estimator for interactive fixed effects in small T panels. .

8.1.4 PANIC

Bai and Ng (2004) suggest what they call a Panel Analysis of Non-stationarity in the Idiosyncratic and Common components (PANIC), which provides a way to analyse unit roots and cointegration. The data are assumed to be generated by

$$\begin{aligned} X_{it} &= c_i + \beta_i t + \lambda_i' F_t + e_{it} \\ F_{mt} &= \alpha_m F_{mt-1} + u_{mt} \\ e_{it} &= \rho_i e_{it-1} + \varepsilon_{it} \end{aligned}$$

Factor m is stationary if $\alpha_m < 1$, the idiosyncratic error e_{it} is stationary if $\rho_i < 1$. If e_{it} is stationary (I(1) X_{it} cointegrate with I(1) factors) then the PCs can consistently estimate the factors, whether they are $I(0)$, $I(1)$ or a mixture. When e_{it} is $I(1)$, this cannot be done, since the first equation is a spurious regression. However Bai and Ng suggest that the data can be differenced and demeaned if there is a trend as above; the first difference of the factors can be estimated by PCs, these can then be cumulated to give the factors and the the idiosyncratic error. Unit root tests can then be applied to these to determine whether the variables are $I(0)$ or $I(1)$. Since removing the F_t has removed the between group dependence, panel unit root tests can be conducted on the e_{it}. When F_t contains $I(1)$ elements, testing that e_{it} is $I(1)$ is a test for X_{it} not cointegrating with the $I(1)$ common factors. They illustrate their procedure by extracting core inflation from 21 component inflation series. Notice that since the factors are orthogonal, completely uncorrelated with each other, they cannot cointegrate. I(1) variables that cointegrate contain a common stochastic trend which cancels out in the linear combination and thus must be correlated. Westerlund and Larsson (2009) examine the issue of pooling the individual PANIC unit root tests. Bai & Ng (2010) extend the procedure. Westerlund (2014) further examines its power properties.

Moon and Perron (2002) and also use factor structures to test for unit roots and Phillips and Sul (2002) use factor structures for homogeneity testing and Phillips and Moon (2003) for testing convergence. Breitung and Das (2008) review testing for unit roots in panels with a factor structure. In all these cases, determining whether the non-stationarity, e.g. the unit root, comes from the dynamics of the observed series being investigated ($\rho_i = 1$) or from cointegration with an $I(1)$ unobserved factor ($\alpha_m = 1$) can be a delicate matter. There is also the difficulty of determining the number of factors and the increased variance from estimating the factor weights. Westerlund (2014) discusses the power of PANIC procedures.

8.1.5 SURE

Suppose that the model is heterogeneous

$$\begin{aligned} y_{it} &= z_t'\alpha_i + x_{it}'\beta_i + f_t'\gamma_i + \varepsilon_{it} \\ i &= 1,2,...,N;\ t = 1,2,...,T \end{aligned}$$

where y_{it} is a scalar dependent variable, z_t is a $k_z\times$ vector of variables that do not differ over groups, e.g. intercept and trend, and x_{it} is a $k_x \times 1$ vector of observed regressors which differ over groups, f_t is a $r \times 1$ vector of unobserved factors and ε_{it} is an unobserved disturbance with $E(\varepsilon_{it}) = 0$, $E\left(\varepsilon_{it}^2\right) = \sigma_i^2$ which is independently distributed across i and t. Estimating

$$y_{it} = z_t'a_i + x_{it}'b_i + v_{it}$$

will give inconsistent estimates of β_i if f_{it} is correlated with x_{it} and inefficient estimates even if f_{it} is not correlated with x_{it}. In the latter case if N is small, the equations can be estimated by SURE, but if N is large relative to T, SURE is not feasible, because the estimated covariance matrix cannot be inverted.

Robertson and Symons (1999, 2007) suggest using the factor structure to obtain an invertible covariance matrix. Their estimator is quite complicated and will not be appropriate if the factors are correlated with the regressors, which may be the case.

8.1.6 Time effects/demeaning

If $\beta_i = \beta$, and there is a single factor which influences each group in the same way, i.e. $\gamma_i = \gamma$ then including time effects, a dummy variable for each period, i.e. the two way fixed effect estimator:

$$y_{it} = \alpha_t + \alpha_i + \beta' x_{it} + u_{it}$$

will estimate $f_t'\gamma = \alpha_t$. This can be implemented by using time-demeaned data $\widetilde{y}_{it} = y_{it} - \overline{y}_t$, where $\overline{y}_t = \sum_{i=1}^{N} y_{it}/N$ and similarly for $\widetilde{x}_{it}$. Unlike SURE the factor does not have to be distributed independently of the x_{it} for this to work.

It is sometimes suggested (e.g. for unit root tests) that demeaned data be used even in the case of heterogeneous slopes. Suppose we have heterogeneous random parameters and the model is

$$\begin{aligned} y_{it} &= f_t + \beta_i' x_{it} + u_{it} \\ \beta_i &= \beta + \eta_i \end{aligned}$$

(including the intercept in x_{it}) averaging over groups for each period we get

$$\overline{y}_t = f_t + \beta \overline{x}_t + \overline{u}_t + N^{-1} \sum_{i=1}^{N} \eta_i' x_{it}$$

noting that

$$\beta_i' x_{it} - \beta \overline{x}_t = \beta_i' \widetilde{x}_{it} + \eta_i' \overline{x}_t$$

demeaning, using $\widetilde{y}_{it} = y_{it} - \overline{y}_t$, gives us

$$\begin{aligned} \widetilde{y}_{it} &= \beta_i' \widetilde{x}_{it} + \widetilde{u}_{it} + e_{it} \\ e_{it} &= \eta_i' \overline{x}_t - N^{-1} \sum_{i=1}^{N} \eta_i' x_{it} \end{aligned}$$

This removes the common factor f_t but has added new terms to the error reflecting the effect of slope heterogeneity. If η_i is independent of the regressors, e_t will have expected value zero and be independent of the regressors, so one can obtain large T consistent estimates of the β_i, but the variances will be larger. One can compare the fit of the panels using the original data y_{it} and the demeaned data $\widetilde{y}_{it}$ to see which effect dominates, i.e. whether the reduction in variance from eliminating f_t is greater or less than the increase in variance from adding e_{it}. This model assumes that the factor has identical effects on each unit. Rather than demeaning, it is usually better to include the means directly.

8.1.7 Orthogonalisation

If the equation is an autoregression, $x_{it} = y_{it-1}$, then the same factors that determine y_{it} determine x_{it}. In these circumstances one can orthogonalise the data before estimation. This involves extracting r principal components of y_{it} as estimates of f_t and using an autoregression of the othogonalised data of the form

$$y_{it} - f_t'\gamma_i = \beta\left(y_{it-1} - f_{t-1}'\gamma_i\right) + v_{it}.$$

The number of factors, r can be determined by some information criteria or test procedure, though the properties of available procedures seem to be sensitive to N and T, and there is no agreement as to the best procedure. This procedure has been primarily used to construct unit root tests that deal with cross-section dependence.

This procedure will not work where the x_{it} are exogenous regressors unless the factors are orthogonal to the regressors. Suppose the γ_i are in fact zero, the factors do not influence y_{it} directly, but influence it through x_{it}

$$x_{it} = f_t'\phi_i + u_{it}$$

orthogonalisation will estimate γ_i by $\beta_i\phi_i$ and remove a large part of the influence of x_{it}. This is rather like seasonally adjusting ice-cream sales and the weather and finding that the weather has no effect on ice-cream sales.

8.2 Applications

8.2.1 Factor structures

Coakley Fuertes and Smith (2006), CFS, consider Purchasing Power Parity, the Fisher relationship and the Feldstein-Horioka relationship. In each case they analyse the factor structure of the data and the residuals from a static model and compare the estimates of a static model from the various techniques.

For PPP, let s_{it} be the logarithm of the nominal exchange rate and $d_{it} = p_{it} - p_t^*$ is the log price differential with the US). According to PPP, exchange rates should reflect movements in the price differential in the long-run

$$s_{it} = \alpha_i + \beta_i d_{it} + u_{it}$$

with $H_0 : \beta_i = 1$.

For Fisher let il_{it} denote annualised long-term nominal interest rates and $\Delta p_{it} = p_{it} - p_{i,t-1}$ the annual inflation rate from period $t-1$ to period t. Assuming $E_t(\Delta p_{i,t+1}) = \Delta p_{it}$, the ex ante real long interest rate can be defined by $rl_{it} = il_{it} - \Delta p_{it}$. The Fisher effect embodies the notion that, in the long run, nominal interest rates are expected to fully reflect inflation that in

$$il_{it} = \alpha_i + \beta_i \Delta p_{it} + u_{it}$$

$H_0 : \beta_i = 1$. In both cases one might expect common factors to arise, for PPP these would include base country effects, the long swings in the real dollar rate, and for the Fisher equation movements in the world real interest rate. The analysis for the PPP and Fisher equations is based on quarterly data for 18 countries over 1973Q1-1998Q4. The panel dimensions for the PPP analysis are $N = 17$ (US is excluded) and $T = 104$ while those for the Fisher regression are $N = 18$ and $T = 100$ (four observations are lost in calculating the annual inflation series $\Delta_4 p_{it}$).

For Feldstein-Horioka (FH) let I_{it} be the share of investment in GDP and S_{it} the share of savings. In a world of unfettered capital mobility, national saving would flow to the countries offering the highest returns and domestic investment would be financed from global capital markets, while with no capital mobility investment is financed by domestic savings so in

$$I_i = \alpha_i + \beta_i S_i + u_i$$

$H_0 : \beta_i = 1$ with zero capital mobility, (see Coakley, Fuertes and Spagnolo, 2004, for further discussion). Their data set for the FH regression uses quarterly observations on national saving, domestic investment and GDP for the period 1980Q1-2000Q4 for 12 OECD countries: Australia, Canada, Finland, France, Italy, Japan, Netherlands, Norway, Spain, Switzerland, UK and US.

Table 8.1 reports the average absolute correlation for each series and the percentage of variation explained by each of the first four PCs. The latter are extracted from the $N \times N$ correlation matrix for each variable. The residual series used are obtained by individual OLS (FE results are similar). As expected, the average absolute cross-section correlations are relatively high.

Table 8.1: Cross-section dependence statistics

	PPP			Fisher			FH		
	e_{it}	d_{it}	$\hat{u}_{it}$	il_{it}	$\Delta_4 P_{it}$	$\hat{u}_{it}$	I_{it}	S_{it}	$\hat{u}_{it}$
$\lvert\rho\rvert$	0.5845	0.8421	0.6717	0.6152	0.6682	0.5542	0.3752	0.3497	0.2617
V_1	0.5602	0.8554	0.7244	0.6558	0.7070	0.6144	0.3971	0.4088	0.2935
V_2	0.3768	0.0910	0.1031	0.1776	0.0930	0.1154	0.2317	0.2507	0.2325
V_3	0.0175	0.0295	0.0563	0.0737	0.0573	0.0724	0.1669	0.1105	0.1728
V_4	0.0163	0.0105	0.0361	0.0151	0.0291	0.0432	0.0514	0.0870	0.0869

$\lvert\rho\rvert$ is the average absolute cross-section correlation. $V(\mathrm{pc}_j)$ is the proportion of variability explained by the jth PC. The PC's are extracted from the correlation matrix. $\hat{u}$ a are the individual OLS residuals.

