

Identifying Distributional Characteristics in Random Coefficients Panel Data Models*

Manuel Arellano
CEMFI, Madrid

Stéphane Bonhomme
CEMFI, Madrid

Revised version: June 2011

Abstract

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

JEL CODE: C23.

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

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

1 Introduction

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

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

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

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

¹For examples of earnings models with individual-specific slopes or profiles, see Lillard and Weiss (1979) or more recently Guvenen (2009).

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

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

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

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

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

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

⁴See Demidenko (2004) for a survey on random-effects (or "mixed") models in statistics. Beran *et al.* (1996) and Hoderlein *et al.* (2010) provide nonparametric treatments of random coefficients models for cross-section data.

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

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

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

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

Our identification proofs are constructive. Thus, they suggest consistent estimators for the distributional quantities of interest. We construct consistent method-of-moment estimators of variances. We also construct nonparametric estimators of the densities of individual effects and errors when covariates are discrete, emphasizing the connection with the literature on nonparametric deconvolution (see for example Carroll and Hall, 1988).

In the last section of the paper we apply this methodology to a matched panel dataset of mothers and births constructed in Abrevaya (2006). We find that the mean smoking effect on birth weight is significantly negative (-160 grams). Moreover, the effect shows substantial heterogeneity across mothers, the effect being very negative (-400 g) below the 20th percentile. In addition, we discuss the validity of the strict exogeneity assumption

in the context of this application. Although the mean effect is not point identified when smoking status is predetermined,⁶ we show that several interesting average effects can be identified and estimated when there are no time-varying regressors. The results suggest that the smoking effect is strongly correlated with smoking choices, justifying the fixed-effects perspective. Moreover, we do not find strong evidence against strict exogeneity on these data.

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

A recent paper by Graham and Powell (2008) studies essentially the same model as we do but their focus is rather different. They are concerned with estimating the expectation of random coefficients whereas our concern is the probability distribution of those coefficients. Their focus is on dealing with continuous regressors that exhibit values that change little across periods (near stayers), by trimming those values under otherwise similar assumptions as Chamberlain (1992), including unlimited serial correlation.⁷ Our focus is in exploiting the opportunities for identifying the distributions of the effects offered by limited serial correlation. Our analysis proceeds either under Chamberlain's or under fixed trimming regularity conditions for simplicity, but it could be extended to the regularity conditions discussed by Graham and Powell. Their contribution and ours are basically orthogonal and complement each other.

In an independent but related contribution, Evdokimov (2009) focuses on situations where T is small and uses deconvolution arguments for identifying and estimating the distribution of individual effects, as we do in this paper. There are several important differences with our approach, however. Evdokimov allows for a scalar individual effect that enters an unspecified structural function, with additively separable idiosyncratic errors. In compar-

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

⁷They also deal with models with as many time periods as random coefficients, which we do not consider.

ison, our approach allows for multidimensional individual fixed effects, though it imposes linearity. Another difference is that identification in Evdokimov relies on conditioning on values of the covariates that are constant between periods, while we impose no restriction on the process of exogenous covariates.

Lastly, related identification strategies for densities have been used in the literature on nonparametric identification and estimation of linear factor models with independent factors. See for example Horowitz and Markatou (1996), Székely and Rao (2000), and Bonhomme and Robin (2010). We contribute to that literature by allowing for correlation patterns that may be natural in applications, individual effects being correlated in an unrestricted way, and errors being possibly serially correlated. We also allow for conditioning covariates.

The rest of the paper is as follows. In Section 2 we present the framework of analysis. Section 3 derives the identifying restrictions on the variances of individual effects and errors. In Section 4, we extend the analysis to the full distributions of effects and errors. We discuss estimation in Section 5, and apply our methodology in Section 6 to study the effect of smoking during pregnancy on birth outcomes. Lastly, Section 7 concludes. Additional results may be found in a supplementary appendix to this paper.⁸

2 The random coefficients model

In this section we describe the model together with some extensions, and list various identification and efficiency bounds results for common parameters and averages of individual effects.

2.1 Model

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

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

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

⁸Available at: http://www.cemfi.es/~bonhomme/Random_appendix.pdf

Assumption 1 (*mean independence*)

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

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

Note that we do not specify the conditional distribution of individual effects. In our “fixed-effects” approach, $\boldsymbol{\gamma}_i$ are random draws from a population, along with y_{it} , \mathbf{z}_{it} and \mathbf{x}_{it} , but their conditional distribution given regressors is left unspecified.⁹ Thus, regressors are strictly exogenous with respect to time-varying errors but endogenous with respect to fixed effects. We will discuss the validity of this assumption in the context of our empirical application in Section 6.

Mean independence will be used to identify the vector of common parameters $\boldsymbol{\delta}$ and the means, variances and covariances of individual-specific parameters $\boldsymbol{\gamma}_i$. When studying the identification of distributions of the effects in Section 4, we will need a stronger assumption of conditional statistical independence.

The identification content of the random coefficients model crucially depends on the amount of variation in covariates \mathbf{X}_i . Throughout the paper we focus on overidentified panel models with $T > q$, thus ensuring that the parameters that are common across individuals are identified. When regressors are discrete, the moments of individual effects will be identified on the subpopulation of individuals for which \mathbf{X}_i has full-column rank. In the following we will denote as \mathbb{S} the subpopulation of individuals for which $\det[\mathbf{X}'_i \mathbf{X}_i] \neq 0$. For example, in our empirical application, \mathbb{S} is the subpopulation of mothers who changed smoking status at least once between births.

The fixed-effects approach delivers identification results for the subpopulation \mathbb{S} only. When covariates are discrete, bounds on moments of individual effects may be obtained following the techniques introduced in Chernozhukov *et al.* (2009). In random coefficients applications, however, outcomes often have large support, and the implied bounds may be uninformative for small T . Another alternative would be to abandon the fixed-effects

⁹This terminology differs from an approach where one conditions on the realized values of the $\boldsymbol{\gamma}_i$'s. In this sense, ours may be interpreted as an “unrestricted correlated random-effects” approach.

approach and restrict the conditional distribution of individual effects given covariates. One possibility would be to reduce dimensionality by introducing a factor structure of fixed-effects factors. As we mentioned in introduction, however, it is often difficult to justify those restrictions in economic applications.

Obtaining identification results for the subpopulation \mathbb{S} only is in the nature of the fixed-effects approach. To illustrate, note that in the absence of common parameters and for a binary treatment x_{it} with coefficient γ_i , the expected outcome change over two periods is $E(\Delta y_{it} | \Delta x_{it}) = E(\gamma_i | \Delta x_{it}) \Delta x_{it}$, which identifies $E(\gamma_i | \Delta x_{it})$ when Δx_{it} is 1 or -1 but not when $\Delta x_{it} = 0$. This is the familiar result that a difference-in-differences assumption only identifies the average treatment effect on the treated. Other discrete treatment contexts (such as matching, instrumental-variables or regression discontinuity) typically only identify average treatment effects for specific subpopulations (of common support units, compliers or units at the discontinuity, respectively). Such subpopulations may or may not be of interest depending on the context of application. Relative to the literature on discrete impacts, our contribution is to provide conditions under which the distribution of impacts and the joint distribution of potential outcomes are nonparametrically identified.

When regressors are continuous and \mathbb{S} has measure one, Chamberlain (1992) noted that the overall mean of individual effects $\mathbf{E}(\gamma_i)$ is point-identified. Our results show that, under suitable assumptions, other features of the distribution are identified. We will estimate moments and distributions of effects on a subpopulation \mathbb{S}_h such that $\det[\mathbf{X}'_i \mathbf{X}_i] > h$, where $h > 0$ is independent of the sample size. In a recent paper, Graham and Powell (2008) let $h = h_N$ tends to zero as N tends to infinity at a suitable rate so that the estimator of $\mathbf{E}(\gamma_i)$ is consistent in the limit. Extending their estimation strategy to other distributional features of γ_i is beyond the scope of this paper.

2.2 Common parameters and averages of individual effects

We start by setting some notation. Firstly, let $\widehat{\gamma}_i$ be the least squares estimate (for $\boldsymbol{\delta}$ known):

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

Next, let us introduce the two following matrices:

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

\mathbf{Q}_i ($T \times T$) is the projection matrix on the orthogonal of the span of the columns of \mathbf{X}_i . \mathbf{Q}_i is a familiar object in least squares algebra, and is symmetric idempotent with rank $T - q$. \mathbf{H}_i ($q \times T$) is simply the least squares operator associated with \mathbf{X}_i . Note that \mathbf{Q}_i is always well-defined irrespective of the rank of \mathbf{X}_i ,¹⁰ while \mathbf{H}_i does not exist outside the subpopulation \mathbb{S} (that is, when $\mathbf{X}_i' \mathbf{X}_i$ is singular).

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

$$\mathbf{Q}_i (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}) = \mathbf{Q}_i \mathbf{v}_i \quad (\text{within-group}), \quad (3)$$

$$\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i = \mathbf{H}_i \mathbf{v}_i \quad (\text{between-group}). \quad (4)$$

While equation (4) expresses the difference between the least-squares estimate of $\boldsymbol{\gamma}_i$ (for known $\boldsymbol{\delta}$) and its true value, equation (3) shows the link between the residuals in the individual-specific least-squares regressions and the population errors.

The next proposition shows the identification of common parameters, and the average of individual effects on the subpopulation \mathbb{S} . All proofs are in Appendix A.

Proposition 1 (*common parameters and mean effects*)

Let Assumption 1 hold. Then:

$$\mathbf{E}(\mathbf{Q}_i (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}) | \mathbf{W}_i) = \mathbf{0} \quad (5)$$

and

$$\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i | \mathbf{W}_i, \mathbb{S}) = \mathbf{E}(\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}). \quad (6)$$

So $\mathbf{E}(\boldsymbol{\gamma}_i | \mathbb{S})$ is identified if $\boldsymbol{\delta}$ is identified. Moreover, $\boldsymbol{\delta}$ is identified if $\mathbf{E}(\mathbf{Z}_i' \mathbf{Q}_i \mathbf{Z}_i)$ has rank $\dim(\boldsymbol{\delta})$.

