UNIVERSITÀ DEGLI STUDI DI VERONA

Poor identification and estimation problems in panel data models with random effects and autocorrelated errors

by

Giorgio Calzolari and Laura Magazzini

January 2009 - #53

WORKING PAPER SERIES

Dipartimento di Scienze Economiche

Poor identification and estimation problems in panel data models with random effects and autocorrelated errors

Giorgio Calzolari* Laura Magazzini†‡

February 10, 2009

Abstract

A dramatically large number of corner solutions occur when estimating by (Gaussian) maximum likelihood a simple model for panel data with random effects and autocorrelated errors. This can invalidate results of applications to panel data with a short time dimension, even in a correctly specified model. We explain this unpleasant effect (usually underestimated, almost ignored in the literature) showing that the expected log-likelihood is nearly flat, thus rising problems of poor identification.

1 Introduction

In a standard linear model for panel data, the residual is decomposed into two terms: an individual specific component, constant over time, and a disturbance term assumed to be homoschedastic and uncorrelated over time and among individuals. This structure generates equal correlation between the error terms of the same unit at different points in time, that does not wipe out as the time lag between the two observations increases. This assumption might not be appropriate in case of economic relationship, where an unobserved shock likely prolongs its effect at subsequent times, causing different patterns of autocorrelation. Ignoring it leads to estimates of the regression coefficient that are consistent but inefficient, and estimated standard errors are biased. In order to take into account this phenomenon, the disturbance can be assumed to follow an AR or MA process, where the difference is in the persistency of the unobserved shocks over time.
We show in this paper the unpleasant behavior of the Gaussian maximum likelihood estimates for a linear panel data model with random effects and AR(1) idiosyncratic noise. The problem, neglected in the literature, can easily invalidate results of applications. We deal with a "correctly specified" Gaussian framework, with various combinations of the three error terms parameters (the variance of the individual random effect, the variance of the idiosyncratic noise, the autoregression parameter). Monte Carlo simulation of the data, followed by maximum likelihood estimation, very often produces a zero estimate of the variance of the individual random effect parameter (corner solution). Correspondingly, there is a shift of the other parameters, whose distrubutions are no more scattered only around their "true" values. Bi-modal distributions become the rule, rather than an exception.

*Department of Statistics, Università di Firenze, <calzolar@ds.unifi.it>

†Department of Economics, Università di Verona, <laura.magazzini@univr.it>

‡An earlier version was presented at the Third Italian Congress of Econometrics and Empirical Economics, Ancona, January 30-31, 2009. We gratefully acknowledge comments and suggestions from E. Battistin, G. Fiorentini, D. Lubian, F. Mealli, G. Millo, P. Paruolo, T. Proietti, C. Rampichini, A. Rossi, A. Sembenelli, E. Sentana, and M. Wagner. but we retain full responsibility for the contents of this paper.

The problem occurs in a dramatically large number of cases when the time dimension is very small, no matter what combination of parameters is adopted (balanced or completely unbalanced values of the two variances, small or large value of the autoregression parameter). Only enormously large numbers (unreasonable) of individuals can compensate very small time dimensions. For moderately small time dimensions (T) the problem is still relevant when the autoregression parameter is large, for any combination of the two variance parameters. For a large value of the autoregression parameter and a variance of the individual random effect smaller than the idiosyncratic variance, the problem still occurs even for moderately large values of T.
For any combination of parameters, the problem becomes less important when the time dimension enlarges, and in practice it always disappears for large T. It is quite relevant for the N and T dimensions used in practical applications.

2 Notation and related literature

The data generating process on the dependent variable y_{it} is expressed as a linear function of a set of independent variables, x_{it}, and an error term, ν_{it}:

$$y_{it} = x'_{it}\beta + \nu_{it} \tag{1}$$

where i denotes the unit (individual, household, firm, country, ...), and the index t denotes the time, with $i = 1, ..., N; t = 1, ..., T$.
When analysing panel data, the error structure can be decomposed into three independent terms:

$$\nu_{it} = \alpha_i + \lambda_t + e_{it} \tag{2}$$

where α_i is the individual effect, representing all the time-invariant (unobserved or unobservable) characteristics of unit i, λ_t is the time effect, representing all the characteristics of time t, invariant across all the cross-sectional units in the sample, and e_{it} is a random term that varies over time and individuals. The error term λ_t is not considered in the analysis, since it can be easily accounted for in a typical (short-T) panel setting by inserting time-dummies in the regression. Thus, we are considering

$$\nu_{it} = \alpha_i + e_{it} \tag{3}$$

In standard settings, the error term e_{it} is assumed to be serially uncorrelated. This assumption is not suited to situations where the effect of unobserved variables vary sistematically over time, as in the case of serially correlated omitted variables or transitory variables whose effect lasts more than one period. In order to take into account these variables and provide a more general autocorrelation scheme, autocorrelation can be considered among e_{it} for the same individual[1].
In this paper we will consider disturbances that are correlated over time, generated by an AR(1) process[2]:

$$e_{it} = \rho e_{it-1} + w_{it} \tag{4}$$

with w_{it} homoschedastic, uncorrelated, and with mean zero. Gaussianity of all the error terms is assumed.
As a result of these assumptions, the variance-covariance matrix of the error term ν_{it} has the following structure:

$$E[\nu_{it}\nu_{js}] = \begin{cases} \sigma_\alpha^2 + \sigma_e^2 & \text{if } i = j, t = s \\ \sigma_\alpha^2 + \rho^{|t-s|}\sigma_e^2 & \text{if } i = j, t \neq s \\ 0 & \text{if } i \neq j \end{cases} \tag{5}$$

[1] Recent research in linear models with random effects is considering serial correlation in the time dimension (Karlsson and Skoglund 2004).
[2] A general error structure has been considered by MaCurdy (1982).

