Arellano, Manuel; Bonhomme, Stéphane

Working Paper

Identifying distributional characteristics in random coefficients panel data models

cemmap working paper, No. CWP22/09

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Arellano, Manuel; Bonhomme, Stéphane (2009) : Identifying distributional characteristics in random coefficients panel data models, cemmap working paper, No. CWP22/09, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2009.2209

This Version is available at:
https://hdl.handle.net/10419/64789

cemmap
centre for microdata methods and practice

Identifying distributional characteristics in random coefficients panel data models

Manuel Arellano
Stéphane Bonhomme

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP22/09

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Identifying Distributional Characteristics in Random Coefficients Panel Data Models*

Manuel Arellano
CEMFI, Madrid

Stéphane Bonhomme
CEMFI, Madrid

This draft: 29 July 2009

Abstract

We study the identification of panel models with linear individual-specific coefficients, when T is fixed. We show identification of the variance of the effects under conditional uncorrelatedness. Identification requires restricted dependence of errors, reflecting a trade-off between heterogeneity and error dynamics. We show identification of the density of individual effects when errors follow an ARMA process under conditional independence. We discuss GMM estimation of moments of effects and errors, and introduce a simple density estimator of a slope effect in a special case. As an application we estimate the effect that a mother smokes during pregnancy on child's birth weight.

JEL CODE: C23.
KEYWORDS: Panel data, random coefficients, multiple effects, nonparametric identification.

*We thank Cristian Bartolucci, Bryan Graham, and Jean-Marc Robin for useful comments. All remaining errors are our own. Research funding from the Spanish Ministry of Science and Innovation, Grant ECO 2008-00280 is gratefully acknowledged.

1 Introduction

Fixed effects methods are a standard way of controlling for endogeneity and/or unobserved heterogeneity in the estimation of common parameters from panel data models. However, sometimes one is willing to treat a model parameter as a heterogeneous quantity (as a "fixed effect") and therefore characteristics of its distribution or the density itself become central objects of interest in estimation.

In a static panel model that is nonlinear in common parameters but linear in random coefficients, the expected value of the random coefficients is fixed-T identified under the assumptions of unrestricted intertemporal distribution of the errors and unrestricted distribution of the effects conditioned on the regressors (Chamberlain, 1992). However, variances and covariances of random coefficients as well as other distributional characteristics are not identified. The reason is that by permitting arbitrary forms of dependence among the errors at all lags, it becomes impossible to separate out what part of the overall time variation is due to unobserved heterogeneity, no matter how long the panel is.

The point of departure of this paper is to consider the identifying content of limited time dependence of time-varying errors. The idea is that we may expect a stronger association between errors that are close to each other than errors that are far apart in time. Moving average and autoregressive processes are convenient implementations of this notion. Subject to limited time series error dependence, alternative identification arrangements become available. In particular variances, higher order moments and densities of random coefficients may be identifiable. We explore such identification trade-offs and provide conditions under which different distributional characteristics are identified. Throughout we adopt a "fixed effects approach" in the sense that the conditional distribution of the random coefficients given explanatory variables is left unrestricted.

A linear random coefficient model is a useful framework of analysis in many microeconometric applications. These include earning dynamics models with individual-specific age profiles and persistent shocks,[1] as well as production function models with firm-specific technological parameters.[2] The estimation of heterogeneous treatment effects is another

[1]For examples of earnings models with individual-specific slopes or profiles, see Lillard and Weiss (1979), Baker (1997), Haider (2001), and Guvenen (2007, 2009).

[2]See for example Mairesse and Grilliches (1990) and Dobbelaere and Mairesse (2008). Other examples can be found in the literature on the education production function and teacher quality (e.g., Aaronson *et al.*, 2007).

area of application. In contrast with the cross-sectional case, panel data on repeated treatments offer the opportunity to estimate a time-invariant distribution of treatment effects across units.[3] For example, in our empirical application, we look at the extent of heterogeneity in the effect of smoking during pregnancy on children outcomes at birth, building on Abrevaya (2006)'s results for mothers with multiple births. There is interest in documenting the determinants of inequality at birth, particularly in relation to policy interventions (e.g. Rosenzweig and Wolpin, 1991) and accounting for heterogeneity in the effects of those determinants is certainly important.

Most statistical approaches to random coefficient models have adopted a random effects perspective, which rules out or restricts the correlation between individual-specific effects and regressors.[4] In economic applications, though, unit-specific effects often represent heterogeneity in preferences or technology, on which economic theory has typically little to say. For this reason, it is often thought (as we do here) that a fixed effects approach, which does not restrict the form of the heterogeneity is preferable.[5] Thus, we regard individual specific parameters as random draws from an unrestricted conditional distribution given regressors.

In an important paper, Chamberlain (1992) derived efficiency bounds for conditional moment restrictions with a nonparametric component, and applied the results to a random coefficient model for panel data. In that model the role of the nonparametric component was played by the conditional expectation of the random coefficients given the regressors. Chamberlain suggested an instrumental-variable estimator of the common parameters and average effects, which attained the bound.

Chamberlain (1992) assumed that time-varying errors were mean independent of individual effects and regressors at all lags and leads (a strict exogeneity assumption). Extending the approach, we consider a similar model with the additional assumption that the autocovariance matrix of the errors conditioned on regressors satisfy moving-average (MA) exclusion restrictions. Non-zero autocovariances are treated as nonparametric functions of regressors. Therefore, they are consistent with an underlying moving average model with unobserved

[3] In a cross-sectional setting only the marginal distributions of potential outcomes may be identified under standard assumptions, to the exclusion of the distribution of gains from treatment (Heckman *et al.*, 1997).

[4] See Demidenko (2004) for a survey on random-effects (or "mixed") models in statistics. Recent work using semi- and nonparametric approaches can be found in Lesaffre and Verbeke (1996), Kleinman and Ibrahim (1998), and Davidian and Zhang (2001).

[5] For example, Cameron and Trivedi (2005, p.777) claim that random coefficient models, although they "are especially popular in the statistics literature (...) are less used in the econometrics literature, because of the reluctance to impose structure on the time-invariant individual-specific fixed effect".

heterogeneity in second-order moments. In this setting, conditional and unconditional variances of effects and errors are point identified, as long as sufficiently many autocovariance restrictions are imposed. For example, identification will require that the order of an MA process be small enough. We also discuss how the results can be generalized to ARMA-type restrictions.

Moreover, we show how Chamberlain's analysis can be extended to obtain a semiparametric efficiency bound for all common parameters and first and second moments of the random coefficients. The result holds for a parametric specification of the error second moments conditioned on regressors and effects, which is either linear in or independent of the effects. We also show how fixed-T consistent and asymptotically normal estimates of these coefficients can be obtained using a system GMM procedure that combines errors in levels with errors in (generalized) deviations. The bound provides guidance on the choice of optimal instruments.

Next, strengthening the mean independence assumption to one of conditional statistical independence between effects and errors given regressors, we study the identification of higher-order moments and distributions. When time-varying errors follow suitably restricted ARMA processes with independent underlying innovations, we obtain fixed-T point identification results for the densities of individual effects and errors. To obtain these results, we first use that (cumulant) independence assumptions lead to higher-order moment restrictions that mimic covariance restrictions. Then we exploit the fact that (statistical) independence assumptions lead to functional restrictions on the second derivatives of log characteristic functions, which are formally analogous to the covariance restrictions. We show that these restrictions nicely extend those for second and higher-order moments, and may be used to establish the identification of distributions.

Our identification proofs are constructive. Thus, they suggest consistent estimators for the distributional quantities of interest. We construct consistent method-of-moment estimators of variances and higher-order moments. We also discuss ways of estimating the densities of individual effects and errors, emphasizing the connection with the literature on nonparametric deconvolution (see Carroll and Hall, 1988, among many other references). As an interesting special case, we consider a model with an heterogeneous intercept and a binary heterogeneous regressor. This corresponds to our empirical application where the smoking effect is heterogeneous across mothers. In this setting, we propose a simple nonparametric

estimator of the density of the mother-specific smoking effect.

In the last section of the paper we apply this methodology to a matched panel dataset of mothers and births constructed in Abrevaya (2006). We find that the mean smoking effect on birthweight is significantly negative (-160 grams). Moreover, the effect shows substantial heterogeneity across mothers, the effect being very negative (-400 g) below the 20th percentile. In addition, we discuss the validity of the strict exogeneity assumption in the context of this application. Although the mean effect is not point identified in this setting,[6] we show that several interesting average effects can be identified and estimated when there are no time-varying regressors. The results suggest that the smoking effect is strongly correlated with smoking choices, justifying the fixed-effects perspective. Moreover, we do not find strong evidence against strict exogeneity on these data.

This paper is related to the literature on the estimation of linear and nonlinear panel data models with fixed effects. A general solution has recently been proposed that relies on reduction of the small-T bias of the maximum likelihood estimator first documented in Neyman and Scott (1948), see Arellano and Hahn (2006) for a survey. Here we show that all marginal effects, including the density of individual-specific effects, are identified for fixed T in a model that is linear in random coefficients. Hence, our approach leads to full elimination of the bias on the quantities of interest.

Related identification strategies for densities have been used in the literature on nonparametric identification and estimation of linear factor models with independent factors. See for example Horowitz and Markatou (1996), Székely and Rao (2000), and Bonhomme and Robin (2009b). We contribute to that literature by allowing for correlation patterns that may be natural in applications, individual effects being correlated in an unrestricted way, and errors being possibly serially correlated. We also allow for conditioning covariates.

The rest of the paper is as follows. In Section 2 we present the framework of analysis. Section 3 derives the identifying restrictions on the variances of individual effects and errors. In Section 4, we extend the analysis to the full distributions of effects and errors. We discuss estimation in Section 5, and apply our methodology in Section 6 to study the effect of smoking during pregnancy on birth outcomes. Lastly, Section 7 concludes.

[6] Chamberlain (1993) and Arellano and Honoré (2001) discuss the lack of identification when regressors are predetermined. Recently, Murtazashvili and Wooldridge (2008) derive conditions under which identification holds in the endogenous case, imposing individual effects to be mean independent of detrended regressors (see also Wooldridge, 2005, for the exogenous case).

2 Preliminaries

2.1 Model and assumptions

We consider a model that relates a vector of T endogenous variables $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$ to a set of regressors $\mathbf{W}_i = [\mathbf{Z}_i, \mathbf{X}_i]$ and a vector of zero-mean error terms $\mathbf{v}_i = (v_{i1}...v_{iT})'$:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\delta} + \mathbf{X}_i\boldsymbol{\gamma}_i + \mathbf{v}_i \quad (i = 1...N)\,. \tag{1}$$

We distinguish two types of regressors: $\mathbf{Z}_i = (\mathbf{z}'_{i1}, ..., \mathbf{z}'_{iT})'$ is a $T \times K$ matrix associated to a vector of K common parameters $\boldsymbol{\delta}$, while $\mathbf{X}_i = (\mathbf{x}'_{i1}, ..., \mathbf{x}'_{iT})'$ is a $T \times q$ matrix associated to a vector of q *unit specific* parameters $\boldsymbol{\gamma}_i$. We start by stating the assumptions.

Assumption 1 ***(mean independence)***

$$\mathbf{E}\left(\mathbf{v}_i \mid \mathbf{W}_i, \boldsymbol{\gamma}_i\right) = \mathbf{0}. \tag{2}$$

Assumption 1 requires $\mathbf{Z}_i$ and $\mathbf{X}_i$ to be *strictly exogenous*.[7] It is possible to treat the case of predetermined or endogenous $\mathbf{Z}_i$'s within the framework of this paper, and we discuss this extension below. However, strict exogeneity of $\mathbf{X}_i$ is essential. If one of the components of $\mathbf{x}_{it}$ is predetermined or endogenous, then the moments of $\boldsymbol{\gamma}_i$ are not point identified in general.

Note that we do not specify the conditional distribution of individual effects. In our "fixed-effect" approach, $\boldsymbol{\gamma}_i$ are random draws from a population, along with y_{it}, $\boldsymbol{z}_{it}$ and $\boldsymbol{x}_{it}$, but their conditional distribution given regressors is left unspecified. Thus, regressors are strictly exogenous with respect to to time-varying errors but endogenous with respect to fixed effects. We will discuss the validity of this assumption in the context of our empirical application in Section 6.

Mean independence will be used to identify the vector of common parameters $\boldsymbol{\delta}$ and the means, variances and covariances of individual-specific parameters $\boldsymbol{\gamma}_i$. When studying the identification of higher-order moments of the effects and their distributions, we will need a stronger assumption.

Assumption 2 ***(conditional statistical independence)***

$$\boldsymbol{\gamma}_i \textit{ and } \mathbf{v}_i \textit{ are statistically independent given } \mathbf{W}_i. \tag{3}$$

[7] Throughout the paper, all (in)equalities conditional on $\mathbf{W}_i$ are understood to hold with probability one. In addition, moments are assumed well-defined (i.e., finite).

Conditional independence restrictions are commonly made in the literature on nonparametric identification and estimation (e.g., Hu and Schennach, 2008, and references therein). Moreover, restriction (3) is in the nature of a fixed-effects approach, where $\boldsymbol{\gamma}_i$ represent individual-specific parameters such as preferences or technology. However, note that Assumption 2 is more restrictive than Assumption 1 as, for example, it rules out the presence of individual effects in the conditional variance of $\mathbf{v}_i$.

Lastly, we will also assume that regressors $\mathbf{X}_i$ are not perfectly collinear *within* each individual sequence of observations.

Assumption 3 ***(absence of multicollinearity)***

$$\text{rank}(\mathbf{X}_i) = q. \tag{4}$$

In particular, Assumption 3 requires that $T \geq q$. This condition is necessary in our approach, as one needs to identify q parameters from a T-dimensional vector of data, for each individual unit. In effect, because of the presence of common parameters, we will need *strictly more* time periods than individual-specific parameters. This requirement shows that the panel dimension is essential in our setting.[8]

Assumption 3 is restrictive as it implies that, when $\mathbf{X}_i$ takes discrete values, the moments of individual effects will be identified on a subpopulation of individuals only. For example, in our empirical application, we will focus on mothers who changed smoking status at least once between births. A related model with continuous $\mathbf{X}_i$'s has been recently studied by Graham and Powell (2008). Their analysis suggests that average effects for the total population of individuals, including individuals for whom (4) is not satisfied, may be consistently estimated using nonparametric methods with trimming.

2.2 Within and between transformations

To motivate our identification analysis, we start by providing an intuition for our approach. Given a vector of common parameters $\boldsymbol{\delta}$, one can estimate each $\boldsymbol{\gamma}_i$ by least squares, yielding:

$$\widehat{\boldsymbol{\gamma}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right).$$

[8] This is very different from situations where restrictions on $\boldsymbol{\gamma}_i$ are imposed, such as independence between $\boldsymbol{\gamma}_i$ and regressors $\mathbf{X}_i$. There, cross-sectional data may be enough for identification (see, e.g., Beran and Hall, 1992, and Hoderlein *et al.*, 2007).

Let us introduce the two following matrices:

$$\begin{aligned}\mathbf{Q}_i &= \mathbf{I}_T - \mathbf{X}_i \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i', \\ \mathbf{H}_i &= \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \mathbf{X}_i'.\end{aligned}$$

$\mathbf{Q}_i$ $(T \times T)$ is the projection matrix on the orthogonal of the span of the columns of $\mathbf{X}_i$. $\mathbf{Q}_i$ is a familar object in least squares algebra, and is symmetric idempotent with rank $T - q$. $\mathbf{H}_i$ $(q \times T)$ is simply the least squares operator associated with $\mathbf{X}_i$.

Left-multiplying (1) by $\mathbf{Q}_i$ and $\mathbf{H}_i$, respectively, we obtain the following equations:

$$\mathbf{Q}_i \left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right) = \mathbf{Q}_i\mathbf{v}_i \qquad \text{(within-group)}, \tag{5}$$

$$\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i = \mathbf{H}_i\mathbf{v}_i \qquad \text{(between-group)}. \tag{6}$$

While equation (6) expresses the difference between the least-squares estimate of $\boldsymbol{\gamma}_i$ (for known $\boldsymbol{\delta}$) and its true value, equation (5) shows the link between the residuals in the individual-specific least-squares regressions and the population errors. We will start from these equations to study the identification of common parameters, the error structure, and the distribution of individual effects.

Two preliminary remarks are in order. First, since $\mathbf{Q}_i$ has rank $T - q$, it is not possible to invert (5) unless some additional restrictions on the time-series process of errors v_{it} are imposed. Second, equation (6) shows that $\widehat{\boldsymbol{\gamma}}_i$ is a noisy estimate of $\boldsymbol{\gamma}_i$. Likewise, any distributional characteristic of $\widehat{\boldsymbol{\gamma}}_i$ (mean, variance, quantile) will be a noisy estimate of the same feature of $\boldsymbol{\gamma}_i$, the identification of which we are after. Importantly, this noise does not vanish when N tends to infinity for fixed T, so unit-by-unit estimates of $\boldsymbol{\gamma}_i$ are not directly informative about the distribution of the underlying effects.

We end this subsection by presenting two simple examples.

Example 1. The first example is a random trend model:

$$y_{it} = \alpha_i + \beta_i t + v_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{7}$$

where v_{it} are serially correlated, for example through an $AR(1)$ process.

Model (7), or a restricted version of it (e.g., with $\beta_i = 0$), is often used to model the dynamics of earnings (see Guvenen, 2009, for a recent reference). In this model the between-

group equations (6) are:

$$\widehat{\alpha}_i = \alpha_i + \overline{v}_i - \frac{\sum_{t=1}^{T}\left(t-\overline{t}\right)\left(v_{it}-\overline{v}_i\right)}{\sum_{t=1}^{T}\left(t-\overline{t}\right)^2}\overline{t} \tag{8}$$

$$\widehat{\beta}_i = \beta_i + \frac{\sum_{t=1}^{T}\left(t-\overline{t}\right)\left(v_{it}-\overline{v}_i\right)}{\sum_{t=1}^{T}\left(t-\overline{t}\right)^2}, \tag{9}$$

whereas the within-group equations (5) are:

$$y_{it} - \widehat{\alpha}_i - \widehat{\beta}_i t = v_{it} - \left(\overline{v}_i - \frac{\sum_{s=1}^{T}\left(s-\overline{t}\right)\left(v_{is}-\overline{v}_i\right)}{\sum_{s=1}^{T}\left(s-\overline{t}\right)^2}\overline{t}\right) - \left(\frac{\sum_{s=1}^{T}\left(s-\overline{t}\right)\left(v_{is}-\overline{v}_i\right)}{\sum_{s=1}^{T}\left(s-\overline{t}\right)^2}\right)t. \tag{10}$$

Example 2. The second example is a model with a binary regressor $s_{ij} \in \{0,1\}$:

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + v_{i\ell}, \quad i = 1,...,N, \quad \ell = 1,...,L. \tag{11}$$

This is the model we use in our empirical application, where $s_{i\ell}$ denotes the smoking status of mother i during the pregnancy of child ℓ, and $y_{i\ell}$ is the birthweight of child ℓ.

Denoting as $n_i = \sum_{\ell=1}^{L} s_{i\ell}$, we obtain:

$$\widehat{\alpha}_i = \alpha_i + \frac{1}{L-n_i}\sum_{\ell=1}^{L}\left(1-s_{i\ell}\right)v_{i\ell} \tag{12}$$

$$\widehat{\beta}_i = \beta_i + \frac{1}{n_i}\sum_{\ell=1}^{L} s_{i\ell}v_{i\ell} - \frac{1}{L-n_i}\sum_{\ell=1}^{L}\left(1-s_{i\ell}\right)v_{i\ell}, \tag{13}$$

and:

$$y_{i\ell} - \widehat{\alpha}_i - \widehat{\beta}_i s_{i\ell} = v_{i\ell} - s_{i\ell}\frac{1}{n_i}\sum_{k=1}^{L} s_{ik}v_{ik} - \left(1-s_{i\ell}\right)\frac{1}{L-n_i}\sum_{k=1}^{L}\left(1-s_{ik}\right)v_{ik}. \tag{14}$$

2.3 Extensions

Nonlinearity in variables and common parameters. Although we discuss identification of the linear model (1), the approach of this paper can be generalized to other settings. A more general formulation is:

$$\mathbf{y}_i = \mathbf{a}(\mathbf{W}_i;\boldsymbol{\theta}) + \mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})\boldsymbol{\gamma}_i + \mathbf{v}_i, \tag{15}$$

where $\boldsymbol{\theta}$ is a vector of common parameters that enter nonlinearly functions $\mathbf{a}$ (which is $T\times 1$) and $\mathbf{B}$ ($T \times q$).

Following Chamberlain (1992), one can consider the generalized within- and between-group equations:

$$\mathbf{Q}_i(\boldsymbol{\theta})(\mathbf{y}_i - \mathbf{a}(\mathbf{W}_i;\boldsymbol{\theta})) = \mathbf{Q}_i(\boldsymbol{\theta})\mathbf{v}_i \qquad \text{(within-group)}, \tag{16}$$

$$\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i = \mathbf{H}_i(\boldsymbol{\theta})\mathbf{v}_i \qquad \text{(between-group)}, \tag{17}$$

where

$$\mathbf{Q}_i(\boldsymbol{\theta}) = \mathbf{I}_T - \mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})\left[\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})'\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})\right]^{-1}\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})', \tag{18}$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = \left(\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})'\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})\right)^{-1}\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})'. \tag{19}$$

$\mathbf{Q}_i(\boldsymbol{\theta})$ and $\mathbf{H}_i(\boldsymbol{\theta})$ are well-defined provided that: $\text{rank}\left[\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})\right] = q$. Because of the within- and between-group equations (16) and (17), the identification analysis of model (15) follows very closely that of the linear model (1). We will indicate the differences in the course of the exposition.

It is instructive to consider examples of model (15). A simple special case is the one-factor model:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \mu_t\alpha_i + v_{it}, \tag{20}$$

where $\mu_1, ..., \mu_T$ are time-varying parameters and α_i is scalar (e.g., Holtz-Eakin *et al.*, 1988). In a wage regression, α_i could be workers' unobserved skills on the labor market, and μ_t their time-varying price. Multiple-equation versions of (20), where $\mathbf{y}_{it}$ is multi-dimensional, could also be considered. Moreover, the model can be generalized to allow for time-varying unobservable individual effects which follow a factor structure (Bai, 2009, Ahn *et al.*, 2007).

Other interesting special cases of (15) are models where the regressors include lags (or leads) of the dependent variable. For example, a first-order autoregressive model:

$$y_{it} = \delta y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\gamma}_i + v_{it}, \quad |\delta| < 1. \tag{21}$$

That (21) is a special case of (15) is seen by writing the reduced-form:

$$y_{it} = \left(\mathbf{x}_{it} + \delta\mathbf{x}_{i,t-1} + ... + \delta^{t-1}\mathbf{x}_{i1}\right)'\gamma_i + \delta^t y_{i0} + v_{it} + \delta v_{i,t-1} + ... + \delta^{t-1}v_{i1},$$

which is of the form (15) with the $(q+1)\times 1$ vector of individual effects: $\widetilde{\boldsymbol{\gamma}}_i = (\boldsymbol{\gamma}_i', y_{i0})'$.

General predetermined variables. Assumption 1 posits the strict exogeneity of $\mathbf{Z}_i$ and $\mathbf{X}_i$ given $\boldsymbol{\gamma}_i$. The critical role of this assumption is to ensure that within and between

errors, $\mathbf{Q}_i\mathbf{v}_i$ and $\mathbf{H}_i\mathbf{v}_i$, have zero conditional mean given all lags and leads of the regressors. However, our approach can be generalized to situations where $\mathbf{Z}_i$ includes predetermined or endogenous variables (although the remainder of the paper assumes strict exogeneity for simplicity). The idea is to replace Assumption 1 for the error $v_{it} = y_{it} - \mathbf{z}_{it}'\boldsymbol{\delta} - \mathbf{x}_{it}'\boldsymbol{\gamma}_i$ with the following generalization:

$$\mathbf{E}\left(v_{it} \mid \mathbf{r}_{i1}, ..., \mathbf{r}_{it}, \mathbf{X}_i, \boldsymbol{\gamma}_i\right) = 0 \quad (t = 1, ..., T)\,, \tag{22}$$

where $\mathbf{r}_{it}$ is a predetermined instrumental variable, which may be external to the model or not. For example, if $\mathbf{r}_{it} = \mathbf{z}_{it}$ the explanatory variable $\mathbf{z}_{it}$ itself is predetermined; if $\mathbf{r}_{it} = \mathbf{z}_{it-1}$ then $\mathbf{z}_{it}$ is contemporaneously endogenous but its lags are predetermined, whereas if $\mathbf{r}_{it}$ is an external instrument $\mathbf{z}_{it}$ is treated as endogenous at all lags.

