Dynamic Panel Data Models I: Covariance Structures and Autoregressions[1]

Class Notes

Manuel Arellano

Revised: October 12, 2009

The methods discussed in this note are motivated by an interest in the time series properties of short panels. Such interest may arise for a variety of reasons. We may be interested in separating out permanent from transitory components of variation as in earnings mobility studies. In another type of applications, we may be able to test theories or identify policy parameters from the mapping between a time series model and a model of individual behaviour. Examples include Hall and Mishkin (1982) and Blundell, Pistaferri, and Preston (2008) on the transmission of income shocks to consumption, and Abowd and Card (1989) on earnings and hours of work in an intertemporal labour supply context. Finally, we may be interested in a predictive distribution for use in some optimization problem under uncertainty. For example, Deaton (1991) used a predictive distribution of future earnings given past earnings to derive optimal consumption paths for consumers who maximize life-cycle expected utility.

1 Dynamic Covariance Structures

1.1 Introduction

A natural extension of the basic error components model is to allow for serial correlation in the time-varying component. This can be achieved by specifying a homogeneous moving average or autoregressive process. We have

$$y_{it} = \eta_i + v_{it}$$

and the covariance matrix of the $T \times 1$ vector y_i is given by:

$$\Omega = V + \sigma_\eta^2 \iota\iota' \tag{1}$$

where V is the $T \times T$ autocovariance matrix of v_{it}. In the basic case, $V = \sigma^2 I_T$. Specification and inference are discussed below. The rest of the introduction is devoted to an informal discussion of the problem of distinguishing between unobserved heterogeneity and dynamics in short panels.

Distinguishing Unobserved Heterogeneity from Genuine Dynamics Let us first consider the identification problem in a panel with $T = 2$. In time series analysis, given a single series of size T $\{y_1, ..., y_T\}$ a first-order autocovariance is calculated as an average of the $T - 1$ products of

[1] This is an abridged version of Part II in Arellano (2003).

observations one period apart: $(T-1)^{-1}\sum_{t=2}^{T} y_t y_{t-1}$. With panel data of size $T=2$, we have N time series with two observations each. In such situation calculating individual *time series autocovariances* is not possible because the time series averages would have just one observation. We can nevertheless calculate a *cross-sectional first-order autocovariance* for the specific two periods available in the panel. This will take the form of an average of the N products of the two observations for each individual: $N^{-1}\sum_{i=1}^{N} y_{i1}y_{i2}$. Thus, when we consider population moments in this context they are to be regarded as population counterparts of cross-sectional moments of the previous type. As for example,

$$E\left(y_{i1}y_{i2}\right) = \operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} y_{i1}y_{i2}. \tag{2}$$

The standard error component model with white noise v_{it} is identified with $T=2$ because

$$Var\left(y_{i1}\right) = Var\left(y_{i2}\right) = \sigma_\eta^2 + \sigma_v^2 \tag{3}$$

$$Cov\left(y_{i1}, y_{i2}\right) = \sigma_\eta^2. \tag{4}$$

In this model all the observed correlation between first and second period data is due to heterogeneity, since for a given individual the sequence of y's is a random white noise process around his specific level η_i. The point to note here is that this *pure heterogeneity model* is observationally equivalent to a *homogeneous model with serial correlation.* For example, if the model is

$$y_{it} = \eta + v_{it} \tag{5}$$

$$v_{it} = \alpha v_{i(t-1)} + \varepsilon_{it}, \tag{6}$$

where η is a constant, $|\alpha|<1$, $\varepsilon_{it} \sim iid\left(0, \sigma_\varepsilon^2\right)$, $v_{i1} \sim iid\left(0, \sigma_v^2\right)$ and $\sigma_v^2 = \sigma_\varepsilon^2/\left(1-\alpha^2\right)$, we have

$$Var\left(y_{i1}\right) = Var\left(y_{i2}\right) = \sigma_v^2 \tag{7}$$

$$Cov\left(y_{i1}, y_{i2}\right) = \alpha\sigma_v^2. \tag{8}$$

In the heterogeneity model the observed autocorrelation ρ_1 is given by

$$\rho_1 = \frac{\lambda}{(1+\lambda)} \tag{9}$$

with $\lambda = \sigma_\eta^2/\sigma_v^2$, whereas in the homogeneous AR(1) model we have

$$\rho_1 = \alpha. \tag{10}$$

If for example the variance of η_i is 4 times the variance of v_{it} in the heterogeneity model, we get $\rho_1 = 4/5 = 0.8$. Exactly the same observed correlation as we would get with a homogeneous AR(1) model with $\alpha = 0.8$. So there is no way to distinguish empirically between the two models from the autocovariance matrix when $T=2$, as long as $\alpha \geq 0$.

With $T = 3$ the previous two models are distinguishable since the heterogeneity model implies

$$Cov\left(y_{i1}, y_{i3}\right) = Cov\left(y_{i1}, y_{i2}\right) \tag{11}$$

whereas the AR(1) model implies

$$Cov\left(y_{i1}, y_{i3}\right) = \alpha Cov\left(y_{i1}, y_{i2}\right). \tag{12}$$

Now the combined model with heterogeneous level and homogeneous AR(1) serial correlation (which allows the intercept η in (5) to be individual specific with variance σ_η^2) is just-identified with

$$Var\left(y_{i1}\right) = Var\left(y_{i2}\right) = Var\left(y_{i3}\right) = \sigma_\eta^2 + \sigma_v^2 \tag{13}$$

$$Cov\left(y_{i1}, y_{i2}\right) = Cov\left(y_{i2}, y_{i3}\right) = \sigma_\eta^2 + \alpha\sigma_v^2 \tag{14}$$

$$Cov\left(y_{i1}, y_{i3}\right) = \sigma_\eta^2 + \alpha^2\sigma_v^2. \tag{15}$$

Pursuing the previous argument, note that with $T = 3$ the heterogeneous AR(1) model will be indistinguishable from a homogeneous AR(2) model. These examples suggest that a non-parametric test of homogeneity will be only possible for large T and N in the absence of structural breaks.

Note that with $T = 2$ the "reduced form" autocovariance matrix contains three free coefficients (two variances and one covariance). Since the heterogeneous AR(1) model also has three parameters α, σ_η^2, and σ_v^2, the order condition for identification is satisfied with equality, but not the rank condition. This is so because the variance equation for the second period will be either redundant or incompatible.

In general, persistence measured from cross-sectional autocorrelation coefficients will combine two different sources. In the AR(1) model with heterogeneous mean we have

$$\rho_1 = \frac{\sigma_\eta^2 + \alpha\sigma_v^2}{\sigma_\eta^2 + \sigma_v^2} = \alpha + \frac{(1-\alpha)\sigma_\eta^2}{\sigma_\eta^2 + \sigma_v^2} = \alpha + \frac{(1-\alpha)\lambda}{(1+\lambda)}, \tag{16}$$

which particularizes to (9) or (10) when either α or λ are equal to zero, respectively.

Often with microdata $\rho_1 \simeq 1$. Nevertheless, a value of ρ_1 close to one may be compatible with many different values of α and λ. For example, fitting the heterogeneous mean AR(1) model to annual employment from a short panel of firms we obtained $\rho_1 = 0.995$, $\alpha = 0.8$ and $\lambda = 36$.

The estimation of autoregressive models with individual effects will be discussed in Section 2. In the remainder of this section we consider time effects, moving average models, and inference from covariance structures.

The previous discussion could have been conducted using moving average instead of autoregressive processes.[2] One advantage of MA over AR processes is that they imply linear restrictions in the autocovariance matrix (e.g. with $T = 3$ the pure MA(1) process implies $Cov\left(y_{i1}, y_{i3}\right) = 0$). The advantages of autoregressive representations are in the possibilities of incorporating certain non-stationary features (like unit roots or nonstationary initial conditions), and the relationship to regression and instrumental-variable settings.

[2] Except for the fact that a pure MA(1) process restricts the range of possible values of ρ_1.

1.2 Time Effects

Often a time series analysis of individual time series will only be meaningful after conditioning on common features. For example, in the empirical consumption model of Hall and Mishkin considered below, the time series properties of consumption and income were investigated after conditioning on trends and demographic characteristics of the household. In other instances, it may be important to remove business cycle or seasonal effects in order to avoid confusion between aggregate and individual specific dynamics. One might consider specifying a regression of y_{it} on some aggregate variables z_t (like GDP growth, the unemployment rate, inflation, or functions of time)

$$y_{it} = \gamma' z_t + y_{it}^I \tag{17}$$

together with a time series model for y_{it}^I. Alternatively, the aggregate component could be specified as a latent common stochastic process y_t^a:

$$y_{it} = y_t^a + y_{it}^I. \tag{18}$$

One would then specify time series models for both y_t^a and y_{it}^I. If $y_t^a \sim iid\left(0, \sigma_a^2\right)$ and y_{it}^I follows the basic error component model, we obtain the *two-way error component model*:

$$y_{it} = y_t^a + \eta_i + v_{it}, \tag{19}$$

whose covariance matrix is given by

$$Var(y) = \sigma_v^2 I_{NT} + \sigma_\eta^2 \left(I_N \otimes \iota_T \iota_T'\right) + \sigma_a^2 \left(\iota_N \iota_N' \otimes I_T\right). \tag{20}$$

where $y = (y_1', ..., y_N')'$, and ι_T and ι_N denote vectors of ones of dimensions T and N. Stochastic modelling of both y_t^a and η_i requires large T and N. In panels with small N and large T the individual effects are treated as parameters.

Time Dummies in Short Panels Conversely, in short panels the number of time series observations is too small to attempt a stochastic modelling of y_t^a. On the other hand, the cross-sectional sample size is large so that the realizations of y_t^a that occur in the sample can be treated as unknown period specific parameters to be estimated. To this end we may specify a set of T time dummies:

$$y_{it} = y^{a\prime} d_t + y_{it}^I \tag{21}$$

where $y^a = (y_1^a, ..., y_T^a)'$ and d_t is a $T \times 1$ vector with one in the t-th position and zero elsewhere.

Note that any aggregate variable z_t will be a linear combination of the time dummies. Thus, if a full set of time dummies is included any aggregate variable will be perfectly colinear with them and hence redundant. If one has a substantive interest in the effects of macro variables, time dummies would not be employed. Indeed, the specification for the macro variables can be regarded as a model for the time dummies. If the substantive interest is in individual dynamics and data are sufficiently informative, however, time dummies afford a robust control for common aggregate effects.