If the explanatory variables x_{it} are stricly exogenous, there is no effect of the x_{it} on the problems we are considering in this paper. We can therefore focus on the error terms parameters only: σ_α^2, σ_e^2, ρ.
Assuming Gaussianity, the likelihood depends only on the variance-covariance matrix, thus only on the three parameters σ_α^2, σ_e^2, ρ.
Autocorrelated disturbances in panel data linear models have been considered for the first time in longitudinal studies of wages and earnings (David 1971, Hause 1973, Lillard and Willis 1978, Lillard and Weiss 1979, MaCurdy 1982, Bhargava, Franzini and Narendranathan 1982).
Lillard and Willis (1978) estimate an earning function with permanent and serially correlated transitory components due to both measured and unmeasured variables. As transitory effects and permanent effects have different economic implications, it is important to separate the permanent and transitory elements of earnings development. In order to estimate model parameters, the authors first apply OLS to the data pooled over individuals and years, and then estimate the variance components and the autocorrelation parameter by applying maximum likelihood to the OLS residuals. A maximum likelihood approach to estimation of the components of the variance matrix is also exploited in Lillard and Weiss (1979). This two-step approach is asymptotically equivalent to full quasi-maximum likehood procedure, robust to failure of the normality hypothesis (MaCurdy 1982).
First order serial correlation in the disturbances within a fixed effect framework was considered by Bhargava et al. (1982), who also proposed a Durbin-Watson type statistics to test the model for serial independence and random walk hypothesis.
A series of LM statistics for testing the presence of serial correlation and individual effect is devised by Baltagi and Li (1995). The proposed LM statistics are invariant to the form of first order serial correlation, i.e. they can be applied both to AR(1) and MA(1) processes. Size and power of the tests are studied by means of Monte Carlo simulation for various combination of the autocorrelation parameter and ratio of the individual effect variance over the amount of the composite error variance. Moreover different sample sizes are considered, changing both N and T, but all experiments are run with $T \geq 10$.
More recently, the one-way error component model with AR(1) disturbances has been applied to study the impact of plant closing on the mean and variance of log earnings (Berry, Gottschalk and Wissoker 1988).
If the assumption of spherical disturbances for e_{it} is violated, as it is in our AR(1) setting, the ordinary formulae for estimating coefficient variances will lead to inconsistent standard errors[3].

3 Identifiability and examples of poor identification

If $T = 2$ the covariance matrix would be

$$\Sigma = Cov(\nu_i) = \begin{bmatrix} \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 + \sigma_e^2\rho \\ \sigma_\alpha^2 + \sigma_e^2\rho & \sigma_\alpha^2 + \sigma_e^2 \end{bmatrix} \tag{6}$$

thus it contains only two *different* elements, from which it is impossible to identify separately the three parameters of the auxiliary model σ_α^2, σ_e^2 and ρ.
$T = 3$ is the smallest possible time dimension; in this case in fact the covariance matrix would

[3]Within the GLS framework, an estimator of the variance exits that is robust to the presence of heteroskedasticity and serial correlation of arbitrary form for fixed T and large N (Kiefer 1980). However, robust estimation can have poor finite sample properties as it requires the estimation of $T(T-1)/2$ parameters (Wooldridge 2002).

be

$$\Sigma = Cov(\nu_i) = \begin{bmatrix} \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 + \sigma_e^2\rho & \sigma_\alpha^2 + \sigma_e^2\rho^2 \\ \sigma_\alpha^2 + \sigma_e^2\rho & \sigma_\alpha^2 + \sigma_e^2 & \sigma_\alpha^2 + \sigma_e^2\rho \\ \sigma_\alpha^2 + \sigma_e^2\rho^2 & \sigma_\alpha^2 + \sigma_e^2\rho & \sigma_\alpha^2 + \sigma_e^2 \end{bmatrix} \tag{7}$$

where the different elements are three, making identification possible. Of course the situation becomes even better for larger values of T, where the presence of higher powers of ρ increases the number of different elements in the matrix.

In practice, however, identification can be very *poor* when ρ is moderately high, even for values that would not be considered *dangerously* close to 1 in a time series context. Some numerical examples could well exemplify the problem.

The following are the covariance matrices produced by different combinations of the three parameters. They have been computed for T up to 20, but are fully displayed for T up to 6; for smaller values of T, the covariance matrix would simply be the first $(T \times T)$ block of each matrix.

$\sigma_\alpha^2 = 0.5$, $\sigma_e^2 = 0.5$, $\rho = 0.9$

$$\begin{bmatrix} 1.000 & 0.950 & 0.905 & 0.864 & 0.828 & 0.795 & ..... \\ 0.950 & 1.000 & 0.950 & 0.905 & 0.864 & 0.828 & ..... \\ 0.905 & 0.950 & 1.000 & 0.950 & 0.905 & 0.864 & ..... \\ 0.864 & 0.905 & 0.950 & 1.000 & 0.950 & 0.905 & ..... \\ 0.828 & 0.864 & 0.905 & 0.950 & 1.000 & 0.950 & ..... \\ 0.795 & 0.828 & 0.864 & 0.905 & 0.950 & 1.000 & ..... \\ .... & .... & .... & .... & .... & .... & ..... \end{bmatrix} \begin{bmatrix} log\ of\ determinant \\ (T = 20) - 44.318; \\ ..... \\ (T = 6) - 11.646; \\ (T = 5) - 9.3154; \\ (T = 4) - 6.9856; \\ (T = 3) - 4.6565; \end{bmatrix}$$