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

Corollary 1 (*projection coefficients*)

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

$$[\mathbf{Var}(\mathbf{F}_i | \mathbb{S})]^{-1} \mathbf{Cov}(\mathbf{F}_i, \boldsymbol{\gamma}_i | \mathbb{S}) = [\mathbf{Var}(\mathbf{F}_i | \mathbb{S})]^{-1} \mathbf{Cov}(\mathbf{F}_i, \widehat{\boldsymbol{\gamma}}_i | \mathbb{S}). \quad (7)$$

¹⁰This is so as long as we consider a generalized formulation of \mathbf{Q}_i as: $\mathbf{Q}_i = \mathbf{I}_T - \mathbf{X}_i \mathbf{X}_i^\dagger$, where \mathbf{X}_i^\dagger is the Moore-Penrose pseudo-inverse of \mathbf{X}_i .

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

$$\mathbf{y}_i = \mathbf{a}(\mathbf{W}_i; \boldsymbol{\theta}) + \mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})\boldsymbol{\gamma}_i + \mathbf{v}_i, \quad (8)$$

where $\boldsymbol{\theta}$ is a vector of common parameters that enter nonlinearly functions \mathbf{a} (which is $T \times 1$) and \mathbf{B} ($T \times q$). We assume that $\mathbf{a}(\mathbf{W}; \boldsymbol{\theta})$ and $\mathbf{B}(\mathbf{W}; \boldsymbol{\theta})$ are continuously differentiable with respect to $\boldsymbol{\theta}$, for each \mathbf{W} .

A simple special case of model (8) is the one-factor model:

$$y_{it} = \mathbf{z}'_{it}\boldsymbol{\delta} + \mu_t\alpha_i + v_{it}, \quad (9)$$

where μ_1, \dots, μ_T are time-varying parameters and α_i is scalar (e.g., Holtz-Eakin *et al.*, 1988). Other interesting special cases of (8) are models where the regressors include lags (or leads) of the dependent variable. For example, a first-order autoregressive model:

$$y_{it} = \delta y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\gamma}_i + v_{it}, \quad |\delta| < 1. \quad (10)$$

That (10) is a special case of (8) is seen by writing the reduced-form:

$$y_{it} = (\mathbf{x}_{it} + \delta\mathbf{x}_{i,t-1} + \dots + \delta^{t-1}\mathbf{x}_{i1})' \boldsymbol{\gamma}_i + \delta^t y_{i0} + v_{it} + \delta v_{i,t-1} + \dots + \delta^{t-1}v_{i1},$$

which is of the form (8) with the $(q+1) \times 1$ vector of individual effects: $\tilde{\boldsymbol{\gamma}}_i = (\boldsymbol{\gamma}'_i, y_{i0})'$.

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

$$\mathbf{Q}_i(\boldsymbol{\theta})(\mathbf{y}_i - \mathbf{a}(\mathbf{W}_i; \boldsymbol{\theta})) = \mathbf{Q}_i(\boldsymbol{\theta})\mathbf{v}_i \quad (\text{within-group}), \quad (11)$$

$$\hat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i = \mathbf{H}_i(\boldsymbol{\theta})\mathbf{v}_i \quad (\text{between-group}), \quad (12)$$

where

$$\mathbf{Q}_i(\boldsymbol{\theta}) = \mathbf{I}_T - \mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})(\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})'\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta}))^{-1}\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})', \quad (13)$$

$$\mathbf{H}_i(\boldsymbol{\theta}) = (\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})'\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta}))^{-1}\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})'. \quad (14)$$

Let $\mathbb{S}_{\boldsymbol{\theta}}$ be the subpopulation of individuals for which $\det[\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})'\mathbf{B}(\mathbf{W}_i; \boldsymbol{\theta})] \neq 0$. We have the next corollary of Proposition 1.

Corollary 2 (Chamberlain's model)

Consider model (8), and suppose that $\mathbf{E}(\mathbf{v}_i | \mathbf{W}_i) = \mathbf{0}$. Then:

$$\mathbf{E}[\mathbf{Q}_i(\boldsymbol{\theta})(y_i - \mathbf{a}(\mathbf{W}_i; \boldsymbol{\theta})) | \mathbf{W}_i] = \mathbf{0}, \quad (15)$$

and

$$\mathbf{E}[\mathbf{H}_i(\boldsymbol{\theta})(y_i - \mathbf{a}(\mathbf{W}_i; \boldsymbol{\theta})) | \mathbf{W}_i, \mathbb{S}_\theta] = \mathbf{E}(\gamma_i | \mathbf{W}_i, \mathbb{S}_\theta), \quad (16)$$

where $\mathbf{Q}_i(\boldsymbol{\theta})$ and $\mathbf{H}_i(\boldsymbol{\theta})$ are given by (13) and (14), respectively. So, $\mathbf{E}(\gamma_i | \mathbb{S}_\theta)$ is identified if $\boldsymbol{\theta}$ is identified.

Extension 2: general predetermined variables. Assumption 1 posits the strict exogeneity of \mathbf{Z}_i and \mathbf{X}_i given γ_i . However, our approach can be generalized to situations where \mathbf{Z}_i includes predetermined or endogenous variables (although the remainder of the paper assumes strict exogeneity for simplicity). The idea is to replace Assumption 1 with the following generalization:

$$\mathbf{E}(v_{it} | \mathbf{r}_{i1}, \dots, \mathbf{r}_{it}, \mathbf{X}_i, \gamma_i) = 0 \quad (t = 1, \dots, T), \quad (17)$$

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

Strict exogeneity of \mathbf{X}_i is an essential ingredient, but as long as this is preserved, nonlinear extensions are also possible. For example, it is possible to consider assumption (17) in conjunction with a model of the form

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

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

When some of the regressors are predetermined, the orthogonality between original errors and conditioning variables in the new assumption is not transmitted to ordinary within errors. The reason is that (17) implies a pattern of sequential orthogonality and each within error depends on the full time series of original errors. However, there is an alternative within transformation that preserves sequential orthogonality, which is provided by a generalization of forward orthogonal deviations (Arellano and Bover, 1995). Let \mathbf{A}_i be a $(T - q) \times T$ upper

triangular decomposition of \mathbf{Q}_i such that $\mathbf{A}'_i \mathbf{A}_i = \mathbf{Q}_i$ and $\mathbf{A}_i \mathbf{A}'_i = \mathbf{I}_{T-q}$. The orthogonal within errors $\mathbf{A}_i \mathbf{v}_i \equiv (v_{i1}^*, \dots, v_{i(T-q)}^*)'$ satisfy assumption (17):

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

Information bound on common parameters and average effects. Chamberlain (1992) obtained the optimal moment conditions of common parameters and average effects for model (8). The moments are optimal in the sense that an estimator based on them attains the semiparametric information bound.

Let us suppose to simplify the notation that \mathbb{S}_θ is the full population of individuals. Following Chamberlain (1992),¹¹ the joint optimal moments for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i)$ can be expressed as

$$\mathbf{E} \left(\begin{array}{c} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}'} [\mathbf{a}_i + \mathbf{B}_i \mathbf{E}(\boldsymbol{\gamma}_i \mid \mathbf{W}_i)] \right\}' \mathbf{A}'_i (\mathbf{A}_i \mathbf{V}_i \mathbf{A}'_i)^{-1} \mathbf{A}_i (\mathbf{y}_i - \mathbf{a}_i) \\ (\mathbf{B}'_i \mathbf{V}_i^{-1} \mathbf{B}_i)^{-1} \mathbf{B}'_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{a}_i - \mathbf{B}_i \boldsymbol{\gamma}) \end{array} \right) = \mathbf{0}, \quad (18)$$

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

For the variance bound to be finite, one needs that $\mathbf{E} \left[(\det(\mathbf{B}'_i \mathbf{V}_i^{-1} \mathbf{B}_i))^{-1} \right] < \infty$. When this condition does not hold, root- N consistent estimation of average effects is not possible. Graham and Powell (2008) refer to this situation as “irregular identification”. In models with slope heterogeneity, this will occur when covariates are very persistent.

3 Identification of second moments

This section discusses the identification of covariance structures.

3.1 Variances of individual effects and errors

To recover the variance of individual effects, we impose restrictions on the conditional variance-covariance matrix of errors \mathbf{v}_i , which we denote as $\boldsymbol{\Omega}_i = \mathbf{Var}(\mathbf{v}_i \mid \mathbf{W}_i)$. To see why restricting $\boldsymbol{\Omega}_i$ is necessary for identification, note that taking second moments in (1) and using Assumption 1 implies the following restrictions in levels:¹²

$$\mathbf{E}[(\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta})(\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta})' \mid \mathbf{W}_i] = \mathbf{X}_i \mathbf{E}(\boldsymbol{\gamma}_i \boldsymbol{\gamma}'_i \mid \mathbf{W}_i) \mathbf{X}'_i + \boldsymbol{\Omega}_i. \quad (19)$$

¹¹The argument is developed in the supplementary appendix to this paper.

¹²Here and throughout the paper, second-order conditional moments of \mathbf{y}_i , $\boldsymbol{\gamma}_i$, and \mathbf{v}_i given \mathbf{W}_i are assumed finite.

Clearly, when $\boldsymbol{\Omega}_i$ is unrestricted, second-order moments of the data are not informative about the second moments of individual effects. This is because, in this case, $\mathbf{E}(\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' | \mathbf{W}_i)$ is absorbed into the unrestricted $\boldsymbol{\Omega}_i$. Identification of the variance of individual effects thus requires restricting the dynamics of errors.