Individual-Specific Trends In the basic error component model there is a heterogeneous constant level of the process. This can be generalized to considering a heterogeneous linear trend:

$$y_{it} = \eta_{0i} + \eta_{1i} t + v_{it} \tag{22}$$

or in vector notation

$$y_i = S\eta_i + v_i \tag{23}$$

where $\eta_i = (\eta_{0i}, \eta_{1i})'$ and S denotes the $T \times 2$ matrix

$$S = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & T \end{pmatrix}.$$

Letting $Var(\eta_i) = \Omega_\eta$, and assuming that $v_{it} \sim iid(0, \sigma^2)$ and independent of η_i, the $T \times T$ covariance matrix of y_i is given by

$$\Omega = S\Omega_\eta S' + \sigma^2 I_T. \tag{24}$$

A necessary condition for identification of Ω_η and σ^2 is that $T \geq 3$. To illustrate the situation, let us consider for $T = 3$ the covariance matrix of the variables y_{i1}, Δy_{i2}, and $\Delta^2 y_{i3}$:

$$y_{i1} = \eta_{0i} + \eta_{1i} + v_{i1} \tag{25}$$

$$\Delta y_{i2} = \eta_{1i} + (v_{i2} - v_{i1}) \tag{26}$$

$$\Delta^2 y_{i3} = v_{i3} - 2v_{i2} + v_{i1}, \tag{27}$$

which provides a non-singular transformation of the original covariance matrix Ω. The two covariance matrices contain the same information, but the transformation simplifies the relationship between the model's parameters and the variances and covariances of the data:

$$Var \begin{pmatrix} y_{i1} \\ \Delta y_{i2} \\ \Delta^2 y_{i3} \end{pmatrix} = \begin{pmatrix} \sigma_{00} + \sigma_{11} + 2\sigma_{01} + \sigma^2 & \sigma_{11} + \sigma_{01} - \sigma^2 & \sigma^2 \\ & \sigma_{11} + 2\sigma^2 & -3\sigma^2 \\ & & 6\sigma^2 \end{pmatrix}. \tag{28}$$

Thus, σ^2 is determined from the variance and covariances in the last column. Given σ^2, σ_{11} can be determined from $Var(\Delta y_{i2})$. Then σ_{01} is determined from $Cov(y_{i1}, \Delta y_{i2})$, and finally σ_{00} is determined from $Var(y_{i1})$.[3]

[3] With $T = 2$, the variances of η_{0i}, η_{1i}, and v_{it} are just identified if η_{0i}, and η_{1i} are assumed to be uncorrelated.

Individual Specific Responses to Aggregate Variables The previous case can be extended to consider individual-specific responses to aggregate variables (like business cycle movements):

$$y_{it} = \eta_i' z_t + v_{it} \tag{29}$$

where z_t denotes a vector of observable aggregate variables, and η_i is a vector of individual specific effects of z_t on y_{it}. Note that for $S = (z_1, ..., z_T)'$ and $Var(\eta_i) = \Omega_\eta$, the variance matrix of y_i is of the same form as (24). Identification in this case will require that z_t has sufficient variation and the dimension of η_i is not too large relative to T.

Time Effects Interacted with Individual Effects Let us now consider a model of the form

$$y_{it} = \eta_i \delta_t + v_{it}. \tag{30}$$

This model can be regarded as specifying an aggregate shock δ_t that has individual-specific effects, or a permanent characteristic η_i that has changing effects over time. The difference with the previous model is that z_t in (29) was known whereas δ_t in (30) is not. Therefore, in a short panel $\delta = (\delta_1, ..., \delta_T)'$ will be treated as a vector of parameters to be estimated.

Assuming that $v_{it} \sim iid(0, \sigma^2)$ independent of η_i, the data covariance matrix takes the form

$$\Omega = \sigma_\eta^2 \delta\delta' + \sigma^2 I_T. \tag{31}$$

This is the structure of the one-factor model of factor analysis. Some scale normalization is required in order to determine δ. Using $\delta'\delta = 1$, it follows that $\sigma_\eta^2 + \sigma^2$ is the largest eigenvalue of Ω and δ is the corresponding eigenvector. Moreover, the remaining $T - 1$ eigenvalues of Ω are equal to σ^2.

Let us illustrate the identification of this type of model by considering a case in which $T = 3$ and the v_{it} are allowed to have period-specific variances σ_t^2. With the normalization $\delta_1 = 1$, the covariance matrix is given by

$$Var\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} = \begin{pmatrix} \sigma_\eta^2 + \sigma_1^2 & \sigma_\eta^2\delta_2 & \sigma_\eta^2\delta_3 \\ & \sigma_\eta^2\delta_2^2 + \sigma_2^2 & \sigma_\eta^2\delta_2\delta_3 \\ & & \sigma_\eta^2\delta_3^2 + \sigma_3^2 \end{pmatrix}. \tag{32}$$

Subject to compatibility, the parameters are just identified and given by

$$\delta_2 = \frac{Cov(y_{i2}, y_{i3})}{Cov(y_{i1}, y_{i3})} \tag{33}$$

$$\delta_3 = \frac{Cov(y_{i2}, y_{i3})}{Cov(y_{i1}, y_{i2})} \tag{34}$$

$$\sigma_\eta^2 = \frac{Cov(y_{i1}, y_{i2})Cov(y_{i1}, y_{i3})}{Cov(y_{i2}, y_{i3})} \tag{35}$$

$$\sigma_t^2 = Var(y_{it}) - \sigma_\eta^2\delta_t^2 \ (t = 1, 2, 3). \tag{36}$$

Note that (33) and (34) can be interpreted as instrumental variable parameters from autoregressive equations. This is a specially useful perspective when v_{it} itself follows an autoregressive process.

1.3 Moving Average Autocovariances

Stationary Models We begin by considering stationary models. Covariance stationarity requires that for all t and j, $Cov\left(y_{it}, y_{i(t-j)}\right)$ does not depend on t:

$$Cov\left(y_{it}, y_{i(t-j)}\right) = \gamma_j. \tag{37}$$

Thus, under stationarity, the $T \times T$ autocovariance matrix of a scalar variable y_{it} depends at most on only T different coefficients $\gamma_0, ..., \gamma_{T-1}$, which implies that it satisfies $T(T+1)/2 - T$ restrictions.

A stationary moving-average structure of order q MA(q) with individual effects will further restrict the coefficients γ_j for $j > q$ to take the same value (corresponding to the variance of the individual effect):

$$\gamma_{q+1} = ... = \gamma_{T-1}. \tag{38}$$

The absence of individual effects will be signaled by the additional restriction that the previous coefficients are equal to zero

$$\gamma_{q+1} = ... = \gamma_{T-1} = 0. \tag{39}$$

Therefore, given stationarity, an MA($T-2$) process (with individual effects) or an MA($T-1$) process (without them) will be observationally equivalent saturated models.[4]

Nonstationary Models Nonstationarity, in the sense of failure of condition (37), may arise for a variety of reasons. Examples include the individual-specific trends and responses to aggregate variables considered above,[5] or nonstationary initial conditions. Moreover, nonstationarity may also arise as a result of unit roots, time-varying error variances (possibly due to aggregate effects), or ARMA models with time-varying coefficients.

Provided $q < T-1$, a nonstationary MA(q) process without permanent effects will satisfy the $(T-q)(T-q-1)/2$ restrictions

$$Cov(y_{it}, y_{i(t-j)}) = 0 \text{ for } j > q. \tag{40}$$

In such model, the elements in the main diagonal of the autocovariance matrix and those in the first q subdiagonals will be free coefficients, except for the symmetry and non-negativity restrictions. Similarly, in a nonstationary MA(q) process with permanent effects the zero elements in the autocovariance matrix will be replaced by a constant coefficient.

[4] A moving average process may also imply inequality restrictions, which are not considered here.

[5] From the point of view of the time series process of a given individual, model (22) introduces a deterministic trend, whereas model (29) is compatible with a stationary process for y_{it} provided z_t is stationary itself. Thus, the immediate reason why (29) is "nonstationary" in our terminology is because we are conditioning on the realizations of z_t.

Multivariate Models The previous considerations can be generalized to a multivariate context. Let y_{it} denote an $m \times 1$ random vector. Then the autocovariance matrix of the vector $y_i = (y'_{i1}, ..., y'_{iT})'$ is of order mT. Under stationarity, for any t and j the $m \times m$ block $Cov\left(y_{it}, y_{i(t-j)}\right)$ does not depend on t:

$$Cov\left(y_{it}, y_{i(t-j)}\right) = \Gamma_j. \tag{41}$$

A stationary vector-MA(q) process with individual effects introduces the restrictions

$$\Gamma_{q+1} = ... = \Gamma_{T-1}. \tag{42}$$

Moreover, if no variable contains individual specific intercepts then also

$$\Gamma_{q+1} = ... = \Gamma_{T-1} = 0. \tag{43}$$

Similar remarks can be made for nonstationary vector-MA specifications.

Abowd and Card (1989) presented an empirical analysis of changes in the logs of annual earnings and hours from three different panels (actually of residuals from regressions of those variables on time dummies and potential experience). For each dataset they found evidence supporting the restrictions implied by a nonstationary MA(2) bivariate process without individual effects.[6] Abowd and Card did not consider the covariance structure of the levels of their variables. They focused on the implications of the time series properties of changes in the variables for life-cycle labour supply models.

Covariance Matrices of Levels and First Differences To examine the relationship between covariance structures in levels and first-differences, let us consider the transformed covariance matrix

$$Var\begin{pmatrix} y_{i1} \\ \Delta y_{i2} \\ \vdots \\ \Delta y_{iT} \end{pmatrix} = \Omega^* = \begin{pmatrix} \omega^*_{11} & \omega^*_{12} & \dots & \omega^*_{1T} \\ \omega^*_{12} & & & \\ \vdots & & \Omega_\Delta & \\ \omega^*_{1T} & & & \end{pmatrix}. \tag{44}$$

The matrix Ω^* is a non-singular transformation of the covariance matrix in levels (so that knowledge of one implies knowledge of the other), and Ω_Δ is the covariance matrix in first differences. Therefore, a model of Ω_Δ is equivalent to a model of the covariance matrix in levels that leaves the coefficients ω^*_{1t} $(t = 1, ..., T)$ unrestricted.

The terms ω^*_{1t} may be informative about the structural parameters in Ω_Δ. If y_{it} follows an MA(q) process with individual effects, Δy_{it} will be an MA($q+1$) process without individual effects. In such a case even if initial conditions are assumed to be nonstationary we would expect

$$\omega^*_{1t} = 0 \text{ for } t > q + 2. \tag{45}$$

Enforcing these restrictions may lead to more efficient estimates of parameters in the structure for Ω_Δ.

[6] Individual effects in the changes of the variables would correspond to individual specific trends in their levels.

1.4 Estimating Covariance Structures

The previous models all specify a structure on a data covariance matrix. It is of some interest to approach identification and inference with reference to a covariance structure, specially when the interest is in estimating the parameters in the structure as opposed to a substantive interest in the probability distribution of the data. In some cases, restrictions on higher-order moments may add identification content, but it is still often useful to know when a parameter of interest in a time series model may or may not be identified from the data covariance matrix alone.

1.4.1 GMM Estimation

Abstracting from mean components for simplicity, suppose the covariance matrix of a $p \times 1$ time series y_i is a function of a $k \times 1$ parameter vector θ given by

$$E(y_i y_i') = \Omega(\theta). \tag{46}$$

If y_i is a scalar time series its dimension will coincide with T, but in the multivariate context $p = mT$.

Vectorizing the expression and eliminating redundant elements (due to symmetry) we obtain a vector of moments of order $r = (p+1)p/2$:

$$vechE\left[y_i y_i' - \Omega(\theta)\right] = E\left[s_i - \omega(\theta)\right], \tag{47}$$

where the *vech* operator stacks by rows the lower triangle of a square matrix.[7]

If $r > k$ and $H(\theta) = \partial\omega(\theta)/\partial\theta'$ has full column rank, the model is overidentified. In that case a standard optimal GMM estimator solves:

$$\widehat{\theta} = \arg\min_c \left[\overline{s} - \omega(c)\right]' \widehat{V}^{-1} \left[\overline{s} - \omega(c)\right] \tag{48}$$

where $\overline{s}$ is the sample mean vector of s_i:

$$\overline{s} = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{49}$$

and $\widehat{V}$ is some consistent estimator of $V = Var(s_i)$. A natural choice is the sample covariance matrix of s_i:

$$\widehat{V} = \frac{1}{N}\sum_{i=1}^{N} s_i s_i' - \overline{s}\overline{s}'. \tag{50}$$

The first-order conditions from the optimization problem are

$$-H(c)'\widehat{V}^{-1}\left[\overline{s} - \omega(c)\right] = 0. \tag{51}$$

[7] If we were interested in considering mean restrictions of the form $E(y_i) = \mu(\theta)$ jointly with covariance restrictions, we could proceed in the same way after redefining the vectors s_i and $\omega(\theta)$ as $s_i = (y_i', [vech(y_i y_i')]')'$ and $\omega(\theta) = (\mu(\theta)', [vech\Omega(\theta)]')'$, respectively.