The first two factors account for 94% of the variation in exchange rates and price differentials and just over 80% of the variation of interest rates and inflation, so clearly common factors are important in these cases. The first two factors tend to account for less of the variation in the residuals than in the variables. In the Fisher case, the factor structures in the dependent and independent variables and residuals are quite similar. In the PPP case they are rather different, the second factor being large in the case of exchange rates but not in price differentials. The factor structure in the FH case is rather different; the first common factor might correspond to the world real interest rate influencing investment in each country.

Next the regressions using a variety of panel estimators. DMG is mean group based on demeaned data, RPC is residual PCs, CS is cross-section on averaged data. Notice that the factors have no effect in the cross-section because they are averaged out. Table 8.2 summarises the results.

Table 8.2: Panel estimators

	PPP		Fisher		FH	
Estimator	$\hat{\beta}$	$se_{\hat{\beta}}$	$\hat{\beta}$	$se_{\hat{\beta}}$	$\hat{\beta}$	$se_{\hat{\beta}}$
POLS	1.2398	0.0151	0.4935	0.0128	0.6139	0.0182
FE	1.1079	0.0176	0.3765	0.0123	0.1850	0.0399
2FE	1.1209	0.0122	0.3701	0.0154	-0.0235	0.0451
FE-RPC	1.1121	0.0111	0.3563	0.0090	0.0649	0.0369
MG	1.1787	0.0901	0.3872	0.0421	0.3276	0.1765
SUR-MG	1.0705	0.0860	0.2967	0.0373	0.3460	0.1570
DMG	1.0876	0.1864	0.3631	0.0519	0.1010	0.2022
MG-RPC	1.1108	0.0986	0.3559	0.0516	0.1832	0.1707
CCE-MG	0.6453	0.1990	0.2714	0.0366	0.0620	0.1989
CS	1.3599	0.1282	0.8859	0.1077	0.6762	0.1095

The relationship between the estimators is very different in these data than in the Monte Carlo simulations repored in CFS(2006). In the simulations the estimates from 2FE, DMG and CCE-MG were almost identical. In these data they are very different. In the PPP the 2FE estimate of β is 1.12, the DMG 1.08 and the CCE-MG 0.64. They are clearly picking up very different things. The variation in the case of the Fisher relation is less. It may be coincidence, but the rankings by size of the estimate in the two examples are quite close. In both cases the cross-section is the largest, then the pooled OLS (which reflects the cross-section variance) then Mean Group. The smallest in both cases is the CCE-MG the next smallest the SURE Mean Group. CCE-MG standard errors are rather larger. The fact that the cross-section estimate is much larger than the others may result from the effect of removing a common factor which is positively correlated with the regressors. The rankings by size of the estimates are slightly different in the FH case.

We should note that in the CCE-MG approach, the slope coefficients on $\bar{y}$ and $\bar{x}$ (standard errors in parentheses) are 1.007 (0.076) and -0.650 (0.399), respectively, for PPP; 0.969 (0.115) and -0.253 (0.103) for the Fisher equation; 0.885 (0.248) and 0.111 (0.424) for the FH equation. An interesting pattern emerges. In all cases the coefficient on $\bar{y}$ is insignificantly different from one, and except for FH, the coefficient of $\bar{x}$ is similar but of opposite sign to the coefficient of x_{it}. This is consistent with them measuring a factor which has similar effects on all countries, like a time fixed effect.

The economic implications of these findings are that it is global factors are important in determining national exchange rates and interest rates, thus it is important to allow for global factors to appear in the equations: all macro-economics should be open-economy macro-economics, this is a theme pursued in the next chapter. There are issues of interpretation. In the Fisher and Feldstein-Horioka regressions, the global factor can be given a reasonably straight forward interpretation as a world real interest rate, which will shift the real interest rate in individual countries and will influence savings and investment in different countries. With PPP it is not clear how the factor should be interpreted, exchange rates are a set of relative prices and it makes no sense to say that all relative prices rise or all countries exchange rates appreciate. What seems to be picked up by the factor is a base country effect, since the exchange rates are all measured against the dollar. This sensitivity to base country effects raises difficulties for the panel PPP example we have used throughout these notes.

8.2.2 Spatial structures

When you have stripped out strong CSD, e.g. by CCE, there may be weak, local or spatial CSD. The traditional approach in spatial econometrics is to specify a priori a spatial weights matrix $\mathbf{W}$. With a large T panel you can look for the correlations directly, rather than having to specify them a priori. You can then look for the weak factors as reflected in the non-zero correlations in correlation matrix of residuals. Regularisation of large covariance matrices refers to the process by which small correlations, below a particular threshold, are set to zero to produce sparse matrices.

Bailey Pesaran and Smith. (2015) use a multiple testing approach and suggest thresholding by setting insignificant correlations to zero taking account of the multiple testing problem. With $n = N(N-1)/2$, independent elelents in the correlation matrix and a desired nominal size p, they suggest setting the required size to p/n. There is also the problem with large correlation matrices, noted above, that if $T < N$ the matrix may not be positive definite. Thresholding does not guarantee that the matrix is positive definite, if a positive definite matrix is required they suggest, shrinking non-zero thresholded correlation coefficients towards an identify matrix. The shrinkage parameter is chosen to just make the matrix positive definite. It makes a difference whether zero correlation implies independence (no non-linear relations) or not.

Bailey, Holly and Pesaran (2016) apply a two stage approach to spatio-temporal dependence with strong and weak CSD to US house prices. The data are changes in log real house prices in N=363 Metropolitan Statistical Areas 1975Q1-2010Q4, T=144. They remove the strong factors in various ways, hierarchical PCs, means etc. After defactoring significant positive and negative correlations are apparent. The usual spatial weighting assumes that all spillovers are positive. But it is possible that migration from A to B could raise house prices in B and depress them in A. Many of the weights from the correlations are similar to those that would arise from a conventional a priori $\mathbf{W}$ Distance matters but it is not all that matters, some areas that are far apart are highly correlated. They also model temporal links through lags.

8.3 Policy Evaluation in Panels

There is a large microeconometric literature on the estimation of treatment effects and recently there have been a number of attempts to export the microeconometric approach to macroeconometric policy evaluation. The micro and macro issues are rather different. For instance, the endogeneity and sample selection bias that arise due to correlated heterogeneity across the units in the micro-treatment case is not a problem in the macro case when the focus of the policy evaluation is on a single unit, and the "policy on/policy off" comparisons are done over time rather than across units. The cross-section dependence that results from strong factors driving all units in a panel means that other units can be used to construct controls that can be used to specify the counterfactual in analysis of a treatment effect or policy evaluation.

Consider a single target, unit 1, subject to intervention at T_0, with post intervention data $t = T_0 + 1, T_0 + 2, ..., T_0 + T_1$, with $T = T_0 + T_1$. Suppose that there are $N - 1$ controls not subject to the intervention and not affected by the intervention in unit 1. In a static model the effect of the intervention is measured as

$$d_{1,T_0+h} = y_{1,T_0+h} - \sum_{i=2}^{N} w_i y_{i,T_0+h};\ h = 1, 2, ..., T_1. \tag{44}$$

There is then an issue over how to choose controls and weights. A popular procedure in this context has been the synthetic control method (SCM). Abadie & Gardeazabal (2003) used it to measure the costs of Basque terrorism, Abadie et al. (2010) considered the California smoking program and Abadie et al. (2015) German reunification. These are all single target cases. Since the package Synth became available on Matlab, R and Stata, others have applied it to multiple target cases. It is used to measure the economic costs of Civil War on many countries by Costalli et al. (2014) and Bove et al (2016), who also use other estimators.

To determine the SCM weights w_i let $\mathbf{x}_{1kt}$ be a set of $k = 1, 2, ..., K$ predictor variables for y_{1t}, with the corresponding variables in the other units given by $\mathbf{x}_{jkt}$, $j = 2, 3, ..., N$. These variables are averaged over the pre-intervention period to get $\overline{\mathbf{x}}_{1k}^{T_0}$ and $\overline{\mathbf{X}}_{k}^{T_0}$ the $N - 1 \times 1$ vector of predictor k in the control group. Then the $N - 1 \times 1$ vector of weights $W = (w_2, w_3, ..., w_N)'$ are chosen to minimize

$$\sum_{k=1}^{K} v_k (\overline{x}_{1k}^{T_0} - W' \overline{X}_{k}^{T_0})^2$$

subject to $\sum_{i=2}^{N} w_i = 1$, $w_i \geq 0$,where v_k is a weight that reflects the relative importance of variable k.

SCM chooses the comparison units to be as similar as possible to the target along the dimensions included in $\mathbf{x}_{ikt}$. The v_k are often chosen by cross-validation, which may be problematic for potentially non-stationary time-series samples. The pre-intervention outcome variable may be included in x_{ikt}; it is argued that matching on the pre-intervention outcomes helps control for the unobserved factors affecting the outcome of interest. Kaul, et al. (2015) argue that with SCM one should never use all pre-intervention outcomes.

In the case of German Reunification, Abadie et al. (2015), use controls and weights w_i of Austria, 0.42, US, 0.22, Japan 0.16, Switzerland 0.11 and Netherlands, 0.09. The synthetic West Germany is similar to the real West Germany in pre 1990 per capita GDP, trade openness, schooling, investment rate and industry share. As they note there may be spillover effects. Since Austria, Switzerland and Netherlands share borders with Germany there is a distinct possibility that their post 1990 values may be influenced by German reunification. Those that are geographically the most similar are most likely to show spillover effects.

In the case of microeconometric treatment effect studies, when the units are only subject to weak factors, SCM is sensible: choose controls that are similar in characteristics to those that are treated: match patients treated with a drug to untreated controls of similar age, sex, background etc. Similarity is usually measured by propensity score, though if there is a single case, e.g. Basque terrorism, one cannot calculate propensity score. With multiple cases, e.g. Civil War, one can. Although there are some significant variables (rich countries have fewer civil wars) it is difficult to predict civil wars. So propensity scores are likely to be inaccurate. In addition the $\mathbf{x}_{jkt}$ are often poor predictors for y_{it}, which is why pre-intervention y_{it} are often used.