MA restrictions. We start by imposing uncorrelatedness restrictions on errors v_{it} . A particular example is a moving average (MA) process of order r , in which case the conditional covariance between v_{it} and $v_{i,t'}$ given \mathbf{W}_i is zero if $|t' - t| > r$.

Formally, we make the following assumption.

Assumption 2 (MA with uncorrelated shocks)

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

$$\text{vec}(\boldsymbol{\Omega}_i) = \mathbf{S}_2 \boldsymbol{\omega}_i. \quad (20)$$

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

$$v_{it} = u_{it} + \theta_{1t} u_{i,t-1} + \dots + \theta_{rt} u_{i,t-r}, \quad t = 1, \dots, T, \quad (21)$$

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

Note that since $\mathbf{Var}(\mathbf{v}_i | \mathbf{W}_i) = \mathbf{E}[\mathbf{Var}(\mathbf{v}_i | \mathbf{W}_i, \boldsymbol{\gamma}_i) | \mathbf{W}_i]$, Assumption 2 is consistent with an underlying moving average model with unobserved heterogeneity of the form

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

for an unspecified function $\boldsymbol{\phi}$ such that $\boldsymbol{\omega}_i = \mathbf{E}[\boldsymbol{\phi}(\mathbf{W}_i, \boldsymbol{\gamma}_i) | \mathbf{W}_i]$, possibly including a larger vector of fixed effects than those present in the conditional mean. In particular, $\theta_{11}, \dots, \theta_{rT}$ in (21) may depend on regressors \mathbf{W}_i , and could also depend on additional individual effects $\boldsymbol{\xi}_i$ as long as $\mathbf{E}(u_{it} | \mathbf{W}_i, \boldsymbol{\gamma}_i, \boldsymbol{\xi}_i) = 0$.

To study when Assumption 2 identifies second-order moments, let us now define the projection matrix on the orthogonal of the span of the columns of $\mathbf{X}_i \otimes \mathbf{X}_i$:

$$\begin{aligned} \mathbf{M}_i &= \mathbf{I}_{T^2} - \left[\mathbf{X}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \right] \otimes \left[\mathbf{X}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \right] \\ &= \mathbf{I}_{T^2} - [\mathbf{I}_T - \mathbf{Q}_i] \otimes [\mathbf{I}_T - \mathbf{Q}_i]. \end{aligned} \quad (22)$$

Note that, as the matrix \mathbf{Q}_i , \mathbf{M}_i is well-defined irrespective of the rank of \mathbf{X}_i .

Left-multiplying (19) by \mathbf{M}_i (in vector form) and using Assumption 2 we obtain:

$$\begin{aligned}\mathbf{M}_i \mathbf{E}[(\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}) \otimes (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}) | \mathbf{W}_i] &= \mathbf{M}_i \mathbf{vec}(\boldsymbol{\Omega}_i) \\ &= \mathbf{M}_i \mathbf{S}_2 \boldsymbol{\omega}_i.\end{aligned}\tag{23}$$

The following identification result is an immediate consequence of (23).¹³

Theorem 1 (*second-order moments*)

Let Assumptions 1 and 2 hold, and suppose that $\boldsymbol{\delta}$ is identified. Suppose also that

$$\text{rank}[\mathbf{M}_i \mathbf{S}_2] = m.\tag{24}$$

Then, $\boldsymbol{\Omega}_i$ and $\mathbf{E}(\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' | \mathbf{W}_i, \mathbb{S})$ are identified.

Remark 1. When (24) holds, one may solve analytically for $\boldsymbol{\Omega}_i$ in (23). For example, in the particular case where errors are i.i.d. homoskedastic with variance σ^2 we have:

$$\sigma^2 = \frac{1}{T - q} \mathbf{E}[(\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta})' \mathbf{Q}_i (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta})].\tag{25}$$

Remark 2. Note that taking variances in (4) implies:

$$\mathbf{Var}(\widehat{\boldsymbol{\gamma}}_i | \mathbb{S}) = \mathbf{Var}(\boldsymbol{\gamma}_i | \mathbb{S}) + \mathbf{E}(\mathbf{H}_i \boldsymbol{\Omega}_i \mathbf{H}_i' | \mathbb{S}).\tag{26}$$

The total variance of the fixed-effects estimates $\widehat{\boldsymbol{\gamma}}_i$ is the sum of two components: the true cross-sectional variation of individual effects, and the contribution of estimation noise for small T . Theorem 1 shows that suitable restrictions on the covariance structure of errors allow to separate these two components, and to recover the variance of $\boldsymbol{\gamma}_i$.

Remark 3. It is interesting to study the order condition associated with the rank condition (24). One can check that

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

¹³To see why the second moment of $\boldsymbol{\gamma}_i$ is conditional on \mathbb{S} in Theorem 1, note that, given $\boldsymbol{\Omega}_i$, $\mathbf{X}_i \mathbf{E}(\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' | \mathbf{W}_i) \mathbf{X}_i'$ is identified by (19). As \mathbf{X}_i has full-column rank in \mathbb{S} , identification of $\mathbf{E}(\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' | \mathbf{W}_i, \mathbb{S})$ follows.

with equality when \mathbf{S}_2 selects all $T(T+1)/2$ non-redundant elements of $\text{vec}(\boldsymbol{\Omega}_i)$.¹⁴ So, the order condition associated with (24) is:

$$\frac{T(T+1)}{2} - \frac{q(q+1)}{2} \geq m. \quad (27)$$

In particular, when errors are MA(r) we need that

$$\frac{T(T+1)}{2} - \frac{q(q+1)}{2} \geq (r+1) \left(T - \frac{r}{2}\right). \quad (28)$$

The left-hand-side in (28) is decreasing in q , while the right-hand side is increasing in r . So, equation (28) highlights a trade-off between heterogeneity and error persistence: the higher the number of individual-specific effects, the smaller the order of the moving-average process compatible with identification for given T .

Lastly, note that, instead of working with the level equations (19), one could work with the subset of *within* equations obtained by taking covariances in (3). The within covariance restrictions do not depend on errors v_{it} being mean independent of individual effects γ_i . In the supplementary appendix, we show that working with within equations alone requires stronger conditions for identification.

AR restrictions. Autoregressive errors are very popular in applied work, and are *not* covered by assumption (20) because autoregressive processes are correlated at all lags. Nevertheless, a similar approach can be adopted to study identification.¹⁵ To see how, consider the following model:

$$v_{it} = \rho_{1t}v_{i,t-1} + \dots + \rho_{pt}v_{i,t-p} + u_{it}, \quad t = p+1, \dots, T, \quad (29)$$

where $\rho_{1,p+1}, \dots, \rho_{pT}$ are unrestricted parameters and $u_{i,p+1}, \dots, u_{iT}$ satisfy Assumption 2. In the case where u_{it} is MA(r), v_{it} given by (29) follows an ARMA(p,r) process.

Letting $\mathbf{u}_i = (v_{i1}, \dots, v_{ip}, u_{i,p+1}, \dots, u_{iT})'$, (29) may be written as $\mathbf{R}(\boldsymbol{\rho}) \mathbf{v}_i = \mathbf{u}_i$, where $\mathbf{R}(\boldsymbol{\rho})$ is a $T \times T$ matrix that depends on $\boldsymbol{\rho} = (\rho_{1,p+1}, \dots, \rho_{pT})'$. Assuming $\mathbf{R}(\boldsymbol{\rho})$ non-singular, we have:

$$\text{vec}(\boldsymbol{\Omega}_i) = [\mathbf{R}(\boldsymbol{\rho}) \otimes \mathbf{R}(\boldsymbol{\rho})]^{-1} \text{vec}(\text{Var}(\mathbf{u}_i | \mathbf{W}_i)).$$

¹⁴A proof of this statement may be found in the supplementary appendix.

¹⁵However, contrary to the moving average case, an autoregressive model with unobserved heterogeneity does not generally imply an autoregressive structure for $\text{Var}(\mathbf{v}_i | \mathbf{W}_i)$.

Let us denote as $\boldsymbol{\xi}$ the free parameters of $\mathbf{Var}(\mathbf{u}_i|\mathbf{W}_i)$, and denote $\boldsymbol{\theta} = (\boldsymbol{\xi}', \boldsymbol{\rho}')'$. Assuming that there is no discontinuity of rank at the truth, a necessary and sufficient condition for local identification is:

$$\text{rank} \left(\mathbf{M}_i \frac{\partial [\mathbf{R}(\boldsymbol{\rho}) \otimes \mathbf{R}(\boldsymbol{\rho})]^{-1} \text{vec}(\mathbf{Var}(\mathbf{u}_i|\mathbf{W}_i))}{\partial \boldsymbol{\theta}'} \right) = \dim(\boldsymbol{\theta}). \quad (30)$$

In particular, leaving covariances involving initial conditions unrestricted, a necessary condition for (30) is:

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

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 .

3.2 Examples

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

$$y_{it} = \alpha_i + \beta_i t + v_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (31)$$

where v_{it} are serially correlated. Model (31), or a restricted version of it (e.g., with $\beta_i = 0$), is often used to model the dynamics of earnings.

Suppose that errors are AR(1):

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

so:

$$y_{it} - \rho y_{i,t-1} = (1 - \rho)\alpha_i + \beta_i(t - \rho(t-1)) + u_{it}.$$

It can be shown that, when ρ is assumed known, $T = 4$, and u_{i2} , u_{i3} and u_{i4} are assumed uncorrelated, their variances are identified from covariance restrictions, together with the covariance matrix of individual effects. This is consistent with the rank condition (24) being satisfied for the transformed model. In contrast, from (30) neither ρ nor the covariance parameters are identified when $T = 4$, though they are (generally) identified when $T = 5$. See the supplementary appendix for details.