The two standard results for large sample inference are, firstly, asymptotic normality of the scaled estimation error

$$\left[\frac{1}{N}H(\widehat{\theta})'\widehat{V}^{-1}H(\widehat{\theta})\right]^{-1/2}\left(\widehat{\theta}-\theta\right)\xrightarrow{d}\mathcal{N}(0,I) \tag{52}$$

and, secondly, the asymptotic chi-square distribution of the minimized estimation criterion (*test statistic of overidentifying restrictions*)

$$S=N\left[\overline{s}-\omega(\widehat{\theta})\right]'\widehat{V}^{-1}\left[\overline{s}-\omega(\widehat{\theta})\right]\xrightarrow{d}\chi^2_{r-k}. \tag{53}$$

Example: Fitting a Homogeneous MA(1) model with T=3 In such case $r=6$ and $k=2$ with $\theta=(\gamma_0,\gamma_1)$ and

$$\Omega=\begin{pmatrix}\gamma_0 & \gamma_1 & 0\\ \gamma_1 & \gamma_0 & \gamma_1\\ 0 & \gamma_1 & \gamma_0\end{pmatrix}. \tag{54}$$

Thus we have

$$s_i=\begin{pmatrix} y_{i1}^2 & y_{i2}y_{i1} & y_{i2}^2 & y_{i3}y_{i1} & y_{i3}y_{i2} & y_{i3}^2\end{pmatrix}' \tag{55}$$

and

$$\omega(\theta)=\begin{pmatrix}\gamma_0\\ \gamma_1\\ \gamma_0\\ 0\\ \gamma_1\\ \gamma_0\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\\ 1 & 0\\ 0 & 0\\ 0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}\gamma_0\\ \gamma_1\end{pmatrix}=H\theta. \tag{56}$$

Since the restrictions are linear, an explicit expression for the GMM estimator is available:

$$\widehat{\theta}=\left(H'\widehat{V}^{-1}H\right)^{-1}H'\widehat{V}^{-1}\overline{s}. \tag{57}$$

Thus, $\widehat{\theta}$ can be obtained as a GLS regression of $\overline{s}$ on H using $\widehat{V}^{-1}$ as weight matrix.

Sometimes using a (possibly parameter-dependent) transformation of the original moments may lead to a simpler estimation problem. One simplification arises when the transformed moments are linear in the parameters whereas the original moments are not. Another simplification is when a subset of the transformed moments are unrestricted, so that one can concentrate on smaller sets of moments and parameters without loss of efficiency (Arellano, 2003, 70-71).

Relationship between GMM and PML If $y_i \sim iid\mathcal{N}[0, \Omega(\theta)]$, the MLE of θ solves[8]

$$\widehat{\theta}_{PML} = \arg\min_{c} \left[\log\det\Omega(c) + \frac{1}{N}\sum_{i=1}^{N} y_i'\Omega^{-1}(c)y_i \right]. \tag{58}$$

If y_i is not assumed normal, $\widehat{\theta}_{PML}$ can be regarded as a Gaussian pseudo maximum likelihood estimator (PML). Defining the selection matrix $\mathcal{D} = \partial vec\Omega/\partial(vech\Omega)'$, the first-order conditions are

$$-H(c)'\left[\mathcal{D}'\left(\Omega^{-1}(c)\otimes\Omega^{-1}(c)\right)\mathcal{D}\right][\overline{s} - \omega(c)] = 0, \tag{59}$$

which are of the same form as those for the GMM problem given in (51).

Under normality, fourth-order moments are functions of second-order moments. Specifically,

$$V^{-1} = \frac{1}{2}\mathcal{D}'\left(\Omega^{-1}(\theta)\otimes\Omega^{-1}(\theta)\right)\mathcal{D}. \tag{60}$$

Thus, under normality an alternative optimal GMM estimator could use a consistent estimate of $\mathcal{D}'\left(\Omega^{-1}(\theta)\otimes\Omega^{-1}(\theta)\right)\mathcal{D}$ as weight matrix. Such estimator does not coincide with $\widehat{\theta}_{PML}$ because in the latter the weight matrix is continuously updated as a function of c in (59), but the two are asymptotically equivalent with or without normality. Under non-normality, they remain consistent and asymptotically normal but they are inefficient for large N relative to the GMM estimator that uses $\widehat{V}^{-1}$ as weight. A PML estimator may still be preferable even under non-normality on finite sample grounds, but if so it is important to base inference on standard errors robust to non-normality.

Testing Nested Restrictions Using Incremental Sargan Tests The test statistic of overidentifying restrictions (53) can be used as an overall specification test against the unrestricted data covariance matrix. Sometimes, however, we are interested in testing additional constraints within a particular covariance structure. For example, we may wish to test for the absence of random effects in a stationary moving average model, or for a stationary moving average against a nonstationary one. Testing of nested restrictions can be accomplished using incremental statistics.

Let the additional constraints under test be $\theta = g(\psi)$, where ψ is another parameter vector of order $s < k$ and each element of g is a twice differentiable function. The GMM estimator of ψ is

$$\widehat{\psi} = \arg\min_{a} \{\overline{s} - \omega[g(a)]\}'\widehat{V}^{-1}\{\overline{s} - \omega[g(a)]\} \tag{61}$$

so that the constrained estimator of θ is $\widehat{\theta}_R = g(\widehat{\psi})$. Moreover, we have

$$S_R = N\left[\overline{s} - \omega(\widehat{\theta}_R)\right]'\widehat{V}^{-1}\left[\overline{s} - \omega(\widehat{\theta}_R)\right] \xrightarrow{d} \chi^2_{r-s}. \tag{62}$$

Finally, the incremental Sargan test statistic S_Δ satisfies

$$S_\Delta = S_R - S \xrightarrow{d} \chi^2_{k-s} \text{ independent of } S. \tag{63}$$

Thus, large values of S_Δ will lead to rejection of the $\theta = g(\psi)$ restrictions.

[8] In the model $y_i = \eta_i\iota + v_i$, unconditional joint normality of y_i can be regarded as the result of both conditional normality given η_i, namely $y_i \mid \eta_i \sim \mathcal{N}(\eta_i\iota, V)$, and normality of η_i: $\eta_i \sim \mathcal{N}(0, \sigma^2_\eta)$, so that $\Omega = \sigma^2_\eta\iota\iota' + V$.

1.5 Illustration: Testing the Permanent Income Hypothesis

Hall and Mishkin (1982) used food consumption and labour income from a PSID sample of $N = 2309$ US households over $T = 7$ years to test the predictions of a permanent income model of consumption behaviour. We use their work as an empirical illustration of dynamic panel covariance structures.

They specified individual means of income and consumption changes as linear regressions on age, age squared, time, and changes in the number of children and adults living in the household. Thus, they were implicitly allowing for unobserved intercept heterogeneity in the levels of the variables, but only for observed heterogeneity in their changes. Deviations from the individual means of income and consumption, denoted $\overline{y}_{it}$ and $\overline{c}_{it}$ respectively, were specified as follows.

Specification of the Income Process Hall and Mishkin assumed that income errors $\overline{y}_{it}$ were the result of two different types of shocks, permanent and transitory:

$$\overline{y}_{it} = y_{it}^L + y_{it}^S. \tag{64}$$

They also assumed that agents were able to distinguish one type of shock from the other and respond to them accordingly. The permanent component y_{it}^L was specified as a random walk

$$y_{it}^L = y_{i(t-1)}^L + \varepsilon_{it}, \tag{65}$$

and the transitory component y_{it}^S as a stationary moving average process

$$y_{it}^S = \eta_{it} + \rho_1 \eta_{i(t-1)} + \rho_2 \eta_{i(t-2)}. \tag{66}$$

A limitation is the lack of measurement error in observed income. That is, a component to which consumption does not respond at all. This is important since measurement error in PSID income is large, but identification would require additional indicators of permanent income.

Specification of the Consumption Process Mean deviations in consumption changes were specified to respond one-to-one to permanent income shocks and by a fraction β to transitory shocks. The magnitude of β will depend on the persistence in transitory shocks (measured by ρ_1 and ρ_2) and on real interest rates. It will also depend on age, but the analysis was simplified by treating it as a constant. This model can be formally derived from an optimization problem with quadratic utility, and constant interest rates that are equal to the subjective discount factor. Since only food consumption is observed, an adjustment was made by assuming a constant marginal propensity to consume food, denoted α. With these assumptions we have

$$\Delta \overline{c}_{it} = \alpha \varepsilon_{it} + \alpha \beta \eta_{it}. \tag{67}$$

In addition, Hall and Mishkin introduced a stationary measurement error in the level of consumption (or transitory consumption that is independent of income shocks) with an MA(2) specification:

$$c_{it}^S = v_{it} + \lambda_1 v_{i(t-1)} + \lambda_2 v_{i(t-2)}. \tag{68}$$

The Resulting Bivariate Covariance Structure Therefore, the model that is taken to the data consists of a joint specification for mean deviations in consumption and income changes as follows:

$$\Delta\overline{c}_{it} = \alpha\varepsilon_{it} + \alpha\beta\eta_{it} + v_{it} - (1-\lambda_1)\, v_{i(t-1)} - (\lambda_1 - \lambda_2)\, v_{i(t-2)} - \lambda_2 v_{i(t-3)} \tag{69}$$

$$\Delta\overline{y}_{it} = \varepsilon_{it} + \eta_{it} - (1-\rho_1)\, \eta_{i(t-1)} - (\rho_1 - \rho_2)\, \eta_{i(t-2)} - \rho_2 \eta_{i(t-3)}. \tag{70}$$

The three innovations in the model are assumed to be mutually independent with constant variances σ_ε^2, σ_η^2 and σ_v^2. Thus, the model contains nine unknown coefficients:

$$\theta = \left(\begin{array}{ccccccccc} \alpha & \beta & \lambda_1 & \lambda_2 & \rho_1 & \rho_2 & \sigma_\varepsilon^2 & \sigma_\eta^2 & \sigma_v^2 \end{array} \right)'.$$

The model specifies a covariance structure for the 12×1 vector

$$w_i = \left(\begin{array}{ccccccccc} \Delta\overline{c}_{i2} & \Delta\overline{c}_{i3} & \cdots & \Delta\overline{c}_{i7} & \Delta\overline{y}_{i2} & \Delta\overline{y}_{i3} & \cdots & \Delta\overline{y}_{i7} \end{array} \right)'$$

$$E\left(w_i w_i'\right) = \Omega(\theta).$$

Let us look in some detail at the form of various elements of $\Omega(\theta)$. We have

$$Var(\Delta\overline{y}_{it}) = \sigma_\varepsilon^2 + 2\left(1 - \rho_1 - \rho_1\rho_2 + \rho_1^2 + \rho_2^2\right)\sigma_\eta^2 \qquad (t = 2, ..., 7) \tag{71}$$

$$Cov(\Delta\overline{y}_{it}, \Delta\overline{y}_{i(t-1)}) = -\left[(1-\rho_1) - (1 - \rho_1 + \rho_2)\,(\rho_1 - \rho_2)\right]\sigma_\eta^2 \tag{72}$$

and also

$$Cov(\Delta\overline{c}_{it}, \Delta\overline{y}_{it}) = \alpha\sigma_\varepsilon^2 + \alpha\beta\sigma_\eta^2 \qquad (t = 2, ..., 7) \tag{73}$$

$$Cov(\Delta\overline{c}_{it}, \Delta\overline{y}_{i(t-1)}) = 0 \tag{74}$$

$$Cov(\Delta\overline{c}_{i(t-1)}, \Delta\overline{y}_{it}) = -\alpha\beta\,(1-\rho_1)\,\sigma_\eta^2. \tag{75}$$

A fundamental restriction of the model is lack of correlation between current consumption changes and lagged income changes, as captured by (74). The model, nevertheless, predicts correlation between current consumption changes and current and future income changes, as seen from (73) and (75).