It is not clear that SCM is as sensible in macroeconometric time-series contexts, where there are strong common factors driving the y_{it}, so prediction from outcomes in other units y_{jt} is more sensible.This is done by Hsiao et al. (2012), using what they call a panel data approach for program evaluation to measure the benefits of political and economic integration of Hong Kong with mainland China. They measure the effect in the same way using (44), but choose the w_i by regression of y_{1t}, growth in Hong Kong on a subset of y_{jt}, $j = 2, 3, ..., N$, growth in the control countries during the pre-intervention period. The subset is chosen by a model selection procedure. They emphasize that Hong Kong is too small for the effects of integration with China to influence any of the control countries. The control group they select contains USA and Taiwan with positive weights and Japan, Korea, Philippines and Taiwan with negative weights. This is sensible in a macroeconometric context, because very different countries can be driven by the same common trends. Hsiao et al. include the US in the controls, not because the US is like Hong Kong, the justification in the SCM procedure, but because US growth is a good predictor of Hong Kong growth. Notice that in micro terminology, the parameter of interest in the macro cases is the effect of treatment on the treated: it makes no sense to consider either the effect of Hong Kong being integrated with West Germany or of East Germany being integrated with China.

The SCM procedure requires common support, the matching variables of all treated units belong to the support of matching variables of control units. This is unlikely to hold for Hong Kong, no other country is like Hong Kong, not even Singapore, the closest comparison. This is not a problem for the prediction method. Abadie et al. criticize the fact that regression methods can give negative weights, but this is to be expected if one interprets the procedure as involving prediction using global factors. Suppose Hong Kong before integration is largely driven by global factor A, the US by factors A and B, and Japan largely by factor B; then the US minus Japan provides an estimate of factor A, which drives Hong Kong. Hsiao et al. (2012) have a comparison of the factor interpretation of their method with that of SCM.

Following Abadie et al. we can write the model in (44)

$$y_{it} = d_{it}c_{it} + y_{it}^N$$

where y_{it}^N is the estimated value in the absence of intervention, equal to the actual value in the absence of intervention, and for a discrete treatment $c_{it} = 1$ if $i = 1$ and $t > T_0$, zero otherwise.

Were we to then model $y_{it}^N = \alpha_i + \boldsymbol{\lambda}_i^{'}\mathbf{f}_t + \varepsilon_{it}$ by a country specific intercept, global factors, $\mathbf{f}_t$ and an error, we would obtain

$$y_{it} = d_{it}c_{it} + \alpha_i + \boldsymbol{\lambda}_i^{'}\mathbf{f}_t + \varepsilon_{it} \tag{45}$$

which is a standard heterogeneous factor augmented panel model and there are a variety of ways of estimating the unobserved factors, $\mathbf{f}_t$ including the interactive fixed effect and common correlated effect estimators.

Gobillon and Magnac (2016) investigate the use of interactive effect or linear factor models in regional policy evaluation. They contrast treatment effect estimates obtained using Bai (2009) with those obtained using difference in differences and synthetic controls and show that difference in differences are generically biased, and derive support conditions for synthetic controls. They use Monte Carlo to compare these estimation methods in small samples. They show that if the true model is a linear factor model synthetic control is equivalent to interactive fixed effects in the interpolation case, (when the matching variables of treated units belong to the support of matching variables of control units, assumed to be convex), but not in the extrapolation case. Chan & Kwok (2016) extends the Pesaran (2006) procedure and extracts principal components from the control group to form factor proxies.

Another recent approach to macro policy evaluation borrows techniques from the micro literature to obtain an estimate of an average treatment effect. Angrist, Jorda and Kuersteiner (2013, AJK), estimate the effect of monetary policy, while Jorda and Taylor (2015) use similar procedures to estimate the effect of fiscal policy in a panel of countries. AJK use local linear projection type estimators to measure the average effect of policy changes on future values of the outcome variables (inflation, industrial production, and unemployment), weighted inversely by policy propensity scores in a way similar to that used to adjust non-random samples. They rely on outcomes averaged across different (possibly heterogenous) policy episodes, they do not use a structural model and their analysis is subject to the Lucas Critique. Their approach requires that the underlying parameters are invariant to policy changes, since it is only policy changes within the same regime that are identified in their framework (see AJK, p.5). In addition, matching estimators of this sort require a lot of data whereas macroeconometric samples tend to be data-poor relative to microeconometric samples. This is reflected in the large confidence bands AJK report around the measures of their estimated effects of target rate changes on macro variables.

None of these procedures are fully dynamic, although AJK allow for lags and Hsiao et al. (2012) allow for serial correlation in calculating standard errors. Pesaran & Smith (2014, 2016) explicitly consider testing for the effect of a policy intervention in a dynamic context, either in the case of a parsimonious reduced form or final form equation, P&S (2016), or in the context of a complete DSGE P&S(2014). They consider two types of intervention: discretionary where there is a deterministic change to the policy variable and rule based where one or more parameters of a stochastic policy rule are changed. P&S (2014) consider both a standard case where all variables in the macroeconometric model, including policy variables, are endogenous and a general case where the DSGE model is augmented by exogenous variables. The latter case accommodates interventions that change exogenous policy parameters, such as a fixed money supply target, or when steady states of some of the variables are changed as occurs when the inflation target is altered.

They are concerned with *ex post* evaluation of a policy intervention in the case of an individual unit (say a country), where time series data are available before as well as after the policy change. The proposed test is based on the difference between the realisations of the outcome variable of interest and counterfactuals obtained assuming no policy change, using only the pre-intervention parameter estimates and solving out for the post-intervention endogenous variables. This test has power against both types of policy change: interventions and shocks, which is useful given that we may be uncertain about the exact nature of the policy change.

In the case of policy intervention the Lucas Critique is an issue for the power of the test, but is not applicable to the construction of the test since the counterfactuals, based on estimates using pre-intervention data, will embody pre-intervention parameters while the realized post-intervention outcomes will embody the effect of the change in the policy parameters and any consequent change in expectations.

9 Within and between unit Cointegration: the GVAR

Cointegration between countries for a scalar y_{it}, log per-capita income has attracted attention in the convergence debate. If N is small, e.g. the $G7$ countries, one can estimate a VAR and use standard methods like Johansen to test for cointegration. However if N is large, e.g. the PWT data this is not feasible, since the 'curse of dimensionality' strikes. There are two broad approaches, shrink the data, e.g. with principal components, or shrink the parameters, e.g. with

Bayesian methods, Banbura et al (2010) or the approach suggested by Chudik and Pesaran (2010, 2011) in the context of what they call an infinite VAR, IVAR. The problem is compounded by the fact that one is rarely looking at a single variable for each unit, but a vector of variables. In this case, there can be both within country cointegration, e.g. between consumption in the UK and income in the UK, and between country cointegration, e.g. between income in the UK and income in the US. In fact many long-run relations, like purchasing power parity, imply between country cointegration. Bannerjee et al (2004) discuss this problem. For a small number of countries, one could construct a vector which includes all the relevant variables for all the relevant countries and look for both sets of cointegration. But this would rapidly become impractical as the number of variables and countries grew and the model tends towards an IVAR.

An alternative approach is to use economic theory and the nature of the interconnections to structure the problem. One approach, which has been widely adopted, follows the Global VAR, GVAR, structure introduced in Pesaran, Schuermann and Weiner (2004) developed in Garratt, Lee, Pesaran and Shin (2007) and Dees, di Mauro, Pesaran and L.V. Smith (2007) DdPS with surveys of applications given in Di Mauro & Pesaran (2013) and Chudik & Pesaran (2016). Although this approach is described in terms of countries, the GVAR structure can be applied to a range of other types of units to approximate an IVAR. Busiere et al. (2011) give an example of modelling over 60 inter-dependent real effective exchange rates to examine whether the introduction of the Euro made a difference to the transmission of shocks.

The various approaches can be combined, Cuaresma, Feldkircher Huber (2014) examine different priors in a Bayesian GVAR.

The estimates discussed are those of DdPS and the exposition partly follows Pesaran and Smith (2006). Dees, Pesaran, Smith and Smith (2009, 2014) relate this structures to the solution of open-economy dynamic stochastic general equilibrium models. We first look at the country specific VARS, which include exogenous foreign variables, proxying global factors. The inclusion of these foreign variables, called star variables, turn the VAR into a VARX*. Then we examine how the VARX* systems for each country can be put together into a GVAR, where y_t is a 134×1 vector (roughly 5 variables for each of 26 countries), there are 63 cointegrating vectors and 71 stochastic trends. Clearly unrestricted estimation of this system is not possible. The restrictions come from a specific factor model.

9.1 VARX*

Suppose there are a set of countries $i = 0, 1, 2, ..., N$, with country 0, say the US, as the numeraire country. The objective in the first stage is to model a particular country, say i, using a VARX*. The second stage puts the indiviudal country models into the GVAR. As an example, a second-order country-specific

VARX*(2,2) model with deterministic trends can be written as

$$\mathbf{x}_{it} = \mathbf{B}_{id}\mathbf{d}_t + \mathbf{B}_{i1}\mathbf{x}_{i,t-1} + \mathbf{B}_{i2}\mathbf{x}_{i,t-2} + \mathbf{B}_{i0}^{*}\mathbf{x}_{it}^{*} + \mathbf{B}_{i1}^{*}\mathbf{x}_{i,t-1}^{*} + \mathbf{B}_{i2}^{*}\mathbf{x}_{i,t-2}^{*} + \mathbf{u}_{it}, \quad (46)$$

where $\mathbf{x}_{it}$ is a $k_i \times 1$ (usually five or six) vector of domestic variables, $\mathbf{x}_{it}^{*}$, a $k_i^{*} \times 1$ vector of foreign variables specific to country i, and $\mathbf{d}_t$ a $s \times 1$ vector of deterministic elements as well as observed common variables such as oil prices, typically $(1, t, p_t^o)'$, but could contain seasonal or break dummy variables. The $\mathbf{x}_{it}^{*}$ are calculated as country specific trade weighted averages of the corresponding variables of the other countries

$$\mathbf{x}_{it}^{*} = \sum_{j=0}^{N} w_{ij}\mathbf{x}_{jt}, \text{ with } w_{ii} = 0,$$

where w_{ij} is the share of country j in the trade (exports plus imports) of country i. The GVAR uses trade weights, but many other weights are possible.

The $\mathbf{x}_{it}^{*}$ are treated as weakly exogenous, an assumption found acceptable, when tested. The VARX* models can be estimated separately for each country, taking into account the possibility of cointegration both within $\mathbf{x}_{it}$ and between $\mathbf{x}_{it}$ and $\mathbf{x}_{it}^{*}$. The foreign variables, $\mathbf{x}_{it}^{*}$ would typically contain the same variables as the domestic variables, $\mathbf{x}_{it}$, thus there is a symmetrical structure to the model which can be given an economic interpretation.