$\sigma_\alpha^2 = 0.4$, $\sigma_e^2 = 0.6$, $\rho = 0.9175$

$$\begin{bmatrix} 1.000 & 0.950 & 0.905 & 0.863 & 0.825 & 0.790 & ..... \\ 0.950 & 1.000 & 0.950 & 0.905 & 0.863 & 0.825 & ..... \\ 0.905 & 0.950 & 1.000 & 0.950 & 0.905 & 0.863 & ..... \\ 0.863 & 0.905 & 0.950 & 1.000 & 0.950 & 0.905 & ..... \\ 0.825 & 0.863 & 0.905 & 0.950 & 1.000 & 0.950 & ..... \\ 0.790 & 0.825 & 0.863 & 0.905 & 0.950 & 1.000 & ..... \\ .... & .... & .... & .... & .... & .... & ..... \end{bmatrix} \begin{bmatrix} log\ of\ determinant \\ (T = 20) - 44.458; \\ ..... \\ (T = 6) - 11.691 \\ (T = 5) - 9.3525 \\ (T = 4) - 7.0139 \\ (T = 3) - 4.6757 \end{bmatrix}$$

$\sigma_\alpha^2 = 0.1$, $\sigma_e^2 = 0.9$, $\rho = 0.9450$

$$\begin{bmatrix} 1.000 & 0.950 & 0.904 & 0.860 & 0.818 & 0.778 & ..... \\ 0.950 & 1.000 & 0.950 & 0.904 & 0.860 & 0.818 & ..... \\ 0.904 & 0.950 & 1.000 & 0.950 & 0.904 & 0.860 & ..... \\ 0.860 & 0.904 & 0.950 & 1.000 & 0.950 & 0.904 & ..... \\ 0.818 & 0.860 & 0.904 & 0.950 & 1.000 & 0.950 & ..... \\ 0.778 & 0.818 & 0.860 & 0.904 & 0.950 & 1.000 & ..... \\ .... & .... & .... & .... & .... & .... & ..... \end{bmatrix} \begin{bmatrix} log\ of\ determinant \\ (T = 20) - 44.418; \\ ..... \\ (T = 6) - 11.689 \\ (T = 5) - 9.3508 \\ (T = 4) - 7.0131 \\ (T = 3) - 4.6754 \end{bmatrix}$$

$\sigma_\alpha^2 = 0.0$, $\sigma_e^2 = 1.0$, $\rho = 0.9505$

$$\begin{bmatrix} 1.000 & 0.950 & 0.903 & 0.859 & 0.816 & 0.776 & ..... \\ 0.950 & 1.000 & 0.950 & 0.903 & 0.859 & 0.816 & ..... \\ 0.903 & 0.950 & 1.000 & 0.950 & 0.903 & 0.859 & ..... \\ 0.859 & 0.903 & 0.950 & 1.000 & 0.950 & 0.903 & ..... \\ 0.816 & 0.859 & 0.903 & 0.950 & 1.000 & 0.950 & ..... \\ 0.776 & 0.816 & 0.859 & 0.903 & 0.950 & 1.000 & ..... \\ .... & .... & .... & .... & .... & .... & ..... \end{bmatrix} \begin{bmatrix} log\ of\ determinant \\ (T = 20) - 44.416; \\ ..... \\ (T = 6) - 11.688 \\ (T = 5) - 9.3508 \\ (T = 4) - 7.0131 \\ (T = 3) - 4.6754 \end{bmatrix}$$

Of course, the matrices and the corresponding determinants (for the same values of T) are not exactly equal, but quite close to each other, thus suggesting that "problems" are likeley to occur. Differences become larger for larger values of T (so that higher powers of ρ could make the difference), thus suggesting that problems will occur more often when the time dimension is small.

4 Expected log-likelihood and identification

Ignoring the exogenous explanatory variables, the log-likelihood of the model (1) is

$$logL(\nu, \theta) = \sum_{i=1}^{N} logL(\nu_i, \theta) = -\frac{N}{2} log(2\pi) - \frac{N}{2} log||B|| - \frac{1}{2} \sum_{i=1}^{N} (\nu_i' B^{-1} \nu_i) \tag{8}$$

where $\theta = [\sigma_\alpha^2, \sigma_e^2, \rho]'$ is the vector of parameters, and B is the $(T \times T)$ covariance matrix of the vector of error terms ν_i (3, 4). As from (5), B is therefore the sum of a matrix whose elements are all σ_α^2, say $\sigma_\alpha^2 \iota\iota'$ (where ι is the $T \times 1$ vector whose elements are all $= 1$) and the covariance matrix of the AR(1) process (4) with the well known Toepliz structure (let's call it $\sigma_e^2 A$)

$$B = \sigma_\alpha^2 \iota\iota' + \sigma_e^2 A = \sigma_\alpha^2 \iota\iota' + \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^{T-1} \\ \rho & 1 & \rho & ... & \rho^{T-2} \\ \rho^2 & \rho & 1 & ... & \rho^{T-3} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & ... & 1 \end{bmatrix} \tag{9}$$

Remembering that the particular structure of matrix A leads to a simple expression of the determinant $|A| = \left(1 - \rho^2\right)^{T-1}$ and of the inverse (a tridiagonal matrix)

$$A^{-1} = \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & -\rho & 0 & ... & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & ... & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & ... & -\rho & 1 \end{bmatrix} \tag{10}$$

the computation of the determinant and inverse of matrix B greatly benefits from the use of Woodbury (or Sherman-Morrison) matrix inversion lemma, both in terms of computation time

as well as (more important) computation accuracy. Its application (tedious but straightforward) gives

$$|B| = \left\{ \left(1-\rho^2\right)^{T-1} + \frac{\sigma_\alpha^2 \left(1-\rho^2\right)^{T-1}}{\sigma_e^2(1+\rho)} \left[T-(T-2)\rho\right] \right\} \left(\sigma_e^2\right)^T \tag{11}$$