Empirical Results Hall and Mishkin estimated their model by Gaussian PML. In the calculation of standard errors no adjustment was made for possible non-normality. They estimated $\widehat{\beta} = 0.3$, which given their estimates of ρ_1 and ρ_2 ($\widehat{\rho}_1 = 0.3$, $\widehat{\rho}_2 = 0.1$) turned out to be consistent with the model only for unrealistic values of real interest rates (above 30 percent). Moreover, they estimated the marginal propensity to consume food as $\widehat{\alpha} = 0.1$, and the moving average parameters for transitory consumption as $\widehat{\lambda}_1 = 0.2$ and $\widehat{\lambda}_2 = 0.1$. The variance of the permanent income shocks was twice as large as that of the transitory shocks: $\widehat{\sigma}_\varepsilon^2 = 3.4$ and $\widehat{\sigma}_\eta^2 = 1.5$.

Finally, they tested the covariance structure focusing on the fundamental restriction of lack of correlation between current changes in consumption and lagged changes in income. They found a negative covariance which was significantly different from zero. They did not consider overall tests of overidentifying restrictions. As a result of this finding they considered an extended version of the model in which a fraction of consumers spent their current income ("Keynesian" consumers).

2 Autoregressive Models with Individual Effects

In this section we discuss the specification and estimation of autoregressive models with individual specific intercepts. We focus on first-order processes for simplicity. We begin by considering the properties of the within-group estimator. In contrast with the static fixed effects model, WG has a small T bias which does not disappear as N becomes large. Next, we consider instrumental variable estimators that are consistent for panels with small T and large N. These estimators use lagged observations as instruments for errors in first differences. Then we discuss the role of assumptions about initial conditions, homoskedasticity, and whether the parameter space includes unit roots or not. Finally, we consider various aspects of inference with VAR panel data models in the context of an empirical application using firm level data on employment and wages.

2.1 Assumptions

Let $\{y_{i0}, y_{i1}, ..., y_{iT}, \eta_i\}_{i=1}^N$ be a random sample[9] such that

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it} \ \ (t = 1, ..., T) \ \ \mid \alpha \mid < 1 \tag{76}$$

$$E\left(v_{it} \mid y_i^{t-1}, \eta_i\right) = 0 \qquad (\textit{Assumption B1}) \tag{}$$

where $y_i^{t-1} = (y_{i0}, y_{i1}, ..., y_{i(t-1)})'$. We observe y_i^T but not the individual intercept η_i, which can be regarded as a missing time-invariant variable with $E(\eta_i) = \eta$ and $Var(\eta_i) = \sigma_\eta^2$.

Thus, this is a model that specifies the conditional mean of y_{it} given its past and a value of η_i. An implication of *B1* is that the errors v_{it} are conditionally serially uncorrelated. Namely,

$$E\left(v_{it} v_{i(t-j)} \mid y_i^{t-1}, \eta_i\right) = 0, \text{ for } j > 0, \tag{77}$$

so that $E\left(v_{it} v_{i(t-j)}\right) = 0$ as well. *B1* also implies lack of correlation between η_i and v_{it} for all t.

Homoskedasticity Assumption *B1* implies that $E\left(v_{it}\right) = 0$ cross-sectionally for any t, but does not restrict the variance of v_{it}. That is, the conditional variance may be some period-specific non-negative function of y_i^{t-1} and η_i

$$E\left(v_{it}^2 \mid y_i^{t-1}, \eta_i\right) = \varphi_t\left(y_i^{t-1}, \eta_i\right), \tag{78}$$

and the unconditional variance may change with t[10]

$$E\left(v_{it}^2\right) = E\left[\varphi_t\left(y_i^{t-1}, \eta_i\right)\right] = \sigma_t^2. \tag{79}$$

[9] We assume for convenience that y_{i0} is observed, so that for each individual we have $T + 1$ observations.

[10] Note that $E\left(v_{it}^2\right)$ is a cross-sectional population mean, as in $\text{plim}_{N\to\infty} N^{-1} \sum_{i=1}^N v_{it}^2$.

Thus, we may consider two different homoskedasticity assumptions: *conditional homoskedasticity*

$$E\left(v_{it}^2 \mid y_i^{t-1}, \eta_i\right) = \sigma_t^2, \qquad (Assumption\ B2)$$

and *time series homoskedasticity*

$$E\left(v_{it}^2\right) = \sigma^2. \qquad (Assumption\ B3)$$

B2 and *B3* may hold in conjunction, but any of them may also occur in the absence of the other.

Assumption *B3* is compatible with individual-specific error variances of the form $E\left(v_{it}^2 \mid \eta_i\right) = \sigma_i^2$. Moreover, since we may not wish to think of σ_i^2 as being exclusively a function of η_i, we could imagine a larger conditioning set of unobserved individual components, leaving the argument unaffected.

Stationarity Assuming $\mid \alpha \mid < 1$, guarantees that the process is stable but not necessarily stationary. Stationarity also requires that the process started in the distant past or, equivalently, that the distribution of initial observations coincides with the steady state distribution of the process.

Solving (76) recursively we obtain

$$y_{it} = \left(\sum\nolimits_{s=0}^{t-1} \alpha^s\right) \eta_i + \alpha^t y_{i0} + \sum\nolimits_{s=0}^{t-1} \alpha^s v_{i(t-s)}. \qquad (80)$$

Furthermore, *B1* implies

$$E\left(y_{it} \mid \eta_i\right) = \left(\sum\nolimits_{s=0}^{t-1} \alpha^s\right) \eta_i + \alpha^t E\left(y_{i0} \mid \eta_i\right), \qquad (81)$$

which for $\mid \alpha \mid < 1$ and large t tends to $\mu_i = \eta_i/(1-\alpha)$. We refer to μ_i as the *steady state mean* for individual i. Thus, stationarity in mean requires

$$E\left(y_{i0} \mid \eta_i\right) = \frac{\eta_i}{(1-\alpha)}, \qquad (Assumption\ B4)$$

in which case all $E\left(y_{it} \mid \eta_i\right)$ are time-invariant and coincide with the steady state mean.

Similarly, under *B1-B3*, for $j \geq 0$ we have

$$Cov\left(y_{it}, y_{i(t-j)} \mid \eta_i\right) = \alpha^{2t-j} Var\left(y_{i0} \mid \eta_i\right) + \alpha^j \left(\sum\nolimits_{s=0}^{t-j-1} \alpha^{2s}\right) \sigma^2, \qquad (82)$$

which for $\mid \alpha \mid < 1$ and large t tends to the *steady state j-th autocovariance* for individual i given by $\alpha^j \sigma^2/(1-\alpha^2)$. Thus, under homoskedasticity, covariance stationarity requires

$$Var\left(y_{i0} \mid \eta_i\right) = \frac{\sigma^2}{(1-\alpha^2)}, \qquad (Assumption\ B5)$$

in which case all $Cov\left(y_{it}, y_{i(t-j)} \mid \eta_i\right)$ are time-invariant and coincide with the steady state autocovariances.

2.2 The Within-Group Estimator

The WG estimator of α is the slope coefficient in an OLS regression of y on lagged y and a full set of individual dummies, or equivalently the OLS estimate in deviations from time means or orthogonal deviations. Letting $y_i = (y_{i1}, ..., y_{iT})'$ and $y_{i(-1)} = (y_{i0}, ..., y_{i(T-1)})'$, the WG estimator of α is

$$\widehat{\alpha}_{WG} = \frac{\sum_{i=1}^{N} y'_{i(-1)} Q y_i}{\sum_{i=1}^{N} y'_{i(-1)} Q y_{i(-1)}} \tag{83}$$

where Q is the WG operator of order T.

The autoregressive equation (76) is of the same form as the static fixed effects model with $x_{it} = y_{i(t-1)}$, but it does not satisfy the strict exogeneity assumption because v_{it} is correlated with future values of the regressor. Indeed, for any value of T

$$E\left(y'_{i(-1)} Q v_i\right) = \sum_{t=1}^{T} E\left[y_{i(t-1)}\left(v_{it} - \overline{v}_i\right)\right] \neq 0 \tag{84}$$

since $y_{i(t-1)}$ is correlated with the average error $\overline{v}_i$ through the terms $v_{i1}...v_{i(t-1)}$. As a consequence, $\widehat{\alpha}_{WG}$ is inconsistent for fixed T as N tends to infinity. The bias will nevertheless tend to zero as T increases since $\text{plim}_{T\to\infty} \overline{v}_i = 0$. Thus, in common with standard time series autoregression, least squares estimation is biased but consistent as T tends to infinity. The problem is that when T is small the biases may be too large to be ignored regardless of the value of N.

The Nickell Bias The form of the bias is important for understanding the environments in which WG can be expected to perform well. For fixed T and large N the bias is

$$\underset{N\to\infty}{\text{plim}}\left(\widehat{\alpha}_{WG} - \alpha\right) = \frac{E\left(y'_{i(-1)} Q v_i\right)}{E\left(y'_{i(-1)} Q y_{i(-1)}\right)}. \tag{85}$$

Under assumptions *B1* and *B3* it can be shown that

$$E\left(y'_{i(-1)} Q v_i\right) = -\sigma^2 h_T(\alpha) \tag{86}$$

where

$$h_T(\alpha) = \frac{1}{(1-\alpha)}\left[1 - \frac{1}{T}\left(\frac{1-\alpha^T}{1-\alpha}\right)\right]. \tag{87}$$

Moreover, if *B4* and *B5* also hold, the denominator of (85) satisfies

$$E\left(y'_{i(-1)} Q y_{i(-1)}\right) = \frac{\sigma^2 (T-1)}{(1-\alpha^2)}\left(1 - \frac{2\alpha h_T(\alpha)}{(T-1)}\right). \tag{88}$$

Thus, through the denominator, the bias depends on the form of initial conditions. The bias formula as given by Nickell (1981) is therefore:

$$\underset{N\to\infty}{\text{plim}}\left(\widehat{\alpha}_{WG} - \alpha\right) = -\frac{\left(1-\alpha^2\right) h_T(\alpha)}{(T-1)}\left(1 - \frac{2\alpha h_T(\alpha)}{(T-1)}\right)^{-1}. \tag{89}$$

The WG bias is of order $1/T$, so that it vanishes as $T \to \infty$, but it may be important for small values of T. When $T = 2$ the bias under stationarity is given by

$$\operatorname*{plim}_{N\to\infty} (\widehat{\alpha}_{WG} - \alpha) = -\frac{(1+\alpha)}{2}, \tag{90}$$

which coincides with the bias of OLS in first differences.

The following table shows the Nickell bias for several values of α and T.

Table 1

WG Bias under Stationarity

$T\backslash\alpha$	0.05	0.5	0.95
2	−0.52	−0.75	−0.97
3	−0.35	−0.54	−0.73
10	−0.11	−0.16	−0.26
15	−0.07	−0.11	−0.17

If $\alpha > 0$ the bias is always negative, and massive with the very small values of T. It becomes smaller in absolute value as T increases, but even when $T = 15$ the bias is still substantial (e.g. 22 percent with $\alpha = 0.5$).