The cointegrating VARX* can be written as a VECM

$$\Delta\mathbf{x}_{it} = \mathbf{B}_{id}\mathbf{d}_t - \mathbf{\Pi}_i\mathbf{z}_{i,t-1} + \mathbf{B}_{i0}^{*}\Delta\mathbf{x}_{it}^{*} + \mathbf{\Gamma}_i\Delta\mathbf{z}_{i,t-1} + \mathbf{u}_{it},$$

where $\mathbf{z}_{it} = (\mathbf{x}_{it}', \mathbf{x}_{it}^{*\prime})'$. Restricting the deterministic and assuming that $rank(\mathbf{\Pi}_i) = r_i < k_i + k_i^{*}$, we have $\mathbf{\Pi}_i = \boldsymbol{\alpha}_i\boldsymbol{\beta}_i'$, where $\boldsymbol{\beta}_i$ is the $(k_i + k_i^{*}) \times r_i$ matrix of the cointegrating coefficients and

$$\Delta\mathbf{x}_{it} = -\boldsymbol{\alpha}_i\boldsymbol{\beta}_i'(\mathbf{z}_{i,t-1} - \boldsymbol{\Upsilon}_i\mathbf{d}_{t-1}) + \mathbf{B}_{i0}^{*}\Delta\mathbf{x}_{it}^{*} + \mathbf{\Gamma}_i\Delta\mathbf{z}_{i,t-1} + \mathbf{\Pi}_i\boldsymbol{\Upsilon}_i\Delta\mathbf{d}_t + \mathbf{u}_{it}, \quad (47)$$

The r_i error correction terms of the model can now be written as

$$\boldsymbol{\xi}_{it} = \boldsymbol{\beta}_i'\mathbf{z}_{it} - \boldsymbol{\beta}_i'\boldsymbol{\Upsilon}_i\mathbf{d}_t = \boldsymbol{\beta}_{ix}'\mathbf{x}_{it} + \boldsymbol{\beta}_{ix*}'\mathbf{x}_{it}^{*} + \boldsymbol{\gamma}_i'\mathbf{d}_t,$$

The $\boldsymbol{\xi}_{it}$ are mean zero $r_i \times 1$ vector of disequilibrium deviations from the long run relationships. In the case of small open economies it is reasonable to assume that the country specific foreign variables are "long run forcing" or $I(1)$ weakly exogenous, and then estimate the VARX* models separately for each country conditional on $\mathbf{x}_{it}^{*}$, taking into account the possibility of cointegration both within $\mathbf{x}_{it}$ and across $\mathbf{x}_{it}$ and $\mathbf{x}_{it}^{*}$. Forecasts and counter-factuals are invariant to the just-identifying restrictions that are used.

All the 26 country-specific (the Euro area being treated as a country) models in the GVAR of DdPS are estimated over 1979Q4-2003Q4 and the lag orders are selected by AIC separately for each country up to a maximum of 2. Different applications have different sets of variables, but in the DdPS version $\mathbf{x}_{it}$ are a subset of the logarithm of real output, inflation, π the exchange rate variable, which is defined as $e_{it} - p_{it}$, where e_{it} is the logarithm of the nominal exchange rate against the dollar; a short interest rate, a long interest rate, and the logarithm of real equity prices, q_{it}. The variables included in the different country models are not always the same, e.g. there are no equity price or long-term interest rate data for some.

As noted above, it is straightforward to test the weak exogeneity assumption for the long-run parameters, of the country specific foreign variables because there are a small number of them. This simply involves running the regressions

$$\Delta \mathbf{x}_{it}^* = \mathbf{c}_{i0}^* + \boldsymbol{\alpha}_i^* \boldsymbol{\xi}_{i,t-1} + \boldsymbol{\Gamma}_i^* \Delta \mathbf{z}_{i,t-1} + \mathbf{u}_{it}^*,$$

and testing that $\boldsymbol{\alpha}_i^* = 0$.

9.2 GVAR

Although estimation is done on a country by country basis, the GVAR model is solved for the world as a whole, taking account of the fact that all the variables are endogenous to the system as a whole. To do this write (46) as

$$\mathcal{A}_{i0}\mathbf{z}_{it} = \mathbf{h}_{i0} + \mathbf{h}_{i1}t + \mathcal{A}_{i1}\mathbf{z}_{it-1} + \mathcal{A}_{i2}\mathbf{z}_{it-2} + \mathbf{u}_{it}, \tag{48}$$

for $i = 0, 1, 2, ..., N$ where $\mathbf{z}_{it}$

$$\mathbf{z}_{it} = \begin{pmatrix} \mathbf{x}_{it} \\ \mathbf{x}_{it}^* \end{pmatrix},$$

and

$$\mathcal{A}_{i0} = (\mathbf{I}_{k_i}, -\mathbf{B}_{i0}^*), \quad \mathcal{A}_{i1} = (\mathbf{B}_{i1}, \mathbf{B}_{i1}^*), \quad \mathcal{A}_{i2} = (\mathbf{B}_{i2}, \mathbf{B}_{i2}^*).$$

The dimensions of $\mathcal{A}_{i0}$, $\mathcal{A}_{i1}$ and $\mathcal{A}_{i2}$ are $k_i \times (k_i + k_i^*)$ and $\mathcal{A}_{i0}$ has full column rank, namely $Rank(\mathcal{A}_{i0}) = k_i$. Also note that

$$\mathbf{z}_{it} = \mathbf{W}_i \mathbf{x}_t,$$

where $\mathbf{x}_t = (\mathbf{x}_{0t}', \mathbf{x}_{1t}', ..., \mathbf{x}_{Nt}')'$ is the $k \times 1$ vector which collects all the endogenous variables of the system, and $\mathbf{W}_i$ is the $(k_i + k_i^*) \times k$ matrix defined by the trade weights w_{ij}. Using this (48) can be written as

$$\mathcal{A}_{i0}\mathbf{W}_i\mathbf{x}_t = \mathbf{h}_{i0} + \mathbf{h}_{i1}t + \mathcal{A}_{i1}\mathbf{W}_i\mathbf{x}_{t-1} + \mathcal{A}_{i2}\mathbf{W}_i\mathbf{x}_{t-2} + \mathbf{u}_{it}, \text{ for } i = 0, 1, 2, ..., N, \tag{49}$$

and the systems stacked to yield the model for $\mathbf{x_t}$

$$\mathbf{H}_0\mathbf{x}_t = \mathbf{h}_0 + \mathbf{h}_1\mathbf{t} + \mathbf{H}_1\mathbf{x}_{t-1} + \mathbf{H}_2\mathbf{x}_{t-2} + \mathbf{u}_t,$$

where

$$\mathbf{H}_j = \begin{pmatrix} \mathcal{A}_{0j}\mathbf{W}_0 \\ \mathcal{A}_{1j}\mathbf{W}_1 \\ \vdots \\ \mathcal{A}_{Nj}\mathbf{W}_N \end{pmatrix}, \ \mathbf{h}_j = \begin{pmatrix} \mathbf{h}_{0j} \\ \mathbf{h}_{1j} \\ \vdots \\ \mathbf{h}_{Nj} \end{pmatrix}, \ \mathbf{u}_t = \begin{pmatrix} \mathbf{u}_{0t} \\ \mathbf{u}_{1t} \\ \vdots \\ \mathbf{u}_{Nt} \end{pmatrix},$$

for $j = 0, 1, 2$. Since $\mathbf{H}_0$ is a known non-singular matrix that depends on the trade weights and parameter estimates, we can obtain the GVAR

$$\mathbf{x}_t = \mathbf{a}_0 + \mathbf{a}_1 \mathbf{t} + \mathbf{G}_1 \mathbf{x}_{t-1} + \mathbf{G}_2 \mathbf{x}_{t-2} + \mathbf{v}_t, \tag{50}$$

where $\mathbf{G}_j = \mathbf{H}_0^{-1}\mathbf{H}_j$, $\mathbf{a}_j = \mathbf{H}_0^{-1}\mathbf{h}_j$, for $j = 0, 1, 2$, and $\mathbf{v}_t = \mathbf{H}_0^{-1}\mathbf{u}_t$. The GVAR can be solved recursively and used for a variety of purposes.

There are no restrictions on the covariance matrix $\mathbf{\Sigma} = \mathbf{E}(\mathbf{v}_t \mathbf{v}_t')$. For each country we have a $k_i \times 1$ vector of estimated residuals $\widehat{\mathbf{u}}_{it}$ from which can be calculated $\widehat{\mathbf{v}}_{it}$ and the elements of the covariance matrix are estimated freely by the $k_i \times k_j$ matrix $\widehat{\mathbf{\Sigma}}_{ij} = \sum_t \widehat{\mathbf{v}}_{it}\widehat{\mathbf{v}}_{jt}'/T$.

The US, the reference country, is treated differently from the other countries. Oil prices are included in the US model as an endogenous variable but included in other country models as weakly exogenous. Exchange rates (in terms of US dollars) are included as endogenous variables in all country models except for the US model. Also all foreign variables are included in the non-US models as weakly exogenous variables, but only foreign real output and foreign inflation are included as weakly exogenous in the US model.

It would be impossible to estimate a VAR of this size, without restrictions: 134 endogenous variables 71 stochastic trends and 63 cointegrating relations. All the roots of the GVAR either lie on or inside the unit circle. The weak exogeneity/long run forcing assumption is rejected only in 5 out of 153 cases. DdPS report the results for various tests of structural stability, the critical values of which are computed using the sieve bootstrap samples obtained from the solution of the GVAR. Evidence of structural instability is found primarily in the error variances (47% of the equations - clustered in the period 1985-1992). Although linear with a simple overall structure, this is a large and complicated model which allows for a large degree of interdependence. There are three routes for between country interdependence: through the impact of the $\mathbf{x}_{it}^*$ variables, oil prices, and through the error covariances. The effects through the $\mathbf{x}_{it}^*$ are generally large, shocks to one country have marked effects on other countries. The between country error covariances are quite small, with the exception of those for the real exchange rate equations, because of the base-country effect, since they are all expressed against the US dollar and this factor has not been removed since, naturally, there is no star exchange rate variable. However, when effective rates are used, which do not reflect the dollar factor, the covariances in the exchange rate equation are close to zero.

An alternative approach would be to estimate the weights as factor loadings directly, e.g. by constructing the factors as principal components, extracted from the pooled set of all the variables in the world economy, or across a given geographical region. In many cases it is difficult to give these estimated factors an economic interpretation. This is a particular problem when there are many variables for many countries, since it may not be obvious how to identify the factors. One has to choose the number of factors and the estimation may induce errors and principal component methods seem to perform worse than *a priori* weights in Monte Carlo studies in a panel context, e.g. Kapetanios and Pesaran

(2005). There is also the problem that a factor which is crucial for one country or region may account for small part of global variance and get ignored, which country specific trade weights avoids.

10 A multi-country new Keynesian (MCNK) Rational Expectations Model

Dees et al (2014) construct a system of small dynamic stochastic general equilibrium, DSGE, models, in which the variables are measured as deviations from steady states. The steady states are long-horizon forecasts of the variables (absent deterministic trends), obtained from the GVAR. These are estimates of multivariate Beveridge-Nelson trends, incorporating stochastic trends and cointegrating vectors, Garratt, Robertson and Wright (2006) . This corresponds to the economic concept of a steady state, the equilibrium to which the system would return in the absence of shocks. Whereas univariate BN trends are not smooth enough, this is not the case with multivariate BN trends. The deviations from steady states (e.g. output gaps), will be $I(0)$ by construction and can be used in DSGE type structural modelling. Unlike the Hodrick-Prescott filter or other purely statistical approaches to trend/cycle decompositions the BN decompositions depend on the cointegrating properties of the GVAR and the long run theory that underlie them.

