Dynamic Panel Data Workshop

Yongcheol Shin, University of York
University of Melbourne

10-12 June 2014

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The panel data are repeated time series observations on the same set of cross-section units. Thus, "pooling" of cross-section and time series data, where there are N cross-section individuals, $i = 1, 2, \ldots, N$, and T time periods, $t = 1, 2, \ldots, T$. Regression model is written as

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + \varepsilon_{it}, \tag{1.1}$$

$i = 1, 2, \ldots, N, t = 1, 2, \ldots, T$, where y_{it} is the value of the dependent variable for cross-section unit i at time t, x_{itj} is the value of the jth explanatory variable for unit i at time t for $j = 1, ..., k$. Let $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itk})$ be a $1 \times k$ vector of regressors (including constant), and $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$ a $k \times 1$ vectors of parameters. Then, (1.1) can be compactly written as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it},\ i = 1, 2, \ldots, N, t = 1, 2, \ldots, T. \tag{1.2}$$

More compactly,

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, i = 1, 2, \ldots, N, \tag{1.3}$$

where

$$\underset{T\times 1}{\mathbf{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \ \underset{T\times k}{\mathbf{x}_i} = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix}, \ \underset{T\times 1}{\boldsymbol{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix},$$

and finally

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.4}$$

where

$$\underset{NT\times 1}{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}, \ \underset{NT\times k}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}, \ \underset{NT\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}.$$

Motivation for use of panel data: The analysis of panel data is the subject of one of active literature in econometrics. See Hsiao (2003) and Baltagi (2008). *First*, we can obtain efficiency gain from using more observations. e.g. Budget study, where y is consumption of some good, $\mathbf{x}$ (prices, income), prices vary over time and (real) income varies over individual. *Second*, we can control the bias of the estimation. e.g. Labor economics, in earnings equation, where y is wage, $\mathbf{x}$ education, age, etc. time invariant unobservables or individual effects, which can be related to individual ability or intelligence.

Sources and types of the panel data

- The Panel Study of Income Dynamics (PSID)collected by the Institute of Social research at the University of Michigan (since 1968). Information about economic status such as income, job, marital status and so on
- The Survey of Income and Program Participation (SIPP, US Department of Commerce) covers shorter time periods.
- The study by Card (1992): Effects of the minimum wage law on employment: Collected information by US States on youth employment, unemployment rates, average wages and other factors for 1976-90.
- Macropanel such as the international data set obtained from the version 5.5 of the Penn World Tables collected by Summers and Heston.

1.1 Models For Pooled Time Series

Here N is relatively small, and T is large enough to run separate regressions for each individual but combining individuals may yield better (more efficient) estimates. Define the $NT \times NT$ covariance matrix,

$$\begin{aligned}\mathbf{V} &= Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= E\begin{bmatrix} \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2' & \dots & \boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_N' \\ \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_2' & \dots & \boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_N' \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\varepsilon}_N\boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_N\boldsymbol{\varepsilon}_2' & \dots & \boldsymbol{\varepsilon}_N\boldsymbol{\varepsilon}_N' \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{11} & \mathbf{v}_{12} & \dots & \mathbf{v}_{1N} \\ \mathbf{v}_{21} & \mathbf{v}_{22} & \dots & \mathbf{v}_{2N} \\ \vdots & \vdots & & \vdots \\ \mathbf{v}_{N1} & \mathbf{v}_{N2} & \dots & \mathbf{v}_{NN} \end{bmatrix},\end{aligned}$$

where the dimension of each block is $T \times T$. Assume that

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V}),$$

and $\mathbf{X}$'s are exogenous. Then, we consider the two basic estimators:

Ordinary least squares (OLS):

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which is unbiased but inefficient; that is,

$$Cov(\hat{\boldsymbol{\beta}}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Generalised least squares (GLS):

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

which is unbiased and efficient; that is,

$$Cov(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

Exercise 1.1.1 *Show that* $Cov(\hat{\boldsymbol{\beta}}_{OLS}) \geq Cov(\hat{\boldsymbol{\beta}}_{GLS})$.

We have different models depending on specifications of $\mathbf{V}$.

1.1.1 Classical regression model

We have ideal conditions such as ε_{it}'s are $iidN(0, \sigma^2)$. Then,

$$\mathbf{V} = \sigma^2\mathbf{I}_{NT},\ GLS = OLS,$$

where $\mathbf{I}_{NT}$ is an $NT \times NT$ identity matrix.

1.1.2 Cross-sectional heteroskedasticity: heterogenous variances

We have ideal conditions except

$$Var(\varepsilon_{it}) = \sigma_i^2 \neq Var(\varepsilon_{jt}) = \sigma_j^2,\ for\ i \neq j,$$

that is, the error variance is allowed to vary across individuals. Then,

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2\mathbf{I}_T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I}_T & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \sigma_N^2\mathbf{I}_T \end{bmatrix}$$

Therefore, the GLS estimator becomes the weighted least squares estimator given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \left(\sum_{i=1}^{N} \sigma_i^{-2}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \sigma_i^{-2}\mathbf{X}_i'\mathbf{y}_i\right). \end{aligned}$$

Noting that σ_i^2's are not observable, feasible GLS estimator is obtained by

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\sum_{i=1}^{N} s_i^{-2} \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} s_i^{-2} \mathbf{X}_i' \mathbf{y}_i \right),$$

where

$$s_i^2 = \frac{1}{T-k} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{GLS} \right)' \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{GLS} \right).$$

The FGLS is consistent, and asymptotically efficient (as $T \to \infty$ and N fixed).

Example 1 *Greene (1997).*

$$I_{it} = \beta_1 + \beta_2 F_{it} + \beta_3 C_{it} + \varepsilon_{it},$$

where $N = 5$, $T = 20$ (1935-1954), I_{it} is gross investment, F_{it} market value of the firm and C_{it} value of plant and equipment.

$$OLS : I = -48.03 + \underset{(.011)}{.106} F + \underset{(.043)}{.305} C, \; R^2 = .78,$$

$$FGLS : I = -36.25 + \underset{(.0074)}{.095} F + \underset{(.030)}{.338} C.$$

1.1.3 Cross-sectional Heteroskedasticity: Correlation across groups

Now, there is correlation across individuals at the same time;

$$Cov(\varepsilon_{it}, \varepsilon_{js}) = \left\{ \begin{array}{c} 0 \text{ for } t \neq s \\ \sigma_{ij} \text{ for } t = s \end{array} \right\}.$$

Then,

$$\mathbf{V} = \begin{bmatrix} \sigma_{11}\mathbf{I}_T & \sigma_{12}\mathbf{I}_T & \dots & \sigma_{1N}\mathbf{I}_T \\ \sigma_{21}\mathbf{I}_T & \sigma_{22}\mathbf{I}_T & \dots & \sigma_{2N}\mathbf{I}_T \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1}\mathbf{I}_T & \sigma_{N2}\mathbf{I}_T & \dots & \sigma_{NN}\mathbf{I}_T \end{bmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I}_T,$$

where the dimension of $\boldsymbol{\Sigma}$ is $N \times N$. This is the same error structure as in "seemingly unrelated regressions model".

FGLS can be obtained as previously: replace unknown σ_{ij} in $\mathbf{V}$ by

$$s_{ij} = \frac{1}{T-k} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{GLS} \right)' \left(\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}_{GLS} \right),$$

so define $\hat{\mathbf{V}} = \mathbf{V}$ accordingly with s_{ij} in place of σ_{ij}. Then,

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X} \right)^{-1} \left(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{y} \right).$$

Again, the GLS estimator is consistent and asymptotically normal as $T \to \infty$ and N fixed.

Example 2 *Greene example continued:*

$$FGLS: I = -30.28 + \underset{(.0068)}{.094}\, F + \underset{(.027)}{.341}\, C.$$

For testing that the off-diagonal elements of $\mathbf{\Sigma}$ are zero; that is, there is no correlation across groups, we use the following LM statistic developed by Breusch and Pagan (1980):

$$LM = T \sum_{i=2}^{N} \sum_{j=1}^{i-1} r_{ij}^2,$$

where r_{ij}^2 is the ijth residual correlation coefficient. Under the (joint) null of $\sigma_{ij} = 0$ for $i, j = 1, 2, ..., N$, and $i \neq j$, as $T \to \infty$,

$$LM \to \chi^2_{\frac{N(N-1)}{2}}.$$

Example 3 *Greene example continued: Using the residuals based on the FGLS estimates given above we find* $LM = 51.32$*, which is far greater than the 95% critical value of* χ^2_{10}*. Hence we may conclude that the simple heteroskedastic model is not general enough for the investment data.*

1.1.4 Autocorrelation (but not correlation across individuals)

See Greene (1997, Section 15.2.3).

Chapter 2

Models for Longitudinal Data

Here we have large N, but small T: hence we use an asymptotic theory as $N \to \infty$ and T fixed.

Example 4 *Panel study for income dynamics (PSID),* $N = 5000, T = 9$.

In principle methods of the previous section could be applied, but problematic because only a few time periods are available. In this case the techniques has been focused on cross sectional variation or heterogeneity. The basic assumption is that time-invariant "individual effect," becomes part of error process: that is, we consider the following error components-based panel,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it},\ i = 1, 2, \ldots, N; t = 1, 2, \ldots, T, \tag{2.1}$$

and $\boldsymbol{\varepsilon}_{it}$ is decomposed as

$$\varepsilon_{it} = \alpha_i + u_{it}, \tag{2.2}$$

where α_i's are called individual effects. Here we assume:

- u_{it}'s are $N(0, \sigma_u^2)$.
- u_{it}' are uncorrelated with $\mathbf{x}_{js}$ for all i, t, j, s, *i.e.*, $\mathbf{x}$'s are exogenous with respect to u.

But, assumptions about α_i vary.

Example 5 *We have observations for* T *time periods on* N *countries. We want to estimate the spillover effect of foreign technology on domestic firm productivity in manufacturing. An error components model describing output in each region with a standard Cobb-Douglas production is given by*

$$q_{it} = \beta_1 k_{it} + \beta_2 l_{it} + \beta_3 m_{it} + \beta_4 f_{it} + \alpha_i + u_{it},$$

where q_{it} is the log of output of domestic firms, k_{it} the log of capital, l_{it} the log of labor, m_{it} the log of material, f_{it} a measure of the influence of foreign firms. A positive spillover is indicated by $\beta_4 > 0$. Why should we allow for the unobserved individual effects α_i? One reason is that an observed positive relationship between output in a region and foreign influence, controlling only for capital, labor and materials, might simply reflect the fact that foreign firms tend to settle in areas that lend themselves to higher productivity; there may be no spillover effect at all. By adding α_i we allow f_{it} to be correlated possibly with features of region, embodied in α_i, that are related to higher productivity. This solution is an improvement over not allowing for α_i.

Example 6 *Let y be net migration into city i at time t. We would like to see whether taxes, housing prices, educational quality and other factors influence population flows. There are certain features of cities, for example geographical characteristics, reputation that could be difficult to model, but are essentially constant over short periods of time. Because the unobservables influence y_{it} and might also be related to local policy and economic variables, it is important to control for them. One model would be*

$$y_{it} = \beta_1 x_{it} + \beta_2 h_{it} + \beta_3 e_{it} + \beta_4 c_{it} + \alpha_i + u_{it},$$

where x are tax rates, h housing prices, e educational quality and c crime rates. Now α_i capture all time-constant (unobservable) differences about cities that might affect migration. Thus, the above regression allows us to estimate

$$E\left(y_{it}|x_{it}, h_{it}, e_{it}, c_{it}, \alpha_i\right),$$

which makes it clear that we are controlling for unobserved city effects when estimating the effects of tax policy on net migration for example.

There are two main approaches to deal with α_i's: fixed effects and random effects.

2.1 Fixed Effects Estimator

Here we treat α_i as fixed, but remember that we do not assume α_i to be uncorrelated with $\mathbf{x}_{it}$. This implies that differences across cross section units can be captured in differences in the constant terms. Notice that the regression of y_{it} on $\mathbf{x}_{it}$ only is biased if α_i is correlated with $\mathbf{x}_{it}$. We want to avoid this bias by using the fixed effects estimation. Defining

$$\mathbf{e}_T = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{T\times 1}, \ \boldsymbol{\alpha}_* = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}_{N\times 1}, \ \mathbf{D} = \begin{bmatrix} \mathbf{e}_T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_T & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e}_T \end{bmatrix}_{NT\times N} = \mathbf{I}_N \otimes \mathbf{e}_T,$$

and

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_2 \\ \vdots \\ \alpha_N \\ \vdots \\ \alpha_N \end{bmatrix}_{NT\times 1} = \begin{bmatrix} \alpha_1 \mathbf{e}_T \\ \alpha_2 \mathbf{e}_T \\ \vdots \\ \alpha_N \mathbf{e}_T \end{bmatrix} = \mathbf{D}\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_* \otimes \mathbf{e}_T,$$

then we have (in T observations for each individual)

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \alpha_i \mathbf{e}_T + \mathbf{u}_i, \tag{2.3}$$

or (in NT observations)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha} + \mathbf{u} = \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha}_* + \mathbf{u}. \tag{2.4}$$

Notice that the bias in regression of y on $\mathbf{X}$ only is due to omission of $\mathbf{D}$ in (2.4). Solution is simply to regress $\mathbf{y}$ on $(\mathbf{X}, \mathbf{D})$. So it is sometimes called least squares dummy variable (LSDV) model; that is,

$$\hat{\boldsymbol{\beta}} = \text{coefficients of } \mathbf{X} \text{ in the regression of } \mathbf{y} \text{ on } (\mathbf{X}, \mathbf{D}).$$

But there is a computational problem for large N, since the dimension of $\mathbf{D}$ is very big (N columns). So we need an alternative formulation. Define the $NT \times NT$ matrices $\mathbf{P}$ and $\mathbf{Q}$ by

$$\mathbf{P} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}',$$

$$\mathbf{Q} = \mathbf{I}_{NT} - \mathbf{P} = \mathbf{I}_{NT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}',$$

then a standard result from least squares algebra says:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Q}\mathbf{y}).$$

Digression on algebra: Notice that

$$\mathbf{e}_T'\mathbf{e}_T = T;\ \mathbf{e}_T\mathbf{e}_T' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ & & \vdots & \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Defining

$$\mathbf{P}_* = \frac{1}{T}\mathbf{e}_T\mathbf{e}_T' = \begin{bmatrix} \frac{1}{T} & \frac{1}{T} & \cdots & \frac{1}{T} \\ \frac{1}{T} & \frac{1}{T} & \cdots & \frac{1}{T} \\ & & \vdots & \\ \frac{1}{T} & \frac{1}{T} & \cdots & \frac{1}{T} \end{bmatrix},$$

then this matrix creates "means" for any $T \times 1$ vector $\mathbf{c} = [c_1, .., c_T]'$;

$$\mathbf{P}_*\mathbf{c} = \begin{bmatrix} \bar{c} \\ \bar{c} \\ \vdots \\ \bar{c} \end{bmatrix} = \bar{c}\mathbf{e}_T,\ \bar{c} = \frac{1}{T}\sum_{t=1}^{T} c_t.$$

Next, define

$$\mathbf{Q}_* = \mathbf{I}_T - \frac{1}{T}\mathbf{e}_T\mathbf{e}_T' = \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \cdots & -\frac{1}{T} \\ & & \vdots & \\ -\frac{1}{T} & -\frac{1}{T} & \cdots & 1-\frac{1}{T} \end{bmatrix};$$

which makes "deviations from means": that is,

$$\mathbf{Q}_*\mathbf{c} = \begin{bmatrix} c_1 - \bar{c} \\ c_2 - \bar{c} \\ \vdots \\ c_T - \bar{c} \end{bmatrix}.$$

These matrices are relevant because

$$\mathbf{P} = \mathbf{I}_N \otimes \mathbf{P}_*,\ \mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_*,$$

which are idempotent,

$$\mathbf{PP} = \mathbf{P},\ \mathbf{QQ} = \mathbf{Q},$$

symmetric

$$\mathbf{P} = \mathbf{P}', \mathbf{Q} = \mathbf{Q}'$$

and orthogonal to each other

$$\mathbf{PQ} = \mathbf{0}.$$

They make "individual means" and "deviation from individual means";

$$\mathbf{Py} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_2 \\ \vdots \\ \bar{y}_N \\ \vdots \\ \bar{y}_N \end{bmatrix}, \ \mathbf{Qy} = \begin{bmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1T} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2T} - \bar{y}_2 \\ \vdots \\ y_{N1} - \bar{y}_N \\ \vdots \\ y_{NT} - \bar{y}_N \end{bmatrix}, \tag{2.5}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}$. In particular, note that

$$\mathbf{Qy} = \mathbf{Q}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha} + \mathbf{u}\right) = \mathbf{QX}\boldsymbol{\beta} + \mathbf{Qu},$$

because

$$\mathbf{Q}\boldsymbol{\alpha} = \mathbf{0}.$$

Therefore, taking deviations from (individual) means removes time-invariant unobservables. Multiplication by $\mathbf{Q}$ (taking deviations from individual means) is often called the "within transformations".

Within estimator This can be obtained by one of the following equivalent expressions:

1. Regression **Qy** on **Qx.**
2. Regression of $y_{it} - \bar{y}_i$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$, $i = 1, ..., N$, $t = 1, ..., T$, where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$.
3. OLS estimation with dummy variables (LSDV).

The within estimator is obtained by[1]

$$\hat{\boldsymbol{\beta}}_W = \left(\mathbf{X}'\mathbf{QX}\right)^{-1}\mathbf{X}'\mathbf{Qy}.$$

Here we should bear in mind that

[1] Using the (double) summation notation, we have

$$\hat{\boldsymbol{\beta}}_W = \left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)'\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)\right\}^{-1}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)'\left(y_{it} - \bar{y}_i\right)\right\}.$$

1. We cannot include any time-invariant regressors.

2. If you estimate by within-estimation and still want $\hat{\alpha}_i$, then

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_W.$$

3. Statistical properties of $\hat{\boldsymbol{\beta}}_W$: unbiased, consistent (as $N \to \infty$ or $T \to \infty$) and asymptotically normal,[2]

$$\sqrt{NT}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \to N\left\{0, \sigma_u^2 \lim_{NT\to\infty}\left(\frac{1}{NT}\mathbf{X}'\mathbf{Q}\mathbf{X}\right)^{-1}\right\}.$$

   The usual inference procedures can be used like t and Wald tests.

In sum, the fixed effects estimation removes potential bias caused by time-invariant unobservables by the within transformation. Cost is that only variation over time (not between individual) is used in estimating $\boldsymbol{\beta}$, which would possibly result in imprecise estimates. Essentially the fixed effects model concentrates on differences within individuals; it is explaining to what extent y_{it} differs from $\bar{y}_i$ and does not explain why $\bar{y}_i$ is different from $\bar{y}_j$. It may be important to realize that $\boldsymbol{\beta}$'s are identified (or consistently estimated) only through within variation of the data.

2.2 Random Effects Estimator

We consider the same basic model, (5.16), but now assume:

- u_{it}'s are $iidN(0, \sigma_u^2)$.
- α_i's are $iidN(0, \sigma_\alpha^2)$.
- α_i's are uncorrelated with u_{jt} for all i, j, t, that is, $E[\alpha_i u_{jt}] = 0$ for all i, j, t.
- α_i and u_{it} are uncorrelated with x_{js} for all i, j, s, t (so $\mathbf{x}$ is exogenous with respect to $\boldsymbol{\alpha}$ and $\mathbf{u}$).

[2] A consistent estimate for σ_u^2 is obtained as the within residual sum of squares divided by $N(T-1)$, that is,

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}_W (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right\}^2.$$

It is also possible to apply the usual degrees of freedom correction in which case k is subtracted from the denominator.

This approach would be appropriate if we believed that sampled cross-sectional units were drawn from a large population. *First*, the OLS estimator is unbiased or consistent, because $\mathbf{x}$ is assumed to be exogenous with respect to $\boldsymbol{\varepsilon} = \boldsymbol{\alpha} + \mathbf{u}$, but inefficient. This model is suitable for case of "pooling", not of "bias reduction" as in the fixed effects model. In general, the GLS estimator will be more efficient. For the GLS estimation we need a more detailed expression for $Cov(\boldsymbol{\varepsilon}) = \mathbf{V}$.

First, consider

$$Cov(\boldsymbol{\varepsilon}_i) = E \begin{bmatrix} \varepsilon_{i1}^2 & \varepsilon_{i1}\varepsilon_{i2} & \cdots & \varepsilon_{i1}\varepsilon_{iT} \\ \varepsilon_{i2}\varepsilon_{i1} & \varepsilon_{i2}^2 & \cdots & \varepsilon_{i2}\varepsilon_{iT} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{iT}\varepsilon_{i1} & \varepsilon_{iT}\varepsilon_{i2} & \cdots & \varepsilon_{iT}^2 \end{bmatrix}_{T\times T}.$$

Under the assumptions given above it is easily seen that

$$E\left(\varepsilon_{it}^2\right) = E\left(\alpha_i^2 + u_{it}^2 + 2\alpha_i u_{it}\right) = \sigma_\alpha^2 + \sigma_u^2,$$
$$E\left(\varepsilon_{it}\varepsilon_{is}\right) = E\left(\alpha_i + u_{it}\right)\left(\alpha_i + u_{is}\right) = \sigma_\alpha^2,\ t \neq s.$$

Hence, for all i,

$$Cov(\boldsymbol{\varepsilon}_i) = \sigma_\alpha^2 \mathbf{e}_T \mathbf{e}_T' + \sigma_u^2 \mathbf{I}_T = \begin{bmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_u^2 \end{bmatrix}.$$

Therefore, the $NT \times NT$ matrix $\mathbf{V}$ can be written as

$$\mathbf{V} = Cov(\boldsymbol{\varepsilon}) = \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Omega} \end{bmatrix} = \mathbf{I}_N \otimes \boldsymbol{\Omega}, \tag{2.6}$$

where $\otimes$ is a Kronecker product.

Now, recall

$$\mathbf{P}_* = \frac{1}{T}\mathbf{e}_T \mathbf{e}_T',\ \mathbf{Q}_* = \mathbf{I}_T - \frac{1}{T}\mathbf{e}_T \mathbf{e}_T',\ \mathbf{P} = \mathbf{I}_N \otimes \mathbf{P}_*,\ \mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_*.$$

Then, $\boldsymbol{\Omega}$ can be rewritten as

$$\begin{aligned} \boldsymbol{\Omega} &= \left(T\sigma_\alpha^2 + \sigma_u^2\right)\frac{1}{T}\mathbf{e}_T \mathbf{e}_T' + \sigma_u^2\left(\mathbf{I}_T - \frac{1}{T}\mathbf{e}_T \mathbf{e}_T'\right) \\ &= \left(T\sigma_\alpha^2 + \sigma_u^2\right)\mathbf{P}_* + \sigma_u^2 \mathbf{Q}_*. \end{aligned}$$

Next,

$$\mathbf{V} = \mathbf{I}_N \otimes \boldsymbol{\Omega} = \left(T\sigma_\alpha^2 + \sigma_u^2\right)\mathbf{P} + \sigma_u^2 \mathbf{Q}, \tag{2.7}$$

where we used

$$\mathbf{P} = \mathbf{I}_N \otimes \mathbf{P}_*,\ \mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_*.$$

Digression on derivation of the inverse of V For the GLS estimation we need to find

$$\mathbf{V}^{-1} = \mathbf{I}_N \otimes \mathbf{\Omega}^{-1} \text{ or } \mathbf{V}^{-1/2} = \mathbf{I}_N \otimes \mathbf{\Omega}^{-1/2}.$$

Using the special nature of $\mathbf{P}$ and $\mathbf{Q}$, it can be shown that[3]

$$\mathbf{V}^{-1} = \frac{1}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{P} + \frac{1}{\sigma_u^2}\mathbf{Q}, \tag{2.8}$$

and then

$$\begin{aligned}\mathbf{V}^{-1/2} &= \frac{1}{\sqrt{T\sigma_\alpha^2 + \sigma_u^2}}\mathbf{P} + \frac{1}{\sqrt{\sigma_u^2}}\mathbf{Q} = \frac{1}{\sigma_u}\left\{\sqrt{\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}}\mathbf{P} + \mathbf{Q}\right\} \\ &= \frac{1}{\sigma_u}\left\{\mathbf{I}_{NT} - \left(1 - \sqrt{\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}}\right)\mathbf{P}\right\}.\end{aligned}$$

Define

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}}, \tag{2.9a}$$

then

$$\mathbf{V}^{-1/2} = \frac{1}{\sigma_u}\left(\mathbf{I}_{NT} - \theta\mathbf{P}\right). \tag{2.10}$$

Notice that the GLS estimator is obtained by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}) \\ &= \left[(\mathbf{V}^{-1/2}\mathbf{X})'(\mathbf{V}^{-1/2}\mathbf{X})\right]^{-1}(\mathbf{V}^{-1/2}\mathbf{X})'(\mathbf{V}^{-1/2}\mathbf{y}),\end{aligned}$$

and therefore $\hat{\boldsymbol{\beta}}_{GLS}$ is obtained from the regression of $\mathbf{V}^{-1/2}\mathbf{y}$ on $\mathbf{V}^{-1/2}\mathbf{X}$. In fact, this is equivalent to the OLS estimation after "θ differences". Since

$$\mathbf{V}^{-1/2}\mathbf{y} = \frac{1}{\sigma_u}\left(\mathbf{I}_{NT} - \theta\mathbf{P}\right)\mathbf{y} = \frac{1}{\sigma_u}\left(\mathbf{y} - \theta\mathbf{P}\mathbf{y}\right),$$

or more precisely

$$\frac{1}{\sigma_u}\left(\mathbf{V}^{-1/2}\mathbf{y}\right)_{it} = \frac{1}{\sigma_u}\left(y_{it} - \theta\bar{y}_i\right),$$

and likewise

$$\frac{1}{\sigma_u}\left(\mathbf{V}^{-1/2}\mathbf{X}\right)_{it} = \frac{1}{\sigma_u}\left(\mathbf{x}_{it} - \theta\bar{\mathbf{x}}_i\right),$$

[3] Similarly, the inverse of $\mathbf{\Omega}$ can be obtained as

$$\mathbf{\Omega}^{-1} = \frac{1}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{P}_* + \frac{1}{\sigma_u^2}\mathbf{Q}_*.$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$. So the GLS estimator is obtained from a regression of

$$(y_{it} - \theta\bar{y}_i) = (\mathbf{x}_{it} - \theta\bar{\mathbf{x}}_i)\,\boldsymbol{\beta} + \tilde{\varepsilon}_{it},$$

where $\tilde{\varepsilon}_{it} = \varepsilon_{it} - \theta\bar{\varepsilon}_i$ (the proportionality constant $\frac{1}{\sigma_u}$ being cancelled out). In other words, a fixed proportion θ of the individual means is subtracted from the data to obtain this transformed model.

We note in passing that

1. We can include time-invariant or individual specific variables. They also get multiplied by $\mathbf{I}_{NT} - \theta\mathbf{P}$.

2. If $\theta = 0$, then the GLS estimator is equivalent to OLS. But, this would occur only if $\sigma_\alpha^2 = 0$.

3. The GLS estimator is consistent as $N \to \infty$ (with T fixed or with $T \to \infty$ such that $\frac{N}{T}$ constant). It is also asymptotically efficient relative to the within estimator with

$$Cov\left(\hat{\boldsymbol{\beta}}_{GLS}\right) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} = \left\{\mathbf{X}'\left(\frac{1}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{P} + \frac{1}{\sigma_u^2}\mathbf{Q}\right)\mathbf{X}\right\}^{-1}.$$

4. The efficiency difference tends to 0 as $T \to \infty$ and $\theta \to 1$. (When $\theta = 1$, $\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_W$.)

Between estimator

We formulate the model in terms of the individual means,

$$\bar{\mathbf{y}}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + \bar{\boldsymbol{\varepsilon}}_i,\ i = 1, ..., N, \tag{2.11}$$

where

$$\underset{T\times 1}{\bar{\mathbf{y}}_i} = \bar{y}_{i\cdot} \otimes \mathbf{e}_T,\ \underset{T\times k}{\bar{\mathbf{x}}_i} = \bar{\mathbf{x}}_{i\cdot} \otimes \mathbf{e}_T$$

$$\bar{\boldsymbol{\varepsilon}}_i = \alpha_i \otimes \mathbf{e}_T + \bar{\mathbf{u}}_i,$$

$$\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it},\ \bar{\mathbf{x}}_{i\cdot} = T^{-1}\sum_{t=1}^{T} \bar{\mathbf{x}}_{it}.$$

Reminding that the matrix $\mathbf{P}$ creates "means" for any conformable vector, then we write (2.11) in matrix form as

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}. \tag{2.12}$$

The OLS estimator in the above regression gives the between estimator,

$$\hat{\boldsymbol{\beta}}_B = \left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}\mathbf{y}, \tag{2.13}$$

which is also unbiased and consistent as $N \to \infty$ under the assumption that $\bar{\mathbf{x}}_i$ is uncorrelated with α_i (such that $E\left(\bar{\mathbf{x}}_i'\bar{\boldsymbol{\varepsilon}}_i\right) = 0$), and with

$$Var\left(\hat{\boldsymbol{\beta}}_B\right) = \left(T\sigma_\alpha^2 + \sigma_u^2\right)\left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1}.$$

The between estimator ignores any information within individuals. The formular in (2.13) will be simplified as: Notice from (2.5) that $\mathbf{Py}$ and $\mathbf{PX}$ can be written as

$$\mathbf{Py} = \bar{\mathbf{y}} \otimes \mathbf{e}_T,\ \mathbf{PX} = \bar{\mathbf{X}} \otimes \mathbf{e}_T,$$

where

$$\underset{N\times 1}{\bar{\mathbf{y}}} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_N \end{bmatrix};\ \underset{N\times k}{\bar{\mathbf{X}}} = \begin{bmatrix} \bar{\mathbf{x}}_{1\cdot} \\ \vdots \\ \bar{\mathbf{x}}_{N\cdot} \end{bmatrix},$$

and thus we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}_B &= \left\{(\bar{\mathbf{x}} \otimes \mathbf{e}_T)'(\bar{\mathbf{x}} \otimes \mathbf{e}_T)\right\}^{-1}(\bar{\mathbf{x}} \otimes \mathbf{e}_T)'(\bar{\mathbf{y}} \otimes \mathbf{e}_T) \qquad (2.14) \\ &= \left(\bar{\mathbf{x}}'\bar{\mathbf{x}} \otimes \mathbf{e}_T'\mathbf{e}_T\right)^{-1}\left(\bar{\mathbf{x}}'\bar{\mathbf{y}} \otimes \mathbf{e}_T'\mathbf{e}_T\right) = \left(\bar{\mathbf{x}}'\bar{\mathbf{x}} \otimes T\right)^{-1}\left(\bar{\mathbf{x}}'\bar{\mathbf{y}} \otimes T\right) \\ &= \left\{\left(\bar{\mathbf{x}}'\bar{\mathbf{x}}\right)^{-1} \otimes T^{-1}\right\}\left(\bar{\mathbf{x}}'\bar{\mathbf{y}} \otimes T\right) = \left(\bar{\mathbf{x}}'\bar{\mathbf{x}}\right)^{-1}\bar{\mathbf{x}}'\bar{\mathbf{y}}, \end{aligned}$$

which is equivalent to the OLS estimator obtained from the following cross-sectional regression:

$$\bar{y}_{i\cdot} = \bar{\mathbf{x}}_{i\cdot}\boldsymbol{\beta} + \bar{\varepsilon}_{i\cdot},\ i = 1, ..., N. \qquad (2.15)$$

The GLS estimator can be shown to be a weighted average of the within estimator $\hat{\boldsymbol{\beta}}_W$ and the between estimator $\hat{\boldsymbol{\beta}}_B$,

$$\hat{\boldsymbol{\beta}}_G = \mathbf{F}\hat{\boldsymbol{\beta}}_W + \left(\mathbf{I}_k - \mathbf{F}\right)\hat{\boldsymbol{\beta}}_B,$$

where

$$\mathbf{F} = \left(\mathbf{X}'\mathbf{Q}\mathbf{X} + \lambda\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Q}\mathbf{X}, \quad \lambda = (1-\theta)^2.$$

This clearly shows that the efficiency gain of the GLS relative to the within estimator comes from the use of between (across individuals) variations. The GLS estimator is the optimal combination of the within and the between estimator. Therefore, it is more efficient than either.

There are some polar cases to consider:

1. If $\lambda = 1$, the GLS is equivalent to the OLS. This would occur only if $\sigma_\alpha^2 = 0$. Thus, the OLS estimator is also a linear combination of the within and the between estimators, but inefficient one.

2. If $\lambda = 0$, the GLS is equivalent to the within estimator. There are two possibilities. The first is $\sigma_u^2 = 0$, in which case all variation across individuals would be due to α_i's, which would be equivalent to the dummy variables used in the fixed effects model. The question of whether they were fixed or random would be moot. The other case is $T \to \infty$.

Feasible GLS estimator

We need to find a consistent (as $N \to \infty$) estimator of

$$\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}},$$

and then $\widehat{\mathbf{V}}$ (see (2.7)). The estimator of σ_u^2 is easily obtained from the within residuals (see (5.4) or (2.4)), denoted $\hat{\sigma}_u^2$ and estimated by

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)-k} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2, \tag{2.16}$$

with the within residuals given by

$$\hat{u}_{it} = (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\,\hat{\boldsymbol{\beta}}_W.$$

From the between regression (2.15) we find

$$\sigma_B^2 = E\left(\bar{\varepsilon}_{i\cdot}^2\right) = E\left(\alpha_i^2 + \bar{u}_{i\cdot}^2 + 2\alpha_i \bar{u}_{i\cdot}\right) = \sigma_\alpha^2 + \frac{1}{T}\sigma_u^2.$$

Hence, the consistent estimator for $\hat{\sigma}_\alpha^2$ is obtained by

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_B^2 - \frac{1}{T}\hat{\sigma}_u^2,$$

where $\hat{\sigma}_B^2$ is the consistent estimator of $\sigma_B^2 = \sigma_\alpha^2 + \frac{1}{T}\sigma_u^2$ obtained by

$$\hat{\sigma}_B^2 = \frac{1}{N-k} \sum_{i=1}^{N} \widehat{\bar{\varepsilon}}_{i\cdot}^2, \tag{2.17}$$

with the between residuals given by[4]

$$\widehat{\bar{\varepsilon}}_{i\cdot} = \bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}}_B.$$

Now, we have

$$\hat{\theta} = 1 - \sqrt{\frac{\hat{\sigma}_{u,W}^2}{\hat{\sigma}_{u,W}^2 + T\hat{\sigma}_\alpha^2}},$$

and the resulting feasible GLS estimates is the random effects estimator,

$$\hat{\boldsymbol{\beta}}_{RE} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}), \tag{2.18}$$

where

$$\hat{\mathbf{V}}^{-1} = \frac{1}{\hat{\sigma}_{u,W}^2 + T\hat{\sigma}_\alpha^2}\mathbf{P} + \frac{1}{\hat{\sigma}_{u,W}^2}\mathbf{Q}.$$

[4] It is also possible to apply degrees of freedom correction in computing $\hat{\sigma}_{u,W}^2$ and $\hat{\sigma}_B^2$.

One thing to note is that the implied estimate of σ_α^2 may be negative in finite samples. Such a negative finding calls the specification of the model into question. See Green, section 16.4.3b.

None of the desirable properties of the random effects estimator relies on $T \to \infty$, although it can be shown that some consistency results follow for T increasing. On the other hand, in this case the fixed effects estimator does rely on T increasing for consistency. See Nickell (1981).

2.3 Fixed Effects or Random Effects

Whether to treat individual effects α_i as fixed or random is not an easy question to answer. The most common view is that the discussion should not be about the true nature of α_i. The appropriate interpretation is that the fixed effects approach is conditional on the values of α_i. This makes sense if the individuals in the sample are 'one of a kind' and cannot be viewed as a random draw from some underlying population. This is probably most appropriate when i denotes, (large) companies or industries.

In contrast the random effects approach is not conditional on the individual α_i's but integrates them out. In this case we are usually not interested in the particular value of some individual's α_i; we just focus on arbitrary individuals that have certain characteristics. The random effects approach allows one to make inference with respect to population characteristics.

One way of formalizing this is noting that the random effects model states that

$$E\left(y_{it}|\mathbf{x}_{it}\right) = \mathbf{x}_{it}\boldsymbol{\beta},$$

while the fixed effects model estimates

$$E\left(y_{it}|\mathbf{x}_{it}\right) = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i.$$

$\boldsymbol{\beta}$'s in these two conditional expectations are the same only if $E\left(\alpha_i|\mathbf{x}_{it}\right) = 0$.

However, even if we are interested in the larger population of individuals, and a random effects framework seems appropriate, the fixed effects estimator may be preferred, since it is likely the case that $\mathbf{x}_{it}$ and α_i are correlated in which the random effects approach, ignoring this correlation, leads to inconsistent estimators. This problem of correlation can be handled only by using the fixed effects approach.

2.3.1 Hausman Test

The general specification test suggested by Hausman (1978) can be used to test the null hypothesis

$$H_0 : \mathbf{x}_{it} \text{ and } \alpha_i \text{ are uncorrelated},$$

against the alternative hypothesis

$$H_1 : \mathbf{x}_{it} \text{ and } \alpha_i \text{ are correlated.}$$

This test is based on an idea that the fixed effects estimator is consistent under both the null and the alternative while the random effects estimator is consistent only under the null but efficient. Let us consider the difference between $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_{RE}$. To evaluate the significance of this difference we need to find its covariance matrix. Under the null the two estimates should not differ significantly, and it can be also shown under the null that

$$\begin{aligned} Var\left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{RE}\right) &= Var\left(\hat{\boldsymbol{\beta}}_W\right) + Var\left(\hat{\boldsymbol{\beta}}_{RE}\right) - Cov\left(\hat{\boldsymbol{\beta}}_W, \hat{\boldsymbol{\beta}}_{RE}\right) - Cov\left(\hat{\boldsymbol{\beta}}_W, \hat{\boldsymbol{\beta}}_{RE}\right) \\ &= Var\left(\hat{\boldsymbol{\beta}}_W\right) - Var\left(\hat{\boldsymbol{\beta}}_{RE}\right), \end{aligned} \tag{2.19}$$

where we used Hausman's essential result that

$$Cov\left(\left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{RE}\right), \hat{\boldsymbol{\beta}}_{RE}\right) = Cov\left(\hat{\boldsymbol{\beta}}_W, \hat{\boldsymbol{\beta}}_{RE}\right) - Var\left(\hat{\boldsymbol{\beta}}_{RE}\right) = 0,$$

or

$$Cov\left(\hat{\boldsymbol{\beta}}_W, \hat{\boldsymbol{\beta}}_{RE}\right) = Var\left(\hat{\boldsymbol{\beta}}_{RE}\right).$$

Consequently, the Hausman test statistic is defined as

$$h = \left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{RE}\right)' \left\{\widehat{Var}\left(\hat{\boldsymbol{\beta}}_W\right) - \widehat{Var}\left(\hat{\boldsymbol{\beta}}_{RE}\right)\right\}^{-1} \left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{RE}\right), \tag{2.20}$$

where $\widehat{Var}\left(\hat{\boldsymbol{\beta}}_W\right)$ and $\widehat{Var}\left(\hat{\boldsymbol{\beta}}_{RE}\right)$ denote the estimates of $Var\left(\hat{\boldsymbol{\beta}}_W\right)$ and $Var\left(\hat{\boldsymbol{\beta}}_{RE}\right)$. Under the null,

$$h \sim \chi^2(k),$$

where k is the number of parameters. The Hausman test thus tests whether the fixed effects and random effects estimators are significantly different. It is also possible to test for a subset of parameters in $\boldsymbol{\beta}$.

2.3.2 Random Effects Correlated with Regressors

Essential difference between the "fixed" and "random" effects is whether or not the individual effects are correlated with regressors. We now consider the case where the random effects are correlated with regressors.

Mundlak (1978) argued that the dichotomy between fixed effects and random effects models disappears if we make the assumption that α_i depend on the mean values of $\mathbf{x}_i$, an assumption he regards as reasonable in many problems. As before, consider the error components model,

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}_i + \mathbf{u}_i, \tag{2.21}$$

but now assume

$$\alpha_i = \bar{\mathbf{x}}_i \boldsymbol{\pi} + w_i,$$

where w_i has the same properties that α_i was assumed to have; that is,

1. w_i's are $iidN(0, \sigma_w^2)$.
2. w_i's are uncorrelated with u_{jt} for all i, j, t; $E[w_i u_{jt}] = 0$ for all i, j, t.
3. w_i's are uncorrelated with x_{jt} for all i, j, t; $E[w_i x_{jt}] = 0$ for all i, j, t.

Then, we rewrite (2.21) as

$$\mathbf{y}_i = \mathbf{x}_i \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\pi} + w_i + \mathbf{u}_i, \tag{2.22}$$

which can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbf{X}\boldsymbol{\pi} + \mathbf{w} + \mathbf{u} = \begin{bmatrix} \mathbf{X} & \mathbf{P}\mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\pi} \end{bmatrix} + \mathbf{w} + \mathbf{u},$$

where we used

$$\boldsymbol{\alpha} = (\mathbf{P}\mathbf{X})\,\boldsymbol{\pi} + \mathbf{w},$$

and $Cov(\mathbf{w} + \mathbf{u}) = \mathbf{V}$ (see (2.6)). Carrying out the GLS estimation, then

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{GLS} \\ \hat{\boldsymbol{\pi}}_{GLS} \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{X} & \mathbf{P}\mathbf{X} \end{bmatrix}' \mathbf{V}^{-1} \begin{bmatrix} \mathbf{X} & \mathbf{P}\mathbf{X} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{X} & \mathbf{P}\mathbf{X} \end{bmatrix}' \mathbf{V}^{-1} \mathbf{y}. \tag{2.23}$$

After some algebra, it can be shown that[5]

$$\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_W = \left(\mathbf{X}'\mathbf{Q}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Q}\mathbf{y}, \tag{2.24}$$

$$\hat{\boldsymbol{\pi}}_{GLS} = \hat{\boldsymbol{\beta}}_B - \hat{\boldsymbol{\beta}}_W = \left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{P}\mathbf{y} - \left(\mathbf{X}'\mathbf{Q}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Q}\mathbf{y}, \tag{2.25}$$

with

$$Var\left(\hat{\boldsymbol{\pi}}_{GLS}\right) = Var\left(\hat{\boldsymbol{\beta}}_B\right) + Var\left(\hat{\boldsymbol{\beta}}_W\right) = \left(T\sigma_w^2 + \sigma_u^2\right)\left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1} + \sigma_u^2 \left(\mathbf{X}'\mathbf{Q}\mathbf{X}\right)^{-1}.$$

This shows that for the linear regression model, the fixed effects is effectively the same as the random effects correlated with all regressors.

The test of $\boldsymbol{\pi} = 0$ can be based on the following statistic:

$$\hat{\boldsymbol{\pi}}'_{GLS} \left[Var\left(\hat{\boldsymbol{\pi}}_{GLS}\right)\right]^{-1} \hat{\boldsymbol{\pi}}_{GLS} \xrightarrow{d} \chi_k^2 \; under \; H_0.$$

[5] The same result can be derived alternatively. Applying the GLS transformation described earlier to (2.22), we have

$$\begin{aligned} y_{it} - \theta \bar{y}_i &= (\mathbf{x}_{it} - \theta \bar{\mathbf{x}}_i)\,\boldsymbol{\beta} + (\bar{\mathbf{x}}_i - \theta \bar{\mathbf{x}}_i)\,\boldsymbol{\pi} + v_{it} \\ &= (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\,\boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\delta} + v_{it}, \end{aligned}$$

where $\boldsymbol{\delta} = (1 - \theta)\,(\boldsymbol{\pi} + \boldsymbol{\beta})$. Using that $\bar{\mathbf{x}}_i$ is orthogonal to $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$, we get the result.

2.4 Alternative IV Estimators

The fixed effects estimator eliminates anything that is time-invariant from the model, which might be a high price to pay for allowing the x variables to be correlated with individual specific heterogeneity α_i. For example, we may be interested in the effect of time invariant variables like gender or schooling on a person's wage. In this section we show that there is no need to restrict attention to the fixed and the random effects only, as it is possible to derive instrumental variables estimators that can be considered an intermediate case between fixed and random effects approach.

We now show that the fixed effects estimator is a special case of an IV estimator. Notice that the fixed effects estimator can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_W &= \left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)\right\}^{-1}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'\left(y_{it}-\bar{y}_i\right)\right\} \\ &= \left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'\mathbf{x}_{it}\right\}^{-1}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'y_{it}\right\} = \hat{\boldsymbol{\beta}}_{IV},\end{aligned}$$

which shows that $\hat{\boldsymbol{\beta}}_W$ has an interpretation of the IV estimator,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + u_{it},\ i = 1, 2, \ldots, N, ; t = 1, 2, \ldots, T,$$

where $\mathbf{x}_{it}$ is instrumented by $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$. Notice that by construction

$$E\left[\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'\alpha_i\right] = 0,$$

so that an IV estimator is consistent provided that

$$E\left[\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)'u_{it}\right] = 0,$$

which is satisfied by our assumption of strict exogeneity of $\mathbf{x}_{it}$. This route may allow us to estimate the effect of time invariant variables in a general context.

To describe this approach, consider the following model:

$$y_{it} = \mathbf{x}_{1,it}\boldsymbol{\beta}_1+\mathbf{x}_{2,it}\boldsymbol{\beta}_2+\mathbf{z}_{1,i}\boldsymbol{\gamma}_1+\mathbf{z}_{2,i}\boldsymbol{\gamma}_2+\alpha_i+u_{it},\ i = 1, 2, \ldots, N, ; t = 1, 2, \ldots, T, \tag{2.26}$$

where we have four different groups of variables; $\mathbf{x}$'s are varying over both time periods and cross-section units, but $\mathbf{z}$'s are varying only over cross-section units and time-invariant.

In addition we assume that the $1 \times k_1$ vector $\mathbf{x}_{1,it}$ and the $1 \times g_1$ vector $\mathbf{z}_{1,i}$ are uncorrelated with α_i:

$$\text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_{1,i}\alpha_i = \mathbf{0},\ \text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{z}}_{1,i}\alpha_i = \mathbf{0},$$

whereas the $1 \times k_2$ vector $\mathbf{x}_{2,it}$ and the $1 \times g_2$ vector $\mathbf{z}_{2,i}$ are correlated with α_i ($k_1 + k_2 = k$ and $g_1 + g_2 = g$). Under these assumptions, the fixed effects estimation provides still consistent estimators for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, but would not identify $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, since time-invariant variables $\mathbf{z}_{1,i}$ and $\mathbf{z}_{2,i}$ are wiped out by the within transformation.

