Identifying Distributional Characteristics in Random Coefficients Panel Data Models

Manuel Arellano
CEMFI, Madrid

Stéphane Bonhomme
CEMFI, Madrid

This draft: September 2008

Abstract

We study the identification of linear panel data models with strictly exogenous regressors and individual-specific coefficients, when the time length of the panel T is fixed. In addition to common parameters and averages of individual effects, we show the identification of the variance of the effects under conditional uncorrelatedness assumptions on error variables. Identification requires the dependence structure of errors to be restricted, reflecting a trade-off between the number of individual-specific parameters and error dynamics. Assuming that effects and errors are independent conditional on regressors, we show the identification of the density of individual effects in cases where errors follow moving averages or ARMA structures with independent innovations. We propose method-of moments estimators of the moments of individual effects and errors, and introduce a simple estimator of the density of the effect of a binary regressor in a special case. We apply the method to estimate the effect that a mother smokes during pregnancy on the weight of her child at birth.

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

1 Introduction

Documenting heterogeneity in behavior and response to interventions is one of the main goals of modern econometrics. For this purpose, compared to cross-sectional data, panel data has an important value-added as it allows to observe the same unit (individual, household, firm...) over time, thereby allowing for the presence of unobserved heterogeneity. The main goal of this paper is 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.

Specifically, we focus on models of the form:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \mathbf{x}_{it}'\boldsymbol{\gamma}_i + v_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T. \tag{1}$$

In this model, the parameter vector $\boldsymbol{\delta}$ is common across individuals, while the vector of *random coefficients* $\boldsymbol{\gamma}_i$ is individual-specific.[1] We assume that the econometrician has data on y_{it}, $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$, and that she does not observe $\boldsymbol{\gamma}_i$ or the error terms v_{it}.

Examples of model (1) are very common in economics. A first example is provided by random trends models, which are obtained when $\mathbf{x}_{it}'\boldsymbol{\gamma}_i = \alpha_i + \beta_i t$ is an individual-specific slope. These models have been used to describe the dynamics of individual earnings (Guvenen, 2008).[2] A second example is given by firm-level production function models (Mairesse and Grilliches, 1990, Dobbelaere and Mairesse, 2008). In a production function approach, it is natural to interpret $\mathbf{x}_{it}$ as inputs, and $\boldsymbol{\gamma}_i$ as technology parameters. A third example arises when $\mathbf{x}_{it}$ is a *treatment* that has heterogeneous effects on outcomes. For example, in rational addiction models (Becker *et al.*, 1994), the price coefficient is the marginal utility of wealth, most likely heterogeneous across individuals (see the discussion in Arellano, 2003, p.131). As another illustration, when measuring the union wage premium it makes sense to allow for heterogeneity in the union status variable, if individuals have different abilities in unionized and non-unionized jobs (Vella and Verbeek, 1998, Lemieux, 1998).

We argue that in all these examples, it is of interest to document the *distribution* of $\boldsymbol{\gamma}_i$. For example, in random trends models of earnings dynamics, knowing the distribution of individual effects (and also of error variables) is necessary if one wants to use the earnings

[1] Consistently with the panel data literature, we refer to i as "individuals", and to t as "time periods".
[2] See also Lillard and Weiss (1979), Baker (1997), Haider (2001) and Guvenen (2007).

process in a consumption model (Guvenen, 2007). More generally, it is often important to estimate the effect of a set of covariates $\mathbf{x}_{it}$ at different quantiles (see the gigantic literature on heterogeneous treatment effects, e.g., Imbens and Angrist, 1994, Heckman and Vytlacil, 2005). In a cross-sectional setting, it is well-known that the distribution of a treatment effect is not point identified (Heckman *et al.*, 1997). In a panel data context, we show that one may have point identification of the full distribution of the effect, if $\mathbf{x}_{it}$ exhibits variation at the individual level. For example, in our application, we will estimate the distribution of mother-specific effects of smoking during pregnancy on children's weight at birth, extending previous work by Abrevaya (2006).

We study the identification of distributional features of $\boldsymbol{\gamma}_i$ under the assumption that the number of time periods T is fixed. Importantly, we make no assumption on the conditional distribution of individual-specific effects given regressors. Hence, we follow a "fixed-effects" approach, treating $\boldsymbol{\gamma}_i$, $i = 1, ..., N$, as random draws from an unknown distribution (Mundlak, 1978). In economic applications, unit-specific effects often represent heterogeneity in preferences or technology, on which economic theory has typically little to say. For this reason, it is important to adopt this minimal approach that does not restrict the form of heterogeneity.[3]

Identification and estimation of common parameters $\boldsymbol{\delta}$ and average individual-specific parameters $\mathbf{E}(\boldsymbol{\gamma}_i)$ in this context has been studied in Chamberlain (1992). The key identifying assumption is strict exogeneity of regressors, which requires errors to be mean independent of regressors at all lags: $\mathbf{E}(v_{it}|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}) = 0$. Strict exogeneity is an important assumption that needs to be carefully examined in the context of each empirical application. When regressors with individual-specific effects are predetermined or endogenous, identification requires to impose restrictions on the distribution of $\boldsymbol{\gamma}_i$ given regressors.[4]

We show the identification of the variance-covariance matrix of $\boldsymbol{\gamma}_i$ under conditional uncorrelatedness assumptions on error variables. We give precise identification conditions under two types of restrictions: moving averages (MA), and autoregressive/ARMA type restrictions. In the particular case where errors are i.i.d. homoskedastic, our results coincide

[3]For example, Cameron and Trivedi (2005, p.777) claim that models with random coefficients of the form (1), 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".

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

with those of Swamy (1970). Importantly, our results show that the variance of individual effects is *not* identified if no assumptions are made on the variance-covariance structure of errors, as in Chamberlain (1992). This defines a clear trade-off, between the amount of unobserved heterogeneity in the model (the dimension of the parameter vector $\boldsymbol{\gamma}_i$) and the dynamic structure of errors. Under additional assumptions, we show that higher-order moments of individual effects and errors, such as skewness and kurtosis, are also identified for fixed T. The same trade-off between dynamics and heterogeneity applies there.

Strengthening the uncorrelatedness assumptions on error variables to conditional independence assumptions, we are able to show the identification of the full distribution of individual effects, and error variables. In particular, this implies the identification of all quantiles of the distribution of individual-specific effects. We treat the case of MA, or more generally ARMA, processes with conditionally independent innovations. This result extends previous work on the identification of factor distributions in independent factor models (Kotlarski, 1967, Székely and Rao, 2000). Importantly, and differently from those papers, we are able to prove *nonparametric* identification of the multivariate conditional distribution of individual effects, without imposing any restrictions on the latter.

Although the main focus of this paper is on identification, we also discuss how to estimate the moments and densities of effects and errors. We propose 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 (e.g., Dasgupta, 2008, Chapter 33). In the case where $\mathbf{x}_{it}'\boldsymbol{\gamma}_i = \alpha_i + \beta_i x_{it}$ with x_{it} scalar binary, and errors are i.i.d., we introduce a simple nonparametric estimator of the density of β_i, using a methodology recently developed in Mallows (2007). We then use this estimator in our application to estimate the effect of smoking on birthweights.

This paper shares common features with three strands of the econometric and statistical literature. Linear panel data models with random coefficients, referred to as *mixed* models, have been extensively studied in statistics. Recent work has tried to treat the distribution of individual effects semi or nonparametrically.[5] Our approach is also connected to the nonparametric identification and estimation of factor distributions in independent factor

[5]See Demidenko (2004), for a survey on mixed models. Harville (1977) and Laird and Ware (1982) are early references. Work using semi and nonparametric approaches can be found in Lesaffre and Verbeke (1996), Kleinman and Ibrahim (1998), and Davidian and Zhang (2001). Related work includes random-coefficient models for cross-section data (Beran and Hall, 1992, Hoderlein *et al.*, 2007).

models.[6] Compared to these two strands of the literature, we take a minimal approach and leave the conditional distribution of individual effects unrestricted.

The paper is also related to the literature on the estimation of 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 Hahn and Newey (2004) and Arellano and Hahn (2006) for a survey. Here we show that all *marginal effects*, including the density of individual-specific effects, are identified for T *fixed* in model (1). Hence, our approach leads to full elimination of the bias on the quantities of interest.

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 moments of individual effects and error variables, and in section 4, method-of-moments estimators of these quantities are proposed. In section 5 we study the nonparametric identification of the densities of effects and errors, discuss estimation in a general context, and propose an estimator in a special case, that we apply in section 6 to Abrevaya's (2006) data on smoking and birth outcomes. Lastly, section 7 concludes.

2 Preliminaries

2.1 Model and assumptions

In most of this paper we consider a model that relates a vector of T endogenous variables $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$ to a set of regressors. For convenience we distinguish two types of regressors: $\mathbf{Z}_i = (z_{i1}...z_{iT})'$ is a $T \times K$ matrix associated to a vector of K parameters $\boldsymbol{\delta}$, while $\mathbf{X}_i = (x_{i1}...x_{iT})'$ is a $T \times q$ matrix associated to a *unit specific* vector of q parameters $\boldsymbol{\gamma}_i$. The linear relationship takes the form:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\delta} + \mathbf{X}_i\boldsymbol{\gamma}_i + \mathbf{v}_i, \quad i = 1...N, \tag{2}$$

where $\mathbf{v}_i = (v_{i1}...v_{iT})'$ is a vector of T error terms.

We start by making assumptions on error variables $\mathbf{v}_i$, specifying their conditional mean given both types of regressor to be zero, thereby assuming that $\mathbf{Z}_i$ and $\mathbf{X}_i$ are *strictly*

[6] See Horowitz and Markatou (1996), Li and Vuong (1998), Hall and Yao (2003), Bonhomme and Robin (2008b), and the literature on Independent Component Analysis surveyed in Hyvärinen *et al.* (2001).

exogenous:[7,8]

$$\mathbf{E}\left(v_{it} \mid \mathbf{Z}_i, \mathbf{X}_i\right) = 0. \tag{3}$$

Endogenous $\mathbf{Z}_i$'s could be dealt with if one had enough instruments to identify/estimate $\boldsymbol{\delta}$. However, strict exogeneity of $\mathbf{X}_i$ is essential. We will discuss the strict exogeneity assumption at the end of this section.

Strict exogeneity alone will typically allow to identify the vector of common parameters $\boldsymbol{\delta}$ and the mean of individual-specific parameters $\mathbf{E}\left(\boldsymbol{\gamma}_i\right)$. In order to achieve the identification of other distributional characteristics of $\boldsymbol{\gamma}_i$, such as variance or quantiles, we will restrict further the distribution of error variables. In the identification analysis of sections 3 and 5, these restrictions will take the form of conditional uncorrelatedness and independence assumptions.

Importantly, we do not specify the conditional distribution of individual effects. This is a distinctive feature of our approach compared to random and mixed-effects models which specify both the distribution of $\mathbf{y}_i$ given $\mathbf{Z}_i$, $\mathbf{X}_i$ and $\boldsymbol{\gamma}_i$, *and* the distribution of $\boldsymbol{\gamma}_i$ given regressors. We adopt a "fixed-effect" approach, which we understand as meaning that $\boldsymbol{\gamma}_i$ are random draws from a population, along with y_{it}, $\boldsymbol{z}_{it}$ and $\boldsymbol{x}_{it}$, but leave their *conditional* distribution given regressors unspecified. See Mundlak (1978), and Wooldridge (2002), p.252, for a similar view of the "fixed-effect" perspective in microeconometric applications. In particular, we leave the correlation between individual effects and regressors unrestricted.

The strict exogeneity assumption (3) needs to be interpreted in view of the lack of restrictions on individual effects. Strict exogeneity requires that regressors $\mathbf{X}_i$ are uncorrelated with errors at all periods. However, regressors are allowed to be correlated with $\boldsymbol{\gamma}_i$ in an unrestrictive manner. In Deschênes and Greenstone (2007) regressors are weather indicators, while the dependent variable is agricultural profit, measured at the county level. There, strict exogeneity is a reasonable assumption only when one accounts for county effects, if only because land quality is likely to be correlated with the weather and to vary from an area to the next while being quite persistent over time. In the case of the union wage premium, endogenous job mobility can invalidate the strict exogeneity asumption (Vella and Verbeek, 1998). Lemieux (1998) exploits plant closing as an indicator of involuntary job mobility to circumvent this problem. In Abrevaya's (2006) study of smoking effects on birthweight, the

[7] All (in)equalities conditional on $\mathbf{Z}_i$ and/or $\mathbf{X}_i$ are understood to hold with probability one.

[8] Throughout the paper, the moments that we use are assumed well-defined (i.e., finite).

assumption fails if women react to a low birthweight by quitting smoking. We will come back to this issue in our use of Abrevaya's data in section 6.

In the course of the analysis, we will also maintain another assumption that requires regressors $\mathbf{X}_i$ not to be collinear *within* each individual sequence of observations, formally:

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

In particular, (4) imposes 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 crucial in our setting. The situation is very different from one 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).

For condition (4) to be satisfied in practice, the sample of observations will often need to be selected. This will be the case in an unbalanced panel where individuals i with $T_i < q$ observations will need to be dropped from the sample. Another instance is when $\mathbf{X}_i$ takes discrete values. For example, in a model with a constant and a binary regressor $\mathbf{x}_i$, both having heterogeneous effects on the dependent variable, sequences with T zeros, or T ones, will be kept out of the analysis. So, the characteristics of interest will typically be identified on a subpopulation of individuals whose x's *change* over time. Hence, in the application to birthweight data in section 6, we will focus on women who changed smoking status between pregnancies. Non-identification of effects on subpopulations of individuals is in common with the Instrumental Variables setting, local treatment effects being identifed on the subpopulation of *compliers* only (Imbens and Angrist, 1994).

In the case where $\mathbf{X}_i$ takes a continuum of values, at first sight the multicollinearity problem does not arise. Indeed, if the determinant $|\mathbf{X}_i'\mathbf{X}_i|$ is non zero with probability one condition (4) will be satisfied. This will often be the case on an a finite sample of observations. However, this view may be seriously misleading in practice. The reason is that if $|\mathbf{X}_i'\mathbf{X}_i|$ is *too close* to zero the corresponding individual-specific parameters will be badly estimated, contaminating estimates of distributional characteristics. It may then make sense to keep only observations for which $|\mathbf{X}_i'\mathbf{X}_i|$ is larger than an empirically determined bandwidth h_N

that shrinks with the sample size. More discussion of this issue is found in Graham and Powell (2008).

2.2 Within and between transformations

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

$$\widehat{\boldsymbol{\gamma}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right). \tag{5}$$

Then, for every q-dimensional parameter vector $\boldsymbol{\gamma}_i$ we can write:

$$\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\gamma}_i = \left[\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta} - \mathbf{X}_i\widehat{\boldsymbol{\gamma}}_i\right] + \left[\mathbf{X}_i\left(\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\right)\right]. \tag{6}$$

The first and second terms on the right-hand side of (6) are *within*-group and *between*-group terms, respectively. In the simple case where $\boldsymbol{\delta}$ is zero and $\mathbf{X}_i$ is a vector of ones, (6) is simply: $y_{it} = (y_{it} - \overline{y}_i) + \overline{y}_i$, where $\overline{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ is the individual mean of y's.