The MCNK model has 131 variables for 33 countries: inflation, $\widetilde{\pi}_{it}$, output, $\widetilde{y}_{it}$, interest rates $\widetilde{r}_{it}$ (except for Saudi Arabia) and real effective exchange rate $\widetilde{re}_{it}$, except for the US (numeraire) $re_{it} = \sum_{j=0}^{N} w_{ij}(e_{it} - e_{jt}) + \sum_{j=0}^{N} w_{ij}p_{jt} - p_{it}$..For each variable there are country-specific foreign trade weighted averages, e.g. $\widetilde{y}_{it}^{*} = \Sigma_{j=0}^{N} w_{ij}\widetilde{y}_{jt}$.

The MCNK model equations are a Phillips Curve (PC)

$$\widetilde{\pi}_{it} = \beta_{ib}\widetilde{\pi}_{i,t-1} + \beta_{if}E_{t-1}\left(\widetilde{\pi}_{i,t+1}\right) + \beta_{iy}\widetilde{y}_{it} + \varepsilon_{i,st},\ i = 0, 1, ..., N,$$

An open economy IS curve

$$\widetilde{y}_{it} = \alpha_{ib}\widetilde{y}_{i,t-1} + \alpha_{if}E_{t-1}\left(\widetilde{y}_{i,t+1}\right) + \alpha_{ir}[\widetilde{r}_{it} - E_{t-1}\left(\widetilde{\pi}_{i,t+1}\right)] + \alpha_{ie}\widetilde{re}_{it} + \alpha_{iy*}\widetilde{y}_{it}^{*} + \varepsilon_{i,dt},\ i = 0, 1, ..., N.$$

Taylor rule (TR)

$$\widetilde{r}_{it} = \gamma_{ib}\widetilde{r}_{i,t-1} + \gamma_{i\pi}\widetilde{\pi}_{it} + \gamma_{iy}\widetilde{y}_{it} + \varepsilon_{i,mt},\ i = 0, 1, ..., N.$$

AR for real exchange rate,

$$\widetilde{re}_{it} = \rho_i\widetilde{re}_{i,t-1} + \varepsilon_{i,et},\ \ |\rho_i| < 1,\ i = 1, 2, ..., N.$$

Estimation is by instrumental variables subject to sign restrictions, e.g. α_{ir}, non positive. The foreign variables aid identification and provide instruments. A determinate solution to the RE model is obtained. The Errors correspond to supply, demand and monetary policy shocks. In the covariance matrix it is assumed that shocks of the same type (e.g. supply shocks) are correlated across countries, but shocks of different types (e.g. supply and demand shocks) are not. It is used to conduct a range of simulations..

11 Concluding comments

11.1 General points

It should be emphasised, that this area is developing very rapidly, with very many interesting and often surprising results emerging. There is also a general pattern of extending issues in the standard time-series literature to panels. Sometimes they extend relatively straightforwardly, sometime not because problems interact. For instance, omitted factors cause apparent structural breaks. Bannerjee and Carrion-I-Silvestre (2015) provide a test for cointegration that allows for both structural breaks and cross-section dependence.

It should also be remembered that theoretical and applied econometrics are very different activities. Theoretical econometrics is a deductive activity where you have no data, know the model and derive properties of estimators and tests conditional on that model. There are right and wrong answers. Applied econometrics is an inductive activity where you do have data, but do not know the model or the questions let alone the answers. The chapter on applied econometrics in Kennedy (2003) is good on these aspects. In applied econometrics one must take account of not merely the statistical theory but also the purpose of the activity and the economic context, which define the parameters of interest. Different models may be appropriate for different purposes, such as forecasting, policy analysis or testing hypotheses and purpose and the.economic context (theory, history, instiutions) should guide the choice of model.

However, even given this there appear to be some general points that applied workers might bear in mind when using large N large T panels.

First, there are a large number of different estimators, differing for instance in their assumptions about heterogeneity, dynamics or cross-section dependence. The dynamic parameters of standard pooled estimators are subject to large potential biases when the parameters differ across groups and the regressors are serially correlated. However, for some purposes, such as forecasting (where parsimony is crucial) or estimating long-run parameters (where the biases may cancel), the pooled estimators may perform well. In these circumstances, it is desirable to use various estimators and if they give very different estimates interpret why they do so.

Second, pooled (or cross-section) regressions can be measuring very different parameters from the averages of the corresponding parameters in time-series regressions. In many cases this difference can be expressed as a consequence of a dependence between the time-series parameters and the regressors. The interpretation of this difference will depend on the theory related to the substantive application. It is not primarily a matter of statistical technique.

Third, the mechanical application of panel unit-root or cointegration tests is to be avoided. To apply these tests requires that the hypotheses involved are interesting in the context of the substantive application, which is again a question of theory rather than statistics.

Fourthly, it is important to test and allow for between group dependence, the CCE estimator is a good start, but you may also need to be able to give the estimates an economic interpretation, which can be difficult.

11.2 A check list of questions

1. Why are you doing this? Purpose is crucial in determining parameters of interest and appropriate estimators, e.g. a good forecasting model may be quite different from a good structural model..
2. Do you know what the variables measure and how they measure it?
3. Have you examined the data carefully?
4. What does economic theory, history and context tell you?
5. Are the data best interpreted as cross-sections or time-series?
6. What do N and T allow you to do?
7. For this N and T what are the properties of the estimators and tests?
8. How different are the different estimators? Can you explain the differences between the estimators?
9. How do you calculate the standard errors?
10. How much structure can you put on the problem?
11. Single equation or system? Structural system or reduced form?
12. Determination of lag lengths and treatment of deterministic terms?
13. How much parameter heterogeneity is there? In what dimensions?
14. What are the orders of integration of the data?
15. If I(1), is there homogeneous, heterogeneous or no cointegration?
16. How many cointegrating vectors? How do you identify the long-run relations?
17. If using a structural model, can you identify the short-run relations?
18. Is there cross-section dependence? How do you interpret it?
19. Are the equations well specified, according to diagnostic tests?
20. Can you interpret the results?

12 References

Abadie, Alberto and Diamond, Alexis and Hainmueller, Jens (2010) Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program, *Journal of the American Statistical Association*,105(490) 493–505}

Abadie, Alberto and Diamond, Alexis and Hainmueller, Jens (2015) Comparative politics and the synthetic control method, *American Journal of Political Science,* 59(2), 495-510.

Abadie, Alberto and Gardeazabal, Javier (2003) The Economic Costs of Conflict: A Case Study of the Basque Country, American Economic Review, 93(1) 113-132.

Aksoy, Y., H. Basso T. Grasl & R P. Smith (2015). Demographic Structure and Macroeconomic Trends, Birkbeck Working Papers in Economics and Finance 1501.

Alvarez, J., & M. Arellano (2003) The Time Series and Cross Section Asymptotics of Dynamic Panel Data Estimators, *Econometrica*, 71, 1121-1159.

Ando T and Bai, J. (2014) A simple new test for slope homogeneity in panel data models with interactive effects, MPRA paper 60795.

Ando T and Bai, J. (2016) panel data models with grouped factor structure under unknown group membership, *Journal of Applied Econometrics, 31(1) 163-191.*

Angrist, J. D., O. Jorda and G. Kuersteiner (2013): Semiparametric estimates of monetary policy effects: string theory revisited, NBER Working Paper No. 19355.

Arellano, M (2003) *Panel Data Econometrics*, Oxford University Press.

Attanasio, O.P., L. Picci & A. Scorcu (2000) Saving Growth and Investment: A Macroeconomic Analysis using a Panel of Countries, *Review of Economics and Statistics*, LXXXII, May, 182-211.

Bai, J (2003) Inferential Theory for Factor Models of Large Dimensions, *Econometrica*, 71(1) 135-172.

Bai, J (2004) Estimating cross-section common stochastic trends in non-stationary panel data, *Journal of Econometrics*, 122, 137-183.

Bai, J. (2009) Panel Data models with interactive fixed effects, *Econometrica*, 77(4) 1229-1279.

Bai, J. (2010) Common breaks in means and variances for panel data, *Journal of Econometrics*, 157, 78-92.

Bai, J (2013) Fixed Effects Dynamic Panel Models, A factor analytic method, *Econometrica*, 81, 285-314.

Bai, J & Ng, S (2002) Determining the number of factors in approximate factor models, *Econometrica*, 70( ) 191-221.

Bai, J. C. Kao & S. Ng (2009) Panel cointegration with global stochastic trends, *Journal of Econometrics* 149, 82-99.

Bai, J & Ng, S (2004) A PANIC attack on unit roots and cointegration, *Econometrica,* 72(4) 1127-1178.

Bai, J & Ng, S (2010) Panel unit root tests with cross-section dependence: a further investigation, *Econometric Theory*, 26, 1088-1114.

Bai, Y. & J. Zhang (2010) Solving the Feldstein-Horioka Puzzle with Financial Frictions, *Econometrica* 78(2) 603-632.

Bailey, Natalia Sean Holly, and M. Hashem Pesaran (2016) A Two Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-Sectional Dependence, *Journal of Applied Econometrics, 31(1) 249-280*.

Bailey, N., G. Kapetanios & M.H. Pesaran (2015) Exponent of cross-sectional dependence: estimation and inference, forthcoming *Journal of Applied Econometrics*.

Bailey, Natalia , M. Hashem Pesaran and L. Vanessa Smith (2015), "A Multiple Testing Approach to the Regularisation of Large Sample Correlation Matrices",

Balazsi, L. L. Matyas and T. Wansbeek (2015) The estimation of multidimensional fixed effects panel data models, CESifo working paper 5251.

Baltagi, B.H. (2008) *Econometric Analysis of Panel Data*, 4th edition New York: Wiley.

Baltagi, B.H. & J.M. Griffin (1984) Short and long-run effects in pooled models, *International Economic Review*, 25, p631-645.

Baltagi, B.H. & J.M. Griffin (1997) Pooled Estimators versus their Heterogeneous Counterparts in the context of dynamic demand for gasoline, *Journal of Econometrics*, 77, 303-327.

Baltagi, B.H,.J.M. Griffin & W. Xiong (2000) To pool or not to pool: homogeneous versus heterogeneous estimators applied to cigarette demand, *Review of Economics and Statistics*, 82, 117-126.

Baltagi B H, G Bresson, James Griffin & Alaine Pirotte (2003) Homogeneous, heterogeneous or shrinkage estimators? Some empirical evidence from French regional gasoline consumption, *Empirical Economics*, 28, 795-811.