Contrary to Assumption 1, the orthogonality between original errors and conditioning variables in the new assumption is not transmitted to ordinary within errors. The reason is that (22) implies a pattern of sequential orthogonality and each within error depends on the full time series of original errors. However, there is an alternative within transformation that preserves sequential orthogonality, which is provided by a generalization of forward orthogonal deviations (Arellano and Bover, 1995). Let $\mathbf{A}_i$ be a $(T-q)\times T$ upper triangular decomposition of $\mathbf{Q}_i$ such that $\mathbf{A}_i'\mathbf{A}_i = \mathbf{Q}_i$ and $\mathbf{A}_i\mathbf{A}_i' = \mathbf{I}_{T-q}$. The orthogonal within errors $\mathbf{A}_i\mathbf{v}_i \equiv \left(v_{i1}^*, ..., v_{i(T-q)}^*\right)'$ satisfy assumption (22):

$$\mathbf{E}\left(v_{it}^* \mid \mathbf{r}_{i1}, ..., \mathbf{r}_{it}, \mathbf{X}_i, \boldsymbol{\gamma}_i\right) = 0 \quad (t = 1, ..., T-q)\,.$$

Strict exogeneity of $\mathbf{X}_i$ is an essential ingredient of the previous argument, but as long as this is preserved, nonlinear extensions are also possible. For example, it is possible to consider assumption (22) in conjunction with a model of the form

$$\mathbf{a}\left(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}\right) = \mathbf{B}\left(\mathbf{X}_i, \boldsymbol{\theta}\right)\boldsymbol{\gamma}_i + \mathbf{v}_i,$$

where the columns of $\mathbf{Y}_i$ contain endogenous and predetermined variables.

3 Identification of first and second moments

In this section we study the identification of common parameters, and means and variances of individual effects and errors.

3.1 Common parameters and averages of individual effects

We start with a proposition which shows the identification of $\boldsymbol{\delta}$ and $\mathbf{E}(\boldsymbol{\gamma}_i)$. All proofs are in Appendix A.

Proposition 1 ***(common parameters and mean effects)***

Let Assumptions 1 and 3 hold. Then:

$$\mathbf{E}\left(\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right) = \mathbf{0} \tag{23}$$

and

$$\mathbf{E}\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right). \tag{24}$$

So $\mathbf{E}(\boldsymbol{\gamma}_i)$ *is identified. Moreover,* $\boldsymbol{\delta}$ *is identified if* $\mathbf{E}(\mathbf{Z}_i'\mathbf{Q}_i\mathbf{Z}_i)$ *has rank* K*, the number of common parameters.*

Proposition 1 shows that $\boldsymbol{\delta}$ can be interpreted as a generalized within-group estimand. Similarly, $\mathbf{E}(\boldsymbol{\gamma}_i)$ can be understood as a *mean-group* estimand. For example, consider a model with a heterogeneous intercept:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \gamma_{i1} + v_{it}. \tag{25}$$

Then, $\boldsymbol{\delta}$ and $\mathbf{E}(\boldsymbol{\gamma}_i)$ satisfy:

$$\begin{aligned}\mathbf{E}\left(y_{it} - \overline{y}_i - \left(\mathbf{z}_{it} - \overline{\mathbf{z}}_i\right)'\boldsymbol{\delta}|\mathbf{Z}_i\right) &= 0,\\ \mathbf{E}\left(\gamma_{i1}\right) &= \mathbf{E}\left(\overline{y}_i - \overline{\mathbf{z}}_i'\boldsymbol{\delta}\right).\end{aligned}$$

Applied researchers often find it useful to regress individual effects estimates $\widehat{\boldsymbol{\gamma}}_i$ on strictly exogenous regressors $\mathbf{F}_i$, see MaCurdy (1981) for an early application. An interesting corollary of Proposition 1 is that the population projection coefficients in the regression of $\widehat{\boldsymbol{\gamma}}_i$ on $\mathbf{F}_i$ are equal to the projection coefficients in the regression of $\boldsymbol{\gamma}_i$ on $\mathbf{F}_i$.

Corollary 1 ***(projection coefficients)***

Let Assumptions 1 and 3 hold. Let also $\mathbf{F}_i$ *be a random vector such that* $\mathbf{E}(v_{it}|\mathbf{W}_i, \mathbf{F}_i) = 0$*. Then:*

$$\left[\mathbf{Var}\left(\mathbf{F}_i\right)\right]^{-1}\mathbf{Cov}\left(\mathbf{F}_i, \boldsymbol{\gamma}_i\right) \;=\; \left[\mathbf{Var}\left(\mathbf{F}_i\right)\right]^{-1}\mathbf{Cov}\left(\mathbf{F}_i, \widehat{\boldsymbol{\gamma}}_i\right). \tag{26}$$

Similar results can be obtained for the more general formulation (15). The next corollary derives moment conditions satified by common parameters $\boldsymbol{\theta}$.

Corollary 2 ***(Chamberlain's model)***

Consider model (15), and suppose that $\mathbf{E}(\mathbf{v}_i|\mathbf{W}_i) = \mathbf{0}$ *and that matrix* $\mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})$ *has rank* q*. Then:*

$$\mathbf{E}\left[\mathbf{Q}_i(\boldsymbol{\theta})\left(\mathbf{y}_i - \mathbf{a}(\mathbf{W}_i;\boldsymbol{\theta})\right)|\mathbf{W}_i\right] = \mathbf{0}, \tag{27}$$

and

$$\mathbf{E}\left[\mathbf{H}_i(\boldsymbol{\theta})\left(\mathbf{y}_i - \mathbf{a}(\mathbf{W}_i;\boldsymbol{\theta})\right)|\mathbf{W}_i\right] = \mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right), \tag{28}$$

where $\mathbf{Q}_i(\boldsymbol{\theta})$ *and* $\mathbf{H}_i(\boldsymbol{\theta})$ *are given by (18) and (19), respectively.*

Corollary 2 provides conditional moment restrictions that may or may not be sufficient to identify $\boldsymbol{\theta}$. For example, consider an AR(1) model with fixed effects without strictly exogenous regressors:

$$y_{it} = \delta y_{i,t-1} + \gamma_{i1} + v_{it}, \quad |\delta| < 1. \tag{29}$$

In the absence of restrictions on the v_{it} process, δ is not identified in model (29). Identification may be achieved by restricting the variance-covariance matrix of $\mathbf{v}_i$ and by exploiting covariance restrictions (Holtz-Eakin *et al.*, 1988, Arellano and Bond, 1991). We will study the identification content of covariance restrictions in the next subsection.

Remark that, once $\boldsymbol{\theta}$ is identified, there is no essential difference between model (1) and model (15). Indeed, one can define $\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{a}(\mathbf{W}_i;\boldsymbol{\theta})$ as the new dependent variable and $\widetilde{\mathbf{X}}_i = \mathbf{B}(\mathbf{W}_i;\boldsymbol{\theta})$ as the new set of regressors, and use the identification results obtained for model (1).

Information bound on common parameters and average effects. Chamberlain (1992) obtained the optimal moment conditions of common parameters and average effects for model (15). The moments are optimal in the sense that an estimator based on them attains the semiparametric information bound.

Following the argument developed in Appendix C, the joint optimal moments for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i)$ can be expressed as

$$\mathbf{E}\begin{pmatrix} \left\{\frac{\partial}{\partial\theta'}\left[\mathbf{a}_i + \mathbf{B}_i\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right)\right]\right\}' \mathbf{A}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) \\ \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{a}_i - \mathbf{B}_i\boldsymbol{\gamma}\right) \end{pmatrix} = \mathbf{0}, \tag{30}$$

where $\mathbf{a}_i = \mathbf{a}(\mathbf{W}_i, \boldsymbol{\theta})$, $\mathbf{B}_i = \mathbf{B}(\mathbf{W}_i, \boldsymbol{\theta})$, $\mathbf{V}_i = \mathbf{Var}(\mathbf{y}_i|\mathbf{W}_i)$, and $\mathbf{A}_i$ is a $(T-q) \times T$ orthogonal decomposition of $\mathbf{Q}_i(\boldsymbol{\theta})$.

3.2 Variances

Variances of individual effects. To recover the variance of individual effects, we impose restrictions on the variance-covariance matrix of errors $\mathbf{v}_i$. For exposition, we start with the case where $\boldsymbol{\Omega}_i = \mathbf{Var}(\mathbf{v}_i|\mathbf{W}_i)$ is known. The following theorem shows that the variance of individual effects is identified under those conditions. The proof is immediate using (6).

Theorem 1 ***(variances of effects)***

Let Assumptions 1 and 3 hold. Then:

$$\mathbf{Var}(\boldsymbol{\gamma}_i|\mathbf{W}_i) = \mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i) - \mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i' \tag{31}$$

and, unconditionally:

$$\mathbf{Var}(\boldsymbol{\gamma}_i) = \mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i) - \mathbf{E}(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'). \tag{32}$$

Theorem 1 shows that the variance-covariance matrix of individual effects is identified given that of errors. In the special case where $\boldsymbol{\Omega}_i = \sigma^2\mathbf{I}_T$, (32) yields:

$$\mathbf{Var}(\boldsymbol{\gamma}_i) = \mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i) - \sigma^2\mathbf{E}\left[(\mathbf{X}_i'\mathbf{X}_i)^{-1}\right]. \tag{33}$$

A familiar expression is obtained in model (25), with a single heterogeneous intercept and classical errors, in which case: $\text{Var}(\gamma_{i1}) = \text{Var}(\overline{y}_i - \overline{\mathbf{z}}_i'\boldsymbol{\delta}) - \sigma^2/T$.

It is instructive to write (32) as

$$\mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i) = \mathbf{Var}(\boldsymbol{\gamma}_i) + \mathbf{E}(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'), \tag{34}$$

which expresses the variance of individual effects estimates as the sum of a between-group and a within-group variance. The between-group term is equal to the variance of individual effects in the population.[9] The within-group variance generally tends to zero when T tends to infinity,[10] but is non zero for fixed T. This clearly decomposes the total variance of $\widehat{\boldsymbol{\gamma}}_i$ into two sources: the true cross-sectional variation of individual effects, and the noise due to T being fixed. It is important to note that the linearity of the model in the individual effects is essential for this result to hold.

[9] This is because $\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i|\boldsymbol{\gamma}_i) = \boldsymbol{\gamma}_i$, so: $\mathbf{Var}(\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i|\boldsymbol{\gamma}_i)) = \mathbf{Var}(\boldsymbol{\gamma}_i)$.

[10] This will be the case if, for any $k > 0$, $\mathbf{X}_i'\mathbf{X}_i/T^k$ and $\mathbf{X}_i'\boldsymbol{\Omega}_i\mathbf{X}_i/T^k$ tend in probability to non zero constants as T tends to infinity. In regular cases like Example 2, we can take $k = 1$. In Example 1, $\widehat{\boldsymbol{\gamma}}_i$ is superconsistent for $\boldsymbol{\gamma}_i$ as T tends to infinity, and we can take $k = 3$.

Variances of errors: MA restrictions. We now turn to the identification of $\mathbf{\Omega}_i$. We will contrast the identifying content of two types of restrictions. Covariance restrictions in *levels* are obtained when using the full variance-covariance matrix of $\mathbf{y}_i$, that is, using Assumption 1:

$$\begin{aligned}\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)'|\mathbf{W}_i\right] &= \mathbf{X}_i\mathbf{E}\left(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i'|\mathbf{W}_i\right)\mathbf{X}_i'+\mathbf{\Omega}_i \\ &\quad +\mathbf{X}_i\mathbf{E}\left(\boldsymbol{\gamma}_i\mathbf{v}_i'|\mathbf{W}_i\right)+\mathbf{E}\left(\mathbf{v}_i\boldsymbol{\gamma}_i'|\mathbf{W}_i\right)\mathbf{X}_i' \\ &= \mathbf{X}_i\mathbf{E}\left(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i'|\mathbf{W}_i\right)\mathbf{X}_i'+\mathbf{\Omega}_i. \end{aligned} \tag{35}$$

The *within* restrictions are obtained from the within-group equation (5), hence:

$$\mathbf{Q}_i\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)'|\mathbf{W}_i\right]\mathbf{Q}_i'=\mathbf{Q}_i\mathbf{\Omega}_i\mathbf{Q}_i'. \tag{36}$$

The within equations (36) are effectively a subset of the level equations (35). However, unlike (35), the within covariance restrictions (36) do not depend on errors v_{it} being mean independent of individual effects $\boldsymbol{\gamma}_i$.

We start by studying the identifying content of restrictions in levels. In vector form, (35) yields:

$$\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\otimes\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right]=\left(\mathbf{X}_i\otimes\mathbf{X}_i\right)\mathbf{E}\left(\boldsymbol{\gamma}_i\otimes\boldsymbol{\gamma}_i|\mathbf{W}_i\right)+\mathbf{vec}\left(\mathbf{\Omega}_i\right). \tag{37}$$

Note that the variance of individual effects is left unrestricted. Let us define the projection matrix on the orthogonal of $\mathbf{X}_i\otimes\mathbf{X}_i$:

$$\begin{aligned}\mathbf{M}_i &= \mathbf{I}_{T^2}-\left[\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\right]\otimes\left[\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\right] \\ &= \mathbf{I}_{T^2}-\left[\mathbf{I}_T-\mathbf{Q}_i\right]\otimes\left[\mathbf{I}_T-\mathbf{Q}_i\right].\end{aligned} \tag{38}$$

Left-multiplying (37) by $\mathbf{M}_i$ we obtain:

$$\mathbf{M}_i\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\otimes\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right]=\mathbf{M}_i\,\mathbf{vec}\left(\mathbf{\Omega}_i\right). \tag{39}$$

As $\mathbf{M}_i$ has rank T^2-q^2,[11] we cannot invert (39) and recover $\mathbf{\Omega}_i$ unless we impose restrictions. We start by imposing uncorrelatedness restrictions on errors v_{it}. A particular example is a moving average (MA) process of order r, in which case the conditional covariance between v_{it} and $v_{i,t+r+1}$ given $\mathbf{W}_i$ is zero for all t.

Formally, we make the following assumption.

[11] Note that $\operatorname{rank}\left(\mathbf{M}_i\right)=\operatorname{Tr}\left(\mathbf{M}_i\right)=T^2-\left[T-\left(T-q\right)\right]^2$.

Assumption 4 *There exists a vector of m parameters* $\boldsymbol{\omega}_i$*, possibly dependent on* $\mathbf{W}_i$*, and a known (selection) matrix* $\mathbf{S}_2$ *such that:*

$$\mathbf{vec}\left(\boldsymbol{\Omega}_i\right) = \mathbf{S}_2\boldsymbol{\omega}_i. \tag{40}$$

Note that since $\mathbf{Var}\left(\mathbf{v}_i|\ \mathbf{W}_i\right) = \mathbf{E}\left[\mathbf{Var}\left(\mathbf{v}_i \mid \mathbf{W}_i, \boldsymbol{\gamma}_i\right) \mid \mathbf{W}_i\right]$, Assumption 4 is consistent with an underlying moving average model with unobserved heterogeneity of the form

$$\mathbf{Var}\left(\mathbf{v}_i \mid \mathbf{W}_i, \boldsymbol{\gamma}_i\right) = \mathbf{S}_2\boldsymbol{\phi}\left(\mathbf{W}_i, \boldsymbol{\gamma}_i\right)$$

for an unspecified function $\boldsymbol{\phi}$ such that $\boldsymbol{\omega}_i = \mathbf{E}\left[\boldsymbol{\phi}\left(\mathbf{W}_i, \boldsymbol{\gamma}_i\right) \mid \mathbf{W}_i\right]$, possibly including a larger vector of fixed effects than those present in the conditional mean.

Assumption 4 contains the case where all errors are conditionally uncorrelated, in which case $m = T$ and $\mathbf{S}_2$ is a selection matrix that has zeros everywhere except at positions $(1,1)$, $(T+2,2)$,..., (T^2,T). More generally, Assumption 4 contains moving-average processes of the form

$$v_{it} = u_{it} + \theta_{1t}u_{i,t-1} + ... + \theta_{rt}u_{i,t-r}, \quad t = 1, ..., T, \tag{41}$$

where $\theta_{11}, ..., \theta_{rT}$ are unrestricted parameters,[12] and $u_{i,1-r}, ..., u_{iT}$ are mutually uncorrelated given regressors. In the MA(r) case, $m = T + T - 1 + ... + T - r = (r+1)(T - r/2)$.

Now, combining (39) and (40) we obtain:

$$\mathbf{M}_i\mathbf{E}\left[\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right) \otimes \left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right] = \mathbf{M}_i\mathbf{S}_2\boldsymbol{\omega}_i. \tag{42}$$

We thus have the following identification theorem.

Theorem 2 ***(variances of errors, information in levels)***

Let Assumptions 1, 3 and 4 hold. Suppose that

$$\text{rank}\left[\mathbf{M}_i\mathbf{S}_2\right] = m. \tag{43}$$

Then matrix $\boldsymbol{\Omega}_i$ *is identified from covariance restrictions in levels (35).*

In the particular case where errors are i.i.d. homoskedastic (and so $m = 0$) we also have the following corollary.

[12] $\theta_{11}, ..., \theta_{rT}$ may depend on regressors $\mathbf{W}_i$, although we omit the i subindex for clarity. They could also depend on individual effects $\boldsymbol{\xi}_i$, as long as $\mathbf{E}\left(u_{it}|\mathbf{W}_i, \boldsymbol{\gamma}_i, \boldsymbol{\xi}_i\right) = 0$.

Corollary 3 ***(variances of errors, i.i.d.)***

If errors are i.i.d. independent of $\mathbf{W}_i$ *with variance* σ^2 *we have*

$$\sigma^2 = \frac{1}{T-q}\mathbf{E}\left[(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})' \mathbf{Q}_i (\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})\right].$$

It is interesting to study the order condition associated with the rank condition (43). One can check that

$$\operatorname{rank}\left[\mathbf{M}_i\mathbf{S}_2\right] \leq \frac{T(T+1)}{2} - \frac{q(q+1)}{2},$$

with equality when $\mathbf{S}_2$ selects all $T(T+1)/2$ non-redundant elements of $\mathbf{vec}\left(\boldsymbol{\Omega}_i\right)$, see Lemma A1 i) in Appendix A. So, the order condition associated with (43) is:

$$\frac{T(T+1)}{2} - \frac{q(q+1)}{2} \geq m. \tag{44}$$

In particular, in the MA(r) case we need that

$$\frac{T(T+1)}{2} - \frac{q(q+1)}{2} \geq (r+1)\left(T - \frac{r}{2}\right). \tag{45}$$

The left-hand-side in (45) is decreasing in q, while the right-hand side is increasing in r. So, equation (45) emphasizes a trade-off between the number of individual-specific effects and the order of the moving-average process.

Working with the within-group equation (5) alone requires stronger conditions for identification, as shown in the following theorem.

Theorem 3 ***(variance of errors, within information)***

Let Assumptions 1, 3 and 4 hold. Suppose that

$$\operatorname{rank}\left[\left(\mathbf{Q}_i \otimes \mathbf{Q}_i\right)\mathbf{S}_2\right] = m. \tag{46}$$

Then matrix $\boldsymbol{\Omega}_i$ *is identified from the within-group covariance restrictions (36) alone.*

The order condition associated with the rank condition (46) is (see Lemma A1 ii) in Appendix A):

$$\frac{(T-q)(T-q+1)}{2} \geq m.$$

Hence, the order condition is more restrictive than the one which was obtained using covariance restrictions in levels, see equation (44).

For example, consider the AR(1) model (29) with a single heterogeneous intercept, and $T = 3$. The autoregressive parameter ρ is not identified from within-group equations alone. However, ρ is identified from covariance restrictions in levels, as the IV estimand in the regression of $(y_{i3} - y_{i2})$ on $(y_{i2} - y_{i1})$ using y_{i1} as instrument.

Variances of errors: AR restrictions. Autoregressive errors are very popular in applied work, and are *not* covered by assumption (40) because autoregressive processes are correlated at all lags. Nevertheless, a similar approach can be adopted to study identification.[13] To see how, consider the following model:

$$v_{it} = \rho_{1t} v_{i,t-1} + ... + \rho_{pt} v_{i,t-p} + u_{it}, \quad t = p+1, ..., T, \tag{47}$$

where $\rho_{1,p+1}, ..., \rho_{pT}$ are unrestricted parameters and $u_{i,p+1}, ..., u_{iT}$ satisfy Assumption 4. In the case where u_{it} is MA(r), v_{it} given by (47) follows an ARMA(p,r) process.

Let $\mathbf{u}_i = (u_{i,p+1}, ..., u_{iT})'$, and let $\mathbf{R}$ be the $(T-p) \times T$ matrix:

$$\mathbf{R} = \begin{pmatrix} -\rho_{p,p+1} & -\rho_{p-1,p+1} & ... & -\rho_{1,p+1} & 1 & 0 & ... & ... & ... & ... & 0 & 0 \\ 0 & -\rho_{p,p+2} & ... & -\rho_{2,p+2} & -\rho_{1,p+2} & 1 & ... & ... & ... & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & -\rho_{p,T-1} & -\rho_{p-1,T-1} & ... & 1 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & -\rho_{pT} & ... & -\rho_{1T} & 1 \end{pmatrix}.$$

Left-multiplying (1) by $\mathbf{R}$ we obtain, as $\mathbf{R}\mathbf{v}_i = \mathbf{u}_i$:

$$\mathbf{R}\mathbf{y}_i = \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta} + \mathbf{R}\mathbf{X}_i\boldsymbol{\gamma}_i + \mathbf{u}_i. \tag{48}$$

Let $\boldsymbol{\omega}_i$ be an $m \times 1$ vector of parameters such that:

$$\mathbf{vec}\left(\mathbf{Var}\left(\mathbf{u}_i|\mathbf{W}_i\right)\right) = \mathbf{S}_2\boldsymbol{\omega}_i.$$

Let also:

$$\widetilde{\mathbf{M}}_i = \mathbf{I}_{(T-p)^2} - \left[\mathbf{R}\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{R}'\mathbf{R}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{R}'\right] \otimes \left[\mathbf{R}\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{R}'\mathbf{R}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{R}'\right].$$

Variance restrictions in model (48) imply that

$$\widetilde{\mathbf{M}}_i\mathbf{E}\left[\left(\mathbf{R}\mathbf{y}_i - \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta}\right) \otimes \left(\mathbf{R}\mathbf{y}_i - \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right] \;=\; \widetilde{\mathbf{M}}_i\mathbf{S}_2\boldsymbol{\omega}_i. \tag{49}$$

Note that, by multiplying by $\mathbf{R}$ we have lost degrees of freedom, as the rank of $\widetilde{\mathbf{M}}_i$ is $[(T-p)^2 - q^2]$ while $\mathbf{M}_i$ has rank $[T^2 - q^2]$. These additional restrictions on the variance-covariance matrix of errors are intuitive, as there are p extra individual-specific parameters to difference out, the initial shocks $v_{i,1-p}, ..., v_{i0}$. Multiplying by $\mathbf{R}$ permits to eliminate these p individual effect. Then, multiplication by $\widetilde{\mathbf{M}}_i$ allows to eliminate the q remaining ones.

[13] However, contrary to the moving average case, an autoregressive model with unobserved heterogeneity does not generally imply an autoregressive structure for $\mathbf{Var}\left(\mathbf{v}_i|\ \mathbf{W}_i\right)$.

It follows from (49) that, for the variances of $u_{i,p+1}, ..., u_{iT}$ and parameters $\rho_{1,p+1}, ..., \rho_{pT}$ to be identified from equation (42) the following rank condition needs to be satisfied:

$$\text{rank}\left(\widetilde{\mathbf{M}}_i \mathbf{S}_2\right) = m. \tag{50}$$

In particular, we need that:

$$\frac{(T-p)(T-p+1)}{2} - \frac{q(q+1)}{2} \geq m.$$

So the maximal q that can be allowed for is inversely related to p. In the case where u_{it} is MA(r), q is inversely related to both p and r.

Before ending this discussion, three remarks are in order. First, contrary to the moving average case, (50) is not strictly sufficient for identification to hold. Indeed, we also need parameters $\rho_{1,p+1}, ..., \rho_{pT}$ to be identified from (49).

Next, one could similarly analyze the case of AR-type restrictions using within information only, as opposed to using restrictions in levels as we have done in this paragraph. The order condition for identification then becomes more restrictive, as it requires that:

$$\frac{(T-p-q)(T-p-q+1)}{2} \geq m.$$

Lastly, the analysis in this section focuses on non-stationary ARMA models. Under stationarity, additional identifying restrictions could be obtained, although non-linear in the autoregressive parameters.