$$B^{-1} = \frac{1}{\sigma_e^2(1-\rho^2)} \begin{bmatrix} 1 & -\rho & 0 & ... & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & ... & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & ... & -\rho & 1 \end{bmatrix}$$

$$- \frac{\sigma_\alpha^2}{\sigma_e^2(1+\rho)\left\{\sigma_e^2(1+\rho)+[T-(T-2)\rho]\sigma_\alpha^2\right\}}$$

$$\begin{bmatrix} 1 & (1-\rho) & (1-\rho) & (1-\rho) & ... & (1-\rho) & 1 \\ (1-\rho) & (1-\rho)^2 & (1-\rho)^2 & (1-\rho)^2 & ... & (1-\rho)^2 & (1-\rho) \\ (1-\rho) & (1-\rho)^2 & (1-\rho)^2 & (1-\rho)^2 & ... & (1-\rho)^2 & (1-\rho) \\ ... & ... & ... & ... & ... & ... & ... \\ (1-\rho) & (1-\rho)^2 & (1-\rho)^2 & (1-\rho)^2 & ... & (1-\rho)^2 & (1-\rho) \\ (1-\rho) & (1-\rho)^2 & (1-\rho)^2 & (1-\rho)^2 & ... & (1-\rho)^2 & (1-\rho) \\ 1 & (1-\rho) & (1-\rho) & (1-\rho) & ... & (1-\rho) & 1 \end{bmatrix} \tag{12}$$

Notice that there is no matrix inversion in the last two equations, so numerical correctness does not depend on routines that perform numerical inversions. If one is not scared by the complexity of the equations above, their use guarentees the accuracy that sometimes we need when suspicion arises about the numerical value of a gradient at the maximum: is it zero or is it not zero? We shall frequently meet gradients whose norm is less than 10^{-6}, and are not zeroes (not to mention cases where 10^{-11} have to be discarded; any "standard" optimization program would achieve convergence at such points, and this might produce misleading results).
Let the vector ν_i be produced by a multivariate normal distribution with zero mean and "true" covariance matrix

$$R = \psi_\alpha^2 \iota\iota' + \psi_e^2 A = \psi_\alpha^2 \iota\iota' + \psi_e^2 \begin{bmatrix} 1 & \phi & \phi^2 & ... & \phi^{T-1} \\ \phi & 1 & \phi & ... & \phi^{T-2} \\ \phi^2 & \phi & 1 & ... & \phi^{T-3} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & ... & 1 \end{bmatrix} \tag{13}$$

Taking expectation of equation (8), the expected value of the log-likelihood computed at any value of $\theta = [\sigma_\alpha^2, \sigma_e^2, \rho]'$ is

$$E\left[logL(\nu,\theta)\right] = N\ E\left[logL(\nu_i,\theta)\right] = -\frac{N}{2}\left[log(2\pi) + log||B|| + tr(B^{-1}R)\right] \tag{14}$$

Some suitable version of the information inequality ensures that, in general, the expression above has *the maximum* at $B = R$. This ensures that, in principle, the model is identified, since the expected likelihood has a unique maximum at the "true" parameter values: $\sigma_\alpha^2 = \psi_\alpha^2$, $\sigma_e^2 = \psi_e^2$, $\rho = \phi$. However, the particular stucture of the matrices R and B can make this maximum very hardly identifiable.

For example, for error terms produced by "true" parameter values $\psi_\alpha^2 = 0.5$, $\psi_e^2 = 0.5$, $\phi = 0.5$ and a "large" time dimension $T = 20$, comparing the expected log-likelihood at the "true" values $\sigma_\alpha^2 = 0.5$, $\sigma_e^2 = 0.5$, $\rho = 0.5$ with the value at $\sigma_\alpha^2 = 0.0$, $\sigma_e^2 = 1.0$, $\rho = 0.75$ the difference is about 50%; so the two points are easily distinguishable. But for $T = 5$ the difference is 6%, and for $T = 3$ (the smallest possible value) the difference is only a bit more than 1%. The risk of confusing the two points becomes not negligeable. The risk will be much higher in the examples that will be discussed later.

5 Expected score and Hessian

We now remember that

$$\frac{\partial log||B||}{\partial B} = B^{-1} \qquad \frac{\partial B^{-1}}{\partial \theta_j} = -B^{-1}\frac{\partial B}{\partial \theta_j}B^{-1} \tag{15}$$

(where θ_j is one of the three elements of the parameters vector $\theta = [\sigma_\alpha^2, \sigma_e^2, \rho]'$). We then observe that derivative of the matrix A with respect to ρ is the matrix whose h, k-th element is $|h-k|\rho^{|h-k|-1}$

$$D = \frac{\partial A}{\partial \rho} = \begin{bmatrix} 0 & 1 & 2\rho & 3\rho^2 & ... & (T-1)\rho^{T-2} \\ 1 & 0 & 1 & 2\rho & ... & (T-2)\rho^{T-3} \\ 2\rho & 1 & 0 & 1 & ... & (T-3)\rho^{T-4} \\ ... & ... & ... & ... & ... & ... \\ (T-2)\rho^{T-3} & ... & ... & ... & ... & 1 \\ (T-1)\rho^{T-2} & ... & ... & ... & ... & 0 \end{bmatrix}$$

The expressions above allow to write in closed form the expected value of the score

$$E\left[\frac{\partial logL(\nu,\theta)}{\partial \theta}\right] = \frac{N}{2}\begin{bmatrix} -\iota'(B^{-1})\iota & + & \iota'(B^{-1}RB^{-1})\iota \\ -tr(B^{-1}A) & + & tr(B^{-1}AB^{-1}R) \\ -tr(B^{-1}D)\sigma_e^2 & + & tr(B^{-1}DB^{-1}R)\sigma_e^2 \end{bmatrix} \tag{16}$$