2.4.1 Hausman and Taylor (1981) IV Estimator

Hausman and Taylor (1981) suggest to estimate (2.26) by IV using the following variables as instruments:

$$\mathbf{x}_{1,it} \text{ for } \mathbf{x}_{1,it},\ \mathbf{x}_{2,it} - \bar{\mathbf{x}}_{2,i} \text{ for } \mathbf{x}_{2,it},\ \mathbf{z}_{1,i} \text{ for } \mathbf{z}_{1,i},\ \text{and } \bar{\mathbf{x}}_{1,i} \text{ for } \mathbf{z}_{2,i},$$

that is, uncorrelated variables $\mathbf{x}_{1,it}$ and $\mathbf{z}_{1,i}$ trivially serve as their own instruments, but $\mathbf{x}_{2,it}$'s are instrumented by their deviation from individual means as in the fixed effects estimation, and finally, $\mathbf{z}_{2,i}$ is instrumented by the individual average of $\mathbf{x}_{1,it}$. Obviously the identification requires that the number of $\mathbf{x}_{1,it}$ is as large as that of $\mathbf{z}_{1,i}$, ($k_1 > g_1$).

The resulting estimator, Hausman-Talyor estimator, also allows us to estimate $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ consistently. If some of time-invariant variables are believed to be correlated with α_i, we require that sufficient time-varying variables that are not correlated with α_i should be included for instruments. In particular, the advantage of the Hausman and Taylor is that one does not have to use external instruments, instruments can be obtained within the model.

There are two versions of the Hausman-Taylor estimator, called HT-IV and HT-GLS, respectively.

HT-IV estimator (consistent but less efficient)

We rewrite (2.26) in the matrix from

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\alpha} + \mathbf{u} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\alpha} + \mathbf{u}, \end{aligned} \tag{2.27}$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')$. We first estimate by $\hat{\boldsymbol{\beta}}_W$ (within estimator), which is still consistent, and consider the following averaged within residuals in matrix form:

$$\mathbf{P}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{Z}\boldsymbol{\gamma} + \left\{\boldsymbol{\alpha} + \mathbf{P}\mathbf{u} + \mathbf{P}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)\right\}, \tag{2.28}$$

where $\mathbf{P}\mathbf{Z} = \mathbf{Z}$ and $\mathbf{P}\boldsymbol{\alpha} = \boldsymbol{\alpha}$. Applying the 2SLS to (2.28) and using the instruments of $(\mathbf{X}_1, \mathbf{Z}_1)$, then we obtain the consistent estimate of $\boldsymbol{\gamma}$ by

$$\hat{\boldsymbol{\gamma}}_W = \left\{\mathbf{Z}'\mathbf{P}_{(\mathbf{X}_1,\mathbf{Z}_1)}\mathbf{Z}\right\}^{-1}\left\{\mathbf{Z}'\mathbf{P}_{(\mathbf{X}_1,\mathbf{Z}_1)}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right\},$$

where

$$\mathbf{P}_{(\mathbf{X}_1,\mathbf{Z}_1)} = (\mathbf{X}_1, \mathbf{Z}_1)\left\{(\mathbf{X}_1, \mathbf{Z}_1)'(\mathbf{X}_1, \mathbf{Z}_1)\right\}^{-1}(\mathbf{X}_1, \mathbf{Z}_1)'$$

is the orthogonal projection operator onto its column space. In sum, $\hat{\boldsymbol{\gamma}}_W$ exists and is consistent as $N \to \infty$ if $k_1 \geq g_2$. We need at least as many instruments $[\mathbf{X}_1, \mathbf{Z}_1]$ as regressors $[\mathbf{Z}_1, \mathbf{Z}_2]$. So you need at least as many $\mathbf{X}_1$'s as $\mathbf{Z}_1$'s.

HT-GLS estimator (consistent and efficient)

Transform (2.27) by $\mathbf{V}^{-1/2}$ as above (θ difference),

$$\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\mathbf{Z}\boldsymbol{\gamma} + \mathbf{V}^{-1/2}(\boldsymbol{\alpha} + \mathbf{u}), \tag{2.29}$$

then do 2SLS using the instruments of $(\mathbf{QX}, \mathbf{Z}_1, \mathbf{PX}_1) = (\mathbf{QX}_1, \mathbf{QX}_2, \mathbf{Z}_1, \mathbf{PX}_1)$ to obtain $\hat{\boldsymbol{\beta}}_{GLS}$ and $\hat{\boldsymbol{\gamma}}_{GLS}$, where $\mathbf{QX}$ are deviations from means of all time-varying variables, $\mathbf{PX}_1$ means of all time-varying variables not correlated with effects, $\mathbf{Z}_1$ time-invariant variables not correlated with effects.

Condition for existence of the estimator now becomes

k_1+k_2+g_1+k_1 (number of instruments) $\geq$ k_1+k_2+g_1+g_2 (number of regressors)

or

$$k_1 \geq g_2,$$

which is the same as for the simple estimator.

Summary of HT estimator

1. $k_1 < g_2$ (underidentification): $\hat{\boldsymbol{\beta}}_W = \hat{\boldsymbol{\beta}}_{GLS}$, but and $\hat{\boldsymbol{\gamma}}_W$ and $\hat{\boldsymbol{\gamma}}_{GLS}$ do not exist.
2. $k_1 = g_2$ (just-identification): $\hat{\boldsymbol{\beta}}_W = \hat{\boldsymbol{\beta}}_{GLS}$, and $\hat{\boldsymbol{\gamma}}_W = \hat{\boldsymbol{\gamma}}_{GLS}$.
3. $k_1 > g_2$ (over-identification): $\hat{\boldsymbol{\beta}}_{GLS}$, $\hat{\boldsymbol{\gamma}}_{GLS}$ are more efficient than $\hat{\boldsymbol{\beta}}_W$, $\hat{\boldsymbol{\gamma}}_W$.

An empirical application: Estimating returns to schooling See Hausman and Taylor (1981).

2.4.2 Further Generalization

Amemiya and McCurdy (1986) and Breusch, Mizon and Schmidt (1989) all consider the same random effects model correlated with some but not all regressors, and suggest a larger set of instruments to improve upon the efficiency of the Hausman and Taylor estimator. A basic question is how

many explanatory variables or some linear combinations must be uncorrelated with the effects in order to improve on the within estimator and to include time-invariant regressors.

Amemiya and McCurdy (1986) suggest the use of the time invariant instruments, $\mathbf{x}_{1,i1} - \bar{\mathbf{x}}_{1,i}$, ..., $\mathbf{x}_{1,iT} - \bar{\mathbf{x}}_{1,i}$, for $\mathbf{z}_{2,i}$. This requires that

$$E\left[\left(\mathbf{x}_{1,it} - \bar{\mathbf{x}}_{1,i}\right)' \alpha_i\right] = 0 \text{ for each } t,$$

which makes sense if the correlation between $\mathbf{x}_{1,it}$ and α_i is due to a time invariant component in $\mathbf{x}_{1,it}$ such that for a given t, $E\left[\mathbf{x}'_{1,it}\alpha_i\right]$ does not depend on t.

Breusch, Mizon and Schmidt (1989) summarise this literature suggesting $\mathbf{x}_{2,i1} - \bar{\mathbf{x}}_{2,i}$, ..., $\mathbf{x}_{2,iT} - \bar{\mathbf{x}}_{2,i}$ as additional instruments for $\mathbf{z}_{2,i}$.

2.5 Extension to two-way error components model

Consider the linear regression model with large N and large T.

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + \lambda_t + u_{it}, \tag{2.30}$$

where α_i denotes the unobservable individual effect and λ_t denotes the unobservable time effect.

2.5.1 The fixed effects model

Regress $\mathbf{y}$ on [$\mathbf{X}$, dummy for individual, dummy for time], which is equivalent to within transformation. Define the following means:

$$y_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it},\ y_{\cdot t} = \frac{1}{N}\sum_{i=1}^{N} y_{it},\ y_{\cdot\cdot} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}.$$

Then, carry out the within transformation:

$$y^w_{it} = y_{it} - y_{i\cdot} - y_{\cdot t} + y_{\cdot\cdot},$$

$$\mathbf{x}^w_{it} = \mathbf{x}_{it} - \mathbf{x}_{i\cdot} - \mathbf{x}_{\cdot t} + \mathbf{x}_{\cdot\cdot}.$$

Then, the within estimator is obtained from the regress y^w_{it} on $\mathbf{x}^w_{it}$. Notice that this transformation removes anything that does not vary over time (eg., α_i) and anything that does not vary over individual (eg., λ_t). The within estimator is unbiased, but consistency depends on.

2.5.2 The random effects model

Treat α_i and λ_t as *iid* draws from $N(0, \sigma^2_\alpha)$ and $N(0, \sigma^2_\lambda)$, and not correlated with $\mathbf{X}$. Then, the OLS estimator is unbiased and consistent as $N \to \infty$ and $T \to \infty$. The GLS estimator is more efficient. For details see Baltagi (2008).

Chapter 3

Dynamic Panels

3.1 Dynamic Panels with Fixed T

Consider a dynamic panel with a lagged dependent variable as regressor,

$$y_{it} = \phi y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{3.1}$$

where we assume an error component specification,

$$\varepsilon_{it} = \alpha_i + u_{it},$$

where $u_i \sim iidN\left(0, \sigma_u^2\right)$.

The motivation here is to distinguish the true dependence of y on lagged y from "spurious" correlation due to unobserved heterogeneity (eg., earning across generations). We also assume for simplicity that $\mathbf{x}_{it}$'s are not correlated with α_i, and $\mathbf{x}_{it}$ is uncorrelated with u_{it} for all $t = 1, ..., T$. Note, however, that α_i is still correlated with $y_{i,t-1}$ since $y_{i,t-1}$ contains α_i. Hence, y_{it-1} is correlated with ε_{it}. This renders the OLS estimator biased and inconsistent even if u_{it}'s are not serially correlated. Thus, OLS and GLS estimators are biased and inconsistent.

Fixed effects would seem to be an obvious possibility, but the within estimator is inconsistent as $N \to \infty$ for fixed T. To illustrate this problem, consider a simple autoregressive model,[1]

$$y_{it} = \phi y_{it-1} + u_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{3.2}$$

where we assume $|\phi| < 1$. The fixed effects estimator is given by

$$\hat{\phi}_W = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - \bar{y}_{i,-1}\right)\left(y_{it} - \bar{y}_i\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - \bar{y}_{i,-1}\right)^2}, \tag{3.3}$$

[1] Extension to a dynamic panel with exogenous variables would be straightforward and in this case we would obtain exactly the same result, at least asymptotically.

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t},\ \bar{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1},$$

and we assume that y_0 exists. Clearly, the within transformation $(y_{i,t-1} - \bar{y}_{i,-1})$ is uncorrelated with α_i, but it is still correlated with u_{it}; that is,

$$Cov\left(y_{i,t-1} - \bar{y}_{i,-1}, u_{it}\right) = -\frac{\sigma_u^2}{T},$$

so the within estimator is biased and inconsistent as $N \to \infty$ for a fixed T. To analyze this, we can substitute (3.2) into (3.3) and get

$$\hat{\phi}_W = \phi + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(u_{it-1} - \bar{u}_{i,-1}\right)\left(y_{it} - \bar{y}_i\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it-1} - \bar{y}_{i,-1}\right)^2}. \tag{3.4}$$

In fact, the fixed effects estimator will be biased of $O\left(T^{-1}\right)$, and in particular, Nickell (1981) has shown that[2]

$$\text{plim}_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(u_{it-1} - \bar{u}_{i,-1}\right)\left(y_{it} - \bar{y}_i\right) = -\frac{\sigma_u^2}{T}\frac{(T-1) - T\phi + \phi^T}{(1-\phi)^2}. \tag{3.5}$$

Therefore, for a typical panel where N is large and T is fixed, the fixed effects estimator is biased and inconsistent.[3] Only if $T \to \infty$, the fixed effects estimator will be consistent for the dynamic error components model.

3.1.1 The Anderson and Hsiao (1981) First-difference IV Estimation

Fortunately, there is a relatively easy way to fix the inconsistency problem. Alternative transformation that wipes out the individual effects yet does not create the above problem in dynamic panels is the first difference transformation. Take first difference of (3.2) to get rid of α_i and obtain

$$\Delta y_{it} = \phi\Delta y_{it-1} + \Delta u_{it},\ t = 2, ..., T, \tag{3.6}$$

where we note that Δu_{it} is an MA(1) process with unit root. The OLS estimator obtained from (4.4) will be inconsistent since Δy_{it-1} and Δu_{it} or more precisely y_{it-1} and u_{it-1} are by definition correlated. Anderson and Hsiao suggested Δy_{it-2} or y_{it-2} as an instrument for Δy_{it-1}. These instruments will not be correlated with Δu_{it} as long as the u_{it} are not serially

[2] See p.328 in Verbeek (2010) for actual magnitude of the bias for a fixed T and for different values of ϕ. See also Phillips and Sul (2003).

[3] The same inconsistency problem occurs with the random effects GLS estimator.

correlated. For example, when using y_{it-2} as an instrument for Δy_{it-1}, then we obtain the following (consistent) IV estimator:[4]

$$\hat{\phi}_{IV} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it-2}\Delta y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it-2}\Delta y_{it-1}}. \tag{3.7}$$

Since $\Delta u_{it} = u_{it} - u_{i,t-1}$, any of $y_{i,s}$, $s \leq t-2$ can be used as legitimate instruments.

3.1.2 The Arellano and Bond (1991) IV-GMM Estimator

It is well-known that imposing more moment conditions increases the efficiency of the estimators provided the additional moment conditions are valid. Arellano and Bond (1991) show that the list of instruments can be extended by exploiting additional moment conditions and letting their number vary with t. In particular, they argue that additional instruments can be obtained if one utilizes the orthogonality conditions that exist between lagged values of y_{it} and u_{it}.

Consider the simple autoregressive panel (3.2) and its first-difference version (4.4). For $t = 3$, we observe (note here $t = 2, ...$, so the observation starts from y_{i1}, not y_{i0})

$$\Delta y_{i3} = \phi \Delta y_{i2} + \Delta u_{i3},$$

and thus y_{i1} is a valid instrument since it is highly correlated with Δy_{i2} but not correlated with Δu_{i3}. Note when $t = 4$, then

$$\Delta y_{i4} = \phi \Delta y_{i3} + \Delta u_{i4},$$

and y_{i2} as well as y_{i1} are valid instruments for Δy_{i3} since both are not correlated with Δu_{i4}. Continuing in this fashion, for the period T, the set of valid instruments becomes $(y_{i1}, y_{i2}, ..., y_{iT-2})$.

Define the $(T-2) \times (1 + 2 + \cdots + T - 2)$ matrix,

$$\mathbf{W}_i = \begin{bmatrix} (y_{i1}) & 0 & \cdots & 0 \\ 0 & (y_{i1}, y_{i2}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (y_{i1}, y_{i2}, ..., y_{iT-2}) \end{bmatrix}, \tag{3.8}$$

as the matrix of instruments, where each row contains the instruments that are valid for a given period. Consequently, the set of all moment conditions

[4] A necessary condition for consistency is that

$$\text{plim}_{N\to\infty} \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=2}^{T} y_{it-2}\Delta u_{it} = 0.$$

can be written concisely as

$$E\left(\mathbf{W}_i'\Delta\mathbf{u}_i\right)=\mathbf{0},$$

where $\Delta\mathbf{u}_i=(\Delta u_3,...,\Delta u_T)'$ or alternatively,

$$E\left(\mathbf{W}_i'\left(\Delta\mathbf{y}_i-\phi\Delta\mathbf{y}_{i,-1}\right)\right)=\mathbf{0},$$

where $\Delta\mathbf{y}_i=(\Delta y_3,...,\Delta y_T)'$ and $\Delta\mathbf{y}_{i,-1}=(\Delta y_2,...,\Delta y_{T-1})'$ are $T\times 2$ vectors, respectively. A total number of moment conditions adds up to $(1+2+\cdots+T-2)$. Next, define the $N(T-2)\times(1+2+\cdots+T-2)$ matrix of instruments by

$$\mathbf{W}=\begin{bmatrix}\mathbf{W}_1\\ \vdots\\ \mathbf{W}_N\end{bmatrix},$$

and rewrite (4.4) in the matrix form as

$$\Delta\mathbf{y}=\phi\Delta\mathbf{y}_{-1}+\Delta\mathbf{u}, \tag{3.9}$$

where

$$\Delta\mathbf{y}=\begin{bmatrix}\Delta\mathbf{y}_1\\ \vdots\\ \Delta\mathbf{y}_N\end{bmatrix}_{N(T-2)\times 1},\Delta\mathbf{y}_{-1}=\begin{bmatrix}\Delta\mathbf{y}_{1,-1}\\ \vdots\\ \Delta\mathbf{y}_{N,-1}\end{bmatrix}_{N(T-2)\times 1},\Delta\mathbf{u}=\begin{bmatrix}\Delta\mathbf{u}_1\\ \vdots\\ \Delta\mathbf{u}_N\end{bmatrix}_{N(T-2)\times 1}.$$

Pre-multiplying (3.9) by $\mathbf{W}'$,

$$\mathbf{W}'\Delta\mathbf{y}=\phi\mathbf{W}'\Delta\mathbf{y}_{-1}+\mathbf{W}'\Delta\mathbf{u}. \tag{3.10}$$

The Arellano and Bond's suggested estimator is the GLS estimator applied to (3.10); that is,

$$\hat{\phi}_{GLS}=\left\{\Delta\mathbf{y}_{-1}'\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\Delta\mathbf{y}_{-1}\right\}^{-1}\left\{\Delta\mathbf{y}_{-1}'\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\Delta\mathbf{y}\right\}, \tag{3.11}$$

where $\mathbf{V}=Var\left(\mathbf{W}'\Delta\mathbf{u}\right)$. They propose the two feasible GLS estimators. First, under the assumption that u_{it} is iid over both i and t, it is easily seen that

$$E\left(\Delta\mathbf{u}_i\Delta\mathbf{u}_i'\right)=\sigma_u^2\left(\mathbf{I}_N\otimes\mathbf{G}\right),$$

where $\Delta\mathbf{u}_i=(\Delta u_3,...,\Delta u_T)'$ and $\mathbf{G}$ is the matrix given by

$$\mathbf{G}=\begin{bmatrix}2&-1&0&\cdots&0&0\\ -1&2&-1&\cdots&0&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&0&\cdots&2&-1\\ 0&0&0&\cdots&-1&2\end{bmatrix}_{(T-2)\times(T-2)}.$$

Then,

$$\mathbf{V} = Var\left(\mathbf{W}'\Delta\mathbf{u}\right) = \sigma_u^2\mathbf{W}'\left(\mathbf{I}_N \otimes \mathbf{G}\right)\mathbf{W}.$$

Therefore, we obtain one-step Arellano and Bond (GMM) estimator by

$$\begin{aligned}\hat{\phi}_{FGLS,1} &= \left\{\Delta\mathbf{y}'_{-1}\mathbf{W}\left[\mathbf{W}'\left(\mathbf{I}_N \otimes \mathbf{G}\right)\mathbf{W}\right]^{-1}\mathbf{W}'\Delta\mathbf{y}_{-1}\right\}^{-1} \\ &\quad \times\left\{\Delta\mathbf{y}'_{-1}\mathbf{W}\left[\mathbf{W}'\left(\mathbf{I_N} \otimes \mathbf{G}\right)\mathbf{W}\right]^{-1}\mathbf{W}'\Delta\mathbf{y}\right\}.\end{aligned} \tag{3.12}$$

Since $\mathbf{G}$ is a fixed matrix, the optimal GMM estimator can be computed in one step if u_{it}'s are assumed to be homoskedastic and exhibit no autocorrelation.

In general, the GMM approach does not impose that u_{it} is iid over both cross-section units and time periods. In this case $\mathbf{V}$ or $\mathbf{V}^{-1}$ can be estimated without imposing these restrictions.[5] Now, we need to replace

$$\mathbf{W}'\left(\mathbf{I}_N \otimes \mathbf{G}\right)\mathbf{W} = \sum_{i=1}^{N}\mathbf{W}_i'\mathbf{G}\mathbf{W}_i,$$

by

$$\mathbf{V}_N = \sum_{i=1}^{N}\mathbf{W}_i'\Delta\mathbf{u}_i\Delta\mathbf{u}_i'\mathbf{W}_i.$$

Since $\Delta\mathbf{u}_i$ is unobservable, we obtain the two-step Arellano and Bond GMM estimator by

$$\hat{\phi}_{FGLS,2} = \left\{\Delta\mathbf{y}'_{-1}\mathbf{W}\widehat{\mathbf{V}}_N^{-1}\mathbf{W}'\Delta\mathbf{y}_{-1}\right\}^{-1}\left\{\Delta\mathbf{y}'_{-1}\mathbf{W}\widehat{\mathbf{V}}_N^{-1}\mathbf{W}'\Delta\mathbf{y}\right\}, \tag{3.13}$$

where

$$\Delta\hat{\mathbf{u}}_i = \Delta\mathbf{y}_i - \hat{\phi}_{FGLS,1}\Delta\mathbf{y}_{-1} \quad \text{and} \quad \widehat{\mathbf{V}}_N^{-1} = \sum_{i=1}^{N}\mathbf{W}_i'\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}_i'\mathbf{W}_i.$$

This GMM estimator requires no knowledge concerning the initial conditions or the distributions of $\mathbf{u}_i$ and α_i. In general, the GMM estimator for ϕ is asymptotically normal with its covariance given by

$$VAR\left(\hat{\phi}_{FGLS,2}\right) = \left\{\Delta\mathbf{y}'_{-1}\mathbf{W}\widehat{\mathbf{V}}_N^{-1}\mathbf{W}'\Delta\mathbf{y}_{-1}\right\}^{-1}.$$

[5] The absence of autocorrelation is necessary for the validity of moment conditions.

GMM estimator in models with exogenous variables

Now we consider a more general dynamic panel model,

$$y_{it} = \phi y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i + u_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{3.14}$$

where $\mathbf{x}_{it}$ is the $k \times 1$ vector of regressors. Its first-difference version becomes

$$\Delta y_{it} = \phi \Delta y_{it-1} + \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}. \tag{3.15}$$

Suppose that the k dimensional regressors $\mathbf{x}_{it}$ are strictly exogenous such that

$$E\left(\mathbf{x}'_{it}u_{is}\right) = \mathbf{0}\ for\ all\ t, s = 1, ..., T,$$

and assume that all $\mathbf{x}_{it}$ are not correlated with the individual effects α_i. Then, all $\mathbf{x}_{it}$ are valid instruments for (3.15) in which case the $1 \times kT$ vector defined by

$$\mathbf{x}_i^* = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$$

should be added to each diagonal element of $\mathbf{W}_i$ in (3.8); that is, we have the $(T-2) \times [(1 + 2 + \cdots + T - 2) + kT(T-2)]$ instrument matrix,

$$\mathbf{W}_i = \begin{bmatrix} (y_{i1}, \mathbf{x}_i^*) & 0 & \cdots & 0 \\ 0 & (y_{i1}, y_{i2}, \mathbf{x}_i^*) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (y_{i1}, y_{i2}, ..., y_{iT-2}, \mathbf{x}_i^*) \end{bmatrix}. \tag{3.16}$$

Writing (3.15) in the matrix form and premultiplying it by $\mathbf{W}'$, we obtain

$$\mathbf{W}'\Delta\mathbf{y} = \phi\mathbf{W}'\Delta\mathbf{y}_{-1} + \mathbf{W}'\Delta\mathbf{X}\boldsymbol{\beta} + \mathbf{W}'\Delta\mathbf{u}, \tag{3.17}$$

where $\Delta\mathbf{y}$, $\Delta\mathbf{y}_{-1}$, $\Delta\mathbf{u}$ are defined just after equation (3.9), and $\Delta\mathbf{X}$ is the stacked $N(T-2) \times k$ matrix of observations on $\Delta\mathbf{x}_{it}$. The two-step GLS estimator can then be obtained by

$$\begin{pmatrix} \hat{\phi}_{FGLS,2} \\ \hat{\boldsymbol{\beta}}_{FGLS,2} \end{pmatrix} = \left(\Delta\mathbf{z}'\mathbf{W}\widehat{\mathbf{V}}_N^{-1}\mathbf{W}'\Delta\mathbf{z}\right)^{-1}\left(\Delta\mathbf{z}'\mathbf{W}\widehat{\mathbf{V}}_{\mathbf{N}}^{-\mathbf{1}}\mathbf{W}'\Delta\mathbf{y}\right), \tag{3.18}$$

where $\mathbf{W} = (\mathbf{W}'_1, ..., \mathbf{W}'_N)'$, $\Delta\mathbf{z} = (\Delta\mathbf{y}_{-1}, \Delta\mathbf{X})$ and $\widehat{\mathbf{V}}_N^{-1}$ is estimated similarly as in (3.13).

Next, if $\mathbf{x}_{it}$ are not strictly exogenous but predetermined such that

$$E(\mathbf{x}_{it}u_{is}) \neq \mathbf{0}\ for\ s < t,$$

(and still assuming that all $\mathbf{x}_{it}$ are not correlated with α_i), then only $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{is-1})$ are valid instruments for (3.15) at period s. Thus, we get the $(T-2) \times$

$[(1+2+\cdots+T-2)+k\,(2+3+\cdots+T-1)]$ matrix of instruments,

$$\mathbf{W}_i = \begin{bmatrix} (y_{i1}, \mathbf{x}_{i1}, \mathbf{x}_{i2}) & 0 & \cdots & 0 \\ 0 & (y_{i1}, y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (y_{i1}, ..., y_{iT-2}, \mathbf{x}_{i1}, ..., \mathbf{x}_{iT-1}) \end{bmatrix}, \tag{3.19}$$

and the two-step GLS estimator of $\begin{pmatrix} \phi \\ \boldsymbol{\beta} \end{pmatrix}$ can be obtained by (3.18), but with the choice of $\mathbf{W}_i$ given by (3.19).

In empirical studies a combination of both exogenous and predetermined regressors may occur rather than two extreme cases, and one can adjust the matrix of instruments $\mathbf{W}$ accordingly. In the case where only subset of $\mathbf{x}_{it}$ are correlated with α_i, then we can also extend the Hausman-Taylor estimation procedure. See Baltagi (2008, section 8.2).

3.1.3 The Arellano and Bover (1995) Study

Arellano and Bover (1995) develop a unifying IV-GMM framework for dynamic panel data models, which also includes the Hausman and Taylor type estimator as a special case. Consider the following static panels:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + \varepsilon_{it},\ i = 1, 2, \ldots, N, ; t = 1, 2, \ldots, T, \tag{3.20}$$

where $\boldsymbol{\beta}$ is $k \times 1$, $\boldsymbol{\gamma}$ is $g \times 1$ and

$$\varepsilon_{it} = \alpha_i + u_{it}.$$

In the vector from,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i = \mathbf{S}_i\boldsymbol{\delta} + \boldsymbol{\varepsilon}_i, \tag{3.21}$$

$$\boldsymbol{\varepsilon}_i = \alpha_i\mathbf{e}_T + \mathbf{u}_i,$$

where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}_{T\times 1},\ \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix}_{T\times k},\ \mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_i \\ \vdots \\ \mathbf{z}_i \end{bmatrix}_{T\times g},\ \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}_{T\times 1},$$

$$\mathbf{S}_i = [\mathbf{X}_i, \mathbf{Z}_i]_{T\times(k+g)},\ \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}_{(k+g)\times 1},\ \mathbf{e}_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{T\times 1},\ \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}_{T\times 1}.$$

Arellano and Bover transform (3.21) using the following nonsingular matrix transformation:

$$\mathbf{H} = \begin{bmatrix} \mathbf{C} \\ \mathbf{e}_T'/T \end{bmatrix}_{T\times T},$$

where $\mathbf{C}$ is any $(T-1)\times T$ matrix of rank $(T-1)$ such that $\mathbf{C}\mathbf{e}_T = 0$. For example, $\mathbf{C}$ could be the first rows of the within group operator (see definition of $\mathbf{Q}_*$) or the first difference operator. Premultiplying $\boldsymbol{\varepsilon}_i$ by $\mathbf{C}$, then we obtain the transformed disturbances,

$$\boldsymbol{\varepsilon}_i^* = \mathbf{H}\boldsymbol{\varepsilon}_i = \begin{bmatrix} \mathbf{C}\boldsymbol{\varepsilon}_i \\ \bar{\varepsilon}_i \end{bmatrix}. \tag{3.22}$$

For example, this class of transformation performs a decomposition between 'within-group' and 'between-group' variation which is helpful in order to implement moment conditions implied by the model. Notice that the first $(T-1)$ transformed disturbances are free of the individual effects α_i by construction. Hence, all exogenous variables are valid instruments for the first $(T-1)$ equations in (3.21).

Define the $1\times(kT+g)$ vector $\mathbf{w}_i = (\mathbf{x}_{i1},...,\mathbf{x}_{iT},\mathbf{z}_i)$, and let $\mathbf{m}_i$ denote the row vector of subset of variables of $\mathbf{w}_i$ assumed to be uncorrelated with α_i such that $\dim(\mathbf{m}_i) = m \geq \dim(\boldsymbol{\gamma}) = g$. Then, a valid instrument matrix becomes

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{w}_i & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & & \mathbf{w}_i & 0 \\ 0 & 0 & 0 & \mathbf{m}_i \end{bmatrix}_{T\times(kT+g+m)}, \tag{3.23}$$

and the moment conditions are given by

$$E\left(\mathbf{W}_i'\mathbf{H}\boldsymbol{\varepsilon}_i\right) = 0. \tag{3.24}$$

Write (3.21) in matrix form,

$$\mathbf{y} = \mathbf{S}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{3.25}$$

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_N \end{bmatrix}_{NT\times(k+g)}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}_{NT\times 1}.$$

Defining the matrix of instruments,

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_N \end{bmatrix}_{NT\times(kT+g+m)},$$

and premultiplying (3.25) by $\mathbf{W}\bar{\mathbf{H}}$, where $\bar{\mathbf{H}} = \mathbf{I}_N \otimes \mathbf{H}$ is a $NT\times NT$ matrix, then we obtain the following complete transformed system:

$$\mathbf{W}'\bar{\mathbf{H}}\mathbf{y} = \mathbf{W}'\bar{\mathbf{H}}\mathbf{S}\boldsymbol{\delta} + \mathbf{W}'\bar{\mathbf{H}}\boldsymbol{\varepsilon}. \tag{3.26}$$

The Arellano-Bover optimal GMM estimator of $\boldsymbol{\delta}$ based on moment condition (3.24) is the GLS estimator applied to (3.26) given by

$$\hat{\boldsymbol{\delta}}_{GLS} = \left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{S}\right)^{-1}\left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{y}\right), \tag{3.27}$$

where

$$Var\left(\boldsymbol{\varepsilon}\right) = \mathbf{I}_N \otimes E\left(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right) = \mathbf{I}_N \otimes \boldsymbol{\Omega},$$

$$\mathbf{V} = Var\left(\mathbf{W}'\bar{\mathbf{H}}\boldsymbol{\varepsilon}\right) = \mathbf{W}'\bar{\mathbf{H}}\left(\mathbf{I}_N \otimes \boldsymbol{\Omega}\right)\bar{\mathbf{H}}'\mathbf{W} = \mathbf{W}'\left(\mathbf{I}_N \otimes \mathbf{H}\boldsymbol{\Omega}\mathbf{H}'\right)\mathbf{W}.$$

The feasible GLS estimator is obtained by replacing $\mathbf{H}\boldsymbol{\Omega}\mathbf{H}'$ or $\mathbf{V}$ by its consistent estimator. First, unrestricted estimator of $\mathbf{H}\boldsymbol{\Omega}\mathbf{H}'$ takes the form,

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\varepsilon}}_i^*\hat{\boldsymbol{\varepsilon}}_i^{*\prime},$$

where $\hat{\boldsymbol{\varepsilon}}_i^*$ are the residuals based on consistent preliminary estimates, in which case we have

$$\hat{\boldsymbol{\delta}}_{FGLS} = \left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\hat{\mathbf{V}}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{S}\right)^{-1}\left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\hat{\mathbf{V}}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{y}\right), \tag{3.28}$$

where

$$\hat{\mathbf{V}} = \mathbf{W}'\left(\mathbf{I}_N \otimes \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\varepsilon}}_i^*\hat{\boldsymbol{\varepsilon}}_i^{*\prime}\right)\right)\mathbf{W}$$

Alternatively, we consider a restricted estimate of $\boldsymbol{\Omega}$,

$$\tilde{\boldsymbol{\Omega}} = \tilde{\sigma}_\alpha^2\mathbf{e}_T\mathbf{e}_T' + \tilde{\sigma}_u^2\mathbf{I}_T,$$

where $\tilde{\sigma}_\alpha^2$ and $\tilde{\sigma}_u^2$ are the consistent estimators of σ_α^2 and σ_u^2 (recall random effects model). Thus, we have

$$\tilde{\boldsymbol{\delta}}_{FGLS} = \left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\tilde{\mathbf{V}}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{S}\right)^{-1}\left(\mathbf{S}'\bar{\mathbf{H}}\mathbf{W}\tilde{\mathbf{V}}^{-1}\mathbf{W}'\bar{\mathbf{H}}\mathbf{y}\right), \tag{3.29}$$

where

$$\tilde{\mathbf{V}} = \mathbf{W}'\left(\mathbf{I}_N \otimes \mathbf{H}\tilde{\boldsymbol{\Omega}}\mathbf{H}'\right)\mathbf{W}.$$

Consider the Hausman and Taylor model again,

$$y_{it} = \mathbf{x}_{1,it}\boldsymbol{\beta}_1 + \mathbf{x}_{2,it}\boldsymbol{\beta}_2 + \mathbf{z}_{1,i}\boldsymbol{\gamma}_1 + \mathbf{z}_{2,i}\boldsymbol{\gamma}_2 + \alpha_i + u_{it},\ i = 1, 2, \ldots, N, ; t = 1, 2, \ldots, T, \tag{3.30}$$

where the $1 \times k_1$ vector $\mathbf{x}_{1,it}$ and the $1 \times g_1$ vector $\mathbf{z}_{1,i}$ are uncorrelated with α_i, but the $1 \times k_{21}$ vector $\mathbf{x}_{2,it}$ and the $1 \times g_2$ vector $\mathbf{z}_{2,i}$ are correlated with α_i ($k_1 + k_2 = k$ and $g_1 + g_2 = g$). Using the Arellano-Bover transformation, it is easily seen that $\mathbf{m}_i$ include the set of and variables $\mathbf{x}_{1,it}$ and $\mathbf{z}_{1,i}$, namely,

$$\mathbf{m}_i = \left(\mathbf{x}_{1,i1}, ..., \mathbf{x}_{1,iT}, \mathbf{z}_{1,i}\right).$$

Then, the Hausman and Taylor is equivalent to $\tilde{\boldsymbol{\delta}}_{FGLS}$ in (3.29).

Because the set of instruments $\mathbf{W}_i$ is block-diagonal, it can be shown that $\hat{\boldsymbol{\delta}}_{FGLS}$ or $\tilde{\boldsymbol{\delta}}_{FGLS}$ is invariant to the choice of the transformation matrix $\mathbf{C}$. Another advantage is that the form of $\Omega^{-\frac{1}{2}}$ (for GLS transformation) need not be known. Hence, this approach generalises the HT, AM and BMS type estimators to a more general form than that of error components, and it is easily extends to the dynamic panels.

Consider the dynamic panels:

$$y_{it} = \phi y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + \varepsilon_{it},\ i = 1, 2, \ldots, N, ; t = 1, 2, \ldots, T, \qquad (3.31)$$

where $\boldsymbol{\beta}$ is $k \times 1$, $\boldsymbol{\gamma}$ is $g \times 1$ and

$$\varepsilon_{it} = \alpha_i + u_{it}.$$

In the vector from,

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\phi + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i = \mathbf{S}_i\boldsymbol{\delta} + \boldsymbol{\varepsilon}_i, \qquad (3.32)$$

$$\boldsymbol{\varepsilon}_i = \alpha_i\mathbf{e}_T + \mathbf{u}_i,$$

where[6]

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}_{T\times 1},\ \mathbf{y}_{i,-1} = \begin{bmatrix} y_{i0} \\ \vdots \\ y_{iT-1} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix}_{T\times k},\ \mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_i \\ \vdots \\ \mathbf{z}_i \end{bmatrix}_{T\times g},\ \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix},$$

$$\mathbf{S}_i = [\mathbf{y}_{i,-1}, \mathbf{X}_i, \mathbf{Z}_i]_{T\times(1+k+g)},\ \boldsymbol{\delta} = \begin{bmatrix} \phi \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}_{(1+k+g)\times 1},\ \mathbf{e}_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},\ \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}.$$

Provided there are enough valid instruments to ensure identification of all parameters, the GMM estimator defined in (3.27) still provides a consistent estimator of $\boldsymbol{\delta}$ in (3.32). Now, the matrix of instruments $\mathbf{W}_i$ is basically the same as in (3.23), where $1 \times (k(T+1) + g)$ vector $\mathbf{w}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, ..., \mathbf{x}_{iT}, \mathbf{z}_i)$ and $\mathbf{m}_i$ is the row vector of subset of variables of $\mathbf{w}_i$ assumed to be uncorrelated with α_i such that $\dim(\mathbf{m}_i) = m \geq \dim(\boldsymbol{\gamma}) = g$. (The only difference is that $t = 0$ is now our first time period observed.) But, notice that $\mathbf{y}_{i,-1}$ is excluded despite its presence in $\mathbf{S}_i$, because $\mathbf{y}_{i,-1}$ is generally regarded as endogenous unless u_{it} is serially uncorrelated (see below). Then, the previous steps of consistent estimation follow.

Finally, consider the case where u_{it} is not serially correlated, so that y_{it-1} is predetermined. In this case we could obtain additional orthogonality restrictions under the additional condition that the transformation matrix

[6] We assume that y_{i0} exists.

C is upper triangular. In this case the transformed error in the equation for period t is independent of α_i and $(u_{i1}, ..., u_{it-1})$ so that $(y_{i0}, y_{i1}, ..., y_{it-1})$ are additional valid instruments. Thus, a valid instrument matrix now becomes

$$\mathbf{W}_i = \begin{bmatrix} (\mathbf{w}_i, y_{i0}) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & & (\mathbf{w}_i, y_{i0}, y_{i1}, ..., y_{it-2}) & 0 \\ 0 & 0 & 0 & \mathbf{m}_i \end{bmatrix}_{T \times (kT+g+m)}, \quad (3.33)$$

and the consistent GMM estimation follows accordingly.

3.1.4 Further Readings

There are a few paper worth reading for further studies. Ahn and Schmidt (1995) observed that the standard IV-GMM estimator suggested by Arellano and Bond (1991) neglects quite a lot of information and is therefore inefficient. They explain how these more moment conditions arise for the simple dynamic model and show how they can be utilized in a GMM framework. Ahn and Schmidt also consider the dynamic model with exogenous regressors and show how one can make efficient use of exogenous variables as instruments. Holtz-Eakin *et. al* (1988) discuss the VAR in panels and Blundell and Bond (1998) examine initial conditions and moment conditions in dynamic panels.

3.2 Dynamic Panels When Both N and T are Large

In recent years there has been an upsurge in the availability and use of panel data sets where both N and T are sufficiently large. For example Summers and Heston's large multi-country panel data has been and still is the focus of much empirical work in the area of macroeconomic growth. Furthermore, as time progresses, micro panel data sets such as the British Household Panel Survey (BHPS), where T is not currently so large, are being updated to incorporate new time series observations as they become available. It is recognized that large panels can be hugely informative about the unknown parameters of economic models and yield very powerful tests of hypotheses nested within these models.

These large N, large T panels raise a number of issues. First of all, since it is possible to estimate a separate regression for each group, which is not possible in the small T case, it is natural to think of heterogeneous panels where the parameters can differ over groups. One can then test for equality of the parameters, rather than having to assume it, as one is forced to do in the small T case. When equality of parameters over groups is tested it is very often rejected and the differences in the estimates between groups can be large. Baltagi and Griffin (1997) discuss the dispersion in

OECD country estimates for gasoline demand functions. Boyd and Smith (2000) review possible explanations for the large dispersion in models of the monetary transmission mechanism for 57 developing countries.

Secondly, since time-series data tend to be non-stationary, determining the order of integration or cointegration of the variables becomes important. Extending the estimation and testing procedures for integrated and cointegrated series to panels is thus a natural development. In fact, the tendency of individual time series tend to reject Purchasing Power Parity has led the emphasis to switch to testing PPP in panels.

Third, one needs to determine the asymptotic properties of standard panel estimators when the data are non-stationary. These properties are rather different from those of single time-series, in particular spurious regression seems to be less of a problem. There is also a question about how to do the asymptotic analysis as both N and T can go to infinity. There are a number of different ways that N and T can go to infinity and the relation between these different ways remains a subject of research, see Smith and Fluertes (2012).

3.2.1 The Mean Group Estimator

The conventional dynamic panel model with small T focuses only on allowing for intercept variation via individual effects. In comparison little attention has been paid to the implications of variation in slope parameters. There are three justifications for analyzing the dynamic heterogeneous panels.

First, the slope heterogeneity does not matter when the primary interest lies in obtaining an unbiased estimator of the average effect of exogenous variables. Zellner (1969) showed that when the regressors are exogenous but their coefficients differ randomly across groups, the pooled estimators such as the fixed and random effects estimator will provide unbiased estimator of the average effect. However, Pesaran and Smith (1995) showed that such results do not extend to dynamic models with lagged dependent variables.

Second, in the case where only relatively small time periods were available, the scope for analysing the slope heterogeneity explicitly appeared limited, e.g. seminal paper by Balestra and Nerlove (1966). Considering now that panels with a reasonable time dimension are available and that the evidence of slope heterogeneity in panels are pervasive, it is a high time to examine the implications of slope heterogeneity directly.

A third possible justification is that the long-run responses which are often the primary focus of analysis are less likely to be subject to slope heterogeneity than the short-run adjustment patterns across groups. Therefore, it is interesting to see how the time-series, cross-section and panel estimates of such long-run coefficients can be compared.

In practice the extent of cross-sectional heterogeneity may be so large as to preclude the use of pooling. An approach that is becoming increasingly

popular in this context is to focus estimation and inference on so called mean group quantities that are "averages" across panel units, see Pesaran and Smith (1995) and Im, Pesaran and Shin (2003).

Consider the following dynamic heterogeneous panels,

$$y_{it} = \phi_i y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta}_i + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{3.34}$$

where we assume:

1. $\mathbf{x}_{it}$ and ε_{is} are uncorrelated each other for all t and s.
2. $\phi_i \sim iidN\left(\phi, \omega_1^2\right)$.
3. $\boldsymbol{\beta}_i \sim iidN\left(\boldsymbol{\beta}, \boldsymbol{\Omega}_2\right)$.
4. ϕ_i and $\boldsymbol{\beta}_i$ are independently distributed with y_{1t},$\mathbf{x}_{it}$ and ε_{it} for all t.
5. The k-dimensional vector $\mathbf{x}_{it}$ are covariance stationary processes.[7]

Under the second and third assumptions this can be regarded as the standard random coefficients model, see Swamy (1970). Under this scenario Pesaran and Smith (1995) have investigated the asymptotic as well as small sample properties of the following three alternative estimators:

1. The pooled (within) estimator, which involves pooling the data by imposing homogeneous slope coefficients, allowing for fixed or random effects.
2. The cross-section (between) estimator, which involves averaging the data over time per each group and estimating the cross-section regression based on the group mean data. Notice that many works on endogenous growth follow this approach, e.g. Barro (1991).
3. The mean group estimator suggested by Pesaran and Smith (1995), which involves estimating separate regression for each group and averaging the individual estimates over groups, that is,

$$\hat{\boldsymbol{\beta}}_{MG} = N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i, \tag{3.35}$$

where $\hat{\boldsymbol{\beta}}_i$ is the OLS estimator of $\boldsymbol{\beta}_i$.

Notice that all the above procedures will provide an estimate of the average coefficient but main difference is that in the first two cases the averaging is implicit while in the third case it is explicit.

[7] In general most of the results follow when $\mathbf{x}_{it}$ are unit root processes.

1. When T is small (even if N is large), all the procedures yield inconsistent estimators.

2. When both T and N are large, both the cross-section (between) estimator and the mean group estimator yield consistent estimators of ϕ and $\boldsymbol{\beta}$.

3. The pooled (within) estimator is not consistent for the expected values of β_i and γ_i even when both T and N are large.[8]

To see that the within estimator is inconsistent we rewrite (3.34) as[9]

$$y_{it} = \phi y_{it-1} + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \tag{3.36}$$

where

$$v_{it} = \varepsilon_{it} + (\lambda_i - \lambda)\, y_{i,t-1} + \mathbf{x}_{it}\,(\boldsymbol{\beta}_i - \boldsymbol{\beta})\,.$$

Then it is easily seen that v_{it} is now correlated with all present and past values of $y_{i,t-1-s}$ and $\mathbf{x}_{it-s}$ for all $s \geq 0$. This correlation renders the OLS estimator inconsistent. Furthermore, the fact that v_{it} is correlated with $y_{i,t-1-s}$ and $\mathbf{x}_{it-s}$ for all $s \geq 0$ rules out the possibility of choosing lagged values of $y_{i,t-1}$ and $\mathbf{x}_{it}$ as legitimate instruments. This composite disturbance v_{it} will also be serially correlated, if x_{it} is serially correlated, as it usually is, and will not be independent of the lagged dependent variable. This heterogeneity bias, which depends on the serial correlation in the x and the variance of the random parameters, can be quite severe.

3.2.2 Pooled mean group estimation in dynamic heterogeneous panels

To date, there have been three alternative estimation procedures for dynamic panels, differing in the relative magnitudes of N and T.

1. (**Small N and large T**) When $N = 1$, the traditional approach was to estimate an autoregressive distributed lag (ARDL) model. For $N > 1$, the seemingly unrelated regression equation (SURE) procedure is often used. The main attraction of the SURE procedure lies in the fact that it allows the contemporaneous error covariances to be freely estimated. However, this is possible only when N is reasonably small relative to T. When N is of the same order of magnitude as T, the case we are interested in, SURE is not feasible.

[8] This inconsistency in heterogenous dynamic panel models was first noted by Robertson and Symonds (1992).

[9] Remind that if $\beta_i = \beta$ and $\gamma_i = \gamma$, then the fixed effect estimators are consistent only as $T \to \infty$, for fixed N. However, they are inconsistent as $N \to \infty$ for fixed T. The latter is the result of the fact that the lagged dependent variable bias arising from the initial conditions, is not removed by increasing N.

2. (**Small T and large N**) Pesaran and Smith (1995) show that the traditional procedures for estimation of pooled models such as the fixed effects, the random effects, the instrumental variables (IV) or the Generalized Method of Moments (GMM) estimators can produce inconsistent, and potentially very misleading estimates of the average values of the parameters unless the slope coefficients are homogeneous. In most panels of this sort, however, tests indicate that these parameters differ significantly across groups.