Likelihood Conditional on η_i In general, the likelihood for one individual conditional on η_i can be sequentially factorized as

$$f(y_i^T \mid \eta_i) = f(y_{i0} \mid \eta_i) \prod_{t=1}^{T} f(y_{it} \mid y_i^{t-1}, \eta_i). \tag{91}$$

If we assume that $y_{it} \mid y_i^{t-1}, \eta_i$ is normally distributed with conditional mean and variance given by assumptions *B1*, *B2* and *B3*, so that

$$y_{it} \mid y_i^{t-1}, \eta_i \sim \mathcal{N}\left(\alpha y_{i(t-1)} + \eta_i, \sigma^2\right), \tag{92}$$

the log-likelihood conditional on η_i and y_{i0} is given by

$$\begin{aligned} \ell_i &= \log \prod_{t=1}^{T} f(y_{it} \mid y_i^{t-1}, \eta_i) \\ &\propto -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_{it} - \alpha y_{i(t-1)} - \eta_i\right)^2. \end{aligned} \tag{93}$$

Clearly, the maximizer of $\sum_{i=1}^{N} \ell_i$ with respect to α, σ^2, and $\eta_1, ..., \eta_N$ gives rise to the WG estimator, which is therefore the Gaussian MLE of α (conditional on y_{i0}) estimated jointly with the individual specific intercepts. Given the large-N-fixed-T inconsistency of WG, this can be regarded as another example of Neyman and Scott's incidental parameter problem.

2.3 Instrumental Variable Estimation

The WG estimator is inconsistent for fixed T because taking differences or deviations to eliminate the effects creates a negative correlation between lagged y's and errors in the transformed equation. However, values of y lagged two periods or more are valid instruments in the equations in first differences. Specifically, an implication of *B1* is that the following $(T-1)T/2$ linear IV moment restrictions hold:

$$E\left[y_i^{t-2}\left(\Delta y_{it} - \alpha \Delta y_{i(t-1)}\right)\right] = 0 \ (t = 2, ..., T). \tag{94}$$

This gives rise to a system of $T-1$ equations with cross-equation restrictions and different instruments valid for different equations, which can be estimated by linear GMM.

Simple IV estimators of this type were first proposed by Anderson and Hsiao (1981). Their proposal was to consider a single moment of the form

$$E\left[\sum\nolimits_{t=2}^{T} y_{i(t-2)}\left(\Delta y_{it} - \alpha \Delta y_{i(t-1)}\right)\right] = 0, \tag{95}$$

or alternatively

$$E\left[\sum\nolimits_{t=3}^{T} \Delta y_{i(t-2)}\left(\Delta y_{it} - \alpha \Delta y_{i(t-1)}\right)\right] = 0. \tag{96}$$

Since (95) and (96) are linear combinations of (94), for large N and fixed T, the "stacked" Anderson-Hsiao IV estimates are asymptotically inefficient relative to GMM based on (94). Stacked IV estimates, however, will remain well defined and consistent regardless of whether T or N or both tend to infinity.

GMM estimators that used all available lags at each period as instruments for the equations in first differences were proposed by Holtz-Eakin, Newey, and Rosen (1988) and Arellano and Bond (1991).

A GMM estimator based on the IV moment conditions (94) takes the form

$$\widehat{\alpha}_{GMM} = \left[\left(\Delta y'_{-1} Z\right) V_N^{-1} \left(Z' \Delta y_{-1}\right)\right]^{-1} \left(\Delta y'_{-1} Z\right) V_N^{-1} \left(Z' \Delta y\right) \tag{97}$$

where $Z'\Delta y = \sum\limits_{i=1}^{N} Z_i' \Delta y_i$, $Z'\Delta y_{-1} = \sum\limits_{i=1}^{N} Z_i' \Delta y_{i(-1)}$, $\Delta y_i = (\Delta y_{i2}, ..., \Delta y_{iT})'$, $\Delta y_{i(-1)} = (\Delta y_{i1}, ..., \Delta y_{i(T-1)})'$ and

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & y_{i0} & y_{i1} & & 0 & & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & \dots & y_{i0} & \dots & y_{i(T-2)} \end{pmatrix}. \tag{98}$$

According to standard GMM theory, an optimal choice of the inverse weight matrix V_N is a consistent estimate of the covariance matrix of the orthogonality conditions $E(Z_i' \Delta v_i \Delta v_i' Z_i)$. Under

conditional and time series homoskedasticity (assumptions *B1*, *B2* and *B3*):[11]

$$E(Z_i'\Delta v_i \Delta v_i' Z_i) = \sigma^2 E(Z_i' D D' Z_i) \tag{99}$$

where D is the $(T-1)\times T$ first-difference matrix operator. Thus, a one-step GMM estimator uses

$$\widehat{V} = \sum_{i=1}^{N} Z_i' D D' Z_i, \tag{100}$$

whereas a two-step GMM estimator uses the robust choice

$$\widetilde{V} = \sum_{i=1}^{N} Z_i' \Delta\widehat{v}_i \Delta\widehat{v}_i' Z_i, \tag{101}$$

where $\Delta\widehat{v}_i$ are one-step GMM residuals.

A heteroskedasticity-robust estimate of the asymptotic variance of one-step GMM can be obtained from the sandwich formula:

$$\widehat{Var}(\widehat{\alpha}_{GMM1}) = \mathcal{M}^{-1}\left[\left(\Delta y_{-1}' Z\right)\widehat{V}^{-1}\widetilde{V}\widehat{V}^{-1}\left(Z'\Delta y_{-1}\right)\right]\mathcal{M}^{-1} \tag{102}$$

where $\mathcal{M} = \left(\Delta y_{-1}' Z\right)\widehat{V}^{-1}\left(Z'\Delta y_{-1}\right)$. Furthermore, an estimate of the asymptotic variance of two-step GMM is given by

$$\widehat{Var}(\widehat{\alpha}_{GMM2}) = \left[\left(\Delta y_{-1}' Z\right)\widetilde{V}^{-1}\left(Z'\Delta y_{-1}\right)\right]^{-1}. \tag{103}$$

Sargan test statistics of the overidentifying restrictions can also be obtained from the minimized two-step GMM criterion as follows:

$$S = \left(\Delta\widetilde{v}' Z\right)\widetilde{V}^{-1}\left(Z'\Delta\widetilde{v}\right). \tag{104}$$

In our case, S will have a limiting chi-square distribution with $[(T-1)T/2]-1$ degrees of freedom. These statistics are widely used as specification diagnostics.

As we shall see below, (94) are not the only restrictions on the data second-order moments implied by the conditional mean independence and homoskedasticity assumptions *B1-B3*, but they are the only ones that are valid in the absence of homoskedasticity or lack of correlation between η_i and v_{it}.

[11] Note that under *B1-B3* a typical block of $E(Z_i'\Delta v_i \Delta v_i' Z_i)$ satisfies

$$E\left(\Delta v_{it}\Delta v_{i(t-j)} y_i^{t-2} y_i^{t-j-2\prime}\right) = E\left(\Delta v_{it}\Delta v_{i(t-j)}\right) E\left(y_i^{t-2} y_i^{t-j-2\prime}\right).$$

2.4 Initial Conditions and Heteroskedasticity

Here we examine the role of assumptions about initial conditions and heteroskedasticity in the estimation of AR models from short panels. We consider three different types of covariance structures. Type 1 relies on stationarity assumptions (*B1* and *B3*-*B5*). Type 2 is the covariance structure assuming an unrestricted joint distribution of y_{i0} and η_i, and time series homoskedasticity (*B1* and *B3*). Finally, Type 3 is the least restrictive covariance structure which allows for both unrestricted initial conditions and time series heteroskedasticity (assumption *B1* only). The choice of auxiliary assumptions matters because of a trade-off between robustness and efficiency in this context.

Estimation Under Stationarity Under assumptions *B1*, *B3*, *B4* and *B5* the first and second-order moments of y_i^T are functions of the four parameters α, σ^2, σ_μ^2, and μ of the form

$$E\left(y_{it}-\mu\right)=0 \ (t=0,1,...,T) \tag{105}$$

$$E\left(y_{it}y_{is}-\omega_{ts}-\mu^2\right)=0 \ (t=0,1,...,T;s=0,1,...,t), \tag{106}$$

where $\mu=E(\mu_i)$, $\sigma_\mu^2=Var(\mu_i)$, and ω_{ts} is the (t,s)th element of the variance matrix of y_i^T given by

$$\omega_{ts}=\sigma_\mu^2+\alpha^{|t-s|}\frac{\sigma^2}{(1-\alpha^2)}. \tag{107}$$

The parameters can be estimated by nonlinear GMM using (105) and (106). Alternatively, it turns out that these nonlinear in parameter moments can be converted into equivalent linear moment equations by transformation and reparameterization (Arellano, 2003). The resulting moments are:

$$\begin{aligned}
E\left[y_i^{t-2}\left(\Delta y_{it}-\alpha\Delta y_{i(t-1)}\right)\right] &= 0 \ (t=2,...,T) && \text{(108a)}\\
E\left[\Delta y_{i(t-1)}\left(y_{it}-\alpha y_{i(t-1)}\right)\right] &= 0 \ (t=2,...,T) && \text{(108b)}\\
E\left(y_{it}-\mu\right) &= 0 \ (t=0,1,...,T) && \text{(108c)}\\
E\left(y_{it}^2-\varphi^2\right) &= 0 \ (t=0,1,...,T) && \text{(108d)}\\
E\left[y_{i0}\left(y_{i1}-\alpha y_{i0}\right)-\psi\right] &= 0 && \text{(108e)}
\end{aligned}$$

where $\varphi^2=\mu^2+\sigma_\mu^2+\sigma^2/(1-\alpha^2)$ and $\psi=(1-\alpha)(\sigma_\mu^2+\mu^2)$. Thus, α, μ, φ^2, and ψ can be estimated by linear GMM using (108a)-(108e). Original parameters can be recovered by undoing the reparameterization. The two sets of GMM estimates will be asymptotically equivalent for optimal choices of the weight matrices.

The orthogonality conditions (108a) coincide with those in (94). The moments (108b) also have a straightforward instrumental variable interpretation: they state that $\Delta y_{i(t-1)}$ has zero mean and is orthogonal to η_i+v_{it}; (108c) and (108d) state the unconditional stationarity of the first and second moments, respectively, and (108e) is an unrestricted moment that determines the variance of the individual effect.

Unrestricted Initial Conditions In the time series context whether a stationary AR model is estimated conditional on the first observation or not does not matter for robustness or asymptotic efficiency. We obtain different estimates but they have similar properties when T is large. In short panels the situation is fundamentally different. An estimator of α obtained under the assumption that $y_{i0} \mid \mu_i$ follows the stationary unconditional distribution of the process will be inconsistent when the assumption is false. Therefore, there is a trade-off between robustness and efficiency, since in short panels the assumption of stationary initial conditions may be very informative about α.

The question is whether initial conditions at the start of the sample are representative of the steady state behaviour of the model or not. In the analysis of country panel data, Barro and Sala-i-Martin (1995) described some examples -like data sets that start at the end of a war or other major historical event- in which one would not expect initial conditions to be distributed according to the steady state distribution of the process. In the case of micro panels, the starting point of the sample may be closer to the start of the process for some units than others. For example, for young workers or new firms initial conditions may be less related to steady state conditions than for older ones.

Taking these considerations into account, a more robust specification is one in which the distribution of y_{i0} given μ_i (or η_i) is left unrestricted.[12] That is, we drop assumptions *B4* and *B5* and let $E(y_{i0} \mid \mu_i)$ and $Var(y_{i0} \mid \mu_i)$ be arbitrary functions of μ_i, while retaining the basic stable specification of equation (76).