Baltagi B H, G Bresson & A Pirotte (2003) Fixed effects, random effects or Hausman-Taylor? A pre-test estimator, *Economics Letters*, 79 p361-369.

Baltagi, B.H, G. Bresson & A Pirotte (2008) To pool or not to pool, p517-546 of Matyas & Sevestre eds *The Econometrics of Panel data*, 3rd edition, Springer-Verlag, Berlin.

Banbura, M. D. Giannone & L. Reichlin (2010) Large Bayesian Vector Autoregressions, *Journal of Applied Econometrics*, 25(1) 71-92.

Banerjee, A (1999) Panel Data, Unit Roots and Cointegration: An Overview, *Oxford Bulletin of Economics and Statistics*, Special Issue on Testing for Unit Roots and Cointegration using Panel Data, Theory and Applications, 61, November, 607-629.

Banerjee, A & J.L. Carrion-i-Silvestre (2015) Cointegration in panel data with structural breaks and cross-section dependence, *Journal of Applied Econometrics*, 30(1), 1-23

Banerjee, A., M.Eberhardt & J.J. Reade (2010) Panel Estimation for Worriers, manuscript.

Banerjee, A., M. Marcellino, & C. Osbat (2004) Some Cautions on the Use of Panel Methods for Integrated Series of Macro-Economic Data, *Econometrics*

Journal, 7, 322-340.

Banerjee, A & M. Marcellino (2009) Factor-Augmented Error Correction Models, in The Methodology and Practice of Econometrics, ed J. Castle, & neil Shephard OUP &CEPR Discussion Paper Series, No 6707.

Bernanke, B.S. J. Boivin & P. Eliasz (2005) Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) approach, *Quarterly Journal of Economics*, Feb 387-422.

Binder, M., C. Hsiao & M.H. Pesaran (2005) Estimation and Inference in Short Panel Vector Autoregression with Unit Roots and Cointegration, *Econometric Theory* 21 (4) 795-837.

Boivin, J., & M. Giannoni (2006) DSGE models in a data rich environment, NBER working paper t0332.

Bond, S. A. Lebleicoglu & F.Schiantarelli (2010) Capital Accoumulation and Growth, a new look at the empirical evidence, *Journal of Applied Econometrics*, 25(7) 1073-1099.

Bond, S. & M. Eberhardt (2013) Accounting for unobserved heterogeneity in panel time series models.

Bonhomme, S & E. Manresa (2012) Grouped patterns of heterogeneity in panel data, CEMFI working paper 1208.

Bove, V, L. Elia and R.P. Smith (2016) On the heterogenous consequences of civil war *Oxford Economic Papers* forthcoming.

Boyd, D & R.P. Smith (2002) Some Econometric Issues in Measuring the Monetary Transmission Mechanism, with an application to Developing Countries, p68-99 of *Monetary Transmission in Diverse Economies*, ed Lavan Mahadeva and Peter Sinclair, Cambridge University Press.

Breitung, J.M. & Meyer (1994) Testing for unit roots using panel data: are wages on different bargaining levels cointegrated, *Applied Economics*, 26, 353-361.

Breitung J. & S. Das (2008) Testing for unit roots in panels with a factor structure, *Econometric Theory*, 24, p88-106.

Breitung J., & M.H. Pesaran (2008) Unit Roots and Cointegration in Panels, in L. Matyas and P. Sevestre, *The Econometrics of Panel Data* (Third Edition), Kluwer Academic Publishers.

Bruno, G.S.F. (2005) Approximating the Bias of the LSDV estimator for dynamic unbalanced panel data models, *Economics Letters*, 87, 361-366.

Brush, J. (2007) Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of US countries. *Economics Letters*, 96 p264-268.

Bun, M.J.G. & J.F. Kiviet (2003) On the diminishing returns of higher order terms in asymptotic expansions of bias, *Economics Letters*, 79, 145-152.

Bun, M.J.G. & M.A. Caree (2006) Bias-corrected estimation in dynamic panel data models with heteroskedasticity, *Economics Letters*, 92, p220-227.

Bussiere, M. A. Chudik & A. Mehl (2011) Does the Euro make a difference: spatio-temporal transmission of global shocks to real effective exchange rates in an infinite VAR, ECB working paper 1292.

Canova, F (2007) *Methods for Applied Macroeconomic Research*, Princeton.

Canova F & M Ciccarelli (2013) Panel vector autoregressive models: a survey, ECB working paper 2013, In *VAR models in macroecomics - New Developments and Applications: Essays in honor of Christopher A Sims*, Vol 32 of Advances in Econometrics, ed Fornby, Killian and Murphy, Emerald.

Cerrato, M., C. dep Peretti, & N. Sarantis (2007) A non-linear panel unit root test under cross-section dependence, London Metropolitan University.

Chan, M.K. & S. Kwok (2016) Policy Evaluation with Interactive Fixed Effects. University of Sydney Economics Working paper 2016-11.

Christiano, Lawrence J. & Eichenbaum, Martin, 1990. "Unit roots in real GNP: Do we know, and do we care?," *Carnegie-Rochester Conference Series on Public Policy*, 32(1), 7-61, January.

Chudik A & M.H. Pesaran, (2011) Infinite dimensional VARs and Factor models, ECB Working paper No 998 *Journal of Econometrics*, vol 163, pp. 4-22.

Chudik, A & M.H. Pesaran (2013) Econometric analysis of high dimension VARs featuring a dominant unit, *Econometrics Review*, vol 32, pp. 592-649.

Chudik, and M. H Pesaran, (2015a) Common Correlated Effects Estimation of Heterogeneous Dynamic Panel Data Models with Weakly Exogenous Regressors *Journal of Econometrics.*188(2), 393-420.

Chudik, A & M.H. Pesaran (2015b) Large Panel Data Models with Cross-Sectional Dependence: A Survey, CESifo WP Number 4371, in *The Oxford Handbook on Panel Data* edited by B. H. Baltagi, Oxford University Press.

Chudik, A & M.H. Pesaran (2016) Theory and Practice of GVAR Modeling, *Journal of Economic Surveys*, 30(1) 165-197.

Chudik, A. M.H. Pesaran & E Tosetti (2011) Weak and strong cross-section dependence and estimation of large panels, *Econometrics Journal 14, C45-C90.*

Chudik, A. Kamiar Mohaddes, M. Hashem Pesaran and Mehdi Raissi, (2016) Long-Run Effects in Large Heterogenous Panel Data Models with Cross-Sectionally Correlated Errors", forthcoming in *Advances in Econometrics* V36, Essays in honor of Aman Ullah.

Coakley, J., A.-M. Fuertes & F. Spagnolo (2004) Is the Feldstein-Horioka puzzle history? *The Manchester School*, 72(5), 569-590.

Coakley J, A-M Fuertes & R.P. Smith (2002) A Principal Components approach to cross-section dependence in panels,10th international conference on panel data, Berlin 2002..

Coakley J, A-M Fuertes & R.P. Smith (2006) Unobserved heterogeneity in panel time series models, *Computational Statistics and Data Analysis,* 50 (9) 2361-2380.

Coakley, J. R.P. Flood, A M. Fuertes & M.P. Taylor (2005) Purchasing Power parity and the theory of general relativity: the first tests, *Journal of International Money and Finance*, 24 293-316.

Cochrane, J.H. (2001) *Asset Pricing*, Princeton University Press.

Costalli, Stefano and Moretti, Luigi and Pischedda, Costantino (2014) The Economic Costs of Civil War: Synthetic Counterfactual Evidence and the Effects of Ethnic Fractionalization, HICN working paper,184.

Cuaresma, J.C. M. Feldkircher & F. Huber (2014) Forecasting with Bayesian Global Vector Autoregressive Models: a comparison of priors, ONB working paper 189

Cubadda, G. A. Hecq & Franz Palm (2008) Macro-panels and reality, *Economics Letters* 99, 537-540.

Dhaene, G. and Koen Jochmans (2015), *Split-panel jackknife estimation of fixed-effect models*, The Review of Economic Studies, forthcoming.

Dhaene, Geert & Koen Jochmans (2015). Bias-corrected estimation of panel vector autoregressions,.hal-01174330.

Dees, S., F. di Mauro, M.H. Pesaran & L.V. Smith (2007) Exploring the international linkages of the Euro area: a global VAR analysis, *Journal of Applied Econometrics*, 22(1), 1-38.

Dees, S., M.H. Pesaran, L.V. Smith & R.P. Smith (2009) Identification of New Keynesian Phillips Curves from a Global Perspective, *Journal of Money Credit and Banking*, 41(7), 1481-1502.

Dees, S., M.H. Pesaran, L.V. Smith & R.P. Smith (2014) Constructing Multi-Country Rational Expectations Models, *Oxford Bulletin of Economics and Statistics*, 76(6) 812–840.

De Silva, S. K. Hadri & R. Tremayne (2009) Panel unit root tests in the presence of cross-sectional dependence: finite sample peformance and an application, *The Econometrics Journal*, 12, 340-366.

Dhaene G and K Jochmans (2012) Split-panel jackknife estimation of fixed effects models, *Review of Economic Studies*, 83(3) 991-1030.

Diebold, F.X. G. Rudebusch & S.B. Arouba (2006) The Macroeconomy and the Yield Curve: a dynamic latent factor approach, *Journal of Econometrics*, 131, 309-338.

Di Mauro, F & M.H. Pesaran eds. (2013) *The GVAR handbook*: Oxford University Press.

Eberhardt, M & F Teal, (2010) Econometrics for Grumblers: A new look at cross-country growth empirics, *Journal of Economic Surveys*, forthcoming.

Elliott, G., T.J. Rothenberg, & J.H. Stock (1996) Efficient Tests for an Autoregressive Unit Root, *Econometrica,* 64, 813-816.

Elliott, G & A. Timmerman (2008) Economic Forecasting, *Journal of Economic Literature*, XLVI(1) 3-56.

Engle, R.F. & C.W.J. Granger (1987) Cointegration and Error Correction: representations, estimation and testing, *Econometrica,* 55, 252-276.

Favero, C.A. M. Marcellini & F. Neglia (2005) Principal Components at Work: The empirical analysis of monetary policy with large data sets, *Journal of Applied Econometrics*, 20 p603-620.

Fernandez-Val, I and M Weidner (2015) Individual and time effects in nonlinear panel models with large N,T, cemmap working paper 17/15.

Forni, M. M.Hallin, M.Lippi & L. Reichlin (2000) The generalised factor model: identification and estimation, *Review of Economics and Statistics*, 82, p540-54.

Forni, M. M.Hallin, M.Lippi & L. Reichlin (2003) The generalised factor model: one sided estimation and forecasting, LEM working paper series 2003/13.

Forni, M. M.Hallin, M.Lippi & L. Reichlin (2005) The generalised factor model, *Journal of the American Statistical Association*, 100, 830-840.

Garratt, A., K. Lee, M.H. Pesaran & Y Shin (2007) *Global and National Macroeconometric Modelling: A Long Run Structural Approach*, Oxford University Press.