3. (**The Bayes and empirical Bayes estimators**) Hsiao, Pesaran and Tahmiscioglu (1998) consider Bayes estimation of short-run coefficients in dynamic heterogenous panels, and establish the asymptotic equivalence of the Bayes estimator (Swamy, 1970)) and the mean group estimator, showing that the mean group estimator is asymptotically normal for large N and large T so long as $\sqrt{N}/T \to 0$ as both N and $T \to \infty$.

When both N and T are sufficiently large, there have been two extreme approaches to analyzing dynamic panels. At one extreme are the traditional pooled estimators, such as the fixed and random effects estimators, where only the intercepts are allowed to differ across groups while all other coefficients and error variances are constrained to be the same. At the other extreme, one can estimate separate equations for each group and examine the distribution of the estimated coefficients across groups. Of particular interest is the mean group estimator by Pesaran and Smith (1995).

There are often good reasons to expect the long-run equilibrium relationships between variables to be similar across groups, due to budget or solvency constraints, arbitrage conditions or common technologies influencing all groups in a similar way. However, the reasons for assuming that short-run dynamics and error variances should be the same tend to be less compelling. On this ground Pesaran, Shin and Smith (1999) provide an intermediate estimator called the pooled mean group estimator (PMGE) which involves both pooling and averaging. This estimator allows the intercepts, short-run coefficients and error variances to differ freely across groups, but only the long-run coefficients are constrained to be the same.

We extend the single time series ARDL modelling to the dynamic panel data model,

$$y_{it} = \sum_{j=1}^{p} \lambda_{ij} y_{i,t-j} + \sum_{j=0}^{q} \mathbf{x}_{i,t-j} \boldsymbol{\delta}_{ij} + \alpha_i + \varepsilon_{it}, \quad \begin{matrix} t = 1, 2, ..., T \\ i = 1, 2, ..., N \end{matrix}, \tag{3.37}$$

where α_i represent the fixed effects, $\mathbf{x}_{i,t}$ is a $1 \times k$ vector of regressors, and λ_{ij}, $\boldsymbol{\delta}_{ij}$ are scalar and $k \times 1$ vector of parameters. It is convenient to work

with the following (unrestricted) error correction form of (3.37):

$$\Delta y_{it} = \phi_i y_{i,t-1} + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{j=1}^{p-1} \lambda_{ij}^* \Delta y_{i,t-j} + \sum_{j=0}^{q-1} \Delta \mathbf{x}_{i,t-j} \boldsymbol{\delta}_{ij}^* + \mu_i + \varepsilon_{it}, \quad (3.38)$$

where

$$\phi_i = -\left(1 - \sum_{j=1}^{p} \lambda_{ij}\right) \quad \text{and} \quad \boldsymbol{\beta}_i = \sum_{j=0}^{q} \boldsymbol{\delta}_{ij}.$$

If we stack the time-series observations for each group, (3.38) can be written as

$$\Delta \mathbf{y}_i = \phi_i \mathbf{y}_{i,-1} + \mathbf{X}_i\boldsymbol{\beta}_i + \sum_{j=1}^{p-1} \lambda_{ij}^* \Delta \mathbf{y}_{i,-j} + \sum_{j=0}^{q-1} \Delta \mathbf{X}_{i,-j} \boldsymbol{\delta}_{ij}^* + \mu_i \boldsymbol{\iota}_T + \boldsymbol{\varepsilon}_i, \quad (3.39)$$

where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}_{T\times 1}, \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix}_{T\times k}, \quad \boldsymbol{\iota}_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{T\times 1}, \quad \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}_{T\times 1}.$$

We now assume:

- Assumption 1. ε_{it}'s are independently distributed across i and t, with zero means, variances $\sigma_i^2 > 0$, and finite fourth-order moments. They are also distributed independently of the regressors, $\mathbf{x}_{it}$.

- Assumption 2. The $ARDL(p, q, q, ..., q)$ model (3.37) is stable in that the roots of

$$\sum_{j=1}^{p} \lambda_{ij} z^j = 1, \; i = 1, 2, ..., N,$$

lie outside the unit circle. This assumption ensures that

$$\phi_i < 0 \; for \; all \; i = 1, 2, ..., N,$$

and hence there exists a long-run relationship between y_{it} and $\mathbf{x}_{it}$ defined by

$$y_{it} = \boldsymbol{\theta}_i \mathbf{x}_{it} + \eta_{it},$$

where $\boldsymbol{\theta}_i = -\boldsymbol{\beta}_i'/\phi_i$ are long-run coefficients and η_{it} is a stationary process.

- Assumption 3. (Long-run homogeneity) The long-run coefficients $\boldsymbol{\theta}_i$ are the same across the groups, i.e.

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}, \; i = 1, 2, ..., N. \quad (3.40)$$

Under Assumptions 2 and 3, (3.39) can be written compactly as

$$\Delta \mathbf{y}_i = \phi_i \boldsymbol{\xi}_i(\boldsymbol{\theta}) + \mathbf{W}_i \boldsymbol{\kappa}_i + \boldsymbol{\varepsilon}_i,\ i = 1, 2, ..., N, \tag{3.41}$$

where we highlight the dependence of the error correction term on $\boldsymbol{\theta}$ as

$$\boldsymbol{\xi}_i(\boldsymbol{\theta}) = \mathbf{y}_{i,-1} - \mathbf{X}_i \boldsymbol{\theta},\ i = 1, 2, ..., N, \tag{3.42}$$

$\mathbf{W}_i = (\Delta \mathbf{y}_{i,-1}, ..., \Delta \mathbf{y}_{i,-p+1}, \Delta \mathbf{X}_i, \Delta \mathbf{X}_{i,-1}, ..., \Delta \mathbf{X}_{i,-q+1}, \boldsymbol{\iota}_T)$, and $\boldsymbol{\kappa}_i = (\lambda^*_{i1}, ..., \lambda^*_{i,p-1}, \boldsymbol{\delta}^{*\prime}_{i0}, \boldsymbol{\delta}^{*\prime}_{i1}, ..., \boldsymbol{\delta}^{*\prime}_{i,q-1}, \mu_i)'$.

We adopt a likelihood approach. Since the parameters of interest are the long-run effects and adjustment coefficients, we directly work with the concentrated log-likelihood function given by

$$\ell_T(\boldsymbol{\varphi}) = -\frac{T}{2} \sum_{i=1}^{N} \ln 2\pi\sigma_i^2 - \frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} ((\Delta \mathbf{y}_i - \phi_i \boldsymbol{\xi}_i(\boldsymbol{\theta}))' \mathbf{H}_i (\Delta \mathbf{y}_i - \phi_i \boldsymbol{\xi}_i(\boldsymbol{\theta})), \tag{3.43}$$

where

$$\mathbf{H}_i = \mathbf{I}_T - \mathbf{W}_i (\mathbf{W}_i' \mathbf{W}_i)^{-1} \mathbf{W}_i'$$

$$\boldsymbol{\varphi} = (\boldsymbol{\theta}', \boldsymbol{\phi}', \boldsymbol{\sigma}')', \quad \boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_N)', \quad \boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2, ..., \sigma_N^2)'.$$

The maximum likelihood (ML) estimators of the long-run coefficients, $\boldsymbol{\theta}$, and the group-specific error-correction coefficients, ϕ_i, will be referred to as the "pooled mean group" (PMG) estimators in order to highlight both the pooling implied by the homogeneity restrictions on the long-run coefficients and the averaging across groups used to obtain means of the short-run parameters of the model.

The Case of Stationary Regressors In this case under fairly standard conditions the consistency and the asymptotic normality of The ML estimators of $\boldsymbol{\varphi}$in (3.43) can be easily established.

The Case of Non-Stationary Regressors The asymptotic analysis in this case is more complicated, but we can still show that the ML estimator of the short-run coefficients $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}$ in the dynamic heterogeneous panel data model (3.41) are $\sqrt{T}$ consistent, and the ML estimator of $\boldsymbol{\theta}$ is T consistent. Furthermore, for a fixed N and as $T \to \infty$, the ML estimator of $\boldsymbol{\psi} = (\boldsymbol{\theta}', \boldsymbol{\phi}')'$ asymptotically has the (mixture) normal distribution.

In sum, for the common long-run coefficients, $\boldsymbol{\theta}$, the pooled ML estimator is consistent so long as $T \to \infty$, irrespective of whether N is large or not, but $\hat{\theta}$ will not be consistent for finite T, even if $N \to \infty$.

Once the pooled ML estimator of the long run parameters, $\hat{\boldsymbol{\theta}}$, is estimated, all other short run coefficients can be consistently estimated by running the individual OLS regressions,

$$\Delta \mathbf{y}_i = \phi_i \hat{\boldsymbol{\xi}}_i + \mathbf{W}_i \kappa_i + error,\ i = 1, ..., N,$$

where

$$\hat{\boldsymbol{\xi}}_i = \mathbf{y}_{i,-1} - \mathbf{X}_i \hat{\boldsymbol{\theta}}.$$

In this case the mean of the error correction coefficients and the other short run parameters can be estimated consistently by the MGE,

$$\hat{\phi}_{MG} = N^{-1} \sum_{i=1}^{N} \hat{\phi}_i \quad \text{and} \quad \hat{\boldsymbol{\kappa}}_{MG} = N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{\kappa}}_i.$$

Empirical Applications: The Consumption Function in the OECD

See Pesaran, Shin and Smith (1999). For further empirical application see Fedderke, Shin and Vaze (2012).

3.3 Estimation and Inference in Panels with Non-stationary Variables

See Panel Time-Series by Smith and Fuertes (2012).

Chapter 4

Threshold Regression Models in Dynamic Panels

4.1 Introduction

We have observed many stylized facts about economic time series as follows:

1. Business cycles are asymmetric in nature, e.g. Burns and Mitchell (1946); namely recessions last longer than expansion.
2. Asset pricing model under noise trading and transaction costs: The larger are the pricing errors, the larger is the expected degree of arbitrage and hence the speedier is the price response to disequilibrium and *vice versa*.
3. Asymmetries are intrinsic to microeconomic behavior. For instance, costs of hiring and firing are asymmetric at the firm level.
4. Asymmetries can result from capital constraints on the goods market.
5. Imperfect competition and\or government interventions cause rigidities on credit, goods and labour markets that affect the dynamics of the economy.

It is increasingly recognised that the implications of linear modes are problematic in dealing with the above observations reflected in various economics and finance applications. In particular, the followings are questionable:

- Linearity, invariance of dynamic multipliers with respect to the size and the sign of the shock and the history of the system.
- Time invariance of the parameters.

Consequently, a great deal of interest has been made in modelling nonlinearities and asymmetries in economic time series.

4.2 Regime Switching Models

Most attention has fallen almost exclusively on regime-switching models, though there is no consensus suggesting a unique approach for specifying econometric models that embed various types of change in regimes.

- Regime shifts are not considered as singular deterministic events but the unobservable regime is assumed to be governed by an exogenous or predetermined stochastic processes. Thus regime shifts of the past are expected to occur in the future in a similar fashion.
- Regime switching models characterise a nonlinear data generating process as piecewise linear by restricting the process to be linear in each regime.
- The models differ in their assumptions concerning the stochastic process generating the regime; TAR, STAR, MS-AR, etc.

4.2.1 Structural break models

Suppose that the structural break occurs at time $t = \tau$ and we have

$$y_t = \left\{ \begin{array}{ll} \alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i} + \varepsilon_t & \text{for } t < \tau \\ \alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i} + \varepsilon_t & \text{for } t \geq \tau \end{array} \right\}, \tag{4.1}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. Then, (4.1) can be written as

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right)\left(1 - I\left(t;\tau\right)\right) + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) I\left(t;\tau\right) + \varepsilon_t,$$

where $I\left(t;\tau\right)$ is the indicator function given by

$$I\left(t;\tau\right) = \left\{ \begin{array}{ll} 0 & \text{for } t < \tau \\ 1 & \text{for } t \geq \tau \end{array} \right\}.$$

Two different assumptions have been made:

The break point τ is known So break is deterministic. To estimate (4.1), split the sample and apply OLS to each regime. Tests of $\beta_{1i} = \beta_{2i}$, $i = 1, ..., p$, will follow the standard χ^2 distribution asymptotically. See Perron (1989) for unit root tests subject to structural breaks.

The break point τ is unknown So break is stochastic and τ needs to be estimated as follow:

$$\begin{aligned} \tau^* &= \arg \min_{\tau \in [0.15T, 0.85T]} RSS\left(\tau\right) \\ &= \arg \min_{\tau \in [0.15T, 0.85T]} \left[\tau \hat{\sigma}_1^2\left(\tau\right) + \left(1 - \tau\right) \hat{\sigma}_2^2\left(\tau\right)\right], \end{aligned}$$

where RSS stands for residual sum of squares, and the grid search is over $\tau \in [0.15T, 0.85T]$ in practice.

Notice that tests of $\beta_{1i} = \beta_{2i}$, $i = 1, ..., p$, does not follow the standard χ^2 distribution asymptotically, but has a nonstandard asymptotic distribution due to the Davies (1987) problem that nuisance parameter (break point τ) is not identified under the null. Most solutions to this problem involve integrating out unidentified parameters from the test statistics. This is usually achieved by calculating test statistics over a grid set of possible values of nuisance parameter and constructing the summary statistics such as sup (maximum) and exponential average, see Andrews and Plobegrer (1994).

Threshold models This is a popular class of nonlinear regime-switching models with each regime determined by observed variables.

Threshold Autoregressive (TAR) model

Now the regime shifts are triggered by an observable, exogenous transition variable x_t crossing threshold c, and consider the two-regime TAR model:

$$y_t = \left\{ \begin{array}{ll} \alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i} + \varepsilon_t & \text{for } x_t \leq c \text{ (regime 1)} \\ \alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i} + \varepsilon_t & \text{for } x_t > c \text{ (regime 2)} \end{array} \right\}, \tag{4.2}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. Alternatively,

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \mathbf{1}\left\{x_t \leq c\right\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) \left(1 - \mathbf{1}\left\{x_t \leq c\right\}\right) + \varepsilon_t, \tag{4.3}$$

where $\mathbf{1}\left\{x_t \leq c\right\}$ is the indicator function.

Self-Exciting Threshold Autoregressive (SETAR) model

If we use as the transition variable a lagged endogenous variable y_{t-d} with delay $d \geq 1$, we obtain the two-regime SETAR model as follow:

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \mathbf{1}\left\{y_{t-d} \leq c\right\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) \left(1 - \mathbf{1}\left\{y_{t-d} \leq c\right\}\right) + \varepsilon_t, \tag{4.4}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$.

Notice that (4.4) can be written alternatively as

$$y_t = \alpha\left(s_t\right) + \sum_{i=1}^{p} \beta_i\left(s_t\right) y_{t-i} + \varepsilon_t, \tag{4.5}$$

where the probability of the unobservable regime 1 is given by

$$\Pr\left(s_t = 1 | S_{t-1}, Y_{t-1}\right) = \mathbf{1}\left\{y_{t-d} \leq c\right\},$$

where $S_{t-1} = \{s_{t-1}, s_{t-2}, ...\}$ and $Y_{t-1} = \{y_{t-1}, y_{t-2}, ...y_{t-p}\}$. This shows that SETAR and MS-AR models can be observationally equivalent, see Carrasco (1994).

Estimation

ML estimation under normality can be carried over the grid search over d and c: select the pair(d, c) that minimises the residual sum of squares:

$$\arg\min_{(d,c)} \sum_{m=1}^{M} T_m \hat{\sigma}_m^2,$$

where T_m and $\hat{\sigma}_m^2$ are the number of observations and the residual variance in regime m. Usually the grid search is restricted such that $\min T_m \geq 0.15T$.

Three-regime SETAR model

We now extend to consider the three-regime SETAR model:

$$\begin{aligned} y_t &= \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \mathbf{1}\{y_{t-d} \leq c_1\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) \mathbf{1}\{c_1 < y_{t-d} \leq c_2\} \\ &\quad + \left(\alpha_3 + \sum_{i=1}^{p} \beta_{3i} y_{t-i}\right) \mathbf{1}\{c_2 < y_{t-d}\} + \varepsilon_t, \end{aligned} \tag{4.6}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$, and c_1 and c_2 are threshold parameters and $c_1 < c_2$.

Example 7 *Trade Cost Model by Sercu, Uppal and van Hulle (1995, Journal of Finance).*

4.2.2 Smooth Transition Autoregressive (STAR) Models

If our aim is to distinguish between the effects of negative and positive deviations (or large and small) from the equilibrium, then TAR models are appropriate. STAR models have attracted more attention. The basic motivation behind it is that prices are expected to adjust more smoothly as is predicted by TAR models. One explanation is: nonlinear asymmetric behavior of heterogeneous market participants will be smoother at the aggregate level.

Granger and Terasvirta (1993) advance the following STAR model:

$$y_t = \left(\alpha_1 + \sum_{i=1}^{p} \beta_{1i} y_{t-i}\right) \{1 - F(z_t; \theta, c)\} + \left(\alpha_2 + \sum_{i=1}^{p} \beta_{2i} y_{t-i}\right) F(z_t; \theta, c) + \varepsilon_t, \tag{4.7}$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. The transition function $F(z_t; \theta, c)$ is a continuous function determining the weights of regime and usually bounded between 0 and 1. c and θ are the threshold and smoothness parameters.

The transition variable z_t can be:

- a lagged endogenous variable ($z_t = y_{t-d}$),
- an exogenous variable ($z_t = x_t$),
- or a function such as ($z_t = g(y_{t-d}, x_t)$).
- For $z_t = t$, we obtain a model with smoothly changing parameters, see Lin and Terasvirta (1994).

The STAR model exhibits:

- two regimes associated with the extreme values of the transition function: $F(z_t; \theta, c) = 1$ and $F(z_t; \theta, c) = 0$;
- transition from one regime to the other is gradual, not abrupt as in TAR;
- the regime occurring at time t is observable and determined by $F(z_t; \theta, c)$.

Logistic Smooth Transition Autoregressive (LSTAR) model

We consider as the transition function in (4.7) the logistic CDF:

$$F(z_t; \theta, c) = \frac{1}{1 + \exp\{-\theta(z_t - c)\}}. \tag{4.8}$$

This model can deal with asymmetric behavior for positive vs negative values of z_t relative to c. We note:

- As $\theta \to \infty$, LSTAR $\to$ TAR, since $F(z_t; \theta, c) = I(z_t > c)$.
- As $\theta \to 0$, LSTAR $\to$ linear AR, since $F(z_t; \theta, c) = 1/2$.

The second order logistic CDF is also considered:

$$F(z_t; \theta, c) = \frac{1}{1 + \exp\{-\theta(z_t - c_1)(z_t - c_2)\}}. \tag{4.9}$$

We note:

- As $\theta \to \infty$, L2STAR $\to$ 3 regime TAR, since $F(z_t; \theta, c) = 1 - I(c_1 < z_t < c_2)$.
- As $\theta \to 0$, L2STAR $\to$ linear AR, since $F(z_t; \theta, c) = 1/2$.

Exponential Smooth Transition Autoregressive (ESTAR) model

We consider as the transition function in (4.7) the exponential function:

$$F(z_t;\theta,c) = 1 - \exp\left\{-\theta(z_t - c)^2\right\}, \tag{4.10}$$

where we assume $\theta \geq 0$ for identification. This model can deal with asymmetric behavior for small vs large deviations of z_t from the threshold c.

The exponential transition function is bounded between zero and 1, *i.e.* $F : \mathbb{R} \to [0,1]$ has the properties:

$$F(0) = 0;\ \lim_{x\to\pm\infty} F(x) = 1,$$

and is symmetrically U-shaped around zero.

As $\theta \to \infty$ and $\theta \to 0$, ESTAR $\to$ linear AR, since $F(z_t;\theta,c) = 1$ and $F(z_t;\theta,c) = 0$, respectively.

Estimation

Nonlinear least squares or ML estimation method via numerical optimisation procedure can be applied, but it also involves the grid search over (d,c) as in TAR models. However, the precise estimation of θ is somewhat difficult in practice.

- For large value of θ, the shape of the logistic function changes only little.
- Accurate estimation of θ requires many observations in the immediate neighborhood of c.
- Insignificance of θ should not be interpreted as evidence against the presence of STAR nonlinearity, see Bates and Watts (1988).

4.2.3 Markov-Switching Autoregressive (MS-AR) Models

Now the regime s_t is generated by a hidden discrete-state homogeneous and ergodic Markov chain:

$$\Pr(s_t|S_{t-1},Y_{t-1}) = \Pr(s_t|S_{t-1};\rho)$$

defined by the transition probabilities,

$$p_{ij} = \Pr(s_{t+1} = j|s_t = j),$$

where $S_{t-1} = \{s_{t-1}, s_{t-2}, ...\}$, $Y_{t-1} = \{y_{t-1}, y_{t-2}, ...y_{t-p}\}$ and ρ are unknown parameters.

The conditional process is a AR(p) model with

- shift in mean (MSM-AR): once-and-for-all jump in time series:

$$y_t - \mu(s_t) = \sum_{i=1}^{p} \beta_i(s_t)(y_{t-i} - \mu(s_{t-i})) + \varepsilon_t, \tag{4.11}$$

- shift in intercept (MSI-AR): smooth adjustment of time series:

$$y_t = \alpha(s_t) + \sum_{i=1}^{p} \beta_i(s_t) y_{t-i} + \varepsilon_t. \tag{4.12}$$

Example 8 *MS-AR models of US GNP. Hamilton (1989) consider the 2-regime MS-AR model for the quarterly growth rate of US GNP:*

$$\Delta y_t - \mu(s_t) = \sum_{i=1}^{4} \beta_i(s_t)(\Delta y_{t-i} - \mu(s_{t-i})) + \varepsilon_t, \tag{4.13}$$

$$\varepsilon_t | s_t \sim iidN\left(0, \sigma^2\right).$$

Two regimes are defined by

$$\mu(s_t) = \left\{ \begin{array}{ll} \mu_1 > 0 & \text{if } s_t = 1 \text{ (expansion)} \\ \mu_2 < 0 & \text{if } s_t = 2 \text{ (contraction)} \end{array} \right\},$$

which are generated by an ergodic Markov chain

$$p_{12} = \Pr(\text{contraction in } t | \text{expansion in } t-1)$$

$$p_{21} = \Pr(\text{expansion in } t | \text{contraction in } t-1)$$

- The statistical analysis of MS-AR models is based on the state-space form. Then, general concepts such as the likelihood principle and a recursive filtering algorithm can be used.
- In contrast to TAR and STAR models MS-AR models include the possibility that the threshold depends on the last regime, i.e., the threshold staying in regime 2 is different from the threshold for switching from regime 1 to regime 2.

4.2.4 Linearity Tests for TAR/STAR Specification

Here the null model is that

$$H_0 : y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t, \tag{4.14}$$

so the model is linear, whilst the alternative models are either

$$H_{1,TAR} : \text{ TAR model given by (4.3)}, \tag{4.15}$$

or

$$H_{1,STAR}: \text{ TAR model given by (4.7).} \tag{4.16}$$

More specifically, against the TAR model we have

$$H_0 : \alpha_1 = \alpha_2 \text{ and } \beta_{1i} = \beta_{2i} \text{ for all } i = 1, ..., p, \tag{4.17}$$

whilst against the STAR model we have

$$H_0 : \theta = 0. \tag{4.18}$$

However, due to the Davies (1987) problem that nuisance parameters in transition function - namely, threshold parameter c in TAR and smoothness parameter θ and threshold parameter in STAR - are not identified under the null, we could not use the standard asymptotic χ^2 distribution.

1. **Sup test approach for TAR models**: We should obtain a supremum of a number of dependent test statistics over the grid over c: $\sup F$, $\sup Wald$, $\sup LR$ and $\sup LM$ tests with nonstandard limiting distribution. To obtain the p-value, we need to run the bootstrapping simulations. See Hansen (1997,2000).

2. **Taylor approximation approach for STAR models**: Approximate smooth transition function with a first-order expansion around $\theta = 0$. Then, using the derived auxiliary regression, we obtain the LM-type tests with a standard χ^2 limiting distribution. See Luukkonen, Saikkonen and Terasvirta (1988) for LSTAR and Saikkonen and Luukkonen (1988) for ESTAR.

4.3 Nonlinear Unit Root Tests in Regime Switching Models

Balke and Fomby (1997) have popularised a joint analysis of nonstationarity and nonlinearity in the context of threshold cointegration. The threshold cointegrating process is defined as a globally stationary process such that it might follow a unit root in the middle regime, but it is dampened in outer regimes. Importantly, they have shown via Monte Carlo experiments that the power of the DF unit root tests falls dramatically with threshold parameters. See also Pippenger and Goering (1993).

As a response, there is a growing literature proposing tests for unit roots against threshold autoregressive (TAR) alternatives, *e.g.* Enders and Granger (1998), Caner and Hansen (2001), Kapetanios, Shin and Snell (2003), Bec, Guay and Guerre (2004) and Kapetanios and Shin (2006).

4.3.1 Unit Root Tests in Two-regime TAR Framework

Enders and Granger (1998) have addressed this issue using a two-regime TAR model with implicitly known threshold value,

$$\Delta y_t = \left\{ \begin{array}{ll} \beta_1 y_{t-1} + u_t & if\ y_{t-1} \leq 0 \\ \beta_2 y_{t-1} + u_t & if\ y_{t-1} > 0 \end{array} \right\},\ t = 1, 2, ..., T, \tag{4.19}$$

and suggested an F-statistic for $\beta_1 = \beta_2 = 0$ in (4.19).

Despite the main aim to derive a more powerful test, their simulation evidence shows that the proposed F test is less powerful than the DF test that ignores the threshold nature of this two regime alternative. But they also provided simulation results showing that the F-test may have higher power than the DF test against the three regime asymmetric TAR models. See also Berben and van Dijk (1999).

There has also been an alternative line of studies. Caner and Hansen (2001) have considered the following two-regime TAR model:

$$\Delta y_t = \boldsymbol{\theta}_1' \mathbf{x}_{t-1} 1_{\{\Delta y_{t-1} \leq r\}} + \boldsymbol{\theta}_2' \mathbf{x}_{t-1} 1_{\{\Delta y_{t-1} > r\}} + e_t,\ t = 1, 2, ..., T, \tag{4.20}$$

where $\mathbf{x}_{t-1} = (y_{t-1}, 1, \Delta y_{t-1}, ..., \Delta y_{t-k})'$, r is an unknown threshold parameter, and e_t is an *iid* error. They have first developed tests for threshold nonlinearity when y_t follows a unit root, and then unit root tests when the threshold nonlinearity is either present or absent. Limitation of this approach is that these tests rely on the stationarity of the transition variable.

4.3.2 Unit Root Tests in Three-regime TAR Framework

Kapetanios and Shin (2006)

Suppose that a univariate series y_t follows the three-regime self-exciting threshold autoregressive (SETAR) model:

$$y_t = \left\{ \begin{array}{ll} \phi_1 y_{t-1} + u_t & if\ y_{t-1} \leq r_1 \\ \phi_0 y_{t-1} + u_t & if\ r_1 < y_{t-1} \leq r_2 \\ \phi_2 y_{t-1} + u_t & if\ y_{t-1} > r_2 \end{array} \right\},\ t = 1, 2, ..., T, \tag{4.21}$$

where u_t is assumed to follow an *iid* sequence with zero mean, constant variance σ_u^2 and finite $4 + \delta$ moments for some $\delta > 0$, r_1 and r_2 are threshold parameters and $r_1 < r_2$. Here, the lagged dependent variable is used as the transition variable with the delay parameter set to 1 for simplicity. The intuitive appeal of the scheme in (4.21) is that it allows the speed of adjustment to vary asymmetrically with regimes. Suppose that

$$\phi_0 \geq 1,\ |\phi_1|,\ |\phi_2| < 1. \tag{4.22}$$

The series are then locally nonstationary, but globally ergodic.

Following the maintained assumption in the literature, we now impose $\phi_0 = 1$ in (4.21), which implies that y_t follows a random walk in the corridor regime. Then, defining $1_{\{.\}}$ as a binary indicator function, (4.21) can be compactly written as

$$\Delta y_t = \beta_1 y_{t-1} 1_{\{y_{t-1} \leq r_1\}} + \beta_2 y_{t-1} 1_{\{y_{t-1} > r_2\}} + u_t, \tag{4.23}$$

where $\beta_1 = \phi_1 - 1$, $\beta_2 = \phi_2 - 1$, and $y_{t-1} 1_{\{y_{t-1} \leq r_1\}}$ and $y_{t-1} 1_{\{y_{t-1} > r_2\}}$ are orthogonal to each other by construction.

We consider the (joint) null hypothesis of unit root as

$$H_0 : \beta_1 = \beta_2 = 0, \tag{4.24}$$

against the alternative hypothesis of threshold stationarity,

$$H_1 : \beta_1 < 0;\ \beta_2 < 0. \tag{4.25}$$

Then, the joint null hypothesis of linear unit root against the nonlinear threshold stationarity can be tested using the Wald statistic denoted by $\mathcal{W}_{(r_1,r_2)}$, which has a nonstandard limiting distribution.

To deal with the Davies problem that threshold parameters r_1 and r_2 are not defined under the null, we consider the supremum, the average and the exponential average of the Wald statistic defined by

$$\mathcal{W}_{\sup} = \sup_{i \in \Gamma} \mathcal{W}^{(i)}_{(r_1,r_2)},\ \mathcal{W}_{\text{avg}} = \frac{1}{\#\Gamma} \sum_{i=1}^{\#\Gamma} \mathcal{W}^{(i)}_{(r_1,r_2)},\ \mathcal{W}_{\exp} = \frac{1}{\#\Gamma} \sum_{i=1}^{\#\Gamma} exp\left(\frac{\mathcal{W}^{(i)}_{(r_1,r_2)}}{2} \right), \tag{4.26}$$

where $\mathcal{W}^{(i)}_{(r_1,r_2)}$ is the Wald statistic obtained from the i-th point of the threshold parameters grid set, Γ and $\#\Gamma$ is the number of elements of Γ.

Unlike the stationary TAR models, the selection of the grid of threshold parameters needs more attention. The threshold parameters r_1 and r_2 usually take on the values in the interval

$$(r_1, r_2) \in \Gamma = \{(r_{1,1}, r_{1,2}), ..., (r_{i,1}, r_{i,2}), ..., (r_{\#\Gamma,1}, r_{\#\Gamma,2})\},$$

where $r_{\min} \leq r_{i,1}$, $i = 1, ...\#\Gamma$, and $r_{\max} \geq r_{i,2}$, $i = 1, ...\#\Gamma$. $r_{\min}$ and $r_{\max}$ are picked so that $\Pr(y_{t-1} < r_{\min}) = \pi_1 > 0$ and $\Pr(y_{t-1} > r_{\max}) = \pi_2 < 1$. The particular choice for π_1 and π_2 is somewhat arbitrary, and in practice must be guided by the consideration that each regime needs to have sufficient observations to identify the underlying regression parameters.

However, since our approach assumes that the coefficient on the lagged dependent variable is set to zero in the corridor regime ($r_1 \leq y_{t-1} < r_2$), we can assign arbitrarily small samples (relative to total sample) to the corridor regime. Notice also that the threshold parameters exist only under the alternative hypothesis in which the process is stationary and therefore

bounded in probability. This observation leads us to make an assumption that the grid for unknown threshold parameters should be selected such that the selected corridor regime be of finite width both under the null and under the alternative. Noticing that a random walk process will stay within a corridor regime of finite width for $O_p\left(\sqrt{T}\right)$ periods only, then setting

$$\pi_1 = \bar{\pi} - c/T^{\delta} \text{ and } \pi_2 = \bar{\pi} + c/T^{\delta},$$

where $\bar{\pi}$ is the sample quantile corresponding to zero and $\delta \geq 1/2$, guarantees that the grid set will be of finite width under the null hypothesis. In practice, c can be chosen so as to give a reasonable coverage of each regime in samples of sizes usually encountered. For example, for $T = 100$ and $\delta = 1/2$, c can be set to 3 to give a 60% coverage of the sample for the grid.

The small sample performance of our suggested tests is compared to that of the DF test via Monte Carlo experiments. We find that both average and exponential average tests have reasonably correct size, but the supremum test tends to display significant size distortions in small samples. As expected, both average and exponential average tests eventually dominate the power of the DF test as the threshold band widens.

KS illustrate the usefulness of our proposed tests by examining the stationarity of bilateral real exchange rates for the G7 countries (excluding France). In sum, our proposed (asymmetry) Wald tests reject the null three times out of five cases, while the DF test rejects the null only once.

Bec, Ben Salem and Carrasco (2004) and Bec, Guay and Gurre (2004)

A three-regime SETAR model (4.21) can be compactly written as

$$\Delta y_t = \beta_1 y_{t-1} 1_{\{y_{t-1} \leq r_1\}} + \beta_0 y_{t-1} 1_{\{r_1 < y_{t-1} < r_2\}} + \beta_2 y_{t-1} 1_{\{y_{t-1} > r_2\}} + u_t, \quad (4.27)$$

where $1_{\{.\}}$ is a binary indicator function, $\beta_1 = \phi_1 - 1$, $\beta_0 = \phi_0 - 1$, $\beta_2 = \phi_2 - 1$, and $y_{t-1}1_{\{y_{t-1} \leq r_1\}}$, $y_{t-1}1_{\{r_1 < y_{t-1} < r_2\}}$, $y_{t-1}1_{\{y_{t-1} > r_2\}}$ are orthogonal to each other by construction. BBC and BGG have considered the three-regime SETAR model (4.27), and proposed the supremum-based Wald test procedure for the joint hypothesis of $\beta_1 = \beta_0 = \beta_2 = 0$ in (4.27).

BBC take the quantile-based approach, assuming that $r = \sqrt{T}\lambda$ where $|r_1| = |r_2| = r$ (symmetric outer regimes).[1] Then they derive the asymptotic distribution of the Wald statistic, denoted by $\mathcal{W}^{BBC}(r)$, for $\beta_1 = \beta_0 = 0$ in (4.27) after imposing $\beta_1 = \beta_2$, which depends on the nuisance parameter, $\lambda/\hat{\sigma}_{LR}$, where $\hat{\sigma}_{LR}$ is the long-run variance of Δy_t obtained under the null. To avoid the Davies problem, they suggest to use the supremum-based tests:

$$\mathcal{W}^{BBC}_{\sup} = \sup_{r \in [r_{\min}, r_{\max}]} \mathcal{W}^{BBC}_i(r). \quad (4.28)$$

[1] The assumption that $r = \sqrt{T}\delta$ guarantees that the probability being in the corridor regime is always positive. On the other if r is fixed, this probability becomes zero.

On the other hand, BGG develop an adaptive consistent unit root tests based on the symmetric three regime TAR model (4.27) with $\beta_1 = \beta_2$ and propose an adaptive choice of the grid set which restricts the grid to remain bounded under the null but to become unbounded under the alternative. They suggest to use the following grid set:

$$r_{\min} = |y|_{(3)} + \frac{\hat{\sigma}_0}{\ell \max(1, t_{ADF})};\ r_{\max} = |y|_{(3)} + \ell \hat{\sigma}_0 \max(1, t_{ADF}), \quad (4.29)$$

where $|y|_{(j)}$'s are the ordered variables of $|y_j|$, $j = 1, ..., T-1$, ℓ is a length parameter to be determined empirically, t_{ADF} is the ADF t-statistic, and $\hat{\sigma}_0^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(y_t - \hat{a} - \hat{\phi} y_{t-1}\right)^2$ with $\hat{a}$ and $\hat{\phi}$ being the OLS estimates. This adaptive choice of the grid set is aimed to boost the power of the tests. First, if the set $\Gamma = [r_{\min}, r_{\max}]$ is small under the null, then the associated critical values of the $\mathcal{W}_{\sup}^{BBC}$ test statistic will be small too. Second, it will make a larger class of alternatives including the linear stationary model. Using the simulation evidence BGG argue that the former provides the most important contribution of the improved power performance of the $\mathcal{W}_{\sup}^{BBC}$ test when the grid set is selected by (4.29).

4.3.3 Unit Root Tests in ESTAR Framework (Kapetanios, Shin and Snell, 2003)

Consider a univariate smooth transition autoregressive of order 1, ESTAR(1) model,

$$y_t = \beta y_{t-1} + \gamma y_{t-1}\left[1 - \exp\left(-\theta y_{t-d}^2\right)\right] + \varepsilon_t, \quad (4.30)$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$, β and γ are unknown parameters, and we assume that $\theta \geq 0$, and $d \geq 1$ is the delay parameter. (4.30) can be conveniently reparameterised as:

$$\Delta y_t = \phi y_{t-1} + \gamma y_{t-1}\left[1 - \exp\left(-\theta y_{t-d}^2\right)\right] + \varepsilon_t, \quad (4.31)$$

where $\phi = \beta - 1$. If θ is positive, then it determines the speed of mean reversion. The representation (4.31) makes economic sense in that many economic models predict that the underlying system tends to display a dampened behavior towards an attractor when it is (sufficiently far) away from it, but that it shows some instability within the locality of that attractor.

We prove under $\theta > 0$ that the condition we need for geometric ergodicity of the model (4.30) or (4.31) is in fact $|\beta + \gamma| < 1$ or $|\phi + \gamma| < 0$.

Remark 1 *The application that motivates our model is that of Sercu et al. (1995) and of Michael et al. (1997). These authors analyse nonlinearities in the PPP relationship. They adopt a null of a unit root for real exchange rates and have an alternative hypothesis of stationarity i.e. the long run PPP. Their theory suggests that the larger the deviation from PPP, the stronger the*

tendency to move back to equilibrium. In the context of our model, this would imply that while $\phi \geq 0$ *is possible, we must have* $\gamma < 0$ *and* $\phi + \gamma < 0$ *for the process to be globally stationary. Under these conditions, the process might display unit root or explosive behaviour in the middle regime for small* y_{t-d}^2*, but for large* y_{t-d}^2*, it has stable dynamics and is geometrically ergodic. They claim that the ADF test may lack power against such stationary alternatives and one of our contributions is to provide an alternative test designed to have a power against such an ESTAR processes.*

Imposing $\phi = 0$ and $d = 1$ gives our specific ESTAR model (4.31) as

$$\Delta y_t = \gamma y_{t-1} \left\{1 - \exp\left(-\theta y_{t-1}^2\right)\right\} + \varepsilon_t. \tag{4.32}$$

Our test directly focuses on a specific parameter, θ, which is zero under the null and positive under the alternative. Hence we test

$$H_0 : \theta = 0, \tag{4.33}$$

against the alternative

$$H_1 : \theta > 0. \tag{4.34}$$

Obviously, testing the null hypothesis (4.45) directly is not feasible, since γ is not identified under the null.

To overcome this problem we follow Luukkonen *et al.* (1988), and derive a t-type test statistic. If we compute a first-order Talyor series approximation to the $ESTAR$ model under the null we get the auxiliary regression

$$\Delta y_t = \delta y_{t-1}^3 + error. \tag{4.35}$$

This suggests that we could obtain the t-statistic for $\delta = 0$ against $\delta < 0$ as

$$t_{NL} = \widehat{\delta} / s.e.\left(\widehat{\delta}\right), \tag{4.36}$$

where $\widehat{\delta}$ is the OLS estimate of δ and $s.e.\left(\widehat{\delta}\right)$ is the standard error of $\widehat{\delta}$. Our test is motivated by the fact that the auxiliary regression is testing the significance of the score vector from the quasi-likelihood function of the $ESTAR$ model, evaluated at $\theta = 0$.

Unlike the case of testing linearity against nonlinearity for the stationary process, the t_{NL} test does not have an asymptotic standard normal distribution. KSS find *inter alia* that under the alternative of a globally stationary ESTAR process, our test has better power in cases where the nonlinear adjustment is relatively important.

KSS also provide an application to *ex post* real interest rates and bilateral real exchange rates from eleven major OECD countries, and in particular find that our proposed test is able to reject a unit root in some cases where the linear ADF tests fails to do so, providing a limited evidence of of nonlinear mean-reversion in both real interest and exchange rates.

4.4 Nonlinear Error Correction Models

Clearly, many stylised facts can be evoked to account for the asymmetric adjusting behavior. In financial markets prices are constrained to persistent short-run disequilibria due to information barriers, transaction costs, noise trading, market segmentation, etc.

- A first strand of the literature is based on a generalisation of the usual concept of cointegration. Notions such as 'attractors', 'transients', 'Lyapunov stability', 'equilibration' have been introduced in an attempt to capture richer dynamics than is allowed by linear cointegration models.
- Another approach aims to clarify the concept of cointegration. If the two processes have the same order of integration, they may be cointegrated if their combination (either linear or nonlinear) is mixing. But, one must impose the bound conditions on the nonlinear functions.
- A third part of the literature is centred on nonlinear co-trending. Nonlinear trends are modelled as general polynomial functions that allows multiple representations of nonlinear trends. Co-trending means that the combination of nonlinear trends provides linear trends.

We assume that the attractor is linear but that the adjustment towards the long-run equilibrium is nonlinear. The NEC model is written as

$$\Delta y_t = \sum_{i=1}^{p} \delta_i' \Delta y_{t-i} + \sum_{i=1}^{q} \gamma_i' \Delta x_{t-i} + \lambda z_{t-1} + f\left(z_{t-1}, \theta\right) + u_t, \tag{4.37}$$

$$\Delta y_t = v_t,$$

$$z_t = y_t - \beta' x_t.$$

Assume that (i) u_t and v_t are mixing processes with finite second-order moments and cross moments; (ii) f is a nonlinear function that is continuously differentiable and satisfies some regularity condition:

$$-1 < \frac{\partial f\left(z_{t-1}, \theta\right)}{\partial z_{t-1}} < 1;$$

(iii) the roots of $\left|1 - \sum_{i=1}^{p} \delta_i L^i\right| = 0$ all lie outside the unit circle; and (iv) u_t is a martingale difference sequence with zero mean and constant variance.

Under this assumption Escribano and Mira (2002) prove that z_t is NED and y_t and x_t are cointegrated. The cointegration hypothesis is tested as:

$$H_0 : f\left(z_{t-1}, \theta\right) = 0 \text{ against } H_1 : f\left(z_{t-1}, \theta\right) \neq 0.$$

H_0 means that the adjustment mechanism is linear. Under H_1: it is not sufficient that $f\left(z_{t-1}, \theta\right) \neq 0$, but this function must characterize an EC mechanism (hence the importance of the stability condition of f).

Estimation of (4.37) can be done in 4-steps:

1. Obtain the OLS estimate of $\boldsymbol{\beta}$ from the regression of y_t on $\mathbf{x}_t$. Construct the estimate of error correction term by $\hat{z}_t = y_t - \hat{\boldsymbol{\beta}}' \mathbf{x}_t$.
2. Substitute $\hat{z}_{t-1}$ for z_{t-1} in $f(z_{t-1}, \theta)$.
3. Use the NLS method to find an estimate of $\boldsymbol{\theta}$.
4. Estimate other coefficients of the model (4.37) by OLS.

In practice the great difficulty lies in finding an appropriate function that satisfies the stability condition defined in Assumption (ii). The following functional forms are employed in Dufrénot and Mignon (2002):

- Logistic Smooth Transition Regression:

$$LF(z_{t-1}) = [1 + \exp(-\theta(z_{t-1} - c))]$$

- Cubic Polynomial:

$$LF(z_{t-1}) = \delta_1 z_{t-1} + \delta_2 z_{t-1}^2 + \delta_3^3 z_{t-1}$$

- Rational Polynomial:

$$LF(z_{t-1}) = \frac{(z_{t-1} + \gamma_1)^3 + \gamma_2}{(z_{t-1} + \gamma_3)^3 + \gamma_4}.$$

Example 9 *Rational or irrational bubbles? Many empirical studies find that there is no cointegration between stock prices and dividends. This may imply that the fluctuations in asset prices are too large to reflect changes occurring in the fundamentals (here dividends). This excess volatility can be regarded as a consequence of the presence of a (possibly) nonstationary bubble. This paves the way for a nonlinear dynamic analysis as to how to add to the usual arbitrage equation a nonlinear component reflecting the complexity of the short term dynamics between both variables.*

4.4.1 Asymmetric TAR NEC Models

See Balke and Fomby (1997).

4.4.2 Asymmetric STR NEC Models

We begin with the following general nonlinear vector error correction model for the $n \times 1$ vector of I(1) stochastic processes, $\mathbf{z}_t$:

$$\Delta \mathbf{z}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{z}_{t-1} + g\left(\boldsymbol{\beta}' \mathbf{z}_{t-1}\right) + \sum_{i=1}^{p} \boldsymbol{\Gamma}_i \Delta \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_t, \ t = 1, 2, ..., T, \qquad (4.38)$$

where $\boldsymbol{\alpha}$ $(n \times r)$, $\boldsymbol{\beta}$ $(n \times r)$ and $\boldsymbol{\Gamma}_i$ $(n \times n)$ are parameter matrices with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of full column rank and $g : \mathbb{R}^r \to \mathbb{R}^n$ is a nonlinear function. See Saikkonen (2004).

We aim to analyse at most one conditional long-run cointegrating relationship between y_t and $\mathbf{x}_t$, and focus on the conditional modelling of the scalar variable y_t given the k-vector $\mathbf{x}_t$ $(k = n-1)$ and the past values of $\mathbf{z}_t$ and $\mathbf{Z}_0$, where we decompose $\mathbf{z}_t = (y_t, \mathbf{x}_t')'$. For this we rewrite (4.38) as

$$\Delta \mathbf{z}_t = \boldsymbol{\alpha} u_{t-1} + g(u_{t-1}) + \sum_{i=1}^{p} \boldsymbol{\Gamma}_i \Delta \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_t, \ t = 1, 2, ..., T, \tag{4.39}$$

where $\boldsymbol{\alpha}$ is an $n \times 1$ vector of adjustment parameters, and

$$u_t = y_t - \boldsymbol{\beta}_x' \mathbf{x}_t, \tag{4.40}$$

with $\boldsymbol{\beta}_x$ being a $k \times 1$ vector of cointegrating parameters.

We now make the following assumption:

- 2**(i)** Partition $\boldsymbol{\alpha} = (\phi, \boldsymbol{\alpha}_x')'$ and $\boldsymbol{\varphi} = (\gamma, \boldsymbol{\varphi}_x')'$ conformably with $\mathbf{z}_t = (y_t, \mathbf{x}_t')'$. Then, $\boldsymbol{\alpha}_x = \boldsymbol{\varphi}_x = \mathbf{0}$.

- 2**(ii)** There is no cointegration among the k-vector of I(1) variables, $\mathbf{x}_t$.

- 2(**iii**) $g(\bullet)$ follows the exponential smooth transition regressive (ESTR) functional form,[1]

$$g(u_{t-1}) = \boldsymbol{\varphi} u_{t-1} \left(1 - e^{-\theta (u_{t-1} - c)^2}\right), \tag{4.41}$$

  where we assume $\theta \geq 0$ for identification purpose and c is a transition parameter.