Illustrations. We first illustrate the results in Example 2 with $L = 3$, which corresponds to our empirical application. We focus on the subpopulation of individuals who have $s_{i\ell} = 1$ only in one period, i.e. such that $n_i = 1$, the analysis being similar for other values of n_i. We assume without loss of generality that $s_{i1} = 1$, and $s_{i2} = s_{i3} = 0$.

In this case, levels restrictions (39) are:

$$\begin{cases} \text{Var}\left(y_{i1}\right) & = & \text{Var}\left(\alpha_i\right) + 2\,\text{Cov}\left(\alpha_i, \beta_i\right) + \text{Var}\left(\beta_i\right) + \text{Var}\left(v_{i1}\right), \\ \text{Var}\left(y_{i2}\right) & = & \text{Var}\left(\alpha_i\right) + \text{Var}\left(v_{i2}\right), \\ \text{Var}\left(y_{i3}\right) & = & \text{Var}\left(\alpha_i\right) + \text{Var}\left(v_{i3}\right), \\ \text{Cov}\left(y_{i1}, y_{i2}\right) & = & \text{Var}\left(\alpha_i\right) + \text{Cov}\left(\alpha_i, \beta_i\right) + \text{Cov}\left(v_{i1}, v_{i2}\right), \\ \text{Cov}\left(y_{i1}, y_{i3}\right) & = & \text{Var}\left(\alpha_i\right) + \text{Cov}\left(\alpha_i, \beta_i\right) + \text{Cov}\left(v_{i1}, v_{i3}\right), \\ \text{Cov}\left(y_{i2}, y_{i3}\right) & = & \text{Var}\left(\alpha_i\right) + \text{Cov}\left(v_{i2}, v_{i3}\right). \end{cases}$$

We see that, when errors are uncorrelated with unrestricted variances, $\text{Var}\left(\beta_i\right)$ and $\text{Var}\left(v_{i1}\right)$ are not separately identified. Although the order condition for identification is

satisfied,[14] the rank condition is not. Remark also that, if we impose the stationarity restriction that all three variances of v_{i1}, v_{i2} and v_{i3} are equal, then they are identified along with the covariance matrix of individual effects.

It is easy to see that $\mathrm{Var}\left(v_{i3} - v_{i2}\right)$ is identified from the within-group restrictions (36) alone. So, if we assume that v_{i2} and v_{i3} are uncorrelated and have equal variance, then $\mathrm{Var}\left(v_{i2}\right) = \mathrm{Var}\left(v_{i3} - v_{i2}\right)/2$ is also identified from those restrictions.

As a second illustration, consider Example 1 with AR(1) errors:

$$v_{it} = \rho v_{i,t-1} + u_{it}.$$

We start by assuming that ρ is known. Applying the $\mathbf{R}$ transformation to equation (7) we obtain:

$$\begin{aligned} \underbrace{y_{it} - \rho y_{i,t-1}}_{y^*_{it}(\rho)} &= (1-\rho)\alpha_i + \beta_i\left(t - \rho(t-1)\right) + u_{it} \\ &= \underbrace{(1-\rho)\alpha_i + \rho\beta_i}_{\alpha^*_i(\rho)} + \underbrace{t(1-\rho)\beta_i}_{t\beta^*_i(\rho)} + u_{it}. \end{aligned}$$

When $T = 4$, we obtain:[15]

$$\left\{\begin{array}{lcl} \mathrm{Var}\left(y^*_{i2}(\rho)\right) & = & \mathrm{Var}\left(\alpha^*_i(\rho)\right) + 4\,\mathrm{Cov}\left(\alpha^*_i(\rho), \beta^*_i(\rho)\right) \\ & & \quad +4\,\mathrm{Var}\left(\beta^*_i(\rho)\right) + \mathrm{Var}\left(u_{i2}\right), \\ \mathrm{Var}\left(y^*_{i3}(\rho) - y^*_{i2}(\rho)\right) & = & \mathrm{Var}\left(\beta^*_i(\rho)\right) + \mathrm{Var}\left(u_{i3} - u_{i2}\right), \\ \mathrm{Var}\left(y^*_{i4}(\rho) - 2y^*_{i3}(\rho) + y^*_{i2}(\rho)\right) & = & \mathrm{Var}\left(u_{i4} - 2u_{i3} + u_{i2}\right), \\ \mathrm{Cov}\left(y^*_{i2}(\rho), y^*_{i3}(\rho) - y^*_{i2}(\rho)\right) & = & \mathrm{Cov}\left(\alpha^*_i(\rho), \beta^*_i(\rho)\right) + 2\,\mathrm{Var}\left(\beta^*_i(\rho)\right) \\ & & \quad + \mathrm{Cov}\left(u_{i2}, u_{i3} - u_{i2}\right), \\ \mathrm{Cov}\left(y^*_{i2}(\rho), y^*_{i4}(\rho) - 2y^*_{i3}(\rho) + y^*_{i2}(\rho)\right) & = & \mathrm{Cov}\left(u_{i2}, u_{i4} - 2u_{i3} + u_{i2}\right), \\ \mathrm{Cov}\left(y^*_{i3}(\rho) - y^*_{i2}(\rho), y^*_{i4}(\rho) - 2y^*_{i3}(\rho) + y^*_{i2}(\rho)\right) & = & \mathrm{Cov}\left(u_{i3} - u_{i2}, u_{i4} - 2u_{i3} + u_{i2}\right). \end{array}\right.$$

It is easy to check that, if u_{i2}, u_{i3} and u_{i4} are assumed uncorrelated, then they are identified from this set of restrictions, together with the covariance matrix of individual effects. This is consistent with the order condition (44) being satisfied in this case.

Note that ρ is not identified from levels equations when $T = 4$. When $T = 5$ we obtain additional identifying restrictions which may suffice for ρ to be identified. For example:

$$\begin{aligned} \mathrm{Var}\left(\beta^*_i(\rho)\right) &= \mathrm{Cov}\left[y^*_{i3}(\rho) - y^*_{i2}(\rho), y^*_{i5}(\rho) - y^*_{i4}(\rho)\right] \\ &= \mathrm{Var}\left(y^*_{i3}(\rho) - y^*_{i2}(\rho)\right) + \frac{1}{2}\,\mathrm{Cov}\left[y^*_{i3}(\rho) - 2y^*_{i2}(\rho), y^*_{i4}(\rho) - 2y^*_{i3}(\rho) + y^*_{i2}(\rho)\right]. \end{aligned}$$

[14] As: $3(3+1)/2 - 2(2+1)/2 = 3$, see equation (45).

[15] $T = 4$ means that we have 3 observations on $y^*_{it}(\rho)$ for given ρ ($t = 2, 3, 4$).

Note that, when $T = 5$, the order condition (44) is satisfied even when u_{it} follows an unrestricted MA(1) process. This suggests that the conditions for identification become rapidly less demanding as T increases.

3.3 Efficiency bounds

Here we show how Chamberlain's analysis can be extended to obtain a joint information bound for common parameters, means and variances of random coefficients, and a parameterization of the variances of errors. Let us write down model (1) as:

$$\mathbf{E}\left(\mathbf{y}_i|\mathbf{W}_i,\boldsymbol{\gamma}_\mathbf{i}\right) = \mathbf{Z}_i\boldsymbol{\delta} + \mathbf{X}_i\boldsymbol{\gamma}_\mathbf{i} \tag{51}$$

together with a specification of the conditional variance of $\mathbf{v}_i$ given $\mathbf{W}_i$ and $\boldsymbol{\gamma}_i$:

$$\mathbf{E}\left(\mathbf{v}_i \otimes \mathbf{v}_i|\mathbf{W}_i,\boldsymbol{\gamma}_\mathbf{i}\right) = \boldsymbol{\omega}_i\left(\boldsymbol{\phi}\right), \tag{52}$$

where $\boldsymbol{\omega}_i$ is a $T^2 \times 1$ vector of functions of a parameter $\boldsymbol{\phi}$, which may also depend on $\mathbf{W}_i$. However, we assume that the variance of $\mathbf{v}_i$ does not depend on $\boldsymbol{\gamma}_i$.[16]

Using (52) together with Assumption 1 we obtain the following expression for the conditional second-order moments of $\mathbf{y}_i$:

$$\begin{aligned}\mathbf{E}\left(\mathbf{y}_i \otimes \mathbf{y}_i|\mathbf{W}_i,\boldsymbol{\gamma}_\mathbf{i}\right) &= \left(\mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{Z}_i\boldsymbol{\delta}\right) + \boldsymbol{\omega}_i\left(\boldsymbol{\phi}\right) + \left(\mathbf{X}_i \otimes \mathbf{Z}_i\boldsymbol{\delta} + \mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{X}_i\right)\boldsymbol{\gamma}_i \\ &\quad + \left(\mathbf{X}_i \otimes \mathbf{X}_i\right)\left(\boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i\right). \end{aligned} \tag{53}$$

Stacking (51) and (53) together yields

$$\mathbf{E}\left(\mathbf{y}_i^*|\mathbf{W}_i,\boldsymbol{\gamma}_i^*\right) = \mathbf{d}\left(\mathbf{W}_i,\boldsymbol{\theta}\right) + \mathbf{R}\left(\mathbf{W}_i,\boldsymbol{\theta}\right)\boldsymbol{\gamma}_i^*, \tag{54}$$

where $\boldsymbol{\theta} = (\boldsymbol{\delta},\boldsymbol{\phi})$, and

$$\boldsymbol{\gamma}_i^* = \begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \end{pmatrix}, \quad \mathbf{d}\left(\mathbf{W}_i,\boldsymbol{\theta}\right) = \begin{pmatrix} \mathbf{Z}_i\boldsymbol{\delta} \\ \left(\mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{Z}_i\boldsymbol{\delta}\right) + \boldsymbol{\omega}_i\left(\boldsymbol{\phi}\right) \end{pmatrix},$$

and

$$\mathbf{R}\left(\mathbf{W}_i,\boldsymbol{\theta}\right) = \begin{pmatrix} \mathbf{X}_i & \mathbf{0} \\ \left(\mathbf{X}_i \otimes \mathbf{Z}_i\boldsymbol{\delta} + \mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{X}_i\right) & \left(\mathbf{X}_i \otimes \mathbf{X}_i\right) \end{pmatrix}.$$

[16] In cases where $\mathbf{E}\left(\mathbf{v}_i \otimes \mathbf{v}_i|\mathbf{W}_i,\boldsymbol{\gamma}_i,\boldsymbol{\xi}_i\right) = \boldsymbol{\Gamma}\left(\mathbf{W}_i,\boldsymbol{\phi}\right)\begin{pmatrix}\boldsymbol{\gamma}_i \\ \boldsymbol{\xi}_i\end{pmatrix}$, we could extend the model and apply a similar approach treating $\boldsymbol{\xi}_i$ as additional random coefficients.

Equation (54), which combines mean and covariance restrictions in levels, is a special case of model (15).[17] Therefore, the optimal moments (and associated semiparametric bound) for $\boldsymbol{\delta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\gamma}^* = \mathbf{E}(\boldsymbol{\gamma}_i^*) = \mathbf{E}\begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \end{pmatrix}$ are of the form given in expression (30).

In particular, using (54) instead of the conditional mean model (51) we obtain in general a tighter bound for the common parameters $\boldsymbol{\delta}$. This is because we have restricted the covariance structure of errors via equation (52). Moreover, if those covariance restrictions do not suffice for $\mathbf{E}(\boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i)$ to be identified, then the information bound for the variance of individual effects will be zero.

4 Identification of distributions

In this section, we discuss the identification of distributions. We start with third and fourth-order moments of errors and individual effects, and then study the identification of their densities.

4.1 Higher-order moments

In applications, it may be of interest to document the skewness and kurtosis of individual effects in addition to mean and variance. It turns out that the model's linearity makes it easy to generalize the previous analysis to higher-order moments.

Definitions. Let $\mathbf{U}$ be an n-dimensional random vector with zero mean and well-defined moments to the fourth-order. We denote by $\boldsymbol{\kappa}_3(\mathbf{U})$ the n^3-dimensional cumulant vector of order 3 whose elements $\kappa_3^{i,j,k}(\mathbf{U})$, for $(i,j,k) \in \{1,...,n\}^3$, are arranged in lexicographic order. Likewise, we denote by $\boldsymbol{\kappa}_4(\mathbf{U})$ the vector of n^4 cumulants of order 4 $\kappa_4^{i,j,k,\ell}(\mathbf{U})$. There is a mapping between moments and cumulants but in our context it is more convenient to work with the latter because of their properties (see Appendix D for further details). In particular, cumulants satisfy a useful multilinearity property. Namely, for any conformable matrix $\mathbf{A}$ we have:

$$\begin{aligned} \boldsymbol{\kappa}_3(\mathbf{AU}) &= (\mathbf{A} \otimes \mathbf{A} \otimes \mathbf{A})\, \boldsymbol{\kappa}_3(\mathbf{U}), \\ \boldsymbol{\kappa}_4(\mathbf{AU}) &= (\mathbf{A} \otimes \mathbf{A} \otimes \mathbf{A} \otimes \mathbf{A})\, \boldsymbol{\kappa}_4(\mathbf{U}). \end{aligned}$$

[17] The only difference is that $\mathbf{E}(\boldsymbol{\gamma}_i^* | \mathbf{W}_i)$ is not fully unrestricted, as its components are first and second moments of the same underlying $\boldsymbol{\gamma}_i$. However, these extra restrictions imply moment *inequalities* that do not affect the bound.

Cumulants of effects. To recover the higher-order cumulants of individual effects we assume that individual effects are independent of errors conditionally on regressors (Assumption 2). Full independence will not be needed to derive the identification results in this subsection. For this purpose, the assumption that $\boldsymbol{\gamma}_i$ and $\mathbf{v}_i$ have zero cross-cumulants of order 3 and 4 will be sufficient. However, full independence will be needed to recover the distribution of individual effects in the next subsection.

Using the between-group equation (6) together with Assumption 2, we obtain that:

$$\begin{aligned}\boldsymbol{\kappa}_3\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) &= \boldsymbol{\kappa}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \boldsymbol{\kappa}_3\left(\mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i\right), \\ &= \boldsymbol{\kappa}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \left(\mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i\right)\boldsymbol{\kappa}_3\left(\mathbf{v}_i|\mathbf{W}_i\right), \end{aligned} \tag{55}$$

and, similarly:

$$\boldsymbol{\kappa}_4\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) = \boldsymbol{\kappa}_4\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \left(\mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i\right)\boldsymbol{\kappa}_4\left(\mathbf{v}_i|\mathbf{W}_i\right). \tag{56}$$

It follows that the conditional cumulants of individual effects are identified given error cumulants. Remark that, as conditional moments can be recovered from conditional cumulants, it follows from these results that conditional and thus unconditional moments of individual effects are also identified.

Consider for example model (25), with a single heterogeneous intercept and i.i.d. errors. We obtain:

$$\begin{aligned}\kappa_3\left(\gamma_{i1}|\mathbf{Z}_i\right) &= \kappa_3\left(\overline{y}_i - \overline{\mathbf{z}}_i'\boldsymbol{\delta}|\mathbf{Z}_i\right) - \frac{\kappa_3\left(v_{it}\right)}{T^2}, \\ \kappa_4\left(\gamma_{i1}|\mathbf{Z}_i\right) &= \kappa_4\left(\overline{y}_i - \overline{\mathbf{z}}_i'\boldsymbol{\delta}|\mathbf{Z}_i\right) - \frac{\kappa_4\left(v_{it}\right)}{T^3}.\end{aligned}$$

Interestingly, (55) and (56) show that the bias on the cumulant of individual effects estimates $\widehat{\boldsymbol{\gamma}}_i$ is of a smaller order of magnitude than the bias on the variance.[18]

Error cumulants. We now turn to the identification of the cumulants of time-varying errors. Taking third- and fourth-order cumulants in model (1), we obtain the following restrictions (in levels):

$$\boldsymbol{\kappa}_3\left(\mathbf{y}_i|\mathbf{W}_i\right) = \left(\mathbf{X}_i \otimes \mathbf{X}_i \otimes \mathbf{X}_i\right)\boldsymbol{\kappa}_3\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) + \boldsymbol{\kappa}_3\left(\mathbf{v}_i|\mathbf{W}_i\right), \tag{57}$$

$$\boldsymbol{\kappa}_4\left(\mathbf{y}_i|\mathbf{W}_i\right) = \left(\mathbf{X}_i \otimes \mathbf{X}_i \otimes \mathbf{X}_i \otimes \mathbf{X}_i\right)\boldsymbol{\kappa}_4\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) + \boldsymbol{\kappa}_4\left(\mathbf{v}_i|\mathbf{W}_i\right). \tag{58}$$

[18] When $\frac{\mathbf{X}_i'\mathbf{X}_i}{T} \xrightarrow{p} constant > 0$ as T tends to infinity, the biases on third- and fourth-order cumulants of individual effects are $O(1/T^2)$ and $O(1/T^3)$, respectively, while the bias on the variance is $O(1/T)$.

As in the case of variances, these systems of equations are singular unless we impose restrictions on the dependence of errors over time. We adopt a similar approach as in (40) and assume that:

$$\boldsymbol{\kappa}_3\left(\mathbf{v}_i|\mathbf{W}_i\right) = \mathbf{S}_3\boldsymbol{\omega}_{3i}, \tag{59}$$

$$\boldsymbol{\kappa}_4\left(\mathbf{v}_i|\mathbf{W}_i\right) = \mathbf{S}_4\boldsymbol{\omega}_{4i}, \tag{60}$$

where $\mathbf{S}_3$ and $\mathbf{S}_4$ are selection matrices and $\boldsymbol{\omega}_{3i}$ and $\boldsymbol{\omega}_{4i}$ are vectors of m_3 and m_4 parameters, respectively, possibly dependent on $\mathbf{W}_i$. Under these assumptions, identification of error cumulants can be shown if rank conditions analog to (43) are satisfied.

To motivate restrictions (59) and (60), let us consider a moving average model of the form (41), where innovations $u_{i,1-r}, ..., u_{iT}$ are now assumed mutually *independent* given regressors. Errors are thus modelled as linear combinations of independent (and not simply uncorrelated) underlying shocks.[19] Because of linearity and independence, it follows that for any time periods t and t' such that v_{it} and $v_{i,t'}$ are independent, the cumulants $\kappa_3^{t,t',s}(\mathbf{v}_i|\mathbf{W}_i)$ and $\kappa_4^{t,t',s,s'}(\mathbf{v}_i|\mathbf{W}_i)$ are zero for all s, s' (see Lemma 1 in Bonhomme and Robin, 2009a). So, these error structures satisfy (59) and (60) for particular selection matrices.

Specifically, in an independent moving average model of order r, v_{it} is conditionally independent of $v_{i,t+r+1}$ for all t. Simple combinatorics then shows that third and fourth-order cumulants depend on $m_3(r)$ and $m_4(r)$ free parameters, respectively, where:

$$\begin{aligned} m_3(r) &= T + 2(T-1) + ... + (r+1)(T-r), \\ m_4(r) &= T + \left(\begin{array}{c} 3 \\ 2 \end{array}\right)(T-1) + ... + \left(\begin{array}{c} r+2 \\ 2 \end{array}\right)(T-r). \end{aligned}$$

It can be shown that the order conditions for identification are, in this case:

$$\left(\begin{array}{c} T+2 \\ 3 \end{array}\right) - \left(\begin{array}{c} q+2 \\ 3 \end{array}\right) \geq m_3(r) \quad , \quad \text{and} \quad \left(\begin{array}{c} T+3 \\ 4 \end{array}\right) - \left(\begin{array}{c} q+3 \\ 4 \end{array}\right) \geq m_4(r). \tag{61}$$

Hence, again, a trade-off between the number of individual-specific effects and the order of the MA process. Interestingly, the order conditions for higher-order cumulants are less stringent than for the variance, compare (61) with (45).

It is also possible to show identification of higher-order moments in autoregressive models of the form (47), if the underlying shocks u_{it} follow an independent moving average model. For that, it suffices to compute cumulants in the equation in quasi-differences (49).

[19] See Rao (1969) and more recently the literature on Independent Component Analysis (ICA) (e.g., Hyvärinen *et al.*, 2001), for references on linear independent factor models in the statistical literature.

Lastly, one can similarly study the identification of higher-order cumulants using within information alone, see equation (5). The conditions for identification then become more restrictive.[20]

Remark on efficiency bounds. The arguments of Subsection 3.3 can be extended to higher-order moments. To do so, consider further extending the model to specify the third-order moments of errors as:

$$\mathbf{E}\left(\mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i | \mathbf{W}_i, \boldsymbol{\gamma}_\mathbf{i}\right) = \boldsymbol{\mu}_{3i}\left(\boldsymbol{\phi}_3\right). \tag{62}$$

We can write third-order moment restrictions as:

$$\begin{aligned}\mathbf{E}\left(\mathbf{y}_i \otimes \mathbf{y}_i \otimes \mathbf{y}_i | \mathbf{W}_i, \boldsymbol{\gamma}_i\right) &= \left(\mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{Z}_i\boldsymbol{\delta}\right) + \mathbf{P}_T\left(\boldsymbol{\omega}_i \otimes \mathbf{Z}_i\boldsymbol{\delta}\right) + \boldsymbol{\mu}_{3i} \\ &\quad + \mathbf{P}_T\left[\left(\mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{Z}_i\boldsymbol{\delta} + \boldsymbol{\omega}_i\right) \otimes \mathbf{X}_i\right]\boldsymbol{\gamma}_i + \mathbf{P}_T\left[\mathbf{Z}_i\boldsymbol{\delta} \otimes \mathbf{X}_i \otimes \mathbf{X}_i\right]\left(\boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i\right) \\ &\quad + \left(\mathbf{X}_i \otimes \mathbf{X}_i \otimes \mathbf{X}_i\right)\left(\boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i\right),\end{aligned}$$

where $\boldsymbol{\omega}_i = \boldsymbol{\omega}_i\left(\boldsymbol{\phi}_2\right)$, $\boldsymbol{\mu}_{3i} = \boldsymbol{\mu}_{3i}\left(\boldsymbol{\phi}_3\right)$, and $\mathbf{P}_T$ denotes the $T^3 \times T^3$ "triplicating" permutation matrix that satisfies, for all $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbf{R}^{3T}$:

$$\mathbf{P}_T\left(\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}\right) = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} + \mathbf{b} \otimes \mathbf{c} \otimes \mathbf{a} + \mathbf{c} \otimes \mathbf{a} \otimes \mathbf{b}.$$

We can then stack first, second, and third-order moment restrictions to obtain:

$$\mathbf{E}\left(\mathbf{y}_i^{3*} | \mathbf{W}_i, \boldsymbol{\gamma}_i^{3*}\right) = \mathbf{d}_3\left(\mathbf{W}_i, \boldsymbol{\theta}_3\right) + \mathbf{R}_3\left(\mathbf{W}_i, \boldsymbol{\theta}_3\right)\boldsymbol{\gamma}_i^{3*}, \tag{63}$$

where $\boldsymbol{\theta}_3 = (\boldsymbol{\delta}, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3)$, and:

$$\boldsymbol{\gamma}_i^{3*} = \begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \end{pmatrix}.$$

Equation (63) still falls into the framework considered in Chamberlain (1992). Note that this approach can be extended to the m-th order, yielding:

$$\mathbf{E}\left(\mathbf{y}_i^{m*} | \mathbf{W}_i, \boldsymbol{\gamma}_i^{m*}\right) = \mathbf{d}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right) + \mathbf{R}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)\boldsymbol{\gamma}_i^{m*}, \tag{64}$$

[20]It can be shown that, in an independent MA(r) model, the order conditions for identification when working with within information are:

$$\begin{pmatrix} T-q+2 \\ 3 \end{pmatrix} \geq m_3(r) \quad , \quad \text{and} \quad \begin{pmatrix} T-q+3 \\ 4 \end{pmatrix} \geq m_4(r).$$

where $\boldsymbol{\theta}_m = (\boldsymbol{\delta}, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, ..., \boldsymbol{\phi}_m)$, with $\boldsymbol{\phi}_3, ..., \boldsymbol{\phi}_m$ a parameterization of error moments up to the m-th order, and where:

$$\boldsymbol{\gamma}_i^{m*} = \begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \\ \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \otimes \boldsymbol{\gamma}_i \\ ... \\ \underbrace{\boldsymbol{\gamma}_i \otimes ... \otimes \boldsymbol{\gamma}_i}_{\text{m times}} \end{pmatrix}.$$

This framework can be used to compute semiparametric efficiency bounds under the independence assumption between individual effects and errors (Assumption 2). We focus on computing bounds for $\boldsymbol{\delta}$, although any moment of individual effects or errors could be analyzed in a similar way.