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

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + v_{i\ell}, \quad i = 1, \dots, N, \quad \ell = 1, \dots, L. \quad (32)$$

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

Let $L = 3$, and consider a “treatment” sequence $(s_{i1}, s_{i2}, s_{i3}) = (1, 0, 0)$. It can be shown that, when errors are uncorrelated with unrestricted variances, $\text{Var}(\beta_i)$ and $\text{Var}(v_{i1})$ are not separately identified from covariance restrictions in levels. Although the order condition for identification is satisfied,¹⁶ the rank condition is not. If we impose the stationarity restriction that all three variances of v_{i1} , v_{i2} and v_{i3} are equal, then they are identified along with the covariance matrix of individual effects. In addition, $\text{Var}(v_{i3} - v_{i2})$ is identified from within restrictions alone.

3.3 Efficiency bounds

Here we show how Chamberlain (1992)’s analysis can be extended to obtain a joint information bound for common parameters, means and variances of random coefficients, and a parameterization of the variances of errors. Let us write down model (1) as:

$$\mathbf{E}(y_i | \mathbf{W}_i, \gamma_i) = \mathbf{Z}_i \boldsymbol{\delta} + \mathbf{X}_i \gamma_i \quad (33)$$

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

$$\mathbf{E}(\mathbf{v}_i \otimes \mathbf{v}_i | \mathbf{W}_i, \gamma_i) = \boldsymbol{\psi}_i(\boldsymbol{\phi}), \quad (34)$$

where $\boldsymbol{\psi}_i$ is a $T^2 \times 1$ vector of functions of a parameter $\boldsymbol{\phi}$, which may also depend on \mathbf{W}_i . However, we assume that the variance of \mathbf{v}_i does not depend on γ_i .¹⁷

Using (34) together with Assumption 1 we obtain the following expression for the conditional second-order moments of \mathbf{y}_i :

$$\begin{aligned} \mathbf{E}(\mathbf{y}_i \otimes \mathbf{y}_i | \mathbf{W}_i, \gamma_i) &= (\mathbf{Z}_i \boldsymbol{\delta} \otimes \mathbf{Z}_i \boldsymbol{\delta}) + \boldsymbol{\psi}_i(\boldsymbol{\phi}) + (\mathbf{X}_i \otimes \mathbf{Z}_i \boldsymbol{\delta} + \mathbf{Z}_i \boldsymbol{\delta} \otimes \mathbf{X}_i) \gamma_i \\ &\quad + (\mathbf{X}_i \otimes \mathbf{X}_i) (\gamma_i \otimes \gamma_i). \end{aligned} \quad (35)$$

¹⁶As: $3(3+1)/2 - 2(2+1)/2 = 3$, see equation (28).

¹⁷In cases where $\mathbf{E}(\mathbf{v}_i \otimes \mathbf{v}_i | \mathbf{W}_i, \gamma_i, \boldsymbol{\xi}_i) = \boldsymbol{\Gamma}(\mathbf{W}_i, \boldsymbol{\phi}) \begin{pmatrix} \gamma_i \\ \boldsymbol{\xi}_i \end{pmatrix}$, we could extend the model and apply a similar approach treating $\boldsymbol{\xi}_i$ as additional random coefficients.

Stacking (33) and (35) together yields

$$\mathbf{E}(\mathbf{y}_i^* | \mathbf{W}_i, \boldsymbol{\gamma}_i^*) = \mathbf{d}(\mathbf{W}_i, \boldsymbol{\theta}) + \mathbf{R}(\mathbf{W}_i, \boldsymbol{\theta}) \boldsymbol{\gamma}_i^*, \quad (36)$$

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

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

and

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

Equation (36), which combines mean and covariance restrictions in levels, is a special case of model (8).¹⁸ Therefore, the optimal moments (and associated semiparametric bound) for $\boldsymbol{\delta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\gamma}^* = \mathbf{E}(\boldsymbol{\gamma}_i^*)$ are of the form given in expression (18).

Lastly, here also finiteness of the variance bound relies on a moment existence condition, namely: $\mathbf{E} \left[(\det(\mathbf{R}_i' [\mathbf{V}_i^*]^{-1} \mathbf{R}_i))^{-1} \right] < \infty$, where $\mathbf{R}_i = \mathbf{R}(\mathbf{W}_i, \boldsymbol{\theta})$ and $\mathbf{V}_i^* = \mathbf{Var}(\mathbf{y}_i^* | \mathbf{W}_i)$. If this moment is not finite, identification of the variance of individual effects is irregular and root- N consistent estimation is not possible.

4 Identification of distributions

In this section, we turn to the identification of the distribution functions of effects and errors.

4.1 Identification result

We will work under the following conditional independence assumption.

Assumption 3 (*conditional statistical independence*)

$$\boldsymbol{\gamma}_i \text{ and } \mathbf{v}_i \text{ are statistically independent given } \mathbf{W}_i. \quad (37)$$

Conditional independence restrictions are commonly made in the literature on nonparametric identification and estimation (e.g., Hu and Schennach, 2008, and references therein).

Moreover, restriction (37) is in the nature of a fixed-effects approach, where $\boldsymbol{\gamma}_i$ represent

¹⁸The only difference is that $\mathbf{E}(\boldsymbol{\gamma}_i^* | \mathbf{W}_i)$ is not fully unrestricted, as its components are first and second moments of the same underlying $\boldsymbol{\gamma}_i$. However, these extra restrictions imply moment *inequalities* that do not affect the bound.

individual-specific parameters such as preferences or technology. However, note that Assumption 3 is more restrictive than mean independence (Assumption 1) as, for example, it rules out the presence of individual effects in the conditional variance of \mathbf{v}_i .

Relaxing conditional independence would require the use of very different methods for identification and estimation. To see why, note that linearity and independence imply that the distribution function of the data may be written as:

$$f_{\mathbf{y}_i|\mathbf{w}_i}(\mathbf{y}|\mathbf{w}) = \int f_{\mathbf{v}_i|\mathbf{w}_i}(\mathbf{y} - \mathbf{z}'\boldsymbol{\delta} - \mathbf{x}'\boldsymbol{\gamma}|\mathbf{w}) f_{\boldsymbol{\gamma}_i|\mathbf{w}_i}(\boldsymbol{\gamma}|\mathbf{w}) d\boldsymbol{\gamma}, \quad (38)$$

where the integral is taken over the support of individual effects. Thus, $f_{\mathbf{y}_i|\mathbf{w}_i}$ is a convolution of $f_{\mathbf{v}_i|\mathbf{w}_i}$ and $f_{\boldsymbol{\gamma}_i|\mathbf{w}_i}$. The analytical identification results and estimators below are based on this property. If one were to relax Assumption 3, the convolution property would be lost, and the problem of recovering $f_{\mathbf{v}_i|\mathbf{w}_i}$ and $f_{\boldsymbol{\gamma}_i|\mathbf{w}_i}$ in (38) would become a challenging inverse problem.

To derive the identification results, it is very convenient to work with *characteristic functions*. Let (\mathbf{Y}, \mathbf{X}) be a pair of random vectors, $\mathbf{Y} \in \mathbf{R}^L$, and let j be a square root of -1 .¹⁹ The conditional characteristic function of \mathbf{Y} given $\mathbf{X} = \mathbf{x}$, is defined as:

$$\Psi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\tau}|\mathbf{x}) = \mathbf{E}(\exp(j\boldsymbol{\tau}'\mathbf{Y})|\mathbf{X} = \mathbf{x}), \quad \boldsymbol{\tau} \in \mathbf{R}^L.$$

Some useful properties of characteristic functions are listed in the supplementary appendix.

We will impose the following technical condition on the characteristic functions of individual effects and errors.

Assumption 4 (*nonvanishing characteristic functions*)

The characteristic functions $\Psi_{\boldsymbol{\gamma}_i|\mathbf{w}_i}$ and $\Psi_{\mathbf{v}_i|\mathbf{w}_i}$ are almost everywhere nonvanishing on \mathbf{R}^q and \mathbf{R}^T , respectively.

To provide some intuition for this assumption note that, since $\mathbf{X}_i\boldsymbol{\gamma}_i$ and \mathbf{v}_i are conditionally independent, we have for all $\boldsymbol{\tau} \in \mathbf{R}^T$:

$$\Psi_{\mathbf{y}_i - \mathbf{z}_i\boldsymbol{\delta}|\mathbf{w}_i}(\boldsymbol{\tau}|\mathbf{W}_i) = \Psi_{\boldsymbol{\gamma}_i|\mathbf{w}_i}(\mathbf{X}_i'\boldsymbol{\tau}|\mathbf{W}_i)\Psi_{\mathbf{v}_i|\mathbf{w}_i}(\boldsymbol{\tau}|\mathbf{W}_i). \quad (39)$$

Hence, when evaluated at a point where $\Psi_{\mathbf{v}_i|\mathbf{w}_i}$ vanishes, equation (39) is not directly informative about $\Psi_{\boldsymbol{\gamma}_i|\mathbf{w}_i}$.

¹⁹We work with the notation $j^2 = -1$ instead of $i^2 = -1$ to avoid confusion with the index of individual units.