We now indicate with F the $T \times T$ matrix $F = \partial D/\partial\rho = \partial^2 A/\partial\rho^2$, whose h, k-th element is $|h-k|(|h-k|-1)\rho^{|h-k|-2}$

$$F = \frac{\partial D}{\partial \rho} = \frac{\partial^2 A}{\partial \rho^2} = \begin{bmatrix} 0 & 0 & 2 & 6\rho & .. & (T-1)(T-2)\rho^{T-3} \\ 0 & 0 & 0 & 2 & .. & (T-2)(T-3)\rho^{T-4} \\ 2 & 0 & 0 & 0 & .. & (T-3)(T-4)\rho^{T-5} \\ 6\rho & 2 & 0 & 0 & .. & (T-4)(T-5)\rho^{T-6} \\ .. & .. & .. & .. & .. & .. \\ (T-2)(T-3)\rho^{T-4} & .. & .. & .. & .. & 0 \\ (T-1)(T-2)\rho^{T-3} & .. & .. & .. & .. & 0 \end{bmatrix}$$

Some tedious but straightforward algebra allows to compute in closed form the second derivatives of the expected log-likelihood (the 3×3 expected Hessian matrix)

$$H = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial\theta\partial\theta'}\right] = \begin{bmatrix} h_{1,1} & h_{1.2} & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & h_{3,2} & h_{3,3} \end{bmatrix} \tag{17}$$

where

$$h_{1,1} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial(\sigma_\alpha^2)^2}\right] = \frac{N}{2}\left[\iota' B^{-1}\iota\iota' B^{-1}\iota - 2\iota' B^{-1}RB^{-1}\iota\iota' B^{-1}\iota\right]$$

$$h_{2,1} = h_{1,2} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial\sigma_\alpha^2\partial\sigma_e^2}\right] = \frac{N}{2}tr\left[B^{-1}\iota\iota' B^{-1}A - 2B^{-1}\iota\iota' B^{-1}AB^{-1}R\right]$$

$$h_{3,1} = h_{1,3} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial\sigma_\alpha^2\partial\rho}\right] = \frac{N}{2}tr\left[B^{-1}\iota\iota' B^{-1}D - 2B^{-1}\iota\iota' B^{-1}DB^{-1}R\right]\sigma_e^2$$

$$h_{2,2} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial(\sigma_e^2)^2}\right] = \frac{N}{2}tr\left[B^{-1}AB^{-1}A - 2B^{-1}AB^{-1}AB^{-1}R\right]$$

$$h_{3,2} = h_{2,3} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial\sigma_e^2\partial\rho}\right]$$

$$= \frac{N}{2}tr\left[B^{-1}DB^{-1}A - 2B^{-1}DB^{-1}AB^{-1}R\right]\sigma_e^2 + tr\left[B^{-1}DB^{-1}R - B^{-1}D\right]$$

$$h_{3,3} = E\left[\frac{\partial^2 logL(\nu,\theta)}{\partial\rho^2}\right]$$

$$= \frac{N}{2}tr\left[B^{-1}DB^{-1}D\sigma_e^2 - 2B^{-1}DB^{-1}DB^{-1}R\sigma_e^2 + B^{-1}FB^{-1}R - B^{-1}F\right]\sigma_e^2$$

With reference to the numerical example of the previous section, we first compute, for several values of T (till the smallest possible value which is $T = 3$), the expected log-likelihood at the "true" values of the parameters 0.5, 0.5, 0.9, where it attains its maximum. We then "constrain" σ_α^2 at different values between the true value (0.5) and 0, and for each of them we maximize the expected log-likelihood with respect to the other two parameters. In Table 1 we display the absolute difference between each constrained maximum and the absolute maximum (that might be relevant for hypotheses testing), as well as the relative difference (that might be relevant when a poor computational accuracy does not guarantee to discriminate between the two). The

			diff.	diff.	diff.
σ_α^2	σ_e^2	ρ	$T = 20$	$T = 5$	$T = 3$
0.4	near 0.6	near 0.91	$N \times .003250$	$N \times .000208$	$N \times .000036$
0.2	near 0.8	near 0.93	$N \times .023391$	$N \times .001072$	$N \times .000185$
0.0	1.0	0.95	$N \times .044034$	$N \times .001909$	$N \times .000329$
			%diff.	%diff.	%diff.
σ_α^2	σ_e^2	ρ	$T = 20$	$T = 5$	$T = 3$
0.4	near 0.6	near 0.91	0.03%	0.01%	0.004%
0.2	near 0.8	near 0.93	0.2%	0.05%	0.02%
0.0	1.0	0.95	0.4%	0.1%	0.04%

Table 1: Constrained maximization at σ_α^2 *versus* absolute maximum at the true values $\psi_\alpha^2 = 0.5, \psi_e^2 = 0.5, \phi = 0.90$

constrained maxima for σ_α^2 fixed at 0.4 or 0.2 are at values of the other two parameters always near the values displayed in the tables, when T varies from 20 to 3. However, when σ_α^2 is fixed at zero, always the constrained maximum is attained for a σ_e^2 equal to the sum of the two "true" variances, and ρ exactly at the average between the "true" value and 1.0: thus, respectively, 1.0 and 0.95 in the tables, but this last rule is valid also for any other values of the "true" parameters. Observing each case in greater detail, one would find that, varying σ_α^2 slowly from the "true" value to zero, the constrained maximum decreases monotonically: faster, if T is large, very very slowly if T is small. The difference between the maximum and the value at zero is only

$N \times 0.000329$ when $T = 3$, thus one would need a value of N of several thousands to discriminate between the two maxima with some kind of criterion (e.g. likelihood ratio). Moreover, in relative terms, the two maxima differ only by 0.04%: an optimization algorithm must be very accurately applied, to be able to discriminate between the two. And what just described is a comparison at the two extremes of the interval for σ_α^2; differences with respect to intermediate values would be even more difficult to appreciate (0.004% between the constrained maxima at $\sigma_\alpha^2 = 0.5$ and $\sigma_\alpha^2 = 0.4$).