Under normality and classical errors there are strong statistical arguments to consider the within-group and between-group likelihoods separately. Indeed, suppose that v_i follows a normal distribution independent of $(\mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\gamma}_i)$ with variance-covariance matrix $\sigma^2\mathbf{I}_T$ (with $\mathbf{I}_T$ the T-by-T identity matrix). In this case the log-likelihood of the data is the sum of the within-group and between-group log-likelihoods. So it is possible to base the estimation of $\boldsymbol{\delta}$ and σ^2 on the within-group likelihood alone, a method sometimes called *restricted* maximum likelihood (Patterson and Thompson, 1971). Alternatively, one can show that the within-group likelihood is proportional to the weighted likelihood in which individual-specific parameters have been integrated out with respect to a "non-informative" uniform prior (Harville, 1974). At an intuitive level, the between-group likelihood is not informative about $\boldsymbol{\delta}$ and σ^2 if the conditional distribution of $\boldsymbol{\gamma}_i$ given the regressors is unrestricted. Arellano (2003, p.26) makes this remark in the context of a model with an heterogeneous intercept.

The previous discussion motivates looking at the following two equations:

$$\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right) = \mathbf{Q}_i\mathbf{v}_i \qquad \text{(within-group)}, \tag{7}$$

$$\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i = \mathbf{H}_i\mathbf{v}_i \qquad \text{(between-group)}, \tag{8}$$

where $\mathbf{Q}_i$ (T-by-T) and $\mathbf{H}_i$ (q-by-T) are the *within-group* and *least-squares* operators, namely:

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

Equations (7) and (8) are readily obtained by left-multiplying (2) by $\mathbf{Q}_i$ and $\mathbf{H}_i$, respectively. While (8) expresses the difference between the least-squares estimate of $\boldsymbol{\gamma}_i$ (for known $\boldsymbol{\delta}$) and its true value, (7) shows the link between the residuals in the individual-specific least-squares regressions and the population errors. Note that these equations do not require either normality or independence of the v's to hold. 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, it is intuitive that errors must be restricted in some way for identification to hold. Moreover, allowing for more parameters to be individual-specific will require a larger number of restrictions on $\mathbf{v}_i$. Formally, the within transformation matrix $\mathbf{Q}_i$ has rank $T-q$ (e.g., Wooldridge, 2002, p.319), so in order to invert (7) one will need to impose a larger number of restrictions on $\mathbf{v}_i$ the larger q is. The second remark concerns the fact that $\widehat{\boldsymbol{\gamma}}_i$ is a noisy estimate of $\boldsymbol{\gamma}_i$. Likewise, any distributional characteristic of $\widehat{\boldsymbol{\gamma}}_i$ (mean, variance, quantiles) 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. For example, in the model with no common parameter and an heterogeneous constant, one has: $\overline{y}_i - \gamma_i = \overline{v}_i$, which is a sample mean of T observations, the variance of which is of the order of magnitude of $1/T$. For this reason, unit-by-unit estimates of $\boldsymbol{\gamma}_i$ are not directly informative on the distribution of the underlying effects.

2.3 Extensions and discussion

Although we discuss identification of the linear model (2), the approach of this paper can be easily generalized to other settings. A more general formulation is (Chamberlain, 1992):

$$\mathbf{y}_i = \mathbf{a}(\mathbf{X}_i; \boldsymbol{\theta}) + \mathbf{B}(\mathbf{X}_i; \boldsymbol{\theta})\boldsymbol{\gamma}_i + \mathbf{v}_i \tag{9}$$

where $\boldsymbol{\theta}$ are common parameters and enter nonlinearly functions $\mathbf{a}$ (which is T-by-1) and $\mathbf{B}$ (T-by-q).

The identification analysis of (9) follows very closely that of the linear model (2). We will indicate the differences in the course of the exposition. It is instructive to consider examples

of (9). A simple version is the one-factor model:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \mu_t\gamma_i + v_{it}, \tag{10}$$

where $\mu_1, ..., \mu_T$ are time-varying parameters and γ_i is scalar (e.g., Holtz-Eakin *et al.*, 1988). In a wage regression, γ_i could be workers' unobserved skills on the labor market, and μ_t their time-varying price. Multiple-equations versions of (10), 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, 2006, Ahn *et al.*, 2007).

A second special case of (9) arises when one wants to limit the number of heterogeneous parameters in a model. For example, instead of:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta}_i + v_{it},$$

one may consider a restricted alternative with *two* individual-specific effects:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\delta} + \widetilde{\alpha}_i + \widetilde{\beta}_i\left(\mathbf{x}_{it}'\mathbf{g}\right) + v_{it}. \tag{11}$$

In practice, having fewer individual-specific parameters could result in more precise estimates. Moreover, in some applications the time dimension may be too small to allow for q effects, but enough to allow for a more parsimonious specification.

As another interesting special case of (9) we mention models where the regressors include lags or leads of the dependent variable. For example, a first-order autoregressive model:

$$y_{it} = \delta y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\gamma}_i + v_{it}, \quad |\delta| < 1. \tag{12}$$

That (12) is a special case of (9) 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 (9) with the $(q+1)$-by-1 vector of individual effects: $\widetilde{\boldsymbol{\gamma}}_i = (\boldsymbol{\gamma}_i', y_{i0})'$. Likewise one could add a lead $y_{i,t+1}$ in (12), as in the rational addiction model of Becker *et al.* (1994). The reduced-form would then have a $q+2$-dimensional vector of individual-specific effects, composed of $\boldsymbol{\gamma}_i$, of the initial value of y_{it} (y_{i0}), and of its *final* value ($y_{i,T+1}$).

Before ending this preliminary section, it is useful to mention a case where the distributional quantities of interest are *not* identified. This happens whenever one of the components

of $\mathbf{x}_{it}$ is predetermined or endogenous, as opposed to strictly exogenous. Chamberlain (1993) and Arellano and Honoré (2001) provide examples. Here we consider the simple model

$$y_{it} = \alpha_i + \beta_i x_{it} + \mathbf{z}_{it}'\boldsymbol{\delta} + v_{it}, \tag{13}$$

where $\mathbf{E}(v_{it}|x_{it}, x_{i,t-1}, ..., \mathbf{Z}_i) = 0$. In other words, x_{it} is predetermined while $\mathbf{z}_{it}$ are strictly exogenous. We focus on the case where $x_{it} \in \{0, 1\}$ is binary, and assume $T = 3$. This corresponds to our application, if the smoking status of a mother is predetermined, see section 6 below.

Following Arellano and Honoré (2001) it is easy to see that, unless the dependence of β_i on $\mathbf{Z}_i$ is restricted, the vector of common parameters $\boldsymbol{\delta}$ is not identified, let alone the mean $\mathbf{E}(\beta_i)$. Moreover, even if $\boldsymbol{\delta}$ is zero, not all of the conditional means of β_i given the sequences of x's are identified. See the appendix for a justification. Hence, in our unrestricted random effects approach, a model with predetermined x's is unidentified. As a final remark, note that if one is ready to assume that β_i is mean independent of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, then identification can be obtained, see section 6.

3 Moment restrictions

In this section we derive moment restrictions on model (2), and study the identification of characteristics of interest. We discuss mean, variance and higher-order moments in turn.

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 (most of them elementary) are in the appendix.

Proposition 1 *Suppose that (3) and (4) hold. We have:*

$$\mathbf{E}\left(\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{Z}_i, \mathbf{X}_i\right) = 0 \tag{14}$$

and

$$\mathbf{E}\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i, \mathbf{X}_i\right) = \mathbf{E}\left(\boldsymbol{\gamma}_i|\mathbf{Z}_i, \mathbf{X}_i\right). \tag{15}$$

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. In the model with only an heterogeneous intercept (and $q = 1$), $\boldsymbol{\delta}$ satisfies:

$$\mathbf{E}\left(y_{it} - \overline{y}_i - \left(\mathbf{z}_{it} - \overline{\mathbf{z}}_i\right)' \boldsymbol{\delta} | \mathbf{Z}_i, \mathbf{X}_i\right) = 0.$$

Likewise, $\mathbf{E}(\boldsymbol{\gamma}_i)$ can be understood as a *mean-group* estimand. In the model with an heterogeneous intercept:

$$\mathbf{E}(\gamma_i) = \mathbf{E}\left(\overline{y}_i - \overline{\mathbf{z}}_i' \boldsymbol{\delta}\right).$$

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 $\mathbf{F}_i$ and in the regression of $\boldsymbol{\gamma}_i$ on $\mathbf{F}_i$ are equal.

Corollary 1 *Let the assumptions in proposition 1 hold. Let also* $\mathbf{F}_i$ *be a random vector such that* $\mathbf{E}(v_{it} | \mathbf{Z}_i, \mathbf{X}_i, \mathbf{F}_i) = 0$. *Then:*

$$\left[\mathrm{Var}(\mathbf{F}_i)\right]^{-1} \mathrm{Cov}(\mathbf{F}_i, \boldsymbol{\gamma}_i) \quad = \quad \left[\mathrm{Var}(\mathbf{F}_i)\right]^{-1} \mathrm{Cov}(\mathbf{F}_i, \widehat{\boldsymbol{\gamma}}_i). \tag{16}$$

Similar results can be obtained for the more general formulation (9). The next corollary derives moment conditions satified by common parameters $\boldsymbol{\theta}$.

Corollary 2 *Consider model (9), and suppose that* $\mathbf{E}(v_{it} | \mathbf{X}_i) = 0$ *and that matrix* $\mathbf{B}(\mathbf{X}_i; \boldsymbol{\theta})$ *has rank q. Then* $\boldsymbol{\theta}$ *satisfies the following moment conditions:*

$$\mathbf{E}\left[\mathbf{Q}_i(\boldsymbol{\theta})\left(\mathbf{y}_i - \mathbf{a}(\mathbf{X}_i; \boldsymbol{\theta})\right) | \mathbf{X}_i\right] = 0, \tag{17}$$

where

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

Corollary 2 provides moment restrictions that may or may not be sufficient to identify $\boldsymbol{\theta}$. Simple cases where they are not sufficient are the one-factor model (10) where there are no regressors ($\boldsymbol{\delta} = \mathbf{0}$), or the AR(1) model with fixed effects (12) where only the constant is individual-specific and there are no common regressors. Then, identification will require to restrict the variance-covariance matrix of errors and to exploit covariance restrictions, see 3.4 below (Holtz-Eakin *et al.*, 1988, Arellano and Bond, 1991). Chamberlain (1992) considers the

identification content of model (9) when no uncorrelatedness assumptions are made on error variables, in which case models with no exogenous regressors are fundamentally unidentified.

Remark that, once $\boldsymbol{\theta}$ is identified, there is no essential difference between model (2) and model (9). Indeed, one can relabel $\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{a}(\mathbf{X}_i; \boldsymbol{\theta})$ as the dependent variable and $\widetilde{\mathbf{X}}_i = \mathbf{B}(\mathbf{X}_i; \boldsymbol{\theta})$ as the set of regressors, and use the identification results obtained for model (2). In particular, the overall and conditional means of individual effects are trivially identified.

A last remark concerns the fact that the moment restrictions on $\boldsymbol{\delta}$ in Proposition 1, and on $\boldsymbol{\theta}$ in Corollary 2, are not unique. For example, in Proposition 1 we could use, for any positive definite T-by-T matrix $\mathbf{W}_i$ possibly dependent on $(\mathbf{Z}_i, \mathbf{X}_i)$:

$$\mathbf{Q}_i^{\mathbf{W}_i} = \mathbf{W}_i^{-1} - \mathbf{W}_i^{-1}\mathbf{X}_i \left(\mathbf{X}_i'\mathbf{W}_i^{-1}\mathbf{X}_i\right)^{-1} \mathbf{X}_i'\mathbf{W}_i^{-1}, \tag{19}$$

in place of $\mathbf{Q}_i$ in equation (14). This choice may have efficiency consequences when turning to estimation, as we will explain in section 4.

3.2 Variances

To recover the variance of individual effects, we make two additional assumptions on model (2). First, we assume that individual effects and errors are uncorrelated given regressors, that is:

$$\mathbf{Cov}\left(\boldsymbol{\gamma}_i, \mathbf{v}_i | \mathbf{Z}_i, \mathbf{X}_i\right) = 0. \tag{20}$$

Condition (20) is satisfied if one assumes that individual effects are strictly exogenous in (2), a natural assumption in a fixed-effect approach (made, e.g., in Chamberlain, 1992):

$$\mathbf{E}\left(v_{it} | \mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\gamma}_i\right) = 0. \tag{21}$$

Second, we impose restrictions on the variance-covariance matrix of errors. For exposition, we start with the case where $\boldsymbol{\Omega}_i = \mathbf{Var}\left(\mathbf{v}_i | \mathbf{Z}_i, \mathbf{X}_i\right)$ is known. The following theorem shows that the variance of individual effects is identified under those conditions. The proof is immediate using (8).

Theorem 1 *Suppose that (3), (4) and (20) hold. Then we have:*

$$\mathbf{Var}\left(\boldsymbol{\gamma}_i | \mathbf{Z}_i, \mathbf{X}_i\right) = \mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i | \mathbf{Z}_i, \mathbf{X}_i\right) - \mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i' \tag{22}$$

and, unconditionally:

$$\mathbf{Var}\left(\boldsymbol{\gamma}_i\right) = \mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i\right) - \mathbf{E}\left(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\right). \tag{23}$$

Theorem 1 shows that the variance-covariance matrix of individual effects is identified given that of error variables. In the special case where $\mathbf{\Omega}_i = \sigma^2 \mathbf{I}_T$, (23) yields:

$$\mathbf{Var}\left(\boldsymbol{\gamma}_i\right) = \mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i\right) - \sigma^2 \mathbf{E}\left[\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]. \tag{24}$$

A familiar expression is obtained in the case where only the constant is heterogeneous, in which case: $\mathrm{Var}\left(\gamma_i\right) = \mathrm{Var}\left(\overline{y}_i - \overline{\mathbf{z}}_i'\boldsymbol{\delta}\right) - \sigma^2/T$ (Arellano, 2003, p.33).

More generally, (23) may be written

$$\mathbf{Var}\left(\widehat{\boldsymbol{\gamma}}_i\right) = \mathbf{Var}\left(\boldsymbol{\gamma}_i\right) + \mathbf{E}\left(\mathbf{H}_i \mathbf{\Omega}_i \mathbf{H}_i'\right), \tag{25}$$

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, because $\widehat{\boldsymbol{\gamma}}_i$ has mean $\boldsymbol{\gamma}_i$. The within-group variance tends to zero when T tends to infinity. 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 to be noted that the linearity of the model (with respect to the individual effects) is crucial for this result to hold.