The assumption that the characteristic function of errors has no real zeros is a common regularity condition in the literature on nonparametric deconvolution. Most well-known families of distributions satisfy this property, the normal being an important special case. However, several distributions (uniform, symmetrically truncated normal) do not satisfy Assumption 4.²⁰

Under Assumption 4 we can take logarithms in (39) and obtain:

$$\ln \Psi_{\mathbf{y}_i - \mathbf{z}_i \boldsymbol{\delta} | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i) = \ln \Psi_{\gamma_i | \mathbf{W}_i}(\mathbf{X}_i' \boldsymbol{\tau} | \mathbf{W}_i) + \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i). \quad (40)$$

Given that conditional second-order moments exist, the log-characteristic functions in this expression are second-order differentiable (e.g., Székely and Rao, 2000). Taking second derivatives we obtain:²¹

$$\begin{aligned} \frac{\partial^2 \ln \Psi_{\mathbf{y}_i - \mathbf{z}_i \boldsymbol{\delta} | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} &= \mathbf{X}_i \left(\frac{\partial^2 \ln \Psi_{\gamma_i | \mathbf{W}_i}(\mathbf{X}_i' \boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \right) \mathbf{X}_i' \\ &\quad + \frac{\partial^2 \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'}, \quad \boldsymbol{\tau} \in \mathbf{R}^T. \end{aligned} \quad (41)$$

Equation (41) nicely extends covariance restrictions to restrictions on the entire distribution of the error terms. Indeed, evaluating (41) at $\boldsymbol{\tau} = \mathbf{0}$ yields:

$$\mathbf{Var}(\mathbf{y}_i - \mathbf{z}_i \boldsymbol{\delta} | \mathbf{W}_i) = \mathbf{X}_i \mathbf{Var}(\gamma_i | \mathbf{W}_i) \mathbf{X}_i' + \boldsymbol{\Omega}_i.$$

Now, as in the case of variances, it is not possible to solve for $\Psi_{\gamma_i | \mathbf{W}_i}$ and $\Psi_{\mathbf{v}_i | \mathbf{W}_i}$ in (41) when the dependence structure of errors is left unrestricted. We study identification under the following assumption.

Assumption 5 (*MA with independent shocks*)

There exists an m -dimensional vector of functions $\boldsymbol{\omega}_i(\boldsymbol{\tau})$ ($\boldsymbol{\tau} \in \mathbf{R}^T$), possibly dependent on \mathbf{W}_i , and a known (selection) matrix \mathbf{S}_2 such that:

$$\mathbf{vec} \left(\frac{\partial^2 \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} \right) = \mathbf{S}_2 \boldsymbol{\omega}_i(\boldsymbol{\tau}), \quad \boldsymbol{\tau} \in \mathbf{R}^T. \quad (42)$$

The spirit of Assumption 5 is similar to the moving-average restrictions of Assumption 3. Indeed, if one imposes (42) at $\boldsymbol{\tau} = \mathbf{0}$ only, then the two assumptions coincide. In particular,

²⁰Recently, Carrasco and Florens (2009) and Evdokimov and White (2010) have studied ways to relax the assumption of nonvanishing characteristic functions in deconvolution contexts.

²¹While $\boldsymbol{\tau} \in \mathbf{R}^T$ denotes a generic argument of $\Psi_{\mathbf{y}_i - \mathbf{z}_i \boldsymbol{\delta} | \mathbf{W}_i}$ (or $\Psi_{\mathbf{v}_i | \mathbf{W}_i}$), $\boldsymbol{\xi} \in \mathbf{R}^g$ denotes a generic argument of $\Psi_{\gamma_i | \mathbf{W}_i}$. Similarly $\boldsymbol{\zeta} \in \mathbf{R}^{T+r}$ in (44) will denote a generic argument of $\Psi_{\mathbf{u}_i | \mathbf{W}_i}$.

the selection matrix in the two assumptions is the same. In addition, while Assumption 3 is consistent with a moving-average process with uncorrelated latent disturbances, Assumption 5 holds when v_{it} follows a moving average process of order r of the form (21), where $u_{i,1-r}, \dots, u_{iT}$ are mutually *independent* given regressors.

To see this last part, note that if $\mathbf{u}_i = (u_{i,1-r}, \dots, u_{iT})'$, then (21) may be written as:

$$\mathbf{v}_i = \mathbf{\Lambda} \mathbf{u}_i, \quad (43)$$

where each element of the $T \times (T + r)$ matrix $\mathbf{\Lambda}$ is either 0 or one of the θ parameters that appear in (21). Taking second derivatives of log-characteristic functions in (43) yields:

$$\frac{\partial^2 \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} = \mathbf{\Lambda} \left(\frac{\partial^2 \ln \Psi_{\mathbf{u}_i | \mathbf{W}_i}(\boldsymbol{\Lambda}' \boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}'} \right) \boldsymbol{\Lambda}', \quad \boldsymbol{\tau} \in \mathbf{R}^T. \quad (44)$$

Because of independence, the central matrix on the right-hand side of (44) is diagonal. So, it follows from the MA structure that $\frac{\partial^2 \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \tau_t \partial \tau_{t'}} = 0$ if $|t' - t| > r$. Hence, Assumption 5 is satisfied for the *same* selection matrix as in Assumption 3.

The following theorem shows that, when Assumption 5 holds and the rank condition (24) is satisfied, the characteristic functions of individual effects and errors are point-identified. As distributions are uniquely determined by their characteristic functions,²² the identification of the conditional distribution functions follows.

Theorem 2 (*distribution functions*)

Let Assumptions 1, 3, 4 and 5 hold, and suppose that $\boldsymbol{\delta}$ is identified. In addition, suppose that the rank condition (24) is satisfied. Then $\Psi_{\mathbf{v}_i | \mathbf{W}_i}$ and $\Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}$ are identified. As a consequence, the distribution functions $f_{\mathbf{v}_i | \mathbf{W}_i}$ and $f_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}$ are identified.

The identification of the second derivatives of log-characteristic functions follows very closely that of variances in Section 3. This is due to the fact that the rank condition for identification, equation (24), is the one that was needed for the identification of variances under MA restrictions.²³ Lastly, note that extensions to autoregressive and ARMA processes with independent underlying innovations can be done along the lines of Section 3.

4.2 Examples

We end this section by considering two examples.

²²This uniqueness result holds for discrete or continuous distributions (e.g., Dudley, 2002, p.303).

²³A technical point is that $\ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}$ is uniquely determined by its second derivative and the fact that $\partial \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\mathbf{0} | \mathbf{W}_i) / \partial \boldsymbol{\tau} = \mathbf{E}(\mathbf{v}_i | \mathbf{W}_i) = \mathbf{0}$.

Kotlarski's model. We start by considering the following simple model with repeated measurements:

$$\begin{cases} y_{i1} &= \alpha_i + v_{i1} \\ y_{i2} &= \alpha_i + v_{i2}, \end{cases}$$

where α_i , v_{i1} and v_{i2} are mutually independent.

In this case, $\mathbf{X}_i = (1, 1)'$, so $\mathbf{M}_i = I_4 - \frac{1}{4}J_4$ (independent of i), where J_4 is the 4×4 matrix of ones. The selection matrix of Assumption 5 is $\mathbf{S}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}'$.

It is easy to see that (41) implies the following restrictions on log-characteristic functions of errors:

$$\begin{aligned} \frac{\partial^2}{\partial t_1^2} \ln \Psi_{v_{i1}}(t_1) &= \frac{\partial^2}{\partial t_1^2} \ln \Psi_{y_i}(t_1, t_2) - \frac{\partial^2}{\partial t_1 \partial t_2} \ln \Psi_{y_i}(t_1, t_2) \\ \frac{\partial^2}{\partial t_2^2} \ln \Psi_{v_{i2}}(t_2) &= \frac{\partial^2}{\partial t_2^2} \ln \Psi_{y_i}(t_1, t_2) - \frac{\partial^2}{\partial t_1 \partial t_2} \ln \Psi_{y_i}(t_1, t_2). \end{aligned}$$

Those are the restrictions used in Kotlarski (1967)'s proof of the nonparametric identification of the marginal distribution functions of α_i , v_{i1} and v_{i2} . Li and Vuong (1998) have used those restrictions in estimation. Horowitz and Markatou (1996) construct a different estimator, using a subset of those restrictions based on the within transformation $y_{i2} - y_{i1} = v_{i2} - v_{i1}$.

This discussion shows that the identification results in this section can be seen as extensions of Kotlarski (1967)'s result. The extension is done in four directions: the multivariate conditional distribution of some components (including the individual effects) is left unrestricted, errors are allowed to be correlated in an ARMA fashion, conditioning regressors are present, and there are unknown common parameters.

Example 2. A similar insight applies to Example 2, where we assume that errors are independent of each other, and where we take $T = 3$ and $\mathbf{s}_i = (1, 0, 0)'$. Here the relevant selection matrix is:

$$\mathbf{S}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}'.$$

Moreover, one obtains, using (41):

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

and:

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

We noted in Section 3 that the rank condition (24) is not satisfied when the distributions of v_{i1} , v_{i2} and v_{i3} are different. It is thus not surprising that $\Psi_{v_{i1}}$ does not appear in this system. Assuming that the distributions of v_{i1} , v_{i2} , and v_{i3} are the same, however, identification of their characteristic function easily follows. Indeed, this model can be seen as an augmented version of Kotlarski's with an extra equation:

$$\begin{cases} y_{i1} &= \alpha_i + \beta_i + v_{i1} \\ y_{i2} &= \alpha_i + v_{i2} \\ y_{i3} &= \alpha_i + v_{i3}. \end{cases}$$

5 Estimation

In this section we discuss estimation of parameters, moments, and densities using an i.i.d. sample $\{\mathbf{y}_i, \mathbf{Z}_i, \mathbf{X}_i\}$, $i = 1, \dots, N$. We will estimate the moments and distributions of individual effects on a subpopulation of individuals. With discrete covariates \mathbf{X}_i , the subpopulation \mathbb{S} contains individuals for which $\det[\mathbf{X}_i' \mathbf{X}_i] \neq 0$. When covariates are continuous, we let $h > 0$ and define \mathbb{S}_h as the subpopulation for which $\det[\mathbf{X}_i' \mathbf{X}_i] > h$. The estimators we propose will be consistent as N tends to infinity, for fixed h . In the rest of this section (and with some abuse of notation) we will simply denote \mathbb{S}_h as \mathbb{S} .²⁴

5.1 Common parameters and average effects

Using (18), the optimal moments for $\boldsymbol{\delta}$ and $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i | \mathbb{S})$ corresponding to model (1) can be written as:

$$\begin{aligned} \mathbf{E} \left[\mathbf{Z}_i' \mathbf{A}_i' (\mathbf{A}_i \mathbf{V}_i \mathbf{A}_i')^{-1} \mathbf{A}_i (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}) \right] &= \mathbf{0} \\ \mathbf{E} \left[(\mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta} - \mathbf{X}_i \boldsymbol{\gamma}) \mid \mathbb{S} \right] &= \mathbf{0}, \end{aligned}$$

²⁴As we argued in Section 2, when covariates are continuous it should be possible to extend the results of Graham and Powell (2008) and let h tend to zero at a suitable rate in order to obtain consistent estimates of moments of individual effects on the whole population of individuals.