A look at first and second derivatives of the expected log-likelihood shows that, starting from three first derivatives (expected score) equal zero at the maximum, the derivative with respect to σ_α^2 is no more zero at the constrained maxima, but nevertheless it is very small even when departing far away from the true value, if T is small. For example, still in the case of the previous tables, when T has the smallest value ($T = 3$), the norm of the (average) expected score (sum of the squared first derivatives) quickly grows from zero to a value 5×10^{-7} as soon as we constrain σ_α^2 to a value slightly smaller than the "true" 0.5; then this norm varies very little (say between 5×10^{-7} and 3×10^{-7}) when σ_α^2 is constrained to values smaller and smaller, till $\sigma_\alpha^2 = 0$. A stopping rule of an optimization algorithm based on such a norm might easily fail and stop at a false maximum.[4]

Second order derivatives could be helpful at this point. The three eigenvalues of the Hessian matrix have the correct sign (negative) when derivatives are computed at the "true" value (or absolute maximum). But one of them is very small in absolute value. Still for the example above, still for the most critical case of $T = 3$, the smallest (in absolute value) eigenvalue is -0.0052 (the largest is -103.8). When σ_α^2 is constrained to zero, two eigenvalues remain negative, while the smallest (in absolute value) is positive ($+0.00033$). Thus, if the numbers were "exact", we might conclude that $\sigma_\alpha^2 = 0$ does not provide a maximum, but something close to a saddle point. But could we really trust in the change of sign of the smallest eigenvalue from -0.0052 to $+0.00033$ when a matrix is so ill conditioned? Maybe we can trust because all derivatives, till second order, are supposed to be very accurate, having been computed analytically and without using any "numerical" routine for matrix inversion, thanks to the use of the Woodbury (or Sherman-Morrison) matrix inversion lemma; for sure, there would be no warranty of this kind if derivatives were computed numerically, as it usually happens. Notice, en passing, that the smallest eigenvalue would remain with the correct negative sign till at some distance from the absolute maximum: at $\sigma_\alpha^2 = 0.3$ the norm is about 5×10^{-7} and all eigenvalues are negative. Easily, a reasonable stopping rule might incorrectly find this as a satisfactory maximum; only some "patologically tight" convergence criteria, as we have applied in our computations, can detect the existence of the problem, with the serious risk of misleading false maxima.

The conclusion of this section seems therefore as follows: the model is identified, but the identification rules are so close to be violated for small values of T that it seems appropriate to speak of "poor" identification.

[4] Playing with numbers, and joking on a problem that might be serious... For the same values of the variances just considered in the example, but with an autoregression parameter 0.99, if the time dimension is very small ($T = 3$), the expected log-likelihoods at the "true" parameter values 0.5, 0.5, 0.99 and false values 0.0, 1.0, 0.995 differ approximately by 0.0001%. Between the "true" parameter values and the false values 0.4, 0.6, 0.99167, the difference is about 0.00001%; in absolute vale, the difference is $N \times 3.5 \times 10^{-7}$, thus suggesting that discrimination between "true" and false maxima by means of some criterion (e.g. likelihood ratio) might require a panel with some 10^7 individuals! In none of the points the norm (sum of the squared elements of the average expected score) is greater than 3×10^{-11}.

6 Log-likelihood and Monte Carlo set up

The discussion in the previous section was concerned with "expected" log-likelihoods, as well as "expected" scores etc. Remarkably small differences have been evidenced for expected log-likelihoods at quite different parameter values. This suggests that "swaps" might easily occur when passing from "expected log-likelihoods" (14) to log-likelihoods computed from a sample (8). It will be enough for the sample log-likelihood around the "true" parameter values to be just a bit smaller than its expectation, and/or sample log-likelihood computed at parameter values far away from the "true" to be just a bit higher, and maximum likelihood estimates would fall in a wrong place. This section aims at showing that, very often, this is indeed the case.
As "poor" identifiability pertains to the estimation of variance components and of the autocorrelation parameter, we will not consider exogenous variable in the simulations. Moreover, when using quasi-maximum likelihood method (robust to distributional assumptions) for the estimation of covariance components, efficient estimation can be achieved by performing a two-step procedure, where on the first step a consistent estimator of regression coefficient is considered, and in the second step maximum likelihood estimation is performed on estimated residuals (MaCurdy 1982).
The model is set as follows:

$$y_{it} = \alpha_i + e_{it}$$

where $\alpha_i \sim IN(0, \sigma^2_\alpha)$, and $e_{it} = \rho e_{i,t-1} + w_{it}$ with $w_{it} \sim IN(0, \sigma^2_e \times (1-\rho^2))$.
By means of a wide Monte Carlo experiment, we shall show that, whatever the values of the two variance parameters, balanced as in the examples above ($\sigma^2_\alpha = 0.5$ and $\sigma^2_e = 0.5$), or quite unbalanced in either way ($\sigma^2_\alpha = 0.1$ and $\sigma^2_e = 0.9$, or viceversa $\sigma^2_\alpha = 0.9$ and $\sigma^2_e = 0.1$), corner solutions for the maxima can occur, providing an estimate $\hat{\sigma}^2_\alpha = 0.0$; correspondingly, the estimate of σ^2_e will compensate the zero, giving a value around the sum of the two "true" variances; the estimate of ρ will be a value close to the average between the "true" value and one. We consider six different sample sizes, with $N = 100$ and $N = 1000$ units, and $T = 3, 5, 20$.