We now turn to the identification of $\mathbf{\Omega}_i$. The within-group equation (7) yields:

$$\mathbf{Q}_i \mathbf{E}\left[\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)\left(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right)' | \mathbf{Z}_i, \mathbf{X}_i\right] \mathbf{Q}_i' = \mathbf{Q}_i \mathbf{\Omega}_i \mathbf{Q}_i'. \tag{26}$$

As $\mathbf{Q}_i$ has rank $T - q$, we cannot invert (26) and recover $\mathbf{\Omega}_i$ unless we impose restrictions. To start with, we restrict a set of pairs of error variables to be conditionally uncorrelated given regressors. A particular example is a moving average (MA) process of order r, the covariances between v_{it} and $v_{i,t+r+1}$ being zero given $\mathbf{Z}_i, \mathbf{X}_i$. Formally, we assume that there exists a vector of m parameters $\boldsymbol{\omega}_i$, possibly dependent on $\mathbf{Z}_i, \mathbf{X}_i$, and a known (selection) matrix $\mathbf{S}_2$ such that:

$$\mathbf{vec}\left(\mathbf{\Omega}_i\right) = \mathbf{S}_2 \boldsymbol{\omega}_i. \tag{27}$$

Condition (27) 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 position $(1,1)$, $(T+2, 2)$,..., (T^2, T). More generally, the assumption contains moving-average processes of the form

$$v_{it} = u_{it} + \theta_{1t} u_{i,t-1} + ... + \theta_{rt} u_{i,t-r}, \quad t = 1, ..., T, \tag{28}$$

where $\theta_{11}, ..., \theta_{rT}$ are unrestricted parameters (possibly dependent on regressors, although we ommit the i subindex for clarity), 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, taking the vector form of (26) yields, using (27):

$$(\mathbf{Q}_i \otimes \mathbf{Q}_i)\, \mathbf{E}\left[(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}) \otimes (\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})\,|\mathbf{Z}_i, \mathbf{X}_i\right] = (\mathbf{Q}_i \otimes \mathbf{Q}_i)\, \mathbf{vec}\,(\boldsymbol{\Omega}_i) \tag{29}$$

$$= (\mathbf{Q}_i \otimes \mathbf{Q}_i)\, \mathbf{S}_2 \boldsymbol{\omega}_i. \tag{30}$$

We thus have the following identification theorem.

Theorem 2 *Suppose that (3), (4) and (27) hold. Suppose also that*

$$\text{rank}\left[(\mathbf{Q}_i \otimes \mathbf{Q}_i)\, \mathbf{S}_2\right] = m. \tag{31}$$

Then matrix $\boldsymbol{\Omega}_i$ is identified from (7) alone.

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

Corollary 3 *If errors are i.i.d. independent of $(\mathbf{Z}_i, \mathbf{X}_i)$ with variance σ^2 we have*

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

When $\mathbf{S}_2$ selects all $T(T+1)/2$ non-redundant elements of $\mathbf{vec}\,(\boldsymbol{\Omega}_i)$, the left-hand side of (31) becomes: $(T-q)(T-q+1)/2$, see Lemma 1 part i) in the appendix. So, the order condition associated with the rank condition (31) is: $(T-q)(T-q+1)/2 \geq m$. In particular, in the MA(r) case we need that

$$\frac{(T-q)(T-q+1)}{2} \geq (r+1)\left(T - \frac{r}{2}\right). \tag{32}$$

The left-hand-side in (32) is decreasing in q, while the right-hand side is increasing in r. So this equation emphasizes a trade-off between the number of individual-specific effects and the order of the moving-average process.

Autoregressive errors are very popular in applied work, and are *not* covered by assumption (27) because errors are correlated at all lags. Nevertheless, a similar approach can be adopted to study identification. To see how, consider the following model:

$$v_{it} = \rho_{1t} v_{i,t-1} + ... + \rho_{pt} v_{i,t-p} + u_{it}, \quad t = p+1, ..., T, \tag{33}$$

where $\rho_{1,p+1}, ..., \rho_{pT}$ are unrestricted parameters and $u_{i,p+1}, ..., u_{iT}$ satisfy assumption (27). In the case where u_{it} is MA(r), v_{it} given by (33) follows an ARMA(p,r) process.

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

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

We have: $\mathbf{R}\mathbf{v}_i = \mathbf{u}_i$, so:

$$\mathbf{R}\mathbf{y}_i = \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta} + \mathbf{R}\mathbf{X}_i\boldsymbol{\gamma}_i + \mathbf{u}_i. \tag{34}$$

Variance restrictions on model (34) imply that

$$\left(\widetilde{\mathbf{Q}}_i \otimes \widetilde{\mathbf{Q}}_i\right) \mathbf{E}\left[(\mathbf{R}\mathbf{y}_i - \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta}) \otimes (\mathbf{R}\mathbf{y}_i - \mathbf{R}\mathbf{Z}_i\boldsymbol{\delta}) \,|\mathbf{Z}_i, \mathbf{X}_i\right] = \left(\widetilde{\mathbf{Q}}_i \otimes \widetilde{\mathbf{Q}}_i\right) \mathbf{S}_2\boldsymbol{\omega}_i, \tag{35}$$

where we have denoted: $\mathbf{vec}\left(\mathbf{Var}\left(\mathbf{u}_i|\mathbf{Z}_i, \mathbf{X}_i\right)\right) = \mathbf{S}_2\boldsymbol{\omega}_i$, where $\boldsymbol{\omega}_i$ is m-by-1, and:

$$\widetilde{\mathbf{Q}}_i = \mathbf{I}_{T-p} - \mathbf{R}\mathbf{X}_i \left(\mathbf{X}_i'\mathbf{R}'\mathbf{R}\mathbf{X}_i\right)^{-1} \mathbf{X}_i'\mathbf{R}'.$$

Note that, by multiplying by $\mathbf{R}$ we have lost p dimensions, as $\widetilde{\mathbf{Q}}_i$ has rank $T-p-q$ while $\mathbf{Q}_i$ has rank $T-q$. 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 $\widetilde{\mathbf{Q}}_i$ allows to eliminate the q remaining ones.

It follows from (35) that, for the variances of $u_{i,p+1}, ..., u_{iT}$ and parameters $\rho_{1,p+1}, ..., \rho_{pT}$ to be identified from equation (30) the following rank condition needs to be satisfied:

$$\text{rank}\left(\left[\widetilde{\mathbf{Q}}_i \otimes \widetilde{\mathbf{Q}}_i\right] \mathbf{S}_2\right) = m. \tag{36}$$

In particular, we need that: $(T-p-q)(T-p-q+1)/2 \geq m$. 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.

Remark that, contrary to the moving average case, (36) is not strictly sufficient for identification to hold. Indeed, we also need parameters $\rho_{1,p+1}, ..., \rho_{pT}$ to be identified from (35). Also, 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.

3.3 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.

To proceed, we need some notation. Let $\mathbf{U} = (U_1, ..., U_n)'$ be a 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

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

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

$$\begin{aligned} \kappa_4^{i,j,k,\ell}(\mathbf{U}) &= \mathbf{E}\left(U_i U_j U_k U_\ell\right) - \mathbf{E}\left(U_i U_j\right)\mathbf{E}\left(U_k U_\ell\right) \\ &\quad -\mathbf{E}\left(U_i U_k\right)\mathbf{E}\left(U_j U_\ell\right) - \mathbf{E}\left(U_i U_\ell\right)\mathbf{E}\left(U_j U_k\right), \quad (i,j,k,\ell) \in \{1, ..., n\}^4. \end{aligned}$$

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 $\kappa_4^{j,j,j,j}(\mathbf{U})/\operatorname{Var}(U_j)^2 + 3$, 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-by-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 will be 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 the results in this subsection to cumulants of order 5 or higher. The first-order cumulant is simply the mean, and the cumulant of order 2 is the variance.

To recover the higher-moments of individual effects we impose a conditional independence restriction on individual effects and errors given regressors:

$$\boldsymbol{\gamma}_i \text{ and } \mathbf{v}_i \text{ are independent given } (\mathbf{Z}_i, \mathbf{X}_i). \tag{37}$$

The conditional independence restriction (37) is in the nature of a fixed-effect approach, where $\boldsymbol{\gamma}_i$ represent individual-specific parameters such as preferences or technology. Full independence (37) 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 section 5. Conditional independence restrictions are commonly made in the literature on nonparametric identification and estimation (e.g., Hu and Schennach, 2008, and references therein).

Using (8) together with (37) we obtain that:

$$\begin{aligned}\boldsymbol{\kappa}_3\left(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i\right) &= \boldsymbol{\kappa}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i\right)-\boldsymbol{\kappa}_3\left(\mathbf{H}_i\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right), \\ &= \boldsymbol{\kappa}_3\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i\right)-\left(\mathbf{H}_i\otimes\mathbf{H}_i\otimes\mathbf{H}_i\right)\boldsymbol{\kappa}_3\left(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right), \qquad (38)\end{aligned}$$

and, similarly:

$$\boldsymbol{\kappa}_4\left(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i\right) = \boldsymbol{\kappa}_4\left(\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i\right)-\left(\mathbf{H}_i\otimes\mathbf{H}_i\otimes\mathbf{H}_i\otimes\mathbf{H}_i\right)\boldsymbol{\kappa}_4\left(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right). \qquad (39)$$

It follows that the conditional cumulants of individual effects given regressors are identified if those of error variables are. In the particular case where only the constant is heterogeneous and errors are i.i.d. we obtain:

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

Remark that, as conditional moments can be recovered from conditional cumulants (Smith, 1995), it follows from these results that conditional and thus unconditional moments of individual effects are also identified.

Interestingly, (38) and (39) show that the bias on the cumulant of individual effects estimates $\widehat{\boldsymbol{\gamma}}_i$ is of the order of magnitude of $1/T^2$ or $1/T^3$, while from (22) the bias on the variance is of order $1/T$. It follows that, while fixed effects estimates have *larger* variance than individual effects in the population, we expect that their skewness and kurtosis will be biased *away* from zero, at least for reasonably large values of T.

Turning to the identification of error cumulants, using (7) together with (37) yields:

$$\boldsymbol{\kappa}_3\left(\mathbf{Q}_i\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{Z}_i,\mathbf{X}_i\right) = \left(\mathbf{Q}_i\otimes\mathbf{Q}_i\otimes\mathbf{Q}_i\right)\boldsymbol{\kappa}_3\left(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right), \qquad (40)$$

$$\boldsymbol{\kappa}_4\left(\mathbf{Q}_i\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{Z}_i,\mathbf{X}_i\right) = \left(\mathbf{Q}_i\otimes\mathbf{Q}_i\otimes\mathbf{Q}_i\otimes\mathbf{Q}_i\right)\boldsymbol{\kappa}_4\left(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right). \qquad (41)$$

As in the case of variances, these systems of equations are singular unless we impose restrictions on the structure of error variables. We adopt a similar approach as in (27) and assume that $\boldsymbol{\kappa}_3(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i) = \mathbf{S}_3\boldsymbol{\omega}_{3i}$, and $\boldsymbol{\kappa}_4(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i) = \mathbf{S}_4\boldsymbol{\omega}_{4i}$, 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 $\mathbf{Z}_i, \mathbf{X}_i$. Under these assumptions, identification of error cumulants can be shown if rank conditions analog to (31) are satisfied.

To motivate these restrictions, let us consider a model of the form (28), where innovations $u_{i,1-r}, ..., u_{iT}$ are now assumed mutually *independent* given regressors. Errors are thus modelled as linear combinations of independent underlying shocks. This modelling has been introduced by Rao (1969) and has recently been popularized by the literature on Independent Component Analysis (ICA) (e.g., Hyvärinen *et al.*, 2001). 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 of $\kappa_3^{t,t',s}(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i)$ and $\kappa_4^{t,t',s,s'}(\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i)$ are zero for all s, s' (see Lemma 1 in Bonhomme and Robin, 2008a). In an independent moving average model, v_{it} is 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:

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

In order for the rank conditions analog to (31) to be satisfied, the necessary order conditions are:

$$\binom{T-q+2}{3} \geq m_3(r) \quad , \quad \text{and} \quad \binom{T-q+3}{4} \geq m_4(r). \tag{42}$$

Hence, again, an apparent 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 (42) with (32).

It is also possible to show identification of higher-order moments in autoregressive models of the form (33), if the underlying shocks u_{it} follow an independent moving average model. For that, a possibility is to compute cumulants in the equation in quasi-differences (35).

3.4 Additional identifying restrictions

So far, the identification analysis has relied on a common strategy: identify common parameters and error moments from the within-group equation (7), and recover the moments

of individual effects from the between-group equation (8). While this strategy has intuitive appeal because it clearly separates the identification of common parameters from that of the individual-specific effects, it may actually lead to a loss of information.

Let us consider the identification of error variances. Instead of working with the variance-covariance matrix of quasi-differenced data as in (26) we can work directly with the variance-covariance matrix of $\mathbf{y}_i$, that is:

$$\mathbf{Var}\left(\mathbf{y}_i|\mathbf{Z}_i,\mathbf{X}_i\right)=\mathbf{X}_i\,\mathbf{Var}\left(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i\right)\mathbf{X}_i'+\boldsymbol{\Omega}_i. \tag{43}$$

In vector form, this yields:

$$\mathbf{vec}\left[\mathbf{Var}\left(\mathbf{y}_i|\mathbf{Z}_i,\mathbf{X}_i\right)\right]=\left(\mathbf{X}_i\otimes\mathbf{X}_i\right)\mathbf{vec}\left[\mathbf{Var}\left(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i\right)\right]+\mathbf{vec}\left(\boldsymbol{\Omega}_i\right). \tag{44}$$

The first term on the right-hand side of (44) is unrestricted, as the variance of individual effects is left unspecified. Let us define the projection matrix on $\mathbf{X}_i$: $\mathbf{P}_i=\mathbf{X}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'$, and the projection matrix on the *orthogonal* of $\mathbf{P}_i\otimes\mathbf{P}_i$:

$$\mathbf{M}_i=\mathbf{I}_{T^2}-\left[\mathbf{P}_i\left(\mathbf{P}_i'\mathbf{P}_i\right)^{-1}\mathbf{P}_i'\right]\otimes\left[\mathbf{P}_i\left(\mathbf{P}_i'\mathbf{P}_i\right)^{-1}\mathbf{P}_i'\right]. \tag{45}$$

Left-multiplying (44) by $\mathbf{M}_i$ yields

$$\mathbf{M}_i\,\mathbf{vec}\left[\mathbf{Var}\left(\mathbf{y}_i|\mathbf{Z}_i,\mathbf{X}_i\right)\right]=\mathbf{M}_i\,\mathbf{vec}\left(\boldsymbol{\Omega}_i\right). \tag{46}$$

Using (14) and (27) we can write (46) as:

$$\mathbf{M}_i\mathbf{E}\left[\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)\otimes\left(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta}\right)|\mathbf{Z}_i,\mathbf{X}_i\right] \;=\; \mathbf{M}_i\mathbf{S}_2\boldsymbol{\omega}_i. \tag{47}$$

Now, suppose that (3), (4) and (27) hold. Suppose also that

$$\mathrm{rank}\left(\mathbf{M}_i\mathbf{S}_2\right)=m. \tag{48}$$

It then follows from (47) that matrix $\boldsymbol{\Omega}_i$ is identified. It is important to notice that condition (48) is *less* restrictive than (31). Indeed, appying Lemma 1 part *ii*) in the appendix one can show that, when $\mathbf{S}_2$ selects all non-redundant elements of $\mathbf{vec}\left(\boldsymbol{\Omega}_i\right)$, the left-hand side in (48) is now: $T(T+1)/2-q(q+1)/2$. The necessary order condition for (48) to hold is thus: $T(T+1)/2-q(q+1)/2\geq m$, or, in the MA(r) case:

$$\frac{T(T+1)}{2}-\frac{q(q+1)}{2}\geq(r+1)\left(T-\frac{r}{2}\right). \tag{49}$$

Interestingly, the left-hand side in (49) corresponds to the *maximum* value of m for which identification may hold under those assumptions. Indeed, covariance restrictions (43) involve $T(T+1)/2$ data covariances, and $q(q+1)/2+m$ unrestricted covariances of the individual effects and error variables.