Assumption 2(i) and (ii) imply that the process $\mathbf{x}_t$ are weakly exogenous and therefore the parameters of interest in (4.43) are variation-free from the parameters in (4.44), see Pesaran *et al.* (2001).

Next, partitioning $\boldsymbol{\varepsilon}_t$ conformably with $\mathbf{z}_t$ as $\boldsymbol{\varepsilon}_t = (\varepsilon_{yt}, \boldsymbol{\varepsilon}_{xt}')'$ and its variance matrix as $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}_{yx} \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}$, we may express ε_{yt} conditionally in terms of $\boldsymbol{\varepsilon}_{xt}$ as

$$\varepsilon_{yt} = \boldsymbol{\sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\varepsilon}_{xt} + e_t, \tag{4.42}$$

where $e_t \sim iid(0, \sigma_e^2)$, $\sigma_e^2 \equiv \sigma_{yy} - \boldsymbol{\sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$ and e_t is uncorrelated with $\boldsymbol{\varepsilon}_{xt}$ by construction. Substituting (4.42) and (4.41) into (4.39), partitioning $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{yi}', \boldsymbol{\Gamma}_{xi}')'$, $i = 1, ..., p$, and under Assumption 2, we obtain the following

conditional nonlinear error correction model for Δy_t and the marginal VAR model for $\Delta \mathbf{x}_t$:

$$\Delta y_t = \phi u_{t-1} + \gamma u_{t-1}\left(1 - e^{-\theta(u_{t-1}-c)^2}\right) + \boldsymbol{\omega}'\Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i'\Delta \mathbf{z}_{t-i} + e_t, \quad (4.43)$$

$$\Delta \mathbf{x}_t = \sum_{i=1}^{p} \boldsymbol{\Gamma}_{xi}\Delta \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_{xt}, \quad (4.44)$$

where $\boldsymbol{\omega} \equiv \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy}$ and $\boldsymbol{\psi}_i' \equiv \boldsymbol{\gamma}_{yi} - \boldsymbol{\omega}'\boldsymbol{\Gamma}_{xi}$, $i = 1, ..., p$.

We call (4.43) the (conditional) nonlinear STR error correction model. The representation (4.43) makes economic sense in that many economic models predict that the underlying system tends to display a dampened behavior towards an attractor when it is (sufficiently far) away from it, but shows some instability within the locality of that attractor.

Testing for Cointegration under STR ECM

To fix ideas for the motivation of the tests, we follow Kapetanios, Shin and Snell (2003, hereafter KSS) and impose $\phi = 0$ in (4.43), implying that u_t follows a unit root process in the middle regime, see also Balke and Fomby (1997) in the context of threshold error correction models. Note that for the operational versions of the tests we suggest below we consider both the case $\phi = 0$ and $\phi \neq 0$. It is then straightforward to show that the test of the null of no cointegration against the alternative of globally stationary cointegration can be based on the null hypothesis of no cointegration as

$$H_0 : \theta = 0, \quad (4.45)$$

against the alternative of nonlinear ESTR cointegration of $H_1 : \theta > 0$, where the positive value of θ determines the stationarity properties of u_t.

We propose a number of operational versions of the cointegration test under the nonlinear STR-ECM framework given by (4.43). To this end we follow Engle and Granger (1987) and take a pragmatic residual-based two step approach. In the first stage, we obtain the residuals, $\hat{u}_t = y_t - \hat{\boldsymbol{\beta}}_x'\mathbf{x}_t$ with $\hat{\boldsymbol{\beta}}_x$ being the OLS estimate of $\boldsymbol{\beta}_x$. In order to overcome the Davies problem that γ in (4.43) is not identified under the null, we follow Luukkonen, Saikkonen and Teräsvirta (1988) and KSS and approximate (4.43) by a first-order Taylor series approximation to $\left(1 - e^{-\theta(u_{t-1}-c)^2}\right)$, while allowing $\phi \neq 0$ under the alternative hypothesis, to get

$$\Delta y_t = \delta_1 u_{t-1} + \delta_2 u_{t-1}^2 + \delta_3 u_{t-1}^3 + \boldsymbol{\omega}'\Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i'\Delta \mathbf{z}_{t-i} + e_t. \quad (4.46)$$

For this model, we consider an F-type test for $\delta_1 = \delta_2 = \delta_3 = 0$ given by

$$F_{NEC} = \frac{(SSR_0 - SSR_1)/3}{SSR_0/(T-4-p)}, \tag{4.47}$$

where SSR_0 and SSR_1 are the sum of squared residuals obtained from the specification with and without imposing the restrictions $\delta_1 = \delta_2 = \delta_3 = 0$ in (4.46), respectively.

There are prior theoretical justifications for restricting the switch point, c to be zero in many economic and financial applications in the ESTR function (4.41), in which case we obtain the following restricted auxiliary testing regression:

$$\Delta y_t = \delta_1 \hat{u}_{t-1} + \delta_2 \hat{u}_{t-1}^3 + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t, \tag{4.48}$$

and obtain the following F-type statistic:

$$F_{NEC}^* = \frac{(SSR_0 - SSR_1)/2}{SSR_0/(T-3-p)}, \tag{4.49}$$

where SSR_0 and SSR_1 are the sum of squared residuals obtained from the specification with and without imposing $\delta_1 = \delta_2 = 0$ in (4.48), respectively.

Finally, under the further assumption that $\phi = 0$ (which is the maintained assumption made in KSS), (4.48) is simplified to

$$\Delta y_t = \delta u_{t-1}^3 + \boldsymbol{\omega}' \Delta \mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\psi}_i' \Delta \mathbf{z}_{t-i} + e_t. \tag{4.50}$$

For this model, we propose a t-type statistic for $\delta = 0$ (no cointegration) against $\delta < 0$ (ESTR cointegration), denoted by t_{NEC}.

The asymptotic distributions of all these tests are nonstandard, and the associated critical values have been tabulated via stochastic simulations.

The small sample performance of the suggested tests is compared to that of the linear EG and Johansen (1995) tests via Monte Carlo experiments. We find that our proposed nonlinear tests have good size and superior power properties compared to the linear tests. In particular, both F_{NEC} and t_{NEC} tests are superior to both linear or nonlinear EG tests when the regressors are weakly exogenous in a cointegrating regression. This supports similar findings made in linear models that the EG test loses power relative to ECM-based cointegration tests because of the loss of potentially valuable information from the correlation between the regressors and the underlying disturbances.

KSS provide an application to investigating the presence of cointegration of asset prices and dividends for eleven stock portfolios allowing for

nonlinear STR adjustment to equilibrium. Interestingly, our new tests are able to reject the null of no cointegration in majority cases whereas the linear EG test rejects only twice. We also estimate adjustment parameters under the alternative, and find that these estimates are well defined in all cases. We further evaluate the impulse response functions of the error correction term with respect to initial impulses of 1-4 standard deviation shocks. The striking finding is that the time taken to recover one half of a one standard deviation shock varies between five and twenty years, whereas the time taken to recover one half of larger shocks varies between just 4 to 18 months. This implies that data periods dominated by extreme volatility may display substantial reversion of prices towards their NPV relationship, while in "calmer" times where the error in the NPV relationship takes on smaller values, the process driving it may well look like a unit root.

4.4.3 MS NEC Models

Psaradakis, Sola and Spagnolo (2004) consider the following single equation-based MS NEC model:

$$y_t + \alpha x_t = z_t,\ z_t = \phi_{s_t} z_{t-1} + \varepsilon_{1t}, \tag{4.51}$$

$$y_t + \beta x_t = u_t,\ u_t = u_{t-1} + \varepsilon_{2t}, \tag{4.52}$$

where $\alpha \neq 0$, $\beta \in R$, $\phi_{s_t} \in (-1, 1]$ and s_t are the latent random variables on $\{0, 1\}$. Suppose that

$$\phi_{s_t} = \phi_0 + (\phi_1 - \phi_0)\, s_t,\ |\phi_0| < 1,\ \phi_1 = 1, \tag{4.53}$$

where $\{s_t\}$ is a homogeneous irreducible and aperiodic Markov chain of order 1with state-space, $S = \{0, 1\}$ and transition probabilities

$$p_{ij} = \Pr\{s_t = j | s_{t-1} = i\},\ i, j \in S. \tag{4.54}$$

Deviations from equilibrium tend to decay to the mean level of 0 as long as $s_t = 0$; otherwise z_t behaves like a nonstationary process. Despite the occasional nonstaitoanry behavior of $\{z_t\}$ when $s_t = 1$, the eq error can be globally stationary, provided that p_{00}, p_{11}, ϕ_0 and ϕ_1 satisfy appropriate restrictions. A necessary and sufficient condition is given by (Franq and Zakoian, 2001)

$$p_{00}\phi_0^2 + p_{11}\phi_1^2 + (1 - p_{00} - p_{11})\,\phi_0^2\phi_1^2 < 1 \text{ and } p_{00}\phi_0^2 + p_{11}\phi_1^2 < 2. \tag{4.55}$$

For an irreducible and aperiodic Markov chain, these conditions are easily satisfied when $|\phi_0| < 1$ and $\phi_1 = 1$. See an application to the relationship between stock prices and dividends in Psaradakis, Sola and Spagnolo (2004).

We could also allow for $\{z_t\}$ to evolve according to the Markov switching ARMA model,

$$z_t = c_{s_t} + \sum_{i=1}^{m} \phi_{s_t}^{(i)} z_{t-i} + \sigma_{s_t}\xi_t + \sum_{j=1}^{q} \psi_{s_t}^{(i)} \sigma_{s_{t-j}} \xi_{t-j}, \tag{4.56}$$

where ξ_t is a white noise with $E\xi_t = 0$ and $E\xi_t^2 = 0$. A sufficient condition for the 2nd order stationarity is that all the eigenvalues of the $2m^2 \times 2m^2$ matrix,

$$\Lambda = \begin{bmatrix} p_{00}\left(\mathbf{\Phi}_0 \otimes \mathbf{\Phi}_0\right) & p_{10}\left(\mathbf{\Phi}_0 \otimes \mathbf{\Phi}_0\right) \\ p_{01}\left(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right) & p_{11}\left(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right) \end{bmatrix}$$

lie on the open disk where

$$\mathbf{\Phi}_h = \begin{bmatrix} \phi_h^{(1)} & \phi_h^{(2)} & \cdots & \phi_h^{(m-1)} & \phi_h^{(m)} \\ 1 & 0 & & 0 & 0 \\ 1 & 0 & & 0 & 0 \\ & & & & \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \; h \in S.$$

Another useful extension is that it is reasonable to expect that the further away from the equilibrium of the system is the higher the probability of switching from an unstable noncorrecting regime to a stable error correcting one. This allows the transition probabilities of the hidden Markov chain to depend on the extent to which the system is out of long-run equilibrium. Therefore,

$$\Pr\left\{s_t = i | s_{t-1} = i, z_{t-1}\right\} = \frac{\exp\left(a_i + b_i z_{t-1}\right)}{1 + \exp\left(a_i + b_i z_{t-1}\right)}, \; i \in S, \tag{4.57}$$

$$\Pr\left\{s_t = j | s_{t-1} = i, z_{t-1}\right\} = 1 - \Pr\left\{s_t = i | s_{t-1} = i, z_{t-1}\right\}, \; i \in S, \; i \neq j \tag{4.58}$$

It is natural to consider testing the null of single-regime/no-coinetgration against the alternative of cointegration with MEC adjustment. The testing problem is nonstandard due to the presence of unit roots and the unidentifiability of the transition probabilities under the null.

Testing for Cointegration under STR ECM

See Hu and Shin (2014).

4.5 Panel Threshold Regression Models

This is a summary of the paper by B. Hansen (1999, JOE, 93: 345-368)

4.5.1 Model

The structural equation is

$$y_{it} = \mu_i + \beta_1' x_{it} I\left(q_{it} \leq \gamma\right) + \beta_2' x_{it} I\left(q_{it} > \gamma\right) + e_{it}, \tag{4.59}$$

which can be written as

$$y_{it} = \mu_i + \beta' x_{it}\left(\gamma\right) + e_{it}, \tag{4.60}$$

where $x_{it}\left(\gamma\right) = \begin{pmatrix} x_{it} I\left(q_{it} \leq \gamma\right) \\ x_{it} I\left(q_{it} > \gamma\right) \end{pmatrix}$ and $\beta = \left(\beta_1', \beta_2'\right)'$. For identification it is required that x_{it} are not time invariant. e_{it} is assumed to be iid, which excludes lagged dependent variables from x_{it}. The analysis is asymptotic with fixed T as $n \to \infty$.

4.5.2 Estimation

Taking averages of (4.59),

$$\bar{y}_i = \mu_i + \beta' \bar{x}_i\left(\gamma\right) + \bar{e}_i, \tag{4.61}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{i=1}^{T} y_{it};\ \bar{x}_i\left(\gamma\right) = \frac{1}{T}\sum_{i=1}^{T} x_{it}\left(\gamma\right) = \frac{1}{T}\sum_{i=1}^{T} \begin{pmatrix} x_{it} I\left(q_{it} \leq \gamma\right) \\ x_{it} I\left(q_{it} > \gamma\right) \end{pmatrix};$$

and taking the difference between (4.60) and (4.61),

$$y_{it}^* = \beta' x_{it}^*\left(\gamma\right) + e_{it}^*, \tag{4.62}$$

where

$$y_{it}^* = y_{it} - \bar{y}_i;\ x_{it}^*\left(\gamma\right) = x_{it}\left(\gamma\right) - \bar{x}_i\left(\gamma\right).$$

Let

$$y_i^* = \begin{bmatrix} y_{i2}^* \\ \vdots \\ y_{iT}^* \end{bmatrix};\ x_i^*\left(\gamma\right) = \begin{bmatrix} x_{i2}^{*\prime}\left(\gamma\right) \\ \vdots \\ x_{iT}^{*\prime}\left(\gamma\right) \end{bmatrix}$$

denote the stacked data with one time period deleted. Then, let

$$Y^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix};\ X^*\left(\gamma\right) = \begin{bmatrix} x_1^*\left(\gamma\right) \\ \vdots \\ x_n^*\left(\gamma\right) \end{bmatrix}$$

denote the data stacked over all individuals. Then,

$$Y^* = X^*\left(\gamma\right)\beta + e^*. \tag{4.63}$$

For given γ, β can be estimated by OLS;

$$\hat{\beta}(\gamma) = \left(X^*(\gamma)' X^*(\gamma)\right)^{-1} X^*(\gamma)' Y^*. \tag{4.64}$$

Chan (1993) and Hansen (1999) recommend estimation of γ by LS:

$$\hat{\gamma} = \arg\min_{\gamma} S_1(\gamma), \tag{4.65}$$

where

$$S_1(\gamma) = \hat{e}^*(\gamma)' \hat{e}^*(\gamma) = Y^{*\prime} \left(I - X^*(\gamma)\left(X^*(\gamma)' X^*(\gamma)\right)^{-1} X^*(\gamma)'\right) Y^* \tag{4.66}$$
$$\hat{e}^*(\gamma) = Y^* - X^*(\gamma)\hat{\beta}(\gamma).$$

Once $\hat{\gamma}$ is obtained,

$$\hat{\beta} = \hat{\beta}(\hat{\gamma});\ \hat{e}^* = \hat{e}^*(\hat{\gamma});$$
$$\hat{\sigma}^2 = \frac{1}{n(T-1)} \hat{e}^{*\prime} \hat{e}^* = \frac{1}{n(T-1)} S_1(\hat{\gamma}). \tag{4.67}$$

Since $S_1(\gamma)$ depends only on γ through the indicator function, the sum of SSE is a step function with most nT steps with the steps occurring at distinct values of the observed threshold variable q_{it}. Thus the minimisation problem can be reduced to searching over the values of γ equalling the (at most nT) distinct values of q_{it} in the sample.

Sort the distinct values of the observations on q_{it}. Eliminate the smallest and largest η%. The remaining N values constitute the values of γ which can be searched for $\hat{\gamma}$. For each of N values regression are estimated yielding the SSE. The smallest value yields the estimate $\hat{\gamma}$. A simplifying shortcut is to restrict search to a smallest set of values of γ. The search may be limited to specific quintiles. This reduces the number of regressions performed in the search. The estimation from such an approximation are likely to be sufficiently precise. For the empirical work we used the grid {1%, 1.25%, 1.5%, 1.75%, 2%,...,99%} which contains 393 quantiles.

4.5.3 Inference

The hypothesis of no threshold is:

$$H_0 : \beta_1 = \beta_2.$$

The FE (4.62) fall in the class of models considered by Hansen (1996) who suggested a bootstrap to simulate the asymptotic distribution of the LR test. Under the null of no threshold, the model is

$$y_{it} = \mu_i + \beta_1' x_{it} + e_{it}, \tag{4.68}$$

after the FE transformation, we have

$$y_{it}^{*} = \beta_1' x_{it}^{*} + e_{it}^{*}, \tag{4.69}$$

from which we obtain: $\tilde{\beta}_1$, $\tilde{e}_{it}^{*}$ and $S_0 = \tilde{e}^{*\prime}\tilde{e}^{*}$. The LR test is based on

$$F_1 = \frac{S_0 - S_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{4.70}$$

Hansen (1996) shows that a bootstrap procedure attains the first-order asymptotic distribution, so p-values are asymptotically valid.

Treat x_{it} and q_{it} as given. Take the residuals, $\hat{e}_{it}^{*}$ and group them by individual: $\hat{e}_i^{*} = (\hat{e}_{i1}^{*}, ..., \hat{e}_{iT}^{*})$. Treat $(\hat{e}_i^{*}, ..., \hat{e}_n^{*})$ as the empirical distribution to be used for bootstrapping. Draw (with replacement) a sample of size n from the empirical distribution and use these errors to create a bootstrap sample under H_0. Using the bootstrap sample estimate the model under the null and the alternative and calculate the bootstrap value of the LR test F_1. Repeat this procedure a large number of times and calculate the percentage of draws for which the simulated statistic exceeds the actual. This is the bootstrap estimate of the asymptotic p-value for F_1 under H_0.

Hansen (1999) argues that the best way to form CI for γ is to form the no-rejection region using the LR test. To test $H_0 : \gamma = \gamma_0$, we have

$$LR_1(\gamma) = \frac{S_1(\gamma) - S_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{4.71}$$

Note that the statistic (4.71) is testing a different hypothesis from (4.70).

Theorem 1. Under Assumptions 1-8 and $H_0 : \gamma = \gamma_0$

$$LR_1(\gamma) \to_d \xi,$$

as $n \to \infty$, where ξ is a random variable with distribution function,

$$P(\xi \leq x) = \left\{1 - \exp\left(-\frac{x}{2}\right)\right\}^2. \tag{4.72}$$

Since the asymptotic distribution in Theorem 1 is pivotal, it may be used to form valid asymptotic CIs. The distribution function (4.72) has the inverse:

$$c(\alpha) = -2\log\left(1 - \sqrt{1-\alpha}\right) \tag{4.73}$$

from which it is easy to calculate critical values.

To form an asymptotic CI for γ the non-rejection region of CI level $1-\alpha$ is the set of values of γ such that $LR_1(\gamma) \leq c(\alpha)$. This is a natural by-product of model estimation. To find LSE of γ the sequence of $S_1(\gamma)$ were calculated. $LR_1(\gamma)$ is a simple renormalization of these numbers and require no further computation.

Chan and Hansen show that the dependence on the threshold estimate is not of first-order asymptotic importance, so inference on β can proceed as if $\hat{\gamma}$ were true value. Hence,

$$\hat{\beta} \overset{a}{\sim} N(\beta, V),$$

where

$$\hat{V} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*}(\hat{\gamma}) x_{it}^{*}(\hat{\gamma})' \right)^{-1} \hat{\sigma}^2.$$

If the errors are allowed to be conditional heteroskedastic, then

$$\hat{V} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*}(\hat{\gamma}) x_{it}^{*}(\hat{\gamma})' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*}(\hat{\gamma}) x_{it}^{*}(\hat{\gamma})' \hat{e}_{it}^{*\,2} \right) \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{*}(\hat{\gamma}) x_{it}^{*}(\hat{\gamma})' \right)^{-1}.$$

4.5.4 Multiple thresholds

The double threshold model takes the form:

$$y_{it} = \mu_i + \beta_1' x_{it} I(q_{it} \leq \gamma_1) + \beta_2' x_{it} I(\gamma_1 < q_{it} \leq \gamma_2) + \beta_3' x_{it} I(q_{it} > \gamma_2) + e_{it}. \tag{4.74}$$

For a given (γ_1, γ_2) the concentrated SSE $S_1(\gamma_1, \gamma_2)$ is straightforward to calculate. The joint LSE of (γ_1, γ_2) are the values which jointly minimisation $S_1(\gamma_1, \gamma_2)$.

In the multiple changepoint model sequential estimation is consistent. In the first stage let $S_1(\gamma)$ be the single threshold SSE and let $\hat{\gamma}_1$ be the estimate that minimisation $S_1(\gamma)$. The analysis of Chong and Bai suggests that $\hat{\gamma}_1$ will be consistent for either γ_1 or γ_2 (depending on which effect is stronger). Fixing the first stage $\hat{\gamma}_1$, the second stage criterion is

$$S_2^r(\gamma_2) = \left\{ \begin{array}{ll} S(\hat{\gamma}_1, \hat{\gamma}_2) & if\ \hat{\gamma}_1 < \gamma_2 \\ S(\hat{\gamma}_2, \hat{\gamma}_1) & if\ \hat{\gamma}_1 > \gamma_2 \end{array} \right\},$$

and the 2nd stage threshold estimate is

$$\hat{\gamma}_2^r = \arg\min_{\gamma_2} S_2^r(\gamma_2).$$

Bai (1997) showed that $\hat{\gamma}_2^r$ is asymptotically efficient but $\hat{\gamma}_1$ is not, and suggested the following refinement estimator. Fixing $\hat{\gamma}_2^r$, define the refinement criterion

$$S_1^r(\gamma_1) = \left\{ \begin{array}{ll} S(\gamma_1, \hat{\gamma}_2^r) & if\ \gamma_1 < \hat{\gamma}_2^r \\ S_1(\hat{\gamma}_2^r, \gamma_1) & if\ \gamma_1 > \hat{\gamma}_2^r \end{array} \right\},$$

and the refinement estimator

$$\hat{\gamma}_1^r = \arg\min_{\gamma_1} S_1^r(\gamma_1).$$

Bai shows that the refinement estimator $\hat{\gamma}_1^r$ is asymptotically efficient and we expect similar results to hold in threshold regression.

The minimisation SSE from the 2nd stage threshold estimate is $S_2^r(\hat{\gamma}_2^r)$ with variance estimate, $\hat{\sigma}^2 = S_2^r(\hat{\gamma}_2^r)/n(T-1)$. Thus an approximate LR test of one vs two thresholds can be based on the statistic:

$$F_2 = \frac{S_1(\hat{\gamma}_1) - S_2^r(\hat{\gamma}_2^r)}{\hat{\sigma}^2}.$$

Since the null asymptotic distribution is non-pivotal we suggest using a bootstrap procedure. Treat x_{it} and q_{it} as given. The bootstrap errors will be drawn from the residuals calculated under the alternative, so the residuals from (4.74). Group the residuals, $\hat{e}_{it}^*$ by individual: $\hat{e}_i^* = (\hat{e}_{i1}^*, ..., \hat{e}_{iT}^*)$ and treat $(\hat{e}_i^*, ..., \hat{e}_n^*)$ as the empirical distribution to be used for bootstrapping. Draw (with replacement) a sample of size n from the empirical distribution. Let $e_i^{\#}$ be a generic $T \times 1$ draw. The dependent variable should be generated under the null of a single threshold:

$$y_{it}^{\#} = \hat{\beta}_1' x_{it} I(q_{it} \leq \hat{\gamma}) + \hat{\beta}_2' x_{it} I(q_{it} > \hat{\gamma}) + e_{it}^{\#}. \tag{4.75}$$

From the bootstrap sample F_2 may be calculated. Repeat this procedure a large number of times to calculate the bootstrap p-value. The null sampling distribution of F_2 depends asymptotically on both γ and β_1, β_2. This leads us to expect that the bootstrap may not produce as accurate CV for F_2 as for F_1, and neither is expected to be second-order accurate.

Let

$$LR_2^r(\gamma) = \frac{S_2^r(\gamma) - S_2^r(\hat{\gamma}_2^r)}{\hat{\sigma}^2},\ LR_1^r(\gamma) = \frac{S_1^r(\gamma) - S_1^r(\hat{\gamma}_1^r)}{\hat{\sigma}^2}.$$

Asymptotic $(1-\alpha)\,\%$ CI for γ_2 and γ_1 are the set of values of γ such that $LR_2^r(\gamma) \leq c(\alpha)$ and $LR_1^r(\gamma) \leq c(\alpha)$.

4.5.5 Investment and financing constraints

We use the multiple threshold regression model:

$$\begin{aligned} I_{it} &= \mu_i + \theta_1 Q_{it-1} + \theta_2 Q_{it-1}^2 + \theta_3 Q_{it-1}^3 + \theta_4 D_{it-1} + \theta_5 Q_{it-1} D_{it-1} \\ &\quad + \beta_1 CF_{it-1} I(D_{it-1} \leq \gamma_1) + \beta_2 CF_{it-1} I(\gamma_1 < D_{it-1} \leq \gamma_2) \\ &\quad + \beta_3 CF_{it-1} I(D_{it-1} > \gamma_2) + e_{it} \end{aligned}$$

where I_{it} is the ratio of investment to capital, Q_{it} is the ratio of total market value to assets, CF_{it} is the ratio of cash flow to assets and D_{it} is the ratio of long term debt to assets, where the stock variables are defined at the end of year.

See Table 5 for the results: what is unexpected is that the firm with the highest debt levels have the smallest coefficient. Also in all three cases the coefficients on cash flows are positive.

4.6 Threshold Autoregressive Models in Dynamic Panels

Increasing availability of large panel data sets, in conjunction with various developments in time series analysis, has prompted more rigorous econometric analyses of dynamic heterogeneous panels. Until recently most econometric analysis has stopped short of studying the issues of nonlinear asymmetric dynamic mechanisms explicitly within a panel data context. Hansen (1999) develops the panel threshold regression model where regression coefficients can take on a small number of different values, depending on the value of other exogenous stationary variable. González, Teräsvirta and van Dijk (2005) generalise this approach and develop a panel smooth transition regression model which allows the coefficients to change gradually from one regime to another. See also Fok, van Dijk and Franses (2005). In a broad context these models are a specific example of the panel data approach that allows coefficients to vary randomly over time and across cross-sectional units as surveyed by Hsiao (2003, Chapter 6). Both approaches are static in nature, though they can be applied to the conventional panel data with large N and fixed T.

In general, there have been a rather small number of studies to adopt these time-series technique into the dynamic panel data model with large N and fixed T, though there is a huge literature on GMM estimation of linear dynamic panels, e.g., Arellano and Bond (1991), Ahn and Schmidt (1995), Arellano and Bover (1995), Blundell and Bond (1998), Blundell, Bond and Windmeijer (2000), Alvarez and Arellano (2003) and Hayakawa (2006). Further, there is no rigorous single study investigating the important issue of nonlinear asymmetric dynamic mechanism in this context. We aim to fill this gap so that we will be to address the issue as how best to model nonlinear asymmetric dynamics mechanism and cross-sectional heterogeneity, simultaneously. This would use the time series techniques advanced by Chan (1993) and Hansen (1996, 2000) with the existing GMM estimation techniques in linear dynamic panels. We develop a threshold autoregressive model in dynamic panels with large N and fixed T and propose various GMM estimation methodologies, namely the FD-GMM, the Level-GMM and the System-GMM estimators. We also provide the bootstrap-based inference procedure for the presence of threshold effects.

4.6.1 Model

Consider the following panel threshold autoregressive model:

$$y_{it} = \phi_1 y_{it-1} 1\left(q_{it} \leq \gamma\right) + \phi_2 y_{it-1} 1\left(q_{it} > \gamma\right) + \varepsilon_{it},\ i = 1, ..., N;\ t = 1, ..., T, \tag{4.76}$$

where y_{it} is a a scalar stochastic variable of interest, $1(\cdot)$ is an indicator function, q_{it} is the transition variable with γ being a threshold parameter, ϕ_1, ϕ_2 are heterogeneous autoregressive parameters associated with different regimes, and ε_{it} consists of the error components

$$\varepsilon_{it} = \alpha_i + v_{it},$$

where α_i is an unobserved individual effect and v_{it} is a zero mean idiosyncratic random disturbance. This is a panel extension of the TAR model popularised by Tong (1990).

We make the following assumptions:

Assumption 1. $\{v_{it}\}$ are iid and independent of η_{it} and y_{i1} with $E(v_{it}) = 0$, $Var(v_{it}) = \sigma^2$ and have the finite 4th moment.

Assumption 2. α_i are iid with $E(\alpha_i) = 0$, $Var(\alpha_i) = \sigma_\alpha^2$ and have the finite 4th moment.

Assumption 3. y_{it} is geometrically ergodic and the initial observations satisfy the mean stationarity condition.

Assumption 4. The threshold variable, q_{it} is stationary and exogenous or predetermined uncorrelated with α_i and v_{it}.

Assumption 5. N is large and T is fixed.

All these assumptions are fairly standard in the literature, e.g. Alvarez and Arellano (2003) and Hansen (1999).

4.6.2 FD-GMM Estimator

It is well-established that the fixed effects estimator of the autoregressive parameter is biased downward, e.g. Nickell (1981). To deal with the correlation of the regressors with individual effects in (4.1), we follow Arellano and Bond (1991) and consider the first-difference transformation. For convenience we define $\lambda_{it}(\gamma) = 1(q_{it} \leq \gamma)$ and write (4.1) as

$$y_{it} = \phi_1 y_{i,t-1} \lambda_{it}(\gamma) + \phi_2 y_{i,t-1}(1 - \lambda_{it}(\gamma)) + \varepsilon_{it}, \tag{4.77}$$

Taking the first difference of (4.77) to get rid of α_i, we obtain

$$\begin{aligned} \Delta y_{it} &= \phi_1 (y_{i,t-1}\lambda_{it}(\gamma) - y_{i,t-2}\lambda_{i,t-1}(\gamma)) \\ &\quad + \phi_2 (y_{i,t-1}(1 - \lambda_{it}(\gamma)) - y_{i,t-2}(1 - \lambda_{i,t-1}(\gamma))) + \Delta v_{it}, \end{aligned} \tag{4.78}$$

for $i = 1, ..., N$ and $t = 2, ..., T$. The OLS estimator from (4.78) is biased since the transformed regressors are correlated with Δv_{it}. To fix this problem we need to find instruments for $(y_{i,t-1}\lambda_{it}(\gamma) - y_{i,t-2}\lambda_{i,t-1}(\gamma))$ and $(y_{i,t-1}(1 - \lambda_{it}(\gamma)) - y_{i,t-2}(1 - \lambda_{i,t-1}(\gamma)))$. The obvious candidates are $y_{i,t-2}\lambda_{i,t-1}(\gamma)$, $y_{i,t-2}(1 - \lambda_{i,t-1}(\gamma))$ and their lagged values. These instruments will not be correlated with Δv_{it} as long as the v_{it} are assumed to be serially uncorrelated.

To further simplify the notations we define

$$\begin{aligned} z_{1it}(\gamma) &= y_{i,t-1}\lambda_{it}(\gamma);\ z_{2it}(\gamma) = y_{i,t-1}(1-\lambda_{it}(\gamma)); \\ \mathbf{z}_{it}(\gamma) &= (z_{1,it}(\gamma), z_{2,it}(\gamma));\ \boldsymbol{\phi} = (\phi_1, \phi_2)', \end{aligned}$$

and write (4.77) and (4.78) respectively as

$$y_{it} = \mathbf{z}_{it}(\gamma)\boldsymbol{\phi} + \varepsilon_{it}, \tag{4.79}$$

$$\Delta y_{it} = \Delta \mathbf{z}_{it}(\gamma)\boldsymbol{\phi} + \Delta v_{it}. \tag{4.80}$$

When investigating the list of instruments by exploiting additional moment conditions, it is straightforward to obtain the following IV matrices for $\Delta z_{1,it}(\gamma)$ and $\Delta z_{2,it}(\gamma)$ respectively for individual $i = 1, ..., N$:

$$\mathbf{W}_{1i}^{d}(\gamma) = \begin{bmatrix} z_{1,i2}(\gamma) & 0 & \cdots & 0 \\ 0 & z_{1,i2}(\gamma), z_{1,i3}(\gamma) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & z_{1,i2}(\gamma), ..., z_{1,i,T-1}(\gamma) \end{bmatrix}, \tag{4.81}$$

$$\mathbf{W}_{2i}^{d}(\gamma) = \begin{bmatrix} z_{2,i2}(\gamma) & 0 & \cdots & 0 \\ 0 & z_{2,i2}(\gamma), z_{2,i3}(\gamma) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & z_{2,i2}(\gamma), ..., z_{2,1i,T-1}(\gamma) \end{bmatrix}, \tag{4.82}$$

where the dimensions of $\mathbf{W}_{1i}^{d}(\gamma)$ and $\mathbf{W}_{2i}^{d}(\gamma)$ are $(T-2) \times m_d$ with $m_d = 0.5(T-2)(T-1)$. Combining them together we obtain the $(T-2) \times 2m_d$ moment matrix for individual $i = 1, ..., N$:

$$\mathbf{W}_{i}^{d}(\gamma) = \left(\mathbf{W}_{1i}^{d}(\gamma), \mathbf{W}_{2i}^{d}(\gamma)\right), \tag{4.83}$$

and the $N(T-2) \times 2m_d$ full matrix of instruments:

$$\mathbf{W}^{d}(\gamma) = \begin{bmatrix} \mathbf{W}_{1}^{d}(\gamma) \\ \vdots \\ \mathbf{W}_{N}^{d}(\gamma) \end{bmatrix}. \tag{4.84}$$

Next, we write (4.80) in the matrix form as

$$\Delta \mathbf{y} = \Delta \mathbf{Z}(\gamma)\boldsymbol{\phi} + \Delta \mathbf{v}, \tag{4.85}$$

where

$$\Delta \mathbf{y} = \begin{bmatrix} \Delta \mathbf{y}_1 \\ \vdots \\ \Delta \mathbf{y}_N \end{bmatrix}_{N(T-2)\times 1}, \ \Delta \mathbf{Z}(\gamma) = \begin{bmatrix} \Delta \mathbf{z}_1(\gamma) \\ \vdots \\ \Delta \mathbf{z}_N(\gamma) \end{bmatrix}_{N(T-2)\times 2}, \Delta \mathbf{v} = \begin{bmatrix} \Delta \mathbf{v}_1 \\ \vdots \\ \Delta \mathbf{v}_N \end{bmatrix}_{N(T-2)\times 1},$$

$$\Delta \mathbf{y}_i = \begin{bmatrix} \Delta y_{i3} \\ \vdots \\ \Delta y_{iT} \end{bmatrix}_{(T-2)\times 1}, \ \Delta \mathbf{z}_i(\gamma) = \begin{bmatrix} \Delta \mathbf{z}_{i3}(\gamma) \\ \vdots \\ \Delta \mathbf{z}_{iT}(\gamma) \end{bmatrix}_{(T-2)\times 2}, \Delta \mathbf{v}_i = \begin{bmatrix} \Delta v_{i3} \\ \vdots \\ \Delta v_{iT} \end{bmatrix}_{(T-2)\times 1},$$

and express the set of all moment conditions concisely as

$$E\left(\mathbf{W}^d(\gamma)' \Delta \mathbf{v}\right) = \mathbf{0}. \tag{4.86}$$

The one-step FD-GMM estimator is then obtained by

$$\begin{aligned} \hat{\boldsymbol{\phi}}_1^d(\gamma) &= \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_1^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{Z}(\gamma)\right\}^{-1} \\ &\quad \times \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_1^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{y}\right\}, \end{aligned} \tag{4.87}$$

where

$$\mathbf{V}_1^d(\gamma) = \sum_{i=1}^{N} \mathbf{W}_i^d(\gamma)' \mathbf{G} \mathbf{W}_i^d(\gamma),$$

and $\mathbf{G}$ is a $(T-2)\times(T-2)$ fixed matrix given by

$$\mathbf{G} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{bmatrix}.$$

If v_{it}'s are assumed to be homoskedastic, then the optimal GMM estimator can be computed in one step. In a more general case where v_{it}'s are heteroskedastic, the weighting matrix should be estimated without imposing these restrictions. Thus the two-step FD-GMM estimator is obtained by

$$\begin{aligned} \hat{\boldsymbol{\phi}}_2^d(\gamma) &= \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_2^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{Z}(\gamma)\right\}^{-1} \\ &\quad \times \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_2^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{y}\right\}, \end{aligned} \tag{4.88}$$

where

$$\mathbf{V}_2^d(\gamma) = \sum_{i=1}^{N} \mathbf{W}_i^d(\gamma)' \Delta \hat{\mathbf{v}}_i(\gamma) \Delta \hat{\mathbf{v}}_i'(\gamma) \mathbf{W}_i^d(\gamma), \ \Delta \hat{\mathbf{v}}_i(\gamma) = \Delta \mathbf{y}_i - \Delta \mathbf{z}_i(\gamma) \hat{\boldsymbol{\phi}}_1^d(\gamma). \tag{4.89}$$

For given γ and for large N, $\boldsymbol{\phi}$ can be consistently estimated by the FD-GMM estimators derived in (4.87) and (4.88), and they are asymptotically normally distributed with covariance matrices given respectively by

$$Var\left(\hat{\boldsymbol{\phi}}_1^d(\gamma)\right) = \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_1^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{Z}(\gamma)\right\}^{-1}, \tag{4.90}$$

$$Var\left(\hat{\boldsymbol{\phi}}_2^d(\gamma)\right) = \left\{\Delta \mathbf{Z}(\gamma)' \mathbf{W}^d(\gamma) \mathbf{V}_2^d(\gamma)^{-1} \mathbf{W}^d(\gamma)' \Delta \mathbf{Z}(\gamma)\right\}^{-1}. \tag{4.91}$$

4.6.3 System-GMM Estimator

Blundell and Bond (1998) demonstrate that the FD-GMM estimator is subject to weak instruments problem especially when the AR coefficient is close to 1 and/or when the variance of the individual effects, σ_α^2 increases relative to the variance of the idiosyncratic error, σ_v^2. They propose the system GMM approach by combining lagged differences as instruments for equations in levels, in addition to lagged levels as instruments for equations in first differences.

We follow Arellano and Bover (1995) and Blundell and Bond (1998) and consider the following additional first-difference moment conditions for the level equations for individual $i = 1, ..., N$:

$$E\left(\varepsilon_{it}\Delta z_{j,i,t-s}\left(\gamma\right)\right) = 0 \text{ for } j = 1, 2;\ t = 3, ..., T \text{ and } 0 \leq s \leq t-3, \quad (4.92)$$

which is satisfied under the mean stationairty of the y_{it} process, see Assumption 3. Using (4.92) we first derive a Level-GMM estimator. In this case we obtain the IV matrices for $z_{1,it}\left(\gamma\right)$ and $z_{2,it}\left(\gamma\right)$ respectively for individual $i = 1, ..., N$:

$$\mathbf{W}_{1i}^{l}\left(\gamma\right) = \begin{bmatrix} \Delta z_{1,i3}\left(\gamma\right) & 0 & \cdots & 0 \\ 0 & \Delta z_{1,i3}\left(\gamma\right), \Delta z_{1,i4}\left(\gamma\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta z_{1,i3}\left(\gamma\right), ..., \Delta z_{1,i,T}\left(\gamma\right) \end{bmatrix}, \quad (4.93)$$

$$\mathbf{W}_{2i}^{l}\left(\gamma\right) = \begin{bmatrix} \Delta z_{2,i3}\left(\gamma\right) & 0 & \cdots & 0 \\ 0 & \Delta z_{2,i3}\left(\gamma\right), \Delta z_{2,i4}\left(\gamma\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta z_{2,i3}\left(\gamma\right), ..., \Delta z_{2,1i,T}\left(\gamma\right) \end{bmatrix}, \quad (4.94)$$

where the dimension of $\mathbf{W}_{1i}^{l}\left(\gamma\right)$ and $\mathbf{W}_{1i}^{l}\left(\gamma\right)$ is $\left(T-2\right) \times m_l$ with $m_l = 0.5\left(T-2\right)\left(T-1\right)$. Combining them together we have

$$\mathbf{W}_{i}^{l}\left(\gamma\right) = \left(\mathbf{W}_{1i}^{l}\left(\gamma\right), \mathbf{W}_{2i}^{l}\left(\gamma\right)\right),\ i = 1, ..., N;\ \mathbf{W}^{l}\left(\gamma\right) = \begin{bmatrix} \mathbf{W}_{1}^{l}\left(\gamma\right) \\ \vdots \\ \mathbf{W}_{N}^{l}\left(\gamma\right) \end{bmatrix}. \quad (4.95)$$

Next, we write the level-equation, (4.79) in the matrix form as

$$\mathbf{y} = \mathbf{Z}\left(\gamma\right)\boldsymbol{\phi} + \boldsymbol{\varepsilon}, \quad (4.96)$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}_{N(T-2)\times 1}, \ \mathbf{Z}\left(\gamma\right) = \begin{bmatrix} \mathbf{z}_1\left(\gamma\right) \\ \vdots \\ \mathbf{z}_N\left(\gamma\right) \end{bmatrix}_{N(T-2)\times 2}, \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}_{N(T-2)\times 1},$$

$$\mathbf{y}_i = \begin{bmatrix} y_{i3} \\ \vdots \\ y_{iT} \end{bmatrix}_{(T-2)\times 1}, \ \mathbf{z}_i(\gamma) = \begin{bmatrix} \mathbf{z}_{i3}(\gamma) \\ \vdots \\ \mathbf{z}_{iT}(\gamma) \end{bmatrix}_{(T-2)\times 2}, \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i3} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}_{(T-2)\times 1}.$$

Therefore, the set of all moment conditions for the level equation, (4.96) can be written concisely as

$$E\left(\mathbf{W}^l(\gamma)'\boldsymbol{\varepsilon}\right) = \mathbf{0}. \tag{4.97}$$

We then obtain the one-step and two-step Level-GMM estimators by

$$\hat{\boldsymbol{\phi}}_1^l(\gamma) = \left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_1^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{Z}(\gamma)\right\}^{-1}\left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_1^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{y}\right\}, \tag{4.98}$$

$$\hat{\boldsymbol{\phi}}_2^l(\gamma) = \left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_2^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{Z}(\gamma)\right\}^{-1}\left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_2^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{y}\right\}, \tag{4.99}$$

where

$$\mathbf{V}_1^l(\gamma) = \sum_{i=1}^{N}\mathbf{W}_i^l(\gamma)'\mathbf{W}_i^l(\gamma),$$

$$\mathbf{V}_2^l(\gamma) = \sum_{i=1}^{N}\mathbf{W}_i^l(\gamma)'\hat{\boldsymbol{\varepsilon}}_i(\gamma)\hat{\boldsymbol{\varepsilon}}_i'(\gamma)\mathbf{W}_i^l(\gamma), \ \hat{\boldsymbol{\varepsilon}}_i(\gamma) = \mathbf{y}_i - \mathbf{z}_i(\gamma)\hat{\boldsymbol{\phi}}_1^l(\gamma).$$

Hence, for given γ and for large N, $\boldsymbol{\phi}$ in (4.96) can be consistently estimated by the Level-GMM estimators derived in (4.98) and (4.99), which are asymptotically normally distributed with covariance matrices given by

$$Var\left(\hat{\boldsymbol{\phi}}_1^l(\gamma)\right) = \left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_1^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{Z}(\gamma)\right\}^{-1}, \tag{4.100}$$

$$Var\left(\hat{\boldsymbol{\phi}}_2^l(\gamma)\right) = \left\{\mathbf{Z}(\gamma)'\mathbf{W}^l(\gamma)\mathbf{V}_2^l(\gamma)^{-1}\mathbf{W}^l(\gamma)'\mathbf{Z}(\gamma)\right\}^{-1}. \tag{4.101}$$

Next, we derive the system GMM estimator combining all the moment conditions in both levels and first-differences equations. Combining (4.85) and (4.96), we obtain the system equations:[2]

$$\mathbf{Y} = \mathbf{X}(\gamma)\boldsymbol{\phi} + \mathbf{u}, \tag{4.102}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_N \end{bmatrix}_{N(T-2)\times 1}, \ \mathbf{X}(\gamma) = \begin{bmatrix} \mathbf{X}_1(\gamma) \\ \vdots \\ \mathbf{X}_N(\gamma) \end{bmatrix}_{N(T-2)\times 2}, \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{bmatrix}_{N(T-2)\times 1},$$

$$\mathbf{Y}_i = \begin{bmatrix} \Delta\mathbf{y}_i \\ \mathbf{y}_i \end{bmatrix}_{2(T-2)\times 1}, \ \mathbf{X}_i(\gamma) = \begin{bmatrix} \Delta\mathbf{z}_i(\gamma) \\ \mathbf{z}_i(\gamma) \end{bmatrix}_{2(T-2)\times 2}, \mathbf{u}_i = \begin{bmatrix} \Delta\mathbf{v}_i \\ \boldsymbol{\varepsilon}_i \end{bmatrix}_{2(T-2)\times 1}.$$

[2]The definition in (4.102) matches the construction of instruments in what follows.