		σ^2_α (with $\sigma^2_\alpha + \sigma^2_e = 1$)					
	ρ	0.1	0.5	0.9	0.1	0.5	0.9
$T = 3$	0.0	17.0%	0%	0%	0%	0%	0%
	0.5	39.0%	12.0%	2.0%	18.0%	0%	0%
	0.9	45.0%	40.0%	34.0%	45.0%	25.0%	5.0%
$T = 5$	0.0	1.0%	0%	0%	0%	0%	0%
	0.5	22.0%	0%	0%	0%	0%	0%
	0.9	43.0%	25.0%	16.0%	41.0%	8.0%	0%
$T = 20$	0.0	0%	0%	0%	0%	0%	0%
	0.5	0%	0%	0%	0%	0%	0%
	0.9	25.6%	0.4%	0%	2.0%	0%	0%
		$N = 100$			$N = 1000$		

Table 2: Share of corner solutions ($\hat{\sigma}^2_\alpha = 0$). 10000 Monte Carlo replications

Corner solutions occur more often when ρ is large and T is small. Few cases occur also when the "true" autoregression parameter is zero (but only if T, N and σ^2_α are small). Many cases occur when the autoregression parameter is large; some cases can be found even with the combination of large T and large N, (but only when σ^2_α is small).

The typical histograms of the three parameters will show: a probability "mass" at zero, for $\hat{\sigma}^2_\alpha$ and a partial histogram around the "true" value; a bimodal distribution for each of the other two parameters. As an example, we report the Monte Carlo distribution of parameters for two different set ups (Figures 1 and 2). Inference would really be problematic!

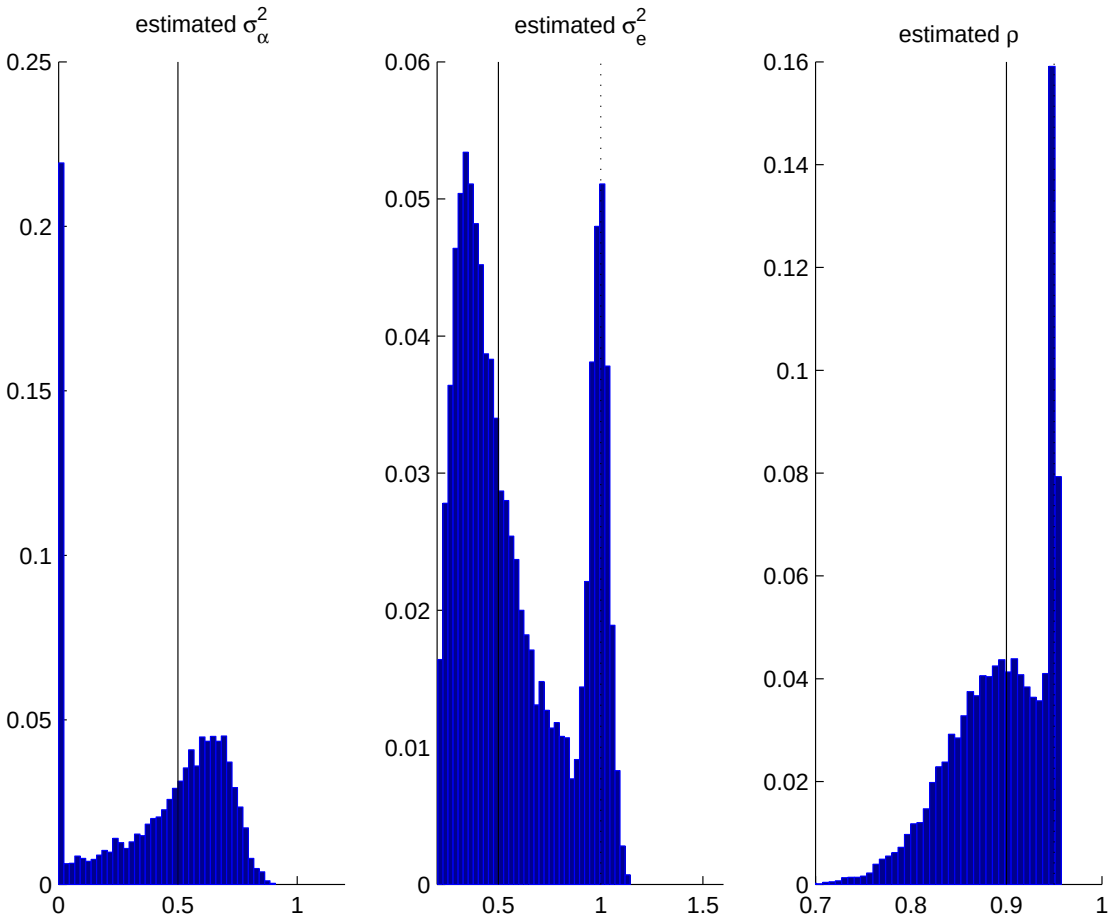

Figure 1: Monte Carlo distribution of estimated parameters: $T = 3$, $N = 1000$, $\sigma^2_\alpha = \sigma^2_e = 0.5$, and $\rho = 0.9$