To illustrate the additional moment restrictions that appear in (46), consider a simple model with $T = 2$, $q = 1$, only an heterogeneous intercept, no common regressors and uncorrelated errors. The within-group equation (7) yields *only one* covariance restriction, namely:

$$\begin{aligned} \text{Var}\left(y_{i2} - y_{i1}\right) &= \text{Var}\left(v_{i2} - v_{i1}\right) \\ &= \text{Var}\left(v_{i1}\right) + \text{Var}\left(v_{i2}\right). \end{aligned}$$

In contrast, (46) yields *two* equations:[9]

$$\begin{aligned} \text{Var}\left(y_{i1}\right) - \text{Cov}\left(y_{i1}, y_{i2}\right) &= \text{Var}\left(v_{i1}\right) \\ \text{Var}\left(y_{i2}\right) - \text{Cov}\left(y_{i1}, y_{i2}\right) &= \text{Var}\left(v_{i2}\right). \end{aligned}$$

So, while (7) allows to identify the sum of the two variances, (46) allows to recover both variances. This argument also applies to higher-order moments. For example, even if errors are i.i.d., the skewness of v_{it} is *not* identified from the within-group equation, because first differences necessarily follow a symmetric distribution (Horowitz and Markatou, 1996). In contrast, using a higher-order version of (46) shows that the skewness is identified if independence between the individual effect and each of the error variables is assumed. Indeed, under that condition the full distribution of individual effects and errors is identified (Kotlarski, 1967). See the discussion in section 5 below.

It may come as a surprise that (7) does not contain all the information about error moments. The reason is that we have made restrictions to achieve the identification of the moments of individual effects. In the case of variances, (20) imposes that individual effects and errors are uncorrelated, while (27) restricts the variance-covariance matrix of errors. Given these restrictions, it is *not* true that the between-group likelihood is uninformative about error moments.

[9] Remark that:

$$\left[\left(\mathbf{M}_i\mathbf{S}_2\right)'\mathbf{M}_i\mathbf{S}_2\right]^{-1}\left(\mathbf{M}_i\mathbf{S}_2\right)'\mathbf{M}_i = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$

It is interesting to compare the two identification strategies. Using the within-group equation (7), the individual effects are differenced out. In contrast, in the second strategy the *moments* of individual effects are removed. In both cases removing the effects yields a standard problem where individual effects do not appear. The discussion in this subsection shows that the second strategy may allow to relax the requirements for identification. When turning to estimation, this will translate into more efficient estimates. Moreover, this second strategy may be applicable to nonlinear models, when differencing out the effects is not possible but one can difference out their distribution, see the conclusion.

4 Estimation, inference and testing

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

4.1 Common parameters and average effects

We start by discussing the estimation of common parameters and mean effects. From (14) $\boldsymbol{\delta}$ can be estimated as:

$$\widehat{\boldsymbol{\delta}} = \left(\sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Q}_i^{\mathbf{W}_i} \mathbf{Z}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Q}_i^{\mathbf{W}_i} \mathbf{y}_i, \tag{50}$$

where $\mathbf{Q}_i^{\mathbf{W}_i}$ is defined by (19). When $\mathbf{W}_i = \mathbf{I}_T$, $\widehat{\boldsymbol{\delta}}$ is the OLS estimator of $\boldsymbol{\delta}$ in (7). When $\mathbf{W}_i = \boldsymbol{\Omega}_i$, the variance-covariance matrix of error variables, $\widehat{\boldsymbol{\delta}}$ coincides with the infeasible GLS estimator of $\boldsymbol{\delta}$, see the appendix for a proof. The asymptotic distribution of $\widehat{\boldsymbol{\delta}}$ is normal with an asymptotic variance given by the standard White-type formula (e.g. Wooldridge, 2002, p.320-321).

Chamberlain (1992) shows that, for the choice $\mathbf{W}_i = \boldsymbol{\Omega}_i$, $\widehat{\boldsymbol{\delta}}$ reaches the semi-parametric efficiency bound for $\boldsymbol{\delta}$ when (21) is assumed. There are two ways of constructing a semi-parametrically efficient feasible version of $\widehat{\boldsymbol{\delta}}$. The first way makes use of the GLS interpretation of the estimator (e.g., Amemiya, 1985, p.186). The second way uses a result of Chamberlain (1992, Proposition 2, p.584) who shows that $\boldsymbol{\Omega}_i$ can be replaced by a positive-definite matrix such that: $\widetilde{\boldsymbol{\Omega}}_i = \mathbf{Var}\left(\mathbf{y}_i | \mathbf{Z}_i, \mathbf{X}_i\right) - \mathbf{X}_i \mathbf{V}_i \mathbf{X}_i'$, where $\mathbf{V}_i$ is *any* positive-definite q-by-q matrix, with no effect on the efficiency of $\widehat{\boldsymbol{\delta}}$. Note that, in order to compute this second estimator, one needs to estimate the conditional variance $\mathbf{Var}\left(\mathbf{y}_i | \mathbf{Z}_i, \mathbf{X}_i\right)$.

Likewise, from (15) a consistent method-of moments estimator of $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i)$ is the weighted mean-group estimator:

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i'\mathbf{W}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{W}_i^{-1}\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right). \tag{51}$$

When $\mathbf{W}_i = \mathbf{I}_T$, $\widehat{\boldsymbol{\gamma}}$ is the *mean-group* estimator of $\boldsymbol{\gamma}$ (e.g., Hsiao and Pesaran, 2006). Chamberlain (1992) shows that $\widehat{\boldsymbol{\gamma}}$ reaches the semi-parametric efficiency bound for $\boldsymbol{\gamma}$ if one chooses $\mathbf{W}_i = \boldsymbol{\Omega}_i$, or $\mathbf{W}_i = \widetilde{\boldsymbol{\Omega}}_i$ which has a feasible counterpart. Estimation in that case requires to estimate the conditional variance $\mathbf{Var}(\mathbf{y}_i|\mathbf{Z}_i, \mathbf{X}_i)$.

It is instructive to compare the mean-group estimator of $\boldsymbol{\gamma}$ given by (51) 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 model (9). A method-of-moment estimator of $\boldsymbol{\theta}$ based on (17) will be consistent. A particular choice for the matrix $\mathbf{Q}_i(\boldsymbol{\theta})$ yields semi-parametric efficiency, see Chamberlain (1992). In this case, $\widehat{\boldsymbol{\gamma}}$ may be estimated in a second step, or jointly with $\boldsymbol{\theta}$.

Chamberlain (1992) emphasizes an important difference between models (2) and (9). Indeed, in the linear model (2) 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 the nonlinear model (9), the fixed effects estimator of $\boldsymbol{\theta}$ is inconsistent in general. In contrast, a method-of-moments estimator based on (17) yields consistent estimates of $\boldsymbol{\theta}$.

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 on the individual effects in the population $\boldsymbol{\gamma}_i$ on $\mathbf{F}_i$. 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. This clearly also applies to the mean-group estimator of the unconditional mean, given by (51). We provide corrected formulas in the appendix.

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 its denominator is the variance of individual effects, which is overestimated by the variance of $\widehat{\boldsymbol{\gamma}}_i$, see (23). Correcting the R^2 requires to consistently estimate the variance of $\boldsymbol{\gamma}_i$, which we discuss next.

4.2 Variance and higher-order moments

We now turn to estimation of variances under the conditions of Theorem 2, that is under MA-type restrictions on the variance-covariance matrix of errors. The extension to autoregressive or ARMA structures is immediate and will not be detailed here. Let $\mathbf{A}^-$ denote any generalized inverse of a matrix $\mathbf{A}$ with full column rank, e.g. $\mathbf{A}^- = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. It follows from (30) that the variance-covariance matrix of errors can be consistently estimated by:

$$\mathbf{vec}\left(\widehat{\mathbf{Var}}(\mathbf{v}_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_2\left[(\mathbf{Q}_i\otimes\mathbf{Q}_i)\,\mathbf{S}_2\right]^-(\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i), \tag{52}$$

where we have denoted: $\widehat{\mathbf{v}}_i = \mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)$. Remark that one gets an alternative expression if the additional restrictions (46) are used, yielding:

$$\mathbf{vec}\left(\widehat{\mathbf{Var}}(\mathbf{v}_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_2\left(\mathbf{M}_i\mathbf{S}_2\right)^-\mathbf{M}_i\left[\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\otimes\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\right], \tag{53}$$

where $\mathbf{M}_i$ is given by (45).

$\widehat{\mathbf{Var}}(\mathbf{v}_i)$ given by (52) will be consistent as long as (27) is satisfied. In the particular case where arrors are i.i.d. with variance σ^2, Corollary 5 motivates estimating σ^2 as:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{N(T-q)}\sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)'\mathbf{Q}_i\left(\mathbf{y}_i - \mathbf{Z}_i\widehat{\boldsymbol{\delta}}\right)\\ &= \frac{1}{N(T-q)}\sum_{i=1}^{N}\widehat{\mathbf{v}}_i'\widehat{\mathbf{v}}_i.\end{aligned} \tag{54}$$

The first-order asymptotic distributions of (52) and (54) are straightforward to derive. Standard arguments show that it coincides with the distribution treating common parameters $\boldsymbol{\delta}$ as known (e.g., Goldberger, 1991, p.103). Interestingly, while $\widehat{\sigma}^2$ is non-negative by construction, $\widehat{\mathbf{Var}}(\mathbf{v}_i)$ in (52) is not necessarily non-negative definite.

Turning to estimation of the variance of individual effects, a consistent estimator based on (23) and (30) is:

$$\begin{aligned}\mathbf{vec}\left(\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right)\right) &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\otimes\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{H}_i\otimes\mathbf{H}_i\right)\mathbf{S}_2\left[\left(\mathbf{Q}_i\otimes\mathbf{Q}_i\right)\mathbf{S}_2\right]^{-}\left[\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\right].\end{aligned} \tag{55}$$

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

$$\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)'-\frac{1}{N(T-q)}\sum_{i=1}^{N}\widehat{\mathbf{v}}_i'\widehat{\mathbf{v}}_i\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \tag{56}$$

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

$$\widehat{\mathbf{Var}}\left(\boldsymbol{\gamma}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)\left(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}}\right)'-\widehat{\sigma}^2\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}, \tag{57}$$

where $\widehat{\sigma}^2$ is given by (54). The estimator given by (57) was introduced by Swamy (1970). Note that it is inconsistent in general if v_{it} is conditionally heteroskedastic. In addition, both estimators given by (56) and (57) will be inconsistent if errors are not mutually uncorrelated given regressors. Moreover, none of the three estimators of the variance of individual effects is non-negative definite by construction.

In practice, it may be important to empirically determine the order of the MA process of error variables. This is of special importance in order to estimate the variance of individual effects, as mispecifying the form of the variance-covariance matrix of errors would result in inconsistent estimates. We suggest a simple strategy for this purpose. For example, in order to test for the presence of independent errors, we propose to estimate $\widehat{\mathbf{Var}}\left(\mathbf{v}_i\right)$ in (52) assuming an MA(1) structure, by choosing the appropriate selection matrix $\mathbf{S}_2$. Then, one can use a Wald test of nullity of the the coefficients of the first off-diagonal of the variance-covariance matrix. This strategy is analogous to the one followed by Arellano and Bond (1991) in models with an heterogeneous intercept.

In specific cases, variance restrictions may be used to estimate common parameters $\boldsymbol{\theta}$ in model (9). Examples are the one-factor model (10), and the AR(1) model with heterogeneous

regressors (12). Conditional mean restrictions (17) can be complemented by the covariance restrictions:

$$\mathbf{Q}_i(\boldsymbol{\theta})\,\mathbf{E}\left[(\mathbf{y}_i-\mathbf{a}(\mathbf{X}_i;\boldsymbol{\theta}))(\mathbf{y}_i-\mathbf{a}(\mathbf{X}_i;\boldsymbol{\theta}))'|\mathbf{X}_i\right]\mathbf{Q}_i(\boldsymbol{\theta})'=\mathbf{Q}_i(\boldsymbol{\theta})\,\boldsymbol{\Omega}_i\mathbf{Q}_i(\boldsymbol{\theta})', \tag{58}$$

where $\mathbf{Q}_i(\boldsymbol{\theta})$ is given by (18). (58) will be uninformative about $\boldsymbol{\theta}$ if the variance-covariance matrix of errors $\boldsymbol{\Omega}_i$ is unrestricted. In that case, (17) may be enough for identification. However, the variance-covariance matrix of errors being unrestricted, the variance of individual effects will *not* be identified. In this context, restricting the covariance structure of errors may have two appealing features: it permits to learn about the variance of effects and the rest of their distribution, and it allows to use more restrictions to estimate $\boldsymbol{\theta}$, namely equation (58), putting less burden on the regressors. A simple case when expoiting (58) is necessary is when $\mathbf{Z}_i, \mathbf{X}_i$ are constant. The estimators of Holtz-Eakin *et al.* (1988) and Arellano and Bond (1991) can be viewed as exploiting those covariance restrictions.