To analyze this case let us introduce the linear projection of y_{i0} on μ_i:

$$y_{i0} = \delta_0 + \delta\mu_i + \varepsilon_{i0} \tag{109}$$

where ε_{i0} is the projection error, so that $E(\varepsilon_{i0}) = 0$ and $Cov(\varepsilon_{i0}, \mu_i) = 0$. Moreover, let $\sigma_0^2 = E(\varepsilon_{i0}^2)$.

Clearly, under assumptions *B4* and *B5* we have that $\delta_0 = 0$, $\delta = 1$ and $\sigma_0^2 = \sigma^2/(1-\alpha^2)$.

In view of (80) and (109), this model can be written as

$$y_{it} = \alpha^t\delta_0 + \left[1 - (1-\delta)\alpha^t\right]\mu_i + \sum\nolimits_{s=0}^{t-1} \alpha^s v_{i(t-s)} + \alpha^t\varepsilon_{i0} \;\; (t = 1, ..., T)\,. \tag{110}$$

This gives rise to a mean-covariance structure that has three additional parameters relative to the stationary model. Means, variances and covariances take now period specific values given by

$$E\left(y_{it}\right) = \alpha^t\delta_0 + \left[1 - (1-\delta)\alpha^t\right]\mu \tag{111}$$

and for $s \leq t$:

$$\omega_{ts} = \left[1 - (1-\delta)\alpha^t\right]\left[1 - (1-\delta)\alpha^s\right]\sigma_\mu^2 + \alpha^{t-s}\left[\sigma^2\left(\sum\nolimits_{j=0}^{s-1}\alpha^{2j}\right) + \alpha^{2s}\sigma_0^2\right]. \tag{112}$$

Consistent estimates of the nonstationary model can be obtained by nonlinear GMM from (111) and (112). As before an alternative equivalent representation is available (Ahn and Schmidt, 1995).

[12] The moving average models of Section 1 assumed stationary initial conditions by specifying y_{i0}=$\mu_i + v_{i0}$.

It turns out that the restrictions implied by *B1* on the data covariance matrix and mean vector can be represented as

$$E\left[y_i^{t-2}\left(\Delta y_{it} - \alpha \Delta y_{i(t-1)}\right)\right] = 0 (t = 2...T) \quad (113a)$$

$$E\left[\left(\Delta y_{i(t-1)} - \alpha \Delta y_{i(t-2)}\right)\left(y_{it} - \alpha y_{i(t-1)} - \eta\right)\right] = 0 (t = 3...T) \quad (113b)$$

$$E\left(y_{it} - \alpha y_{i(t-1)} - \eta\right) = 0 (t = 1...T) \quad (113c)$$

In addition, time series homoskedasticity (assumption *B3*) implies

$$E\left[\left(y_{it} - \alpha y_{i(t-1)} - \eta\right)^2 - \sigma_u^2\right] = 0 \ (t = 1, ..., T), \quad (114)$$

where $\sigma_u^2 = \sigma_\eta^2 + \sigma^2$. Finally, the following four unrestricted moments determine the first and second moments of y_{i0}, $c_0 = Cov\left(y_{i0}, \eta_i\right)$ and $c_1 = Cov\left(y_{i1}, \eta_i\right)$:[13]

$$E\left(y_{i0} - \mu_0\right) = 0 \quad (115a)$$

$$E\left[y_{i0}^2 - \varphi_0^2\right] = 0 \quad (115b)$$

$$E\left[y_{i0}\left(y_{i2} - \alpha y_{i1} - \eta\right) - c_0\right] = 0 \quad (115c)$$

$$E\left[y_{i1}\left(y_{i2} - \alpha y_{i1} - \eta\right) - c_1\right] = 0. \quad (115d)$$

In this representation, coefficients related to initial conditions and the individual effect variance can be ignored since they only appear through unrestricted moments. Thus, optimal GMM estimates of α, σ_u^2, and η can be obtained from the moment conditions (113a) to (114) alone.

Time Series Heteroskedasticity In time series, estimators of autoregressive models under the assumption of homoskedasticity remain consistent when the assumption is false. This is not so in short panels. GMM or PML estimators of α in any of the two previous models will be inconsistent for fixed T as N tends to infinity if the unconditional variances of the errors vary over time.

PML estimators of the conditional mean parameters obtained under the assumption of conditional homoskedasticity, however, are robust to conditional heteroskedasticity in short panels, as long as the restrictions implied by the pseudo likelihood on the unconditional covariance matrix of the data are satisfied. The same is true of GMM estimates of the corresponding covariance structures.

Therefore, unless one has a substantive interest in modelling conditional variances, robust estimates of α can be obtained in conjunction with the unconditional variances of the errors. Conversely, if one is interested in modelling dispersion in the conditional distributions of $y_{it} \mid y_i^{t-1}, \mu_i$, the use of estimators of (possibly time-varying) unconditional error variances σ_t^2 as estimates of the conditional variances may result in misspecification.

Time series heteroskedasticity may arise as a result of the presence of aggregate effects in the conditional variance of the process. Thus, time varying σ's may occur in conjunction with a stationary

[13] Under stationary initial conditions $c_0 = \sigma_\eta^2/(1-\alpha)$ and $c_0 = c_1$.

idiosyncratic process, and even with a stationary aggregate effect. In the latter situation, time series heteroskedasticity would just reflect the fact that in a short panel we condition on the values of the aggregate effects that occur in the sample, and these enter the conditional variance.

So we also consider a model in which the unconditional variances of the errors are allowed to vary over time in an arbitrary way, hence relaxing assumption *B3*. In combination with unrestricted initial conditions, this gives rise to a covariance structure characterized by the $(T+4)\times 1$ parameter vector $(c_0, c_1, \sigma_0^2, \sigma_1^2, ..., \sigma_T^2, \alpha)$. In terms of the moment conditions (113a) to (115d) the only modification is that (114) now becomes a set of unrestricted moments

$$E\left[\left(y_{it} - \alpha y_{i(t-1)} - \eta\right)^2 - \sigma_{ut}^2\right] = 0 \ (t = 1, ..., T), \tag{116}$$

where $\sigma_{ut}^2 = \sigma_\eta^2 + \sigma_t^2$, so that the only restrictions implied by the model are (113a)-(113c).

2.5 Mean Stationarity

We have seen that assumptions about initial conditions present a trade-off between robustness and efficiency in the context of short panels. The trade-off is particularly acute for autoregressive models with roots close to the unit circle, since in such case the IV moment conditions (94) may be very weak.

Here we discuss a model that enforces mean stationarity but leaves variances unrestricted. That is, we assume that assumptions *B1* and *B4* hold but not necessarily *B2*, *B3* or *B5*. So that we have

$$E(y_{it} \mid \mu_i) = \mu_i \ (t = 0, 1, ..., T)$$

This implies that $\delta_0 = 0$ and $\delta = 1$ in (109). Note that the mean stationarity assumption does not refer to the start of an individual's process, but to the first observations in the actual sample.

Under mean stationarity the covariance between y_{it} and μ_i does not depend on t, so that $c_0 = c_1 = (1-\alpha)\sigma_\mu^2$, and Δy_{it} is uncorrelated with μ_i. In view of the discussion in the previous section, the implication is that for this model both sets of IV conditions (108a) and (108b) for errors in differences and levels, respectively, are valid. Adding the mean conditions (108c), the full list of restrictions implied by mean stationarity on the data first and second moments is:[14]

$$E\left[y_i^{t-2}\left(\Delta y_{it} - \alpha \Delta y_{i(t-1)}\right)\right] = 0 \ (t = 2, ..., T) \tag{117}$$

$$E\left[\Delta y_{i(t-1)}\left(y_{it} - \alpha y_{i(t-1)} - \eta\right)\right] = 0 \ (t = 2, ..., T) \tag{118}$$

$$E\left(y_{it} - \alpha y_{i(t-1)} - \eta\right) = 0 \ (t = 1, ..., T) \tag{119}$$

$$E\left(y_{i0} - \mu\right) = 0. \tag{120}$$

So full covariance-information linear estimation of α is possible for this model using a GMM estimator that combines instruments in levels for equations in differences with instruments in differences for equations in levels (Arellano and Bover, 1995; Blundell and Bond, 1998).

[14] Mean stationarity also implies $E\left[\Delta y_{i(t-j)}\left(y_{it} - \alpha y_{i(t-1)}\right)\right] = 0 \ (t = 2, ..., T; j = 2, ..., t-1)$, but these moments are redundant given (117) and (118) since they are linear combinations of them.

Levels & Differences GMM Linear Estimators Moments (117)-(119) can be written as

$$E\begin{pmatrix} Z_i'Du_i \\ Z_{\ell i}'u_i \end{pmatrix} \equiv E\left(Z_i^{\dagger\prime}\overline{\mathcal{H}}u_i\right) = 0 \tag{121}$$

where $u_i = y_i - \alpha y_{i(-1)}$, Z_i is the matrix of instruments given in (98) for equations in differences, $Z_{\ell i}$ is the matrix of instruments for equations in levels that takes the form

$$Z_{\ell i} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \Delta y_{i1} & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & 1 & \Delta y_{i2} & & 0 & 0 \\ \vdots & & \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & \Delta y_{i(T-1)} \end{pmatrix}, \tag{122}$$

$\overline{\mathcal{H}}$ is the $(2T-1)\times T$ selection matrix $\overline{\mathcal{H}} = (D', I_T)'$, where D is the $(T-1)\times T$ first difference matrix operator, and $Z_i^\dagger$ is a block diagonal matrix with blocks Z_i and $Z_{\ell i}$. The model specifies different instruments for a system of $T-1$ equations in first differences followed by T equations in levels.

Letting $X_i = (y_{i(-1)}, \iota)$, a GMM estimator of α and η will be of the form:

$$\begin{pmatrix} \widehat{\alpha}_S \\ \widehat{\eta}_S \end{pmatrix} = \left[\left(\sum_{i=1}^N X_i'\overline{\mathcal{H}}'Z_i^\dagger\right) A_N \left(\sum_{i=1}^N Z_i^{\dagger\prime}\overline{\mathcal{H}}X_i\right)\right]^{-1} \left(\sum_{i=1}^N X_i'\overline{\mathcal{H}}'Z_i^\dagger\right) A_N \left(\sum_{i=1}^N Z_i^{\dagger\prime}\overline{\mathcal{H}}y_i\right). \tag{123}$$

An optimal two-step choice for A_N can be obtained from the inverse of a consistent estimate of the moment covariance matrix $E\left(Z_i^{\dagger\prime}\overline{\mathcal{H}}u_iu_i'\overline{\mathcal{H}}'Z_i^\dagger\right)$. However, unlike in IV estimation in differences (97), this matrix depends on unknown parameters even under conditional homoskedasticity. As a result, in this case there is no one-step efficient GMM estimator under "classical" errors.[15]

2.6 Unit Roots

So far we have considered stable models or models in which a unit root is a feature of the specification, like the integrated moving average process used by Hall and Mishkin (1982) to model household income. In the Hall and Mishkin's example the random walk component is the device used to model permanent income shocks. In such context the empirical interest is in measuring how large the random walk component is relative to the stationary component, rather than testing for its presence.

Sometimes the presence or absence of unit roots is a central feature of the model of interest so that unit root testing is not warranted. In this section, however, we assume an interest in testing the unit root hypothesis, and examine the nature of the problem in short panels with unobserved heterogeneity.