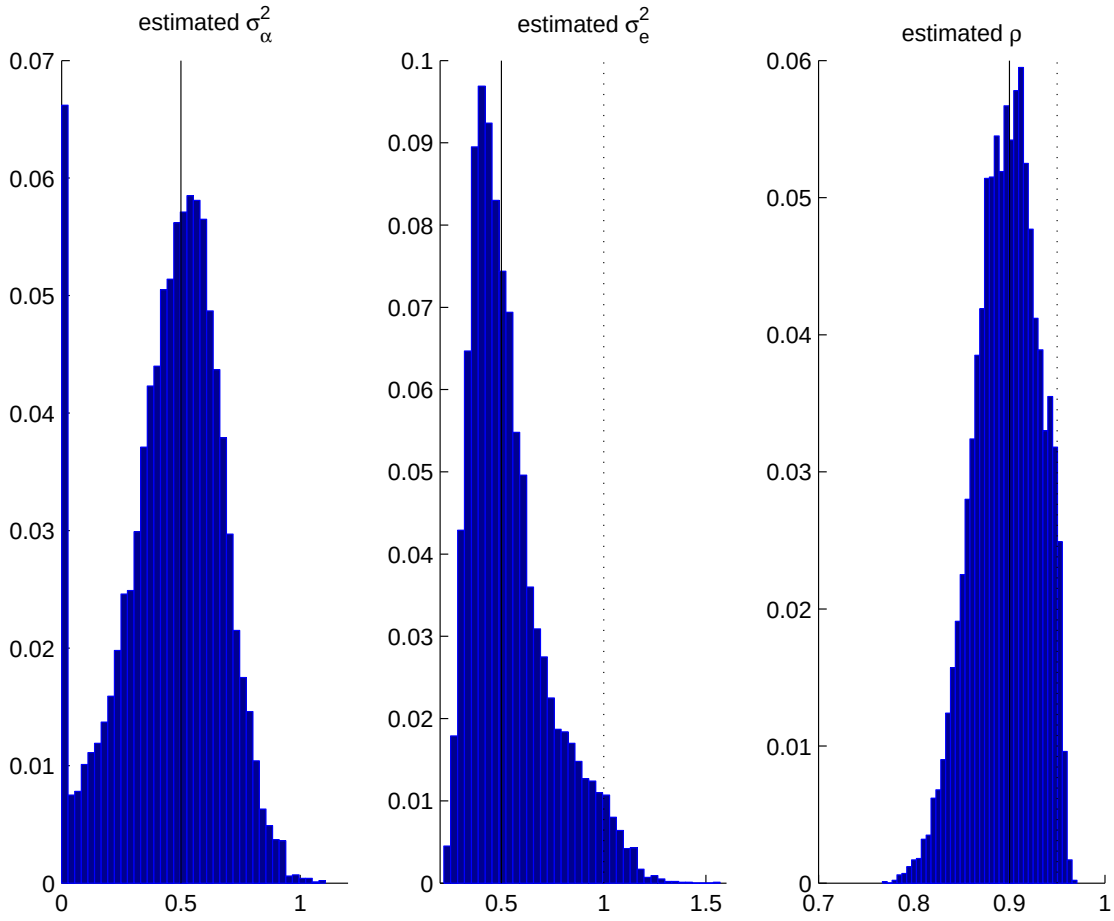

Figure 2: Monte Carlo distribution of estimated parameters: $T = 10$, $N = 100$, $\sigma_\alpha^2 = \sigma_e^2 = 0.5$, and $\rho = 0.9$

7 Summary

This paper has shown that a problem of *poor* identifiability arises in the case of Gaussian likelihood function in the context of random effect panel data model with autocorrelated disturbances. The standard random effect model is not apt at describing situations where the correlation of residuals for the same unit over time decreases as the time lag increases. This is likely the case when studying economic relationship, where the effect of transitory variables typically lasts more than one period. Ignoring correlation leads to estimates of the regression coefficient that are consistent but inefficient, and estimated standard errors are biased. In order to take into account this phenomenon, the disturbance can be assumed to follow an AR or MA process, where the difference is in the persistency of the unobserved shocks over time.
We show the unpleasant behavior of the Gaussian maximum likelihood estimates for a linear panel data model with random effects and AR(1) idiosyncratic noise. By means of a wide Monte Carlo experiment, we show that corner solutions (zero estimate of the variance of the individual random effect) are encountered in a large fraction of the experiments. When this is the case, the other parameters of the model shift, and the distributions are no more scattered only around their "true" values, rather they are bi-modal. Unless the sample involves an (unreasonable) enormous number of units, the problem is relevant for the time dimensions usually encountered in panel data analysis. Even if the probability of a corner solution decreases as the time dimension enlarges, the problem is quite relevant for the N and T dimensions used in practical applications.

References

Baltagi, B. and Li, Q.: 1995, Testing AR(1) against MA(1) Disturbances in an Error Component Model, *Journal of Econometrics* **68**, 133–151.

Berry, S., Gottschalk, P. and Wissoker, D.: 1988, An Error Components Model of the Impact of Plant Closing on Earnings., *Review of Economics & Statistics* **70**(4), 701–07.

Bhargava, A., Franzini, L. and Narendranathan, W.: 1982, Serial Correlation and the Fixed Effects Model, *The Review of Economic Studies* **49**(4), 533–549.

David, M.: 1971, Lifetime Income Variability and Income Profiles, *Proceedings of the Annual Meeting of the American Statistical Association*, pp. 285–92.

Hause, J.: 1973, The Covariance Structure of Earnings and the On the Job Training Hypothesis, *NBER Working Paper* .

Karlsson, S. and Skoglund, J.: 2004, Maximum-likelihood based inference in the two-way random effects model with serially correlated time effects, *Empirical Economics* **29**(1), 79–88.

Kiefer, N.: 1980, Estimation of Fixed Effect Models for Time Series of Cross-Sections with Arbitrary Intertemporal Covariances, *Journal of Econometrics* **14**, 195–202.

Lillard, L. and Weiss, Y.: 1979, Components of Variation in Panel Earnings Data: American Scientists 1960-1970, *Econometrica* **47**(2), 437–54.

Lillard, L. and Willis, R.: 1978, Dynamic Aspects of Earning Mobility, *Econometrica* **46**(5), 985–1012.

MaCurdy, T.: 1982, The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis, *Journal of Econometrics* **18**(1), 83–114.

Wooldridge, J.: 2002, *Econometric Analysis of Cross Section and Panel Data*, MIT Press.