Consider the increasing sequence of moment conditions (64), for $m = 2, 3, ...$ Let $\mathbf{V}^{(m)}$ be the efficiency bound on the asymptotic variance for $\boldsymbol{\delta}$ obtained from the first m of those moment conditions. $\mathbf{V}^{(m)}$ can be computed using Chamberlain's (1992) results. Following the discussion in Appendix C, $\mathbf{V}^{(m)}$ is the efficiency bound corresponding to the conditional moment restriction:

$$\mathbf{E}\left[\mathbf{A}_{mi}\left(\mathbf{y}_i^{m*} - \mathbf{d}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)\right)|\mathbf{W}_i\right] = 0,$$

where $\mathbf{A}_{mi}$ is a generalized orthogonal deviation operator such that:

$$\mathbf{A}_{mi}'\mathbf{A}_{mi} = \mathbf{I} - \mathbf{R}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)\left(\mathbf{R}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)'\mathbf{R}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)\right)^{-1}\mathbf{R}_m\left(\mathbf{W}_i, \boldsymbol{\theta}_m\right)'.$$

The sequence $\mathbf{V}^{(m)}$ being nonincreasing in the semi-definite sense (as a larger m means that a larger number of moment conditions is used), we can define the limit:[21]

$$\mathbf{V}^{(\infty)} = \lim_{m \to +\infty} \mathbf{V}^{(m)}.$$

Let $\mathbf{V}_0$ be the semiparametric bound for $\boldsymbol{\delta}$ under independence. Clearly, as $\mathbf{V}_0 \leq \mathbf{V}^{(m)}$ for all m, it follows that $\mathbf{V}_0 \leq \mathbf{V}^{(\infty)}$.

Newey (2004) studies under which conditions, in a given model, the asymptotic variance of the optimal GMM estimator based on an increasing sequence of conditional moment conditions tends to the semiparametric bound, that is, when $\mathbf{V}_0 = \mathbf{V}^{(\infty)}$. He finds that for this to hold, a *spanning* condition is sufficient. This condition requires that the restrictions

[21] See Lemma B.1 in Newey (2004).

imposed by the moment conditions are equivalent to those imposed by the semiparametric model.

Intuitively, we expect a *spanning* condition to hold in our case, as the increasing sequence of moment conditions (64) exhausts all the restrictions implied by independence. We therefore conjecture that $\mathbf{V}_0 = \mathbf{V}^{(\infty)}$.

4.2 Densities

We now turn to the identification of the densities of effects and errors. We work under Assumption 2, which requires conditional independence between $\mathbf{v}_i$ and $\boldsymbol{\gamma}_i$.

To derive the identification results, it is very convenient to work with *characteristic functions*. Let $(\mathbf{Y}, \mathbf{X})$ be a pair of random vectors, $\mathbf{Y} \in \mathbf{R}^L$, and let j be a square root of -1.[22] The conditional characteristic function of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, is defined as:

$$\Psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = \mathbf{E}\left(\exp(j\mathbf{t}'\mathbf{Y})|\mathbf{x}\right), \quad \mathbf{t} \in \mathbf{R}^L.$$

Some useful properties of characteristic functions are discussed in Appendix D.

Densities of individual effects. The following theorem shows that, if the distribution of the error terms is known, then the characteristic function, and hence the distribution, of individual effects is identified.

Theorem 4 ***(characteristic functions of effects)***

Let Assumptions 2 and 3 hold. Suppose that the characteristic function of $\mathbf{v}_i$ *given* $\mathbf{W}_i$ *is nonvanishing on* $\mathbf{R}^T$*. Then we have, for all* $\boldsymbol{\tau} \in \mathbf{R}^q$*:*

$$\Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) = \frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)} \tag{65}$$

and, unconditionally:

$$\Psi_{\boldsymbol{\gamma}_i}(\boldsymbol{\tau}) = \mathbf{E}\left(\frac{\exp\left(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i\right)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\right). \tag{66}$$

The assumption that the characteristic function of errors has no real zeros is very common in the literature on nonparametric deconvolution; see Schennach (2004) and references therein. For example, the characteristic function of the normal distribution has no (real or complex) zeros.

[22] We work with the notation $j^2 = -1$ instead of $i^2 = -1$ to avoid confusion with the index of individual units.

We immediately obtain the following corollary, which shows that the logarithm of the characteristic function of $\boldsymbol{\gamma}_i$ given regressors is identified under similar conditions.

Corollary 4 ***(cumulants of effects)***

Suppose in addition to the assumptions of Theorem 4 that the characteristic function of $\boldsymbol{\gamma}_i$ *given* $\mathbf{W}_i$ *is almost everywhere nonvanishing on* $\mathbf{R}^q$*. Then we have, for all* $\boldsymbol{\tau} \in \mathbf{R}^q$*:*

$$\ln \Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) = \ln \Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) - \ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i). \tag{67}$$

Using Corollary 4, we can see that the identification result for the distribution of individual effects is a generalization of the results that we have obtained for the first moments. Indeed, taking second-order derivatives in (67) evaluated at $\boldsymbol{\tau} = \mathbf{0}$ we obtain the covariance restrictions (31). Taking third and fourth-order derivatives yields the restrictions for third- and fourth-order cumulants (55) and (56), respectively.

Applying the inverse Fourier transform we obtain the following corollary.

Corollary 5 ***(density of effects)***

Under the assumptions of Theorem 4 we have, for all q-dimensional vector $\boldsymbol{\gamma}$*:*

$$f_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i) = \frac{1}{(2\pi)^q} \int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma}) \frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)} \mathbf{d}\boldsymbol{\tau} \tag{68}$$

and, unconditionally:

$$f_{\boldsymbol{\gamma}_i}(\boldsymbol{\gamma}) = \frac{1}{(2\pi)^q} \int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma}) \mathbf{E}\left(\frac{\exp(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)} \right) \mathbf{d}\boldsymbol{\tau}. \tag{69}$$

Corollary 5 shows the identification of the conditional and unconditional densities of individual effects. To interpret this result, we use a large-T approximation, which relies on the fact that the distribution of $\mathbf{H}_i\mathbf{v}_i$ is approximately normal for large T. We obtain (see Appendix A for a derivation):[23]

$$\begin{aligned} f_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i) = {} & f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i) \\ & -\frac{1}{2} \operatorname{Tr}\left(\mathbf{H}_i \boldsymbol{\Omega}_i \mathbf{H}_i' \frac{\partial^2 f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right) + O_p\left(\frac{1}{T^2} \right), \end{aligned} \tag{70}$$

where Tr() is the trace operator.

[23] In order to derive the asymptotic expansion, we assume that $\frac{\mathbf{X}_i'\mathbf{X}_i}{T} \xrightarrow{p} constant > 0$ as T tends to infinity.

In model (25) with a single heterogeneous intercept, no exogenous regressors, and i.i.d. errors with variance σ^2, this yields:

$$f_{\gamma_i}(\gamma) \quad = \quad f_{\widehat{\gamma}_i}(\gamma) - \frac{\sigma^2}{2T}\frac{d^2 f_{\widehat{\gamma}_i}(\gamma)}{d\gamma^2} + O\left(\frac{1}{T^2}\right). \tag{71}$$

Equation (71) is intuitive: in regions of high curvature (such as the mode of the distribution), the density of fixed effects estimates understates the density of population effects.

Densities of time-varying errors. We now consider the identification of the distribution of the error terms. It is convenient to define the following object:

$$\boldsymbol{\kappa}_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = -\mathbf{vec}\left(\frac{\partial^2 \ln \Psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x})}{\partial \mathbf{t}\partial \mathbf{t}'}\right).$$

$\boldsymbol{\kappa}_{\mathbf{Y}|\mathbf{X}}$ is well-defined if the variance of $\mathbf{Y}$ given $\mathbf{X}$ exists (e.g., Székely and Rao, 2000). Moreover:

$$\boldsymbol{\kappa}_{\mathbf{Y}|\mathbf{X}}(\mathbf{0}|\mathbf{x}) = \mathbf{vec}\left(\mathbf{Var}\left(\mathbf{Y}|\mathbf{X}\right)\right).$$

The function $\boldsymbol{\kappa}_{\mathbf{Y}|\mathbf{X}}$ will be useful to extend covariance equalities to equalities involving the full distribution of the random variables.

Assumption 2 implies that, provided that the corresponding characteristic functions do not vanish then, for any $\mathbf{t} \in \mathbf{R}^T$:

$$\begin{aligned} \ln \Psi_{\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) &= \ln \Psi_{\mathbf{X}_i\boldsymbol{\gamma}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) + \ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) \\ &= \ln \Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\mathbf{X}_i'\mathbf{t}|\mathbf{W}_i) + \ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i). \end{aligned} \tag{72}$$

Taking (minus) second derivatives we obtain, in vector form:

$$\boldsymbol{\kappa}_{\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) \quad = \quad (\mathbf{X}_i \otimes \mathbf{X}_i)\,\boldsymbol{\kappa}_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\mathbf{X}_i'\mathbf{t}|\mathbf{W}_i) + \boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i).$$

So, left-multiplying by $\mathbf{M}_i$ (which projects on the orthogonal of $\mathbf{X}_i \otimes \mathbf{X}_i$), this yields:

$$\mathbf{M}_i\boldsymbol{\kappa}_{\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) = \mathbf{M}_i\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i), \quad \mathbf{t} \in \mathbf{R}^T, \tag{73}$$

where $\mathbf{M}_i$ is given by (38).

Equation (73) nicely extends covariance restrictions to restrictions on the entire distribution of the error terms. Indeed, evaluating (73) at $\mathbf{t} = \mathbf{0}$ yields the covariance restrictions in levels (39). Now, as in the case of variances, $\mathbf{M}_i$ having rank $T^2 - q^2$ it is not possible to invert (73) unless the dependence structure of errors is restricted.

We study identification under the assumption that errors follow an independent moving average process of order r of the form (41), where $u_{i,1-r}, ..., u_{iT}$ are mutually independent given regressors. Extensions to autoregressive and ARMA processes with independent underlying innovations can be done along the lines of Section 3.

Lemma A2 in Appendix A shows that, in an independent MA model, the partial derivatives of the log characteristic function of errors are zero for all indices t and t' such that v_{it} and $v_{it'}$ are independent. It follows that there exists an m-dimensional vector of functions $\boldsymbol{\omega}_i(\mathbf{t})$ ($\mathbf{t} \in \mathbf{R}^T$), possibly dependent on regressors, such that:

$$\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i) = \mathbf{S}_2\boldsymbol{\omega}_i(\mathbf{t}), \quad \mathbf{t} \in \mathbf{R}^T. \tag{74}$$

The selection matrix $\mathbf{S}_2$ is the same that appeared in the covariance restrictions (40). Indeed, (74) evaluated at $\mathbf{t} = \mathbf{0}$ yields (40). In particular, $m = (r+1)(T - r/2)$.

Combining (73) with (74), we obtain the folowing identification theorem.

Theorem 5 ***(characteristic function of errors)***

Let Assumptions 1, 2 and 3 hold. Suppose that (74) holds, and that the rank condition (43) is satisfied. Lastly, suppose that the conditional characteristic function of errors $\Psi_{\mathbf{v}_i|\mathbf{W}_i}$ is non-vanishing on $\mathbf{R}^T$. Then $\Psi_{\mathbf{v}_i|\mathbf{W}_i}$ is identified from the restrictions in levels (72).

The identification of $\ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}$ comes from the fact that its second derivatives are identified, and that both the log-characteristic function and its first derivatives are zero at $\mathbf{t} = \mathbf{0}$. This last part comes from the first derivative of the log-characteristic function at the origin being the mean of the random variable, which is zero because of Assumption 1.

Note that the rank condition for identification, equation (43), is the one that was needed for the identification of error variances in Section 3. Remark also that Theorem 5 implies the identification of the density of errors, using the inverse Fourier transformation, as in Corollary 5 above.

To summarize the results so far, we have obtained the nonparametric identification of the distributions of individual effects and time-varying errors under two main conditions: the independence of effects and errors, and conditional independence restrictions on errors that are sufficiently spaced. These results extend Kotlarski (1967) and Székely and Rao (2000) to cases where conditioning regressors are present, and the multivariate conditional distribution of some components (including the individual effects) is left unrestricted.

To end the discussion of identification, we remark that the identification of time-varying errors could be similarly studied in the context of the independent MA models (74), if the within-group information alone is used. Doing so, and as in the case of variances, we would require more restrictive order and rank conditions for identification to hold. The formal analysis is somewhat more involved and is not presented here.

Illustration. Let us consider again Example 2 with $T = 3$, with $\mathbf{s}_i = (1, 0, 0)'$. We assume that errors v_{i1}, v_{i2} and v_{i3} are independent of each other. It can be shown that the levels restrictions on error distributions imply that $\Psi_{v_{i2}}$ and $\Psi_{v_{i3}}$ are identified. Indeed, using (73) we have, for example:

$$\begin{aligned} 3\frac{\partial^2}{\partial t_2^2}\ln\Psi_{v_{i2}}(t_2) - \frac{\partial^2}{\partial t_3^2}\ln\Psi_{v_{i3}}(t_3) &= 3\frac{\partial^2}{\partial t_2^2}\ln\Psi_{y_i}(t_1,t_2,t_3) - 2\frac{\partial^2}{\partial t_2 \partial t_3}\ln\Psi_{y_i}(t_1,t_2,t_3) \\ &\quad -\frac{\partial^2}{\partial t_3^2}\ln\Psi_{y_i}(t_1,t_2,t_3), \\ 3\frac{\partial^2}{\partial t_3^2}\ln\Psi_{v_{i3}}(t_3) - \frac{\partial^2}{\partial t_2^2}\ln\Psi_{v_{i2}}(t_2) &= -\frac{\partial^2}{\partial t_2^2}\ln\Psi_{y_i}(t_1,t_2,t_3) - 2\frac{\partial^2}{\partial t_2 \partial t_3}\ln\Psi_{y_i}(t_1,t_2,t_3) \\ &\quad +3\frac{\partial^2}{\partial t_3^2}\ln\Psi_{y_i}(t_1,t_2,t_3), \end{aligned}$$

from which it follows that $\frac{\partial^2}{\partial t_2^2}\ln\Psi_{v_{i2}}(t_2)$ and $\frac{\partial^2}{\partial t_3^2}\ln\Psi_{v_{i3}}(t_3)$ are identified. Hence, using that $\frac{\partial}{\partial t_2}\ln\Psi_{v_{i2}}(0) = 0$ (as $\mathbf{E}(v_{i2}) = 0$) and $\ln\Psi_{v_{i2}}(0) = 0$, and similarly for v_{i3}, it follows that $\ln\Psi_{v_{i2}}(t_2)$ and $\ln\Psi_{v_{i3}}(t_3)$ are identified.

This discussion shows the identification of the distributions of v_{i2} and v_{i3}.[24] Remark that the distribution of v_{i1} is not identified. This is not surprising as the rank condition for identification (43) is not satisfied in this case. However, the error distribution is identified when errors are assumed stationary.

It is interesting to contrast the identification result using the equations in levels, with the one using only the within-group information (5). In that case, only the distribution of $v_{i3} - v_{i2}$ is identified. So, even in the stationary case, the distribution of v_{i2} (which is equal to that of v_{i3}) is not identified in general. For example, the third-order cumulant $\kappa_3(v_{i2})$ is not identified, as:

$$\kappa_3(v_{i3} - v_{i2}) = \kappa_3(v_{i3}) - \kappa_3(v_{i2}) = 0.$$

[24] In this case, the identification of the distributions of v_{i2} and v_{i3} is also a direct application of a theorem due to Kotlarski (1967).

It can be shown that, if in addition to stationarity error distributions are assumed symmetric around zero, then error distributions are identified.[25]

5 Estimation

In this section we first briefly discuss estimation of parameters and moments of interest, using a i.i.d. sample $\{\mathbf{y}_i, \mathbf{Z}_i, \mathbf{X}_i\}$, $i = 1, ..., N$. Then, we discuss how to estimate densities.

5.1 Common parameters and average effects

We start by discussing the estimation of common parameters and mean effects. Using (30), the optimal moments for $\boldsymbol{\delta}$ and $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i)$ corresponding to model (1) can be written as:

$$\mathbf{E}\left(\begin{array}{c} \mathbf{Z}_i'\mathbf{A}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right) \\ \left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\gamma}\right) \end{array}\right) = \mathbf{0},$$

where $\mathbf{A}_i$ is a $(T-q)\times T$ orthogonal decomposition of $\mathbf{Q}_i = \mathbf{I}_T - \mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'$, and where $\mathbf{V}_i = \mathbf{Var}\left(\mathbf{y}_i|\mathbf{W}_i\right)$. Thus, given any conformable matrix $\boldsymbol{\Psi}_i$, $\boldsymbol{\delta}$ can be estimated as:

$$\widehat{\boldsymbol{\delta}} = \left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{A}_i'\left(\mathbf{A}_i\boldsymbol{\Psi}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i\mathbf{Z}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{A}_i'\left(\mathbf{A}_i\boldsymbol{\Psi}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i\mathbf{y}_i. \tag{75}$$

When $\boldsymbol{\Psi}_i = \mathbf{I}_T$, $\widehat{\boldsymbol{\delta}}$ is the OLS estimator of $\boldsymbol{\delta}$ in the within-group equations (5). When $\boldsymbol{\Psi}_i$ is such that $\left(\mathbf{A}_i\boldsymbol{\Psi}_i\mathbf{A}_i'\right)^{-1} = \left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}$, $\widehat{\boldsymbol{\delta}}$ coincides with the infeasible GLS estimator of $\boldsymbol{\delta}$. To construct a feasible version of $\widehat{\boldsymbol{\delta}}$ that is semiparametric efficient, the quantity $\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i' = \mathbf{E}\left(\mathbf{A}_i\mathbf{v}_i\mathbf{v}_i'\mathbf{A}_i'|\mathbf{W}_i\right)$ needs to be replaced by a consistent estimator. Note that $\mathbf{A}_i\mathbf{v}_i = \mathbf{A}_i\mathbf{y}_i - \mathbf{A}_i\mathbf{Z}_i\boldsymbol{\delta}$. Therefore, this is a standard application of semiparametric GLS as in Robinson (1987).[26]

Likewise, a consistent method-of-moments estimator of $\boldsymbol{\gamma}$ is the weighted mean-group estimator:

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\boldsymbol{\Psi}_i^{-1}\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right). \tag{76}$$

When $\boldsymbol{\Psi}_i = \mathbf{I}_T$, $\widehat{\boldsymbol{\gamma}}$ is simply the mean-group estimator of $\boldsymbol{\gamma}$ (e.g., Hsiao and Pesaran, 2006). In view of the discussion in Appendix C, when $\boldsymbol{\Psi}_i$ is such that $\left(\mathbf{X}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\boldsymbol{\Psi}_i^{-1} =$

[25] This is because, in this case $\Psi_{v_{i2}}(t) = \left[\Psi_{v_{i3}-v_{i2}}(t)\right]^{1/2}$, see Horowitz and Markatou (1996). $\Psi_{v_{i2}}(t)$ is real and strictly positive because, by assumption, v_{i2} is symmetric and $\Psi_{v_{i2}}$ does not vanish.

[26] If $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$ (conditional homoskedasticity of $\mathbf{v}_i$ with respect to $\mathbf{W}_i$), a feasible GLS estimator that replaces $\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'$ with $\mathbf{A}_i\widetilde{\boldsymbol{\Omega}}\mathbf{A}_i'$, where $\widetilde{\boldsymbol{\Omega}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)'$, would be asymptotically efficient.

$\left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}$ the variance matrix of $\widehat{\boldsymbol{\gamma}}$ attains the efficiency bound.[27]

It is instructive to compare the mean-group estimator of $\boldsymbol{\gamma}$ given by (76) with the *pooled OLS* estimator

$$\widetilde{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right).$$

Consistency of $\widetilde{\boldsymbol{\gamma}}$ requires lack of correlation between $\mathbf{X}_i$ and $(\mathbf{X}_i(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}) + \mathbf{v}_i)$. This is true if the individual effects $\boldsymbol{\gamma}_i$ are independent of $\mathbf{X}_i$, but not with correlated effects in general. In contrast, the mean-group estimator $\widehat{\boldsymbol{\gamma}}$ is still consistent when effects and regressors are correlated.

A similar approach may be adopted to deal with Chamberlain's model given by equation (15). A method-of-moment estimator of $\boldsymbol{\theta}$ based on (27) will be consistent. A particular choice for the matrix $\mathbf{Q}_i(\boldsymbol{\theta})$ or its orthogonal decomposition yields semiparametric efficiency (see Appendix C).

Chamberlain (1992) emphasizes an important difference between the linear model (1) and the more general formulation (15). Indeed, in the linear model (1) the estimator $\widehat{\boldsymbol{\delta}}$ coincides with the joint fixed effects estimator of $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_N$, see Cornwell and Schmidt (1987). In contrast, in the nonlinear model (15), the fixed effects estimator of $\boldsymbol{\theta}$ is inconsistent in general, but a method-of-moments estimator based on (27) yields a consistent estimate for $\boldsymbol{\theta}$.[28]

Turning to projection coefficients, Corollary 1 shows that the coefficients estimates obtained when regressing fixed effects estimates:

$$\widehat{\boldsymbol{\gamma}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right),$$

on a set of strictly exogenous regressors $\mathbf{F}_i$, yields consistent estimates for the coefficients of the projection of the population individual effects $\boldsymbol{\gamma}_i$ on the regressors $\mathbf{F}_i$. However, because common parameters $\widehat{\boldsymbol{\delta}}$ have been estimated beforehand, the standard errors of the estimates of the projection coefficients need to be corrected. In particular, this point applies to the mean-group estimator of the unconditional mean $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i)$, given by (76). We provide corrected formulas in Appendix B.

[27] Thus, feasible semiparametric efficient estimation of mean effects requires to estimate the conditional variance $\mathbf{Var}(\mathbf{y}_i|\mathbf{W}_i)$.

[28] The key difference is the dependence of $\mathbf{B}(\mathbf{W}_i, \boldsymbol{\theta})$ on the common parameters. In such a situation we can see from (30) that optimal estimation requires not only estimates of $\mathbf{V}_i$, but also of $\mathbf{E}(\boldsymbol{\gamma}_i|\mathbf{W}_i)$.

Interestingly, the regression-provided R^2 in the regression of $\widehat{\boldsymbol{\gamma}}_i$ on $\mathbf{F}_i$ is inconsistent for the population R^2 in the regression of $\boldsymbol{\gamma}_i$ on $\mathbf{F}_i$, with a downward bias. The reason is that the denominator of the R^2 is the variance of individual effects, which is overestimated by the variance of $\widehat{\boldsymbol{\gamma}}_i$, see (32). In order to compute a correct R^2, we need to consistently estimate the variance of $\boldsymbol{\gamma}_i$, which we discuss next.

5.2 Variances and higher-order moments

Variances. We now turn to estimation of variances under the conditions of Theorem 4, that is under MA-type restrictions on the variance matrix of errors. The extension to autoregressive or ARMA structures presents no difficulty and will not be detailed here. In the following, $\mathbf{A}^-$ denotes any generalized inverse of a full-column rank matrix $\mathbf{A}$, e.g. $\mathbf{A}^- = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$.

The following estimator of the unconditional variance matrix of errors, based on (42), uses covariance restrictions in levels:

$$\mathbf{vec}\left(\widehat{\mathbf{Var}}(\mathbf{v}_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_2\left(\mathbf{M}_i\mathbf{S}_2\right)^-\mathbf{M}_i\left(\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\right), \tag{77}$$

where $\mathbf{M}_i$ is given by (38), and where we have denoted: $\widehat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}$.

Alternatively, the following estimation uses only the within information:

$$\mathbf{vec}\left(\widetilde{\mathbf{Var}}(\mathbf{v}_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_2\left[\left(\mathbf{Q}_i\otimes\mathbf{Q}_i\right)\mathbf{S}_2\right]^-\left(\mathbf{Q}_i\widehat{\mathbf{v}}_i\otimes\mathbf{Q}_i\widehat{\mathbf{v}}_i\right). \tag{78}$$

$\widehat{\mathbf{Var}}(\mathbf{v}_i)$ given by (77) will be consistent as long as (40) is satisfied. In the particular case where errors are i.i.d. with variance σ^2, Corollary 3 motivates estimating σ^2 as:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{N(T-q)}\sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)'\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\\ &= \frac{1}{N(T-q)}\sum_{i=1}^{N}\widehat{\mathbf{v}}_i'\mathbf{Q}_i\widehat{\mathbf{v}}_i.\end{aligned} \tag{79}$$

The first-order asymptotic distributions of (77), (78), and (79) can be easily derived. Standard arguments show that they coincide with the distribution treating common parameters $\boldsymbol{\delta}$ as known. Interestingly, while $\widehat{\sigma}^2$ is non-negative by construction, $\widehat{\mathbf{Var}}(\mathbf{v}_i)$ in (77) and (78) are not necessarily non-negative definite.