A similar approach can be used to estimate higher-order moments. Using (38) together with the independent MA restriction with selection matrix $\mathbf{S}_3$, the vector of third-order cumulants of error variables can be estimated as:

$$\widehat{\boldsymbol{\kappa}}_3(\mathbf{v}_i)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_3\left[(\mathbf{Q}_i\otimes\mathbf{Q}_i\otimes\mathbf{Q}_i)\,\mathbf{S}_3\right]^{-}\left[\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\right]. \tag{59}$$

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

$$\begin{aligned}\widehat{\boldsymbol{\kappa}}_3(\boldsymbol{\gamma}_i) &= \frac{1}{N}\sum_{i=1}^{N}(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}})\otimes(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}})\otimes(\widehat{\boldsymbol{\gamma}}_i-\widehat{\boldsymbol{\gamma}})\\ &\quad-\frac{1}{N}\sum_{i=1}^{N}(\mathbf{H}_i\otimes\mathbf{H}_i\otimes\mathbf{H}_i)\,\mathbf{S}_3\left[(\mathbf{Q}_i\otimes\mathbf{Q}_i\otimes\mathbf{Q}_i)\,\mathbf{S}_3\right]^{-}\left[\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\otimes\widehat{\mathbf{v}}_i\right].\end{aligned} \tag{60}$$

Estimating fourth-order cumulants of errors and effects is more complicated. The reason is that conditional cumulants involve *products* of conditional expectations, to which the law of iterated expectations does not apply. For example, in the scalar case (where $\mathbf{E}(v|x)=0$):

$$\begin{aligned}\mathbf{E}\left[\kappa_4(v|x)\right] &= \mathbf{E}\left[\mathbf{E}\left(v^4|x\right)-3\mathbf{E}\left(v^2|x\right)^2\right]\\ &= \mathbf{E}\left(v^4\right)-3\mathbf{E}\left[\mathbf{E}\left(v^2|x\right)^2\right].\end{aligned}$$

When regressors are discrete and take a small number of values, conditional versions of the fourth-order analogs to (59) and (60) can be estimated for all the values of the regressors,

recovering conditional moments and hence unconditional ones by aggregation. When regressors take many values, however, nonparametric estimation of conditional expectation functions of the type $\mathbf{E}(v^2|x)$ will be necessary.

We end this section by noting that the whole analysis so far has been conditional on the form of heterogeneity in the model: q was assumed known, as well as which regressors have an heterogeneous impact across individuals. In practice it may be of interest to test for the presence of heterogeneity. Bonhomme (2008), generalizing a result by Orme and Yamagata (2006), shows that the critical values of a standard F-test of the null hypothesis that some of the regressors have commmon impacts across individuals remain valid under non-normality of the errors when N tends to infinity, for fixed T. Importantly, to apply this test one must make assumptions on the variance-covariance structure of error variables. It is not possible to test for the presence of heterogeneity under arbitrary correlation of the errors, because heterogeneous models are typically equivalent to homogeneous models with serial correlation (see Arellano, 2003, p.58).

5 Distributions

5.1 Identification

As in the case of the variance and higher-order moments, we impose two types of restrictions on model (2). First, we assume full independence between individual effects and errors given regressors, see (37). Second, we will restrict the dependence between error variables in a similar way as in section 4.

To derive the identification results, it will be 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.[10] The conditional characteristic function of $\mathbf{Y}$ given $\mathbf{X}$, given $\mathbf{X} = \mathbf{x}$, is defined as:

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

The following properties of characteristic functions will be useful (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*:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^L} \int \exp\left(-j\mathbf{t}'\mathbf{y}\right) \Psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) \mathbf{dt}. \tag{61}$$

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

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:

$$\Psi_{\mathbf{Y}_1+\mathbf{Y}_2|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = \Psi_{\mathbf{Y}_1|\mathbf{X}}(\mathbf{t}|\mathbf{x})\Psi_{\mathbf{Y}_2|\mathbf{X}}(\mathbf{t}|\mathbf{x}). \tag{62}$$

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

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

Theorem 3 *Suppose that (4) and (37) hold. Suppose also that the characteristic function of* $\mathbf{v}_i$ *given* $\mathbf{Z}_i, \mathbf{X}_i$ *is nonvanishing on* $\mathbf{R}^T$*. Then we have, for all* $\boldsymbol{\tau} \in \mathbf{R}^q$*:*

$$\Psi_{\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i) = \frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}{\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)} \tag{63}$$

and, unconditionally:

$$\Psi_{\boldsymbol{\gamma}_i(\boldsymbol{\tau})} = \mathbf{E}\left(\frac{\exp\left(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i\right)}{\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}\right). \tag{64}$$

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.

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 *Suppose in addition to the assumptions of Theorem 3 that the characteristic function of* $\boldsymbol{\gamma}_i$ *given* $\mathbf{X}_i$ *and* $\mathbf{Z}_i$ *is almost everywhere nonvanishing on* $\mathbf{R}^q$*. Then we have, for all* $\boldsymbol{\tau} \in \mathbf{R}^q$*:*

$$\log \Psi_{\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i) = \log \Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i) - \log \Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i). \tag{65}$$

Corollary (5) shows that the identification result for the distribution of effects is a generalization of the result for the first moments. Indeed, taking second-order derivatives in (65) at $\boldsymbol{\tau} = \mathbf{0}$ yields (22). Taking third and fourth-order derivatives yield (38) and (39), respectively.

Applying the inverse Fourier transform (61) we obtain the following corollary.

Corollary 5 *Under the assumptions of theorem 3 we have, for all q-dimensional vector* $\boldsymbol{\gamma}$*:*

$$f_{\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\gamma}|\mathbf{Z}_i,\mathbf{X}_i) = \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\frac{\Psi_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}{\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}\mathbf{d}\boldsymbol{\tau} \tag{66}$$

and, unconditionally:

$$f_{\boldsymbol{\gamma}_i}(\boldsymbol{\gamma}) = \frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\mathbf{E}\left(\frac{\exp(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i)}{\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}\right)\mathbf{d}\boldsymbol{\tau}. \tag{67}$$

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 the appendix for a derivation):

$$\begin{aligned} f_{\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\gamma}|\mathbf{Z}_i,\mathbf{X}_i) &= f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\gamma}|\mathbf{Z}_i,\mathbf{X}_i) \\ &\quad -\frac{1}{2}\operatorname{Tr}\left(\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\frac{\partial^2 f_{\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i}(\boldsymbol{\gamma}|\mathbf{Z}_i,\mathbf{X}_i)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right) + O_p\left(\frac{1}{T^2}\right), \end{aligned} \tag{68}$$

where Tr() is the trace operator. In the simple case where there are no common regressors, only the constant is heterogeneous, and v_{it} is i.i.d. with variance σ^2, this yields:

$$f_{\gamma_i}(\gamma) = f_{\widehat{\gamma}_i}(\gamma) - \frac{\sigma^2}{2T}\frac{d^2 f_{\widehat{\gamma}_i}(\gamma)}{d\gamma^2} + O\left(\frac{1}{T^2}\right). \tag{69}$$

Equation (68) allows a natural interpretation of the bias correction: in regions of high curvature (such as the mode of the distribution), the density of estimated effects understates the density of true effects. We shall illustrate this intuition in the application below.

We now consider the identification of the distribution of error variables. We define (minus one times) the vector of second derivatives of the log characteristic function of $\mathbf{Y}$ given $\mathbf{X}$ as:

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

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

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

Using (7) and (37), we obtain, assuming that the characteristic function of errors is non-vanishing on $\mathbf{R}^T$:

$$\boldsymbol{\kappa}_{\mathbf{Q}_i(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta})|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{t}|\mathbf{Z}_i,\mathbf{X}_i) = (\mathbf{Q}_i\otimes\mathbf{Q}_i)\,\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{Q}_i'\mathbf{t}|\mathbf{Z}_i,\mathbf{X}_i), \quad \mathbf{t}\in\mathbf{R}^T. \tag{70}$$

We study identification under the assumption that errors follow an independent moving average process of the form (28), where $u_{i,1-r}, ..., u_{iT}$ are mutually independent given regressors. Extensions to autoregressive and ARMA processes with independent underlying innovations is immediate. Lemma 2 in the appendix shows that, in an independent MA model, the partial derivatives of the log characteristic function of error variables are zero for all indices t and t' such that v_{it} and $v_{it'}$ are independent. It follows that there exists a m-dimensional vector of functions $\boldsymbol{\omega}_i(\mathbf{t})$ $(\mathbf{t} \in \mathbf{R}^T)$, possibly dependent on regressors, such that:

$$\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{t}|\mathbf{Z}_i, \mathbf{X}_i) = \mathbf{S}_2\boldsymbol{\omega}_i(\mathbf{t}), \quad \mathbf{t} \in \mathbf{R}^T. \tag{71}$$

The selection matrix $\mathbf{S}_2$ is the *same* that appeared in the covariance restrictions (27). Indeed, (71) evaluated at $\mathbf{t} = \mathbf{0}$ yields (27). Moreover, $m = (r+1)(T - r/2)$.

Let us denote as $\mathbf{Q}_{i1}, ..., \mathbf{Q}_{iT}$ the columns of matrix $\mathbf{Q}_i$. Clearly, (71) implies that $\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}$ is identified on the vector space spanned by $\mathbf{Q}_{i1}, ..., \mathbf{Q}_{iT}$. In order for identification to hold everywhere we need to make another assumption on the structure of $\mathbf{Q}_i$.

Assumption 1 *Let $(t_1, t_2) \in \{1, ..., T\}^2$. Every square submatrix of $\mathbf{Q}_i$ whose indices (s, s') are such that v_{is} and $v_{is'}$ are neither independent of v_{it_1} nor of v_{it_2} is non-singular.*

Let $\overline{m}(t_1, t_2)$ be the number of elements $s \in \{1, ..., T\}$ such that v_{is} is neither independent of v_{it_1} nor of v_{it_2}. Let also $\overline{m} = \max\{\overline{m}(t_1, t_2), (t_1, t_2)\}$. In particular, Assumption 1 requires that: $\text{rank}(\mathbf{Q}_i) \geq \overline{m}$. For an independent MA($r$) process, the maximum $\overline{m}(t_1, t_2)$ is attained for $t_1 = t_2 = T/2$, possibly rounded to an integer value, in which case: $\overline{m} = 2r + 1$. The order condition associated with Assumption 1 is thus: $T - q \geq 2r + 1$.

Combining (70) with (71) and using Assumption 1 we obtain the folowing identification theorem.

Theorem 4 *Suppose that (3), (4), (31), (71) and Assumption 1 hold. Suppose also that the characteristic function of errors given regressors, $\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}$, is non-vanishing on $\mathbf{R}^T$. Then $\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}$ is identified from (7) alone.*

The identification of $\Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}$ 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 is because the first derivative of the log-characteristic function at the origin is the

mean of the random variable, which is zero by (3). Hence also the identification of the density of error variables, using inverse Fourier transform as in Corollary 5.

To summarize the results so far, we have obtained the nonparametric identification of the distributions of individual effects and errors under two main conditions: the independence of effects and error variables, and conditional independence restrictions on errors that are sufficiently spaced. These results extend the ones in Kotlarski (1967) and Székely and Rao (2000) to cases where fully unrestricted individual effects, as well as conditioning regressors, are present.

To end the discussion of identification, we remark that within-group and between-group equations could be combined in a similar way as in subsection 3.4, differencing out the *distribution* of individual effects instead of the individual effects themselves. Using (37) it is easy to show that, if the characteristic functions of errors and individual effects have no real zeros, the following equation holds:

$$\mathbf{M}_i \boldsymbol{\kappa}_{\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta} | \mathbf{Z}_i, \mathbf{X}_i} \left(\mathbf{t} | \mathbf{Z}_i, \mathbf{X}_i \right) = \mathbf{M}_i \boldsymbol{\kappa}_{\mathbf{v}_i | \mathbf{Z}_i, \mathbf{X}_i} \left(\mathbf{t} | \mathbf{Z}_i, \mathbf{X}_i \right), \quad \mathbf{t} \in \mathbf{R}^T, \tag{72}$$

where $\mathbf{M}_i$ is given by (45). Evaluating (72) at $\mathbf{t} = \mathbf{0}$ yields (46). As in the case of variances, there are more identifying restrictions in (72) than in (70), which was derived from the within-group equation (7) alone. In particular, the rank and order conditions for identification can be relaxed when working with these equations. Namely, it can be shown that (31) is sufficient for the distribution of error variables to be identified from (72).

5.2 Estimation

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. We start with error variables. A natural possibility is to assume a flexible parametric family for errors, for example using normal mixtures, possibly allowing for conditional heteroskedasticity of a restricted form with respect to the regressors. Ghosal and Van der Vaart (2001, 2007) provide results on the ability of normal mixtures to approximate unknown densities. Imposing a parametric structure should not be seen as a severe limitation if the conditions of the identification theorems are satisfied, as their conclusions refer to the *nonparametric identification* of the distributions. Note that it is easy to implement maximum likelihood estimation when work-

ing with the within-group equations (7).[11] Following this approach, however, it does not seem straightforward to use the information contained in the additional moment restrictions (72).

Instead of postulating a parametric model for error variables, it may be possible to estimate their densities nonparametrically using characteristic-function based methods that have been proposed in the literature. Horowitz and Markatou (1996) estimate the distribution of error variables from within-group equations in a simple model with an individual-specific intercept and symmetric errors. Delaigle *et al.* (2008) have studied an alternative estimator for that model. Hall and Yao (2003) and Li and Vuong (1998) have proposed other estimators for the same model, the second one being generalized to independent multi-factor models by Bonhomme and Robin (2008b). We are not aware of extensions of these methods to deal with the presence of conditioning variables.

Once the density of errors (or their characteristic function) has been estimated, there remains to estimate the density of individual effects. The identifying equation (67) 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. There has been considerable work on nonparametric deconvolution in the statistics literature. In standard settings, many estimators are now available: standard Fourier inversion with kernel (Carroll and Hall, 1988, among many other references), wavelets (Fan and 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 theretical 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).

In the case of model (2), implementing a deconvolution approach to estimate $\boldsymbol{\gamma}_i$ is complicated by the presence of the conditioning regressors. A natural estimator based on (67)

[11] In practice, it can be useful to transform (7) into a system of $T-q$ equations, instead of T equations. To do that, left-multiply (7) by the $(T-q)$-by-T Cholesky root of $\mathbf{Q}_i$, say $\mathbf{A}_i$, that satisfies $\mathbf{Q}_i = \mathbf{A}_i'\mathbf{A}_i$.

is:

$$\widehat{f}_{\gamma_i}(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi)^q}\int \exp(-j\boldsymbol{\tau}'\boldsymbol{\gamma})\frac{\exp(j\boldsymbol{\tau}'\widehat{\boldsymbol{\gamma}}_i)}{\widehat{\Psi}_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{H}_i'\boldsymbol{\tau}|\mathbf{Z}_i,\mathbf{X}_i)}K_N(\boldsymbol{\tau})\mathbf{d}\boldsymbol{\tau}, \tag{73}$$

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

A potential problem with (73) is that, even if we expect $\widehat{f}_{\gamma_i}$ 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 $\mathbf{Z}_i$, $\mathbf{X}_i$, $i = 1, ..., N$. So the estimator could behave badly in the presence of strong heteroskedastocity (see Delaigle and Meister, 2008, for a related argument). Modifying and studying nonparametric deconvolution estimators to estimate the distribution of individual effects in model (2) is definitely outside of the scope of this paper. However, in simple cases and under more restrictive assumptions, existing estimators can be used for estimation. This is what we do in the next subsection.