We consider the three different versions of the system GMM estimator, denoted $\hat{\boldsymbol{\phi}}^{all}(\gamma)$, $\hat{\boldsymbol{\phi}}^{\min}(\gamma)$ and $\hat{\boldsymbol{\phi}}^{BB}(\gamma)$, respectively, e.g. Hayakawa (2006). First, $\hat{\boldsymbol{\phi}}^{all}(\gamma)$ uses all the moment conditions given by (4.86) and (4.97). In this case we have the following full moment matrix:

$$\mathbf{W}_i^{all}(\gamma) = \begin{bmatrix} \mathbf{W}_i^d(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_i^l(\gamma) \end{bmatrix},\ i = 1, ..., N;\ \mathbf{W}^{all}(\gamma) = \begin{bmatrix} \mathbf{W}_1^{all}(\gamma) \\ \vdots \\ \mathbf{W}_N^{all}(\gamma) \end{bmatrix}. \tag{4.103}$$

Secondly, $\hat{\boldsymbol{\phi}}^{\min}(\gamma)$ employs only the minimum necessary $2(T-2)$ moment conditions in levels and first-differences equations for individual $i = 1, ..., N$:

$$\mathbf{W}_{ji}^{d,\min}(\gamma) = \begin{bmatrix} z_{j,i2}(\gamma) & 0 & \cdots & 0 \\ 0 & z_{j,i3}(\gamma) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & z_{j,i,T-1}(\gamma) \end{bmatrix},\ j = 1, 2, \tag{4.104}$$

$$\mathbf{W}_{ji}^{l,\min}(\gamma) = \begin{bmatrix} \Delta z_{j,i3}(\gamma) & 0 & \cdots & 0 \\ 0 & \Delta z_{j,i4}(\gamma) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta z_{j,1i,T}(\gamma) \end{bmatrix},\ j = 1, 2. \tag{4.105}$$

We thus have the following moment matrix:

$$\mathbf{W}_i^{\min}(\gamma) = \begin{bmatrix} \mathbf{W}_i^{d,\min}(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_i^{l,\min}(\gamma) \end{bmatrix},\ i = 1, ..., N;\ \mathbf{W}^{\min}(\gamma) = \begin{bmatrix} \mathbf{W}_1^{\min}(\gamma) \\ \vdots \\ \mathbf{W}_N^{\min}(\gamma) \end{bmatrix}. \tag{4.106}$$

Thirdly, $\hat{\boldsymbol{\phi}}^{BB}(\gamma)$ combines the full set of moment conditions, $\mathbf{W}_i^d(\gamma)$ in differences equations and a nonredundant subset of moment conditions, $\mathbf{W}_i^{l,\min}(\gamma)$ in levels equation. We then have:

$$\mathbf{W}_i^{BB}(\gamma) = \begin{bmatrix} \mathbf{W}_i^d(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_i^{l,\min}(\gamma) \end{bmatrix},\ i = 1, ..., N;\ \mathbf{W}^{BB}(\gamma) = \begin{bmatrix} \mathbf{W}_1^{BB}(\gamma) \\ \vdots \\ \mathbf{W}_N^{BB}(\gamma) \end{bmatrix}. \tag{4.107}$$

Therefore, the one-step and two-step System-GMM estimators for $h = all$, min and BB are obtained by[3]

$$\hat{\boldsymbol{\phi}}_1^h(\gamma) = \left\{\mathbf{X}(\gamma)'\mathbf{W}^h(\gamma)\mathbf{V}_1^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{X}(\gamma)\right\}^{-1}\left\{\mathbf{X}(\gamma)'\mathbf{W}^h(\gamma)\mathbf{V}_1^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{Y}\right\}, \tag{4.108}$$

[3] It is easily seen that the system GMM estimator is equivalent to the linear combination of FD-GMM and Level-GMM estimators where the weights are different for different moment matrices employed, e.g. Blundell et al. (2000).

$$\hat{\boldsymbol{\phi}}_2^h(\gamma) = \left\{\mathbf{X}(\gamma)'\mathbf{W}^h(\gamma)\mathbf{V}_2^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{X}(\gamma)\right\}^{-1}\left\{\mathbf{X}(\gamma)\mathbf{W}^h(\gamma)\mathbf{V}_2^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{Y}\right\}, \tag{4.109}$$

where

$$\mathbf{V}_1^h(\gamma) = \sum_{i=1}^{N}\mathbf{W}_i^h(\gamma)'\mathbf{H}\mathbf{W}_i^h(\gamma), \ \mathbf{H} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{T-2} \end{bmatrix},$$

$$\mathbf{V}_2^h(\gamma) = \sum_{i=1}^{N}\mathbf{W}_i^h(\gamma)'\hat{\mathbf{u}}_i^h(\gamma)\hat{\mathbf{u}}_i^{h\prime}(\gamma)\mathbf{W}_i^h(\gamma), \ \hat{\mathbf{u}}_i^h(\gamma) = \mathbf{Y}_i - \mathbf{X}_i(\gamma)\hat{\boldsymbol{\phi}}_1^h(\gamma).$$

For given γ and for large N, $\boldsymbol{\phi}$ in (4.102) can be consistently estimated by the System-GMM estimators derived above, and they are asymptotically normally distributed with covariance matrices given respectively

$$Var\left(\hat{\boldsymbol{\phi}}_1^h(\gamma)\right) = \left\{\mathbf{X}(\gamma)'\mathbf{W}^h(\gamma)\mathbf{V}_1^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{X}(\gamma)\right\}^{-1}, \ h = all, \min, BB, \tag{4.110}$$

$$Var\left(\hat{\boldsymbol{\phi}}_2^h(\gamma)\right) = \left\{\mathbf{X}(\gamma)'\mathbf{W}^h(\gamma)\mathbf{V}_2^h(\gamma)^{-1}\mathbf{W}^h(\gamma)'\mathbf{X}(\gamma)\right\}^{-1}, \ h = all, \min, BB \tag{4.111}$$

4.6.4 Estimation of and Testing for Threshold Effects

We have developed the optimal estimation procedure for the threshold autoregressive model in dynamic panels under the implicit assumption that the value of the threshold parameter, γ is given. This section will address consistent estimation of γ and develop the bootstrap-based testing procedure for the null of no threshold effects in dynamic panels. For convenience we focus on the case of the FD-GMM estimator.[4]

We obtain the consistent estimator of γ by (Chan, 1993; Hansen, 1999)

$$\hat{\gamma} = \arg\min_{\gamma} Q_1(\gamma), \tag{4.112}$$

where $Q_1(\gamma)$ is the generalised minimum distance measure given by

$$Q_1(\gamma) = \Delta\hat{\mathbf{v}}(\gamma)'\mathbf{W}^d(\gamma)\mathbf{V}_2^d(\gamma)^{-1}\mathbf{W}^d(\gamma)'\Delta\hat{\mathbf{v}}(\gamma), \tag{4.113}$$

$$\Delta\hat{\mathbf{v}}(\gamma) = \Delta\mathbf{y} - \Delta\mathbf{Z}(\gamma)\hat{\boldsymbol{\phi}}_2^d(\gamma).$$

Once $\hat{\gamma}$ is obtained, we obtain

$$\hat{\boldsymbol{\phi}}_2^d = \hat{\boldsymbol{\phi}}_2^d(\hat{\gamma}); \ \Delta\hat{\mathbf{v}} = \Delta\hat{\mathbf{v}}(\hat{\gamma}); \ \hat{\sigma}^2 = \frac{1}{N(T-2)}Q_1(\hat{\gamma}). \tag{4.114}$$

Since $Q_1(\gamma)$ depends only on γ through the indicator function, the sum of SSE is a step function with most $N(T-2)$ steps with the steps occurring at

[4]Estimation of the threshold parameter and test of threshold effects in the cases of Level-GMM and System-GMM estimators proceed exactly as described in this section.

distinct values of the observed threshold variable q_{it}. Thus the minimisation problem can be reduced to searching over the values of γ equalling the distinct values of q_{it} in the sample. In practice we need to truncate the smallest and largest 10% for example. The remaining values constitute the values of γ which can be searched for $\hat{\gamma}$. For each of these values regression are estimated yielding the SSE and the smallest value yields the estimate $\hat{\gamma}$ and $\hat{\boldsymbol{\phi}}_2^d = \hat{\boldsymbol{\phi}}_2^d(\hat{\gamma})$.

We follow Hansen (1996) and develop a bootstrap procedure to simulate the asymptotic distribution of the LR test statistic for the null hypothesis of no threshold:

$$H_0 : \phi_1 = \phi_2.$$

Under the null of no threshold, the model (4.80) reduces to

$$y_{it} = \phi_1 y_{i,t-1} + \varepsilon_{it}. \tag{4.115}$$

Taking the FD transformation, we have

$$\Delta y_{it} = \phi_1 \Delta y_{i,t-1} + \Delta v_{it}, \tag{4.116}$$

from which we obtain the linear one-step and two-step FD-GMM estimators denoted $\tilde{\phi}_1^d$ and $\tilde{\phi}_2^d$, respectively. Then we obtain the the generalised minimum distance measure under the null by

$$Q_0 = \Delta\tilde{\mathbf{v}}'\tilde{\mathbf{W}}^d \left(\tilde{\mathbf{V}}_2^d\right)^{-1} \tilde{\mathbf{W}}^{d\prime}\Delta\tilde{\mathbf{v}}, \tag{4.117}$$

where $\Delta\tilde{\mathbf{v}} = \Delta\mathbf{y} - \tilde{\phi}_{1,2}^d \Delta\mathbf{y}_{-1}$ and $\tilde{\mathbf{W}}^d$ and $\tilde{\mathbf{V}}_2^d$ are the corresponding instrument matrix and optimal weighting matrix for the two-step linear FD-GMM estimator. The LR test statistic is then given by

$$LR = \frac{Q_0 - Q_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{4.118}$$

The bootstrap p-value of the LR statistic is evaluated as follows:[5] We first take the residuals, $\Delta\hat{\mathbf{v}}(\hat{\gamma}) = \left(\Delta\hat{\mathbf{v}}_1(\hat{\gamma})', ..., \Delta\hat{\mathbf{v}}_N(\hat{\gamma})'\right)'$ with $\Delta\hat{\mathbf{v}}_i(\hat{\gamma}) = \left(\Delta\hat{v}_{i3}(\hat{\gamma}), ..., \Delta\hat{v}_{iT}(\hat{\gamma})\right)'$ and treat them as the empirical distribution to be used. We then generate the jth bootstrap sample residual vector, denoted $\Delta\mathbf{v}^{(j)}$ by drawing (with replacement) from the empirical distribution and use these errors to create a bootstrap sample under H_0,

$$\Delta y_{it}^{(j)} = \tilde{\phi}_1 \Delta y_{i,t-1}^{(j)} + \Delta v_{it}^{(j)},\ j = 1, ..., B, \tag{4.119}$$

for $i = 1, ..., N$ and $t = 3, ..., T$, where $\tilde{\phi}_1$ is the two-step GMM estimator obtained from (4.116) and we treat the initial values y_{i1} and y_{i2} as given.

[5] Hansen (1996) shows that a bootstrap procedure attains the first-order asymptotic distribution, so p-values are asymptotically valid.

Using the bootstrap sample generated in (4.119) we estimate the model under the null and under the alternative and calculate the bootstrap value of the LR test at each replication. We set the number of replications, $B = 1000$ and calculate the percentage of draws for which the simulated statistic exceeds the actual one. This is the bootstrap estimate of the asymptotic p-value for LR under H_0.

4.6.5 Asymmetric capital structure adjustments: New evidence from dynamic panel threshold models

Dang, Kim and Shin (2012) develop a dynamic panel threshold model of capital structure to test the dynamic trade-off theory, allowing for asymmetries in firms' adjustments toward target leverage. Our novel estimation approach is able to consistently estimate heterogeneous speeds of adjustment in different regimes as well as to properly test for the threshold effect. We consider several proxies for adjustment costs that affect the asymmetries in capital structure adjustments and find evidence that firms with large financing imbalance (or a deficit), large investment or low earnings volatility adjust faster than those with the opposite characteristics. Firms not only adjust at different rates but also seem to adjust toward heterogeneous leverage targets. Moreover, we document a consistent pattern that firms undertaking quick adjustment are over-levered with a financing deficit and rely heavily on equity issues to make such adjustment.

Dynamic capital structure adjustment models The conventional econometric specification to model firms' adjustment toward target leverage takes the form of a partial adjustment process:

$$\Delta \ell_{it} = \delta \left(\ell_{it}^{*} - \ell_{it-1} \right) + v_{it}$$

where ℓ_{it} and ℓ_{it}^{*} denote the actual (observed) and target leverage ratios for firm i at time t. v_{it} is an error component. δ is the speed of adjustment that measures how fast firms move toward their target leverage. Target leverage can be considered as a unique ratio determined by firms' characteristics as:[6]

$$\ell_{it}^{*} = \beta' x_{it}$$

where x_{it} denotes the $k \times 1$ vector of exogenous factors determining target leverage with β being the structural parameters.

We turn to the one-stage procedure by combining the above two equations:

$$\ell_{it} = \phi \ell_{it-1} + \pi' x_{it} + v_{it}$$

[6] Here, we follow the literature and consider the five most commonly-used determinants of leverage, namely (asset) tangibility, growth opportunities, non-debt tax shields, profitability and firm size.

where $\phi = 1-\delta$, $\pi = \delta\beta$ and v_{it} is an one-way error component that includes the individual firm fixed effects:

$$v_{it} = \alpha_i + e_{it}$$

Both the short-run dynamics, $\hat{\phi}$, and the long-run coefficients, $\hat{\beta} = \frac{\hat{\pi}}{1-\hat{\phi}}$, can be jointly estimated in one stage.

Firms adjust at different rates according to the position of their actual leverage relative to targets as well as the costs of their adjustment. To capture this dynamic trade-off behavior, we develop the regime-switching, dynamic threshold model:

$$\ell_{it} = \left(\phi_1 \ell_{it-1} + \pi_1' x_{it}\right) 1_{\{q_{it} \leq c\}} + \left(\phi_2 \ell_{it-1} + \pi_2' x_{it}\right) 1_{\{q_{it} > c\}} + v_{it}$$

where $1_{\{.\}}$ is an indicator function taking the value 1 if the event is true and 0 otherwise. Model represents an important extension of the (linear) partial adjustment model, in that it allows for short-run asymmetries in two AR(1) parameters (ϕ_1 and ϕ_2), the implied speeds of adjustment ($\delta_1 = 1 - \phi_1$ and $\delta_2 = 1 - \phi_2$), and the short run coefficients (π_1 and π_2) as well as long-run asymmetries in the target leverage (β_1 and β_2), conditional on the transition variable, q_{it}, and the threshold parameter, c. For simplicity, the transition variable, q_{it}, is assumed to be stationary and exogenous.

Threshold partial adjustment models We derive the GMM estimators and describe how the threshold parameter is estimated and its confidence intervals are constructed. The fixed-effects (FE) estimates of ϕ_1 and ϕ_2 are biased downward because the regressors are correlated with (unobserved) firm fixed effects, α_i (Nickell, 1981). This suggests that the FE estimator of the speeds of adjustment, δ_1 and δ_2 is biased upward.

To address this issue, we follow the GMM literature. Specifically, we combine time series techniques on threshold modeling (Caner and Hansen, 2004; Hansen, 2000) with the existing GMM literature (Alvarez and Arellano, 2003). We first rewrite:

$$\ell_{it} = \left(\phi_1 \ell_{1,it-1}(c) + \pi_1' x_{1,it}(c)\right) + \left(\phi_2 \ell_{2,it-1}(c) + \pi_2' x_{2,it}(c)\right) + v_{it},\ v_{it} = \alpha_i + e_{it}$$

where $\ell_{1,it-1}(c) = \ell_{it-1} 1_{\{q_{it} \leq c\}}$, $\ell_{2,it-1}(c) = \ell_{it-1} 1_{\{q_{it} > c\}}$. Next, to deal with the correlation between the regressors and the firm fixed effects, we use the first-difference transformation:

$$\Delta\ell_{it} = \left(\phi_1 \Delta\ell_{1,it-1}(c) + \pi_1' \Delta x_{1,it}(c)\right) + \left(\phi_2 \Delta\ell_{2,it-1}(c) + \pi_2' \Delta x_{2,it}(c)\right) + \Delta e_{it}$$

However, applying the pooled OLS estimator still produces biased estimates since $\Delta\ell_{1,it-1}(c)$ and $\Delta\ell_{2,it-1}(c)$ are correlated with Δe_{it}. Hence, we need to find their instruments that satisfy the orthogonal condition with Δe_{it}.

Two obvious candidates for these instruments are $\ell_{1,it-2}(c)$ and $\ell_{2,it-2}(c)$, as commonly used in the (just-identified) instrumental variable estimation approach (AH-IV) (Anderson and Hsiao, 1982).

To improve the efficiency of the AH-IV estimator, we follow (Arellano and Bond, 1991) and consider lagged values of $\ell_{1,it-2}(c)$ and $\ell_{2,it-2}(c)$ as additional instruments. We next construct the full GMM instrument matrices for $\Delta\ell_{1,it-1}(c)$ and $\Delta\ell_{2,it-1}(c)$, denoted $W_{1i}(c)$ and $W_{2i}(c)$, for $i = 1, \ldots, N$ and $j = 1, 2$:

$$W_{1i}(c) = \begin{bmatrix} \ell_{j,i1}(c) & & & \\ & \ell_{j,i1}(c), \ell_{j,i2}(c) & & \\ & & & \ell_{j,iT-2}(c), \ell_{j,iT-2}(c) \end{bmatrix}$$

We express the model in the matrix form:

$$\Delta\ell = Z_1(c)\theta_1 + Z_2(c)\theta_2 + \Delta e = Z(c)\theta + \Delta e$$

where $Z_1(c) = (\Delta\ell_{1,-1}(c), \Delta X_1(c))$, $Z_2(c) = (\Delta\ell_{2,-1}(c), \Delta X_2(c))$, $Z(c) = (Z_1(c), Z_2(c))$, $\theta_1 = (\phi_1, \pi_1')'$, $\theta_2 = (\phi_2, \pi_2')'$, $\theta = (\theta_1', \theta_2')'$, $\Delta\ell = (\Delta\ell_1', ..., \Delta\ell_N')'$, $\Delta\ell_i = (\Delta\ell_{i2}, ..., \Delta\ell_{iT})'$, $\Delta\ell_{j,-1}(c) = \left(\Delta\ell_{j1,-1}(c), ..., \Delta\ell_{jN,-1}(c)\right)'$, $\Delta\ell_{ji,-1}(c) = \left(\Delta\ell_{ji1}(c), ..., \Delta\ell_{ji,T-1}(c)\right)'$, $\Delta X_j(c) = \left(\Delta X_{j1}(c), ..., \Delta X_{jN}(c)\right)'$, $\Delta X_{ji}(c) = (\Delta X_{ji2}(c), ..., \Delta X_{jiT}(c))'$ for $j = 1, 2$.

We can construct the associated instrument matrix for $Z(c)$ as the following $N(T-2) \times \{(T-2)(T-1) + 2K\}$ matrix:

$$W(c) = \begin{bmatrix} W_1(c) \\ \vdots \\ W_N(c) \end{bmatrix}, \; W_i(c) = (W_{1i}(c), \Delta X_{1i}(c), W_{2i}(c), \Delta X_{2i}(c)), \; i = 1, ..., N$$

By employing the moment conditions, $E\left(W(c)'\Delta e\right) = 0$, we obtain a GMM estimator of θ (given a threshold parameter value, c) as:

$$\hat{\theta}(c) = \left[Z(c)'W(c)V(c)^{-1}W(c)'Z(c)\right]^{-1}\left[Z(c)'W(c)V(c)^{-1}W(c)'\Delta\ell\right]$$

The GMM theory suggests that an optimal (inverted) weighting matrix, $V(c)$, be given by the covariance matrix of the orthogonality conditions, $E\left(W(c)'\Delta e\right) = 0$.

Next, we derive the GMM estimator in two cases, with homoscedasticity or heteroscedasticity. First, if e_{it} is independent and has homoscedastic variance, σ^2, the GMM estimator can be simply computed in one step. The covariance matrix of $E\left(W(c)'\Delta e\right) = 0$ is given by:

$$E\left(W_i(c)'\Delta e_i \Delta e_i' W_i(c)\right) = W_i(c)'GW_i(c)$$

where G is a $(T-2)\times(T-2)$ fixed matrix with 2's on the main diagonal, -1's on the next sub-diagonals, and zeros otherwise. Thus, we obtain the one-step GMM estimator by:

$$\hat{\theta}_{GMM1}(c)=\left[Z(c)'W(c)V_{GMM1}(c)^{-1}W(c)'Z(c)\right]^{-1}\left[Z(c)'W(c)V_{GMM1}(c)^{-1}W(c)'\Delta\ell\right]$$

where $V_{GMM1}(c)=\sum_i W_i(c)'GW_i(c)$.

If e_{it} is heteroscedastic, the one-step GMM estimator is inefficient. In this more general case, we consider the following robust estimator of the covariance matrix:

$$V_{GMM2}(c)=\sum_i W_i(c)'\Delta\hat{e}_i\Delta\hat{e}_i'W_i(c)$$

where $\Delta\hat{e}_i=\Delta\ell_i(c)-Z_i(c)\hat{\theta}_{GMM1}(c)$ is the $(T-2)\times 1$ vector of residuals obtained from the one-step GMM estimation. We then obtain an efficient two-step GMM estimator by:

$$\hat{\theta}_{GMM2}(c)=\left[Z(c)'W(c)V_{GMM2}(c)^{-1}W(c)'Z(c)\right]^{-1}\left[Z(c)'W(c)V_{GMM2}(c)^{-1}W(c)'\Delta\ell\right]$$

Next, the threshold parameter, c, can be consistently estimated as:

$$\hat{c}=\arg\min_{c\in C}Q(c)$$

where C is the grid set and $Q(c)$ is the generalized distance measure:

$$Q(c)=\left\{\frac{1}{N}W(c)'\Delta\hat{e}(c)\right\}'\left\{\frac{1}{N}\hat{V}_{GMM2}(c)\right\}^{-1}\left\{\frac{1}{N}W(c)'\Delta\hat{e}(c)\right\}$$

where $\Delta\hat{e}(c)=\Delta\ell-Z(c)\hat{\theta}_{GMM2}(c)$. Since the model is linear in θ for each c, we use a practical grid search algorithm to find a consistent threshold estimate, $\hat{c}$, over a grid set that consists of the support of the transition variable, q. Following the literature, we use two cut-off points at the 15th and 85th percentiles to avoid potential extreme values of the transition variable while ensuring there is a sufficient number of observations in each regime.

Under the maintained assumption that the transition variable, q_{it}, is stationary and exogenous, the GMM estimators of $\theta(c)$ are asymptotically independent of the threshold estimate such that inference on θ can proceed as if $\hat{c}$ were the true value, e.g., Hansen (1999, 200) and (Caner and Hansen, 2004). Hence, it is easily seen that the asymptotic distribution of $\hat{\theta}_{GMM2}(c)$ is normal with the covariance matrix estimated by:

$$Var\left(\hat{\theta}_{GMM2}(c)\right)=\left[Z(c)'W(c)V_{GMM2}(c)^{-1}W(c)'Z(c)\right]^{-1}$$

4.7 Dynamic Panels with Threshold Effect and Endogeneity

Seo and Shin (2014) address an important and challenging issue as how best to model nonlinear asymmetric dynamics and cross-sectional heterogeneity, simultaneously, in the dynamic threshold panel data framework, in which both threshold variable and regressors are allowed to be endogenous. Depending on whether the threshold variable is strictly exogenous or not, we propose two different estimation methods: first-differenced two-step least squares and first-differenced GMM. The former exploits the fact that the threshold variable is strictly exogenous to achieve the super-consistency of the threshold estimator. We provide asymptotic distributions of both estimators. The bootstrap-based test for the presence of threshold effect as well as the exogeneity test of the threshold variable are also developed. Monte Carlo studies provide a support for our theoretical predictions. Finally, using the UK and the US company panel data, we provide two empirical applications investigating an asymmetric sensitivity of investment to cash flows and an asymmetric dividend smoothing.

4.7.1 The Model

Consider the following dynamic panel threshold regression model:

$$y_{it} = \left(1, x_{it}'\right) \phi_1 1\left(q_{it} \leq \gamma\right) + \left(1, x_{it}'\right) \phi_2 1\left(q_{it} > \gamma\right) + \varepsilon_{it},\ i = 1, ..., n;\ t = 1, ..., T, \tag{4.1}$$

where y_{it} is a scalar stochastic variable of interest, x_{it} is the $k_1 \times 1$ vector of time-varying regressors, that may include the lagged dependent variable, $1\left(\cdot\right)$ is an indicator function, and q_{it} is the transition variable. γ is the threshold parameter, and ϕ_1 and ϕ_2 the slope parameters associated with different regimes. The regression error, ε_{it} consists of the error components:

$$\varepsilon_{it} = \alpha_i + v_{it}, \tag{4.2}$$

where α_i is an unobserved individual fixed effect and v_{it} is a zero mean idiosyncratic random disturbance. In particular, v_{it} is assumed to be a martingale difference sequence,

$$\mathrm{E}\left(v_{it} | \mathcal{F}_{t-1}\right) = 0,$$

where $\mathcal{F}_t$ is a natural filtration at time t. It is worthwhile to mention that we do not assume x_{it} or q_{it} to be measurable with respect to $\mathcal{F}_{t-1}$, thus allowing endogeneity in both the regressor, x_{it} and the threshold variable, q_{it}. But, as will be shown, efficient estimation depends on whether q_{it} is exogenous or not. As we will consider the asymptotic experiment under large n with a fixed T, the martingale difference assumption is just for expositional simplicity. The sample is generated from random sampling across i.

A leading example of interest is the self-exciting threshold autoregressive (SETAR) model popularized by Tong (1990), in which case we have x_{it} consisting of the lagged y_{it}'s and $q_{it} = y_{i,t-1}$.

We allow for both "fixed threshold effect" and "diminishing or small threshold effect" for statistical inference for the threshold parameter, γ by defining (e.g. Hansen, 2000):

$$\delta = \delta_n = \delta_0 n^{-\alpha} \text{ for } 0 \leq \alpha < 1/2. \tag{4.3}$$

It is well-established in the linear dynamic panel data literature that the fixed effects estimator of the autoregressive parameters is biased downward (e.g. Nickell, 1981). To deal with the correlation of the regressors with individual effects in (4.1) and (4.2), we follow Arellano and Bond (1991) and consider the first-difference transformation of (4.1) as follows:

$$\Delta y_{it} = \beta' \Delta x_{it} + \delta' X_{it}' \mathbf{1}_{it}(\gamma) + \Delta \varepsilon_{it}, \tag{4.4}$$

where Δ is the first difference operator, $\underset{k_1 \times 1}{\beta} = \left(\phi_{12}, ..., \phi_{1,k_1+1}\right)'$, $\underset{(k_1+1)\times 1}{\delta} = \phi_2 - \phi_1$, and

$$\underset{2\times(1+k_1)}{X_{it}} = \begin{pmatrix} (1, x_{it}') \\ \left(1, x_{i,t-1}'\right) \end{pmatrix} \quad \text{and} \quad \underset{2\times 1}{\mathbf{1}_{it}(\gamma)} = \begin{pmatrix} 1(q_{it} > \gamma) \\ -1(q_{it-1} > \gamma) \end{pmatrix}.$$

Let $\theta = \left(\beta', \delta', \gamma\right)'$ and assume that θ belongs to a compact set, $\Theta = \Phi \times \Gamma \subset \mathbb{R}^k$, with $k = 2k_1 + 2$. It is worthwhile to note that the transformed model, (4.4) consists of 4 regimes, which are generated by two threshold variables, q_{it} and q_{it-1}. This change in the model characteristic is relevant in inference using the least squares estimation as discussed in Section 4.7.3.

The OLS estimator obtained from (4.4) is not unbiased since the transformed regressors are now correlated with $\Delta\varepsilon_{it}$. To fix this problem we need to find an $l \times 1$ vector of instrument variables, $\left(z_{it_0}', ..., z_{iT}'\right)'$ for $2 < t_0 \leq T$, such that either

$$\mathrm{E}\left(z_{it_0}' \Delta\varepsilon_{it_0}, ..., z_{iT}' \Delta\varepsilon_{iT}\right)' = 0, \tag{4.5}$$

or, for each $t = t_0, ..., T$,

$$\mathrm{E}\left(\Delta\varepsilon_{it} | z_{it}\right) = 0. \tag{4.6}$$

Notice that z_{it} may include lagged values of (x_{it}, q_{it}) and lagged dependent variables if not included in x_{it} or q_{it}. The number of instruments may be different for each time t.

4.7.2 Estimation

Depending upon whether q_{it} is endogenous or not and whether the conditional moment restriction (4.6) holds or not, we will develop different estimation methods.

FD-GMM

We allow for the threshold variable q_{it} to be endogenous, and develop a two-step GMM estimation. To this end we consider the $l \times 1$ vector of the sample moment conditions:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta),$$

where

$$g_i(\theta) = \begin{pmatrix} z_{it_0}\left(\Delta y_{it_0} - \beta'\Delta x_{it_0} - \delta' X'_{it_0}\mathbf{1}_{it_0}(\gamma)\right) \\ \vdots \\ z_{iT}\left(\Delta y_{iT} - \beta'\Delta x_{iT} - \delta' X'_{iT}\mathbf{1}_{iT}(\gamma)\right) \end{pmatrix}. \tag{4.7}$$

Also, let $g_i = g_i(\theta_0) = \left(z'_{it_0}\Delta\varepsilon_{it_0}, ..., z'_{iT}\Delta\varepsilon_{iT}\right)'$ and $\Omega = \mathrm{E}(g_i g'_i)$ where Ω is assumed to be finite and positive definite. For a positive definite matrix, W_n such that $W_n \xrightarrow{p} \Omega^{-1}$, let

$$\bar{J}_n(\theta) = \bar{g}_n(\theta)' W_n \bar{g}_n(\theta). \tag{4.8}$$

Then, the GMM estimator of θ is given by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \bar{J}_n(\theta). \tag{4.9}$$

Since the model is linear in ϕ for each γ and the objective function $\bar{J}_n(\theta)$ is not continuous in γ, the grid search algorithm is more practical. Let

$$\bar{g}_{1n} = \frac{1}{n}\sum_{i=1}^{n} g_{1i}, \quad \text{and} \quad \bar{g}_{2n}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} g_{2i}(\gamma),$$

where

$$\underset{l\times 1}{g_{1i}} = \begin{pmatrix} z_{it_0}\Delta y_{it_0} \\ \vdots \\ z_{iT}\Delta y_{iT} \end{pmatrix}, \quad \underset{l\times(k-1)}{g_{2i}(\gamma)} = \begin{pmatrix} z_{it_0}\left(\Delta x_{it_0}, \mathbf{1}_{it_0}(\gamma)' X_{it_0}\right) \\ \vdots \\ z_{iT}\left(\Delta x_{iT}, \mathbf{1}_{iT}(\gamma)' X_{iT}\right) \end{pmatrix}.$$

Then, the GMM estimator of β and δ, for a given γ, is given by

$$\left(\hat{\beta}(\gamma)', \hat{\delta}(\gamma)'\right)' = \left(\bar{g}_{2n}(\gamma)' W_n\, \bar{g}_{2n}(\gamma)\right)^{-1} \bar{g}_{2n}(\gamma)' W_n\, \bar{g}_{1n}.$$

Denoting the objective function evaluated at $\hat{\beta}(\gamma)$ and $\hat{\delta}(\gamma)$ by $\hat{J}_n(\gamma)$, we obtain the GMM estimator of θ by

$$\hat{\gamma} = \underset{\gamma\in\Gamma}{\operatorname{argmin}}\, \hat{J}_n(\gamma), \quad \text{and} \quad \left(\hat{\beta}', \hat{\delta}'\right)' = \left(\hat{\beta}(\hat{\gamma})', \hat{\delta}(\hat{\gamma})'\right)'.$$

The asymptotic property of the GMM estimator, $\hat{\gamma}$, which will be presented in Section 4.7.3, is different from the conventional least squares estimator, e.g. Chan (1993) and Hansen (2000).

The two-step optimal GMM estimator is obtained as follows:

1. Estimate the model by minimizing $\bar{J}_n(\theta)$ with either $W_n = I_l$ or

$$W_n = \begin{pmatrix} \frac{2}{n}\sum_{i=1}^n z_{it_0} z_{it_0}' & \frac{-1}{n}\sum_{i=1}^n z_{it_0} z_{it_0+1}' & 0 & \cdots \\ \frac{-1}{n}\sum_{i=1}^n z_{it_0+1} z_{it_0}' & \frac{2}{n}\sum_{i=1}^n z_{it_0+1} z_{it_0+1}' & \ddots & \ddots \\ 0 & \ddots & \ddots & \frac{-1}{n}\sum_{i=1}^n z_{iT-1} z_{iT}' \\ \vdots & \ddots & \frac{-1}{n}\sum_{i=1}^n z_{iT} z_{iT-1}' & \frac{2}{n}\sum_{i=1}^n z_{iT} z_{iT}' \end{pmatrix} \tag{4.10}$$

   and collect residuals, $\widehat{\Delta\varepsilon}_{it}$.

2. Estimate the parameter θ by minimizing $\bar{J}_n(\theta)$ with

$$W_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{g}_i \hat{g}_i' - \frac{1}{n^2}\sum_{i=1}^n \hat{g}_i \sum_{i=1}^n \hat{g}_i' \right)^{-1}, \tag{4.11}$$

   where $\hat{g}_i = \left(\widehat{\Delta\varepsilon}_{it_0} z_{it_0}', ..., \widehat{\Delta\varepsilon}_{iT} z_{iT}' \right)'$.

FD-2SLS

This subsection considers the case where the threshold variables, q_{it} and $q_{i,t-1}$ in (4.4), are exogenous and the conditional moment restriction (4.6) holds. That is, z_{it} includes q_{it} and $q_{i,t-1}$. In this case, we can improve upon the GMM estimator presented above. In particular, the threshold estimate, $\hat{\gamma}$ can achieve the efficient rate of convergence, as obtained in the classical regression model (e.g. Hansen, 2000), and the slope estimate, $\hat{\phi}$ can achieve the semi-parametric efficiency bound (Chamberlain, 1987) under conditional homoskedasticity as if the true threshold value, γ_0, is known. This strong result can be obtained since the two sets of estimators are asymptotically independent.

We consider two cases for the reduced form regression – the regression of endogenous regressors on the instrumental variables: the first type of the reduced form is a general non-linear regression where unknown parameters can be estimated by the standard $\sqrt{n}$ rate, and the second type is the threshold regression with a common threshold.

The second case was also considered by Caner and Hansen (2004), albeit in the single equation setup. Their approach consists of three steps; the first two steps yield an estimate of the threshold value and the third step performs the standard GMM for the linear regression within each subsample divided by the threshold. However, this split-sample GMM approach does not work with the panel data with a time varying threshold variable, q_{it}, because it generates multiple regimes with cross regime restrictions. Furthermore, their approach is not fully efficient. In this regard, we will develop a more efficient estimation algorithm for the threshold value below.

Reduced Form Here we consider general non-linear regressions for the reduced form and later provide the asymptotic variance formula that corrects the estimation error stemming from the reduced form regression. This is practically relevant since the linear projection in the reduced form invalidates the consistency of $\hat{\theta}$ when the structural form is the threshold regression, e.g. Yu (2013) .

The first-differenced model, (4.4) with the conditional moment condition, (4.6) and the exogeneity of q, implies the following regression of Δy_{it} on z_{it}:

$$\mathrm{E}\left(\Delta y_{it}|z_{it}\right) = \beta'\mathrm{E}\left(\Delta x_{it}|z_{it}\right) + \delta'\mathrm{E}\left(X_{it}'|z_{it}\right)\mathbf{1}_{it}\left(\gamma\right). \tag{4.12}$$

Assume that the reduced form regressions are given by, for each t,

$$\mathrm{E}\left(\begin{array}{c} 1, x_{it}' \\ 1, x_{it-1}' \end{array}|z_{it}\right) = \left(\begin{array}{c} 1, F_{1t}'\left(z_{it};b_{1t}\right) \\ 1, F_{2t}'\left(z_{it};b_{2t}\right) \end{array}\right) = \underset{2\times(1+k_1)}{F_t\left(z_{it};b_t\right)}, \tag{4.13}$$

where $b_t = \left(b_{1t}', b_{2t}'\right)'$ is an unknown parameter vector and F_t is a known function. Also let

$$H_t\left(z_{it};b_t\right) = \mathrm{E}\left(\Delta x_{it}|z_{it}\right) = F_{1t}\left(z_{it};b_t\right) - F_{2t}\left(z_{it};b_t\right).$$

For instance, Caner and Hansen (2004) consider the linear regression and the threshold regression for F_t. If $x_{it-1} \in z_{it}$, then $F_{2t} = x_{1t-1}$.

Note that there are two regressions for x_{it} due to the first difference transformation and the possibility that z_{it} varies over time. Furthermore, it is not sufficient to consider the regression $\mathrm{E}\left(\Delta x_{it}|z_{it}\right)$ only, due to the threshold effect in the structural form (4.12) .

The above representation in (4.12) and (4.13) motivates the following two-step estimation procedure:

1. For each t, estimate the reduced form, (4.13) by the least squares, and obtain the parameter estimates, $\hat{b}_t,\ t = t_0, ..., T$, and the fitted values, $\hat{F}_{it} = F_t\left(z_{it};\hat{b}_t\right)$.

2. Estimate θ by

$$\min_{\theta\in\Theta}\hat{\mathbb{M}}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=t_0}^{T} e_{it}\left(\theta,\hat{b}_t\right)^2, \tag{4.14}$$

   where

$$e_{it}\left(\theta,b_t\right) = \Delta y_{it} - \beta' H_t\left(z_{it};b_t\right) - \delta' F_t\left(z_{it};b_t\right)'\mathbf{1}_{it}\left(\gamma\right).$$

   This step can be done simply by the grid search as the model is linear in β and δ for a fixed γ. Thus, $\hat{\beta}\left(\gamma\right)$ and $\hat{\delta}\left(\gamma\right)$ can be obtained from the pooled OLS of Δy_{it} on $\hat{H}_{it}$ and $\hat{F}_{it}'\mathbf{1}_{it}\left(\gamma\right)$, and $\hat{\gamma}$ is defined as the minimizer of the profiled sum of squared errors, $\hat{\mathbb{M}}_n\left(\gamma\right)$.

This procedure produces a rate-optimal estimator for γ, implying that β and δ can be estimated as if γ_0 were known. In the special case with $T = t_0$, we end up estimating a linear regression model with a conditional moment restriction. The above two-step estimation yields the optimal estimate for β and δ provided that the model is conditionally homoskedastistic, i.e., $\mathrm{E}\left(\Delta\varepsilon_{it}^2|z_{it}\right) = \sigma^2$, see Chamberlain (1987). While it requires to estimate the conditional heteroskedasticity to fully exploit the implications of the conditional moment restriction, (4.6) under more general setup, it is reasonable to employ our two-step estimator and robustify the standard errors for the heteroskedasticity. We will provide a heteroskedasticity-robust standard error for $\hat{\beta}$ and $\hat{\delta}$. Further, the standard error is corrected for the estimation error in the first step estimation of b.

Threshold Regression in Reduced Form Suppose that z_{it} includes 1 and x_{it-1}, 1 being the first element of z_{it}, and

$$\begin{aligned} x_{it} &= \Gamma_{1t} z_{it} 1\{q_{it} \leq \gamma\} + \Gamma_{2t} z_{it} 1\{q_{it} > \gamma\} + \eta_{it}, \\ \mathrm{E}\left(\eta_{it}|z_{it}\right) &= 0. \end{aligned} \tag{4.15}$$

This implies that

$$\begin{aligned} \Delta y_{it} &= \lambda_{1t}' z_{it} 1\{q_{it} \leq \gamma\} + \lambda_{2t}' z_{it} 1\{q_{it} > \gamma\} \\ &\quad -\lambda_{3t}' z_{it} 1\{q_{it-1} \leq \gamma\} - \lambda_{4t}' z_{it} 1\{q_{it-1} > \gamma\} + e_{it}, \\ \mathrm{E}\left(e_{it}|z_{it}\right) &= 0. \end{aligned} \tag{4.16}$$

The parameters are subject to the constraints: $\lambda_{1t}' = \left(0, \beta'\Gamma_{1t}\right)$, $\lambda_{2t}' = \left(\delta_1, \phi_{22}'\Gamma_{2t}\right)$, $\lambda_{3t}' z_{it} = \beta' x_{it-1}$, and $\lambda_{4t}' z_{it} = \phi_{22}' x_{it-1} - \delta_1$. Also, $e_{it} = \Delta\varepsilon_{it} + \eta_{it}'\left(\beta + 1\{q_{it} > \gamma\}\delta_2\right)$. Since the estimates of λ and γ are asymptotically independent, we do not impose these constraints on λ to estimate γ to simplify the exposition.

Thus, we estimate the model as follows:

1. Estimate γ by the pooled least square of (4.16), which can be done by the grid search,[7] and denote the estimate by $\tilde{\gamma}$.

2. Fix γ at $\tilde{\gamma}$ and estimate Γ_{jt}, $j = 1, 2$, in (4.15) by the OLS, for each t.

3. Estimate β and δ in (4.12) by the OLS with γ and the reduced form parameters fixed at the estimates obtained from the preceding steps. Denote these estimates by $\tilde{\beta}$ and $\tilde{\delta}$.

[7] That is, fix γ and obtain $\tilde{e}_{it}(\gamma)$ and $\tilde{\lambda}_{jt}(\gamma)$, $j = 1, ..., 4$ by the OLS for each t. Then, $\tilde{\gamma}$ is the minimizer of the profiled sum of squared errors, $\sum_{i,t} \tilde{e}_{it}^2(\gamma)$ and $\tilde{\lambda}_{jt} = \tilde{\lambda}_{jt}(\tilde{\gamma})$, $j = 1, ..., 4$.

Remark 2 *Our approach is different from that of Caner and Hansen, who estimate the threshold parameter separately in the reduced and the structural form. Their approach introduces dependence between the separate threshold estimates, which invalidates their asymptotic distribution.*[8] *Intuitively, the estimation error in the first step affects the second step estimation of γ since the true thresholds are restricted to be the same in both reduced and structural forms.*

4.7.3 Asymptotic Distributions

This section develops asymptotic theories for the estimators presented in the previous section. There are two frameworks in the literature. One is the diminishing threshold assumption (Hansen, 2000) and the other the fixed threshold assumption (Chan, 1993). For the GMM estimator we present the asymptotics that accommodates both setups and for the 2SLS we develop the asymptotic distribution only under Hansen's framework. We also discuss the estimation of unknown quantities in the asymptotic distributions such as the asymptotic variances and the normalizing factors when an estimator is not asymptotically normal.

FD-GMM

Partition $\theta = \left(\theta_1', \gamma\right)'$, where $\theta_1 = \left(\beta', \delta'\right)'$. As the true value of δ is δ_n, the true values of θ and θ_1 are denoted by θ_n and θ_{1n}, respectively. And define

$$\underset{l\times k_1}{G_\beta} = \begin{bmatrix} -\mathrm{E}\left(z_{it_0}\Delta x_{it_0}'\right) \\ \vdots \\ -\mathrm{E}\left(z_{iT}\Delta x_{iT}'\right) \end{bmatrix}, \quad \underset{l\times(k_1+1)}{G_\delta}(\gamma) = \begin{bmatrix} -\mathrm{E}\left(z_{it_0}\mathbf{1}_{it_0}(\gamma)' X_{it_0}\right) \\ \vdots \\ -\mathrm{E}\left(z_{iT}\mathbf{1}_{iT}(\gamma)' X_{iT}\right) \end{bmatrix},$$

and

$$\underset{l\times 1}{G_\gamma(\gamma)} = \begin{bmatrix} \left\{\mathrm{E}_{t_0-1}\left[z_{it_0}(1, x_{it_0-1})'\,|\gamma\right] p_{t_0-1}(\gamma) - \mathrm{E}_{t_0}\left[z_{it_0}(1, x_{it_0})'\,|\gamma\right] p_{t_0}(\gamma)\right\}\delta_0 \\ \vdots \\ \left\{\mathrm{E}_{T-1}\left[z_{iT}(1, x_{iT-1})'\,|\gamma\right] p_{T-1}(\gamma) - \mathrm{E}_{T}\left[z_{iT}(1, x_{iT})'\,|\gamma\right] p_{T}(\gamma)\right\}\delta_0 \end{bmatrix},$$

where $\mathrm{E}_t[\cdot|\gamma]$ stands for the conditional expectation given $q_{it} = \gamma$ and $p_t(\cdot)$ denotes the density of q_{it}.

The true value of β is fixed at β_0 while that of δ depends on n, for which we write $\delta_n = \delta_0 n^{-\alpha}$ for some $0 \leq \alpha < 1/2$ and $\delta_0 \neq 0$. θ_n are interior points of Θ. Furthermore, Ω is finite and positive definite.

[8] Lemma 1 in Caner and Hansen (2004) requires more restrictions. More specifically, their (A.7) is true only when the threshold estimate is n-consistent, which is not the case in the maintained diminishing threshold parameter setup. Accordingly, the high-level assumption (17) in their Assumption 2 is no longer satisfied.

This is a standard assumption for the threshold regression model as in Hansen (2000).

(i) The threshold variable, q_{it} has a continuous and bounded density, p_t such that $p_t(\gamma_0) > 0$ for all $t = 1, ..., T$; (ii) $\mathrm{E}_t\left(z_{it}\left(x'_{it}, x'_{i,t-1}\right)|\gamma\right)$ is continuous at γ_0, where $\mathrm{E}_t(\cdot|\gamma) = \mathrm{E}(\cdot|q_{it} = \gamma)$ and $\mathrm{E}_t\left(z_{it}\left(x'_{it}, x'_{i,t-1}\right)|\gamma\right)\delta_0 \neq 0$ for some t.

The smoothness assumption on the distribution of the threshold variable and some conditional moments are standard. However, we do not require the discontinuity of the regression function at the change point. In other words, the distribution of GMM estimator of the unknown threshold is invariant to the continuity of the regression function at the change point. This is a novel feature of the GMM. Heuristically, the GMM criterion function can be viewed as an extreme form of smoothing in the sense of Seo and Linton (2007). As a consequence, we do not need a prior knowledge on the continuity of the model to make inference for the threshold model.

Let $G = (G_\beta, G_\delta(\gamma_0), G_\gamma(\gamma_0))$. Then, assume that G is of the full column rank.

This is a standard rank condition in GMM. Then, we have:

Theorem 10 *Under Assumptions 4.7.3-4.7.3, as* $n \to \infty$,

$$\begin{pmatrix} \sqrt{n}\begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\delta} - \delta_n \end{pmatrix} \\ n^{1/2-\alpha}(\hat{\gamma} - \gamma_0) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right).$$

The asymptotic variance matrix contains δ_0, and the convergence rate of $\hat{\gamma}$ hinges on the unknown quantity, α. These two quantities cannot be consistently estimated in separation, but they cancel out in the construction of t-statistic. Thus, confidence intervals for θ can be constructed in the standard manner. Let

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}'_i - \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}'_i\right),$$

where $\hat{g}_i = g_i\left(\hat{\theta}\right)$ and

$$\hat{G}_\beta = \begin{bmatrix} -\frac{1}{n}\sum_{i=1}^{n} z_{it_0}\Delta x'_{it_0} \\ \vdots \\ -\frac{1}{n}\sum_{i=1}^{n} z_{iT}\Delta x'_{iT} \end{bmatrix}, \quad \hat{G}_\delta = \begin{bmatrix} -\frac{1}{n}\sum_{i=1}^{n} z_{it_0}\mathbf{1}_{it_0}(\hat{\gamma})' X_{it_0} \\ \vdots \\ -\frac{1}{n}\sum_{i=1}^{n} z_{iT}\mathbf{1}_{iT}(\hat{\gamma})' X_{iT} \end{bmatrix}.$$

Then, G_γ may be estimated by the standard Nadaraya-Watson kernel estimator: that is, for some kernel K and bandwidth h (e.g. the Gaussian

kernel and Silverman's rule of thumb), let

$$\hat{G}_\gamma = \begin{bmatrix} \frac{1}{nh}\sum_{i=1}^{n} z_{it_0}\left[(1, x_{it_0-1})' K\left(\frac{\hat{\gamma}-q_{it_0-1}}{h}\right) - (1, x_{it_0})' K\left(\frac{\hat{\gamma}-q_{it_0}}{h}\right)\right]\hat{\delta} \\ \vdots \\ \frac{1}{nh}\sum_{i=1}^{n} z_{iT}\left[(1, x_{iT-1})' K\left(\frac{\hat{\gamma}-q_{iT-1}}{h}\right) - (1, x_{iT})' K\left(\frac{\hat{\gamma}-q_{iT}}{h}\right)\right]\hat{\delta} \end{bmatrix}. \tag{4.17}$$

See Hardle and Linton (1994) for more detailed discussion on the choice of kernel K and bandwidth h.