Turning to estimation of the variance of individual effects, a consistent estimator based on (32) and (42) is:

$$\begin{aligned}\mathbf{vec}\left(\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right)\right) &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\otimes\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{H}_i\otimes\mathbf{H}_i\right)\mathbf{S}_2\left(\mathbf{M}_i\mathbf{S}_2\right)^{-}\mathbf{M}_i\left[\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\right]. \qquad (80)\end{aligned}$$

Note that, as in the case of the variance of errors, the variance estimator $\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right)$ in (80) is not necessarily non-negative definite.

In the case where errors are i.i.d. but not necessarily homoskedastic, an alternative estimator is:

$$\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)'-\frac{1}{N(T-q)}\sum_{i=1}^{N}\widehat{\mathbf{v}}_i'\mathbf{Q}_i\widehat{\mathbf{v}}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \qquad (81)$$

Lastly, if in addition errors are assumed homoskedastic then we can estimate the variance of $\boldsymbol{\gamma}_i$ by:

$$\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)'-\widehat{\sigma}^2\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}, \qquad (82)$$

where $\widehat{\sigma}^2$ is given by (79). The estimator given by (82) was introduced by Swamy (1970). Note that it is inconsistent in general if v_{it} is conditionally heteroskedastic. In addition, both estimators given by (81) and (82) will be inconsistent if errors are not mutually uncorrelated given regressors.

Remark 1 (testing the covariance structure of errors). In practice, it may be important to empirically determine the order of the MA process of the error terms. This is of special importance in order to estimate the variance of individual effects, as misspecifying the form of the variance matrix of errors would result in inconsistent estimates. This can be done easily using the above results, as we now explain.

Let $\mathbf{S}_2$ be a selection matrix with m columns, and suppose that one wants to test

$$H_0 : \mathbf{vec}\left(\boldsymbol{\Omega}_i\right) = \mathbf{S}_2\boldsymbol{\omega}_i$$

against an unrestricted alternative. Using (39) we have, under H_0:

$$\mathbf{M}_i\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\otimes\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right] = \mathbf{M}_i\mathbf{S}_2\left(\mathbf{M}_i\mathbf{S}_2\right)^{-}\mathbf{M}_i\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\otimes\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{W}_i\right]. \qquad (83)$$

This suggests to consider a test of significance of the following quantity:

$$\widehat{\mathcal{T}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{G}_i\mathbf{M}_i \left(\boldsymbol{I}_{T^2} - \mathbf{S}_2 \left(\mathbf{M}_i\mathbf{S}_2\right)^{-} \mathbf{M}_i\right) \left(\widehat{\mathbf{v}}_i \otimes \widehat{\mathbf{v}}_i\right),$$

where $\mathbf{G}_i$ is a $\left(\frac{T(T+1)}{2} - \frac{q(q+1)}{2}\right) \times T^2$ matrix such that $\mathbf{M}_i\mathbf{D}_T = \mathbf{G}_i'\mathbf{C}_i$, with $\mathbf{D}_T$ the duplication matrix (Magnus and Neudecker, 1988, p.49), and $\mathbf{C}_i$ a full row matrix.[29]

The minimum chi-square statistic then satisfies:

$$\widehat{\mathcal{T}}'\widehat{\mathcal{V}}^{-1}\widehat{\mathcal{T}} \xrightarrow{d} \chi^2_d,$$

where $d = \frac{T(T+1)}{2} - \frac{q(q+1)}{2} - m$, and where the matrix $\widehat{\mathcal{V}}$ depends on fourth-order moments of the data.

This strategy may be interpreted as an extension of the test of covariance structures proposed in Abowd and Card (1989) to random coefficient models. In particular, it is immediate to extend the approach to sequentially test various MA structures, starting with the less restrictive one (e.g., testing MA(q), then MA(q-1), etc...). However, a distinctive feature of our test relative to Abowd and Card is that it also incorporates information in levels (see the discussion in Arellano, 2003, p.67).

Remark 2 (efficient estimation of variances). We have seen in Subsection 3.3 that model (1) with parametric covariance restrictions on errors can be put into the framework of Chamberlain (1992), where the parameters of interest are common parameters, mean and variances of individual effects, and variances of errors.

Guided by the form of the optimal moments, we can consider estimators $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\phi}}\right)$ that solve the following estimating equations (using the notation of Subsection 3.3):

$$\frac{1}{N}\sum_{i=1}^{N} \left\{\frac{\partial}{\partial \boldsymbol{\theta}'}\left[\mathbf{d}\left(\mathbf{W}_i, \boldsymbol{\theta}\right) + \mathbf{R}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\mathbf{h}_i\right]\right\}' \mathbf{A}_i' \left(\mathbf{A}_i\boldsymbol{\Psi}_i\mathbf{A}_i'\right)^{-1} \mathbf{A}_i \left[\mathbf{y}_i^* - \mathbf{d}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\right] = 0$$

for some choice of $\boldsymbol{\Psi}_i$ and $\mathbf{h}_i$. The matrix $\mathbf{A}_i$ depends on $\boldsymbol{\theta}$ and is an orthogonal decomposition of $\mathbf{I} - \mathbf{R}_i\left(\mathbf{R}_i'\mathbf{R}_i\right)^{-1}\mathbf{R}_i'$, where $\mathbf{R}_i$ is a shorthand for $\mathbf{R}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)$.

When $\boldsymbol{\Psi}_i$ is such that $\mathbf{A}_i\boldsymbol{\Psi}_i\mathbf{A}_i' = \mathbf{A}_i\mathbf{Var}\left(\mathbf{y}_i^* | \mathbf{W}_i\right)\mathbf{A}_i'$ and $\mathbf{h}_i = \mathbf{E}\left(\boldsymbol{\gamma}_i^* \mid \mathbf{W}_i\right)$, the estimator $\widehat{\boldsymbol{\theta}}$ attains the asymptotic variance bound. A feasible version will replace population by estimated quantities. In particular, note that the conditional mean $\mathbf{E}\left(\boldsymbol{\gamma}_i^* \mid \mathbf{W}_i\right)$ can be

[29] Note that transformation by $\mathbf{G}_i$ eliminates redundancies.

expressed in terms of observable quantities since:

$$\mathbf{E}\left(\boldsymbol{\gamma}_i^* \mid \mathbf{W}_i\right) = \mathbf{E}\left[\left(\mathbf{R}_i'\mathbf{R}_i\right)^{-1}\mathbf{R}_i'\left(\mathbf{y}_i^* - \mathbf{d}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\right) \mid \mathbf{W}_i\right].$$

Likewise, the optimal moments result suggests estimators of $\boldsymbol{\gamma}^* = \mathbf{E}\left(\boldsymbol{\gamma}_i^*\right)$ of the form

$$\widehat{\boldsymbol{\gamma}}^* = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{R}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{R}_i\right)^{-1}\mathbf{R}_i'\boldsymbol{\Psi}_i^{-1}\left[\mathbf{y}_i^* - \mathbf{d}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\right].$$

The estimator $\widehat{\boldsymbol{\gamma}}^*$ attains the efficiency bound when $\boldsymbol{\Psi}_i$ satisfies:

$$\left(\mathbf{R}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{R}_i\right)^{-1}\mathbf{R}_i'\boldsymbol{\Psi}_i^{-1} = \left(\mathbf{R}_i'\mathbf{Var}\left(\mathbf{y}_i^*|\mathbf{W}_i\right)^{-1}\mathbf{R}_i\right)^{-1}\mathbf{R}_i'\mathbf{Var}\left(\mathbf{y}_i^*|\mathbf{W}_i\right)^{-1}.$$

Remark 3 (higher-order moments). The identification analysis in section 4.1 directly suggests an estimation approach for conditional higher order cumulants of error terms and fixed effects. Using moment restrictions in levels (57) and (58), together with the independent moving-average restrictions (59) and (60), we see that the vectors of third- and fourth-order conditional cumulants of time-varying errors can be estimated as:

$$\begin{aligned}
\widehat{\boldsymbol{\kappa}}_3\left(\mathbf{v}_i|\mathbf{W}_i\right) &= \mathbf{S}_3\left[\mathbf{M}_i^{(3)}\mathbf{S}_3\right]^{-}\mathbf{M}_i^{(3)}\widehat{\boldsymbol{\kappa}}_3\left(\mathbf{y}_i|\mathbf{W}_i\right), \\
\widehat{\boldsymbol{\kappa}}_4\left(\mathbf{v}_{ii}|\mathbf{W}_i\right) &= \mathbf{S}_4\left[\mathbf{M}_i^{(4)}\mathbf{S}_4\right]^{-}\mathbf{M}_i^{(4)}\widehat{\boldsymbol{\kappa}}_4\left(\mathbf{y}_i|\mathbf{W}_i\right),
\end{aligned}$$

where $\mathbf{M}_i^{(3)}$ and $\mathbf{M}_i^{(4)}$ are analogs of $\mathbf{M}_i$ for third- and fourth-order restrictions, respectively. For example: $\mathbf{M}_i^{(3)}$ has rank $T^3 - q^3$ and satisfies $\mathbf{M}_i^{(3)}\left(\mathbf{X}_i \otimes \mathbf{X}_i \otimes \mathbf{X}_i\right) = \mathbf{0}$. In addition, $\widehat{\boldsymbol{\kappa}}_3\left(\mathbf{y}_i|\mathbf{W}_i\right)$ and $\widehat{\boldsymbol{\kappa}}_4\left(\mathbf{y}_i|\mathbf{W}_i\right)$ denote nonparametric estimates of the conditional cumulants of the data.

Third- and fourth-order conditional cumulants of individual effects can be estimated by:

$$\begin{aligned}
\widehat{\boldsymbol{\kappa}}_3\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) &= \widehat{\boldsymbol{\kappa}}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \left(\mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i\right)\mathbf{S}_3\left[\mathbf{M}_i^{(3)}\mathbf{S}_3\right]^{-}\mathbf{M}_i^{(3)}\widehat{\boldsymbol{\kappa}}_3\left(\mathbf{y}_i|\mathbf{W}_i\right), \\
\widehat{\boldsymbol{\kappa}}_4\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) &= \widehat{\boldsymbol{\kappa}}_4\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \left(\mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i \otimes \mathbf{H}_i\right)\mathbf{S}_4\left[\mathbf{M}_i^{(4)}\mathbf{S}_4\right]^{-}\mathbf{M}_i^{(4)}\widehat{\boldsymbol{\kappa}}_4\left(\mathbf{y}_i|\mathbf{W}_i\right),
\end{aligned}$$

where $\widehat{\boldsymbol{\kappa}}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right)$ and $\widehat{\boldsymbol{\kappa}}_4\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right)$ are nonparametric estimates of the conditional cumulants of the fixed-effects estimates.

The unconditional third-order cumulants of error terms can be directly obtained without involving nonparametric conditional expectation terms as follows:

$$\widehat{\boldsymbol{\kappa}}_3\left(\mathbf{v}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_3\left[\mathbf{M}_i^{(3)}\mathbf{S}_3\right]^{-}\mathbf{M}_i^{(3)}\left(\widehat{\mathbf{v}}_i \otimes \widehat{\mathbf{v}}_i \otimes \widehat{\mathbf{v}}_i\right).$$

However, this is not the case of fourth-order cumulants and of cumulants of random coefficients, due to the nonlinearity of their mapping with moments.

Finally, recall that in Subsection 4.1 we obtained the optimal instruments for common parameters and unconditional moments of fixed effects for a model with a parametric specification of the conditional higher order moments of transitory errors. This suggests that at least in this case it is possible to obtain asymptotically efficient estimates of unconditional moments of fixed effects (and therefore also cumulants), which do not depend on nonparametric quantities.

Illustration. Consider again Example 2 with $L = 3$, for a sequence of covariates $s_{i1} = 1$, $s_{i2} = 0$, $s_{i3} = 0$. Assume in addition that $v_{i\ell}$ are i.i.d. Using the within information, only the moments of $v_{i3} - v_{i2} = y_{i3} - y_{i2}$ are identified. So the third-order cumulant of $v_{i\ell}$ is not identified, unless we assume that $v_{i\ell}$ is symmetric (in which case it is zero). The fourth-order cumulant of $v_{i\ell}$ can be estimated by

$$\widehat{\kappa}_4 (v_{i\ell}) = \frac{1}{2}\widehat{\kappa}_4 (y_{i3} - y_{i2}), \tag{84}$$

where the right-hand side in (84) is simply an empirical fourth-order cumulant. Using (84) and the symmetry assumption, one can estimate the cumulants of α_i and β_i.

In this example, it is possible to compute simple estimates of the cumulants of β_i that do not require the symmetry assumption. Indeed, taking first differences we get:

$$\begin{aligned} y_{i1} - y_{i2} &= \beta_i + v_{i1} - v_{i2}, \\ y_{i2} - y_{i3} &= v_{i2} - v_{i3}. \end{aligned}$$

This motivates computing the estimators:

$$\widehat{\kappa}_3 (\beta_i) = \widehat{\kappa}_3 (y_{i1} - y_{i2}) - \widehat{\kappa}_3 (y_{i2} - y_{i3}), \tag{85}$$

$$\widehat{\kappa}_4 (\beta_i) = \widehat{\kappa}_4 (y_{i1} - y_{i2}) - \widehat{\kappa}_4 (y_{i2} - y_{i3}). \tag{86}$$

5.3 Densities

General solutions. Although the main focus of this paper is on identification, in this subsection we discuss ways to estimate the densities of individual effects and errors. A possibility is to estimate the densities of errors and individual effects jointly, for example using sieve maximum likelihood (Ai and Chen, 2003, Hu and Schennach, 2008). A difficulty

with this approach is that one should account for the conditioning on possibly continuous regressors $\mathbf{W}_i$. For this reason, a sequential approach might be preferable.

Starting with error terms, a possibility is to assume a flexible parametric family for errors, for example using normal mixtures.[30] Ghosal and Van der Vaart (2001, 2007) provide results on the ability of normal mixtures to approximate unknown densities. Imposing a flexible parametric structure should not be seen as a severe limitation if the conditions of the identification theorems are satisfied. Note that it is easy to implement maximum likelihood estimation when working with the within-group equations (5). Following this approach, however, it does not seem straightforward to use the information contained in the restrictions in levels (73).

Instead of postulating a parametric model for errors, it may be possible to estimate their densities nonparametrically using characteristic-function based methods that have been proposed in the literature. For example, Horowitz and Markatou (1996) estimate the distribution of errors from within-group equations in a simple model with an individual-specific intercept and symmetric errors, see also Li and Vuong (1998), Hall and Yao (2003), Delaigle *et al.* (2008), and Bonhomme and Robin (2009b) for related approaches in similar or more general models. We are not aware of extensions of these methods to deal with the presence of conditioning covariates.

Once the density of errors (or their characteristic function) has been estimated, there remains to estimate the density of individual effects. The identifying equation (69) of Corollary 5 suggests that one could use kernel deconvolution techniques, replacing the expectation by a sample mean and trimming the integral to ensure convergence. Formally, we could consider

$$\widehat{f}_{\boldsymbol{\gamma}_i}(\boldsymbol{\gamma}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\frac{\exp(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i)}{\widehat{\Psi}_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}K_N(\boldsymbol{\tau})\mathbf{d}\boldsymbol{\tau}, \tag{87}$$

where $\widehat{\Psi}_{\mathbf{v}_i|\mathbf{W}_i}$ is an estimate of the characteristic function of errors, and $K_N(\boldsymbol{\tau})$ is a truncation factor, depending on the sample size N, whose values go to zero when $|\boldsymbol{\tau}|$ tends to infinity. $K_N(\boldsymbol{\tau})$ is supported on a cube $[-T_N, T_N]^q$, where T_N diverges to infinity with N (see Delaigle and Gijbels, 2004, for examples of functions K_N).

There has been considerable work on nonparametric deconvolution in the statistics literature. In standard settings, many estimators are now available: standard Fourier inversion with trimming (Carroll and Hall, 1988, among many other references), wavelets (Fan and

[30] Possibly allowing for conditional heteroskedasticity of a restricted form with respect to the regressors.

Koo, 2002) and recently the Tikhonov-regularization technique of Carrasco and Florens (2007). These estimators have typically low convergence rates, especially if the errors in the regression have *smoother* distributions than the one of the variable to be estimated (Fan, 1991). The smoothness of a distribution refers to the thinness of the tails of its characteristic function: the thinner the tails, the smoother the characteristic function. In cases where errors follow a "supersmooth" distribution such as the normal, asymptotic convergence rates may be as slow as logarithmic. Despite these slow theoretical rates, existing simulation evidence is rather encouraging, especially if the bandwidth or trimming parameters that these estimators require are well chosen (Delaigle and Gijbels, 2004).

However, a potential problem with the deconvolution estimator (87) is that, even if we expect $\widehat{f}_{\boldsymbol{\gamma}_i}$ to converge to the density of individual effects when N gets large, its convergence rate will be governed by the *smoothest* of all the distributions of $\mathbf{H}_i\mathbf{v}_i$ given $\mathbf{W}_i$, $i = 1, ..., N$. So the estimator could behave badly in the presence of strong heteroskedasticity (see Delaigle and Meister, 2008, for a related argument). Modifying and studying nonparametric deconvolution estimators to estimate the distribution of individual effects in model (1) is outside of the scope of this paper. However, in simple cases under more restrictive assumptions, existing estimators can be used for estimation, as we now explain.

A special case. We now discuss estimation of the distributions in the special case of Example 2, which is the setting of our empirical application in the next section:

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + v_{i\ell}, \quad \ell = 1, ..., L. \tag{88}$$

Including strictly exogenous regressors poses no difficulty, as common parameters can be estimated beforehand.

Consider a sequence $\mathbf{s} = (s_1, ..., s_L)'$, and consider all individuals having sequence $\mathbf{s}_i = \mathbf{s}$. If $\mathbf{s}$ consists of L zeros, or of L ones, then α_i and β_i are unidentified. We thus focus on the cases where s's change over time. Define, for $m \in \{-1, 0, 1\}$

$$\mathcal{L}_m(\mathbf{s}) \;=\; \{(i, \ell) \in \{1, ..., N\} \times \{2, ..., L\}, \mathbf{s}_i = \mathbf{s}, \Delta s_{i\ell} = m\},$$

where $\Delta s_{i\ell} = s_{i\ell} - s_{i,\ell-1}$.

We have:

$$\Delta y_{i\ell} = \Delta v_{i\ell}, \quad \text{for all } (i,\ell) \in \mathcal{L}_0(\mathbf{s}), \tag{89}$$

$$\Delta y_{i\ell} = \beta_i + \Delta v_{i\ell}, \quad \text{for all } (i,\ell) \in \mathcal{L}_1(\mathbf{s}), \tag{90}$$

$$-\Delta y_{i\ell} = \beta_i - \Delta v_{i\ell}, \quad \text{for all } (i,\ell) \in \mathcal{L}_{-1}(\mathbf{s}). \tag{91}$$

We assume that errors are i.i.d. given $\mathbf{s}_i = \mathbf{s}$. This implies that all $\Delta v_{i\ell}$, $\ell = 2, ..., L$, have the same distribution. So one can interpret (90) and (91) as simple deconvolution equations, where the left-hand side is the sum of the unobserved β_i, and the independent error $\pm\Delta v_{i\ell}$, and where, because of equation (89), we also observe a random sample from $\Delta v_{i\ell}$.

Having reformulated the problem of estimating the distribution of β_i in model (88) as a simple deconvolution problem, it is now possible to use any existing deconvolution technique to estimate its density nonparametrically. For example, the characteristic function of $\Delta v_{i\ell}$ given $\mathbf{s}_i = \mathbf{s}$ could be estimated as:

$$\widehat{\Psi}_{\Delta v_{i\ell}|\mathbf{s}_i=\mathbf{s}}(\tau|\mathbf{s}_i = \mathbf{s}) = \frac{1}{L_0(\mathbf{s})} \sum_{(i,\ell)\in\mathcal{L}_0(\mathbf{s})} \exp(j\tau\Delta y_{i\ell}),$$

where $L_m(\mathbf{s})$ is the number of observations in $\mathcal{L}_m(\mathbf{s})$. Thus, the characteristic function of β_i given $\mathbf{s}_i = \mathbf{s}$ could be estimated as:[31]

$$\widehat{\Psi}_{\beta_i|\mathbf{s}_i=\mathbf{s}}(\tau|\mathbf{s}_i = \mathbf{s}) = \frac{1}{2} \frac{\frac{1}{L_1(\mathbf{s})} \sum_{(i,\ell)\in\mathcal{L}_1(\mathbf{s})} \exp(j\tau\Delta y_{i\ell}) + \frac{1}{L_{-1}(\mathbf{s})} \sum_{(i,\ell)\in\mathcal{L}_{-1}(\mathbf{s})} \exp(-j\tau\Delta y_{i\ell})}{\frac{1}{L_0(\mathbf{s})} \sum_{(i,\ell)\in\mathcal{L}_0(\mathbf{s})} \exp(j\tau\Delta y_{i\ell})}.$$

Then the density of β_i given $\mathbf{s}_i = \mathbf{s}$ could be recovered by inverse Fourier transformation (with trimming).

In the application below, we will use another approach to estimate the density of β_i. We use a method due to Mallows (2007), which has a number of attractive features. It relies on (simulated) samples rather than on characteristic functions or densities, and thus does not require to select a bandwidth or truncation parameter. In addition, the method is very simple to implement, and it shows a very good behavior in simulation experiments, compared to standard kernel deconvolution. We present the algorithm, along with some illustrative simulations, in Appendix E.

Lastly, note that the approach taken in the case of Example 2 allows us to estimate the density of β_i, but not the density of α_i. It is certainly of interest, in many cases, to estimate

[31] We have used that, because of the i.i.d. assumption, $\Delta v_{i\ell}$ and $-\Delta v_{i\ell}$ have the same distribution given $\mathbf{s}_i = \mathbf{s}$.

the distribution of the effect of a binary treatment, as for example in our application. However, knowing the joint density of (α_i, β_i) is also useful, if only to estimate the distributions of potential outcomes in the model.[32] For this reason, it is of interest to develop extensions of the deconvolution approach that allow to estimate the distribution of individual effects and errors in general linear panel data models like model (1). We are currently extending the approach in Mallows (2007) to accomodate three types of extensions: how to deal with multivariate effects, to treat the error distributions nonparametrically, and how to allow for continuous conditioning covariates. This is done in a companion paper (Arellano and Bonhomme, in progress).

6 Application

6.1 Model and data

We study the effect of smoking during pregnancy on birth outcomes, building on Abrevaya (2006). We estimate the following model:

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + \mathbf{z_{i\ell}}'\boldsymbol{\delta} + v_{i\ell}, \quad i = 1, ..., N, \quad \ell = 1, ..., L, \tag{92}$$

where i and ℓ index mothers and children, respectively.

In this equation, the dependent variable $y_{i\ell}$ is the weight at birth of child ℓ of mother i, $s_{i\ell}$ is the smoking status of mother i when she was pregnant of child ℓ ($s_{i\ell} = 1$ indicating that the mother was smoking), and $\mathbf{z}_{i\ell}$ gathers other determinants of birthweight.

Weight at birth strongly correlates with outcomes later in life. For this reason, the determinants of birthweight have been extensively studied.[33] Abrevaya (2006), using a panel data approach, finds strong negative effects of smoking on birthweights. He assumes that β_i is homogeneous across mothers in (92). Here we take advantage of the panel dimension to account for heterogeneity in the smoking effect.

The parameters α_i and β_i in model (92) are mother-specific effects. They stand for persistent health characteristics of the mother, which could be partly genetic. It is possible

[32] In Example 2, potential outcomes take the form:

$$\begin{aligned} y_{i\ell}(0) &= \alpha_i + v_{i\ell}, \\ y_{i\ell}(1) &= \alpha_i + \beta_i + v_{i\ell}. \end{aligned}$$

[33] See Rosensweig and Wolpin (1991) for a study of various determinants. Studies of the effect of smoking during pregnancy on birthweight are Permutt and Hebel (1989), and Evans and Ringel (1999).

to interpret model (92) as describing a production function, the output of which being the child and the producer being the mother. The production technology is then represented by the mother-specific characteristics α_i and β_i. These characteristics are supposed to stay constant between births. In addition, they may be correlated with smoking status. In particular, a mother could decide not to smoke if she knows that her children will suffer from it (i.e., if she has a very negative β_i).