To end this discussion of estimations, remark that the proposed strategy is sequential, starting with the density of error variables and then recovering the density of individual effects by deconvolution. An alternative is to estimate errors and effects *jointly*. A natural candidate would be to use sieve maximum likelihood (Ai and Chen, 2003, Hu and Schennach, 2008). The main difficulty with this approach is that one should account for the conditioning on possibly continuous $\mathbf{Z}_i, \mathbf{X}_i$.

5.3 A special case

We now discuss a special case of model (2), where $\mathbf{X}_i$ has two components: a constant, and a single binary regressor $\mathbf{x}_i$, where $x_{it} \in \{0, 1\}$. The model is:

$$y_{it} = \alpha_i + \beta_i x_{it} + \mathbf{z}_{it}'\boldsymbol{\delta} + v_{it}. \tag{74}$$

We study identification and estimation of the distribution of β_i in this model. In several applications it is of interest to know the distribution of the effect of a binary "treatment". In the application, x_{it} will a smoking/non smoking indicator.

We asume that $\boldsymbol{\delta}$ is identified, and work with $\widetilde{y}_{it} = y_{it} - \mathbf{z}_{it}'\boldsymbol{\delta}$. Consider a sequence $\mathbf{x}_i = (x_{i1}, ..., x_{iT})'$. Given a sequence of T zeros, or T ones, α_i and β_i are unidentified. We

thus focus on the subpopulation of individual units whose x's change over time. We assume that T is at least 3. So for a given sequence $\mathbf{x}_i$ there exist at least three indices t_1, t_2, t_3 such that: $x_{it_1} = x_{it_2}$, and $x_{it_1} \neq x_{it_3}$. Hence:

$$\widetilde{y}_{it_2} - \widetilde{y}_{it_1} = v_{it_2} - v_{it_1}, \tag{75}$$

$$\widetilde{y}_{it_3} - \widetilde{y}_{it_1} = (x_{it_3} - x_{it_1})\,\beta_i + v_{it_3} - v_{it_1}. \tag{76}$$

We assume that, for every sequence of x's, errors are i.i.d. It is possible to relax this assumption, by imposing conditional homoskedasticity of the errors instead of restrictions on their dynamics, the approach being very similar. In the i.i.d. case, $v_{it_2} - v_{it_1}$ and $v_{it_3} - v_{it_1}$ have the same distribution. Moreover, $x_{it_3} - x_{it_1}$ is either 1 or -1. So one can interpret (76) as a simple deconvolution equation, where the left-hand side is the sum of the unobserved β_i, and the independent error $v_{it_3} - v_{it_1}$. Moreover, the distribution of the latter is also that of $\widetilde{y}_{it_2} - \widetilde{y}_{it_1}$, and could be estimated by a simple kernel density estimator.

Having reformulated the estimation of the distribution of β_i in (74) as a simple deconvolution problem, it is now possible to use any existing deconvolution technique to estimate its density nonparametrically. In the application, we will use a method due to Mallows (2007). The method is based on simulation, and does not require to select a bandwidth. In addition, it is very simple to implement and hence could be of interest to practitioners. Lastly, our experiments show very good behavior in simulations, compared to standard kernel deconvolution.

Starting with two vectors A and C, sorted in ascending order, Mallows' algorithm aims at finding a vector B such that the sum of random draws from B and C yields a random draw from A. The algorithm works as follows, starting with a guess B_0 for B:

1. Permute the vector B_0 randomly, this yields $\widetilde{B}_1$.
2. Let $\widetilde{A}_1$ be the permutation of A sorted according to $\widetilde{B}_1 + C$.
3. Set $B_1 = \widetilde{A}_1 - C$. Iterate.

In our experiments, the algorithm always converged to a stationary chain after a short "burn-in" period, less than 10 initial iterations for a total of 1000.[12]

[12]Note that, for this algorithm to work, A and C must have the same size. If this is not the case, one may replace them by m bootstrap draws with replication from A and C, respectively, where m is the desired common size. In this way, we can use all the restrictions of the type (75) and (76) in one single algorithm.

6 Application

6.1 Model and data

Following Abrevaya (2006), we study the effect of smoking during pregnancy on birth outcomes. Abrevaya uses the Natality Data Sets for the US for the years 1990 and 1998. As there are no unique identifyers 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 (denoted by the index j) are matched to mothers (i). We estimate the following model:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \mathbf{z_{ij}}'\boldsymbol{\delta} + v_{ij}, \quad j = 1, ..., J_i. \tag{77}$$

In this equation, the dependent variable y_{ij} is the weight at birth of child j of mother i. x_{ij} is the smoking status of mother i when she was pregnant of child j, $x_{ij} = 1$ indicating that the mother was smoking. $\mathbf{z}_{ij}$ gathers other 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).

α_i and β_i in model (77) are mother-specific effects. They partly represent genetic endowments of the mother. A possible interpretation of (77) is as a production function, the "output" being the child and the "producer" being the mother. The "production technology" is then represented by the characteristics of the mother, 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 (3) requires that mothers will not stop smoking because one of their children had a low birthweight, i.e that the shocks v_{ij} are uncorrelated to 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 feed-back effects on the choice of inputs. We will try to relax the strict exogeneity assumption at the end of this section.

Abrevaya (2006) estimates a restricted version of (77), where β_i is homogeneous among individuals. To allow for heterogeneity, identification requires that $J_i \geq 3$. For this reason we focus on mothers who had at least 3 children during the period (1989-1998). In the dataset, J_i is exactly equal to 3 for all i. In addition, we need x_{ij} to vary for every mother. So we only consider mothers who changed smoking status between the three births. The final sample contains 1445 mothers.[13]

6.2 Results

We first check if there actually *is* heterogeneity of mother's responses to smoking in the data. An F-test of the null hypothesis that the β_i, $i = 1, ..., N$, are all equal, has a p-value of 0 (the F-statistic has a value of 1.32 for $(1444, 1437)$ degrees of freedom). Recall that the test is asymptotically valid when N tends to infinity for fixed T, even under non-normality (Bonhomme, 2008). This indicates the presence of heterogeneous β's in the sample we study.

Next, we estimate common parameters $\boldsymbol{\delta}$ in (77). For this, we use the generalized within-group estimator (50), 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 prenetal visits and the quality of prenatal care are never significant. The only significant covariate is the gender of the child, boys having higher weigth at birth.

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 (50) with $\mathbf{W}_i = \mathbf{I}_T$. The data is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births.

[13]Descriptive statistics show that this subsample is intermediate between the subsamples of women who always smoked, or never smoked. For example, women who smoke more are younger on average, and their children have lower weight at birth.

To interpret the mother-specific effects, we regress them on a set of covariates: the education of the mother, her married status, and the mean of the smoking indicators over the three births. Results are given in Table 2. Standard errors were corrected as explained in 4.1. 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, and justifies the fact of treating this effect as unobserved mother heterogeneity. Remark that the R^2 need to be corrected, see 4.1. For this, it is necessary to estimate the variance of the effects, which are the results that we present next. For comparison, the uncorrected R^2 are .055 and .005 for α_i and β_i, respectively.

Table 2: 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 data is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births.

Table 3 shows the estimates of the moments of α_i and β_i. The mean smoking effect of

-161 grams, computed by (51), is close to the fixed-effect estimate (i.e., imposing homogeneity of the β's in model 77) of -144 g found by Abrevaya (2006, Table IV).

Turning to variances, rows numbered (1) in Table 3 show the estimates of the Swamy formula, see (57). Both α_i and β_i show substantial dispersion. In particular, the standard deviation of β_i is 313 g. 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 Swamy formula yields a correlation of $-.47$ between α_i and β_i. So, mothers who have better (genetic) characteristic and have children with higher weight at birth (higher α_i) are *more* affected by smoking when they smoke (lower β_i).

Having 3 observations per mother requires to impose strong restrictions on the variance-covariance matrix of error variables. Indeed, (7) is a system of rank 1 $(= T - q)$, making it necessary to suppose that errors are i.i.d. Using the additional covariance restrictions (46) one can slighty relax the i.i.d. assumption. Rows numbered (2) in Table 3 show variance estimates when one permits the variances of errors for the first, second and third children to be different. It is easy to check that one cannot leave those three variances unrestricted, however. In rows numbered (2) we impose that the variance of errors for the jth child is $a + bj$, where a and b are scalars that we estimate from an empirical counterpart of (46). 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. We also tried to allow for limited correlation between errors, using (46), and found similar results. This suggests that the i.i.d. assumption is not rejected on these data.

We now comment the results for higher-order moments. Rows numbered (3) in Table 3 show the result of the estimation of skewness and kurtosis under the i.i.d. assumption, using the within-group equations (7). Clearly, the skewness of error variables is *not* identified from these equations. To estimate the moments of individual effects, 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 significant from the ones of the normal distribution (0 and 3, respectively).[14] Now, as explained in (3.4), the within-group equations do not contain all the information about error moments. Indeed, using first differences (75) and (76) it is easy to compute consistent estimates of the skewness and kurtosis of β_i that do

[14] 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).

not depend on errors to be symmetric. We show these estimates in (4) in Table 3. In that case also, the skewness and kurtosis are not significant from those of the normal.

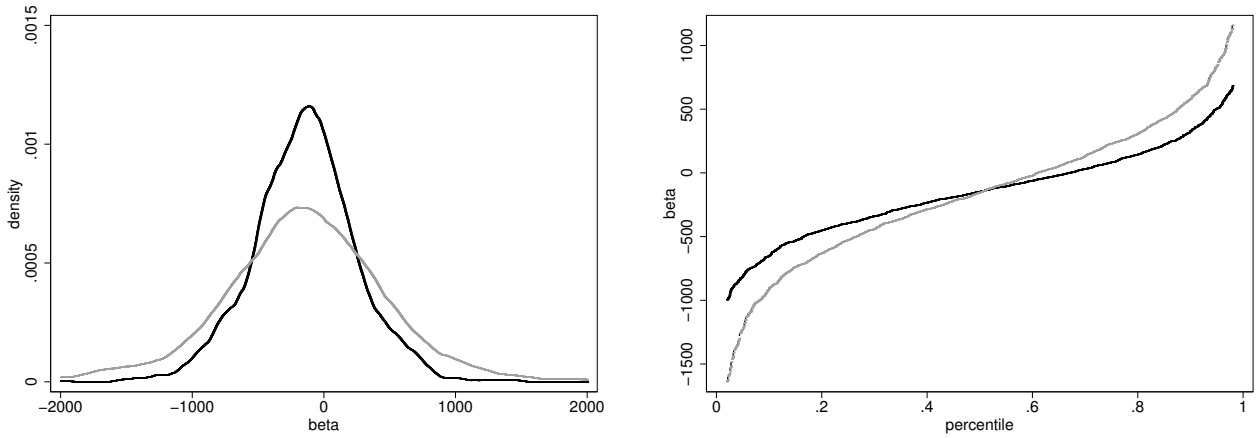
Table 3: Moments of α_i and β_i

Moment		Estimate	Standard error
		α_i	
Mean		2782	435
Variance	(1)	127647	15161
Variance	(2)	120423	24155
Skewness	(3)	-1.67	.428
Kurtosis	(3)	7.12	2.28
		β_i	
Mean		-161	17.0
Variance	(1)	98239	21674
Variance	(2)	85673	34550
Skewness	(3)	-1.29	.909
Skewness	(4)	-1.06	1.25
Kurtosis	(3)	-.34	7.84
Kurtosis	(4)	7.50	7.10
		α_i, β_i	
Covariance	(1)	-52661	14375
Covariance	(2)	-45437	24165

Note: Estimates of moments of α_i and β_i. The data is the "Matched panel data #3" in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births. (1) refers to the Swamy variance (57), (2) uses the additional restrictions (46) and allows for unconditional heteroskedasticity, (3) restricts the skewness of error variables to be zero, (4) corresponds to estimation in first differences.

We then use the strategy outlined in 5.3, based on Mallows' (2007) algorithm, to estimate the density of β_i nonparametrically. The results are shown in Figure 1, where we also plot the estimated density of the least-square estimates $\widehat{\beta}_i$ for comparison (in light print). The bottom graph in Figure 1 shows the quantile estimates of β_i and $\widehat{\beta}_i$. Correcting for sample noise in the estimation of the density of the smoking effect leads to a strikingly different picture. The density of β_i has much lower variance than that of $\widehat{\beta}_i$, and a much higher mode, consistently with (68). Comparison with the normal (not shown) shows some evidence of "peakedness" of β_i. In addition, our method allows to estimate the smoking effects at different quantiles. When corrected for the presence of sample noise, the effect is mostly

Figure 1: Density and quantiles of β_i (dark) and $\widehat{\beta}_i$ (light)



Note: Density estimate obtained using Mallows' (2007) simulation algorithm (for β_i), and standard kernel (for $\widehat{\beta}_i$). Quantile estimates obtained using inversion of the cumulative distribution function.

negative (up to percentile 75), and reaches very negative values for some mothers (around 500 g at percentile 20).

6.3 Predeterminedness of smoking behavior

The previous results have been derived under the assumption that the smoking status is strictly exogenous. To relax the strict exogeneity assumption, we need to make assumptions on the correlation between individual effects and regressors in order to perserve point identification of the moments, see subsection 2.3.

We consider model (77) and assume that β_i is independent of regressors $x_{ij}, \mathbf{z}_{ij}$ and errors v_{ij}, but make no assumption on α_i. Taking first differences yields:

$$\Delta y_{ij} = \beta_i \Delta x_{ij} + \Delta \mathbf{z_{ij}}' \boldsymbol{\delta} + \Delta v_{ij}, \tag{78}$$

where $\Delta y_{ij} = y_{ij} - y_{i,j-1}$.

Predeterminedness of x_{ij} means that: $\mathbf{E}\left(v_{ij} | x_{ij}, x_{i,j-1}, ..., \mathbf{Z}_i\right) = 0$. Under this condition, we show in the appendix that consistent estimates of $\beta = \mathbf{E}\left(\beta_i\right)$ and $\boldsymbol{\delta}$ can be computed using an Instrumental Variables (IV) regression of Δy_{ij} on Δx_{ij} and $\Delta \mathbf{z_{ij}}$, using as instruments $x_{i,j-1}, ..., \mathbf{Z}_i$. We call the coefficient estimates $\widehat{\beta}$ and $\widehat{\boldsymbol{\delta}}$. Next, assuming that v_{ij} is independent of $x_{i,j-1}, ..., \mathbf{Z}_i$ (which is a stronger assumption than predeterminedness) we show that the

variance of β_i can be consistently estimated by an IV regression of $\left(\Delta y_{ij} - \widehat{\beta}\Delta x_{ij} - \Delta \mathbf{z}_{ij}'\widehat{\boldsymbol{\delta}}\right)^2$ on $(\Delta x_{ij})^2$ and a constant, using as instruments $x_{i,j-1}, ..., \mathbf{Z}_i$.