Furthermore, let $\hat{V}_s = \hat{\Omega}^{-1/2}\left(\hat{G}_\beta, \hat{G}_\delta\right)$ and $\hat{V}_\gamma = \hat{\Omega}^{-1/2}\hat{G}_\gamma$. Then, the asymptotic variance-covariance matrix for the regression coefficient, $\theta_1 = \left(\beta', \delta'\right)'$ can be consistently estimated by

$$\left(\hat{V}_s'\hat{V}_s - \hat{V}_s'\hat{V}_\gamma\left(\hat{V}_\gamma'\hat{V}_\gamma\right)^{-1}\hat{V}_\gamma'\hat{V}_s\right)^{-1},$$

while the t-statistic for $\gamma = \gamma_0$ defined by

$$t = \frac{\sqrt{n}\left(\hat{\gamma} - \gamma_0\right)}{\hat{V}_\gamma'\hat{V}_\gamma - \hat{V}_\gamma'\hat{V}_s\left(\hat{V}_s'\hat{V}_s\right)^{-1}\hat{V}_s'\hat{V}_\gamma},$$

converges to the standard normal distribution. Therefore, the confidence intervals can be constructed as in the standard GMM case.

Alternatively, the standard nonparametric bootstrap, which resamples across i with replacement, can be employed to construct the confidence intervals. See Section 4.7.4 for further details.

FD-2SLS

This section presents the asymptotic theory for the 2SLS estimator of θ. A few technical issues arise in the two-step estimation in the panel data such as the multiple threshold variables, which is a consequence of the first difference transformation. We begin with the case where the reduced form is the regular nonlinear regression and the reduced form parameter estimates are asymptotically normal. Next, we consider the case where the reduced form also follows the threshold regression.

Since some elements of x_{it} may belong to z_{it}, in which case the reduced form is identity, and some elements of $E\left(x_{it}|z_{it}\right)$ may be identical to $E\left(x_{it}|z_{it+1}\right)$ for some t, we collect all distinct reduced form regression functions, F_t, $t = t_0, ..., T$, that are not identities, and denote it as $F\left(z_i, b\right)$, where z_i and b are the collections of all distinct elements of z_{it} and b_t, $t = t_0, ..., T$. Accordingly, we denote the collection of the corresponding elements of x_{it}'s by $\S_i$, and write the reduced form as the multivariate cross

section regression as follows:

$$\begin{aligned} \S_i &= F(z_i, b) + \eta_i, \\ \mathrm{E}(\eta_i | z_i) &= 0. \end{aligned} \tag{4.18}$$

Let $\hat{b}$ denote the least squares estimate, and we follow the convention that $F_i(b) = F(z_i, b)$, $F_i = F(z_i, b_0)$, $\hat{F}_i = F\left(z_i, \hat{b}\right)$, etc, where b_0 indicates the true value of b, when there is no confusion. We consider two cases explicitly. The first case is where $\hat{b}$ is asymptotically normal and the second is the threshold regression.

Reduced Form This section considers the reduced forms, which allow for stochastic linearization and thus the asymptotic normality of reduced form parameter estimates. We directly assume the asymptotic normality of $\hat{b}$ and the existence of a matrix-valued influence function, $\mathbb{F}$ below. More primitive conditions to yield this asymptotic normality of $\hat{b}$ are provided in the Appendix. Notice that $|A|$ denotes the Euclidean norm if A is a vector, and the vector induced norm if A is a matrix.

There exists a matrix-valued function, $\mathbb{F}(z_i, b)$ such that $\mathrm{E}|\mathbb{F}_i|^2 < \infty$ and

$$\sqrt{n}\left(\hat{b} - b_0\right) = \left(\mathrm{E}\mathbb{F}_i\mathbb{F}_i'\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{F}_i \eta_i + o_p(1).$$

We begin with this high-level assumption because our main goal is to illustrate how the estimation error in the first step affects the asymptotic distribution of the estimator of the regression coefficients, β and δ and of the threshold parameter, γ in the second step. We introduce some more notations. Recall the functions introduced in Section 4.7.2 and let

$$\underset{(2k_1+1)\times 1}{\Xi_{it}(\gamma, b_t)} = \begin{bmatrix} H_{it}(b_t) \\ F_{it}(b_t)' \mathbf{1}_{it}(\gamma) \end{bmatrix},$$

for each t, and

$$\underset{(2k_1+1)\times(T-t_0+1)}{\Xi_i(\gamma, b)} = \left(\Xi_{it_0}(\gamma, b_{t_0}), ..., \Xi_{iT}(\gamma, b_T)\right).$$

Also, let e_i be the vector stacking $\left\{\Delta\varepsilon_{it} + \beta_0'\left(\Delta x_{it} - \mathrm{E}\left(\Delta x_{it} | z_{it}\right)\right)\right\}_{t=t_0}^{T}$. Then, define

$$\underset{(2k_1+1)\times(2k_1+1)}{M_1(\gamma)} = \mathrm{E}\left[\Xi_i(\gamma)\Xi_i(\gamma)'\right], \quad \text{and} \quad \underset{(2k_1+1)\times(2k_1+1)}{V_1(\gamma)} = A(\gamma)\Omega(\gamma, \gamma)A(\gamma)',$$

where

$$\begin{aligned} \underset{((2k_1+1)+k_b)\times((2k_1+1)+k_b)}{\Omega(\gamma_1, \gamma_2)} &= \mathrm{E}\left[\begin{pmatrix} \Xi_i(\gamma_1) e_i, \\ \mathbb{F}_i \eta_i \end{pmatrix} \left(e_i' \Xi_i'(\gamma_2), \eta_i' \mathbb{F}_i'\right)\right], \\ \underset{(2k_1+1)\times((2k_1+1)+k_b)}{A(\gamma)} &= \left(I_{(2k_1+1)}, \ -\mathrm{E}\left[\frac{\partial}{\partial b'} \sum_{t=t_0}^{T} \left(H_{it}'\beta_0\right) \Xi_{it}(\gamma)\right] \left(\mathrm{E}\mathbb{F}_i\mathbb{F}_i'\right)^{-1}\right). \end{aligned}$$

For the asymptotic distribution of $\hat{\gamma}$, we introduce:

$$\begin{aligned} M_2(\gamma) &= \sum_{t=t_0}^{T} \left[\mathrm{E}_t \left[\left(\left(1, F_{1,it}'\right) \delta_0 \right)^2 |\gamma \right] p_t(\gamma) + \mathrm{E}_{t-1} \left[\left(\left(1, F_{2,it}'\right) \delta_0 \right)^2 |\gamma \right] p_{t-1}(\gamma) \right], \\ V_2(\gamma) &= \sum_{t=t_0}^{T} \left(\mathrm{E}_t \left[\left(e_{it} \left(1, F_{1,it}'\right) \delta_0 \right)^2 |\gamma \right] p_t(\gamma) + \mathrm{E}_{t-1} \left[\left(e_{it} \left(1, F_{2,it}'\right) \delta_0 \right)^2 |\gamma \right] p_{t-1}(\gamma) \right) \\ &\quad + 2 \sum_{t=t_0}^{T-1} \mathrm{E}_t \left[e_{it} e_{it+1} \left(1, F_{1,it}'\right) \delta_0 \left(1, F_{2,it+1}'\right) \delta_0 |\gamma \right] p_t(\gamma). \end{aligned}$$

Following the convention, we write $V_j = V_j(\gamma_0)$ and $M_j = M_j(\gamma_0)$ for $j = 1, 2$.

We further assume:

The true value of β is fixed at β_0 while that of δ depends on n, for which we write $\delta_n = \delta_0 n^{-\alpha}$ for some $0 < \alpha < 1/2$ and $\delta_0 \neq 0$.

If $\alpha = 0$, the asymptotic distribution for $\hat{\gamma}$ is different from the one obtained here. However, the convergence rate result in the proof of the theorem is still valid.

(i) The threshold variable, q_{it} has a continuous and bounded density, p_t, such that $p_t(\gamma_0) > 0$ for all $t = 1, ..., T$; (ii) $\mathrm{E}_t(w_{it}|\gamma)$ is continuous at γ_0 for all t, and non-zero for some t, where w_{it} is either $\left(e_{it} \left(1, F_{1,it}'\right) \delta_0 + e_{it+1} \left(1, F_{2,it+1}'\right) \delta_0 \right)^2$, $\left(\left(1, F_{1,it}'\right) \delta_0 \right)^2$, or $\left(\left(1, F_{2,it}'\right) \delta_0 \right)^2$.

For some $\epsilon > 0$ and some $\zeta > 0$, $\mathrm{E}\left(\sup_{t \leq T, |b - b_0| < \epsilon} |e_{it} F_t(z_{it}, b_t)|^{2+\zeta} \right) < \infty$. For all $\epsilon > 0$, $\mathrm{E}\left(\sup_{t \leq T, |b - b_0| < \epsilon} |e_{it} (F_t(z_{it}, b_t) - F_t(z_{it}))|^{2+\zeta} \right) = O\left(\epsilon^{2+\zeta}\right)$.

The minimum eigenvalue of the matrix, $\mathrm{E}\Xi_{it}(\gamma) \Xi_{it}'(\gamma)$ is bounded below by a positive value for all $\gamma \in \Gamma$ and $t = 1, ..., T$.

The asymptotic confidence intervals can be constructed by inverting a test statistic. In particular, Hansen (2000) advocates the LR inversion for the construction of confidence intervals for the threshold value, γ_0, for which we define the LR statistic as

$$LR_n(\gamma) = n \frac{\hat{\mathbb{M}}_n(\gamma) - \hat{\mathbb{M}}_n(\hat{\gamma})}{\hat{\mathbb{M}}_n(\hat{\gamma})}.$$

Then, we present the main asymptotic results for the 2SLS estimator and the LR statistic in the following Theorem:

Theorem 11 *Let Assumptions 4.7.3-4.7.3 hold. Then,*

$$\sqrt{n} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\delta} - \delta_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, M_1^{-1} V_1 M_1^{-1}\right), \tag{4.19}$$

and

$$n^{1-2\alpha}\frac{M_2^2}{V_2}\left(\hat{\gamma}-\gamma_0\right)\xrightarrow{d}\underset{r\in\mathbb{R}}{\operatorname{argmin}}\left(\frac{|r|}{2}-W\left(r\right)\right), \tag{4.20}$$

where $W\left(r\right)$ *is a two-sided standard Brownian motion and it is independent of the limit variate in* (4.19)*. Furthermore,*

$$\frac{M_2\sigma_e^2}{V_2}\ LR\left(\gamma_0\right)\xrightarrow{d}\inf_{r\in\mathbb{R}}\left(|r|-2W\left(r\right)\right),$$

where $\sigma_e^2=\mathrm{E}\left(e_{it}^2\right)$*.*

Note that the first step estimation error does not affect the asymptotic distribution of $\hat{\gamma}$, while it contributes the asymptotic variance of $\hat{\beta}$ and $\hat{\delta}$ through Ω. Estimation of the asymptotic variances of $\hat{\beta}$ and $\hat{\delta}$ is standard, i.e. the same as in the linear regression due to the asymptotic independence. The asymptotic distribution for $\hat{\gamma}$ in (4.20) is symmetric around zero and has a known distribution function,

$$1+\sqrt{x/2\pi}\exp\left(-x/8\right)+\left(3/2\right)\exp\left(x\right)\Phi\left(-3\sqrt{x}/2\right)-\left(\left(x+5\right)/2\right)\Phi\left(\sqrt{x}/2\right),$$

for $x\geq 0$, where Φ is the standard normal distribution function. See Bhattacharya and Brockwell (1976). The unknown normalizing factor $n^{2\alpha}V_2^{-1}M_2^2$ can be estimated by $\hat{V}_2^{-1}\hat{M}_2^2$, where

$$\begin{aligned}
\hat{M}_2 &= \sum_{t=t_0}^{T}\frac{1}{nh}\sum_{i=1}^{n}\left[\left(\left(1,\hat{F}_{1,it}'\right)\hat{\delta}\right)^2 k\left(\frac{q_{it}-\hat{\gamma}}{h}\right)+\left(\left(1,\hat{F}_{2,it}'\right)\hat{\delta}\right)^2 k\left(\frac{q_{it-1}-\hat{\gamma}}{h}\right)\right],\\
\hat{V}_2 &= \sum_{t=t_0}^{T}\frac{1}{nh}\sum_{i=1}^{n}\left(\left(\hat{e}_{it}\left(1,\hat{F}_{1,it}'\right)\hat{\delta}\right)^2 k\left(\frac{q_{it}-\hat{\gamma}}{h}\right)+\left(\hat{e}_{it}\left(1,\hat{F}_{2,it}'\right)\hat{\delta}\right)^2 k\left(\frac{q_{it-1}-\hat{\gamma}}{h}\right)\right)\\
&\quad +2\sum_{t=t_0}^{T-1}\frac{1}{nh}\sum_{i=1}^{n}\hat{e}_{it}\hat{e}_{it+1}\left(1,\hat{F}_{1,it}'\right)\hat{\delta}\left(1,\hat{F}_{2,it+1}'\right)\hat{\delta}k\left(\frac{q_{it}-\hat{\gamma}}{h}\right).
\end{aligned}$$

The normalization factor, $V_2^{-1}M_2\sigma_e^2$ for the LR statistic can be estimated by $\hat{V}_2^{-1}\hat{M}_2\hat{\sigma}_e^2$, where $\hat{\sigma}_e^2=\left(n\left(T-t_0+1\right)\right)^{-1}\sum_{i=1}^{n}\sum_{t=t_0}^{T}\hat{e}_{it}^2$. Notice that it becomes 1 under the leading case of conditional homoskedasticity and the martingale difference sequence assumption for e_{it}. Hansen (2000) provides the distribution function of the asymptotic distribution of the LR_n statistic, which is $\left(1-e^{-x/2}\right)^2$.

Threshold Regression in Reduced Form Now, consider the case where the reduced form is a threshold regression, (4.15). The estimator, $\hat{\theta}$ is obtained from the three-step procedure following (4.15). Despite the difference in the estimation procedure, the asymptotic distributions of $\hat{\theta}$ can be presented by a slight modification of Theorem 11. In particular, the reduced form regression (4.18) in the beginning of Section 4.7.3 is characterized by the regression (4.15) given in Section 4.7.2.

Corollary 12 *Let Assumption 4.7.3 hold and* $\lambda_j = \left(\lambda'_{jt_0}, ..., \lambda'_{jT}\right)'$, $j = 1, ..., 4$. *Assume that* $\lambda_1 - \lambda_2 = n^{-\alpha}\delta_1$ *for some non-zero vector* δ_1. *Assumptions, 4.7.3 and 4.7.3, hold with* $F_{1,it} = \Gamma_{1t} z_{it} 1\{q_{it} \leq \gamma\} + \Gamma_{2t} z_{it} 1\{q_{it} > \gamma\}$ *and* $F_{2,it} = x_{it-1}$. *Furthermore, assume that* $\mathrm{E}|z_{it}|^4 < \infty$ *and* $\mathrm{E}e_{it}^4 < \infty$. *Then, the asymptotic distribution of* $\tilde{\theta}$ *is the same as in Theorem 11.*

4.7.4 Testing

Testing for Linearity

The preceding asymptotic results provide ways to make inference for unknown parameters and their functions. However, it is well-established that the test for linearity or threshold effects requires us to develop the different asymptotic theory due to the presence of unidentified parameters under the null hypothesis (e.g. Davies, 1977). Specifically, we consider the null hypothesis of interest as

$$\mathcal{H}_0 : \delta_0 = 0, \quad \text{for any } \gamma \in \Gamma, \tag{4.21}$$

against the alternative

$$\mathcal{H}_1 : \delta_0 \neq 0, \quad \text{for some } \gamma \in \Gamma.$$

Then, a natural test statistic for the null hypothesis, $\mathcal{H}_0$ is

$$\mathrm{supW} = \sup_{\gamma \in \Gamma} W_n(\gamma),$$

where $W_n(\gamma)$ is the standard Wald statistic for each fixed γ, that is,

$$W_n(\gamma) = n\hat{\delta}(\gamma)' \hat{\Sigma}_\delta(\gamma)^{-1} \hat{\delta}(\gamma),$$

where $\hat{\delta}(\gamma)$ is the estimate of δ, given γ by either the FD-GMM or FD-2SLS, and $\hat{\Sigma}_\delta(\gamma)$ is the corresponding consistent asymptotic variance estimator for $\hat{\delta}(\gamma)$. In the FD-GMM case, we employ $\hat{\Sigma}_\delta(\gamma) = R\left(\hat{V}_s(\gamma)\hat{V}_s(\gamma)\right)^{-1} R'$, where $\hat{V}_s(\gamma)$ is computed as in Section 4.7.3 with $\hat{\gamma} = \gamma$ and $R = \left(\mathbf{0}_{(k_1+1)\times k_1}, I_{k_1+1}\right)$. In the FD-2SLS case, we can simply use the same formula for the estimation of the asymptotic variance of $\hat{\delta}(\gamma)$ since the estimation error in γ does not affect the estimation of δ. The supremum type statistic is an application of the union-intersection principle commonly used in the literature, e.g. Hansen (1996), and Lee et al. (2011).

The limiting distribution of supW depends on the associated estimation methods. If δ were estimated by FD-2SLS, as is well-known in the literature, the limit is the supremum of the square of a Gaussian process with some unknown covariance kernel, yielding non-pivotal asymptotic distribution. In case of the FD-GMM, the Gaussian process is given by a simpler covariance kernel, though it seems not easy to pivotalize the statistic.

Theorem 13 (*i*) *Consider the FD-GMM estimation. Let* $G(\gamma) = (G_\beta, G_\delta(\gamma))$ *and* $D(\gamma) = G(\gamma)' \Omega^{-1} G(\gamma)$. *Suppose that* $\inf_{\gamma\in\Gamma} \det(D(\gamma)) > 0$ *and Assumption 4.7.3* (*i*) *holds. Then, under the null* (4.21), *we have*

$$\text{supW} \xrightarrow{d} \sup_{\gamma\in\Gamma} Z' G(\gamma)' D(\gamma)^{-1} R' \left[RD(\gamma)^{-1} R' \right]^{-1} RD(\gamma)^{-1} G(\gamma) Z,$$

where $Z \sim \mathcal{N}\left(0, \Omega^{-1}\right)$.
(*ii*) *Consider the 2SLS estimation. Suppose that Assumptions,* **??**, **??**, *4.7.3*(*i*), *4.7.3, and 4.7.3, hold. Then, under the null* (4.21),

$$\text{supW} \xrightarrow{d} \sup_{\gamma\in\Gamma} B(\gamma)' M_1(\gamma)^{-1} R' \left[RM_1(\gamma)^{-1} V_1(\gamma) M_1(\gamma)^{-1} R' \right]^{-1} RM_1(\gamma)^{-1} B(\gamma),$$

where $B(\gamma)$ *is a mean-zero Gaussian process with the covariance kernel,* $A(\gamma_1)\Omega(\gamma_1,\gamma_2)A(\gamma_2)'$.

When the reduced form is also a threshold regression, however, our test can be performed based on the model, (4.16). A null model in this case might be that both reduced form and the structural equations are linear for all t; that is,

$$\mathcal{H}_0' : \lambda_{1t} - \lambda_{2t} = \lambda_{3t} - \lambda_{4t} = 0, \quad \text{for all } \gamma \in \Gamma \text{ and } t = t_0, ..., T. \tag{4.22}$$

Repeating the discussion therein, the model, (4.16) is estimated by the pooled OLS for each γ and as such the construction of supW statistic is standard (e.g. Hansen, 1996).

These limiting distributions are not asymptotically pivotal and critical values cannot be tabulated. We bootstrap or simulate the asymptotic critical values or p-values, see Hansen (1996) for the latter. Here we describe the bootstrap procedure in details.

Let $\hat{\theta}$ be either the FD-GMM or the FD-2SLS estimator and construct

$$\widehat{\Delta\varepsilon_{it}} = \Delta y_{it} - \Delta x_{it}'\hat{\beta} - \hat{\delta}' X_{it}' \mathbf{1}_{it}(\hat{\gamma}),$$

for $i = 1, ..., n$, and $t = t_0, ..., T$. Then,

1. Let i^* be a random draw from $\{1, ..., n\}$, and $X_{it}^* = X_{i^*t}$, $q_{it}^* = q_{i^*t}$, $z_{it}^* = z_{i^*t}$ and $\Delta\varepsilon_{it}^* = \widehat{\Delta\varepsilon_{i^*t}}$. Then, generate
$$\Delta y_{it}^* = \Delta x_{it}^{*\prime}\hat{\beta} + \Delta\varepsilon_{it}^* \quad \text{for} \quad t = t_0, ..., T.$$
2. Repeat step 1 n times, and collect $\{(\Delta y_{it}^*, X_{it}^*, q_{it}^*, z_{it}^*) : i = 1, ..., n; t = t_0, ..., T\}$.
3. Construct the supW statistic, say supW*, from the bootstrap sample using the same estimation method for $\hat{\theta}$.
4. Repeat steps 1-3 B times, and evaluate the bootstrap p-values by the frequency of supW* that exceeds the sample statistic, supW.

Note that when simulating the bootstrap samples, the null hypothesis is imposed in step 1.

Testing for Exogeneity

In this section we describe how to test for the exogeneity of the threshold variable. Recently, Kapetanios (2010) develops the exogeneity test of the regressors in threshold regression models, following the general principle of the Hausman (1978) test (e.g. Pesaran et al., 1999). Similarly, we can develop the Hausman type testing procedure for the validity of the null hypothesis that the threshold variable is exogenous. Indeed, this is a straightforward by-product obtained by combining FD-GMM and FD-2SLS estimation methods and their asymptotic results.

Specifically, we propose the following t-statistic for the null hypothesis that GMM estimate of the unknown threshold, $\hat{\gamma}_{GMM}$, is equal to the 2SLS estimate, $\hat{\gamma}_{2SLS}$:

$$t_H = \frac{\sqrt{n}\left(\hat{\gamma}_{GMM} - \hat{\gamma}_{2SLS}\right)}{\hat{V}_\gamma'\hat{V}_\gamma - \hat{V}_\gamma'\hat{V}_s\left(\hat{V}_s'\hat{V}_s\right)^{-1}\hat{V}_s'\hat{V}_\gamma},$$

where the denominator is derived as in Section 4.1. Notice that

$$\hat{\gamma}_{2SLS} = \gamma_0 + o_p\left(n^{-1/2}\left(\hat{V}_\gamma'\hat{V}_\gamma - \hat{V}_\gamma'\hat{V}_s\left(\hat{V}_s'\hat{V}_s\right)^{-1}\hat{V}_s'\hat{V}_\gamma\right)\right)$$

due to its super-consistency. Then, it is easily seen that the asymptotic distribution of the t-statistic is the standard normal under the null hypothesis of strict exogeneity of the threshold variable, q_{it}.

4.7.5 Monte Carlo Experiments & Empirical Applications

See Seo and Shin (2014).

4.8 Bootstrap-based Bias Corrected Within Estimation of Threshold Regression Models in Dynamic Panels

Dang, Kim and Shin (2010) propose new estimation procedure to analyze asymmetric threshold effects in dynamic panels with unobserved individual effects when the number of time periods is fixed by combining nonlinear threshold regression techniques with FD-GMM estimation techniques. Kim and Shin (2014) advance an alternative approach to an analysis of dynamic panels with threshold effects, called the bias corrected within estimation procedure based on an iterative bootstrap mechanism by extending the approach in linear dynamic panels applied by Everaert and Pozzi (2007). Considering that the over-fitting bias and the weak instrument problem associated with the FD-GMM estimators will become more serious in the threshold dynamic panels, we expect that the proposed estimation procedure will

achieve the higher efficiency relative to the FD-GMM estimators. Monte Carlo simulation exercises confirm the validity of our proposed approach. In an application to the dynamic threshold version of Tobin's Q investment function using the UK company panel data, we are able to find strong evidence in favor of nonlinear dynamic threshold effects in firm's investment function.

4.8.1 Model

Consider the following dynamic panel threshold regression model:

$$y_{it} = \left(\phi_1 y_{it-1} + \boldsymbol{\beta}_1' \mathbf{x}_{it}\right) 1\left(q_{it} \leq \gamma\right) + \left(\phi_2 y_{it-1} + \boldsymbol{\beta}_2' \mathbf{x}_{it}\right) 1\left(q_{it} > \gamma\right) + \varepsilon_{it}, \tag{4.23}$$

for $i = 1, ..., N;\ t = 1, ..., T$, where y_{it} is a a scalar stochastic dependent variable, $\mathbf{x}_{it}$ is a $k \times 1$ vector of weakly exogenous variables, $1(\cdot)$ is an indicator function, q_{it} is the transition variable with γ being a threshold parameter, ϕ_i and $\boldsymbol{\beta}_i$ for $i = 1, 2$ are the corresponding heterogeneous parameters associated with two different regimes, and ε_{it} consists of the error components,[9]

$$\varepsilon_{it} = \alpha_i + v_{it},$$

where α_i is an unobserved individual effect and v_{it} is a zero mean idiosyncratic random disturbance. This is a panel extension of the dynamic threshold regression model in time series.

We make the following assumptions:
Assumption 1. $\{v_{it}\}$ are iid and independent of η_{it} and y_{i1} with $E(v_{it}) = 0$, $Var(v_{it}) = \sigma^2$ and have the finite 4th moment.
Assumption 2. α_i are iid with $E(\alpha_i) = 0$, $Var(\alpha_i) = \sigma_\alpha^2$ and have the finite 4th moment.
Assumption 3. y_{it} is geometrically ergodic and the initial observations satisfy the mean stationarity condition. Stability condition, $|\phi_i| < 1$ for $i = 1, 2$ or global stability condition, $|\phi_1 + \phi_2| < 2$.
Assumption 4. Exogenous variables, x_{it} are either I(0) or I(1), correlated with α_i but not correlated with v_{it}.The threshold variable, q_{it} is stationary and exogenous or predetermined uncorrelated with α_i and v_{it}.
Assumption 5. N is large and T is fixed.

All these assumptions are fairly standard in the literature, e.g. Alvarez and Arellano (2003) and Hansen (1999).

To simplify the notations we write (4.23) as

$$y_{it} = \boldsymbol{\delta}_1' \mathbf{z}_{1,it}(\gamma) + \boldsymbol{\delta}_2' \mathbf{z}_{2,it}(\gamma) + \varepsilon_{it} = \boldsymbol{\delta}' \mathbf{z}_{it}(\gamma) + \varepsilon_{it}, \tag{4.24}$$

where

$$\mathbf{z}_{1,it}(\gamma) = \begin{bmatrix} y_{i,t-1} \\ \mathbf{x}_{it} \end{bmatrix} \times 1\left(q_{it} \leq \gamma\right);\ \mathbf{z}_{2,it}(\gamma) = \begin{bmatrix} y_{i,t-1} \\ \mathbf{x}_{it} \end{bmatrix} \times 1\left(q_{it} > \gamma\right);$$

[9] The extension to the panels with the time-specific dummy effects is straightforward.

$$\mathbf{z}_{it}(\gamma) = \begin{bmatrix} \mathbf{z}_{1,it}(\gamma) \\ \mathbf{z}_{2,it}(\gamma) \end{bmatrix};\ \boldsymbol{\delta}_1 = \begin{bmatrix} \phi_1 \\ \boldsymbol{\beta}_1 \end{bmatrix};\ \boldsymbol{\delta}_2 = \begin{bmatrix} \phi_2 \\ \boldsymbol{\beta}_2 \end{bmatrix};\ \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{bmatrix}.$$

Taking the within transformation of (4.24), we obtain

$$\tilde{y}_{it} = \boldsymbol{\delta}'\tilde{\mathbf{z}}_{it}(\gamma) + \tilde{v}_{it}, \tag{4.25}$$

where

$$\tilde{y}_{it} = y_{it} - \bar{y}_i;\ \tilde{\mathbf{z}}_{it}(\gamma) = \mathbf{z}_{it}(\gamma) - \bar{\mathbf{z}}_i(\gamma);\ \tilde{v}_{it} = v_{it} - \bar{v}_i;$$
$$\bar{y}_i = T^{-1}\sum_{i=1}^{T} y_{it};\ \bar{\mathbf{z}}_i(\gamma) = T^{-1}\sum_{i=1}^{T}\begin{pmatrix} \mathbf{z}_{1,it}(\gamma) \\ \mathbf{z}_{2,it}(\gamma) \end{pmatrix}.$$

Next, we write (4.25) in the matrix form:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}(\gamma)\boldsymbol{\delta} + \tilde{\mathbf{v}}, \tag{4.26}$$

where

$$\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_N \end{bmatrix}_{NT\times 1},\ \tilde{\mathbf{Z}}(\gamma) = \begin{bmatrix} \tilde{\mathbf{z}}_1(\gamma) \\ \vdots \\ \tilde{\mathbf{z}}_N(\gamma) \end{bmatrix}_{NT\times 2(k+1)},\ \tilde{\mathbf{v}} = \begin{bmatrix} \tilde{\mathbf{v}}_1 \\ \vdots \\ \tilde{\mathbf{v}}_N \end{bmatrix}_{NT\times 1},$$

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \tilde{y}_{i1} \\ \vdots \\ \tilde{y}_{iT} \end{bmatrix}_{T\times 1},\ \tilde{\mathbf{z}}_i(\gamma) = \begin{bmatrix} \tilde{\mathbf{z}}'_{i1}(\gamma) \\ \vdots \\ \tilde{\mathbf{z}}'_{iT}(\gamma) \end{bmatrix}_{T\times 2(k+1)},\ \tilde{\mathbf{v}}_i = \begin{bmatrix} \tilde{v}_{i1} \\ \vdots \\ \tilde{v}_{iT} \end{bmatrix}_{T\times 1}.$$

For given γ and for large N and large T, $\boldsymbol{\delta}$ can be consistently estimated by the following within estimator:

$$\hat{\boldsymbol{\delta}}(\gamma) = \left(\tilde{\mathbf{Z}}(\gamma)'\tilde{\mathbf{Z}}(\gamma)\right)^{-1}\tilde{\mathbf{Z}}(\gamma)'\tilde{\mathbf{y}}. \tag{4.27}$$

Suppose that the consistent estimator of the threshold parameter is available and denoted, $\hat{\gamma}$. Then inference on $\boldsymbol{\delta}$ can proceed if $\hat{\gamma}$ were true value. Hence,

$$\hat{\boldsymbol{\delta}}(\gamma) \overset{a}{\sim} N(\boldsymbol{\delta}, \boldsymbol{\Omega}),$$

where the consistent estimate of $\boldsymbol{\Omega}$ can be obtained either by

$$\hat{\boldsymbol{\Omega}} = \hat{\sigma}^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{z}}_{it}(\hat{\gamma})\tilde{\mathbf{z}}_{it}(\hat{\gamma})'\right)^{-1},$$

if the errors are assumed to be iid or by

$$\hat{\mathbf{V}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{z}}_{it}(\hat{\gamma})\tilde{\mathbf{z}}_{it}(\hat{\gamma})'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{z}}_{it}(\hat{\gamma})\tilde{\mathbf{z}}_{it}(\hat{\gamma})'\hat{\tilde{v}}^2_{it}(\hat{\gamma})\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{z}}_{it}(\hat{\gamma})\tilde{\mathbf{z}}_{it}(\hat{\gamma})'\right)^{-1},$$

where $\hat{\tilde{v}}^2_{it}(\hat{\gamma}) = \tilde{y}_{it} - \hat{\tilde{\boldsymbol{\delta}}}'(\hat{\gamma})\tilde{\mathbf{z}}_{it}(\gamma)$, if the errors are allowed to be conditional heteroskedastic.

4.8.2 Bootstrap-based Bias Corrected Within Estimator

It is well-established in the linear dynamic panels that the fixed effects estimator of the autoregressive parameter is biased downward for fixed T (Nickell, 1981). Recently, in order to deal with the correlation of the regressors with individual effects, Dang, Kim and Shin (2012) adopt the FD-GMM and System-GMM approaches, and propose the extended GMM methodologies for consistently estimating the regime-specific parameters of the dynamic panel threshold regression model. Seo and Shin (2014) extend this framework further by allowing for both regressors and the threshold variable to be endogenous.

Here, we follow Everaert and Pozzi (2007), and develop an alternative bias corrected estimator based on an iterative bootstrap procedure. It is well-established that the GMM estimators tend to have relatively large standard errors and are subject to substantial small sample biases due to too many moment conditions, especially as T rises, e.g. Ziliak (1997). In this regard, Everaert and Pozzi (2007) propose the bootstrap-based bias correcting algorithm, which aims to reduce the bias of the within estimator while maintaining its higher efficiency relative to GMM estimators. Furthermore, Dang, Kim and Shin (2013) conduct a comprehensive empirical and simulation study in the corporate capital-structure, so as to compare and evaluate the finite-sample performance of all of the existing estimators for estimating both the long-run target leverage relationship and the associated speed of adjustment, in fixed T panels with unobserved individual effects, and document strong evidence that the bootstrap bias corrected within estimator outperforms the FD-GMM estimators in terms of estimation accuracy and efficiency. Another main advantage of the bootstrap algorithm is that this approach can be applied in a straightforward manner to the complex model with higher-order lagged regressors in which analytic corrections are not easily available.

For convenience we define $\hat{\boldsymbol{\delta}}(\gamma) = \hat{\boldsymbol{\delta}}$ without loss of generality. When T is fixed, it is easily seen that

$$E\left(\hat{\boldsymbol{\delta}}\right) \neq \boldsymbol{\delta}.$$

Suppose that using the repeated sampling experiment we are able to generate a sequence of J biased estimates, $\hat{\boldsymbol{\delta}}_1^*(\boldsymbol{\delta}), ..., \hat{\boldsymbol{\delta}}_J^*(\boldsymbol{\delta})$. Then, it follows that

$$E\left(\hat{\boldsymbol{\delta}}\right) = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} \hat{\boldsymbol{\delta}}_j^*(\boldsymbol{\delta}).$$

It is then clear that $\bar{\boldsymbol{\delta}}$ will be an unbiased estimator of $\boldsymbol{\delta}$ if the following condition holds:

$$\hat{\boldsymbol{\delta}} = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} \hat{\boldsymbol{\delta}}_j^*(\bar{\boldsymbol{\delta}}). \tag{4.28}$$

In other words, if we would sample repeatedly from a population with parameters $\bar{\boldsymbol{\delta}}$ and calculate the estimate $\hat{\boldsymbol{\delta}}_j^* \left(\bar{\boldsymbol{\delta}}\right)$ in each sample, $\bar{\boldsymbol{\delta}}$ is an unbiased estimator of $\boldsymbol{\delta}$ if the average of $\hat{\boldsymbol{\delta}}_j^* \left(\bar{\boldsymbol{\delta}}\right)$, $j = 1, ..., J$ corresponds to the FE estimate, $\hat{\boldsymbol{\delta}}$ based on the original data. See also Tanazaki (2004).

A bias corrected estimate of $\boldsymbol{\delta}$ can be obtained by searching over the parameter space until $\bar{\boldsymbol{\delta}}$ is found that satisfies (4.28). This search is implemented using the following iterative bootstrap algorithm: The core of this algorithm consists of a bootstrap procedure which simulates the distribution of the FE estimator when sampling from (4.24) with the initial values of a parameter vector, denoted $\tilde{\boldsymbol{\delta}}_{(0)}$.

1. Estimate the individual effects by

$$\tilde{\boldsymbol{\alpha}} = (T-1)^{-1} \mathbf{D}' \left(\mathbf{y} - \mathbf{Z}\tilde{\boldsymbol{\delta}}_{(0)}\right),$$

where $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_N)'$, $\mathbf{D} = \mathbf{I}_N \otimes \mathbf{e}_{T-1}$, $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\mathbf{e}_{T-1} = (1, ..., 1)'$ is the $(T-1) \times 1$ vector of ones. Then estimate the residual vector by

$$\tilde{\mathbf{v}} = \tilde{\mathbf{y}} - \tilde{\mathbf{Z}}\tilde{\boldsymbol{\delta}}_{(0)},$$

or rescale them as (see MacKinnon, 2002)

$$\tilde{v}_{it}^* = \sqrt{\frac{T-1}{T-2}} \left(\frac{\tilde{v}_{it}}{\sqrt{m_{it}}} - \frac{1}{T-1} \sum_{s=2}^{T} \frac{\tilde{v}_{is}}{\sqrt{m_{is}}} \right),$$

where m_{it} is the ith diagonal element of the idempotent matrix, $\mathbf{M} = \mathbf{I}_{N(T-1)} - \tilde{\mathbf{Z}}' \left(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\right)^{-1} \tilde{\mathbf{Z}}'$.

2. We generate the bth bootstrap samples for $b = 1, ..., B$ as follows:

   (a) Draw the bth bootstrap sample residual vector, denoted $\tilde{\mathbf{v}}^{(b)}$, from either $\tilde{\mathbf{v}}$ or $\tilde{\mathbf{v}}^*$ and generate a bootstrap sample by

$$\mathbf{y}^{(b)} = \mathbf{Z}^{(b)}\tilde{\boldsymbol{\delta}}_{(0)} + \mathbf{D}\tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{v}}^{(b)},$$

   where $\mathbf{z}_{it}^{(b)}(\gamma) = \left(z_{1,it}^{(b)}(\gamma), z_{2,it}^{(b)}(\gamma)\right) = \left\{y_{i,t-1}^{(b)} 1(q_{it} \leq \gamma), y_{i,t-1}^{(b)} 1(q_{it} > \gamma)\right\}$ and we condition on the initial value, y_{i1} such that $y_{i1}^{(b)} = y_{i1}$.

   (b) Compute the FE estimator by (4.27), namely

$$\tilde{\boldsymbol{\delta}}^{(b)} \left(\tilde{\boldsymbol{\delta}}_{(0)}\right) = \left(\tilde{\mathbf{Z}}^{(b)\prime}\tilde{\mathbf{Z}}^{(b)}\right)^{-1} \tilde{\mathbf{Z}}^{(b)\prime}\tilde{\mathbf{y}}^{(b)}.$$

3. Repeat the step 2 B times and calculate the empirical mean by

$$\bar{\boldsymbol{\delta}}_{(0)} = B^{-1} \sum_{b=1}^{B} \tilde{\boldsymbol{\delta}}^{(b)} \left(\tilde{\boldsymbol{\delta}}_{(0)} \right).$$

Define the difference between $\hat{\boldsymbol{\delta}}$ and $\bar{\boldsymbol{\delta}}_{(0)}$ by

$$\mathbf{b}_{(0)} = \hat{\boldsymbol{\delta}} - \bar{\boldsymbol{\delta}}_{(0)}.$$

(a) If $\mathbf{b}_{(0)} = \mathbf{0}$, $\tilde{\boldsymbol{\delta}}_{(0)}$ will be an unbiased estimator of $\boldsymbol{\phi}$ by the condition, (4.28).

(b) Otherwise update

$$\tilde{\boldsymbol{\delta}}_{(k+1)} = \tilde{\boldsymbol{\delta}}_{(k)} + \mathbf{b}_{(k)},\ k = 0, 1, 2, ...$$

and iterate the bootstrap procedures outlined in 1-3 until the condition, (4.28) is satisfied.

Following Everaert and Pozzi (2007), we will set the number of bootstrap samples B equal to 1000 and use the convergence criterion $\left|\mathbf{b}_{(k)}\right| < 0.005$ and set the upper bound on the number of iterations equal to 50. We also use the FE estimator as $\tilde{\boldsymbol{\delta}}_{(0)}$.

We now discuss how to generate the residual vector, $\tilde{\mathbf{v}}^{(b)}$ in details. Resampling $\tilde{\mathbf{v}}^{(b)}$ in a nonparametric way has the advantage that it does not require an explicit distributional assumption for $\mathbf{v}$. As we may also allow for temporal dependence in $\mathbf{v}$ we consider two alternative resampling schemes. First, when v_{it} is assumed to be iid across i and over t, we resample from

$$\tilde{\mathbf{v}}_i^{(b)} = \left(\tilde{v}^*_{i_1,t_2}, ..., \tilde{v}^*_{i_N,t_T} \right);\ i = 1, ..., N,$$

where the vectors of indices $(i_1, ..., i_N)$ and $(t_2, ..., t_T)$ are obtained by drawing with replacement randomly from $(1, ..., N)'$ and $(2, ..., T)'$, respectively. Second, if ε_{it} exhibits temporal dependence, e.g. conditional heteroskedasticity, we will use the wild bootstrap (Goncalves and Kilian, 2004) and resample from

$$\tilde{\mathbf{v}}_i^{(b)} = \left(\tau_{i2} \tilde{v}^*_{j2}, ..., \tau_{iT} \tilde{v}^*_{jT} \right);\ i = 1, ..., N,$$

where the index j is drawn with replacement from $(1, ..., N)'$ and τ_{it} is a binomial random variable with mean 0 and variance 1 that takes on value -1 and 1 respectively with probability 1/2. The advantage is that it is asymptotically valid for either T, N or both grow large. We then collect all the resampled residual vectors in

$$\tilde{\mathbf{v}}^{(b)} = \left(\tilde{\mathbf{v}}_1^{(b)\prime}, ..., \tilde{\mathbf{v}}_N^{(b)\prime} \right).$$

4.8.3 Estimation of and Testing for Threshold Effects

We have developed the optimal estimation procedure for the threshold autoregressive model in dynamic panels under the implicit assumption that the value of the threshold parameter, γ is given. This section will address consistent estimation of γ and develop the bootstrap-based testing procedure for the null of no threshold effects in dynamic panels.

We follow Chan (1993) and Hansen (1999) and obtain the consistent estimator of γ in by

$$\hat{\gamma} = \arg\min_{\gamma} Q_1(\gamma), \tag{4.29}$$

where $Q_1(\gamma)$ is the generalised minimum distance measure given by

$$Q_1(\gamma) = \hat{\mathbf{v}}(\gamma)' \hat{\mathbf{v}}(\gamma), \tag{4.30}$$

where $\hat{\mathbf{v}}(\gamma) = \tilde{\mathbf{y}} - \tilde{\mathbf{Z}}(\gamma)\hat{\boldsymbol{\delta}}^B(\gamma)$ and $\hat{\boldsymbol{\delta}}^B(\gamma)$ is the bootstrap-bias corrected estimator given γ. Once $\hat{\gamma}$ is obtained, we obtain

$$\hat{\boldsymbol{\delta}}^B = \hat{\boldsymbol{\delta}}^B(\hat{\gamma});\ \hat{\mathbf{v}} = \hat{\mathbf{v}}(\hat{\gamma});\ \hat{\sigma}^2 = \frac{1}{N(T-2)} Q_1(\hat{\gamma}). \tag{4.31}$$

Since $Q_1(\gamma)$ depends only on γ through the indicator function, this is a step function with most $N(T-2)$ steps with the steps occurring at distinct values of the observed threshold variable q_{it}. Thus the minimisation problem can be reduced to searching over the values of γ equalling the distinct values of q_{it} in the sample. In practice we need to truncate the smallest and largest 10% for example. The remaining values constitute the values of γ which can be searched for $\hat{\gamma}$. For each of these values regression are estimated yielding the SSE and the smallest value yields the estimate $\hat{\gamma}$ and $\hat{\boldsymbol{\delta}}^B = \hat{\boldsymbol{\delta}}^B(\hat{\gamma})$.

We follow Hansen (1996) and develop a bootstrap procedure to simulate the asymptotic distribution of the LR test statistic for the null hypothesis of no threshold:

$$H_0 : \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2.$$

Under the null of no threshold, the model (4.23) reduces to

$$y_{it} = \phi_1 y_{it-1} + \boldsymbol{\beta}_1' \mathbf{x}_{it} + \varepsilon_{it} = \boldsymbol{\delta}_1' \mathbf{z}_{it}^* + \varepsilon_{it},$$

where $\mathbf{z}_{it}^* = (y_{it-1}, \mathbf{x}_{it}')'$. Taking the within transformation, we have

$$\tilde{y}_{it} = \boldsymbol{\delta}_1' \tilde{\mathbf{z}}_{it}^* + \tilde{v}_{it}, \tag{4.32}$$

where $\tilde{\mathbf{z}}_{it}^* = \mathbf{z}_{it}^* - T^{-1}\sum_{i=1}^{T} \mathbf{z}_{it}^*$, from which we obtain the bootstrap-bias corrected estimator $\tilde{\boldsymbol{\delta}}_1^B$, and the corresponding generalised minimum distance measure:

$$Q_0 = \tilde{\tilde{\mathbf{v}}}' \tilde{\tilde{\mathbf{v}}},$$

where $\widetilde{\widetilde{\mathbf{v}}} = \left\{\widetilde{\widetilde{v}}_{it}\right\}$ with $\widetilde{\widetilde{v}}_{it} = \tilde{y}_{it} - \tilde{\boldsymbol{\delta}}_1^{B\prime}\tilde{\mathbf{z}}_{it}^*$. The LR test statistic is then given by

$$LR = \frac{Q_0 - Q_1(\hat{\gamma})}{\hat{\sigma}^2}. \tag{4.33}$$

We take the residuals, $\hat{\mathbf{v}}_i = \hat{\mathbf{v}}_i(\hat{\gamma})$ (see (4.30)), group them by individual $i = 1, ..., N$, namely $\hat{\mathbf{v}}_i = (\hat{v}_{i2}, ..., \hat{v}_{iT})'$ and treat $(\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_N)$ as the empirical distribution to be used for bootstrapping. We then draw (with replacement) a sample of size N from the empirical distribution and use these errors to create a bootstrap sample under H_0 and under the alternative, separately. Treating the initial value y_{i1} as given, we use the bootstrap sample and estimate the model under the null and under the alternative and calculate the bootstrap value of the LR test at each replication. We repeat this procedure a large number of times, e.g. 1000 times and calculate the percentage of draws for which the simulated statistic exceeds the actual. This is the bootstrap estimate of the asymptotic p-value for LR under H_0. Hansen (1996) shows that a bootstrap procedure attains the first-order asymptotic distribution, so p-values are asymptotically valid.

4.8.4 Empirical Application: To be filled.

Many studies in empirical corporate finance have employed dynamic panel data models to examine the dynamic behavior of corporate financial policy variables. However, a major difficulty in using these models is determining how to obtain consistent and efficient estimates, especially in short panels of company data, in the likely presence of (1) unobserved heterogeneity and endogeneity, (2) residual serial correlation, or (3) fractional dependent variables. Which estimators should researchers use in these contexts?