Garratt, A. D. Robertson & S. Wright (2006) Permanent versus transitory components and economic fundamentals, *Journal of Applied Econometrics*, 21, 521-542.

Gengenbach, C. J-P Urbain & J. Westerlund (2015) Error correction testing in panels with common stochastic trends, manuscript.

Gengenbach, C. F.C. Palm & J-P Urbain (2009) Panel unit root tests in the presence of cross-sectional dependencies: comparisons and implications for modelling, *Econometric Reviews*, 29 111-145

Gobillon, L. & T. Magnac (2016) Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls, *Review of Economics and Statistics*, 98(3) 535-551..

Gourieroux, C., P.C.B. Philips & J. Yu (2010) Indirect Inference for Dynamic Panel Models, *Journal of Econometrics* 157 p68-77.

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

Gutierrez, Luciano (2003) On the power of panel cointegration tests: a monte carlo comparison, *Economics Letters* 80, p105-111.

Hahn, J. and G. Kursteiner (2002) Asymptotically unbiased inference for a dynamic panel with fixed effects when both n and T are large, *Econometrica* 70, 1639-1657.

Hall S, S Lavarova & G Urga (1999) A principal components analysis of common stochastic trends in heterogeneous panel data: some Monte Carlo Evidence, *Oxford Bulletin of Economics and Statistics* 61, 749-767.

Han, C, P.C.B. Phillips & D. Sul (2014) X-differencing and dynamic panel model estimation, *Econometric Theory 30(1), 201-251.*

Han C. & P.C.B. Phillips (2011) First difference MLE and dynamic panel estimation, Cowles Foundation DP 1780.

Harris D. D.I. Harvey, S.J. Leybourne & N.D. Sakkas (2010) Local Asymptotic Power of the Im-Pesaran-Shin panel unit root test and the impact of initial observations, *Econometric Theory*, 26, 311-324.

Hausman JA (1978) Specification tests in econometrics, *Econometrica*, 49, 1377-1398.

Hausman J A & W E Taylor (1981) Panel data and unobservable individual effects, *Econometrica*, 49, 1377-1398.

Hayakawa, K., M. H Pesaran and L. V. Smith (2014), Transformed Maximum Likelihood Estimation of Short Dynamic Panel Data Models with Interactive Effects CAFE Research Paper No. 14.06, May 2014

Heston, Alan Robert Summers & Bettina Aten (2009) Penn World Table Version 6.3, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, August 2009.

Holly, S, M.H. Pesaran & T. Yamagata (2010) A spatial temporal model of house prices in the USA, *Journal of Econometrics*, 158(1) 160-173.

Hsiao, C. (1986) *Analysis of Panel Data,* Cambridge: Cambridge University Press.

Hsiao, C. (2003) Analysis of Panel Data, 2/e, Cambridge: Cambridge University Press.

Hsiao, C., H.S. Ching, and K. W. Shui (2012), A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong kong with mainland China, Journal of Applied Econometrics, 27(5) p705–740.

Hsiao, C. & M.H. Pesaran (2008) Random Coefficient Panel Data Models, in Matyas and Sevestre (eds). *The Econometrics of Panel Data*, 3rd edition.

Hsiao, C., M.H. Pesaran & A. K. Tahmiscioglu (1999) Bayes Estimation of short-run Coefficients in Dynamic Panel Data Models, in *Analysis of Panels and Limited Dependent Variable Models: A Volume in Honour of G.S. Maddala* edited by C. Hsiao, K. Lahiri, L-F Lee and M.H. Pesaran, Cambridge: Cambridge University Press.

Hsiao, C., M.H. Pesaran & A. K. Tahmiscioglu (2002) Maximum Likelihood estimation of fixed effect dynamic panel data models covering short time-periods, *Journal of Econometrics*, 109, 107-150.

Hsiao, C., M.H. Pesaran & A. Pick (2012) Diagnostic Tests of Cross Section Independence for Limited Dependent Variable Panel Data Models", *Oxford Bulletin of Economics and Statistics*, vol 74, no 2, pp. 253-277.

Hsiao C & J Zhang (2015) IV, GMM or Likelihood approach to estimate dynamic panel models when either N or T or both are large, *Journal of Econometrics*, 187, 312-322.

Im, K.S., M.H. Pesaran & Y. Shin (2003) Testing for unit roots in heterogenous panels, *Journal of Econometrics,* 115, July, 53-74.

Imbs, J. H. Mumtaz, M.O. Ravn & H. Rey (2005) PPP Strikes back: aggregation and the real exchange rate, *Quarterly Journal of Economics*, CXX(1), 1-43.

Jang, M.J., & D. W. Shin (2005) Comparison of panel unit root tests under cross sectional dependence, *Economics Letters*, 89, 12-17.

Johansen, S. (1988) Statistical Analysis of Cointegrating Vectors, *Journal of Economic Dynamics and Control*, 12, 231-254.

Jorda O. & A.M.Taylor (2015) The time for austerity: estimating the average treatment effect of fiscal policy. *Economic Journal*, 126(1) 219-255.

Kao, C (1999) Spurious Regression and Residuals based tests for cointegration in panel data, *Journal of Econometrics*, 90, 1-44.

Kapetanios, G. (2007) Dynamic Factor Extraction of Cross-sectional Dependence in panel unit root tests, *Journal of Applied Econometrics*, 22, p313-338.

Kapetanios, G. J. Mitchell & Y. Shin (2014) A nonlinear panel data model of cross-section dependence, *Journal of Econometrics*, 179, 134-157.

Kapetanios, G. & M.H. Pesaran (2007) Alternative approaches to estimation and inference in large multifactor panels: small sample results with an application to modelling asset returns, in G. Phillips and E. Tzavalis (eds) *The*

Refinement of Econometric Estimation and Test Procedures, Cambridge University Press.

Kapetanios, G., M.H. Pesaran & T. Yamagata (2011) Panels with non-stationary multifactor error structures, Journal of Econometrics, 160(2) 326-348.

Kapetanios, G, Y. Shin & A Snell (2003) Testing for a Unit Root in the Nonlinear STAR framework, *Journal of Econometrics*, 112, 359-379.

Karlsson S & Lothgren M (2000) On the power and interpretation of panel unit root tests, *Economics Letters*, 66 p249-55.

Kaul, Ashok, Stefan Klossner, Gregor Pfeifer, Manuel Schieler (2015) Synthetic Control Methods: Never use all pre-intervention outcomes as economic predictors, mimeo

Kennedy, P. (2003) *A Guide to Econometrics*, 5th edition, Blackwell.

Kuersteiner, G.M. and I. R. Prucha (2015) Dynamic Spatial Panel Models: Networks, Common Shocks and Sequential Exogeneity, CESifo Working Paper No5445.

Kwiatkowski, D., P.C. Phillips, P. Schmidt & Y. Shin (1992) Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics*, 54, 159-178.

Lagana G. & A. Mountford (2005) Measuring Monetary Policy in the UK: A Factor-Augmented Vector Autoregression Model Approach, *Manchester School, Supplement*, 77-98.

Larsson, R., J. Lyhagen & M. Lothgren (2001) Likelihood-based Cointegration Tests in Heterogeneous panels, *Econometrics Journal*, 4 109-142.

Lazarova, S. L. Trapani & G. Urgan (2007) Common Stochastic Trends and Aggregation in Heterogeneous Panels, *Econometric Theory* 23, 89-105.

Lee, K., M.H. Pesaran & R.P. Smith (1997) Growth and Convergence in a Multi-country Empirical Stochastic Solow Model, *Journal of Applied Econometrics, 12, 4, 357-392.*

Levin, A. and C.-F. Lin & C.-S.J. Chu (2002) Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties, *Journal of Econometrics*, 108, 1-24.

Maddala, G.S. & I.-M. Kim (1998) *Unit Roots, Cointegration and Structural Change*, Cambridge: Cambridge University Press.

Maddala, G.S. & S. Wu (1999) A Comparative Study of Unit Root Tests with panel data and a new simple test, *Oxford Bulletin of Economics and Statistics*, Special Issue on Testing for Unit Roots and Cointegration using Panel Data, Theory and Applications, 61, November, 631-652.

Madsen, E. (2005) Estimating cointegrating relations from a cross section *Econometrics Journal*, 8, p380-405.

Matyas, L. & P. Sevestre (Eds.) (1996) *The Econometrics of Panel Data*, 2nd edition, Dordrecht: Kluwer Academic Publishers.

Matyas, L. & P. Sevestre (Eds.) (2005) *The Econometrics of Panel Data*, 3rd edition, Dordrecht: Kluwer Academic Publishers.

McCoskey, S. & C. Kao (1998) A Residual Based Test of the Null of Cointegration in Panel Data, *Econometric Reviews*, 17, 57-84.

McKenzie, D.J. (2004) Asymptotic theory for heterogeneous dynamic psuedo-panels, *Journal of Econometrics*, 120, 235-262.

Moon H.R. & B. Perron (2004) Testing for a unit root in panels with dynamic factors, *Journal of Econometrics*, 122, 81-126.

Moon H.R. & B. Perron (2007) An empirical analysis of nonstationarity in a panel of interest rates with factors, *Journal of Applied Econometrics*, 22, 383-400.

Moon H.R. & B. Perron (2012) Beyond panel unit root tests: using multiple testing to determine the non-stationarity properities of individual series in a panel. *Journal of Econometrics*, 169, 29-33.

Moon H.R. B. Perron and PCB Phillips (2007) Incidental trends and the power of unit root tests, *Journal of Econometrics*, 141, 416-459.

Moon H.R. B. Perron and PCB Phillips (2015) Incidental parameters and dynamic panel modelling, *Oxford Handbook of Panel Data.*

Moon HR and M Weidner (2015) Linear Regression for Panel with unknown number of factors as interactive fixed effects, *Econometrica* forthcoming.

Mundlack, Y. (2005) Economic Growth: lessons from two centuries of US agriculture, *Journal of Economic Literature*, Dec, p989-1024.

Nickel S (1981) Biases in dynamic models with fixed effects, *Econometrica*, 49, 1417-1426.

Oxley, L. & M. McAleer (1999) *Practical Issues in Cointegration Analysis*, Oxford: Basil Blackwell.

Onatski, Alexei (2009) Testing hypotheses about the number of factors in large factor models. *Econometrica* 77(5) 1447-1479.

Paap, R, W. Wang and X. Zhang (2015) To pool or not to pool: what is a good strategy?

Pedroni, P (1999) Critical values for cointegration tests in heterogenous panels with multiple regressors, *Oxford Bulletin of Economics and Statistics*, Special Issue on Testing for Unit Roots and Cointegration using Panel Data, Theory and Applications, 61, November, 653-670.

Pedroni, P. (2004) Panel Cointegration; Asymptotic and Finite Sample Properties of Pooled Time-series Tests with applications to the PPP hypothesis, *Econometric Theory*, 3, 579-625.