First of all, the null and alternative hypotheses need to be specified. We begin by considering, as the alternative hypothesis, the stable AR model with unrestricted initial conditions and time series

[15] An arbitrary but convenient choice of one-step weight matrix is given by the inverse of $N^{-1}\sum_{i=1}^N Z_i^{\dagger\prime}\overline{\mathcal{H}\mathcal{H}}'Z_i^\dagger$.

heteroskedasticity. As for the null, we consider a random walk without drift. The model is

$$y_{i0} = \delta_0 + \delta\mu_i + v_{i0} \tag{124a}$$

$$y_{it} = \alpha y_{i(t-1)} + (1-\alpha)\mu_i + v_{it}. \tag{124b}$$

Thus, when $\alpha = 1$ we have

$$y_{it} = y_{i(t-1)} + v_{it}, \tag{125}$$

so that heterogeneity only plays a role in the determination of the starting point of the process. The implication is that when $\alpha = 1$ only the variance of y_{i0} is identified in the covariance matrix of (y_{i0}, μ_i).

An alternative specification of the null would be a random walk with an individual specific drift

$$y_{it} = y_{i(t-1)} + \eta_i + v_{it}, \tag{126}$$

but this is a model with heterogeneous linear growth that would be more suited for comparisons with stationary models including individual trends.

Under the null $H_0 : \alpha = 1$, pooled OLS in levels is a consistent and asymptotically normal estimator of α in (125) for fixed T and large N. Therefore, a standard t-ratio statistic can be used to perform a fixed-T-large-N one-sided test of the unit root hypothesis against the alternative of a stable process.

The IV Estimator with Unit Roots When $\alpha = 1$ the rank condition of the IV moments (94) fails. The y_i^{t-2} are uncorrelated with $\Delta y_{i(t-1)}$ when $\alpha = 1$ since in such case $\Delta y_{i(t-1)}$ is the innovation in period $(t-1)$. That is, for $j \geq 2$ we have

$$\begin{aligned} Cov(y_{i(t-j)}, \Delta y_{i(t-1)}) &= -(1-\alpha)\left[Cov(y_{i(t-j)}, y_{i(t-2)}) - Cov(y_{i(t-j)}, \mu_i)\right] \\ Cov(y_{i(t-j)}, \Delta y_{it}) &= \alpha Cov(y_{i(t-j)}, \Delta y_{i(t-1)}). \end{aligned}$$

Hence, when $\alpha = 1$, the IV moments (94) are not only satisfied for the true value of α, but also for any other value. The implication is that GMM estimators based on (94) are inconsistent when $\alpha = 1$.

Mean Stationarity and Unit Roots Let us now consider a model that specifies mean stationarity when $|\alpha| < 1$ and a random walk without drift when $\alpha = 1$. This is model (124a)-(124b) with the restriction $\delta = 1$, which gives rise to

$$y_{it} = \mu_i + v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1} + \alpha^t v_{i0}. \tag{127}$$

Here σ_μ^2 and σ_0^2 are not separately identified when $\alpha = 1$, but the rank condition from moments (117)-(118) is satisfied. As noted by Arellano and Bover (1995), the reason is that when $\alpha = 1$ we have

$$Cov\left(\Delta y_{i(t-1)}, y_{i(t-1)}\right) = \sigma_t^2 \tag{128}$$

which will be non-zero as long as $\sigma_t^2 > 0$. So the IV moments for the errors in levels (118) ensure the determination of α when α equals one. The implication is that GMM estimators of the mean stationary model remain consistent when $\alpha = 1$.

2.7 Estimating and Testing VAR's for Firm Employment and Wages

- In this section we discuss inference with autoregressive models in an empirical illustration.
- We consider autoregressive employment and wage equations estimated from the panel of firms used by Alonso-Borrego and Arellano (1999).
- This is a balanced panel of 738 Spanish manufacturing companies, for which there are available annual observations for the period 1983-1990.
- We consider various specializations of a bivariate VAR(2) model for the logs of employment and wages, denoted n_{it} and w_{it} respectively.
- Individual and time effects are included in both equations.
- The form of the model is

$$n_{it} = \delta_{1t} + \alpha_1 n_{i(t-1)} + \alpha_2 n_{i(t-2)} + \beta_1 w_{i(t-1)} + \beta_2 w_{i(t-2)} + \eta_{1i} + v_{1it} \tag{129}$$

$$w_{it} = \delta_{2t} + \gamma_1 w_{i(t-1)} + \gamma_2 w_{i(t-2)} + \lambda_1 n_{i(t-1)} + \lambda_2 n_{i(t-2)} + \eta_{2i} + v_{2it}. \tag{130}$$

Univariate AR Estimates for Employment

- We begin by obtaining alternative estimates of a univariate AR(1) model for employment (setting $\alpha_2 = \beta_1 = \beta_2 = 0$).
- Table 2 compares OLS estimates in levels, first-differences, and within-groups with those obtained by GMM using as instruments for the equation in first differences all lags of employment up to $t-2$. The results are broadly consistent with what would be expected for an AR data generation process with unobserved heterogeneity.
- Taking GMM estimates as a benchmark, OLS in levels is biased upwards, and WG and OLS in differences are biased downwards, with a much larger bias in the latter.
- The one- and two-step GMM estimates in the 4-th and 5-th columns, respectively, are based on the sample moments $b_N(\beta) = (b'_{3N}, ..., b'_{8N})'$, where β is the 7×1 parameter vector $\beta = (\alpha, \Delta\delta_3, ..., \Delta\delta_8)'$ and

$$b_{tN} = \frac{1}{738} \sum_{i=1}^{738} \begin{pmatrix} 1 \\ n_i^{t-2} \end{pmatrix} \left(\Delta n_{it} - \Delta\delta_t - \alpha \Delta n_{i(t-1)} \right) \ (t = 3, ..., 8). \tag{131}$$

  $b_N(\beta)$ contains 27 orthogonality conditions, so that there are 20 overidentifying restrictions.
- These are tested with the Sargan statistic. There is a contrast between the value of the one-step Sargan statistic (35.1), which is too high for a chi-square with 20 degrees of freedom, and the robust two-step statistic which is much smaller (15.5).

- This should not be taken as evidence against the overidentifying restrictions, but as an indication of the presence of conditional heteroskedasticity.

- Column 6 in Table 2 reports two-step GMM estimates of an AR(2) model. Since one cross-section is spent in constructing the second lag, the two orthogonality conditions in b_{3N} are lost, so we are left with 25 moments. There is a second autoregressive coefficient but $\Delta\delta_3$ is lost, so the total number of parameters is unchanged.

- Finally, the last column in Table 2 presents continuously updated GMM estimates of the AR(2) model. They use the same moments as GMM2, but the weight matrix is continuously updated.

Table 2

Univariate AR Estimates for Employment

	OLS-levels	OLS-dif.	WG	GMM1	GMM2	GMM2	C.U. GMM2
$n_{i(t-1)}$	0.992	0.054	0.69	0.86	0.89	0.75	0.83
	(0.001)	(0.026)	(0.025)	(0.07)	(0.06)	(0.09)	(0.09)
$n_{i(t-2)}$						0.04	0.03
						(0.02)	(0.02)
Sargan	–	–	–	35.1	15.5	14.4	13.0
(d.f.)				(20)	(20)	(18)	(18)
m_1	2.3	−0.6	−9.0	−8.0	−7.6	−6.0	
m_2	2.2	2.3	0.6	0.5	0.5	0.3	

$N = 738$, $T = 8$, $1983 - 1990$. Heteroskedasticity robust standard errors in parentheses. Time dummies included in all equations.

- From the orthogonality conditions above only first-differences of time effects are directly estimated. The initial time effect can be estimated as

$$\widehat{\delta}_3 = \frac{1}{738}\sum_{i=1}^{738}(y_{i3} - \widehat{\alpha}_1 y_{i2} - \widehat{\alpha}_2 y_{i1}) \tag{132}$$

and, given estimates of their changes, the rest can be estimated recursively from $\widehat{\delta}_t = \widehat{\Delta\delta}_t + \widehat{\delta}_{t-1}$ $(t = 4, ..., 8)$.

- Given the large cross-sectional sample size, the realizations of the time effects in the data can be accurately estimated, but with only 6 time series observations we do not have enough information to consider a stochastic model for δ_t.

- On the other hand, individual effects can be estimated as

$$\widehat{\eta}_i = \frac{1}{T-2}\sum_{s=3}^{T}\widehat{u}_{is} \tag{133}$$

where $\widehat{u}_{is} = y_{is} - \widehat{\delta}_s - \widehat{\alpha}_1 y_{i(s-1)} - \widehat{\alpha}_2 y_{i(s-2)}$.

- Here the situation is the reverse. Since the $\widehat{\eta}_i$ are averages of just $T-2=6$ observations, they will typically be very noisy estimates of realizations of the effects for particular firms.

- However, the variance of η_i can still be consistently estimated for large N.

- Optimal estimation of σ_η^2 and the σ_t^2 requires consideration of the data covariance structure, but noting that the errors in levels $u_{it} \equiv \eta_i + v_{it}$ satisfy $Var(u_{it}) = \sigma_\eta^2 + \sigma_t^2$ and $Cov(u_{it}, u_{is}) = \sigma_\eta^2$, simple consistent estimates can be obtained as:

$$\widehat{\sigma}_\eta^2 = \frac{2}{T(T-1)}\sum_{t=2}^{T}\sum_{s=1}^{t-1}\widehat{Cov}\left(\widehat{u}_{it}, \widehat{u}_{is}\right) \tag{134}$$

$$\widehat{\sigma}_t^2 = \widehat{Var}(\widehat{u}_{it}) - \widehat{\sigma}_\eta^2. \tag{135}$$

- For the AR(2) employment equation Alonso-Borrego and Arellano reported $\widehat{\sigma}_\eta^2 = .038$ and $T^{-1}\sum_{t=1}^{T}\widehat{\sigma}_t^2 = .01$. Thus, variation in firm specific intercepts was approximately 4 times larger than the average random error variance.

- In this example time dummies are important for the model to be accepted by the data. Without them, GMM2 estimates of the AR(2) employment equation in first differences yielded a Sargan statistic of 59.0 ($d.f.18$) without constant, and of 62.7 ($d.f.18$) with constant. Thus, implying a sound rejection of the overidentifying restrictions.

- For the firms in our data set, average growth of employment during the 7 year period 1984-90 is 1 percent, but this is the result of almost no growth in the first two years, 1 percent growth in 1986, 2 percent in 1987-89 and zero or negative growth in 1990.

- Given such pattern, it is not surprising that we reject the restrictions imposed by the cross-sectional orthogonality conditions with a common intercept or a linear trend.

Bivariate VAR Estimates for Employment and Wages

- For the rest of the tutorial we focus on the bivariate model (129)-(130) since it allows us to illustrate a richer class of problems.

- Table 3 presents OLS in levels and GMM2 in differences for employment (columns 1 and 2), and wages (columns 4 and 5).

Table 3

VAR Estimates

	Employment			Wages		
	OLS-levels	GMM2 dif.	GMM2 lev.&dif.	OLS-levels	GMM2 dif.	GMM2 lev.&dif.
$n_{i(t-1)}$	1.11	0.84	1.17	0.08	−0.04	0.08
	(0.03)	(0.09)	(0.03)	(0.03)	(0.10)	(0.03)
$n_{i(t-2)}$	−0.12	−0.003	−0.13	−0.07	0.05	−0.06
	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
$w_{i(t-1)}$	0.14	0.08	0.13	0.78	0.26	0.78
	(0.03)	(0.08)	(0.02)	(0.03)	(0.11)	(0.02)
$w_{i(t-2)}$	−0.11	−0.05	−0.11	0.18	0.02	0.08
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
$\chi^2_{ce}(2)$	41.7	7.2	43.7	26.1	3.3	10.4
p-value	0.00	0.03	0.00	0.00	0.19	0.006
Sargan	–	36.9	61.2	–	21.4	64.2
(d.f.)		(36)	(48)		(36)	(48)
p-value		0.43	0.096		0.97	0.06
m_1	−0.6	−6.8	−8.0	0.05	−5.7	−9.5
m_2	1.6	0.2	1.3	−2.7	0.5	−0.6

$N = 738$, $T = 8$, $1983 - 1990$. Heteroskedasticity robust standard errors in parentheses. Time dummies included in all equations.