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

$$\widehat{\boldsymbol{\delta}} = \left(\sum_{i=1}^N \mathbf{Z}'_i \mathbf{A}'_i (\mathbf{A}_i \mathbf{\Psi}_i \mathbf{A}'_i)^{-1} \mathbf{A}_i \mathbf{Z}_i \right)^{-1} \sum_{i=1}^N \mathbf{Z}'_i \mathbf{A}'_i (\mathbf{A}_i \mathbf{\Psi}_i \mathbf{A}'_i)^{-1} \mathbf{A}_i \mathbf{y}_i. \quad (45)$$

When $\mathbf{\Psi}_i = \mathbf{I}_T$, $\widehat{\boldsymbol{\delta}}$ is the OLS estimator of $\boldsymbol{\delta}$ in the within-group equations (3). When $\mathbf{\Psi}_i$ is such that $(\mathbf{A}_i \mathbf{\Psi}_i \mathbf{A}'_i)^{-1} = (\mathbf{A}_i \mathbf{V}_i \mathbf{A}'_i)^{-1}$, $\widehat{\boldsymbol{\delta}}$ coincides with the infeasible GLS estimator of $\boldsymbol{\delta}$. To construct a feasible version that is semiparametric efficient, the quantity $\mathbf{A}_i \mathbf{V}_i \mathbf{A}'_i = \mathbf{E}(\mathbf{A}_i \mathbf{v}_i \mathbf{v}'_i \mathbf{A}'_i | \mathbf{W}_i)$ needs to be replaced by a consistent estimator. Note that $\mathbf{A}_i \mathbf{v}_i = \mathbf{A}_i \mathbf{y}_i - \mathbf{A}_i \mathbf{Z}_i \boldsymbol{\delta}$. Therefore, this is a standard application of semiparametric GLS as in Robinson (1987).²⁵

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

$$\widehat{\boldsymbol{\gamma}} = \frac{\sum_{i=1}^N d_i(h) (\mathbf{X}'_i \mathbf{\Psi}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{\Psi}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}})}{\sum_{i=1}^N d_i(h)}, \quad (46)$$

where $d_i(h) = \mathbf{1}\{\det(\mathbf{X}'_i \mathbf{X}_i) > h\}$. In view of the discussion in the supplementary appendix, when $\mathbf{\Psi}_i$ is such that $(\mathbf{X}'_i \mathbf{\Psi}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{\Psi}_i^{-1} = (\mathbf{X}'_i \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{V}_i^{-1}$, the variance matrix of $\widehat{\boldsymbol{\gamma}}$ attains the efficiency bound.²⁶

A similar approach may be adopted to deal with Chamberlain's model given by equation (8). A method-of-moment estimator of $\boldsymbol{\theta}$ based on (15) will be consistent. A particular choice for the matrix $\mathbf{Q}_i(\boldsymbol{\theta})$ or its orthogonal decomposition yields semiparametric efficiency. Note that, in contrast, the fixed-effects estimator of $\boldsymbol{\theta}$ is inconsistent in general.²⁷

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

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

on a set of strictly exogenous regressors \mathbf{F}_i , yields consistent estimates for the coefficients of the projection of the population individual effects $\boldsymbol{\gamma}_i$ on the regressors \mathbf{F}_i . However, because

²⁵If $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$ (conditional homoskedasticity of \mathbf{v}_i with respect to \mathbf{W}_i), a feasible GLS estimator that replaces $\mathbf{A}_i \mathbf{V}_i \mathbf{A}'_i$ with $\mathbf{A}_i \widetilde{\boldsymbol{\Omega}} \mathbf{A}'_i$, where $\widetilde{\boldsymbol{\Omega}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}}) (\mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}})'$, would be asymptotically efficient.

²⁶Thus, feasible semiparametric efficient estimation of mean effects requires to estimate the conditional variance $\mathbf{Var}(\mathbf{y}_i | \mathbf{W}_i)$.

²⁷The key difference with the linear model (1)– where the fixed-effects estimator is consistent– is the dependence of $\mathbf{B}(\mathbf{W}_i, \boldsymbol{\theta})$ on the common parameters. In such a situation we can see from (18) that optimal estimation requires not only estimates of \mathbf{V}_i , but also of $\mathbf{E}(\boldsymbol{\gamma}_i | \mathbf{W}_i)$.

common parameters $\widehat{\boldsymbol{\delta}}$ have been estimated beforehand, the standard errors of the estimates of the projection coefficients need to be corrected. In particular, this point applies to the mean-group estimator of the unconditional mean $\boldsymbol{\gamma} = \mathbf{E}(\boldsymbol{\gamma}_i|\mathbb{S})$, given by (46). We provide corrected formulas in the supplementary 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 the denominator of the R^2 is the variance of individual effects, which is overestimated by the variance of $\widehat{\boldsymbol{\gamma}}_i$. In order to compute a correct R^2 , we need to consistently estimate the variance of $\boldsymbol{\gamma}_i$, which we discuss next.

5.2 Variances

We now turn to estimation of variances under the conditions of Theorem 1, that is under MA-type restrictions on the variance matrix of errors. The extension to autoregressive or ARMA structures presents no difficulty and will not be detailed here. In the following, \mathbf{A}^\dagger denotes the Moore-Penrose inverse of \mathbf{A} .

The following estimator of the unconditional variance matrix of errors, based on (23), uses covariance restrictions in levels:

$$\text{vec}\left(\widehat{\mathbf{Var}}(\mathbf{v}_i)\right) = \frac{1}{N} \sum_{i=1}^N \mathbf{S}_2 (\mathbf{M}_i \mathbf{S}_2)^\dagger \mathbf{M}_i (\widehat{\mathbf{v}}_i \otimes \widehat{\mathbf{v}}_i), \quad (47)$$

where \mathbf{M}_i is given by (22), and where we have denoted: $\widehat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}}$.

$\widehat{\mathbf{Var}}(\mathbf{v}_i)$ given by (47) will be consistent as long as (20) is satisfied. In the particular case where errors are i.i.d. with variance σ^2 , (25) motivates estimating σ^2 as:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N(T-q)} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}})' \mathbf{Q}_i (\mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}}) \\ &= \frac{1}{N(T-q)} \sum_{i=1}^N \widehat{\mathbf{v}}_i' \mathbf{Q}_i \widehat{\mathbf{v}}_i. \end{aligned} \quad (48)$$

The first-order asymptotic distributions of (47) and (48) can be easily derived. Standard arguments show that they coincide with the distribution treating common parameters $\boldsymbol{\delta}$ as known. Note that, while $\widehat{\sigma}^2$ is non-negative by construction, $\widehat{\mathbf{Var}}(\mathbf{v}_i)$ in (47) is not necessarily non-negative definite.

Turning to estimation of the variance of individual effects, a consistent estimator based

on (26) is:

$$\begin{aligned} \text{vec} \left(\widehat{\mathbf{Var}}(\gamma_i | \mathbb{S}) \right) &= \frac{\sum_{i=1}^N d_i(h) (\hat{\gamma}_i - \hat{\gamma}) \otimes (\hat{\gamma}_i - \hat{\gamma})}{\sum_{i=1}^N d_i(h)} \\ &\quad - \frac{\sum_{i=1}^N d_i(h) (\mathbf{H}_i \otimes \mathbf{H}_i) \mathbf{S}_2 (\mathbf{M}_i \mathbf{S}_2)^\dagger \mathbf{M}_i [\hat{\mathbf{v}}_i \otimes \hat{\mathbf{v}}_i]}{\sum_{i=1}^N d_i(h)}. \end{aligned} \quad (49)$$

Note that, as in the case of the variance of errors, the variance estimator $\widehat{\mathbf{Var}}(\gamma_i | \mathbb{S})$ in (49) is not necessarily non-negative definite.

In the special case where $\boldsymbol{\Omega}_i = \sigma^2(\mathbf{W}_i)\mathbf{I}_T$, an alternative estimator is:

$$\widehat{\mathbf{Var}}(\gamma_i | \mathbb{S}) = \frac{\sum_{i=1}^N d_i(h) (\hat{\gamma}_i - \hat{\gamma}) (\hat{\gamma}_i - \hat{\gamma})'}{\sum_{i=1}^N d_i(h)} - \frac{\sum_{i=1}^N d_i(h) \hat{\mathbf{v}}_i' \mathbf{Q}_i \hat{\mathbf{v}}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1}}{(T - q) \sum_{i=1}^N d_i(h)}. \quad (50)$$

Lastly, if in addition $\sigma^2(\mathbf{W}_i) = \sigma^2$ is independent of \mathbf{W}_i then we can estimate the variance of γ_i by:

$$\widehat{\mathbf{Var}}(\gamma_i | \mathbb{S}) = \frac{\sum_{i=1}^N d_i(h) (\hat{\gamma}_i - \hat{\gamma}) (\hat{\gamma}_i - \hat{\gamma})'}{\sum_{i=1}^N d_i(h)} - \hat{\sigma}^2 \frac{\sum_{i=1}^N d_i(h) (\mathbf{X}_i' \mathbf{X}_i)^{-1}}{\sum_{i=1}^N d_i(h)}, \quad (51)$$

where $\hat{\sigma}^2$ is given by (48). The estimator in (51) was introduced by Swamy (1970). Note that it is inconsistent in general if v_{it} is conditionally heteroskedastic. In addition, both estimators given by (50) and (51) will be inconsistent if errors are not mutually uncorrelated given regressors.