However, strict exogeneity (Assumption 1) requires that mothers will not change their smoking behavior because one of their children had a low birthweight, as the shocks $v_{i\ell}$ are assumed uncorrelated with the sequence of smoking statuses. This assumption will fail to hold if for example mothers do not know their α_i and β_i before they have had a child, and learning takes place over time. This is a common concern when estimating any type of production function, where there can be feedback effects on the choice of inputs. We will try to relax the strict exogeneity assumption at the end of this section.

Data. We use a sample of mothers from Abrevaya (2006). Abrevaya uses the Natality Data Sets for the US for the years 1990 and 1998. As there are no unique identifiers in these data, he develops a method to match mothers to children, in particular focusing on pairs of states of birth (for mother and child) that have a small number of observation. Abrevaya carefully documents the possible errors caused by this matching strategy. We will use the "matched panel #3", which is likely to be less contaminated by matching error.

This results in a panel dataset where children are matched to mothers. The determinants $\mathbf{z}_{i\ell}$ gathers determinants of birthweights that present between-children variation: the gender of the child, the age of the mother at the time of birth, dummy variables indicating the existence of prenatal visits, and the value of the "Kessner" index of the quality of prenatal care (see Abrevaya, 2006, p.496).

To allow for heterogeneity, we focus on mothers who had at least 3 children during the period (1989-1998). In the dataset, the number of children is exactly 3 for every mother. In addition, we need the smoking indicator $s_{i\ell}$ to vary (at least once) for every mother. So we only consider mothers who changed smoking status between the three births. The final sample contains 1445 mothers.[34]

[34]Descriptive statistics show that this subsample is somewhere in-between the subsample of women who always smoked, and the one of women who never smoked. For example, women who smoke during a larger number of pregnancies are younger on average, and their children have lower weight at birth.

6.2 Results

Testing for heterogeneity. As a preliminary exercise, and to motivate the subsequent results, we start by testing that β_i is heterogeneous in (92). Bonhomme (2008) shows that, although a standard F-test that of the null hypothesis:

$$H_0 : \beta_i = \beta \text{ for all } i = 1, ..., N,$$

is *not* valid under non-normality when N tends to infinity, a simple rescaling of the F statistic is asymptotically distributed as a standard normal under the null. In our case, the F-statistic has a value of 1.32 (for $(1444, 1437)$ degrees of freedom), and the rescaled F-statistic has a value of 4.47. This indicates the presence of heterogeneous β's in the sample we study.

Common parameters. Next, we estimate common parameters $\boldsymbol{\delta}$ in (92). For this, we use the generalized within-group estimator (75), with the identity as weighting matrix. The results are shown in Table 1. Although they have the expected signs, the variables indicating the number of prenatal visits and the quality of prenatal care are never significant. The only significant covariate is the gender of the child, boys having higher birthweight.

Table 1: Estimates of common parameters $\boldsymbol{\delta}$

Variable	Estimate	Standard error
Male	130	22.8
Age	39.0	32.0
Age-sq	-.638	.577
Kessner=2	-82.0	52.7
Kessner=3	-159	81.9
No visit	-18.0	124
Visit=2	83.2	53.9
Visit=3	136	99.2

Note: Estimates of $\boldsymbol{\delta}$ using (75) with $\boldsymbol{\Psi}_i = \mathbf{I}_T$. The dataset is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births (1445 mothers). Standard errors are clustered at the mother level.

Average effects. We now turn to mother-specific effects. Table 2 shows the estimates of the moments of α_i and β_i. The mean smoking effect, computed using the mean-group

formula (76) with the identity as weighting matrix, is -161 grams. This represents a negative and significant effect of smoking on birthweight. Note that this value is close to the fixed-effects estimate obtained by Abrevaya: -144 g, when imposing homogeneity of the β's in model (92). In comparison, the mean of α_i is 2782 g, significant.

Table 2: Moments of α_i and β_i

Moment	Estimate	Standard error
	Means	
Mean α_i	2782	435
Mean β_i	-161	17.0
	Variances (i.i.d. errors)	
Variance α_i	127647	15161
Variance β_i	98239	21674
Covariance (α_i, β_i)	-52661	14375
	Variances (non stationary errors)	
Variance α_i	120423	24155
Variance β_i	85673	34550
Covariance (α_i, β_i)	-45437	24165
	Higher-order moments (within)	
Skewness α_i	-1.67	.428
Skewness β_i	-1.29	.909
Kurtosis α_i	7.12	2.28
Kurtosis β_i	-.34	7.84
	Higher-order moments (first differences)	
Skewness β_i	-1.06	1.25
Kurtosis β_i	7.50	7.10

Note: Estimates of moments of α_i and β_i. The dataset is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births (1445 mothers). See the text for an explanation of the various estimators reported. Standard errors are clustered at the mother level.

To interpret the mother-specific effects, we estimate the projection coefficients in a regression of α_i and β_i on a set of mother-specific characteristics: the education of the mother, her marital status, and the mean of the smoking indicators over the three births. Results are given in Table 3. The coefficient estimates are simply calculated by regressing the fixed-

effects estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ on the mother-specific covariates. Standard errors are corrected as explained in Subsection 5.1.

Table 3 shows that black mothers have children with lower birthweight, however, they seem to be less sensitive to smoking. Also, the children of mothers who smoke more have on average lower birthweights. The R^2 in the regressions are .113 and .021 for α_i and β_i, respectively. This shows that observed covariates explain little of the variation in β_i.[35] One can interpret this finding as a motivation for treating β_i as unobserved mother heterogeneity.

Table 3: Regression of α_i and β_i on mother-specific characteristics

Variable	Estimate	Standard error
	α_i	
High-school	15.1	42.7
Some college	38.5	55.3
College graduate	58.7	72.1
Married	3.51	34.6
Black	-364	54.0
Mean smoking	-161	83.9
Constant	2879	419
	R^2= .113	
	β_i	
High-school	-15.9	42.8
Some college	-15.9	42.8
College graduate	64.5	63.8
Married	31.9	41.8
Black	132	60.6
Mean smoking	-49.8	101
Constant	-172	67.1
	R^2= .021	

Note: Estimates of projection coefficients of of α_i and β_i on mother-specific characteristics. The dataset is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births (1445 mothers). Standard errors are clustered at the mother level.

[35]Remark that the R^2 needs to be corrected, as explained in Subsection 5.1. For comparison, the uncorrected R^2 are .055 and .005 for α_i and β_i, respectively.

Variances. We now turn to variances of mother-specific effects. Rows 3 to 5 in Table 2 show the estimates of the coefficients of the variance matrix of (α_i, β_i) obtained from the levels restrictions, see (80), assuming that errors are i.i.d. given covariates.[36] Given the i.i.d. assumption, the estimates are numerically equal to those using the Swamy formula (82).

Both α_i and β_i show substantial dispersion. In particular, the standard deviation of β_i is 313 g.[37] This can be compared to the standard deviation of 628 g of the least squares estimates $\widehat{\beta}_i$. So in this example, removing the sample noise due to the very small number of observations per mother (3 children) leads to a drastic decrease in the variance. In addition, the estimate of the correlation between α_i and β_i is $-.47$. Given those estimates, the standard deviation of $\alpha_i + \beta_i$ is estimated to be 347 g, compared to 357 g for α_i. This means that the two potential outcomes, for smokers and non smokers, have roughly the same variance.

Having three observations per mother, we need to impose strong restrictions on the variance matrix of errors in order to preserve identification. Using restrictions in levels (39), one can slighty relax the i.i.d. assumption. Rows 6 to 8 in Table 2 show variance estimates under a weaker assumption, which permits the variances of errors for the first, second and third children to be different. As we saw in Subsection 3.2, one cannot leave those three variances unrestricted, however. In rows 6 to 8 we impose that the variance of errors for the jth child is $a + bj$, where a and b are scalars.[38] The results show that the variances of α_i and β_i are not much affected. For example, the standard deviation of β_i is now 292 g. This suggests that the i.i.d. assumption is not rejected on these data.[39]

Higher-order moments. Results for higher-order moments are reported in rows 9 to 14 of Table 2. Rows numbered 9 to 12 in the table show the result of the estimation of skewness and kurtosis under the i.i.d. assumption, using the within-group equations (5). The skewness of errors is not identified from these equations, and we assume that errors are symmetrically distributed. The results show that α_i is negatively skewed and kurtotic, while the skewness and kurtosis of β_i are not significantly different from the ones of the normal distribution (0 and 3, respectively).[40]

[36] Hence, the selection matrix in (80) is $\mathbf{S}_2 = \text{vec}\,(\mathbf{I}_3)$.

[37] Interestingly, when including the number of cigarettes smoked during pregnancy as an additional control, the average smoking effect drops to -135 g, but the standard deviation remains almost unchanged.

[38] Technically, this translates into a different selection matrix $\mathbf{S}_2$ in (80).

[39] We also tried to allow for limited correlation between errors, using (39), and found similar results.

[40] In order to estimate the asymptotic standard errors of higher-order moments we have used the nonparametric bootstrap clustered at the mother level (500 replications).

As we saw in Subsection 5.2, it is possible to compute a simple estimator of the moments of β_i that does not depend on the symmetry assumption, see equations (85) and (86). These estimates (aggregated over smoking sequences) are shown in rows 13 and 14 of Table 2. In that case also, the skewness and kurtosis are not significantly different from those of the normal.

Density and quantiles. Lastly, we present the estimates of the density and quantile function of the smoking effect β_i, estimated using Mallows' (2007) deconvolution algorithm as explained in Subsection 5.3. The results are shown on the left column of Figure 1. For comparison, density and quantile estimates of the least squares estimates $\widehat{\gamma}_i$ are reported on the right column of the figure.

We see that correcting for sample noise in the estimation has strong effects on density and quantile estimates. The density of β_i is much less dispersed than that of $\widehat{\beta}_i$, and its mode is much higher. This last finding is consistent with equation (70) above, which suggested that the density of fixed effects estimates typically underestimates the truth at the mode.

In addition, our approach allows to estimate the smoking effect at various quantiles. When corrected for the presence of sample noise, the effect is mostly negative (up to percentile 75), and reaches very negative values for some mothers (around 400 g at percentile 20). This points to strong heterogeneity in the effect, suggesting that the cost of smoking (in terms of children outcomes) is very high for some mothers.

6.3 Predeterminedness of smoking behavior

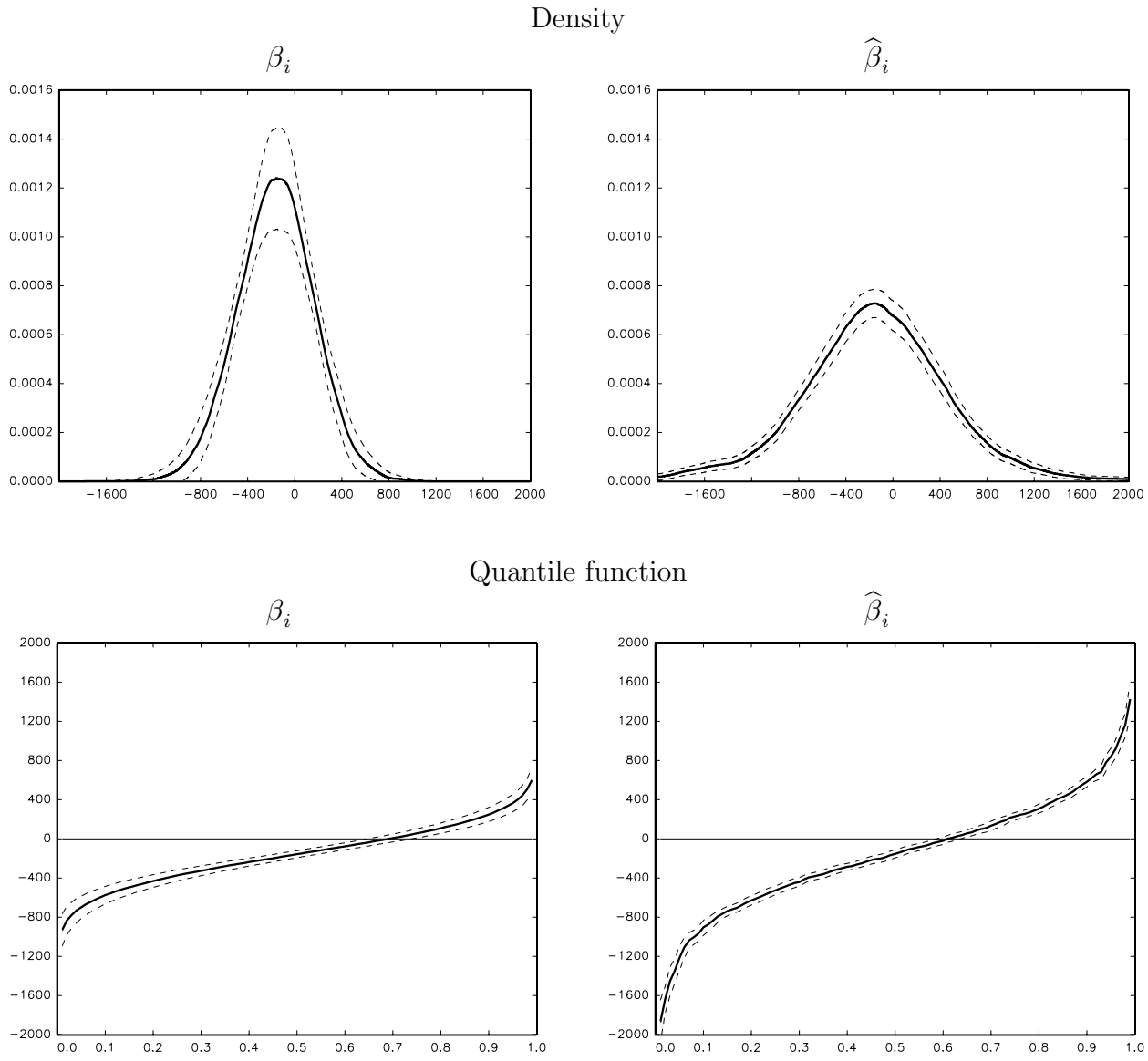
The previous results have been derived under the assumption that the smoking status is strictly exogenous. We now relax the strict exogeneity assumption and assume that smoking is predetermined in model (92), that is:

$$\mathbf{E}\left(v_{i\ell}|\alpha_i, \beta_i, s_{i\ell}, s_{i,\ell-1}, ...\right) = 0. \tag{93}$$

Condition (93) is less restrictive than the strict exogeneity condition (Assumption 1). In particular, (93) could hold in contexts where mothers react to an unexpected birth outcome by changing their smoking behavior.

We consider a simple version of the model without exogenous time-varying regressors. Including time-varying regressors reduces the possibilities of point identification of effects

Figure 1: Density and quantile estimates of the smoking effect



Note: The left column shows the density and quantile function estimates of the smoking effect β_i, obtained using Mallows' (2007) deconvolution algorithm. The right column shows density and quantiles of the fixed effects estimates $\widehat{\beta}_i$. Densities were estimated using a Gaussian kernel with Silverman's rule of thumb for the bandwidth. Thick solid lines represent point estimates, dashed lines show 95% bootstrapped pointwise confidence bands (clustered at the mother level, 300 replications).

of interest, requiring to restrict the correlation between individual effects and regressors. Taking differences between child ℓ and child $m < \ell$ we have:

$$y_{i\ell} - y_{im} = \beta_i \left[s_{i\ell} - s_{im} \right] + v_{i\ell} - v_{im}. \tag{94}$$

It turns out that interesting average effects are point identified in this framework under the predeterminedness condition (93). To see why, remark that, for $k = 0, 1$:

$$\begin{aligned} \mathbf{E}\left(y_{i\ell} - y_{im} | s_{im} = k \right) &= \mathbf{E}\left(\beta_i \left[s_{i\ell} - s_{im} \right] | s_{im} = k \right) + \mathbf{E}\left(v_{i\ell} - v_{im} | s_{im} = k \right) \\ &= \mathbf{E}\left(\beta_i \left[s_{i\ell} - s_{im} \right] | s_{im} = k \right), \end{aligned}$$

where we have used that, because of (93), both $v_{i\ell}$ and v_{im} are mean independent of s_{im}.

Moreover, using that $s_{i\ell}$ can take only two values:

$$\mathbf{E}\left(\beta_i \left[s_{i\ell} - s_{im} \right] | s_{im} = k \right) = (1 - 2k) \Pr\left(s_{i\ell} = 1 - k | s_{im} = k \right) \mathbf{E}\left(\beta_i | s_{im} = k, s_{i\ell} = 1 - k \right).$$

Hence, the following average effects are identified:

$$\mathbf{E}\left(\beta_i | s_{im} = k, s_{i\ell} = 1 - k \right) = (1 - 2k) \frac{\mathbf{E}\left(y_{i\ell} - y_{im} | s_{im} = k \right)}{\Pr\left(s_{i\ell} = 1 - k | s_{im} = k \right)}. \tag{95}$$

Table 4: Average smoking effects under predeterminedness

	Predetermined		Strictly exogenous		
Smoking sequence	Estimate	Standard error	Estimate	Standard error	Number obs.
$(0, 1, .)$	-85.0	43.0	-117	28.9	482
$(1, 0, .)$	-221	36.4	-189	28.8	460
$(., 0, 1)$	-168	38.0	-150	28.0	452
$(., 1, 0)$	-139	45.9	-151	33.9	386
$(0, ., 1)$	-123	33.9	-146	25.8	599
$(1, ., 0)$	-218	37.7	-213	29.3	511

Note: Estimates of the mean of β_i in model (92) without exogenous regressors, for various smoking sequences. For example, $(0, 1, .)$ refers to mothers who did not smoke during the pregnancy of their first child, and smoked while pregnant of their second child. Estimates in column 1 are computed under predeterminedness of the smoking status, while estimates in column 3 are computed under strict exogenity. Standard errors are clustered at the mother level.

We report empirical estimates of (95) in Table 4, for various values of m, ℓ and k. In the same table (column 3), we report the estimates calculated under strict exogeneity.[41] Table

[41] That is, computing the mean of $\widehat{\beta}_i$ on the various sequences of smoking statuses.

4 shows a wide dispersion of average effects estimates between types of smoking sequences. For example, the mean smoking effect is -221 g for mothers who quitted smoking between the first and second child, while it is -85.0 g for mothers who started to smoke during the second pregnancy, the difference between the two estimates being significant at 1%. A similarly striking difference can be observed for women who changed their smoking status between the first and third pregnancies (effects of -218 g and -123 g, respectively). The effects for the second to third pregnancies are not statistically different (see rows 3 and 4).

These findings are consistent with mothers taking into account their own effect of smoking on children outcomes (their β_i) when deciding whether to smoke or not. Moreover, they reinforce the evidence that the smoking effect is heterogeneous across mothers, in a setting where smoking choices are predetermined.

Another interesting result from Table 4 is that, though quantitatively distinct, the results obtained under predeterminedness and strict exogeneity of smoking behavior are qualitatively similar. For example, under strict exogeneity the mean effect is -189 g for mothers who quitted smoking between the first and second child, while it is -117 g for mothers who started to smoke during the second pregnancy, the difference being significant at 5%. Indeed, none of the effects obtained under strict exogeneity is statistically different from the one obtained under predeterminedness (for a given smoking sequence) at the 5% level.[42] This suggests that the strict exogeneity assumption is not unreasonable on these data.

7 Conclusion

Documenting heterogeneity in behavior and response to interventions is one of the main goals of modern econometrics. For this purpose, panel data have an important value-added compared to (single or repeated) cross-sectional data. The reason is that by observing the same units (individuals, households, firms...) over time, it is possible to allow for the presence of unobserved heterogeneity with a clear empirical content. The main goal of this paper has been to derive conditions under which the distribution of heterogeneous components can be consistently estimated in a class of panel data models with multiple sources of heterogeneity.

In many microeconomic applications, it is of interest to estimate the distributions of individual-specific effects. We have provided fixed-T identification results for variances and more generally densities of random coefficients and time-varying errors, in linear panel data

[42] The only significant difference at the 10% level is the one for the sequence $(1, 0, .)$.

models with strictly exogenous regressors. Distributional characteristics of individual effects (other than the mean) are not identified under the assumptions of unrestricted intertemporal distribution of the errors and unrestricted distribution of the effects conditioned on the regressors. In our results we have exploited the identifying content of limited time dependence of time varying errors.

In addition, we have proposed fixed-T consistent estimators of variances, as well as a nonparametric estimator of the density of the individual effect of a binary regressor in a model with i.i.d. errors. Constructing consistent density estimators in more general settings is important. We plan to pursue this task in another paper (Arellano and Bonhomme, in progress).

It is also of interest to relax some of the model's assumptions, in particular strict exogeneity is a concern in many applications. Our analysis of the effect of smoking on birthweight suggests that, in cases where regressors are predetermined instead of strictly exogenous, some average effects may still be point identified. Chernozhukov *et al.* (2009) obtain similar results in some nonlinear panel data models. This seems an interesting route for further research.

References

[1] Aaronson, D., L. Barrow, and W. Sander (2007): "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics*, 25, 95-135.

[2] Abrevaya, J. (2006): "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," *Journal of Applied Econometrics*, vol. 21(4), 489-519.

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

[4] Ahn, S.C., Y.H. Lee, and P. Schmidt (2007): "Panel Data Models with Multiple Time-Varying Effects," unpublished manuscript.

[5] Ai, C., and X. Chen (2003): "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795-1844.

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

[7] 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.

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

[9] Arellano, M., and J. Hahn (2006): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Cambridge University Press.

[10] Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 5, North Holland, Amsterdam.

[11] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229-1279.

[12] Baker, M. (1997): "Growth-rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings," *Journal of Labor Economics,* 15, 338–375.

[13] Beran, R., and Hall, P. (1992): "Estimating Coefficient Distributions in Random Coefficient Regression," *Annals of Statistics*, 20, 1110-1119.

[14] Bonhomme, S. (2008): "A Test of Homogeneity in Random Coefficients Panel Data Models," unpublished manuscript.

[15] Bonhomme, S., and J. M. Robin (2009a): "Consistent Noisy Independent Component Analysis," *Journal of Econometrics*, 149(1), 12-25.

[16] Bonhomme, S., and J. M. Robin (2009b): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," forthcoming *Review of Economic Studies*.

[17] Cameron, C., and P.K. Trivedi (2005): *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.

[18] Carrasco, M., and J.P. Florens (2007): "Spectral Method for Deconvolving a Density," unpublished manuscript.

[19] Carroll, R. J., and P. Hall (1988): "Optimal rates of Convergence for Deconvoluting a Density," *Journal of the American Statistical Association*, 83, 1184-1186.

[20] Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression", *Econometrica*, 60, 567–596.

[21] Chamberlain, G. (1993): "Feedback in Panel Data Models", unpublished manuscript, Department of Economics, Harvard University.

[22] Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2009): "Identification and Estimation of Marginal Effects in Nonlinear Panel Models," unpublished manuscript.

[23] Comon, P. (1994): "Independent Component Analysis, a New Concept?," *Signal Processing*, 36(3), 287-314.

[24] Cornwell, C., and P. Schmidt (1987): "Models for which the MLE and the Conditional MLE Coincide", unpublished manuscript, Michigan State University.

[25] Davidian, M., and D. Zhang (2001): "Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data," *Biometrics*, 57, 795-802.

[26] Delaigle, A., and I. Gijbels (2004): "Comparison of Data-Driven Bandwidth Selection Procedures in Deconvolution Kernel Density Estimation," *Computational Statistics and Data Analysis*, 45, 249-267.

[27] Delaigle, A., P. Hall, and A. Meister (2008): "On Deconvolution with Repeated Measurements," *Annals of Statistics*, 36, 665-685.

[28] Delaigle, A., and A. Meister (2008): "Density Estimation with Heteroscedastic Error," *Bernoulli*, 14, 562-579.

[29] Demidenko, E. (2004): *Mixed Models. Theory and Applications*, John Wiley & Sons.

[30] Dobbelaere, S., and J. Mairesse (2007): "Panel Data Estimates of the Production Function and Product and Labor Market Imperfections", unpublished manuscript.

[31] Evans, W. N., and J. S. Ringel (1999): "Can Higher Cigarette Taxes Improve Birth Outcomes?," *Journal of Public Economics*, 72, 135-154.

[32] Fan, J. Q. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of statistics*, 19, 1257–1272.

[33] Fan, J., and J.Y. Koo (2002): "Wavelet Deconvolution," *IEEE transactions on Information Theory*, Vol. 48, 3, 734-747.