We then apply this approach to Abrevaya's (2006) dataset. It is *not* possible to test for the strict exogeneity of the smoking indicator in the model with two mother-specific effects, because mothers have at most three children in the data. However, a simple regression of birthweight on current and *future* smoking status with mother-specific effects, with no regressors, yields a coefficient of -35.9 with a standard error of 25.3 for smoking status during the next pregnancy (i.e., $x_{i,j+1}$). This provides some evidence of predeterminedness, although the coefficient is not significant at 10%. Accounting for smoking to be predetermined, we then estimate the mean of β_i to be -158 with a standard error of 27.4. This is very close to the mean effect reported in Table 3. The variance is estimated to be 293239 with a standard error of 187448. This corresponds to a standard deviation of 541 g, and is quite higher than the one reported in Table 3. However, the two variances are not statistically different. Despite the limitation of this exercise, due to the small number of children per mother and the relatively small sample size, this confirms that there is considerable heterogeneity in the effect of smoking during pregnancy on birthweight.

7 Conclusion

We have derived conditions under which the distribution of heterogeneous components can be consistently estimated in a class of panel data models with multiple sources of heterogeneity. For our identification results to apply, three main conditions must be met: the model must be linear in the individual effects, regressors with individual-specific coefficients must be strictly exogenous, and the dependence structure of error variables must be restricted. Under these conditions, we prove the nonparametric identification of the full distribution of the effect of a covariate, and of that of error variables. Crucial for this result to hold is the panel data setting, which allows to observe the same individual with various values of the covariate.

We have proposed a nonparametric estimator of the density of individual-specific effects in a special case. Extending this approach to more general settings where $\mathbf{x}_{it}$ is continuous and errors are not i.i.d. is not immediate, as the available statistical methods should be extended to account for the conditioning on regressors. There seems to be interesting work to do in that direction.

Relaxing the model assumptions also seems important. In a companion paper, we study

a method to deal with nonlinear panel data models with continuous dependent variables. The main insight comes from subsection 3.4 above: even if it is not possible to difference the individual effects out in a nonlinear setting, it may be possible to difference out their *distribution*. This approach should be applicable to a class of models with predetermined regressors.

Lastly, in models where the dependent variable is not continuous, one is very likely to loose fixed-T identification. A fixed-T approach would then consist in characterizing the *identified bounds* of the distributional features of interest (Honoré and Tamer, 2006), requiring a very different analysis.

References

[1] 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.

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

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

[4] Amemiya, T. (1985): *Advanced Econometrics*, Blackwell, Oxford.

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

[6] Arellano, M., and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277-297.

[7] Arellano, M., and 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.

[8] 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.

[9] Bai, J. (2006): "Panel Data Models with Interactive Fixed Effects," *mimeo.*

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

[11] Becker, G.S., Grossman, M., and K.M. Murphy (1994): "An Empirical Analysis of Cigarette Addiction," *American Economic Review*, vol. 84(3), 396-418.

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

[13] Bonhomme, S. (2008): "A Test of Homogeneity in Random Coefficients Panel Data Models," *mimeo.*

[14] Bonhomme, S., and J. M. Robin (2008a): "Consistent Noisy Independent Component Analysis," *mimeo.*

[15] Bonhomme, S., and J. M. Robin (2008b): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," *mimeo.*

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

[17] Carrasco, M., and J.P. Florens (2007): "Spectral Method for Deconvolving a Density," *mimeo.*

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

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

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

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

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

[23] Dasgupta, A. (2008): *Asymptotic Theory of Statistics and Probability*, Springer Texts in Statistics.

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

[25] 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.

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

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

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

[29] Deschênes, O., and M. Greenstone (2007): "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather," *American Economic Review*, 97(1), 354-385.

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

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

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

[33] 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.

[34] 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.

[35] Goldberger, A.S. (1991): *A Course in Econometrics*, Harvard University Press.

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

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

[38] Guvenen, F. (2008): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, forthcoming.

[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] Harville, D.A. (1974): "Optimal Procedures for some Constrained Selection Problems," *Journal of the American Statistical Association,* 69, 446-452.

[43] Harville, D.A. (1977): "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association,* 72, 320-340.

[44] 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.

[45] Heckman, J.J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, vol. 73(3), 669-738.

[46] Hoderlein, S., Klemelä, J., and E. Mammen (2007): "Reconsidering the Random Coefficient Model", *mimeo.*

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

[48] Honore, B., and E. Tamer (2006): "Bounds on Parameters in Dynamic discrete-Choice Models,"*Econometrica*, 74(3), 611-629.

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

[50] 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).

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

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

[53] Imbens, G.W., and J.D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467-75.

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

[55] 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.

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

[57] Laird, N.M., and J.H. Ware (1982): "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

[58] Lemieux, T. (1998): "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection," *Journal of labor Economics*, 16, 261–291.

[59] 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.

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

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

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

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

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

[65] 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.

[66] 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.

[67] Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46(1), 69-85.

[68] 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.

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

[70] Orme, C.D., and T. Yamagata (2006): "The Asymptotic Distribution of the F-test Statistic for Individual Effects," *Econometrics Journal*, 9(3), 404-422.

[71] Patterson, H.D., and R. Thompson (1971): "Recovery of Interblock Information when Block Sizes are Unequal," *Biometrika*, 58, 545-554.

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

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

[74] Smith, P.J. (1995): "A Recursive Formulation of the Old Problem of Obtaining Moments from Cumulants and Vice Versa," *The American Statistician*, 49(2), 217-218.

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

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

[77] Vella, F., and M. Verbeek (1998): "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics*, 13, 163-183.

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

[79] 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

Non-identification in the predetermined case. Here we emphasize the lack of identification of $\boldsymbol{\delta}$ and $\mathbf{E}(\beta_i)$ in model (13). The two identifying equations for these quantities are:

$$\begin{aligned}\mathbf{E}(\Delta y_{i3}|x_{i1},x_{i2},\mathbf{Z}_i) &= \mathbf{E}(\beta_i\Delta x_{i3}|x_{i1},x_{i2},\mathbf{Z}_i)+\Delta\mathbf{z}_{i3}'\boldsymbol{\delta}\\ &= \mathbf{P}(\Delta x_{i3}=1|x_{i1},x_{i2},\mathbf{Z}_i)\,\mathbf{E}(\beta_i|\Delta x_{i3}=1,x_{i1},x_{i2},\mathbf{Z}_i)\\ &\quad -\mathbf{P}(\Delta x_{i3}=-1|x_{i1},x_{i2},\mathbf{Z}_i)\,\mathbf{E}(\beta_i|\Delta x_{i3}=-1,x_{i1},x_{i2},\mathbf{Z}_i)+\Delta\mathbf{z}_{i3}'\boldsymbol{\delta},\end{aligned}$$

and:

$$\begin{aligned}\mathbf{E}(\Delta y_{i2}|x_{i1},\mathbf{Z}_i) &= \mathbf{E}(\beta_i\Delta x_{i2}|x_{i1},\mathbf{Z}_i)+\Delta\mathbf{z}_{i2}'\boldsymbol{\delta}\\ &= \mathbf{P}(\Delta x_{i2}=1|x_{i1},\mathbf{Z}_i)\,\mathbf{E}(\beta_i|\Delta x_{i2}=1,x_{i1},\mathbf{Z}_i)\\ &\quad -\mathbf{P}(\Delta x_{i2}=-1|x_{i1},\mathbf{Z}_i)\,\mathbf{E}(\beta_i|\Delta x_{i2}=-1,x_{i1},\mathbf{Z}_i)+\Delta\mathbf{z}_{i2}'\boldsymbol{\delta}.\end{aligned}$$

Clearly, $\boldsymbol{\delta}$ and $\mathbf{E}(\boldsymbol{\beta}_i)$ are not identified unless one imposes restrictions on the conditional mean of $\boldsymbol{\beta}_i$ given $\mathbf{X}_i$ and $\mathbf{Z}_i$.

Proof of Proposition 1. Assumption (4) implies that $\mathbf{H}_i$ and $\mathbf{Q}_i$ exist. We have, using (3):

$$\mathbf{E}(\mathbf{Q}_i(\mathbf{y}_i-\mathbf{Z}_i\boldsymbol{\delta})|\mathbf{Z}_i,\mathbf{X}_i) = \mathbf{E}(\mathbf{Q}_i\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i)$$

Likewise, again using assumption (3):

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

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

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

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, (15) implies that $\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i)=\mathbf{E}(\boldsymbol{\gamma}_i)$, so:

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

The conclusion follows.

Proof of Corollary 2. Clearly:

$$\mathbf{Q}_i(\boldsymbol{\theta})\,(\mathbf{y}_i-\mathbf{a}(\mathbf{X}_i,\boldsymbol{\theta})) = \mathbf{Q}_i(\boldsymbol{\theta})\,\mathbf{v}_i.$$

The conclusion comes from: $\mathbf{E}(\mathbf{v}_i|\mathbf{X}_i)=\mathbf{0}$.

Proof of Theorem 1.

$$\begin{aligned}\mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i|\mathbf{Z}_i,\mathbf{X}_i) &= \mathbf{Var}(\boldsymbol{\gamma}_i+\mathbf{H}_i\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i)\\ &= \mathbf{Var}(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i)+\mathbf{Var}(\mathbf{H}_i\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i)\\ &= \mathbf{Var}(\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i)+\mathbf{H}_i\boldsymbol{\Omega}_i\mathbf{H}_i'\end{aligned}$$

where we have used (20) in the second equality. Hence (22). Unconditionally we have:

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

Proof of Corollary 3. In the particular case where errors are i.i.d. independent of $(\mathbf{Z}_i, \mathbf{X}_i)$ with variance σ^2 a solution is obtained by applying the trace operator to:

$$\mathbf{E}((\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})') = \sigma^2\mathbf{E}(\mathbf{Q}_i\mathbf{Q}_i') = \sigma^2\mathbf{E}(\mathbf{Q}_i).$$

As $\mathrm{Tr}(\mathbf{Q}_i) = T - q$, and $\mathrm{Tr}\left((\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})'\right) = (\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})'(\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta})$, this yields:

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

A lemma for section 3.

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

i) $\mathrm{rank}\{[(\mathbf{I}_n - \mathbf{P}) \otimes (\mathbf{I}_n - \mathbf{P})]\mathbf{D}_n\} = \frac{(n-p)(n-p+1)}{2}$,

ii) $\mathrm{rank}[(\mathbf{I}_{n^2} - \mathbf{P}\otimes\mathbf{P})\mathbf{D}_n] = \frac{n(n+1)}{2} - \frac{p(p+1)}{2}$.

Proof. Part *i*). Because of idempotence: $\mathrm{rank}(\mathbf{I}_n - \mathbf{P}) = n - p$. 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, $\{\mathbf{v}_i \otimes \mathbf{v}_j, (i,j) \in \{1,...,p\}^2\}$ forms a linearly independent family. So does $\{\mathbf{v}_i \otimes \mathbf{v}_j, (i,j) \in \{1,...,p\}^2, i \leq j\}$. As this family has $(n-p)(n-p+1)/2$ elements, the conclusion follows.

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

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

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

Hence, because $\mathbf{D}_n$ has full column rank, the rank of: $(\mathbf{I}_{n^2} - \mathbf{P}\otimes\mathbf{P})\mathbf{D}_n$ is equal to that of: $\mathbf{B}_n = \mathbf{I}_{\frac{n(n+1)}{2}} - \mathbf{D}_n^-(\mathbf{P}\otimes\mathbf{P})\mathbf{D}_n$. 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): $\mathrm{rank}(\mathbf{B}_n) = \mathrm{Tr}(\mathbf{B}_n)$. Now:

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

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

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

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

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

This ends the proof. ■

Proof of Equation (50). We have:

$$\mathbf{Q}_i\mathbf{y}_i = \mathbf{Q}_i\mathbf{Z}_i\boldsymbol{\delta} + \mathbf{Q}_i\mathbf{v}_i.$$

Let $\mathbf{Q}_i = \mathbf{A}_i'\mathbf{A}_i$, where $\mathbf{A}_i$ is $(T-q)$-by-T, with rank $T-q$. As $\mathbf{A}_i'$ is full-column rank we have:

$$\mathbf{A}_i\mathbf{y}_i = \mathbf{A}_i\mathbf{Z}_i\boldsymbol{\delta} + \mathbf{A}_i\mathbf{v}_i.$$

Then $\mathbf{Var}\left(\mathbf{A}_i\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i\right) = \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i'$. So

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

Now:

$$\mathbf{A}_i'\left(\mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i'\right)^{-1}\mathbf{A}_i = \mathbf{Q}_i^{\boldsymbol{\Omega}_i}.$$

To see this, note that letting $\overline{\mathbf{A}}_i = \mathbf{A}_i\boldsymbol{\Omega}_i^{1/2}$ and $\overline{\mathbf{X}}_i = \boldsymbol{\Omega}_i^{-1/2}\mathbf{X}_i$, the previous equation can be written as

$$\overline{\mathbf{A}}_i'\left(\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i'\right)^{-1}\overline{\mathbf{A}}_i = \mathbf{I}_T - \overline{\mathbf{X}}_i\left(\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\right)^{-1}\overline{\mathbf{X}}_i'.$$

This is because

$$\begin{aligned}
\mathbf{I}_T &= \begin{pmatrix}\mathbf{I}_{T-q} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_q\end{pmatrix}\\
&= \begin{pmatrix}\left(\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i'\right)^{-1}\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i' & \mathbf{0}\\ \mathbf{0} & \left(\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\right)^{-1}\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\end{pmatrix}\\
&= \begin{pmatrix}\left(\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i'\right)^{-1}\overline{\mathbf{A}}_i\\ \left(\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\right)^{-1}\overline{\mathbf{X}}_i'\end{pmatrix}\begin{pmatrix}\overline{\mathbf{A}}_i' & \overline{\mathbf{X}}_i\end{pmatrix}
\end{aligned}$$

as $\overline{\mathbf{A}}_i\overline{\mathbf{X}}_i = \mathbf{0}$, so that:

$$\begin{aligned}
\mathbf{I}_T &= \begin{pmatrix}\overline{\mathbf{A}}_i' & \overline{\mathbf{X}}_i\end{pmatrix}\begin{pmatrix}\left(\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i'\right)^{-1}\overline{\mathbf{A}}_i\\ \left(\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\right)^{-1}\overline{\mathbf{X}}_i'\end{pmatrix}\\
&= \overline{\mathbf{A}}_i'\left(\overline{\mathbf{A}}_i\overline{\mathbf{A}}_i'\right)^{-1}\overline{\mathbf{A}}_i + \overline{\mathbf{X}}_i\left(\overline{\mathbf{X}}_i'\overline{\mathbf{X}}_i\right)^{-1}\overline{\mathbf{X}}_i'.
\end{aligned}$$