To address this important research question, Dang, Kim and Shin (2014) investigate two classes of advanced econometric techniques for dynamic panel data models, including (1) the instrumental variable (IV) approach, the first-difference GMM (FD-GMM), the system GMM(SYS-GMM), and the long difference GMM (LD- and LDP-GMM), and (2) the bias-corrected estimators, based on either an analytical approach (LSDVC) or an iterative bootstrap procedure (BC). Further, we consider an augmented Tobit estimator (DPF) that accounts for the fractional nature of the dependent variable. We conduct Monte Carlo simulation experiments and present two empirical applications, to capital structure and cash holdings, to examine the relative performance of the estimators.

Our simulation and empirical results show that the bias-corrected estimators, BC and LSDVC, are generally most appropriate and robust for estimating dynamic panel data models in empirical corporate finance. These methods can estimate the autoregressive coefficient and the coefficients on the explanatory variables with the most accuracy and efficiency. In our simulations, they are also robust to changes in the key control parameters,

including the relative magnitude of the fixed effects and the relative explanatory power of the regressors. Of these two methods, BC is generally preferable to LSDVC because it performs well even in regressions with autocorrelation, and in models with higher lag orders, such as the ARDL(2,1) model. In the specific case where the dependent variable is a ratio, and censored at 0 and 1, BC and LSDVC may still provide reasonable estimates, although at a high level of censoring researchers should consider using DPF instead. In our empirical applications using capital structure and cash holdings, we find that BC and LSDVC produce the most plausible estimates of the speeds of dynamic leverage and cash adjustments, as well as the most reasonable estimates of the coefficients on the explanatory variables. Our analysis thus provides additional insights into empirical research studying target leverage and cash adjustment behaviors. Our results, obtained using these appropriate estimators, suggest that firms adjust toward their target leverage at a moderate rate, between 26% and 28%, but move toward their target cash holdings more quickly, with a speed of 48%.

We find that the IV/GMM estimators generally perform poorly in our simulation and empirical studies. Their estimates of the autoregressive coefficient tend to be biased and unreliable. Moreover, these methods are very sensitive to the presence of unobserved heterogeneity, endogeneity, and, in particular, serially correlated errors. Overall, our study suggests that the IV/GMM estimators should be used with extreme care.

While our comprehensive simulations have systematically examined the most important issues in the estimation of dynamic panel data models, we cannot account for all of the specific settings and minor issues encountered in the vast empirical research literature. In addition, while our simulation results should generalize to many areas of corporate finance, our empirical applications are restricted to two topics, namely capital structure and cash holdings. Hence, it would be useful for future research to verify our simulation and empirical findings in other areas of corporate finance.

4.9 Further Issues

- Threshold Error Correction Models in Dynamic Panels with Homogeneous Long-run Relationship
- PMG Estimation of Threshold Dynamic (Heterogeneous) Panels
- Smooth transition regression in dynamic heterogeneous panels
- Markov switching panel data models
- Cross Section Dependence

Chapter 5

Cross Sectionally Correlated Panels

Cross-section dependence, CSD, seems pervasive in panels, it seems rare that the covariance of the errors is zero. 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 Baltagi, 2005) 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.

5.1 Overview on Cross-section Dependence

This section mainly consists of the summary of Chapters 7 and 8 in Smith and Fuertes (2012).

5.1.1 Representations of CSD

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. Spatial models give the $N \times 1$ vector of

errors the structure

$$e_t = W\varepsilon_t$$

where ε_t is cross-sectionally independent and W is a known (possibly time-varying) matrix, reflecting, for instance, whether the units share a common border. This can be used to represent spatial autoregressive, moving average or error component models. Pesaran and Tosetti (2009) discuss the links between the various forms of CSD and give more precise definitions.

Factor models have the errors reflect a vector of unobserved common factors

$$y_{it} = z_t'\alpha_i + \beta_i' x_{it} + \gamma_i' f_t + \varepsilon_{it}$$

where y_{it} is a scalar dependent variable, z_t is a $k_z \times 1$ vector of variables that do not differ over units, e.g. intercept and trend, x_{it} is a $k_x \times 1$ vector of observed regressors which differ over units, f_t is an $r \times 1$ vector of unobserved factors, which may influence each unit differently and which may be correlated with the x_{it}, 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 (possibly) t. The covariance between the errors $e_{it} = \gamma_i' f_t + \varepsilon_{it}$ is determined by the factor loadings γ_i. Notice that if f_t is correlated with x_{it}, as is likely in many economic applications such as global cycles, then not allowing for CSD by omitting f_t causes the estimates of β_i to be biased and inconsistent.

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

$$y_t = \Phi y_{t-1} + u_t$$

and consider approximations to this structure as $N \to \infty$ (Chudik and Pesaran, 2009).

Kapetanios, Mitchell and Shin (2010) have an interesting non-linear model of endogenous cross-section dependence, which can capture aspects of herding.

5.1.2 Weak and strong CSD

With weak CSD, the dependences are local and decline with N. This could be the case with spatial correlations, where each cross-section unit is correlated with near neighbors but not others; with strong CSD the dependences influence all units. The distinction can be expressed in various ways. Suppose the elements of y_t are stationary, e.g. growth rates, and the weighted average of the elements $\bar{y}_t = \sum_{i=1}^N y_{it}/N$, where the weights are 'granular', go to zero as $N \to \infty$. Then with weak CSD the variance of $\bar{y}_t$ goes to zero as $N \to \infty$. If there is strong CSD it does not, for instance there may be a global cycle in $\bar{y}_t$. If there is weak CSD the influence of the factors, $\sum_{i=1}^N \gamma_i^2$ 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 are bounded as $N \to \infty$. If there is strong dependence, the largest eigenvalue goes to infinity with N. 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 $\alpha < 1$.

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, f_t. If f_t is omitted and the correlation between f_t and x_{it} changes, this will change the estimate of β_i giving the appearance of structural change. Similarly, an omitted factor may give the impression of non-linearity.[1] Since unobserved factors play a major role in the treatment of CSD, we begin by discussing the estimation of such factors. The implications for estimation are different depending on whether f_t are merely regarded as nuisance parameters that we wish to control for in order to get better estimates of or whether they are the parameters of interest: one wishes to estimate f_t as variables of economic interest in their own right.

5.1.3 The correlated common effect 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:

$$y_{it} = z_t'\alpha_i + x_{it}'\beta_i + \delta_{0i}\bar{y}_t + \delta_i'\bar{x}_t + u_{it}$$

To see the motivation, assume a single factor and average (33) across units to give:

$$\bar{y}_t = z_t'\bar{\alpha} + \bar{x}_t'\bar{\beta} + \bar{\gamma}\bar{f}_t + \bar{\varepsilon}_t + N^{-1}\sum\left(\beta_i - \bar{\beta}\right)' x_{it}$$

$$\bar{f}_t = \frac{1}{\bar{\gamma}}\left\{\bar{y}_t - z_t'\bar{\alpha} - \bar{x}_t'\bar{\beta} - \bar{\varepsilon}_t - N^{-1}\sum\left(\beta_i - \bar{\beta}\right)' x_{it}\right\}$$

so the $\bar{y}_t$ and $\bar{x}_t$ provide a proxy for the unobserved factor. Notice that the covariance between $\bar{y}_t$ 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 $\bar{\gamma}$ or the vector equivalent, is non-zero.

This formulation assumes heterogeneous coefficients, there are homogeneous versions. There are sometimes economic reasons for adding averages,

[1] 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.

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 (Bai, 2009) and Westerlund & Urbain (2011) examine why this should be the case.

5.1.4 Uses of factor models

One can distinguish two different types of problem that the factor models are used for. The Pesaran approach to the role of unobserved factors in panel data models is primarily motivated by the need to allow for "error" cross-sectional dependence. The aim is to estimate the (mean) coefficient of x_{it} allowing for the possibility of error cross-sectional dependence and/or missing unobserved effects, irrespective of whether f_t are I(0) or I(1).

Alternatively, in some applications it might be relevant to view the unobserved factor as "missing" (omitted) common effects. An example of the latter is "technology" in the aggregate production function. In modelling error cross-sectional dependence the error of each cross section unit should have mean zero (otherwise the model suffers from omitted variables) and could be serially correlated. The errors could also be I(1). But if the aim is to test for cointegration between observables, y_{it} and x_{it} (which could also contain observed common effects such as oil prices), and if we maintain the possibility that the errors of the relationship between y_{it} and x_{it} can be I(1), then it is clear that y_{it} and x_{it} cannot cointegrate.

One could, hypothesize instead that y_{it} and x_{it} and f_t are cointegrated where f_t is an "unobserved" factor. Even if such a possibility existed, it may not be relevant if the economic relation of interest is between y_{it} and x_{it}; e.g. in the case of PPP or UIP.

In the case where the common factor represents a missing variable, such as the technological variable in the production function, the null of interest to the economist is in fact that y_{it} (log output per man $f_t$hour) k_{it} (log capital per man hour) and f_t (global technology) are cointegrated. The role of f_t is not to model error cross-sectional dependence, but is an integral part of the model explanation, which happens to be unobserved. In this case one could try to obtain proxies for f_t - some in the literature assume that f_t is a linear trend with a stationary component, others assume that it is a latent variable and use HP type filter to identify and measure it. In the context of growth convergence one might estimate f_t by the cross sectional average of y_{it} over i. But, given the unobserved nature of f_t, there will be some degree of arbitrariness associated with these choices. For example, how can we establish that f_t is I(1) or trend stationary? Not knowing whether f_t is I(1) or trend stationary, how can we test that y_{it}, k_{it} and f_t are cointegrated.[2]

[2] In the case of testing for panel unit roots, the cross-sectionally augmented CADF test, CADF is a joint test of a unit root and a stationary f_t.

5.2 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}, $j = 1, ..., r$:

$$x_{it} = \lambda_{i0} + \sum_{j=1}^{r} \lambda_{ij} f_{jt} + e_{it}$$

λ_{ij} are often called factor loadings, the e_{it} are 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.

5.2.1 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.

5.2.2 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 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 parametric 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. From the late 1940s the errors in equations model cam to dominate. The basic approach to measuring unobserved variables by the principal components (PCs) of a data matrix was developed by Hotelling (1933), then a Professor of Economics. Later, Richard Stone (1947) used this method 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 played relatively little role in the development of econometrics, which focussed on the errors in equations model. There are some exceptions, such as the use of it by the Adelman's to measure development, and factor interpretations of Friedman's permanent

income but they were relatively rare. 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.

5.3 Calculating Principal Components

5.3.1 Static Models

Suppose that we have a $T \times N$ data matrix, X with typical element x_{it}, observations on a variable for units $i = 1, ..., N$ and periods $t = 1, ..., T$. 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 factors stacked in the $T \times r$ matrix F. In matrix notation,

$$X = F\Lambda + E$$

where Λ is an $r \times N$ 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, say ML to estimate the factor loadings, but we will use a different approach and 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

$$\begin{aligned} L &= a_1'X'Xa_1 - \phi_1(a_1'a_1 - 1) \\ \frac{\partial L}{\partial a_1} &= 2X'Xa_1 - 2\phi_1 a_1 = 0 \\ X'Xa_1 &= \phi_1 a_1 \end{aligned}$$

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:

$$X'XA = \Phi A$$

where A is the matrix of eigenvectors and 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

$$X = F_1A_1' + F_2A_2'$$

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_2A_2' as the idiosyncratic factors corresponding to the e_{it} in (35). While it is an abuse of this notation, we will usually write F_1 as F and F_2A_2' as E.

5.3.2 Dynamic Models

Suppose that t does represent time and we write the factor model in time series form:

$$x_t = \Lambda f_t + e_t$$

where x_t is an $N \times 1$ vector, an $N \times r$ matrix of loadings, f_t a $r \times 1$ vector of factors and e_t an $N \times 1$ vector of errors. In using PCs to calculate the factors we have ignored all the information in the lagged values of 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

$$x_t = \Lambda^* f_t^* + e_t^*$$

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,

$$x_t = \Lambda(L) f_t + e_t^s$$

where $\Lambda(L)$ is a pth order lag polynomial. This may be less efficient in the sense that $r < rp$: 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 x_t comes from an autoregression in x_t, dynamics in f_t or serial correlation in 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.

5.3.3 Issues in using PCs

How to choose r

How many factors to use, i.e. determining r; depends on statistical criteria, 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. Recently there has been work on information criteria that can be used when both N and T are large.

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

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

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:

$$IC_1(r) = \log(V(r)) + r\left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right)$$

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

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.

Having chosen r denote the estimates

$$\tilde{e}_{it} = x_{it} - \tilde{\lambda}_i' \tilde{f}_t$$

$$\tilde{E} = X - \tilde{\Lambda}\tilde{F}$$

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

Kapetanios (2004a) suggests using the largest eigenvalue to choose the number of factors. Onatski (2009) suggests another function of the largest eigenvalues of the spectral density matrix at a specified frequency. The statistical properties of the various tests, information criteria and other methods of choosing r for economic data is still a matter of research. Of course the choice of r will depend not just on statistical criteria but also the purpose of the exercise and the context.

How to choose N

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). Notice that in forecasting we would need to update our estimates of F_t, and perhaps r the number of factors, as T our sample size changes.

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 $\tilde{\Lambda} = P^{-1}\Lambda$ and factors are $\tilde{F} = FP$. 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.

These are needed to provide some interpretation of the factors.[3]

The same identification issue arises in simple regression. For

$$y = X\beta + u$$

is observationally equivalent to the reparameterisation

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

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.

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.

The most obvious candidate for imposed weights is to use equal weights, a simple average (perhaps after having standardised the variable to have mean zero and variance one). It was noted above that in many cases the

[3] 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.

first PC seems to have roughly equal weights and thus behave like an average or sum.

Alternatively, economic theory may suggest suitable weights. For instance, effective exchange rates for country i (weighted averages of exchange rates with all other countries) use trade weights: exports plus imports of i with j as a share of total exports plus imports of country i). PCs might give a lot of weight to a set of countries which have very volatile exchange rates (which would account for a lot of the variance) even though country i does not trade with them. Measures of core inflation give zero weights to the inflation rates of certain volatile components of total expenditure, while a PC might give a high weight to those volatile components because they account for a lot of the variance. Monte Carlo evaluation of estimators that allow for CSD indicate the methods that use imposed weights, like the CCE estimator, often do much better than estimators that rely on estimating the number of PCs and their weights.

Explanation using PCs

Suppose the model of interest is

$$y = X\beta + u$$

where β is an $N \times 1$ vector and you wish to reduce the dimension of 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

$$y = XAA'\beta + u$$

$$y = F\delta + u$$

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

$$y = F_1\delta_1 + F_2\delta_2 + u$$

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 $\delta_2 = 0$ is testable (as long as $N < T$, which it may not be in some applications) and should be tested if it can be. This has been suggested as a way of dealing with multicollinearity, or choosing a set of instruments, however there are some problems. First, it is quite possible 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. Secondly unless 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.

5.3.4 Factor Augmented VARS, FAVARs

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. These time paths are called 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 an $M \times 1$ vector of observed focus variables Y_t, a $K \times 1$ vector

of unobserved factors F_t with a VAR structure

$$\begin{pmatrix} F_t \\ Y_t \end{pmatrix} = A(L) \begin{pmatrix} F_{t-1} \\ Y_{t-1} \end{pmatrix} + v_t$$

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

$$X_t = \Lambda F_t + e_t$$

where F_t are estimated as the principal components of X_t, which may include Y_t. Notice there is an identification problem, since $X_t = \Lambda F_t + e_t = \Lambda P P^{-1} F_t + e_t = \Lambda^* F_t^* + e_t$. It is common to use an arbitrary statistical assumption to identify the loadings as eigenvectors, but other assumptions are possible. The standard practice used by most of these authors is to difference the observable data so that they are stationary, I(0). The factors are therefore also stationary. As noted above differencing loses levels information about the level relationships, but if one does not difference one has to take account of cointegration etc.

The argument is that (a) a small number of factors can account for a large proportion of the variance of X_t and thus parsimoniously reduce omitted variable bias in the VAR ; (b) the factor structure for X_t allows one to calculate impulse response functions for all the elements of X_t in response to a (structural) shock in Y_t transmitted through 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 X, (2) estimate the VAR treating the PCs as measures of F_t variables along with the M observed focus variables 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 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 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 X_t, but do not help in explaining 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.

5.4 Estimation of Cross Sectionally Dependent Panels

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, appears to be CCE type estimators. However, there are a number of issues of interpretation.

It is common in a lot of time-series applications of PCs to transform the data to make it stationary before calculating PCs, e.g. by differencing. If one is trying to measure a stationary unobservable, e.g a. global trade cycle, this is clearly sensible. It is equally clearly not sensible if one is trying to measure a non-stationary unobservable, e.g. a global trend. Even in the stationary case it is important that transformations beyond differencing be considered, stationary transformations of levels variables, such as interest rate spreads, may also contain valuable information. We now describe main techniques for dealing with cross-section dependence.

5.4.1 SURE

Suppose that the model is heterogeneous:

$$y_{it} = z_t'\alpha_i + x_{it}'\beta_i + f_t'\gamma_i + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T$$

where y_{it} is a scalar dependent variable, z_t is a k_z 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 an $r \times 1$ vector of unobserved factors 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 t. Estimating

$$y_{it} = z_t'\alpha_i + x_{it}'b_i + v_{it},\ i = 1, ..., N,\ t = 1, ..., T$$

will give inconsistent estimates of β_i if f_t is correlated with x_{it} and inefficient estimates even if f_t 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.

5.4.2 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 $\tilde{y}_{it} = y_{it} - \bar{y}_t$, where $\bar{y}_t = N^{-1}\sum_{i=1}^{N} y_{it}/N$ and similarly for $\tilde{x}_{it}$: Unlike SURE the factor does not have to be distributed independently of 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

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

noting that

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

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

$$\tilde{y}_{it} = \beta_i' \tilde{x}_{it} + \tilde{u}_{it} + e_{it}$$

$$e_{it} = \eta_i' \bar{x}_t - N^{-1}\sum_{i=1}^{N} \eta_i' x_{it}$$

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 $\tilde{y}_{it}$ 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.

5.4.3 Including Means, the CCE estimator

The correlated common effect estimator of Pesaran (2006), discussed earlier, suggests including the means of y_{it} and x_{it} as additional regressors, to remove the effect of the factors, which are treated as nuisance parameters. The CCE procedure is simple to apply, can handle multiple factors which are I(0) or I(1), 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

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

These clearly hold for the mean:

$$w_i = \frac{1}{N};\ \sum_{i=1}^{N} |w_i| = 1\ and\ \sum_{i=1}^{N} w_i\gamma_i = N^{-1}\sum_{i=1}^{N} \gamma_i \neq 0$$

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. Kapetanios, Pesaran and Yamagata (2011) show that this procedure is robust to a wide variety of data generation processes including unit roots. See also Kapetanios and Pesaran (2007).

5.4.4 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

$$x_{it} = c_i + \beta_i t + \lambda_i' F_t + e_{it}$$

$$F_{mt} = \alpha_m F_{mt-1} + u_{mt}$$

$$e_{it} = \rho_i e_{it-1} + \varepsilon_{it}$$

Factor α_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 idiosyncratic error. Unit root tests can then be applied to these to determine whether the variables are I(0) or I(1). Since removing F_t has removed the between group dependence, panel unit root tests can be conducted on 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. Breitung and Das (2008) review testing for unit roots in panels with a factor structure. In all these cases, determining whether 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.

5.4.5 Residual Principal components

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

$$y_{it} = b_i' x_{it} + e_{it}$$

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

$$y_{it} = \beta_i' x_{it} + c_i' \hat{f}_t + v_{it}$$

Assume that the x_{it} are generated by

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

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

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

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 $\phi_{3i} = 0$. Otherwise using inconsistent estimates of $\hat{e}_{it}$ 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 x_t or if the factor is perfectly correlated with x_t. If as N gets large the factor is perfectly correlated with x_t, then it is obviously sensible to use $\bar{x}_t$.

5.4.6 Interactive fixed effects

Bai (2009) considers the model:

$$Y_{it} = X_{it}'\beta + \lambda_i' F_t + \varepsilon_{it};$$

$$Y = X\beta + F\Lambda' + \varepsilon$$

where X_{it}' is a $p \times 1$ vector of observable regressors and F_t is an $r \times 1$ vector of unobserved factors. This model, is similar to that of the previous subsection, the difference being it assumes homogeneous. Bai interprets it as a generalisation of the usual additive two way fixed effect model, e.g. when $\lambda_i' = \lambda_i$, 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.

5.4.7 Further remarks

Coakley, Fuertes and Smith (2006) conduct Monte Carlo experiments to contrast the finite sample properties of a large number of estimators in the context of cross-section dependence. Attention is confined to static models with strong exogeneity to establish some baseline results. The results for the different settings suggests:

1. Averaging across spurious regressions gives an unbiased measure of β in the presence of cross-sectional dependence also. This result is of interest since the asymptotic theory in Phillips and Moon (1999) builds on the assumption of cross-section independence.

2. The efficiency of the CS estimator improves in the I(1) case.

3. The CCE does very well across the board.

Monte Carlo results should be treated with some caution since the DGP used in the simulation may not reflect factors typical in real applications. See Smith and Furetes (2012) for more details and the empirical applications.

5.5 Panel Gravity Models in the Presence of Cross Section Dependence

5.5.1 Overview on the Euro's Trade Effects

Recently, there has been an intense policy debate on the Euro effects on trade flows between Euro and non-Euro nations. Baldwin (2006) offers an extensive survey, covering the infamous Rose (2000)'s huge trading effect over 200% as well as most recent studies reporting the relatively smaller effects. It is widely acknowledged that the Rose's estimate of the currency union effect on trade is severely (upward) biased. In particular, his estimates are heavily inflated by the presence of small (e.g. Ireland, Panama) or very small (e.g. Kiribati, Greenland, Mayotte) countries. An important issue is why a currency union raises trades so much. Thus, it is unclear whether one can uncover similar findings for the European monetary union involving the substantially large economies such as Germany and France.

The gravity model popularised by Rose (2000) attempts to provide the main link between trade flows and trade barriers, though his original approach has attracted the number of strong criticisms. The main critiques are classified as follows: inverse causality or endogeneity; missing or omitted variables; and incorrect model specification (nonlinearity or threshold effects). Nowadays, the general consensus is, once these methodological issues have been accommodated appropriately, that the currency union effect seems to be far less than those reported earlier by Rose and others, especially using the larger dataset including numerous smaller countries. We still find that the range of the estimated Euro effects is very wide from 2% to more than 70%.

There may be other third factors, such as common language, colonial history, and political/institutional link, that may influence both currency choice and trade link. In this regard, high correlations reported in earlier studies may be spurious as an artifact of reverse causality. A related issue is how the currency union is formed. Countries who decide to join a currency union are self-selected on the basis of distinctive features shared by countries that have been EU members during the pre-Euro period. Hence, countries are likely to foster integration by enhancing standards of harmonization and reducing regulatory barriers. To address this issue, a number of studies have employed different techniques such as Heckman selection and instrumental variables, though they still obtained the substantial Euro effects on trade.

A more important issue is omitted variables bias. Omitted pro-bilateral

trade variables are likely to be correlated with the currency union dummy, as the formation of currency unions is not random, but rather driven by some factors which are likely to be omitted from the gravity regression. The implication is that the Euro effect will capture general economic integration among the member states, not merely the currency effect. Several studies tried to reduce the endogenous effect of currency unions by introducing country-pair and year fixed effects in the gravity regression.

Anderson and van Wincoop (2003) propose the 'micro foundation' of the gravity equation by introducing the multilateral resistance terms, which are relative trade barriers -the bilateral barrier relative to average trade barriers that both countries face with all their trading partners. The empirical gravity literature has simply added so-called remoteness variables, which are defined as a weighted average distance from all trading partners with the weights being based on the size of the trading partners, though such atheoretical remoteness indices fail to capture any of the relative trade barriers in a coherent manner. Hence, the standard gravity model is seriously lacking if multilateral resistance terms and/or trade costs are ignored or seriously misspecified.

Furthermore, Baldwin (2006) stresses an importance of taking into account time-varying multilateral resistance terms, and criticises the conventional fixed effect estimation technique because many of omitted pair-specific variables clearly reflect time-varying factors such as multilateral trade costs. The use of time-invariant effects only may still leave a times-series trace in the residual, which is likely to be correlated with the currency union dummy. An incomplete account of time variation in multilateral resistance terms is likely to cause omitted-variable bias. Therefore, once such time-varying effects are explicitly incorporated in the gravity model, the impact of currency-union would be greatly reduced. A number of studies have attempted to capture time-varying effects when estimating the Euro's trade effect.

In sum, a large number of existing studies have established the importance of appropriately taking into account unobserved and time-varying multilateral resistance and bilateral heterogeneity, simultaneously. This immediately raises another important issue of cross-section dependence among trade flows, which has been so far neglected. Only recently, Herwartz and Weber (2010) propose to capture both multilateral resistance terms and omitted trade costs via unobserved time-varying country-pair specific random walk factors, and develop the Kalman-filter extension of the gravity model. They find that aggregate trade (export) within the Euro area increases between 2000 and 2002 by 15 to 25 percent compared with trade with non-members of the Euro area due to a decrease in long-lasting trade costs. More importantly, Behrens et al. (2012) derive a quantity-based structural gravity equation system in which both trade flows and error terms are allowed to be cross-sectionally correlated, and propose the modified spatial

techniques by adopting a broader definition of the spatial weight matrix, called the interaction matrix, which can be derived directly from theoretical model. By controlling for cross-sectional interdependence and thus directly capturing multilateral resistance, they find that the measured Canada–US border effects are significantly lower than paradoxically large estimates reported by McCallum (1995). Behrens et al. (2012) also argue that their approach - unconstrained linearized gravity equation with cross-sectionally correlated trade flows - is better suited than the two-stage gravity equation system with nonlinear constraints in unobservable price indices advanced by Anderson and von Wincoop (2003).

Taken together, all of the above discussions may suggest that an Euro effect on trade is expected to be smaller in the future than previously thought once multilateral resistance term is well-captured via the cross-sectional interdependence of trade flows. In retrospect, Serlenga and Shin (2007) is the first paper to introduce the cross-section dependence into the panel gravity model, and to provide consistent estimation procedure for both time-varying and time-invariant regressors. Employing the dataset over the period 1960-2001, SS find that the introduction of a common currency does not exert any significant effect on intra-EU trade, though their sample covers only three years' data since the introduction of the Euro in 1999. Given the availability of a longer sample, we wish to redress this important issue by extending the cross-sectionally dependent panel gravity model and addressing all of the issues related to unobserved and time-varying multilateral resistance and bilateral heterogeneity as surveyed above.

The fixed effect estimation has been most popular in the literature on gravity models, though it fails to estimate coefficients on time-invariant variables such as distance, since the within transformation wipes out those variables. Another important issue is how to extend the fixed effect specification into a more general case with individual-specific and time-varying effects, both of which affect bilateral trade flows. The multilateral resistance function and trade costs are not only difficult to measure, but also likely to vary over time. A number of approaches have been proposed. Simply, fixed time dummies or time trends are added as a proxy for time-varying effects in the gravity equation to allow to time trend coefficients to be heterogeneous across country-pairs. Alternatively, some studies include *ad hoc* regional remoteness indices, although these indices have no theoretical foundation (Behrens et al., 2012).

5.5.2 Extended HT estimation

We now consider a more generalized panel data model advanced by SS and Baltagi (2010):

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\gamma}'\mathbf{z}_i + \boldsymbol{\pi}_i'\mathbf{s}_t + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{5.1}$$

$$\varepsilon_{it} = \alpha_i + \boldsymbol{\varphi}_i' \boldsymbol{\theta}_t + u_{it}, \tag{5.2}$$

where $\mathbf{x}_{it} = (x_{1,it}, ..., x_{k,it})'$ is a $k \times 1$ vector of variables that vary across individuals and over time periods, $\mathbf{s}_t = (s_{1,t}, ..., s_{s,t})'$ is an $s \times 1$ vector of observed time-specific factors, $\mathbf{z}_i = (z_{1,i}, ..., z_{g,i})'$ is a $g \times 1$ vector of individual-specific variables, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_g)'$ and $\boldsymbol{\pi}_i = (\pi_{1,i}, ..., \pi_{s,i})'$ are conformably defined column vectors of parameters, α_i is an individual effect that might be correlated with explanatory variables $\mathbf{x}_{it}$ and $\mathbf{z}_i$, $\boldsymbol{\theta}_t$ is the $c \times 1$ vector of unobserved common factors with conformable parameter vector, $\boldsymbol{\varphi}_i = (\varphi_{1,i}, ..., \varphi_{c,i})'$, and u_{it} is a zero mean idiosyncratic uncorrelated random disturbance.

The distinguishing feature of this model is that it allows for both observed and unobserved time effects both of which are cross-sectionally correlated. Both factors are expected to provide good proxy for any remaining complex time-varying patterns associated with multilateral resistance and globalisation trends, e.g. Mastromarco et al. (2013). Notice that the cross-section dependence in (5.1) is explicitly allowed through heterogeneous loadings, $\boldsymbol{\varphi}_i$, see Pesaran (2006) and Bai (2009).[4] It is easily seen that most specifications of the gravity equation in the literature can be expressed as a variation of the model given by (5.1) and (5.2).[5] Hence, this factor-based cross-sectionally dependent panel gravity model is expected to capture the time-varying pattern of unobserved trading effects, such as the multilateral resistance, in a robust manner.

The conventional panel data estimators of (5.1) would be seriously biased without properly accommodating the cross-sectionally dependent factor structure given by (5.2). To explicitly address this issue, we consider the two leading approaches proposed by Pesaran (2006) and Bai (2009). Following Pesaran (2006) we first derive the cross-sectionally augmented regression of (5.1) as follows:

$$y_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + \boldsymbol{\gamma}' \mathbf{z}_i + \boldsymbol{\lambda}_i' \mathbf{f}_t + \alpha_i^* + u_{it}^*, \; i = 1, ..., N, \; t = 1, ..., T, \tag{5.3}$$

where $\mathbf{f}_t = (\mathbf{s}_t', \bar{y}_t, \bar{\mathbf{x}}_t')' \left\{ = (f_{1,t}, ..., f_{\ell,t})' \right\}$ is the $\ell \times 1$ vector of augmented time-specific factors with $\ell = s + 1 + k$ and $\boldsymbol{\lambda}_i = (\lambda_{1,i}, ..., \lambda_{\ell,i})'$, $\bar{y}_t = N^{-1} \sum_{i=1}^N y_{it}$, $\bar{\mathbf{x}}_t = N^{-1} \sum_{i=1}^N \mathbf{x}_{it}$, $\boldsymbol{\lambda}_i' = \left(\boldsymbol{\pi}_i' - (\varphi_i / \bar{\varphi}) \, \bar{\boldsymbol{\pi}}', (\varphi_i / \bar{\varphi}), - (\varphi_i / \bar{\varphi}) \, \boldsymbol{\beta}' \right)'$ with $\bar{\varphi} = N^{-1} \sum_{i=1}^N \varphi_i$ and $\bar{\boldsymbol{\pi}} = N^{-1} \sum_{i=1}^N \boldsymbol{\pi}_i$, $\alpha_i^* = \alpha_i - (\varphi_i / \bar{\varphi}) \, \bar{\alpha} - (\varphi_i / \bar{\varphi}) \, \boldsymbol{\gamma}' \bar{\mathbf{z}}$ with $\bar{\alpha} = N^{-1} \sum_{i=1}^N \alpha_i$ and $\bar{\mathbf{z}} = N^{-1} \sum_{i=1}^N \mathbf{z}_i$, and $u_{it}^* = u_{it} - (\varphi_i / \bar{\varphi}) \, \bar{u}_t$ with $\bar{u}_t = N^{-1} \sum_{i=1}^N u_{it}$. Using (5.3), we can derive the CCEP

[4] Chudik et al. (2011) show that these factor models exhibit the strong form of cross-sectional dependence since the maximum eigenvalue of the covariance matrix for ε_{it} tends to infinity at rate N. On the other hand spatial econometric models, developed by Behrens et al. (2012), display the weak form of cross-sectional dependence, which can be represented by an infinite number of weak factors and no idiosyncratic error.

[5] For example, the specification employed by Bun and Klaassen (2007) is obtained by simply including t as one element in $\mathbf{s}_t$, but without unobserved factors, $\boldsymbol{\theta}_t$.

estimator of $\boldsymbol{\beta}$ in (5.4) by (5.4) below. Alternatively, $\boldsymbol{\beta}$ can be consistently estimated by the principal component (PC) estimator proposed by Bai (2009). Chudik et al. (2011) show that the PC estimator is similar to the CCEP estimator, except that the cross section averages are replaced by the estimated common factors $\left(\hat{\boldsymbol{\theta}}_t\right)$, which are obtained using the Bai and Ng (2002) procedure.[6] In this case we have $\mathbf{f}_t = \left(\mathbf{s}_t', \hat{\boldsymbol{\theta}}_t'\right)'$ in (5.3). Specifically, the CSD-consistent estimator of $\boldsymbol{\beta}$ is given by[7]

$$\hat{\boldsymbol{\beta}}_{CSD} = \left(\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{M}_T \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{M}_T \mathbf{y}_i\right), \ \hat{\boldsymbol{\beta}}_{CSD} = \hat{\boldsymbol{\beta}}_{CCEP} \text{ or } \hat{\boldsymbol{\beta}}_{PC} \quad (5.4)$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})'$, $\mathbf{M}_T = \mathbf{I}_T - \mathbf{H}_T \left(\mathbf{H}_T' \mathbf{H}_T\right)^{-1} \mathbf{H}_T'$, $\mathbf{H}_T = (\mathbf{1}_T, \mathbf{f})$, $\mathbf{1}_T = (1, ..., 1)'$ and $\mathbf{f} = (\mathbf{f}_1', ..., \mathbf{f}_T')'$.

Both CCEP and PC estimators are unable to estimate the coefficients, $\boldsymbol{\gamma}$ on time-invariant variables because they are the extended fixed effect estimators. In this regard, SS combine the CCEP estimation with the instrumental variables estimation proposed by Hausman and Taylor (1981, HT), and develop the CCEP-HT estimator. Baltagi (2010) further proposes the CCEP-AM estimator by employing the additional instrument variables proposed by Amemiya and MaCurdy (1986, AM). It is then straightforward to consider further additional set of instrument variables proposed by Breusch, Mizon and Schmidt (1989, BMS), which we denote by the CCEP-BMS estimator. We can also develop the corresponding counterparts, using the Bai's PC estimator, which we denote by PC-HT, PC-AM and PC-BMS estimators, respectively.

Conformable with Hausman and Taylor (1981), we decompose $\mathbf{x}_{it} = (\mathbf{x}_{1it}', \mathbf{x}_{2it}')'$ and $\mathbf{z}_i = (\mathbf{z}_{1i}', \mathbf{z}_{2i}')'$, where $\mathbf{x}_{1it}$, $\mathbf{x}_{2it}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors, and $\mathbf{z}_{1i}$, $\mathbf{z}_{2i}$ are $g_1 \times 1$ and $g_2 \times 1$ vectors. Then, we can estimate $\boldsymbol{\gamma}$ consistently using instrumental variables in the following regression:

$$d_{it} = \boldsymbol{\gamma}_1' \mathbf{z}_{1i} + \boldsymbol{\gamma}_2' \mathbf{z}_{2i} + \alpha_i^* + u_{it}^* = \mu + \boldsymbol{\gamma}' \mathbf{z}_i + \varepsilon_{it}^*, \ i = 1, ..., N, \ t = 1, ..., T, \quad (5.5)$$

where $d_{it} = y_{it} - \hat{\boldsymbol{\beta}}_{CSD}' \mathbf{x}_{it} - \boldsymbol{\lambda}_i' \mathbf{f}_t$, $\mu = E(\alpha_i^*)$ and $\varepsilon_{it}^* = (\alpha_i^* - \mu) + u_{it}^*$ is a zero mean process by construction. In matrix notation we have:

$$\mathbf{d} = \mu \mathbf{1}_{NT} + \mathbf{Z}_1 \boldsymbol{\gamma}_1 + \mathbf{Z}_2 \boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}^*, \quad (5.6)$$

where $\mathbf{d} = (\mathbf{d}_1', ..., \mathbf{d}_N')'$, $\mathbf{d}_i = (d_{i1}, ..., d_{iT})'$, $\mathbf{Z}_j = \left(\left(\mathbf{z}_{j1}' \otimes \mathbf{1}_T\right)', ..., \left(\mathbf{z}_{jN}' \otimes \mathbf{1}_T\right)'\right)'$, $j = 1, 2$, $\mathbf{1}_{NT} = (\mathbf{1}_T', ..., \mathbf{1}_T')'$, $\mathbf{1}_T = (1, ..., 1)'$, and $\boldsymbol{\varepsilon}^* = \left(\boldsymbol{\varepsilon}_1^{*\prime}, ..., \boldsymbol{\varepsilon}_N^{*\prime}\right)'$ with

[6] After applying the within transformation of the model, (5.1), we can extract the factors from the within residuals in an iterative manner.

[7] Under fairly standard regularity conditions, Pesaran (2006) and Bai (2009) prove that as $(N, T) \to \infty$ jointly, $\hat{\boldsymbol{\beta}}_{CSD}$ is consistent and follows the asymptotic normal distribution.

$\boldsymbol{\varepsilon}_i^* = (\varepsilon_{i1}^*, ..., \varepsilon_{iT}^*)'$. Replacing $\mathbf{d}$ by its consistent estimate, $\hat{\mathbf{d}} = \left\{\hat{d}_{it}, i = 1, ..., N,\ t = 1, ..., T, \right\}$, where $\hat{d}_{it} = y_{it} - \hat{\boldsymbol{\beta}}'_{CSD}\mathbf{x}_{it} - \hat{\boldsymbol{\lambda}}'_i\mathbf{f}_t$ and $\hat{\boldsymbol{\lambda}}_i$ are the OLS estimators of $\boldsymbol{\lambda}_i$ consistently estimated from the regression of $\left(y_{it} - \hat{\boldsymbol{\beta}}'_{CSD}\mathbf{x}_{it}\right)$ on $(1, \mathbf{f}_t)$ for $i = 1, ..., N$, we now have:

$$\hat{\mathbf{d}} = \mu\mathbf{1}_{NT} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}^+ = \mathbf{C}\boldsymbol{\delta} + \boldsymbol{\varepsilon}^+, \tag{5.7}$$

where $\boldsymbol{\varepsilon}^+ = \boldsymbol{\varepsilon}^* + \left(\hat{\mathbf{d}} - \mathbf{d}\right)$, $\mathbf{C} = (\mathbf{1}_{NT}, \mathbf{Z}_1, \mathbf{Z}_2)$ and $\boldsymbol{\delta} = (\mu, \boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)'$.

To deal with nonzero correlation between $\mathbf{Z}_2$ and $\boldsymbol{\alpha}$, we need to find the $NT \times (1 + g_1 + h)$ matrix of instrument variables:

$$\mathbf{W} = [\mathbf{1}_{NT}, \mathbf{Z}_1, \mathbf{W}_2],$$

where $\mathbf{W}_2$ is an $NT \times h$ matrix of instrument variables for $\mathbf{Z}_2$ with $h \geq g_2$ for identification. First, we follow SS and consider the $NT \times (k_1 + \ell)$ HT instrument matrix given by

$$\mathbf{W}_2^{HT} = \left[\mathbf{P}\mathbf{X}_1, \mathbf{P}\hat{\boldsymbol{\xi}}_1, \mathbf{P}\hat{\boldsymbol{\xi}}_2, ..., \mathbf{P}\hat{\boldsymbol{\xi}}_\ell\right]$$

where $\mathbf{P} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ is the $NT \times NT$ idempotent matrix with $\mathbf{D} = \mathbf{I}_N \otimes \mathbf{1}_T$, $\mathbf{I}_N$ being an $N \times N$ identity matrix, and $\hat{\boldsymbol{\xi}}_j = \left(\hat{\lambda}_{j,1}\mathbf{f}'_j, \hat{\lambda}_{j,2}\mathbf{f}'_j, ..., \hat{\lambda}_{j,N}\mathbf{f}'_j\right)'$, $j = 1, ..., \ell$, where $\mathbf{f}_j = (f_{j,1}, ..., f_{j,T})'$ with $\hat{\lambda}_{j,i}$ being consistent estimate of heterogenous factor loading, $\lambda_{j,i}$. Next, we follow Baltagi (2010) and derive the $NT \times (k_1 + \ell + Tk_1 + T\ell)$ AM instrument matrix by

$$\mathbf{W}_2^{AM} = \left[\mathbf{W}_2^{HT}, (\mathbf{Q}\mathbf{X}_1)^*, \left(\mathbf{Q}\hat{\boldsymbol{\xi}}_1\right)^*, \left(\mathbf{Q}\hat{\boldsymbol{\xi}}_2\right)^*, ..., \left(\mathbf{Q}\hat{\boldsymbol{\xi}}_\ell\right)^*\right] \tag{5.8}$$

where $\mathbf{Q} = \mathbf{I}_{NT} - \mathbf{P}$ and $(\mathbf{Q}\mathbf{X}_1)^* = (\mathbf{Q}\mathbf{X}_{11}, \mathbf{Q}\mathbf{X}_{12}, ..., \mathbf{Q}\mathbf{X}_{1T})$ is the $NT \times k_1T$ matrix with $\mathbf{Q}\mathbf{X}_{1t} = (\mathbf{Q}\mathbf{X}_{11t}, ..., \mathbf{Q}\mathbf{X}_{1kt})'$.[8] Finally, it is straightforward to derive the $NT \times (k_1 + \ell + Tk_1 + T\ell + Tk_2)$ BMS instrument matrix by

$$\mathbf{W}_2^{BMS} = \left[\mathbf{W}_2^{AM}, (\mathbf{Q}\mathbf{X}_2)^*\right]$$

where $(\mathbf{Q}\mathbf{X}_2)^* = (\mathbf{Q}\mathbf{X}_{21}, \mathbf{Q}\mathbf{X}_{12}, ..., \mathbf{Q}\mathbf{X}_{2T})$.[9]

To derive the consistent estimator of $\boldsymbol{\delta}$, we premultiply $\mathbf{W}'$ by (5.7)

$$\mathbf{W}'\hat{\mathbf{d}} = \mathbf{W}'\mathbf{C}\boldsymbol{\delta} + \mathbf{W}'\boldsymbol{\varepsilon}^+. \tag{5.9}$$

Therefore, the GLS estimator of $\boldsymbol{\delta}$ is obtained by

$$\hat{\boldsymbol{\delta}}_{GLS} = \left[\mathbf{C}'\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\mathbf{C}\right]^{-1}\mathbf{C}'\mathbf{W}\mathbf{V}^{-1}\mathbf{W}'\hat{\mathbf{d}}, \tag{5.10}$$

[8] Notice that the rank of $(\mathbf{Q}\mathbf{X}_1)^*$ is $(T-1)k_1$, because only $(T-1)$ deviations from means are (linearly) independent since each variable (see BMS). Similarly for $\left(\mathbf{Q}\hat{\boldsymbol{\xi}}_1\right)^*$, ..., $\left(\mathbf{Q}\hat{\boldsymbol{\xi}}_\ell\right)^*$.

[9] As before, the rank of $(\mathbf{Q}\mathbf{X}_2)^*$ is only $(T-1)k_2$.

where $\mathbf{V} = Var\left(\mathbf{W}'\boldsymbol{\varepsilon}^{+}\right)$. To obtain the feasible GLS estimator we replace $\mathbf{V}$ by its consistent estimator. In practice, estimates of $\boldsymbol{\delta}$ and $\mathbf{V}$ can be obtained iteratively until convergence, see also SS for further details.

Notice that the HT-IV estimator employs only the mean of $\mathbf{X}_1$ to be uncorrelated with the effects, α_i^* whereas the AM-IV estimator exploits such moment conditions to be held at every time period. Hence, the validity of the AM instruments requires a stronger exogeneity assumption for $\mathbf{X}_1$, under which the AM-IV estimator is more efficient than HT-IV. Furthermore, the BMS instruments require uncorrelatedness of $\mathbf{X}_2$ with α_i^* separately at every point in time. The validity of the AM and the BMS instruments can be easily tested via the Hausman statistics testing for the difference between HT-IV and AM-IV estimators and between AM-IV and BMS-IV estimators as follows:

$$\begin{aligned} H_{AM} &= \left(\hat{\boldsymbol{\delta}}_{AM} - \hat{\boldsymbol{\delta}}_{HT}\right)' \left[Var\left(\hat{\boldsymbol{\delta}}_{HT}\right) - Var\left(\hat{\boldsymbol{\delta}}_{AM}\right)\right]^{-1} \left(\hat{\boldsymbol{\delta}}_{AM} - \hat{\boldsymbol{\delta}}_{HT}\right) \\ H_{BMS} &= \left(\hat{\boldsymbol{\delta}}_{BMS} - \hat{\boldsymbol{\delta}}_{AM}\right)' \left[Var\left(\hat{\boldsymbol{\delta}}_{AM}\right) - Var\left(\hat{\boldsymbol{\delta}}_{BMS}\right)\right]^{-1} \left(\hat{\boldsymbol{\delta}}_{BMS} - \hat{\boldsymbol{\delta}}_{AM}\right) \end{aligned}$$

both of which follow the asymptotic χ^2_g distribution with the degree of freedom g being the number of coefficients tested.

5.5.3 MSS (2013) extension

Recently, an investigation of unobserved and time-varying multilateral resistance and omitted trade determinants has assumed a prominent role in order to measure the Euro effects on trades precisely. We implement two methodologies: the factor-based gravity model by Serlenga and Shin (2013) and the spatial-based techniques by Behrens, Ertur and Kock (2012, BEK), both of which allow trade flows and error terms to be cross-sectionally correlated. Applying these approaches to the dataset over 1960-2008 for 190 country-pairs of 14 EU and 6 non-EU OECD countries, we find that the Euro impact estimated by the factor-based model amounts to 4-5% only, far less than 20% estimated by the spatial-based model. The cross-section dependency test results also confirm that the factor-based model is more appropriate in accommodating correlation between regressors, and unobserved individual and time effects. Overall we may conclude that the trade-creating effects of the Euro should be viewed in the proper historical and multilateral perspective rather than in terms of the formation of a monetary union as an isolated event.