[34] Ghosal, S., and A.W. Van der Vaart (2001): "Rates of Convergence for Bayes and Maximum Likelihood Estimation for Mixture of Normal Densities", *Annals of Statistics*, 29, 1233–1263.

[35] Ghosal, S., and A.W. Van der Vaart (2007): "Posterior Convergence Rates of Dirichlet Mixtures of Normal Distributions at Smooth Densities", *Annals of Statistics*, 35, 697–723.

[36] Graham, B.S., and J.L. Powell (2008): "Identification and Estimation of Irregular Correlated Random Coefficient Models," unpublished manuscript.

[37] Guvenen, F. (2007): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review*, 97, 687–712.

[38] Guvenen, F. (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12, 58-79.

[39] Hahn, J. and W.K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica*, 72, 1295–1319.

[40] Haider, S.J. (2001): "Earnings Instability and Earnings Inequality of Males in the United States," *Journal of Labor Economics,* 19, 799-836.

[41] Hall, P., and Q. Yao (2003): "Inference in Components of Variance Models with Low Replications," *Annals of Statistics*, 31, 414-441.

[42] Heckman, J.J., J.N. Smith, and N. Clements (1997), "Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts," *Review of Economic Studies*, 64, 487-536.

[43] Hoderlein, S., Klemelä, J., and E. Mammen (2007): "Reconsidering the Random Coefficient Model", unpublished manuscript.

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

[45] Horowitz, J. L., and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, 63, 145–168.

[46] Hsiao, C., and H. Pesaran (2006): "Random Coefficient Panel Data Models." In: L. Matyas and P. Sevestre (eds), *The Econometrics of Panel Data*. Kluwer Academic Publishers (forthcoming).

[47] Hu, Y., and S.M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76(1), 195-216.

[48] Hyvärinen, A., J. Karhunen and E. Oja (2001): *Independent Component Analysis*, John Wiley & Sons, New York.

[49] Kleinman, K., and J.G. Ibrahim (1998): "A Semi-Parametric Bayesian Approach to the Random Effects Model," *Biometrics*, 54, 921-938.

[50] Kofidis, E., and P.A. Regalia (2000): "Tensor Approximation and Signal Processing Applications," in: *AMS Conf. on Structured Matrices in Operator Theory, Numerical Analysis, Control, Signal and Image Processing*, AMS Publ.

[51] Kotlarski, I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69-76.

[52] Lesaffre, E, and G. Verbeke (1996): "A linear mixed-effects model with heterogeneity in the random-effects population," *Journal of the American Statistical Association*, 91, 217-221.

[53] Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

[54] Lillard, L., and Y. Weiss (1979): "Components of Variation in Panel Earnings Data: American Scientists, 1960-70," *Econometrica*, Vol.47, 437-454.

[55] Lindgren, B.W. (1993): *Statistical Theory*, Chapman & Hall, New York.

[56] MaCurdy, T. (1981): "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, 89, 1059-1085.

[57] Magnus, J.R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, Chichester.

[58] Mairesse, J., and Z. Griliches (1990): "Heterogeneity in Panel Data: Are there Stable Production Functions?," in: Champsaur, P., Deleau, M., Grandmont, J.M., Laroque, G., Guesnerie, R., Henry, C., Laffont, J.J., Mairesse, J., Monfort, A., Younes, Y. (Eds.), *Essays in Honor of Edmond Malinvaud*, vol. 3, Cambridge, MA: MIT Press.

[59] Mallows, C. (2007): "Deconvolution by Simulation," in: Liu, R., Strawderman, W., and C.H. Zhang (Eds.), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, Beachwood, Ohio, USA: Institute of Mathematical Statistics.

[60] Murtazashvili, I., and J.M. Wooldridge (2008): "Fixed effects instrumental variables estimation in correlated random coefficient panel data models," *Journal of Econometrics*, vol. 142(1), 539-552.

[61] Newey, W. (2004): "Efficient Semiparametric Estimation via Moment Restrictions," *Econometrica*, 72(6), 1877-1897.

[62] Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1–32.

[63] Permutt, T., and J. R. Hebel (1989): "Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight," *Biometrics*, 45, 619-622.

[64] Rao, C.R. (1969): "A Decomposition Theorem for Vector Variables with a Linear Structure," *Annals of Mathematical Statistics,* 40, 1845–1849.

[65] Robinson, P. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 855-891.

[66] Rosensweig, M.R., and K.I. Wolpin (1991): "Inequality at Birth : The Scope for Policy Intervention," *Journal of Econometrics*, 50, 205-228.

[67] Schennach, S. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.

[68] Swamy, P. A. (1970): "Efficient Inference in a Random Coefficient Model," *Econometrica,* 38, 311–323.

[69] Székely, G.J., and C.R. Rao (2000): "Identifiability of Distributions of Independent Random Variables by Linear Combinations and Moments," *Sankhyä*, 62, 193-202.

[70] Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

[71] Wooldridge, J.M. (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *The Review of Economics and Statistics*, vol. 87(2), 385-390.

APPENDIX

A Proofs

A.1 Proofs of Section 3

Proposition 1. Assumption 3 implies that $\mathbf{H}_i$ and $\mathbf{Q}_i$ exist. We have, using (2):

$$\mathbf{E}\left(\mathbf{Q}_i(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})|\mathbf{W}_i\right) = \mathbf{E}\left(\mathbf{Q}_i\mathbf{v}_i|\mathbf{W}_i\right)$$

Likewise, again using assumption (2):

$$\mathbf{E}\left(\mathbf{H}_i(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})|\mathbf{W}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i + \mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right).$$

Corollary 1. Using that $\mathbf{E}(\mathbf{v}_i|\mathbf{W}_i, \mathbf{F}_i) = \mathbf{0}$ it is immediate to see that:

$$\mathbf{E}\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i, \mathbf{F}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i + \mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i, \mathbf{F}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i, \mathbf{F}_i\right).$$

By the law of iterated expectations we obtain:

$$\mathbf{E}\left(\mathbf{F}_i\widehat{\boldsymbol{\gamma}}_i'\right) = \mathbf{E}\left(\mathbf{F}_i\boldsymbol{\gamma}_i'\right).$$

Lastly, (24) implies that $\mathbf{E}\left(\widehat{\boldsymbol{\gamma}}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i\right)$, so:

$$\mathbf{Cov}\left(\mathbf{F}_i, \boldsymbol{\gamma}_i\right) = \mathbf{E}\left(\mathbf{F}_i\boldsymbol{\gamma}_i'\right) - \mathbf{E}\left(\mathbf{F}_i\right)\mathbf{E}\left(\boldsymbol{\gamma}_i'\right) = \mathbf{E}\left(\mathbf{F}_i\widehat{\boldsymbol{\gamma}}_i'\right) - \mathbf{E}\left(\mathbf{F}_i\right)\mathbf{E}\left(\widehat{\boldsymbol{\gamma}}_i'\right) = \mathbf{Cov}\left(\mathbf{F}_i, \widehat{\boldsymbol{\gamma}}_i\right).$$

The conclusion follows.

Corollary 2. Similar to the proof of Proposition 1.

Theorem 1.

$$\begin{aligned}\mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) &= \mathbf{Var}\left(\boldsymbol{\gamma}_i + \mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i\right)\\ &= \mathbf{Var}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) + \mathbf{Var}\left(\mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i\right)\\ &= \mathbf{Var}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right) + \mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\end{aligned}$$

where we have used Assumption 1 in the second equality. Hence (31). Unconditionally we have:

$$\begin{aligned}\mathbf{Var}\left(\boldsymbol{\gamma}_i\right) &= \mathbf{E}\left(\mathbf{Var}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right)\right) + \mathbf{Var}\left(\mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right)\right)\\ &= \mathbf{E}\left[\mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i\right) - \mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\right] + \mathbf{Var}\left(\mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{W}_i\right)\right)\\ &= \mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i\right) - \mathbf{E}\left(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\right).\end{aligned}$$

Corollary 3. Taking the trace in (39) we obtain:

$$\begin{aligned}\mathrm{Tr}\,\mathbf{E}\left[\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)' - \left(\mathbf{I}_T - \mathbf{Q}_i\right)\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)'\left(\mathbf{I}_T - \mathbf{Q}_i\right)|\mathbf{W}_i\right]\\ = \mathrm{Tr}\left(\boldsymbol{\Omega}_i\right) - \mathrm{Tr}\left(\left(\mathbf{I}_T - \mathbf{Q}_i\right)\boldsymbol{\Omega}_i\left(\mathbf{I}_T - \mathbf{Q}_i\right)\right).\end{aligned}$$

In the particular case where errors are i.i.d. independent of $\mathbf{W}_i$ with variance σ^2, this yields:

$$\mathbf{E}\left[\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)'\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right) - \left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)'\left(\mathbf{I}_T - \mathbf{Q}_i\right)\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)\right] = (T-q)\sigma^2,$$

where we have use that $\mathrm{Tr}\left(\mathbf{Q}_i\right) = T - q$. Hence:

$$\sigma^2 = \frac{1}{T-q}\mathbf{E}((\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})'\,\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)).$$

Lemma A1 *Let* $\mathbf{P}$ *be a symmetric idempotent* $n \times n$ *matrix with rank* p*. Let* $\mathbf{D}_n$ *be the* $n^2 \times n(n+1)/2$ duplication *matrix that transforms* $\text{vech}(\mathbf{A})$ *into* $\text{vec}(\mathbf{A})$*, for any* $n \times n$ *matrix* $\mathbf{A}$ *(Magnus and Neudecker, 1988, p.49). Then:*

i) $\quad \text{rank}\left[\left(\mathbf{I}_{n^2} - \mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n\right] = \frac{n(n+1)}{2} - \frac{p(p+1)}{2}.$

ii) $\quad \text{rank}\left\{\left[\left(\mathbf{I}_n - \mathbf{P}\right) \otimes \left(\mathbf{I}_n - \mathbf{P}\right)\right]\mathbf{D}_n\right\} = \frac{(n-p)(n-p+1)}{2},$

Proof. Part *i*). The proof uses results from Magnus and Neudecker (1988, MN hereafter). From MN's Theorem 13 p.49-50 we have:

$$\begin{aligned}\left(\mathbf{I}_{n^2} - \mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n &= \mathbf{D}_n\mathbf{D}_n^-\left(\mathbf{I}_{n^2} - \mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n \\ &= \mathbf{D}_n\left(\mathbf{I}_{\frac{n(n+1)}{2}} - \mathbf{D}_n^-\left(\mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n\right),\end{aligned}$$

where $\mathbf{D}_n^- = \left(\mathbf{D}_n'\mathbf{D}_n\right)^{-1}\mathbf{D}_n'$ denotes the Moore-Penrose generalized inverse of $\mathbf{D}_n$.

Hence, because $\mathbf{D}_n$ has full column rank, the rank of: $\left(\mathbf{I}_{n^2} - \mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n$ is equal to that of: $\mathbf{B}_n = \mathbf{I}_{\frac{n(n+1)}{2}} - \mathbf{D}_n^-\left(\mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n$. But, using equations (14) and (15) in MN (Theorem 13 p.50) it is easy to show that $\mathbf{B}_n$ is idempotent. So, using MN's Theorem 21 (p.20): $\text{rank}\left(\mathbf{B}_n\right) = \text{Tr}\left(\mathbf{B}_n\right)$. Now:

$$\begin{aligned}\text{Tr}\left(\mathbf{D}_n^-\left(\mathbf{P} \otimes \mathbf{P}\right)\mathbf{D}_n\right) &= \text{Tr}\left(\mathbf{D}_n\mathbf{D}_n^-\left(\mathbf{P} \otimes \mathbf{P}\right)\right) \\ &= \frac{1}{2}\text{Tr}\left(\mathbf{P} \otimes \mathbf{P}\right) + \frac{1}{2}\text{Tr}\left(\mathbf{K}_n\left(\mathbf{P} \otimes \mathbf{P}\right)\right) \\ &= \frac{p^2}{2} + \frac{1}{2}\text{Tr}\left(\mathbf{K}_n\left(\mathbf{P} \otimes \mathbf{P}\right)\right),\end{aligned}$$

where $\mathbf{K}_n$ is the *commutation* matrix (MN, p.47). Let $\mathbf{E}_{ij}$ be a $n \times n$ matrix with zeros everywhere, except a one at position (i,j). Let also $\mathbf{P} = \left[p_{ij}\right]_{(i,j)}$.

$$\begin{aligned}\text{Tr}\left(\mathbf{K}_n\left(\mathbf{P} \otimes \mathbf{P}\right)\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{vec}\left(\mathbf{E}_{ij}\right)'\mathbf{K}_n\left(\mathbf{P} \otimes \mathbf{P}\right)\mathbf{vec}\left(\mathbf{E}_{ij}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{vec}\left(\mathbf{E}_{ij}\right)'\mathbf{vec}\left(\mathbf{P}\mathbf{E}_{ij}'\mathbf{P}'\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij}p_{ji} \\ &= \sum_{i=1}^{n}p_{ii} = p,\end{aligned}$$

where the next to last equality comes from idempotence of $\mathbf{P}$. So:

$$\text{Tr}\left(\mathbf{B}_n\right) = \frac{n(n+1)}{2} - \frac{p^2}{2} - \frac{p}{2}.$$

This ends the proof.

Part *ii*). Because of idempotence: $\text{rank}\left(\mathbf{I}_n - \mathbf{P}\right) = n - p$. Let $\mathbf{v}_1, ..., \mathbf{v}_p$ be a basis of the vector space spanned by the columns of $\mathbf{I}_n - \mathbf{P}$. Clearly, $\left\{\mathbf{v}_i \otimes \mathbf{v}_j, (i,j) \in \{1, ..., p\}^2\right\}$ forms a linearly independent family. So does $\left\{\mathbf{v}_i \otimes \mathbf{v}_j, (i,j) \in \{1, ..., p\}^2, i \leq j\right\}$. As this family has $(n-p)(n-p+1)/2$ elements, the conclusion follows.

■

A.2 Proofs of Section 4

Theorem 4. Let $\boldsymbol{\tau} \in \mathbf{R}^q$. Using (6) and Assumption 2 we obtain:

$$\begin{aligned}\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) &= \Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\Psi_{\mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) \\ &= \Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i).\end{aligned}$$

If $\Psi_{\mathbf{v}_i}$ is almost everywhere nonvanishing we obtain (65). Moreover, (66) follows from taking expectations:

$$\begin{aligned}\Psi_{\boldsymbol{\gamma}_i}(\boldsymbol{\tau}) &= \mathbf{E}\left(\Psi_{\boldsymbol{\gamma}_i|\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\right) \\ &= \mathbf{E}\left(\frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\right) \\ &= \mathbf{E}\left(\frac{\mathbf{E}\left(\exp\left(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i\right)|\mathbf{W}_i\right)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\right) \\ &= \mathbf{E}\left(\frac{\exp\left(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i\right)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\right).\end{aligned}$$

Theorem 5. Clearly, because of (43), (73) and Assumption 4: $\boldsymbol{\omega}_i(\mathbf{t})$, $\mathbf{t} \in \mathbf{R}^T$, is identified. Hence $\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{t}|\mathbf{W}_i)$ is identified for all $\mathbf{t} \in \mathbf{R}^T$.

By successive integration and using that, because of Assumption 1:

$$\frac{\partial \ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{0}|\mathbf{W}_i)}{\partial \mathbf{t}} = \mathbf{E}(\mathbf{v}_i|\mathbf{W}_i) = \mathbf{0},$$

and that, because of the definition of a characteristic function:

$$\ln \Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{0}|\mathbf{W}_i) = 0,$$

it follows that the characteristic function of errors is identified.

Corollary 5. Inverse Fourier transformation yields:

$$\begin{aligned}f_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i) &= \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\Psi_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\mathbf{d}\boldsymbol{\tau} \\ &= \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\mathbf{d}\boldsymbol{\tau}.\end{aligned}$$

The unconditional result is similarly obtained.

Proof of equation (70). Under regularity conditions, and provided that $\frac{\mathbf{X}_i'\mathbf{X}_i}{T} \xrightarrow{p} constant > 0$ as T tends to infinity, we have, for all $\boldsymbol{\tau} \in \mathbf{R}^q$:

$$\begin{aligned}\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i) &= \Psi_{\mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i) \\ &= \exp\left[-\frac{1}{2}\boldsymbol{\tau}'\,\mathbf{Var}\left(\mathbf{H}_i\mathbf{v}_i|\mathbf{W}_i\right)\boldsymbol{\tau} + O_p\left(\frac{1}{T^2}\right)\right] \\ &= \exp\left[-\frac{1}{2}\boldsymbol{\tau}'\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\boldsymbol{\tau} + O_p\left(\frac{1}{T^2}\right)\right].\end{aligned}$$

So:

$$\begin{aligned} f_{\boldsymbol{\gamma}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i) &= \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)}{\Psi_{\mathbf{v}_i|\mathbf{W}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{W}_i)}\mathbf{d}\boldsymbol{\tau} \\ &= \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\exp\left[\frac{1}{2}\boldsymbol{\tau}'\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\boldsymbol{\tau}+O_p\left(\frac{1}{T^2}\right)\right]\mathbf{d}\boldsymbol{\tau} \\ &= \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\tau}|\mathbf{W}_i)\left[1+\frac{1}{2}\boldsymbol{\tau}'\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\boldsymbol{\tau}\right]\mathbf{d}\boldsymbol{\tau}+O_p\left(\frac{1}{T^2}\right) \\ &= f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i)-\frac{1}{2}\operatorname{Tr}\left(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\frac{\partial^2 f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{W}_i}(\boldsymbol{\gamma}|\mathbf{W}_i)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right)+O_p\left(\frac{1}{T^2}\right), \end{aligned}$$

where the last equality comes from taking second derivatives in (68).

A lemma. Here we extend Lemma 1 in Bonhomme and Robin (2009a). Consider an independent factor model: $\mathbf{Y}=\boldsymbol{\Lambda}\mathbf{X}$, where $\mathbf{Y}=(Y_1,...,Y_L)'$, $\mathbf{X}=(X_1,...,X_L)'$, $\boldsymbol{\Lambda}$ is a matrix of $L\times S$ parameters (possibly dependent on conditioning covariates), and the S components of the vector $\mathbf{X}$ are independent (also possibly conditionally). Note that L can be less than S. We assume that the variances of $\mathbf{X}_s$ (and thus also of $\mathbf{Y}_\ell$) are finite.

Lemma A2 *Let $(i,j)\in\{1,...,L\}^2$ such that Y_i and Y_j are independent. Then:*

$$\frac{\partial^2 \ln \Psi_{\mathbf{Y}}(\mathbf{t})}{\partial t_i \partial t_j}=0, \quad \mathbf{t}\in\mathbf{R}^L.$$

Proof. We denote the elements of $\boldsymbol{\Lambda}$ as λ_{is}, $i=1,...,L$, $s=1,...,S$. It follows from independence that:

$$\frac{\partial^2 \ln \Psi_{\mathbf{Y}}(\mathbf{t})}{\partial t_i \partial t_j} = \sum_{s=1}^{S}\lambda_{is}\lambda_{js}\left(\frac{\partial^2 \ln \Psi_{X_s}\left(\sum_{i'=1}^{L}\lambda_{i's}t_{i'}\right)}{\partial\tau^2}\right).$$

By the Darmois theorem (Comon, 1994, p.306), as Y_i and Y_j are independent it follows that, for all s, either $\lambda_{is}\lambda_{js}=0$, or X_s is Gaussian.

When X_s is Gaussian: $\frac{\partial^2 \ln \Psi_{X_s}(\sum \lambda_{i's}t_{i'})}{\partial\tau^2}=\frac{\partial^2 \ln \Psi_{X_s}(0)}{\partial\tau^2}$ is constant, independent of $\mathbf{t}$. So we have:

$$\begin{aligned} \frac{\partial^2 \ln \Psi_{\mathbf{Y}}(\mathbf{t})}{\partial t_i \partial t_j} &= \sum_{s=1}^{S}\lambda_{is}\lambda_{js}\left(\frac{\partial^2 \ln \Psi_{X_s}(0)}{\partial\tau^2}\right) \\ &= \operatorname{Cov}(Y_i,Y_j) \\ &= 0. \end{aligned}$$

This end the proof.

■

B Consistent standard errors for the linear projection coefficients

The regression coefficients in:

$$\gamma_{\ell i}=\mathbf{F}_i'\boldsymbol{\pi}_\ell+\xi_{\ell i}, \quad \ell=1,...,q \tag{B1}$$

where $\mathbf{F}_i$ is such that $\mathbf{E}(\mathbf{v}_i|\mathbf{W}_i, \mathbf{F}_i) = \mathbf{0}$, are given by

$$\boldsymbol{\pi}_\ell = \left[\mathbf{E}\left(\mathbf{F}_i\mathbf{F}_i'\right)\right]^{-1}\mathbf{E}\left(\mathbf{F}_i\boldsymbol{\gamma}_{\ell i}\right), \tag{B2}$$

and a root-N-consistent estimator of $\boldsymbol{\pi}_\ell$ is

$$\widehat{\boldsymbol{\pi}}_\ell \;=\; \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i\mathbf{F}_i'\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i\widetilde{\boldsymbol{\gamma}}_{\ell i}, \tag{B3}$$

where, if $\mathbf{h}_{i\ell}'$ denotes the ℓth row of matrix $\mathbf{H}_i$:

$$\widetilde{\boldsymbol{\gamma}}_{\ell i} \;\equiv\; \mathbf{h}_{i\ell}'\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right).$$

We have:

$$\begin{aligned}\widetilde{\boldsymbol{\gamma}}_{\ell i} &= \mathbf{h}_{i\ell}'\left(\mathbf{Z}_i\boldsymbol{\delta} + \mathbf{X}_i\boldsymbol{\gamma}_i + \mathbf{v}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\\ &= \mathbf{F}_i'\boldsymbol{\pi}_\ell + \xi_{\ell i} - \mathbf{h}_{i\ell}'\mathbf{Z}_i\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) + \mathbf{h}_{i\ell}'\mathbf{v}_i.\end{aligned}$$

Hence, letting $\boldsymbol{\Psi}_N = N^{-1}\sum_{i=1}^{N}\mathbf{F}_i\mathbf{F}_i'$ we have

$$\boldsymbol{\Psi}_N\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i\xi_{\ell i}\right) - \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i\mathbf{h}_{i\ell}'\mathbf{Z}_i\right)\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) + \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_i\mathbf{h}_{i\ell}'\mathbf{v}_i\right).$$

Also

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Q}_i\mathbf{Z}_i\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Q}_i\mathbf{v}_i. \tag{B4}$$

It is easily shown (e.g., Wooldridge, 2002, p.321 for a special case) that a consistent estimator of $\mathbf{Avar}\left[\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right)\right]$ is given by:

$$\boldsymbol{\Psi}_N^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{a}_i\mathbf{a}_i'\right)\boldsymbol{\Psi}_N^{-1},$$

where

$$\mathbf{a}_i = \mathbf{F}_i\left(\mathbf{h}_{i\ell}'\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right) - \mathbf{F}_i'\widehat{\boldsymbol{\pi}}_\ell\right) - \left(\sum_{j=1}^{N}\mathbf{F}_j\mathbf{h}_{j\ell}'\mathbf{Z}_j\right)\left(\sum_{j=1}^{N}\mathbf{Z}_j'\mathbf{Q}_j\mathbf{Z}_j\right)^{-1}\mathbf{Z}_i'\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right).$$

C Computing Chamberlain's semiparametric bound

Model and notation. Consider the general panel model that is linear in fixed effects but nonlinear in variables and common parameters:

$$\mathbf{y}_i = \mathbf{a}\left(\mathbf{W}_i, \boldsymbol{\theta}\right) + \mathbf{B}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\boldsymbol{\gamma} + \mathbf{B}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)\boldsymbol{\varepsilon}_i + \mathbf{v}_i$$

$$\mathbf{E}\left(\mathbf{v}_i \mid \mathbf{W}_i, \boldsymbol{\gamma}_i\right) = \mathbf{0}, \qquad \mathbf{E}\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{0},$$

where $\boldsymbol{\gamma} = \mathbf{E}\left(\boldsymbol{\gamma}_i\right)$ and $\boldsymbol{\varepsilon}_i = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}$. For shortness, write $\mathbf{B}_i = \mathbf{B}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)$ and $\mathbf{a}_i = \mathbf{a}\left(\mathbf{W}_i, \boldsymbol{\theta}\right)$. Moreover, let $\mathbf{Var}\left(\mathbf{y}_i \mid \mathbf{W}_i\right) = \mathbf{V}_i$, $\mathbf{Var}\left(\mathbf{v}_i \mid \mathbf{W}_i\right) = \boldsymbol{\Omega}_i$, and $\mathbf{Var}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{W}_i\right) = \boldsymbol{\Sigma}_i$. Thus,

$$\mathbf{V}_i = \mathbf{B}_i\boldsymbol{\Sigma}_i\mathbf{B}_i' + \boldsymbol{\Omega}_i$$

The interest is in the optimal estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ following Chamberlain (1992).