Consistent standard errors for the linear projection coefficients. The regression coefficients in:

$$\gamma_{\ell i} = \mathbf{F}_i'\boldsymbol{\pi}_\ell + \xi_{\ell i}, \quad \ell = 1, ..., q \tag{A1}$$

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

$$\boldsymbol{\pi}_\ell = \left[\mathbf{E}\left(\mathbf{F}_i\mathbf{F}_i'\right)\right]^{-1}\mathbf{E}\left(\mathbf{F}_i\boldsymbol{\gamma}_{\ell i}\right), \tag{A2}$$

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

$$\widehat{\boldsymbol{\pi}}_\ell = \left(\frac{1}{N}\sum_{i=1}^N \mathbf{F}_i\mathbf{F}_i'\right)^{-1}\frac{1}{N}\sum_{i=1}^N \mathbf{F}_i\widetilde{\gamma}_{\ell i}, \tag{A3}$$

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

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

We have:

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

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

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = \left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Q}_i\mathbf{Z}_i\right)^{-1}\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Q}_i\mathbf{v}_i. \tag{A4}$$

Combining the two expressions we get

$$\begin{aligned}\boldsymbol{\Psi}_N\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right) &= \left(\frac{1}{N}\sum_{i=1}^N \mathbf{F}_i\xi_{\ell i}\right) + \begin{pmatrix}\boldsymbol{\Phi}_N & \mathbf{I}\end{pmatrix}\frac{1}{N}\sum_{i=1}^N \begin{pmatrix}\mathbf{Z}_i'\mathbf{Q}_i\\ \mathbf{F}_i\mathbf{h}_{i\ell}'\end{pmatrix}\mathbf{v}_i\\ &= \left(\frac{1}{N}\sum_{i=1}^N \mathbf{F}_i\xi_{\ell i}\right) + \overline{\boldsymbol{\Phi}}_N\frac{1}{N}\sum_{i=1}^N \mathbf{C}_{i\ell}\mathbf{v}_i\end{aligned}$$

where

$$\boldsymbol{\Phi}_N = -\left(\frac{1}{N}\sum_{i=1}^N \mathbf{F}_i\mathbf{h}_{i\ell}'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Q}_i\mathbf{Z}_i\right)^{-1}$$

and $\overline{\boldsymbol{\Phi}}_N = \begin{pmatrix}\boldsymbol{\Phi}_N & \mathbf{I}\end{pmatrix}$, and

$$\mathbf{C}_{i\ell} = \begin{pmatrix}\mathbf{Z}_i'\mathbf{Q}_i\\ \mathbf{F}_i\mathbf{h}_{i\ell}'\end{pmatrix}.$$

Note that if $\mathbf{v}_i$ is uncorelated with the effects given the regressors, i.e. if either (20) or (21) holds given $\mathbf{Z}_i, \mathbf{X}_i, \mathbf{F}_i$:

$$\mathbf{E}\left(\mathbf{F}_i\xi_{\ell i}\mathbf{v}_i'\mathbf{C}_{i\ell}'\right) = \mathbf{0}.$$

Therefore, the asymptotic variance of $\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right)$ is

$$\mathbf{Avar}\left[\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right)\right] = \boldsymbol{\Psi}_0^{-1}\mathbf{E}\left(\xi_{\ell i}^2\mathbf{F}_i\mathbf{F}_i'\right)\boldsymbol{\Psi}_0^{-1} + \boldsymbol{\Psi}_0^{-1}\overline{\boldsymbol{\Phi}}_0\mathbf{E}\left(\mathbf{C}_{i\ell}\boldsymbol{\Omega}_i\mathbf{C}_{i\ell}'\right)\overline{\boldsymbol{\Phi}}_0'\boldsymbol{\Psi}_0^{-1} \tag{A5}$$

where $\boldsymbol{\Psi}_0 = \mathbf{E}\left(\mathbf{F}_i\mathbf{F}_i'\right)$, $\overline{\boldsymbol{\Phi}}_0 = \begin{pmatrix} \boldsymbol{\Phi}_0 & \mathbf{I} \end{pmatrix}$, and $\boldsymbol{\Phi}_0 = \mathbf{E}\left(\mathbf{F}_i\mathbf{h}_{i\ell}'\mathbf{Z}_i\right)\left[\mathbf{E}\left(\mathbf{Z}_i'\mathbf{Q}_i\mathbf{Z}_i\right)\right]^{-1}$.

The term $\mathbf{E}\left(\xi_{\ell i}^2\mathbf{F}_i\mathbf{F}_i'\right)$ cannot be directly estimated because $\gamma_{\ell i}$ is unobservable. Let us consider the following estimator that would be produced by a regression routine:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \widetilde{\xi}_{\ell i}^2\mathbf{F}_i\mathbf{F}_i' &= \frac{1}{N}\sum_{i=1}^N \left(\widetilde{\gamma}_{\ell i} - \mathbf{F}_i'\widehat{\boldsymbol{\pi}}_\ell\right)^2\mathbf{F}_i\mathbf{F}_i' \\
&= \frac{1}{N}\sum_{i=1}^N \left[\xi_{\ell i} + \mathbf{h}_{i\ell}'\mathbf{v}_i - \mathbf{F}_i'\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right) - \mathbf{h}_{i\ell}'\mathbf{Z}_i\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right)\right]^2\mathbf{F}_i\mathbf{F}_i' \\
&\xrightarrow{p} \mathbf{E}\left(\xi_{\ell i}^2\mathbf{F}_i\mathbf{F}_i'\right) + \mathbf{E}\left(\mathbf{h}_{i\ell}'\boldsymbol{\Omega}_i\mathbf{h}_{i\ell}\mathbf{F}_i\mathbf{F}_i'\right).
\end{aligned}$$

Thus,

$$\mathbf{E}\left(\xi_{\ell i}^2\mathbf{F}_i\mathbf{F}_i'\right) = \mathbf{E}\left[\left(\mathbf{h}_{i\ell}'\left[\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right] - \mathbf{F}_i'\boldsymbol{\pi}_\ell\right)^2\mathbf{F}_i\mathbf{F}_i'\right] - \mathbf{E}\left(\mathbf{h}_{i\ell}'\boldsymbol{\Omega}_i\mathbf{h}_{i\ell}\mathbf{F}_i\mathbf{F}_i'\right),$$

and

$$\begin{aligned}
\mathbf{Avar}\left[\sqrt{N}\left(\widehat{\boldsymbol{\pi}}_\ell - \boldsymbol{\pi}_\ell\right)\right] &= \boldsymbol{\Psi}_0^{-1}\mathbf{E}\left[\left(\mathbf{h}_{i\ell}'\left[\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}\right] - \mathbf{F}_i'\boldsymbol{\pi}_\ell\right)^2\mathbf{F}_i\mathbf{F}_i'\right]\boldsymbol{\Psi}_0^{-1} \\
&\quad + \boldsymbol{\Psi}_0^{-1}\left[\overline{\boldsymbol{\Phi}}_0\mathbf{E}\left(\mathbf{C}_{i\ell}\boldsymbol{\Omega}_i\mathbf{C}_{i\ell}'\right)\overline{\boldsymbol{\Phi}}_0' - \mathbf{E}\left(\mathbf{h}_{i\ell}'\boldsymbol{\Omega}_i\mathbf{h}_{i\ell}\mathbf{F}_i\mathbf{F}_i'\right)\right]\boldsymbol{\Psi}_0^{-1}.
\end{aligned} \tag{A6}$$

The conclusion is that ordinary robust standard errors obtained when regressing $\widetilde{\boldsymbol{\gamma}}_{\ell i}$ on $\mathbf{F}_i$ are inconsistent whith a bias term provided by the second term on the right-hand side of (A6). In the special case where there is no $\mathbf{Z}_i$ and all parameters are individual-specific in (2), the bias term is zero. So the inconsistency is due to the fact that $\boldsymbol{\delta}$ is estimated.

Lastly, 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

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

Proof of theorem 3. Let $\boldsymbol{\tau} \in \mathbf{R}^q$. Using (8) and assumption (37) we obtain:

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

If $\Psi_{\mathbf{v}_i}$ is almost everywhere nonvanishing we obtain (63). Moreover, (64) follows from taking expectations:

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

Proof of corollary 5. Inverse Fourier transformation yields:

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

The unconditional result is similarly obtained.

Proof of equation (68). Under regularity conditions, we have, for all $\boldsymbol{\tau}\in\mathbf{R}^q$:

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

So:

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

where the last equality comes from taking second derivatives in (66).

A lemma for subsection 5.1. Here we extend Lemma 1 in Bonhomme and Robin (2008a). 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-by-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 2 *Let $(i,j)\in\{1,...,L\}^2$ such that Y_i and Y_j are independent. Then:*

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

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

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

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

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

■

Proof of theorem 4. Clearly, because of (31), (70) and (71): $\boldsymbol{\omega}_i(\mathbf{Q}_i'\mathbf{t})$, $\mathbf{t} \in \mathbf{R}^T$, is identified. Moreover, Lemma 2 shows that $\partial \log \Psi_{\mathbf{v}_i}(\mathbf{t})/\partial t_1$ depends only on the indices t_2 such that v_{it_1} and v_{it_2} are not independent. Hence, $\partial^2 \log \Psi_{\mathbf{v}_i}(\mathbf{t})/\partial t_1 \partial t_2$ depends only on the indices s such that v_{is} and v_{it_1} are not independent, and v_{is} and v_{it_2} are not independent. Let us call $\mathcal{I}$ the set of such indices s. Let $\omega_{i,t_1,t_2}(\mathbf{t})$ be the element of $\boldsymbol{\omega}_i(\mathbf{t})$ corresponding to the cross-derivative with indices (t_1, t_2). $\omega_{i,t_1,t_2}(\mathbf{t})$ is only a function of t_s, for s in $\mathcal{I}$. So, $\omega_{i,t_1,t_2}(\mathbf{Q}_i'\mathbf{t})$ is a function of $\mathbf{Q}_{is}'\mathbf{t}$, for s in $\mathcal{I}$. Now, Assumption 1 implies that the submatrix of $\mathbf{Q}_i'\mathbf{Q}_i = \mathbf{Q}_i$ with indices in $\mathcal{I}$ is non singular. Hence $\omega_{i,t_1,t_2}(\mathbf{t})$ is identified for all $\mathbf{t} \in \mathbf{R}^T$. Repeating the argument for all t_1, t_2 yields the identification of $\boldsymbol{\omega}_i(\mathbf{t})$ for all $\mathbf{t} \in \mathbf{R}^T$.

It follows that $\boldsymbol{\kappa}_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{t}|\mathbf{Z}_i, \mathbf{X}_i)$ is identified for all $\mathbf{t} \in \mathbf{R}^T$. By successive integration and using that, because of (3):

$$\frac{\partial \log \Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{0}|\mathbf{Z}_i, \mathbf{X}_i)}{\partial \mathbf{t}} = \mathbf{E}(\mathbf{v}_i|\mathbf{Z}_i, \mathbf{X}_i) = \mathbf{0},$$

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

$$\log \Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{0}|\mathbf{Z}_i, \mathbf{X}_i) = 0,$$

it follows that the characteristic function of errors is identified.

Proof of equation (72). Let $\mathbf{t} \in \mathbf{R}^T$. Using (8) and assumption (37) we obtain:

$$\log \Psi_{\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{t}|\mathbf{Z}_i, \mathbf{X}_i) = \log \Psi_{\boldsymbol{\gamma}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{X}_i'\mathbf{t}|\mathbf{Z}_i, \mathbf{X}_i) + \log \Psi_{\mathbf{v}_i|\mathbf{Z}_i,\mathbf{X}_i}(\mathbf{t}|\mathbf{Z}_i, \mathbf{X}_i).$$

Taking second derivatives and left-multiplying by $\mathbf{M}_i$ yields (72).

Predeterminedness, subsection 6.3. Taking conditional expectations in (78) yields:

$$\begin{aligned} \mathbf{E}(\Delta y_{ij}|x_{i,j-1}, ..., \mathbf{Z}_i) &= \mathbf{E}(\beta_i \Delta x_{ij}|x_{i,j-1}, ..., \mathbf{Z}_i) + \Delta \mathbf{z_{ij}}'\boldsymbol{\delta} \\ &= \mathbf{E}(\beta_i)\,\mathbf{E}(\Delta x_{ij}|x_{i,j-1}, ..., \mathbf{Z}_i) + \Delta \mathbf{z_{ij}}'\boldsymbol{\delta}. \end{aligned}$$

So, consistent estimates of $\beta = \mathbf{E}(\beta_i)$ and $\boldsymbol{\delta}$ are given by an Instrumental Variables (IV) regression of Δy_{ij} on Δx_{ij} and $\Delta \mathbf{z_{ij}}$, using as instruments $x_{i,j-1}, ..., \mathbf{Z}_i$. We call the coefficient estimates $\widehat{\beta}$ and $\widehat{\boldsymbol{\delta}}$.

We can proceed similarly to recover the variance of β_i. We have:

$$\Delta y_{ij} - \beta \Delta x_{ij} - \Delta \mathbf{z_{ij}}' \boldsymbol{\delta} = (\beta_i - \beta) \Delta x_{ij} + \Delta v_{ij},$$

so:

$$\begin{aligned} \mathbf{E}\left[\left(\Delta y_{ij} - \beta \Delta x_{ij} - \Delta \mathbf{z_{ij}}' \boldsymbol{\delta}\right)^2 | x_{i,j-1}, ..., \mathbf{Z}_i\right] &= \mathrm{Var}(\beta_i)\, \mathbf{E}\left[(\Delta x_{ij})^2 | x_{i,j-1}, ..., \mathbf{Z}_i\right] \\ &\quad + \mathbf{E}\left[(\Delta v_{ij})^2 | x_{i,j-1}, ..., \mathbf{Z}_i\right], \end{aligned}$$

where we have used that

$$\mathbf{E}\left((\beta_i - \beta) | \Delta x_{ij}, v_{ij}, x_{i,j-1}, ..., \mathbf{Z}_i\right) = 0.$$

Assuming that v_{ij} is independent of $x_{i,j-1}, ..., \mathbf{Z}_i$ (which is a stronger assumption than predeterminedness) we obtain:

$$\mathbf{E}\left[\left(\Delta y_{ij} - \beta \Delta x_{ij} - \Delta \mathbf{z_{ij}}' \boldsymbol{\delta}\right)^2 | x_{i,j-1}, ..., \mathbf{Z}_i\right] = \mathrm{Var}(\beta_i)\, \mathbf{E}\left[(\Delta x_{ij})^2 | x_{i,j-1}, ..., \mathbf{Z}_i\right] + \mathbf{E}\left[(\Delta v_{ij})^2\right].$$

It follows that the variance of β_i can be consistently estimated by regressing $\left(\Delta y_{ij} - \widehat{\beta} \Delta x_{ij} - \Delta \mathbf{z}_{ij}' \widehat{\boldsymbol{\delta}}\right)^2$ on $(\Delta x_{ij})^2$ and a constant, using as instruments $x_{i,j-1}, ..., \mathbf{Z}_i$.