Remark 1 (testing the covariance structure of errors). In practice, one may wish to construct a specification test of the order of the MA process of the error terms. This is of special importance in order to estimate the variance of individual effects, as misspecifying the form of the variance matrix of errors would result in inconsistent estimates. A test of Assumption 2 may be based on the following overidentifying restrictions, which are an immediate consequence of (23):

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

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

Let us assume to simplify the notation that \mathbb{S} is the full population. Guided by the form of the optimal moments, we can consider estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\phi}})$ that solve the following estimating equations (using the notation of Subsection 3.3):

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

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

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

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

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

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

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

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

5.3 Densities

When covariates are discrete, the analytical formulas based on characteristic functions that we derived in Section 4 may be used to construct consistent estimators of distributions, as we now explain.

When errors v_{it} are independent over time, Assumption 5 holds for the $T^2 \times T$ selection matrix \mathbf{S}_2 that has zeros everywhere except at positions $(1, 1), (T + 2, 2), \dots, (T^2, T)$. We thus estimate the second derivatives of log-characteristic functions of errors as:

$$\frac{d^2 \ln \widehat{\Psi}_{v_{it} | \mathbf{w}_i}(\boldsymbol{\tau}_t | \mathbf{w})}{d\boldsymbol{\tau}^2} = \mathbf{m}'_{it} \mathbf{vec} \left(\frac{\partial^2 \ln \widehat{\Psi}_{\widehat{\mathbf{v}}_i | \mathbf{w}_i}(\boldsymbol{\tau} | \mathbf{w})}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} \right), \quad \boldsymbol{\tau} = (\tau_1, \dots, \tau_T)' \in \mathbf{R}^T, \quad (52)$$

where \mathbf{m}'_{it} is the t th row of the $T \times T^2$ matrix $(\mathbf{M}_i \mathbf{S}_2)^\dagger \mathbf{M}_i$, and where

$$\widehat{\Psi}_{\widehat{\mathbf{v}}_i | \mathbf{w}_i}(\boldsymbol{\tau} | \mathbf{w}) = \frac{\sum_{i=1}^N \mathbf{1}\{\mathbf{W}_i = \mathbf{w}\} \exp(j \boldsymbol{\tau}' \widehat{\mathbf{v}}_i)}{\sum_{i=1}^N \mathbf{1}\{\mathbf{W}_i = \mathbf{w}\}} \quad (53)$$

is the empirical characteristic function of $\widehat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{Z}_i \widehat{\boldsymbol{\delta}}$, in a cell \mathbf{w} .²⁸

The characteristic function of v_{it} is then estimated by successive integration, as:

$$\widehat{\Psi}_{v_{it}|\mathbf{w}_i}(\tau|\mathbf{w}) = \exp\left(\int_0^\tau \int_0^\nu \frac{d^2 \ln \widehat{\Psi}_{v_{it}|\mathbf{w}_i}(v|\mathbf{w})}{d\tau^2} dv d\nu\right), \quad \tau \in \mathbf{R}, \quad (54)$$

where $d^2 \ln \widehat{\Psi}_{v_{it}|\mathbf{w}_i}/d\tau^2$ is given by (52). Note that, by construction, $\widehat{\Psi}_{v_{it}|\mathbf{w}_i}(0|\mathbf{w}) = 1$ and $d \ln \widehat{\Psi}_{v_{it}|\mathbf{w}_i}(0|\mathbf{w})/d\tau = 0$.

Next, we estimate the characteristic function of γ_i as:

$$\widehat{\Psi}_{\gamma_i|\mathbf{w}_i}(\boldsymbol{\xi}|\mathbf{w}) = \frac{\widehat{\Psi}_{\widehat{\gamma}_i|\mathbf{w}_i}(\boldsymbol{\xi}|\mathbf{w})}{\widehat{\Psi}_{\mathbf{v}_i|\mathbf{w}_i}(\mathbf{H}'_i \boldsymbol{\xi}|\mathbf{w})}. \quad (55)$$

When errors are independent, $\widehat{\Psi}_{\mathbf{v}_i|\mathbf{w}_i}(\boldsymbol{\tau}|\mathbf{w}) = \prod_{t=1}^T \widehat{\Psi}_{v_{it}|\mathbf{w}_i}(\tau_t|\mathbf{w})$, where $\widehat{\Psi}_{v_{it}|\mathbf{w}_i}$ is given by (54). Equation (55), which relies on statistical independence (Assumption 3), relates the characteristic function of individual effects γ_i to that of least squares estimates $\widehat{\gamma}_i$.

More generally, when v_{it} follows an MA process with latent disturbances u_{is} , where $s = 1 - r, \dots, T$, a formula similar to (52) may be used to compute estimates of second derivatives $d^2 \ln \widehat{\Psi}_{u_{is}|\mathbf{w}_i}/d\zeta^2$, and thus characteristic function estimates $\widehat{\Psi}_{u_{is}|\mathbf{w}_i}$ by successive integration. In this case, $\widehat{\Psi}_{\mathbf{v}_i|\mathbf{w}_i}(\boldsymbol{\tau}|\mathbf{w}) = \widehat{\Psi}_{\mathbf{u}_i|\mathbf{w}_i}(\widehat{\boldsymbol{\Lambda}}' \boldsymbol{\tau}|\mathbf{w})$, where $\widehat{\boldsymbol{\Lambda}}$ is a consistent estimate of the $T \times (T + r)$ matrix $\boldsymbol{\Lambda}$ that defines the MA structure, see (43).²⁹

Given estimates of the characteristic functions, densities of absolutely continuous individual effects and errors may be estimated by inverse Fourier transformation (with trimming).³⁰ A nonparametric kernel deconvolution estimator of the conditional density of γ_i is:

$$\widehat{f}_{\gamma_i|\mathbf{w}_i}(\gamma|\mathbf{w}) = \frac{1}{(2\pi)^q} \int_{\mathbf{R}^q} K_N(\boldsymbol{\xi}) \exp(-j\boldsymbol{\xi}'\gamma) \widehat{\Psi}_{\gamma_i|\mathbf{w}_i}(\boldsymbol{\xi}|\mathbf{w}) d\boldsymbol{\xi}, \quad (56)$$

where $\widehat{\Psi}_{\gamma_i|\mathbf{w}_i}$ is given by (55), and where $K_N(\boldsymbol{\xi})$ is a truncation factor depending on the sample size N whose values go to zero when $|\boldsymbol{\xi}|$ tends to infinity. An example is $K_N(\boldsymbol{\xi}) = \mathbf{1}\{\boldsymbol{\xi} \in [-T_N, T_N]^q\}$, where T_N diverges to infinity with N .

The asymptotic properties of nonparametric kernel deconvolution estimators are well studied (Carroll and Hall, 1988, Fan, 1991). Although we do not derive the asymptotic

²⁸Note that (52) provides overidentifying restrictions that may be used to improve precision, as it holds for any vector $\boldsymbol{\tau} \in \mathbf{R}^T$ with t th element equal to τ_t . A possible choice is $\boldsymbol{\tau} = (0, \dots, 0, \tau_t, 0, \dots, 0)'$.

²⁹A consistent estimate $\widehat{\boldsymbol{\Lambda}}$ may be obtained using a minimum-distance approach that exploits the covariance restrictions derived in Section 3.

³⁰Note that, while the identification result of Theorem 2 holds for discrete or continuous random variables, the application of the inverse Fourier transformation requires continuity.

properties of the density estimator (56) in this paper, consistency and rates of convergence may be obtained using results from the existing literature (in particular, Bonhomme and Robin, 2010). As can be seen from (55), the rates will depend on the speed at which $\Psi_{\mathbf{v}_i|\mathbf{w}_i}(\mathbf{H}'_i\xi|\mathbf{w})$ tends to zero as $|\xi|$ tends to infinity.

A special case. Consider the the special case of Example 2, which is the setting of our empirical application in the next section:

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + v_{i\ell}, \quad \ell = 1, \dots, L, \quad (57)$$

where we assume that errors $v_{i\ell}$ are i.i.d. given (s_{i1}, \dots, s_{iL}) . Including strictly exogenous regressors poses no difficulty, as common parameters can be estimated beforehand.³¹

Consider a covariates sequence $\mathbf{s} = (s_1, \dots, s_L)'$. We focus on sequences where $s_{i\ell}$ changes over time. We have:

$$\Delta y_{i\ell} = \Delta v_{i\ell}, \quad \text{if } \Delta s_{i\ell} = 0, \quad (58)$$

$$\Delta s_{i\ell} \Delta y_{i\ell} = \beta_i + \Delta s_{i\ell} \Delta v_{i\ell}, \quad \text{if } \Delta s_{i\ell} \neq 0. \quad (59)$$

As errors are i.i.d. given $\mathbf{s}_i = \mathbf{s}$, it follows that all $\pm \Delta v_{i\ell}$, $\ell = 2, \dots, L$, have the same distribution. So one can interpret (59) as a simple deconvolution equation, where the right-hand side is the sum of the unobserved β_i , and the independent error $\pm \Delta v_{i\ell}$. In addition, because of equation (58), we also observe a random sample from $\Delta v_{i\ell}$. In particular, the characteristic function of $\Delta v_{i\ell}$ may be simply estimated as the empirical characteristic function of $\Delta y_{i\ell}$, on the subsample of observations such that $\Delta s_{i\ell} = 0$.³²

Continuous covariates. The restrictions that motivate the above estimators are conditional on covariates. When covariates are continuously distributed, the conditioning must be dealt with. A natural approach would be to rely on nonparametric kernel regression estimators in the construction of the empirical characteristic functions, see (53). However, the asymptotic properties of these conditional deconvolution estimators are *not* a direct application of existing results in the literature, and are beyond the scope of this article.³³

³¹In particular, common regressors \mathbf{Z}_i in (1) may be continuously distributed. The important requirement here is that \mathbf{X}_i (\mathbf{s}_i in the application) be discrete.