Optimal estimation of common parameters. Define the idempotent matrix

$$\mathbf{Q}_i = \mathbf{I}_T - \mathbf{B}_i \left(\mathbf{B}_i'\mathbf{B}_i\right)^{-1} \mathbf{B}_i'$$

and let $\mathbf{A}_i$ be a $(T-q) \times T$ semi-triangular matrix such that $\mathbf{Q}_i = \mathbf{A}_i'\mathbf{A}_i$ and $\mathbf{A}_i\mathbf{A}_i' = \mathbf{I}_{T-q}$.

All information about $\boldsymbol{\theta}$ is contained in the $(T-q)$ conditional moments

$$\mathbf{E}\left(\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) \mid \mathbf{W}_i\right) = \mathbf{0}.$$

The conditional variance matrix of the transformed residuals is

$$\mathbf{E}\left[\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right)\left(\mathbf{y}_i - \mathbf{a}_i\right)'\mathbf{A}_i' \mid \mathbf{W}_i\right] = \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i' = \mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'.$$

The corresponding optimal instruments are

$$\mathbf{E}\left[\mathbf{D}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\left(\mathbf{A}_i\mathbf{y}_i - \mathbf{A}_i\mathbf{a}_i\right)\right] = \mathbf{0},$$

where

$$\mathbf{D}_i = \mathbf{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}'}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) \mid \mathbf{W}_i\right].$$

We show below that

$$\mathbf{D}_i = -\mathbf{A}_i\left(\frac{\partial\mathbf{a}_i}{\partial\boldsymbol{\theta}'} + \sum_{j=1}^{q}\frac{\partial\mathbf{b}_{ji}}{\partial\boldsymbol{\theta}'}\mathbf{E}\left(\gamma_{ji} \mid \mathbf{W}_i\right)\right), \tag{C5}$$

where $\mathbf{B}_i = \left(\mathbf{b}_{1i}, ..., \mathbf{b}_{qi}\right)$, and $\boldsymbol{\gamma}_i = \left(\gamma_{1i}, ..., \gamma_{qi}\right)'$. Therefore, the optimal moment for $\boldsymbol{\theta}$ is

$$\mathbf{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left[\mathbf{a}_i + \mathbf{B}_i\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right)\right]'\mathbf{A}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\left(\mathbf{A}_i\mathbf{y}_i - \mathbf{A}_i\mathbf{a}_i\right)\right] = \mathbf{0}. \tag{C6}$$

Proof of (C5). We need $\partial\mathbf{A}_i/\partial\theta_k$. First note that the partial derivatives of $\mathbf{Q}_i$ are given by

$$\frac{\partial\mathbf{Q}_i}{\partial\theta_k} = -\mathbf{Q}_i\frac{\partial\mathbf{B}_i}{\partial\theta_k}\left(\mathbf{B}_i'\mathbf{B}_i\right)^{-1}\mathbf{B}_i' - \mathbf{B}_i\left(\mathbf{B}_i'\mathbf{B}_i\right)^{-1}\frac{\partial\mathbf{B}_i'}{\partial\theta_k}\mathbf{Q}_i. \tag{C7}$$

To see the connection between $\mathbf{dQ}_i$ and $\mathbf{dA}_i$ note that

$$\mathbf{dQ}_i = \mathbf{A}_i'\left(\mathbf{dA}_i\right) + \left(\mathbf{dA}_i'\right)\mathbf{A}_i$$

$$\left(\mathbf{dA}_i\right)\mathbf{A}_i' + \mathbf{A}_i\left(\mathbf{dA}_i'\right) = \mathbf{0},$$

so that

$$\mathbf{A}_i\mathbf{dQ}_i = \left(\mathbf{dA}_i\right) + \mathbf{A}_i\left(\mathbf{dA}_i'\right)\mathbf{A}_i = \left(\mathbf{dA}_i\right) - \left(\mathbf{dA}_i\right)\mathbf{A}_i'\mathbf{A}_i = \left(\mathbf{dA}_i\right)\mathbf{B}_i\left(\mathbf{B}_i'\mathbf{B}_i\right)^{-1}\mathbf{B}_i'.$$

Post-multiplying by $\mathbf{B}_i$, the partial derivatives satisfy

$$\mathbf{A}_i\frac{\partial\mathbf{Q}_i}{\partial\theta_k}\mathbf{B}_i = \frac{\partial\mathbf{A}_i}{\partial\theta_k}\mathbf{B}_i.$$

Finally, inserting (C7) and noting that $\mathbf{A}_i\mathbf{B}_i = \mathbf{0}$ it turns out that

$$\frac{\partial\mathbf{A}_i}{\partial\theta_k}\mathbf{B}_i = -\mathbf{A}_i\frac{\partial\mathbf{B}_i}{\partial\theta_k}. \tag{C8}$$

Now, to see that (C5) holds note that

$$\mathbf{D}_i = \mathbf{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}'}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) \mid \mathbf{W}_i\right] = \mathbf{E}\left[\begin{array}{ccc} \frac{\partial}{\partial \theta_1}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) & \cdots & \frac{\partial}{\partial \theta_K}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right), \end{array} \mid \mathbf{W}_i\right]$$

and using (C8) we obtain the k-th column of $\mathbf{D}_i$ as follows

$$\begin{aligned}
\mathbf{E}\left[\frac{\partial}{\partial \theta_k}\mathbf{A}_i\left(\mathbf{y}_i - \mathbf{a}_i\right) \mid \mathbf{W}_i\right] &= \left(\frac{\partial \mathbf{A}_i}{\partial \theta_k}\right)\mathbf{E}\left(\mathbf{y}_i - \mathbf{a}_i \mid \mathbf{W}_i\right) - \mathbf{A}_i\left(\frac{\partial \mathbf{a}_i}{\partial \theta_k}\right) \\
&= \left(\frac{\partial \mathbf{A}_i}{\partial \theta_k}\right)\mathbf{E}\left(\mathbf{B}_i\boldsymbol{\gamma}_i + \mathbf{v}_i \mid \mathbf{W}_i\right) - \mathbf{A}_i\left(\frac{\partial \mathbf{a}_i}{\partial \theta_k}\right) \\
&= \left(\frac{\partial \mathbf{A}_i}{\partial \theta_k}\right)\mathbf{B}_i\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right) - \mathbf{A}_i\left(\frac{\partial \mathbf{a}_i}{\partial \theta_k}\right) \\
&= -\mathbf{A}_i\left(\frac{\partial \mathbf{a}_i}{\partial \theta_k} + \frac{\partial \mathbf{B}_i}{\partial \theta_k}\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right)\right) \\
&= -\mathbf{A}_i\left(\frac{\partial \mathbf{a}_i}{\partial \theta_k} + \sum_{j=1}^{q}\frac{\partial \mathbf{b}_{ji}}{\partial \theta_k}\mathbf{E}\left(\gamma_{ji} \mid \mathbf{W}_i\right)\right).
\end{aligned}$$

Optimal estimation of expected fixed effects. Using matrix inversion formulas, we obtain the following expressions linking $\mathbf{V}_i^{-1}$ and $\boldsymbol{\Omega}_i^{-1}$, which will be used below:

$$\boldsymbol{\Omega}_i^{-1} = \mathbf{V}_i^{-1} + \mathbf{V}_i^{-1}\mathbf{B}_i\left(\boldsymbol{\Sigma}_i^{-1} - \mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}$$

$$\left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1} = \boldsymbol{\Sigma}_i + \left(\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i\right)^{-1}.$$

Suppose for the sake of the argument that $\boldsymbol{\theta}$ is known so that $\mathbf{w}_i = \mathbf{y}_i - \mathbf{a}_i$ and $\mathbf{B}_i$ are observable. The model implies the following moments:

$$\mathbf{E}\left[\left(\mathbf{C}_i'\mathbf{B}_i\right)^{-1}\mathbf{C}_i'\left(\mathbf{w}_i - \mathbf{B}_i\boldsymbol{\gamma}\right)\right] = \mathbf{0},$$

for some $\mathbf{C}_i$. So we consider the asymptotic distribution of estimators of the form

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{C}_i'\mathbf{B}_i\right)^{-1}\mathbf{C}_i'\mathbf{w}_i.$$

The scaled estimation error satisfies

$$\sqrt{N}\left(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mathbf{C}_i'\mathbf{B}_i\right)^{-1}\mathbf{C}_i'\mathbf{v}_i \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Upsilon}\right),$$

where

$$\boldsymbol{\Upsilon} = \mathbf{E}\left(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right) + \mathbf{E}\left[\left(\mathbf{C}_i'\mathbf{B}_i\right)^{-1}\mathbf{C}_i'\boldsymbol{\Omega}_i\mathbf{C}_i\left(\mathbf{B}_i'\mathbf{C}_i\right)^{-1}\right].$$

An optimal choice of $\mathbf{C}_i$ satisfies

$$\left(\mathbf{C}_i'\mathbf{B}_i\right)^{-1}\mathbf{C}_i' = \left(\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1},$$

which leads to

$$\boldsymbol{\Upsilon} = \mathbf{E}\left(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right) + \mathbf{E}\left[\left(\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i\right)^{-1}\right], \tag{C9}$$

or

$$\boldsymbol{\Upsilon} = \mathbf{Var}\left[\mathbf{E}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{W}_i\right)\right] + \mathbf{E}\left[\left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\right]. \tag{C10}$$

One optimal choice is $\mathbf{C}_i' = \mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}$. To characterize the range of optimal choices, let us define $\boldsymbol{\Psi}_i$ for some $q \times q$ matrix $\mathbf{K}_i \geq \mathbf{0}$ such that:

$$\boldsymbol{\Psi}_i^{-1} = \mathbf{V}_i^{-1} + \mathbf{V}_i^{-1}\mathbf{B}_i\mathbf{K}_i\left[\mathbf{I} - \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)\mathbf{K}_i\right]^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}$$

Note that setting $\mathbf{K}_i = \boldsymbol{\Sigma}_i$ we have $\boldsymbol{\Psi}_i = \boldsymbol{\Omega}_i$. However, while $\boldsymbol{\Psi}_i$ depends on $\mathbf{K}_i$ the quantity $\left(\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}$ does not:[43]

$$\left(\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1} = \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1} = \left(\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Omega}_i^{-1}$$

The conclusion is that an optimal moment uses $\mathbf{C}_i' = \mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}$, and all optimal instruments of the form $\left(\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}$ are the same regardless of the value of $\mathbf{K}_i$. Thus, we can set $\mathbf{K}_i = \mathbf{0}$ without lack of generality and use $\mathbf{C}_i' = \mathbf{B}_i'\mathbf{V}_i^{-1}$.

Therefore, the form of an estimator that attains the bound is

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{w}_i,$$

which is numerically identical for all permissible values of $\mathbf{K}_i$.

The optimal moment conditions for $\boldsymbol{\gamma}$ can be written as

$$\mathbf{E}\left[\left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}\left(\mathbf{w}_i - \mathbf{B}_i\boldsymbol{\gamma}\right)\right] = \mathbf{0}. \tag{C11}$$

Joint optimal moments: system GMM. It is easy to see that the optimal moments for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, (C6) and (C11) respectively, are uncorrelated:

$$\mathbf{E}\left(\left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}\left(\mathbf{B}_i\boldsymbol{\varepsilon}_i + \mathbf{v}_i\right)\mathbf{v}_i'\mathbf{A}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i\frac{\partial}{\partial\boldsymbol{\theta}'}\left[\mathbf{a}_i + \mathbf{B}_i\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right)\right]\right) = \mathbf{0}.$$

Therefore, the optimal moments for estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are:

$$\mathbf{E}\left(\begin{array}{c}\frac{\partial}{\partial\boldsymbol{\theta}}\left[\mathbf{a}_i + \mathbf{B}_i\mathbf{E}\left(\boldsymbol{\gamma}_i \mid \mathbf{W}_i\right)\right]'\mathbf{A}_i'\left(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'\right)^{-1}\left(\mathbf{A}_i\mathbf{y}_i - \mathbf{A}_i\mathbf{a}_i\right) \\ \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{a}_i - \mathbf{B}_i\boldsymbol{\gamma}\right)\end{array}\right) = \mathbf{0}.$$

D Multivariate cumulants and characteristic functions

Here we collect some standard definitions and properties of cumulants and characteristic functions that are used in the paper, with a view to make the discussion as self-contained as possible.

Cumulants. Let $\mathbf{U} = (U_1, ..., U_n)'$ be an n-dimensional random vector with zero mean and well-defined moments to the fourth-order. We define its *cumulant vector of order 3* as the n^3-dimensional vector $\boldsymbol{\kappa}_3(\mathbf{U})$ whose elements $\kappa_3^{i,j,k}(\mathbf{U})$, for $(i,j,k) \in \{1, ..., n\}^3$, are arranged in lexicographic order and are such that

$$\kappa_3^{i,j,k}(\mathbf{U}) = \mathbf{E}\left(U_iU_jU_k\right), \quad (i,j,k) \in \{1, ..., n\}^3.$$

[43] Note that

$$\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1} = \left[\mathbf{I} - \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)\mathbf{K}_i\right]^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}$$

$$\mathbf{B}_i'\boldsymbol{\Psi}_i^{-1}\mathbf{B}_i = \left[\mathbf{I} - \left(\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i\right)\mathbf{K}_i\right]^{-1}\mathbf{B}_i'\mathbf{V}_i^{-1}\mathbf{B}_i.$$

Likewise, we define $\boldsymbol{\kappa}_4(\mathbf{U})$ whose n^4 elements are

$$\begin{aligned}\kappa_4^{i,j,k,\ell}(\mathbf{U}) &= \mathbf{E}\left(U_iU_jU_kU_\ell\right) - \mathbf{E}\left(U_iU_j\right)\mathbf{E}\left(U_kU_\ell\right) \\ &\quad -\mathbf{E}\left(U_iU_k\right)\mathbf{E}\left(U_jU_\ell\right) - \mathbf{E}\left(U_iU_\ell\right)\mathbf{E}\left(U_jU_k\right), \quad (i,j,k,\ell) \in \{1,...,n\}^4.\end{aligned}$$

For a nonzero mean random vector $\mathbf{V}$, we define $\boldsymbol{\kappa}_3(\mathbf{V}) = \boldsymbol{\kappa}_3\left(\mathbf{V} - \mathbf{E}\left(\mathbf{V}\right)\right)$, and we similarly define $\boldsymbol{\kappa}_4(\mathbf{V})$.

The *skewness* of U_j ($i \in \{1,...,n\}$) and its *kurtosis* are given by: $\kappa_3^{j,j,j}(\mathbf{U})/\operatorname{Var}(U_j)^{3/2}$ and $\left[\kappa_4^{j,j,j,j}(\mathbf{U})/\operatorname{Var}(U_j)^2\right]+3$, respectively. We may similarly define conditional cumulants by replacing the expectations in these formulas by conditional expectations.

Cumulants satisfy a multilinearity property, and can be interpreted as tensors (Kofidis and Regalia, 2000). Namely, for any $s \times n$ matrix $\mathbf{A}$ we have:

$$\begin{aligned}\boldsymbol{\kappa}_3(\mathbf{AU}) &= (\mathbf{A}\otimes\mathbf{A}\otimes\mathbf{A})\,\boldsymbol{\kappa}_3(\mathbf{U}), \\ \boldsymbol{\kappa}_4(\mathbf{AU}) &= (\mathbf{A}\otimes\mathbf{A}\otimes\mathbf{A}\otimes\mathbf{A})\,\boldsymbol{\kappa}_4(\mathbf{U}).\end{aligned}$$

Moreover, cumulants of the sums of *independent* random variables satisfy: $\boldsymbol{\kappa}_3(\mathbf{U}+\mathbf{V}) = \boldsymbol{\kappa}_3(\mathbf{U}) + \boldsymbol{\kappa}_3(\mathbf{V})$, and: $\boldsymbol{\kappa}_4(\mathbf{U}+\mathbf{V}) = \boldsymbol{\kappa}_4(\mathbf{U}) + \boldsymbol{\kappa}_4(\mathbf{V})$. Because of these properties, it is sometimes more convenient to work with cumulants than with moments, although there exists a mapping between the two.

Here we have only defined cumulants of order 3 and 4. We could easily generalize these results to cumulants of order 5 or higher. The first-order cumulant is simply the mean, and the cumulants of order 2 are the variances and covariances.

Characteristic functions. Let $(\mathbf{Y},\mathbf{X})$ be a pair of random vectors, $\mathbf{Y} \in \mathbf{R}^L$, and let j be a square root of -1. The conditional characteristic function of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, is defined as:

$$\Psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = \mathbf{E}\left(\exp(j\mathbf{t}'\mathbf{Y})|\mathbf{x}\right), \quad \mathbf{t} \in \mathbf{R}^L.$$

We make use of the following properties of characteristic functions in the paper (e.g., Lindgren, 1993, p.128-131). First, there exists a mapping between the (conditional) characteristic function and the (conditional) density, the so-called *inverse Fourier transform*:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^L}\int \exp\left(-j\mathbf{t}'\mathbf{y}\right)\Psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x})\mathbf{dt}. \tag{D12}$$

This means that all the information about a random variable is contained in its characteristic function. Second, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent given $\mathbf{X}$ then:

$$\Psi_{\mathbf{Y}_1+\mathbf{Y}_2|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = \Psi_{\mathbf{Y}_1|\mathbf{X}}(\mathbf{t}|\mathbf{x})\Psi_{\mathbf{Y}_2|\mathbf{X}}(\mathbf{t}|\mathbf{x}). \tag{D13}$$

Lastly, cumulants (when they exist) can be obtained from the successive derivatives of the logarithm of the characteristic function (also called cumulant generating function) evaluated at $\mathbf{t} = \mathbf{0}$.

E Mallows' algorithm (2007)

The algorithm. The model is:

$$A_i = B_i + C_i,$$

where B_i and C_i are independent of each other. Two unrelated random samples from A_i and C_i are available, that we denote as $\mathbf{A}$ and $\mathbf{C}$, respectively. We assume that $\mathbf{A}$ and $\mathbf{C}$ are sorted in ascending order. The objective of the algorithm is to draw approximate random samples from B_i.

The algorithm is as follows.

1. Start with $\mathbf{B}_0 = \text{sort}\{\mathbf{A} - \mathbf{C}\}$.
2. Start step one. Permute $\mathbf{B}_0$ randomly, this yields $\widetilde{\mathbf{B}}_0$.
3. Let $\widetilde{\mathbf{A}}_0$ be the permutation of $\mathbf{A}$ sorted according to $\widetilde{\mathbf{B}}_0 + \mathbf{C}$.
4. Set $\mathbf{B}_1 = \text{sort}\{\widetilde{\mathbf{A}}_0 - \mathbf{C}\}$. Go to step two.

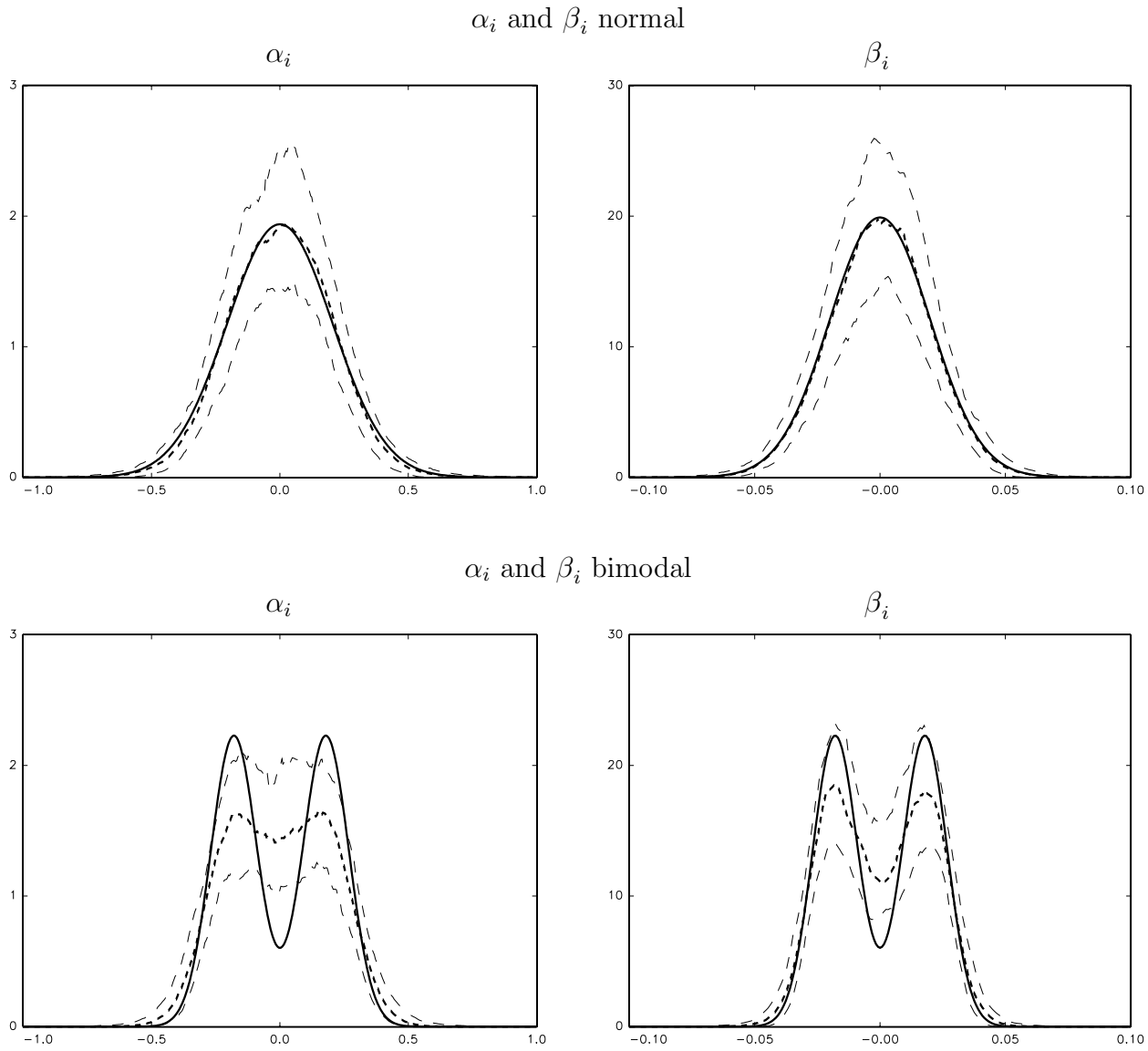
In our experiments, the algorithm always converged to a stationary chain after a short "burn-in" period. In practice, we removed the first 500 initial iterations out of total of 2000.

Lastly, note that, for this algorithm to work, $\mathbf{A}$ and $\mathbf{C}$ must have the same size. If this is not the case, one may replace them by m bootstrap draws with replication from $\mathbf{A}$ and $\mathbf{C}$, respectively, where m is the desired common size. In the application $\mathbf{A}$ is twice the size of $\mathbf{C}$. We simply used the stacked vector $[\mathbf{C}', \mathbf{C}']'$ instead of $\mathbf{C}$.

Illustration. We here briefly present some simulation results, which suggest that Mallows' algorithm works well in practice. The Monte-Carlo design is the random trend model of Example 1, with stationary AR(1) errors. The parameter values are chosen to roughly replicate the results by Guvenen (2009) on PSID data: T is 20, N is 1000, ρ is .80, the variances of α_i and β_i are .04 and .0004, respectively, their covariance is $-.001$, and u_{it} has variance .02. Lastly, we use two different designs for the marginal distributions of α_i and β_i: the normal, and a symmetric bimodal normal mixture with two components.

We apply Mallows' algorithm to equations (8) and (9). We use a random sample from the errors $\mathbf{v}_i$. The densities of α_i and β_i are then estimated using a Gaussian kernel with a rule-of-thumb bandwidth. Figure E1 shows the results of 100 simulations. We observe that the estimator is unbiased in the case where α_i and β_i are normal. Interestingly, the confidence bands are very thin in the tails of the density. In the bimodal case, the estimation is somewhat worse. However, the estimator succeeds at replicating the bimodality of the latent variables.

Figure E1: Density estimates on simulated data using Mallows' algorithm



Note: α_i and β_i are obtained using Mallows' (2007) deconvolution algorithm. The DGP is that of Example 1 with parameters roughly chosen to match Guvenen (2009). Thick line is the pointwise median across simulations, dashed lines are the 10%-90% pointwise confidence bands.