$\chi^2_{ce}(2)$ is a Wald test statistic of the joint significance of cross effects.

- The table also contains GMM estimates that combine levels and differences, but these will be discussed below in conjunction with testing for mean stationarity.

- In line with the univariate results, the OLS estimates in levels for both equations are markedly different to GMM2 in differences, and imply a substantially higher degree of persistence, which is consistent with the presence of heterogeneous intercepts.

- The GMM estimates use as instruments for the equations in first differences all the available lags of employment and wages up to $t-2$. With $T = 8$, a second-order VAR and time dummies, there are 36 overidentifying restrictions for each equation. Neither of the Sargan test statistics provide evidence against these restrictions.

- It may be possible to improve the efficiency by jointly estimating the two equations. Optimal joint GMM estimates would use a weight matrix that takes into account the correlation between the moment conditions of the employment and wage equations.

Testing for Residual Serial Correlation

- If the errors in levels are serially independent, those in first differences will exhibit first- but not second-order serial correlation.
- Moreover, the first-order serial correlation coefficient should be equal to -0.5.
- In this regard, an informal but often useful diagnostic is provided by the inspection of the autocorrelation matrix for the errors in first differences.
- Serial correlation matrices for employment and wages based on GMM residuals in first-differences are shown in Table 4, broadly conforming to the expected pattern.

Table 4

(a) GMM1 (dif.) Residual Serial Correlation Matrix for Employment

$$\begin{pmatrix} 1. & & & & \\ -.53 & 1. & & & \\ .10 & -.49 & 1. & & \\ -.04 & -.015 & -.46 & 1. & \\ -.015 & .04 & -.08 & -.44 & 1. \end{pmatrix}$$

(b) GMM1 (dif.) Residual Serial Correlation Matrix for Wages

$$\begin{pmatrix} 1. & & & & \\ -.51 & 1. & & & \\ .03 & -.33 & 1. & & \\ .004 & -.035 & -.42 & 1. & \\ .009 & .00 & -.03 & -.39 & 1. \end{pmatrix}$$

- Formal tests of serial correlation are provided by the m_1 and m_2 statistics reported in Table 3for the VAR model (and also in Table 2 for the univariate results).
- They are asymptotically distributed as $\mathcal{N}(0,1)$ under the null of no autocorrelation, and have been calculated from residuals in first differences (except for OLS in levels).
- So if the errors in levels were uncorrelated, we would expect $m1$ to be significant, but not $m2$, as is the case for the GMM2-dif estimates for employment and wages.

- The m_j statistics (Arellano and Bond, 1991) are moment tests of significance of the average j-th order autocovariance r_j:

$$r_j = \frac{1}{T-3-j} \sum_{t=4+j}^{T} r_{tj} \tag{136}$$

where $r_{tj} = E\left(\Delta v_{it} \Delta v_{i(t-j)}\right)$. Their null is $H_0 : r_j = 0$ and they are given by

$$m_j = \frac{\widehat{r}_j}{SE(\widehat{r}_j)} \tag{137}$$

where $\widehat{r}_j$ is the sample counterpart of r_j based on first-difference residuals $\widehat{\Delta v}_{it}$ and $\widehat{r}_{tj} = N^{-1} \sum_{i=1}^{N} \widehat{\Delta v}_{it} \widehat{\Delta v}_{i(t-j)}$.

- The estimates in Table 3 are based on the assumption that given individual and time effects n_{it} and w_{it} only depend on the past two observations. Provided T is sufficiently large, the m_j statistics can be used to test assumptions on lag length.

Testing for Stationarity in Mean of Initial Observations

- We turn to consider GMM estimates that combine levels and differences, as shown in columns 3 (employment) and 6 (wages) of Table 3.

- For the employment equation, estimates are based on the following 40 moments for errors in differences:

$$b_{tN}^{d} = \sum_{i=1}^{738} \begin{pmatrix} n_i^{t-2} \\ w_i^{t-2} \end{pmatrix} \left(\Delta n_{it} - \Delta \delta_{1t} - \alpha_1 \Delta n_{i(t-1)} - \alpha_2 \Delta n_{i(t-2)} - \beta_1 \Delta w_{i(t-1)} - \beta_2 \Delta w_{i(t-2)}\right) \tag{138}$$

$(t = 4, ..., 8)$,

together with 6 moments for the period-specific constants:

$$b_{tN}^{c} = \sum_{i=1}^{738} \left(n_{it} - \delta_{1t} - \alpha_1 n_{i(t-1)} - \alpha_2 n_{i(t-2)} - \beta_1 w_{i(t-1)} - \beta_2 w_{i(t-2)}\right) \ (t = 3, ..., 8), \tag{139}$$

and 12 additional moments for errors in levels:

$$b_{tN}^{\ell} = \sum_{i=1}^{738} \begin{pmatrix} \Delta n_{i(t-1)} \\ \Delta w_{i(t-1)} \end{pmatrix} \left(n_{it} - \delta_{1t} - \alpha_1 n_{i(t-1)} - \alpha_2 n_{i(t-2)} - \beta_1 w_{i(t-1)} - \beta_2 w_{i(t-2)}\right) \tag{140}$$

$(t = 3, ..., 8)$.

The moments are functions of the 10×1 parameter vector $\beta = (\ \delta_3, ..., \delta_8, \alpha_1, \alpha_2, \beta_1, \beta_2)$, so that there are 48 overidentifying restrictions.

- The estimates for the wage equation were obtained in exactly the same manner.
- Employment and wage changes lagged two periods or more are not used as instruments for the equations in levels because they are redundant given those already included.
- We report two-step robust estimates whose weight matrix is based on the kind of one-step residuals described above.
- Note that, contrary to what we would expect under mean stationarity, the combined levels & differences GMM estimates in both equations are closer to the OLS-levels estimates than to GMM in differences.
- A test of the moment restrictions (140) is a test of whether, given an aggregate time effect, the mean of the distribution of initial observations and the mean of the steady state distribution coincide.
- This can be done by computing incremental Sargan test statistics. Specifically, under the null of mean stationarity, the difference between the $lev.\&dif.$ and the $dif.$ Sargan statistics would be asymptotically distributed as a χ^2 with 12 degrees of freedom.
- Since we obtain $\Delta S_n = 24.3$ (p-val. 0.0185) for employment, and $\Delta S_w = 42.8$ (p-val. 0.00) for wages, the null is rejected for the two equations, although somewhat more marginally so in the case of employment.

Testing for the Presence of Unobserved Heterogeneity

- In the absence of unobserved heterogeneity OLS in levels are consistent estimates, but more generally estimation (eg. of the employment equation) could be based on the following 60 sample moments

$$b^*_{tN} = \sum_{i=1}^{738} \begin{pmatrix} 1 \\ n_i^{t-1} \\ w_i^{t-1} \end{pmatrix} \left(n_{it} - \delta_{1t} - \alpha_1 n_{i(t-1)} - \alpha_2 n_{i(t-2)} - \beta_1 w_{i(t-1)} - \beta_2 w_{i(t-2)}\right) \tag{141}$$

$(t = 3, ..., 8)$.

- Given the 46 moments in (138) and (139), (141) adds the following 14 moments:

$$b^h_{3N} = \sum_{i=1}^{738} \begin{pmatrix} n_{i1} \\ n_{i2} \\ w_{i1} \\ w_{i2} \end{pmatrix} (n_{i3} - \delta_{13} - \alpha_1 n_2 - \alpha_2 n_{i1} - \beta_1 w_{i2} - \beta_2 w_{i1}) \tag{142}$$

$$b_{tN}^{h} = \sum_{i=1}^{738} \begin{pmatrix} n_{i(t-1)} \\ w_{i(t-1)} \end{pmatrix} \left(n_{it} - \delta_{1t} - \alpha_1 n_{i(t-1)} - \alpha_2 n_{i(t-2)} - \beta_1 w_{i(t-1)} - \beta_2 w_{i(t-2)} \right) \qquad (143)$$

$(t = 4, ..., 8)$.

- Thus a test for the validity of the moments (142) and (143) can be regarded as testing for the presence of unobserved heterogeneity.

- This can be done by calculating combined GMM estimates based on (138), (139), (142) and (143) -or equivalently levels-GMM estimates based on (141)- and obtaining the corresponding incremental Sargan tests relative to GMM in differences.

- The resulting estimates for employment and wages are very close to OLS, and both incremental tests reject the absence of unobserved heterogeneity. The incremental Sargan statistics ($d.f. = 14$) take the values $\Delta S_n^h = 36.0$ (p-val. 0.001) for employment, and $\Delta S_w^h = 47.2$ (p-val. 0.00) for wages.

Testing for Granger Non-Causality with and without Heterogeneity

- The hypothesis that employment does not Granger-cause wages conditional on individual and time effects imposes the restrictions $\lambda_1 = \lambda_2 = 0$. Conversely, to test whether wages Granger-cause employment we examine the validity of $\beta_1 = \beta_2 = 0$.

- The testing of these restrictions is of some interest in our example because a version of model (129)-(130) in which the wage equation only includes its own lags can be regarded as the reduced form of an intertemporal labour demand model under rational expectations (as in Sargent, 1978).

- Wald test statistics of the joint significance of cross-effects are reported in Table 3 for the two equations. For the GMM2 estimates in first-differences we find that wages Granger-cause employment, but employment does not Granger-cause wages.

- An interesting point is that conditioning on individual effects is crucial for this result. As shown in Table 3, if the tests were based upon the OLS estimates in levels, the hypothesis that employment does not Granger-cause wages would be clearly rejected. This illustrates how lack of control of individual heterogeneity could result in a spurious rejection of non causality.

- Moreover, Granger non-causality would also be rejected using the estimates that impose mean stationarity of the initial observations. Thus, in short panels assumptions about initial conditions also matter for the assessment of non causality.

References

[1] Abowd, J.M. and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes", *Econometrica*, 57, 411-445.

[2] Ahn, S. and P. Schmidt (1995): "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-27.

[3] Alonso-Borrego, C. and M. Arellano (1999): "Symmetrically Normalized Instrumental-Variable Estimation Using Panel Data", *Journal of Business & Economic Statistics*, 17, 36-49.

[4] Anderson, T. W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components", *Journal of the American Statistical Association*, 76, 598-606.

[5] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press, Oxford.

[6] 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.

[7] Arellano, M. and O. Bover (1995): "Another Look at the Instrumental-Variable Estimation of Error-Components Models", *Journal of Econometrics*, 68, 29-51.

[8] Barro, R. J. and X. Sala-i-Martin (1995): *Economic Growth*, McGraw-Hill, New York.

[9] Blundell, R. and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models", *Journal of Econometrics*, 87, 115-143.

[10] Blundell, Richard, Luigi Pistaferri, and Ian Preston (2008): "Consumption Inequality and Partial Insurance", *American Economic Review*, 98, 1887-1921.

[11] Deaton, A. (1991): "Saving and Liquidity Constraints", *Econometrica*, 59, 1221-1248.

[12] Hall, R. and F. Mishkin (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", *Econometrica*, 50, 461-481.

[13] Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371-1395.

[14] Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects", *Econometrica*, 49, 1417-1426.