Perron, P (2006). Dealing with Structural Breaks, chapter 8, p278-352 of T.C. Mills and K. Patterson, *Palgrave Handbook of Econometrics*, Vol 1, Palgrave.

Pesaran, M.H. (2006) Estimation and Inference in Large Heterogeneous Panels with a multifactor error structure, *Econometrica,* 74(4) 967-1012.

Pesaran, M.H. (2007) A simple panel unit root test in the presence of cross section dependence, *Journal of Applied Econometrics,* 22(2).p265-312.

Pesaran, M.H. (2012) On the Interpretation of Panel Unit Root Tests", *Economics Letters*, vol 116, pp. 545-546.

Pesaran M.H. (2015) *Time-series and panel data econometrics for macroeconomics and finance*, Oxford University Press.

Pesaran, M.H. (2015a) Testing Weak Cross-Sectional Dependence in Large Panels *Econometric Reviews,* 34, p1089-1117.

Pesaran, M.H., N. U. Haque & S. Sharma (2000) Neglected heterogeneity and dynamics in Cross-country Savings Regressions, chapter 3 p53-82 of J. Krishnakumar & E. Rouchetti (Eds) *Panel Data Econometrics - Future Directions: Papers in Honour of Prof. Balestra*, Contributions to Economic Analysis, Elsevier Science.

Pesaran M.H. T Schuermann & S.M. Weiner, (2004) Modelling Regional Interdepencies using a Global Error Correcting Macroeconometric Model, *Journal of Business and Economic Statistics*, 22 (2) 129-162.

Pesaran, M.H. Schuerman & L.V. Smith (2009) Forecasting financial variables with Global VARs *International Journal of Forecasting*, 25(4) 642-675.

Pesaran, M.H. Y. Shin & R.J. Smith (2001) Bounds Testing approaches to the Analysis of Level Relationships, *Journal of Applied Econometrics*, 16, 289-326.

Pesaran, M.H., Y. Shin & R.P. Smith (1999) Pooled Mean Group Estimation of Dynamic Heterogenous Panels, *Journal of the American Statistical Association*, 94, 621-634.

Pesaran, M.H. & R.P. Smith (1995) Estimating Long-run relationships from Dynamic Heterogenous Panels, *Journal of Econometrics, 68, 79-113.*

Pesaran, M.H. & R.P. Smith (1998) Structural Analysis of Cointegrating VARs, *Journal of Economic Surveys*, 12, 471-506, reprinted in Oxley and McAleer (1999).

Pesaran, M.H. & R.P. Smith (2006) Macroeconometric Modelling with a global perspective, *Manchester School*, Supplement, 24-49.

Pesaran M.H. & Smith R.P. (2014) Tests of policy ineffectiveness in macroeconometrics CAFÉ research paper 14.07 and Birkbeck Working paper in Economics and Finance 1405, June 2014.

Pesaran M.H. & Smith R.P. (2016).Counterfactual analysis in macroeconometrics: an empirical investigation into the effects of quantitative easing, *Research in Economics*, 70 (2) June 2016 p262-280.

Pesaran, M.H. L.V. Smith & T. Yamagata (2013) Panel Unit Root Tests in the Presence of a Multifactor Error Structure, *Journal of Econometrics*, vol 175, no 2, pp. 94-115.

Pesaran, M.H. R.P. Smith & K.S. Im (1996) Dynamic Linear Models for Heterogenous Panels p145-195 of Matyas and Sevestre (1996).

Pesaran, M.H. R.P. Smith, T. Yamagata & L. Hvozdyk (2009) Pairwise Tests of Purchasing Power Parity , *Econometric Reviews,* 28(6) 1-27.

Pesaran, M.H. & E. Tosetti (2011) Large Panels with Common Factors and Spatial Correlations, *Journal of Econometrics*, 161(2), 182-202.

Pesaran M.H., A. Ullah & T. Yamagata (2006) A bias adjusted LM test of error cross-section independence, forthcoming *Econometrics Journal*.

Pesaran M.H. & T. Yamagata (2008) Testing slope homogeneity in large panels, *Journal of Econometrics, 142, 50-93.*.

Pesaran, M.H. and Qiankun Zhou (2014) Estimation of Time-invariant Effects in Static Panel Data Models,

Pesaran, M.H. and Qiankun Zhou (2015) To pool or not to pool, Phillips, P.C.B. (1986) Understanding Spurious Regressions in Economics, *Journal of Econometrics,* 33, 311-340.

Phillips, P.C.B. (2004) Challenges of trending time-series econometrics, *Cowles Foundation Discussion Paper* 1472.

Phillips, P.C.B. & Z. Xiao (1998) A Primer on Unit Root Testing, *Journal of Economic Surveys*, 12, 423-470, reprinted in Oxley & McAleer (1999).

Phillips, P.C.B., & H.R. Moon (1999) Linear Regression Limit Theory for Nonstationary Panel Data, *Econometrica,* 67,5, 1057-1112.

Phillips, P.C.B., & H.R. Moon (2000) Nonstationary Panel Data Analysis: an overview of some recent developments, *Econometric Reviews* 19, 3, 263-286.

Phillips, P.C.B. & D. Sul (2003) Dynamic Panel Testing and Homogeneity Testing under Cross section dependence, *The Econometrics Journal*, 6, 217-259..

Phillips, P.C.B. & D. Sul (2003) The elusive empirical shadow of growth convergence, Cowles Foundation Discussion Paper 1398.

Phillips, P.C.B. & D. Sul (2007) Bias in Dynamic Panel Estimation with Fixed Effects incidental trends and cross-section dependence, *Journal of Econometrics*, 137, p162-188.

Quah, D (1994) Exploiting Cross-section Variation for Unit Root Inference in Dynamic Data, *Economics Letters*, 44, 9-19.

Robertson, D. & J. Symons (1992) Some Strange Properties of Panel Data Estimators, *Journal of Applied Econometrics*, 7, 175-89.

Robertson, D. & J. Symons (1999) Factor Residuals in Panel Regressions: a suggestion for estimating panels allowing for cross-sectional dependence, mimeo University of Cambridge

Robertson, D. & J. Symons (2007) Maximum likelihood factor analysis with rank deficient sample covariance matrices, *Journal of Multivariate Analysis*, 98, 813-828.

Roodman, D (2009) A note on the theme of too many instruments, *Oxford Bulletin of Economics and Statistics*, 71(1) 135-158.

Sarafidis, V. T. Yamagata & D. Robertson (2009) A Test for Cross Section Dependence for a linear dynamic panel data model with regressors, *Journal of Econometrics*, 148, 149-161.

Sarno L & M P Taylor (2002) *The Economics of Exchange Rates*, Cambridge University Press.

Shin, Y. & A. Snell (2006) Mean group tests for stationarity in heterogeneous panels, *Econometrics Journal,* 9, 123-158.

Smith L.V & A Galesi (2011) GVAR toolbox. https://sites.google.com/site/gvarmodelling/gvar-toolbox

Smith R.P & G. Zoega, (2008) Global Factors, Unemployment Adjustment and the Natural Rate *Economics – The Open Access, Open Assessment E-Journal*, Vol 2 2008-22 July.

Stock, J.H. & M. W. Watson (2005) Implications of Dynamic Factor Models for VAR Analysis, *NBER Working Paper* 11467, SW.

Stock, J.H. & M. W. Watson (2006) Heteroskedasticity Robust Standard Errors for Fixed Effects Panel Data Regression, NBER Working Paper T0323.

Stock J, Watson M. (2010) Dynamic Factor Models. In: Clements MP, Henry DF *Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press; 2010.

Stock J, Watson M. (2016) Factor Models and Structural Vector Autoregressions in Macroeconomics, *Handbook of Macroeconomics*, ed J. B. Taylor.

Stone, J.R.N. (1947) On the interdependence of blocks of transactions, *Supplement to the Journal of the Royal Statistical Society*, 11, 1-31.

Su, L., Z. Shi and PCB Phillips (2014) Identifying Latent Structures in panel data, Cowles Foundation Discussion Paper 1965.

Swamy, P.A.V.B. (1970) Efficient Inference in a Random Coefficient Regression Model, *Econometrica* 38, 311-323.

Taylor, A.M. (2001) Potential Pitfalls for the Purchasing Power Parity Puzzle? Sampling and Specification Biases in Mean-Reversion Tests of the Law of One Price, *Econometrica*, 69, 473-498.

Toda, H. Y and T. Yamamoto (1995). Statistical inferences in vector autoregressions with possibly integrated processes. J*ournal of Econometrics*, 66, 225-250.

Trapani, L. & G. Urga (2010) Micro versus macro cointegration in heterogeneous panels, *Journal of Econometrics*, 155(1), 1-18.

Vega, S.H. & J.P. Elhorst (2013) On spatial econometric models, spillover effects and W, presented at the 53rd ERSA converence.

Wagner, M. & J. Hlouskova (2010) The performance of panel cointegration methods: results from a large scale simulation study. *Econometric Reviews*, 29(2), 182-223.

Wasserstein R.L. & N.A. Lazar (2016) The ASA's statement on p values: context, process and purpose, *The American Statistician,*

Westerlund, J. (2006) Testing for panel cointegration with a level break. *Economics Letters*, 91, 27-33.

Westerlund, J. (2007) Testing for error correction in panel data, *Oxford Bulletin of Economics and Statistics*, 69(6) p709-748.

Westerlund, J. (2005) New Simple Tests for Panel Cointegration, *Econometric Reviews*, 24, p297-316.

Westerlund, J. (2013) Rethinking the univariate approach to panel unit root testing: using covariates to resolve the incidental trend problem.

Westerlund J (2014) The power of PANIC, forthcoming *Journal of Econometrics*

Westerlund, J. and S.A. Basher (2008) Mixed Signals among tests for panel cointegration, *Economic Modelling*, 25, p128-136.

Westerlund J and J Breitung (2013) Lessons from a decade of IPS and LLC, *Econometric Reviews*, 32(5-6), 547-591.

Westerlund J & W Hess (2011) A new poolability test for cointegrated panels, *Journal of Applied Econometrics*, 26: 56-88.

Westerlund J. & R. Larsson (2009) A note on the pooling of individual PANIC unit root tests, *Econometric Theory*, 25, 1851-1868.

Westerlund J and M Norkute (2014) A factor analytic method to interactive effects dynamic panel models with and without a unit root.

Westerlund J. & J-P Urbain (2014) Cross-sectional averages versus principal components,

Wooldridge, J.M. (2010) *Econometric Analysis of Cross-section and Panel Data*, 2nd edition MIT press.

Zellner, A. (1962) An Efficient Method of Estimating Seemingly Unrelated Regressions and Test for Aggregation Bias, *Journal of the American Statistical Association*, 57, 348-368.

Zellner, A (1969) On the aggregation problem; A new approach to a troublesome problem, in K.A. Fox et al. eds *Economic models, estimation and risk programming; Essays in honor of Gerhard Tintner* (Springer-Verlag, Berlin) 365-378.