Alternatively, we now investigate the issue of CSD among trade flows through employing the spatial techniques. This approach assumes that the structure of cross section correlation is related to the location and the distance among units on the basis of a pre-specified weight matrix. Hence, cross section correlation is represented mainly by means of a spatial process, which

explicitly relates each unit to its neighbours. A number of approaches for modeling spatial dependence has been suggested in the spatial literature. The most popular ones are the Spatial Autoregressive (SAR), the Spatial Moving Average (SMA), and the Spatial Error Component (SEC) specifications. The spatial panel data model is estimated using the maximum likelihood (ML) or the generalized method of moments (GMM) techniques (*e.g.*, Elhorst, 2011). We follow BEK and consider a spatial panel data gravity (SARAR) model, which combines a spatial lagged variable and a spatial autoregressive error term:

$$y_{it} = \rho y_{it}^{*} + \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\gamma}'\mathbf{z}_i + \tilde{\alpha}_i + v_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{5.11}$$

$$v_{it} = \lambda v_{it}^{*} + u_{it} \tag{5.12}$$

where $y_{it}^{*} = \sum_{j \neq i}^{N} w_{ij} y_{jt}$ is the spatial lagged variable, and $v_{it}^{*} = \sum_{j \neq i}^{N} w_{ij} v_{jt}$ is the spatial autoregressive error term, w_{ij}'s are the spatial weight with the row-sum normalisation, $\sum_i w_{ij} = 1$, and u_{it} is a zero mean idiosyncratic disturbance with constant variance. This approach is especially designed to deal with CSD across both variables and error terms in which ρ is the spatial lag coefficient and λ refers to the spatial error component coefficient. These coefficients capture the spatial spillover effects and measure the influence of the weighted average of neighboring observations on cross section units. Chudik *et al.* (2011) show that a particular form of a weak cross dependent process arises when pairwise correlations take non-zero values only across finite units that do not spread widely as the sample size rises. A similar case occurs in the spatial processes, where the local dependency exists only among adjacent observations. In particular, Pesaran and Tosetti (2011) show that spatial processes commonly used, such as the SAR or the SMA process, can be represented by a process with an infinite number of weak factors and no idiosyncratic error terms.

Both the factor- and the spatial-based models cannot estimate the coefficients, $\boldsymbol{\gamma}$ on time-invariant variables in the presence of fixed effects. In this regard, we follow SS and combine these estimators with the instrumental variables estimation. We denote such estimators by the PCCE-HT, PCCE-AM, PCCE-BMS, PC-HT, PC-AM, PC-BMS, SARAR-HT, SARAR-AM, and SARAR-BMS estimators, respectively.

In particular, we follow LeSage and Pace (2009), and discuss the estimation results for the spatial gravity model in terms of direct and indirect effects. To this end we rewrite (5.11):

$$\mathbf{y}_t = \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t,\ t = 1, ..., T \tag{5.13}$$

where $\mathbf{y}_t = (y_{1t}, ..., y_{Nt})'$, $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{N}$ is the $N \times N$ spatial weight matrix, $\mathbf{X}_t = (\mathbf{x}_{1t}', ..., \mathbf{x}_{Nt}')$ is the $N \times k$ matrix of time-varying regressors, $\mathbf{Z} = (\mathbf{z}_1', ..., \mathbf{z}_N')$ is the $N \times g$ matrix of time-invariant regressors,and $\boldsymbol{\varepsilon}_t =$

$(\varepsilon_{1t}, ..., \varepsilon_{Nt})'$ with $\varepsilon_{it} = \tilde{\alpha}_i + v_{it}$. We then rewrite (5.13) as

$$\mathbf{y}_t = (\mathbf{I}_N - \rho\mathbf{W})^{-1} (\mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t). \tag{5.14}$$

Then, the impacts of a change in the rth time-varying regressor corresponds to the following $N \times N$ matrix of partial derivatives:

$$\frac{\partial \mathbf{y}_t}{\partial X_{rt}} = (\mathbf{I}_N - \rho\mathbf{W})^{-1} \beta_r, \ r = 1, ..., k \tag{5.15}$$

Notice that diagonal elements of (5.15) (direct impacts), are different across cross-section units; off-diagonal terms (indirect impacts) differ from zero, and the matrix is not symmetric. We now have N direct effects and $N(N-1)$ indirect effects. To avoid such an interactive heterogeneity issue, LeSage and Pace (2009) suggest to employ only three scalar measures to summarise information contained in the matrix (5.15): the average of the N diagonal elements as a measure of direct effects, the average of the $N(N-1)$ off-diagonal elements as the average of the cumulative indirect effects and the average total effect as the mean of total effects.

The estimation results for the factor-based model provide the following stylised findings: First, the impacts of distance and common language on trade are significantly negative and positive whereas the border impact is insignificant. Further investigation of their time-varying coefficients reveals that border and language effects started to fall more sharply after 1999. More importantly, we find that both the Euro and the custom union impacts on trade amounts to 4-5% and 11% only. These findings support the thesis that the potential trade-creating effects of the Euro should be viewed in terms of the proper historical and multilateral perspective rather than simply in terms of the formation of a monetary union as an isolated event. Next, the estimation results for the spatial-based gravity model indicate that the impacts of the Euro and the custom union on trade rises to 20% and 30%, respectively, both significantly higher than those obtained by the PCCE and the PC estimators. Furthermore, the CD test results confirm that the factor-based model is able to better accommodate correlation between regressors, unobserved individual and time effects. This evidence highlights an importance of appropriately controlling for CSD in the panel gravity models of trade flows through the use of both observed and unobserved factors in order to account for time-varying multilateral resistance, trade costs and globalisation trends.

5.6 A Nonlinear Panel Data Model of Cross-Sectional Dependence

Kapetanios, Mitchell and Shin (2014) propose a nonlinear panel data model which can endogenously generate both 'weak' and 'strong' cross-sectional dependence. The model's distinguishing characteristic is that a given agent's

behaviour is influenced by an aggregation of the views or actions of those around them. The model allows for considerable flexibility in terms of the genesis of this herding or clustering type behaviour. At an econometric level, the model is shown to nest various extant dynamic panel data models. These include panel AR models, spatial models, which accommodate weak dependence only, and panel models where cross-sectional averages or factors exogenously generate strong, but not weak, cross sectional dependence. An important implication is that the appropriate model for the aggregate series becomes intrinsically nonlinear, due to the clustering behaviour, and thus requires the disaggregates to be simultaneously considered with the aggregate. We provide the associated asymptotic theory for estimation and inference. This is supplemented with Monte Carlo studies and two empirical applications which indicate the utility of our proposed model as a vehicle to model different types of cross-sectional dependence.

We propose nonlinear panel data models. The distinguishing characteristic is the use of unit-specific aggregates/summaries of past values of variables relating to other units that are 'close' in some sense to a given unit, to model that unit. The nature of the models is dynamic:

$$x_{i,t} = \rho \sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) x_{j,t-1} + \epsilon_{i,t}, \quad i = 1, ...N, \quad t = 1, ...T, \tag{5.16}$$

where $x_{-i,t} = (x_{1,t}, x_{2,t}, ..., x_{i-1,t}, x_{i+1,t}, ..., x_{Nt})'$ and $\sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) = 1$. This form of the model is extremely general and simply signifies that $x_{i,t}$ depends, possibly in a nonlinear fashion depending on how w_{ij} is parameterised, on weighted averages of past values of $x_t = (x_{1,t}, ..., x_{Nt})'$, where the weights depend on x_{t-1}. We split x_{t-1} into $x_{-i,t-1}$ and $x_{i,t-1}$ to emphasise the potentially special role of the own lag of $x_{i,t}$ in the specification. One particular motivation is structural and follows from the claim that it mimics structural interactions between economic units. Another, more econometric, justification simply notes that this model can accommodate generic forms of cross-sectional dependence, including evolving clusters.

The model in (5.16) is extremely general as it encompasses a wide variety of nonlinear specifications. We consider a number of particular nonlinear specifications for the construction of the unit specific aggregates. We place particular emphasis on specifications where the weights depend on x_{t-1} only through distances of the form $|x_{j,t-1} - x_{i,t-1}|$. We choose a particular specification of this type that is easy to analyse, based on a threshold mechanism, to illustrate the class of models we focus on. This model nests a variety of dynamic panel data models, such as panel data AR models and panel models where cross-sectional averages are used to pick up cross-sectional dependence (e.g., Pesaran, 2006). Interestingly, it is also closely related to factor models, that have received considerable attention recently (Bai, 2009).

Our models provide an intuitive means by which many forms of cross-

sectional dependence can arise in a large panel dataset comprised of variables of a 'similar' nature that relate to different agents/units. These variables might be the disaggregates underlying often studied macroeconomic or financial aggregates, such as economy-wide inflation or the S&P500 index. In particular, the model allows these different economic units to cluster; and for these clusters (including their number) to evolve over time. Such clustering also has implications when modelling and forecasting the aggregate of these units.

The degree of cross-sectional dependence in our models can vary, from a case where it is similar to standard factor models, for which the largest eigenvalue of the variance covariance matrix of the data tends to infinity at rate N, to the case of very weak or no factor structure where this eigenvalue is bounded as $N \to \infty$. Of course, all intermediate cases can arise as well. Our work can be viewed as a particular instance of a large dimensional VAR, but for the fact that our model is intrinsically nonlinear in nature.

We provide an analysis of the stochastic version of the model; and allow for both threshold and smooth transition type nonlinearities. Our model constitutes, to the best of our knowledge, the first attempt to introduce endogenous cross-sectional dependence into a panel modelling framework. Our work concentrates on the case where $|\rho| < 1$, and subsequently our model has particular stationarity properties. We do briefly discuss the 'unit root' case where $|\rho| = 1$, but our treatment is indicative and not comprehensive. We defer detailed treatment to future research.

5.6.1 Model

We propose a particular dynamic panel model for a multitude of agents. Let $x_{i,t}$ denote the variable of interest, such as the agent's income or the agent's view of the future value of some macroeconomic variable, at time t, for agent i. Then, we specify

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \epsilon_{i,t}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{5.17}$$

where

$$m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right),$$

$\{\epsilon_{i,t}\}_{t=1}^{T}$ is an error process, $\mathcal{I}(.)$ is the indicator function and $-1 < \rho < 1$. Verbally, the above model states that $x_{i,t}$ is influenced by the cross-sectional average of a selection of $x_{j,t-1}$ and in particular that the relevant $x_{j,t-1}$ are those that lie closest to $x_{i,t-1}$. The model involves a K nearest neighbour mechanism except that it is in the data generating process and not as a technique to estimate an unknown function. This formalises the intuitive

idea that people are affected more by those with whom they share common views or behaviour. The model may be equally viewed as a descriptive model of agents' behaviour, reflecting the fact that 'similar' agents are affected by 'similar' effects, or as a structural model of agents' views whereby agents use the past views of other agents, similar to them to form their own views. The interaction term in (5.17) may then be thought of capturing the (cross-sectional) local average or common component of their views. This idea of commonality has various clear, motivating and concrete examples in a variety of social science disciplines, such as psychology and politics. In economics and finance, the herding could be rational (imitative herding) or irrational.

A deterministic form of the above model has been analysed previously in the mathematical and system engineering literature. In particular, they have analysed a continuous form of the restricted version of (5.17) given by

$$x_{i,t} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right) x_{j,t-1}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{5.18}$$

where $m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right)$. To the best of our knowledge, we are the first to introduce a stochastic term to this type of model and to allow for an unknown value of the threshold parameter.

The model, (5.17), bears considerable resemblance to threshold autoregressive (TAR) models analysed in the time-series literature. However, unlike straightforward extensions of TAR models to a panel setting, whereby individual units/agents would not interact through the nonlinear specification, the nonlinearity in (5.17) is inherently cross-sectional in nature; this provides for the development of a dynamic network effect. In deterministic contexts this has been shown to generate interesting behaviour like clustering.

5.6.2 Special cases

It is interesting to note the nature of restricted versions of the above model, obtained by taking extreme values of the threshold parameter. By setting $r = 0$, we obtain a simple panel autoregressive model

$$x_{i,t} = \rho x_{i,t-1} + \epsilon_{i,t}. \tag{5.19}$$

On the other hand, letting $r \to \infty$, we obtain the following model

$$x_{i,t} = \frac{\rho}{N} \sum_{j=1}^{N} x_{j,t-1} + \epsilon_{i,t}, \tag{5.20}$$

where past cross-sectional averages of opinions inform, in similar fashions, current opinions. Recently, the use of such cross-sectional averages has been

advocated by Pesaran (2006) as a means of modelling cross-sectional dependence in the form of unobserved factors. However, unlike these models where the use of cross-sectional averages is an approximation to the unknown model, in our case this is a limiting case of a 'structural' nonlinear model.

It is important to investigate the statistical properties of our model. A number of results, stated and proved in the appendix, provide help in this respect. Intuitively, as we show in Appendix, (5.17) is geometrically ergodic, and therefore asymptotically stationary, if $|\rho| < 1$. This allows for the analysis of estimators along traditional lines.

5.6.3 Cross-sectional dependence and factor models

In the factor literature the behaviour of the covariance matrix of $x_t = (x_{1,t}, ..., x_{N,t})'$ is considered. Factor models have the property that both the maximum eigenvalue and the row/column sum norm of the covariance matrix tend to infinity at rate N, as $N \to \infty$. In contrast, for other models of cross-sectional dependence such as, for example, spatial AR or MA models, these quantities are bounded, implying that they exhibit much lower degrees of cross-sectional dependence than factor models.

We show that the column sum norm of the variance covariance matrix of x_t when x_t follows (5.17) is $O(N)$. Thus, the model is more similar to factor models than spatial AR or MA models. Interestingly, there are versions of (5.17) that resemble spatial models more than factor models. Another finding is that (5.20) implies a variance covariance matrix for x_t with a column sum norm that is $O(1)$. This is surprising, given the similarity that cross-sectional average schemes have with factor models as detailed in Pesaran (2006). In our case no exogenous factors exist and the cross-sectional average is a primitive term that exists in the structure of the model.

It is important to restate here differences between our model and a factor model. When a dataset has pronounced cross-sectional dependence exhibited by, say, exploding eigenvalues or the column sum norm associated with its covariance matrix, then a factor model should offer some fit, irrespective of the structural form giving rise to this cross-sectional dependence. Principal components, in particular, nonparametrically construct linear combinations of the variables that capture (strong) cross-sectional dependence, whatever its genesis. But when the data generating process resembles our model, such that clusters emerge endogenously and their number varies over time, a large number of factors may be required; and the number needed may also have to change over time. Factor models are intrinsically reduced form; they focus on modelling cross-sectional dependence using an exogenously given number of unobserved factors. Since our model nests (5.20), it is not surprising that it can approximate a factor model when $r \to \infty$. On the other hand, our model has a clear parametric structure, which is a fea-

ture shared by some classes of dynamic spatial model. But our models are more general than spatial models, in the sense that the weighting schemes are estimated endogenously, rather than assumed *ex ante*. Furthermore, it is worth noting that the factor model cannot accommodate the weak cross-sectional dependence seen in spatial models, in contrast to the extensions of our nonlinear model. These extensions demonstrate that the nonlinear model can, in general, be seen to lie between the two extremes characterised by weakly cross-sectionally dependent spatial models and strongly cross-sectionally dependent factor models.

5.6.4 General suggestions on the empirical applications including the herding

Inflation expectations Here, we have considered the basic model with fixed effects as:

$$\pi_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1} + \epsilon_{i,t} \tag{5.21}$$

where π is the one-quarter ahead CPI inflation rate forecast, $\nu_i \sim iid(0, \sigma_\nu^2)$, and obtained the within estimator of ρ along with the consistent estimator of r. We can provide some economic interpretations for $\hat{\rho}$ and $\hat{r}$ in terms of persistence or inertia and the relative distance of similarity.

Notable exceptions in the current application are to estimate two extreme models, denoted PAR and CSA, respectively:

$$\pi_{i,t} = \nu_i + \rho\pi_{i,t-1} + \epsilon_{i,t} \tag{5.22}$$

$$\pi_{i,t} = \nu_i + \rho\bar{\pi}_{t-1} + \epsilon_{i,t} \tag{5.23}$$

It is interesting to see how the estimates of ρ differ for each of three models, along other statistical measures as discussed in the case with the the stock return application. Assuming that the overall performance of the model, (5.21), is superior and after carrying out further CDS analyses with the exogenous factor structure, we then move to estimate the extension or generalisation as discussed in the model of the form, namely,

$$\pi_{i,t} = \nu_i + \rho_1\tilde{\pi}_{i,t-1} + \rho_2\tilde{\pi}^c_{i,t-1} + \epsilon_{i,t} \tag{5.24}$$

where $\tilde{\pi}_{i,t-1}$ and $\tilde{\pi}^c_{i,t-1}$ are the respective cross-section averages related to similar and dissimilar forecasters given by

$$\tilde{\pi}_{i,t-1} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1}$$

$$\tilde{\pi}^{c}_{i,t-1} = \frac{1}{N - m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| > r\right) \pi_{j,t-1}$$

In this context, we can also test the hypothesis of $\rho_2 = 0$ (non-informational contents arising from dissimilar forecasters). If not, what's the prior implication of the signs of ρ_2? In other words, in forming the own forecast, how does each forecaster use any (past) information from others? At least in this context, we can avoid any contemporaneous endogeneity issue?

Next issue is to find a way of decomposing the (partial) aggregate parameter, ρ in (5.21) or $\rho_1$in (5.24) into the own effect and the neighbor effect. One obvious candidate is to consider

$$\pi_{i,t} = \nu_i + \rho_0 \pi_{i,t-1} + \rho_1 \hat{\pi}_{i,t-1} + \epsilon_{i,t} \tag{5.25}$$

$$\pi_{i,t} = \nu_i + \rho_{10} \pi_{i,t-1} + \rho_{11} \hat{\pi}_{i,t-1} + \rho_2 \tilde{\pi}^{c}_{i,t-1} + \epsilon_{i,t} \tag{5.26}$$

where

$$\hat{\pi}_{i,t-1} = \frac{1}{\hat{m}_{i,t}} \sum_{j=1, j\neq i}^{N} \mathcal{I}\left(|\pi_{i,t-1} - \pi_{j,t-1}| \leq r\right) \pi_{j,t-1}$$

In the spatial modelling literature, Anselin et al. (2008) distinguish spatial dynamic models into four categories based on the following general time-space-dynamic specification:

$$x_{i,t} = \rho_0 x_{i,t-1} + \rho_1 \sum_{j\neq i} w_{ij} x_{j,t-1} + \beta \sum_{j\neq i} w_{ij} x_{j,t} + v_i + \epsilon_{i,t} \tag{5.27}$$

Here, $\sum_{j\neq i} w_{ij} y_{jt}$ and $\sum_{j\neq i} w_{ij} y_{jt-1}$ are *a first order spatial lag* and its time-lagged value, respectively. The parameter, ρ_0, captures serial dependence of x_{it}, β represents the intensity of a contemporaneous spatial effect ,and ρ_1 captures *space time autoregressive dependence*. Most studies focus on the stable case. with $|\rho_0 + \rho_1 + \beta| < 1$, but Lee and Yu (2007) also develop the unit root analysis of a spatial dynamic panel). This general specification, (5.27), includes various special cases of spatial lag models on panel data discussed in the literature:

- if $\rho_0 = \rho_1 = 0$, we obtain a 'pure-space recursive' model in which dependence results from the neighborhood locations in the previous time period;
- if $\rho_1 = 0$, the model is reduced to a 'time space recursive' model in which dependence relates to both the location itself ($x_{i,t-1}$) and its neighbors in the previous time period $\sum_{j\neq i} w_{ij} x_{j,t-1}$;
- if $\beta = 0$, we obtain a 'time space simultaneous' model which includes the time lag ($x_{i,t-1}$) and the spatial lag, $\sum_{j\neq i} w_{ij} x_{j,t}$;

- finally, if $\rho_0 = \beta = 0$, we are dealing with a spatial autoregressive model on panel data, while if $\rho_0 = \rho_1 = \beta = 0$ we obtain a 'simple' dynamic model.

According to Anselin (2001) and Abreu et al. (2005), the addition of a spatially lagged dependent variable causes simultaneity and endogeneity problems and thus a candidate consistent estimator should lie between the OLS and within estimates.

Notice that our model, (5.25), is similar to the time-space recursive model considered in Korniotis (2010), who apply this model to investigate the issue of internal vs external habit formation using the annual consumption data for the U.S. states, and find that state consumption growth is not significantly affected by its own (lagged) consumption growth but it is affected by lagged consumption growth of nearby states. Notice that the weight w_{ij} measures the importance of $x_{j,t-1}$ on x_{it}. The weights are observed quantities, which are known to the econometrician, and they are therefore exogenous. Because the spatial lag, $\sum_{j=1}^{N} w_{ij}x_{j,t-1}$, is a weighted average of past consumption choices of other cross-sectional units, it is the measure of the catching-up habit. The weights w_{ij} are organized in the $N \times N$ spatial matrix $\mathbf{W}$.

Currently, there is a trade-off between the model, (5.25), and the time space recursive model employed by Korniotis (2010). In (5.25), the neighbors are selected endogenously but the equal weights are imposed to the selected neighbors. By contrast, in the time space recursive model, the neighbors are selected more or less exogenously but the weights, though not time-varying mostly, are sometimes given in a flexible manner. In any case the application of the model, (5.25) to similar issue of the consumption habit formation will provide an interesting insight.[10]

Unless $\rho_2 = 0$, the model (5.26) should be more general.

Alternative decomposition I am considering how to relate or modify the Sias' (2004) approach to an analysis of herding. The idea is to estimate ρ from (5.21), and find a way to decompose:

$$\rho = \rho_{own} + \rho_{neighbor}$$

following the the Sias' approach. But the analogy is not quite one-to-one. The potential advantage of this approach is the possible robustness of this measure which can also be used for a finite T, as well.

[10] Of course, we can allow the weights to be inversely proportional to the distance once the threshold parameter is consistently estimated. See further discussions below.

The few more suggestions As discussed above, we can allow different weights to the selected neighbors as follows: We now generalise (5.21)

$$x_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) w_{ij} x_{j,t-1} + \epsilon_{i,t} \tag{5.28}$$

where we may consider the following weights

$$w_{ij} = \frac{d_{ij}^{-2}}{\sum_{j=1}^{N} d_{ij}^{-2}}, \; d_{ij} = |x_{i,t-1} - x_{j,t-1}|$$

(then, how we define w_{ii}, just normalised to 1?). The estimation can be done in two steps: first, the consistent estimate of r is obtained from (5.21). Then, construct the weights and the associated cross-section averages and estimate the model, (5.28). Or possibly more complicated due to the grid search over r. If successful, then our approach is more general than the spatial model as discussed above.

Next, we may consider the following extension of (5.21):

$$x_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(\left|x_{t-1}^{\max} - x_{j,t-1}\right| \leq r\right) x_{j,t-1} + \epsilon_{i,t} \tag{5.29}$$

where $x_{t-1}^{\max} = \max_j x_{j,t-1}$, such that the distance is measured with respect to the best performer rather than the unit i. The alternative functional type can also be considered. This may be the better example that those discussed on p.19.

From the empirical point of view, the following consideration may be useful: Suppose that the distance between $x_{i,t-1}$ and $x_{j,t-1}$ or more generally between q_{it} and q_{jt}, can be regarded as the sort of similarity measure, and that the parameter may measure the impact of the certain policy. We then estimate the value of ρ's under different values of r, and make a 2-dimensional plot to investigate whether the relationship between ρ and r is monotonic or nonlinear and so on. This approach may be related to the recent GMM analytic approach where the sample moment condition is not equal to zero for over-identified case.

5.7 Modelling Technical Efficiency in Cross Sectionally Dependent Stochastic Frontier Panels

Mastromarco, Serlenga and Shin (2014) propose a unified framework for accommodating both time- and cross-section dependence in modelling technical efficiency in stochastic frontier models. In particular, we adopt the multi-step procedure advanced by Bailey et al. (2013) within the nonlinear

panel data model advanced by Kapetanios et al. (2014, KMS). This extended KMS approach enables us to deal with both weak and strong forms of cross section dependence by introducing exogenously driven common factors and an endogenous threshold selection mechanism. Using the dataset of 26 OECD countries over the period 1970-2010, we provide the satisfactory estimation results for the production technology parameters and the associated efficiency ranking of individual countries. We find positive spillover effect on efficiency, supporting the hypothesis that knowledge spillover is more likely to be induced by technological proximity. Furthermore, our approach enables us to identify efficiency clubs endogenously in the stochastic frontier analysis.

We address an issue of modelling cross section dependence in the stochastic frontier analysis (SFA), and begin with the standard Cobb-Douglas production function:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{5.30}$$

where y_{it} is a logarithm of output of country i at time t, $\mathbf{x}_{it}$ a $k \times 1$ vector of (logged) production inputs, $\boldsymbol{\beta}$ a $k \times 1$ vector of structural parameters, and ε_{it} is the composite stochastic errors including the idiosyncratic disturbance (v_{it}) and time varying (logged) technical inefficiency (u_{it}):

$$\varepsilon_{it} = v_{it} - u_{it}. \tag{5.31}$$

Mastromarco *et al.* (2013) introduce the importance of a factor-based production function which takes into account of strong cross section dependence, and propose the panel stochastic frontier model with unobserved time-varying factors for modelling the time-varying technical inefficiency, u_{it}:

$$u_{it} = \alpha_i + \boldsymbol{\lambda}_i'\mathbf{f}_t,\ i = 1, ..., N,\ t = 1, ..., T, \tag{5.32}$$

where α_i is (unobserved) time-invariant individual effects, and $\mathbf{f}_t$ is an $r \times 1$ vector of unobserved factors that are expected to provide a proxy for nonlinear and complex trending patterns associated with globalisation and the business-cycle. This factor approach clearly accommodates strong cross section dependence (CSD). As will be clear, we observe the pervasive evidence of strong CSD among technical inefficiency, u_{it}.

Recent literature emphasises that the individual country's total factor productivity (TFP) is likely to be significantly affected by economic performance of neighboring or frontier countries. To allow for such spatial dependence structure, Ertur and Koch (2007) develop a growth model in which technological interdependency is specified through spatial externalities as the knowledge in one country produces externalities that may spillover into other countries. They provide evidence that the spatially augmented Solow model can produce the better prediction of the important role played by

spillover effects in international growth and convergence. In particular, the productivity shocks in SFA are assumed to be spatially correlated, typically, as follows:

$$\boldsymbol{\varepsilon}_t = \rho \mathbf{W} \boldsymbol{\varepsilon}_t + \mathbf{e_t},\ t = 1, ..., T, \tag{5.33}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$, $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{N}$ is the $N \times N$ spatial weight matrix with diagonal elements equal to zero, ρ is a spatial autoregressive parameter, and $\mathbf{e_t} = (e_{1t}, ..., e_{Nt})$ the vector of zero-mean idiosyncratic disturbances. The key assumption here is that the elements in $\mathbf{W}$ are selected exogenously on the basis of geographic or economic proximity measures such as contiguity, physical/economic/climatic distances or dissimilarities.

The spatial models can only control for weak CSD whilst the factor-based models can allow for strong CSD (Pesaran and Tosetti, 2011). In this regard the spatial-based approach is likely to produce biased estimates in the presence of strong CSD. While spatial autoregressive models which control for weak forms of CSD are generally estimated by MLE, Pesaran (2006) and Bai (2009) develop two alternative consistent estimation methodologies in the presence of strong CSD. Pesaran proposes the Pooled Common Correlated Effects (PCCE) by approximating unobserved common factors by cross-sectional averages of dependent and independent variables. Bai allows regressors to be correlated with both factors and loadings by including additive and interactive fixed effects, and proposes an maximum likelihood estimation method (IPC) in which unobservable common factors can be consistently estimated by the principal components. These studies clearly suggest that factors can play an important role in the cross sectionally correlated panels.

Nevertheless, factor-based models impose an assumption that the strong CSD is mainly driven by an exogenously given unobserved factors. Recently, KMS propose an alternative approach that allows the structure of CSD to be determined endogenously. In this paper, we combine factor-based model with the KMS approach and propose a consistent estimator of time varying efficiency which controls for both strong and weak CSD. Bai and Pesaran impose a common structure behind efficiency whilst KMS allow the efficiency cluster to be determined endogenously.

5.7.1 The model

The distinguishing feature of our model is the use of unit-specific aggregates, which summaries of past values of efficiency, and connects the units that are close to the technology frontier (the best units). The product of a country i at time t, Y_{it}, is determined by the levels of labor input and private capital, L_{it} and K_{it}. It is also affected by the Hicks-neutral multi-factor productivity TFP. The production function is expressed:

$$Y_{it} = TFP_{it} F(L_{it}, K_{it}), \tag{5.34}$$

where TFP_{it} depends on the technological progress. In the econometric specification, we hypothesize that countries differ for the efficiency in factor usage. The TFP_{it} component can be decomposed into the level of technology A_{it}, a measurement error w_{it}, and the efficiency measure τ_{it} with $0 < \tau_{it} \leq 1$:

$$TFP_{it} = A_{it}\tau_{it}w_{it}. \tag{5.35}$$

By writing (5.34) in log form:

$$y_{it} = \alpha + \beta_1 k_{it} + \beta_2 l_{it} - u_{it} + v_{it}, \tag{5.36}$$

with the two-way error components structure given by

$$\varepsilon_{it} = v_{it} - u_{it}, \tag{5.37}$$

where $v_{it} = \ln w_{it}$ and $u_{it} = -\ln(\tau_{it})$ is the term measuring the (time-varying) technical inefficiency.

Specifically, we propose an inefficiency model which might be given a behavioural interpretation such that innovators consider the behaviour of other agents as:

$$u_{it} = \alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t. \tag{5.38}$$

where

$$\tilde{u}_{it}(r) = \frac{1}{m_{it}}\sum_{j=1}^{N} I\left(\left|u_{t-1}^* - u_{jt-1}\right| \leq r\right) u_{jt-1}, \tag{5.39}$$

and r is the threshold parameter that is determined endogenously and u_{t-1}^* is the efficiency of the best performing country and $m_{it} = \sum_{j=1}^{N} I\left(\left|u_{t-1}^* - u_{jt-1}\right| \leq r\right)$. Model (5.38) extends (5.32) by including $\tilde{u}_{it}(r)$. This interaction term may then be thought of capturing the cross sectional local average of the best practices or common technology.

The specification in (5.38) explicitly allows the dynamics of technical inefficiency to be interacted spatially and enables us to address the spatial spillover effects such as the diffusion of new technologies. *A priori*, we expect that such externalities can be captured by a negative value of ρ. As a by-product of the KMS approach, we can also identify the heterogeneous technology clubs that may vary over time and across cross-section units; the frontier cluster formed by technology leading countries and the other group of countries substantially below the frontier.

To determine the production frontier, defined as the maximum attainable output by given level of inputs, the inefficiency should be zero. We follow Schmidt and Sickles (1984), Kumbhakar (1990) and attempt to measure individual inefficiency:

$$e_{it} = \max_i\left(u_{it}\right) - \left(u_{it}\right) = \max_i\left(\alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t\right) - \left(\alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t\right) \tag{5.40}$$

5.7.2 Econometric estimation

We discuss in details how to estimate the proposed model (5.36-5.38) in SFA. For convenience we rewrite the model as follows:

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{5.41}$$

$$\varepsilon_{it} = v_{it} - u_{it}, \tag{5.42}$$

$$u_{it} = \alpha_i + \rho\tilde{u}_{it}(r) + \boldsymbol{\lambda}_i'\mathbf{f}_t, \tag{5.43}$$

$$\tilde{u}_{it}(r) = \frac{1}{m_{it}}\sum_{j=1}^{N} I\left(\left|u_{t-1}^* - u_{jt-1}\right| \leq r\right) u_{jt-1}, \tag{5.44}$$

where α_i is (unobserved) individual-specific effect, $\mathbf{f}_t$ is an $r \times 1$ vector of unobserved factors and $\boldsymbol{\lambda}_i$ is an $r \times 1$ vector of the heterogeneous loading; $\tilde{u}_{it}(r)$ represents a cluster effect which is equal to the average efficiency of countries which are close to the frontier where $u_{t-1}^* = \min_j (u_{jt-1})$ and v_{it} is an idiosyncratic disturbance.

To obtain consistent estimate of inefficiency in equation (5.40), we first estimate $\hat{\boldsymbol{\beta}}$ in (5.41) by PCCE or IPC, and derive $\hat{\varepsilon}_{it} = y_{it} - \mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{PCCE,IPC}$ and $\hat{v}_{it} = v_{it} - \left(\hat{\boldsymbol{\beta}}_{PCCE,IPC} - \boldsymbol{\beta}\right)\mathbf{x}_{it} = v_{it} + o_p(1)$. Then, by normalizing with respect to the maximum we get a first proxy of inefficiency as $\hat{e}_{it} = \max_i (\hat{\varepsilon}_{it}) - (\hat{\varepsilon}_{it})$. Next, following KMS, we consider the standard estimation procedure for a threshold model, where by a grid of values for r is constructed. Then for all values on that grid the model is estimated by least squares to obtain estimates of ρ. More specifically, we estimate $\hat{r}$ and $\hat{\rho}$ jointly by minimising the following criterion function:

$$\mathbf{V}(r, \rho) = \min_{r,\rho}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{e}_{it} - \rho\frac{1}{m_{it}}\sum_{j=1}^{N} I\left(\left|\hat{e}_{t-1}^* - \hat{e}_{jt-1}\right| \leq r\right)\hat{e}_{jt-1}\right)^2. \tag{5.45}$$

The time-varying individual technical inefficiencies can be therefore consistently estimated by

$$\hat{e}_{it} = \max_i (\hat{u}_{it}) - (\hat{u}_{it}) = \max_i\left(\hat{\alpha}_i + \hat{\rho}\tilde{u}_{it}(\hat{r}) + \hat{\boldsymbol{\lambda}}_i'\mathbf{f}_t\right) - \left(\hat{\alpha}_i + \hat{\rho}\tilde{u}_{it}(\hat{r}) + \hat{\boldsymbol{\lambda}}_i'\mathbf{f}_t\right) \tag{5.46}$$

Finally, we will convert $\hat{e}_{it}$ to the time-varying individual technical efficiency by

$$\hat{\tau}_{it} = \exp(-\hat{e}_{it}). \tag{5.47}$$

For empirical implementations, we follow Bailey *et al.* (2013) who propose a multi-step procedure to deal with both strong and weak forms of CSD as follows:

1. Test for the existence of CSD by applying the Pesaran (2013) CD test;
2. If the null of CSD is rejected, strong CSD is controlled for by applying the factor model.
3. The Pesaran (2013) CD test applied to the (de-factored) residuals again;
4. If the null of no CSD is still rejected, apply spatial or network modelling to the residuals.

In particular, we adopt the multi-step procedure advanced by Bailey *et al.* (2013) within the KMS nonlinear panel data model. This extended KMS approach enables us to deal with both strong and weak forms of cross section dependence jointly by combining the (exogenously driven) factor-based approach with an endogenous threshold efficiency regime selection mechanism in a rather flexible manner.

5.8 Further Issues

This summarises the concluding remarks by Smith and Fuertes (2012)

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.

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. 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, institutions) 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, one should be very careful about using standard pooled estimators such as FE to estimate dynamic models, including lagged dependent variables, from panel data. The dynamic parameters 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. 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.

Currently we are working on the STARDL, GVAR & so on...

A general 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. Single equation or system? Structural system or reduced form?
10. How much parameter heterogeneity is there? In what dimensions?
11. If I(1), is there homogeneous, heterogeneous or no cointegration?
12. How many cointegrating vectors? How do you identify the long-run relations?

13. If using a structural model, can you identify the short-run relations?
14. Can you interpret the results?

Chapter 6

References

Ahn, S.C. and P. Schmidt (1995): "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics* 68, 5-27.

Amemiya, T. and T. McCurdy (1986): "Instrumental Variables Estimation of an Error Components Model," *Econometrica* 54: 869-880.

Anderson, J. and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review* 93: 170-92.

Arellano, M (2003) Panel Data Econometrics, Oxford University Press.

Arellano, M. and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58, 277-297.

Arellano, M. and O. Bover (1995): "Another Look at the Instrumental Variable Estimation of Error Components Models," *Journal of Econometrics* 68, 29-51.

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 nonstationary 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 & 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.

Bailey, N., S. Holly and H.M. Pesaran (2013): "Modelling Spatial Dependence with Pairwise Correlations" unpublished manuscript.

Bailey, N., G. Kapetanios & M.H. Pesaran (2012) Exponent of cross-sectional dependence: estimation and inference, mimeo.

Baldwin, R.E. and D. Taglioni (2006): "Gravity for Dummies and Dummies for Gravity Equations," NBER Working Paper 12516.

Balke, N.S. and T.B. Fomby (1997), "Threshold Cointegration,"International Economic Review 38, 627-645.

Bec, F., M. Ben Salem and M. Carrasco (2004), "Tests of Unit-root versus Threshold Specification with an Application to the PPP," Journal of Business and Economic Statistics 22, 382-395.

Baltagi, B.H. (2010): "Narrow Replication of Serlenga and Shin (2007) Gravity Models of Intra-EU Trade: Application of the PCCE-HT Estimation in Heterogeneous Panels with Unobserved Common Time-specific Factors," *Journal of Applied Econometrics*, Replication Section 25: 505-506.

Baltagi, B.H. (2008) Econometric Analysis of Panel Data, 4th edition New York: Wiley.

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 & A Pirotte (2003) Fixed effects, random effects or Hausman-Taylor? A pre-test estimator, Economics Letters, 79 p361-369.

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.

Behrens, K., C. Ertur and W. Kock (2012): "Dual Gravity: Using Spatial Econometrics To Control For Multilateral Resistance," *Journal of Applied Econometrics* 27: 773-794.

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.

Blundell, R. and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics* 87, 115-143.

Boivin, J., & M. Giannoni (2006) DSGE models in a data rich environment, NBER working paper t0332.

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.

Brav, A., J. Graham, C. Harvey and R. Michaely (2005): "Payout Policy in the 21st Century," *Journal of Financial Economics* 77, 483-527.

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.

Breusch, T., G. Mizon and P. Schmidt (1989): "Efficient Estimation Using Panel Data," *Econometrica* 57: 695-700.

Bruno, G.S.F. (2005) Approximating the Bias of the LSDV estimator for dynamic unbalanced panel data models, Economics Letters, 87, 361-366.

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.

Caner, M. and B.E. Hansen (2004): "Instrumental Variable Estimation of a Threshold Model," *Econometric Theory* 20, 813-843.

Cerrato, M., C. dep Peretti, & N. Sarantis (2007) A non-linear panel unit root test under cross-section dependence, London Metropolitan University.

Chudik A & M.H. Pesaran, (2011a) Infinite dimensional VARs and Factor models, ECB Working paper No 998 Journal of Econometrics forthcoming.

Chudik, A & M.H. Pesaran (2011b) Econometric analysis of high dimension VARs featuring a dominant unit, Cambridge.

Chudik, A. M.H. Pesaran & E Tosetti (2011) Weak and strong cross-section dependence and estimation of large panels, Econometrics Journal 14, C45-C90.

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.

Dang, V.A., M. Kim and Y. Shin (2012): "Asymmetric Capital Structure Adjustments: New Evidence from Dynamic Panel Threshold Models," *Journal of Empirical Finance* 19, 465-482.

Dang, V.A., M. Kim and Y. Shin (2014): Asymmetric Adjustment toward Optimal Capital Structure: Evidence from a Crisis, forthcoming in *International Review of Financial Analysis.*

Dang, V.A., M. Kim and Y. Shin (2014): In search of robust methods for dynamic panel data models in empirical corporate finance, *mimeo.* University of York.

Davies, R.B. (1977): "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika* 64, 247-254.

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.

Eberhardt, M & F Teal, (2010) Econometrics for Grumblers: A new look at cross-country growth empirics, Journal of Economic Surveys, forthcoming.

Elliott, G & A. Timmerman (2008) Economic Forecasting, Journal of Economic Literature, XLVI(1) 3-56.

Ertur, C. and W. Koch (2007): "Growth, Technological Interdependence and Spatial Externalities: Theory and Evidence," *Journal of Applied Econometrics*, 22: 1033–1062.

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.

Fedderke, J., Y. Shin and P. Vaze (2012) Trade, Technology and Wage Inequality in the South African Manufacturing. Oxford Bulletin of Economics and Statistics 74: 808-830.

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 (2006, 2012) Global and National Macroeconometric Modelling: A Long Run Structural Approach, Oxford University Press.

Gengenbach, C. J-P Urbain & J. Westerlund (2009) Error correction testing in panels with global 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

González, A., T. Teräsvirta and D. van Dijk (2005): "Panel Smooth Transition Model and an Application to Investment Under Credit Constraints," Working Paper, Stockholm School of Economics.

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.

Hansen, B.E. (1996): "Inference when a Nuisance Parameter is not Identified under the Null Hypothesis," *Econometrica* 64, 414-30.

Hansen, B.E. (1999): "Threshold Effects in Non-dynamic Panels: Estimation, Testing and Inference," *Journal of Econometrics* 93, 345-368.

Hansen, B.E. (2000): "Sample Splitting and Threshold Estimation," *Econometrica* 68, 575-603.

Hansen, B.E. (2011): "Threshold Autoregression in Economics," *Statistics and Its Interface* 4, 123-127.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica* 46, 1251–1271.

Harris, R.D.F. & E. Tzavalis (1999a) Inference for Unit Roots in Dynamic Panels Where the Time Dimension is fixed, Journal of Econometrics, 91, 201-226.

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. (2012): "The Asymptotic Properties of the System GMM Estimator in Dynamic Panel Data Models when Both N and T are Large," *mimeo.*, Hiroshima University.

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.

Hotelling, H (1933) Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24.

Hsiao, C. (2003) Analysis of Panel Data, 2/e, Cambridge: Cambridge University Press.

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.

Hu, L. and Y. Shin (2014) Testing for Cointegration in Markov Switching Error Correction Models, forthcoming in Vol. 33 of *Advances in Econometrics*: Essays in Honor of Peter C.B. Phillips

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.

Kapetanios, G. (2004) A New Method for determining the number of factors in factor models with large datasets, mimeo Queen Mary, University of London.

Kapetanios, G. (2007) Dynamic Factor Extraction of Cross-sectional Dependence in panel unit root tests, Journal of Applied Econometrics, 22, p313-338.

Kapetanios, G. (2010): "Testing for Exogeneity in Threshold Models," *Econometric Theory* 26, 231-259.

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 nonstationary 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.

Kapetanios, G., Y. Shin and A. Snell (2006) Testing for Cointegration in Nonlinear Smooth Transition Error Correction Models, Econometric Theory 22, 279-303.

Kapetanios, G., J. Mitchell and Y. Shin (2014): "A Nonlinear Panel Data Model of Cross-Sectional Dependence," *Journal of Econometrics,* 179 (2014): 134-157.

Lagana G. & A. Mountford (2005) Measuring Monetary Policy in the UK: A Factor-Augmented Vector Autoregression Model Approach, Manchester School, Supplement, 77-98.

LeSage J.P. and R.K. Pace (2009): *Introduction to Spatial Econometrics.* CRC Press Taylor.

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.

Mastromarco, C., L. Serlenga and Y. Shin (2012): Is Globalization Driving Economic Growth? A Threshold Stochastic Frontier Panel Data Modeling Approach. *Review of International Economics* 20: 563-579.

Mastromarco, C., L. Serlenga and Y. Shin (2013): Globalisation and Technological Convergence in EU," *Journal of Productivity Analysis*, 40: 15-29.

Mastromarco, C., L. Serlenga and Y. Shin (2014): Multilateral Resistance and Euro Effects on Trade Flows, forthcoming in *Advances in Spatial Science*, Springer.

Mastromarco, C., L. Serlenga and Y. Shin (2014): Modelling Technical Efficiency in Cross Sectionally Dependent Stochastic Frontier Panels.

mimeo. University of York.

Michael, P., R.A. Nobay and D.A. Peel (1997), "Transactions Costs and Nonlinear Adjustment in Real Exchange Rates: An Empirical Investigation,"Journal of Political Economy 105, 862-879.

Mitchell, J., K. Mouratis & M.Weale (2005) An assessment of factor based economic indicators: a comparison of factor and regression based preliminary estimates of euro-area GDP growth, NIESR.

Mundlack, Y. (2005) Economic Growth: lessons from two centuries of US agriculture, Journal of Economic Literature, Dec, p989-1024.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica* 49, 1417-1426.

Onatski, Alexei (2009) Testing hypotheses about the number of factors in large factor models. Econometrica 77(5) 1447-1479.

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. (2013): "Testing Weak Cross-Sectional Dependence in Large Panels," Cambridge Working Paper 1208.

Pesaran M.H. T Schuermann & S.M. Weiner, (2004) Modelling Regional Interdependency 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 (2006) Macroeconometric Modelling with a global perspective, Manchester School, Supplement, 24-49.

Pesaran, M.H. L.V. Smith & T. Yamagata (2012) Panel Unit Root Tests in the Presence of a Multifactor Error Structure, mimeo Cambridge.

Pesaran, M.H. & E. Tosetti (2011) Large Panels with Common Factors and Spatial Correlations, Journal of Econometrics, 161(2), 182-202.

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. & 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 (2006) Bias in Dynamic Panel Estimation with Fixed Effects incidental trends and cross-section dependence, Journal of Econometrics, 137, p162-188.

Psaradakis, Z., M. Sola and F. Spagnolo (2004), "On Markov Error-Correction Models With an Application to Stock Prices and Dividends," Journal of Applied Econometrics 19, 69-88.

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.

Rose, A. (2000): "Currency Unions and Trade: The Effect is Large," *Economic Policy* 33: 449-61.

Saikkonen, P. (2005), "Stability Results for Nonlinear Error Correction Models," Journal of Econometrics 127, 69-81.

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.

Serlenga, L. and Y. Shin (2007): "Gravity Models of Intra-EU Trade: Application of the PCCE-HT Estimation in Heterogeneous Panels with Unobserved Common Time-specific Factors," *Journal of Applied Econometrics* 22: 361-381.

Serlenga, L. and Y. Shin (2013): "The Euro Effect on Intra-EU Trade: Evidence from the Cross Sectionally Dependent Panel Gravity Models," *mimeo*. University of York.

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 www.cfap.jbs.cam.ac.uk/research/gvartoolbox/c

Smith R.P & A. Fuertes (2012) Panel Time-Series. cemmap course

Smith R.P & G. Zoega, (2005) Unemployment Investment and Expected Global Returns: a panel FAVAR approach, Birkbeck Working Paper 524.

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.

Stone, J.R.N. (1947) On the interdependence of blocks of transactions, Supplement to the Journal of the Royal Statistical Society, 11, 1-31.

Swamy, P.A.V.B. (1970) Efficient Inference in a Random Coefficient Regression Model, Econometrica 38, 311-323.

Trapani, L. & G. Urga (2010) Micro versus macro cointegration in heterogeneous panels, Journal of Econometrics, 155(1), 1-18.

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. & R. Larsson (2009) A note on the pooling of individual PANIC unit root tests, Econometric Theory, 25, 1851-1868.

Westerlund J. & J-P Urbain (2011) Cross-sectional averages or principal components, Meteor Research Memorandum 11/053.

Zilak, J. (1997), "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators," *Journal of Business and Economic Statistics* 15, 419-431.