³²The previous simplification facilitates the estimation of the marginal density of β_i . If the joint density of (α_i, β_i) is sought, the general discussion applies.

³³Evdokimov (2009) recently derived the rate of convergence of a conditional nonparametric deconvolution estimator in a model with a scalar individual effect.

6 The effect of smoking on birth weight

In this section we study the effect of smoking during pregnancy on birth outcomes, building on Abrevaya (2006). Stata codes are available online.³⁴

6.1 Model and data

We estimate the following model:

$$y_{i\ell} = \alpha_i + \beta_i s_{i\ell} + \mathbf{z}_{i\ell}' \boldsymbol{\delta} + v_{i\ell}, \quad i = 1, \dots, N, \quad \ell = 1, \dots, L, \quad (60)$$

where i and ℓ index mothers and children, respectively.

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

Weight at birth strongly correlates with outcomes later in life. For this reason, the determinants of birth weight have been extensively studied.³⁵ Abrevaya (2006), using a panel data approach, finds strong negative effects of smoking on birth weights. He assumes that β_i is homogeneous across mothers in (60). Here we take advantage of the panel dimension to account for heterogeneity in the smoking effect.

The parameters α_i and β_i in model (60) are mother-specific effects. They stand for persistent health characteristics of the mother, which could be partly genetic. It is possible to interpret model (60) as describing a production function, the output of which being the child and the producer being the mother. The production technology is then represented by the mother-specific characteristics α_i and β_i . These characteristics are supposed to stay constant between births. In addition, they may be correlated with smoking status. In particular, a mother could decide not to smoke if she knows that her children will suffer from it (i.e., if she has a very negative β_i).

However, strict exogeneity (Assumption 1) requires that mothers will not change their smoking behavior because one of their children had a low birth weight, as the shocks $v_{i\ell}$ are assumed uncorrelated with the sequence of smoking statuses. This assumption will fail to hold if for example mothers do not know their α_i and β_i before they have had a child, and learning takes place over time. This is a common concern when estimating any type of

³⁴At: http://www.cemfi.es/~bonhomme/Random_codes.zip.

³⁵See Rosensweig and Wolpin (1991) for a study of various determinants. Studies of the effect of smoking during pregnancy on birth weight are Permutt and Hebel (1989), and Evans and Ringel (1999).

production function, where there can be feedback effects on the choice of inputs. We will attempt to relax the strict exogeneity assumption at the end of this section.

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

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

To allow for heterogeneity, we focus on mothers who had at least 3 children during the period (1989-1998). In the dataset, the number of children is exactly 3 per mother. In addition, we need the smoking indicator s_{il} to vary (at least once). So we only consider mothers who changed smoking status between the three births. The final sample contains 1445 mothers.³⁶

6.2 Results

Common parameters. We first estimate common parameters $\boldsymbol{\delta}$ in (60). For this, we use the generalized within-group estimator (45), with the identity as weighting matrix. The results are shown in Table 1. Although they have the expected signs, the variables indicating the number of prenatal visits and the quality of prenatal care are never significant. The only significant covariate is the gender of the child, boys having higher birth weight.

Average effects. We now turn to mother-specific effects. Table 2 shows the estimates of the moments of α_i and β_i . The mean smoking effect, computed using the mean-group formula (46) with the identity as weighting matrix, is -161 grams. This represents a negative

³⁶Descriptive statistics show that this subsample is somewhere in-between the subsample of women who always smoked, and the one of women who never smoked. For example, women who smoke during a larger number of pregnancies are younger on average, and their children have lower weight at birth.

Table 1: Estimates of common parameters δ

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 δ using (45) with $\Psi_i = \mathbf{I}_T$. The dataset is the “Matched panel data #3” in Abrevaya (2006). The sample only includes mothers who had three children and changed smoking status between births (1445 mothers). Standard errors are clustered at the mother level.

and significant effect of smoking on birth weight. Note that this value is close to the fixed-effects estimate obtained by Abrevaya: -144 g, when imposing homogeneity of the β ’s in model (60). In comparison, the mean of α_i is 2782 g.

To assess the predictability of the mother-specific effects, we estimate the projection coefficients in a regression of α_i and β_i on a set of mother-specific characteristics: the education of the mother, her marital status, and the mean of the smoking indicators over the three births.³⁷ Results are given in Table 3.³⁸

The estimates from the linear projection measure by how much the mother-specific effects α_i and β_i correlate with observed covariates. As a result, they have no causal interpretation. For example, Table 3 shows that black mothers have children with lower birth weight (lower α_i), though they seem to be *less* sensitive to smoking (*higher* β_i). Also, the children of mothers who smoke more have on average a lower α_i . 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 .³⁹ One can interpret this finding as a motivation for treating β_i as unobserved mother heterogeneity.

³⁷None of the covariates in Table 3 varies across births. This is due to the way mother-births pairs are matched in the dataset. See Abrevaya (2006).

³⁸The coefficient estimates are simply calculated by regressing the fixed-effects estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ on the mother-specific covariates. Standard errors are corrected as explained in Subsection 5.1.

³⁹Remark that the R^2 needs to be corrected, as explained in Subsection 5.1. For comparison, the uncorrected R^2 are .055 and .005 for α_i and β_i , respectively.

Table 2: Moments of α_i and β_i

Moment	Estimate	Standard error
Means		
Mean α_i	2782	435
Mean β_i	-161	17.0
Variances (i.i.d. errors)		
Variance α_i	127647	15161
Variance β_i	98239	21674
Covariance (α_i, β_i)	-52661	14375
Variances (non stationary errors)		
Variance α_i	120423	24155
Variance β_i	85673	34550
Covariance (α_i, β_i)	-45437	24165

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

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

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

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

Variiances. We now turn to variances of mother-specific effects. Rows 3 to 5 in Table 2 show the estimates of the coefficients of the variance matrix of (α_i, β_i) obtained from the levels restrictions, see (49), assuming that errors are i.i.d. given covariates.⁴⁰ Given the i.i.d. assumption, the estimates are numerically equal to those using the Swamy formula (51).

Both α_i and β_i show substantial dispersion. In particular, the standard deviation of β_i is 313 g.⁴¹ This can be compared to the standard deviation of 628 g of the least squares estimates $\widehat{\beta}_i$. So in this example, removing the sample noise due to the very small number of observations per mother (3 children) leads to a drastic decrease in the variance. In addition, the estimate of the correlation between α_i and β_i is -0.47 . Given those estimates, the standard deviation of $\alpha_i + \beta_i$ is estimated to be 347 g, compared to 357 g for α_i . This means that the two potential outcomes, for smokers and non smokers, have roughly the same variance.

Having three observations per mother, we need to impose strong restrictions on the variance matrix of errors in order to preserve identification. Using restrictions in levels, one can slightly relax the i.i.d. assumption. Rows 6 to 8 in Table 2 show variance estimates under a weaker assumption, which permits the variances of errors for the first, second and third children to be different. As we saw in Subsection 3.1, one cannot leave those three variances unrestricted, however. In rows 6 to 8 we impose that the variance of errors for the j th child is $a + bj$, where a and b are scalars.⁴² The results show that the variances of α_i and β_i are not much affected. For example, the standard deviation of β_i is now 292 g. This suggests that the i.i.d. assumption is not strongly rejected on these data.

Density and quantiles. Lastly, we present the estimates of the density and quantile function of the smoking effect β_i . In Section 5.3 we argued that, when covariates are discrete, densities of individual effects and errors may be estimated using nonparametric kernel deconvolution techniques.

One specific feature of our application, however, is the small signal-to-noise ratio (that is, $\text{Var}(\beta_i)/\text{Var}(\Delta v_{i\ell})$), roughly 23% according to our estimates. In Appendix B, we calibrate a simulation exercise to our empirical results. We find that kernel deconvolution estimates of the density are substantially biased when the signal-to-noise ratio is low. In addition, in the same section of the appendix we describe a deconvolution algorithm recently proposed

⁴⁰Hence, the selection matrix in (49) is $\mathbf{S}_2 = \text{vec}(\mathbf{I}_3)$.

⁴¹Interestingly, when including the number of cigarettes smoked during pregnancy as an additional control, the average smoking effect drops to -135 g, but the standard deviation remains almost unchanged.

⁴²Technically, this translates into a different selection matrix \mathbf{S}_2 in (49).

by Mallows (2007), and we compare its performance relative to kernel deconvolution. This simple algorithm, based on elementary simulations and permutations of the sample, shows remarkably good results. For this reason, the results we report on the left column of Figure 1 use Mallows' algorithm. For comparison, density and quantile function of the least squares estimates $\widehat{\beta}_i$ are reported on the right column of the figure.

We see that correcting for sample noise in the estimation has strong effects on density and quantile estimates. The density of β_i is much less dispersed than that of $\widehat{\beta}_i$, and its mode is much higher. In addition, our approach allows to estimate the smoking effect at various quantiles. When corrected for the presence of sample noise, the effect is mostly negative (up to percentile 75), and reaches very negative values for some mothers (around 400 g at percentile 20). This points to strong heterogeneity in the effect, suggesting that the cost of smoking (in terms of children outcomes) is very high for some mothers.

A surprising result apparent from the left column of Figure 1 is that for a high percentage of mothers (roughly 30%) smoking has a *positive* effect. This result could be due to misspecification. For example, errors v_{il} could be non i.i.d., in which case our correction to estimate the dispersion of β_i would be incorrect. With only three births per mother, there is little that can be done to overcome this problem. Another possibility is that smoking status is not strictly exogenous. The last part of this section focuses on this possibility.

6.3 Predeterminedness of smoking behavior

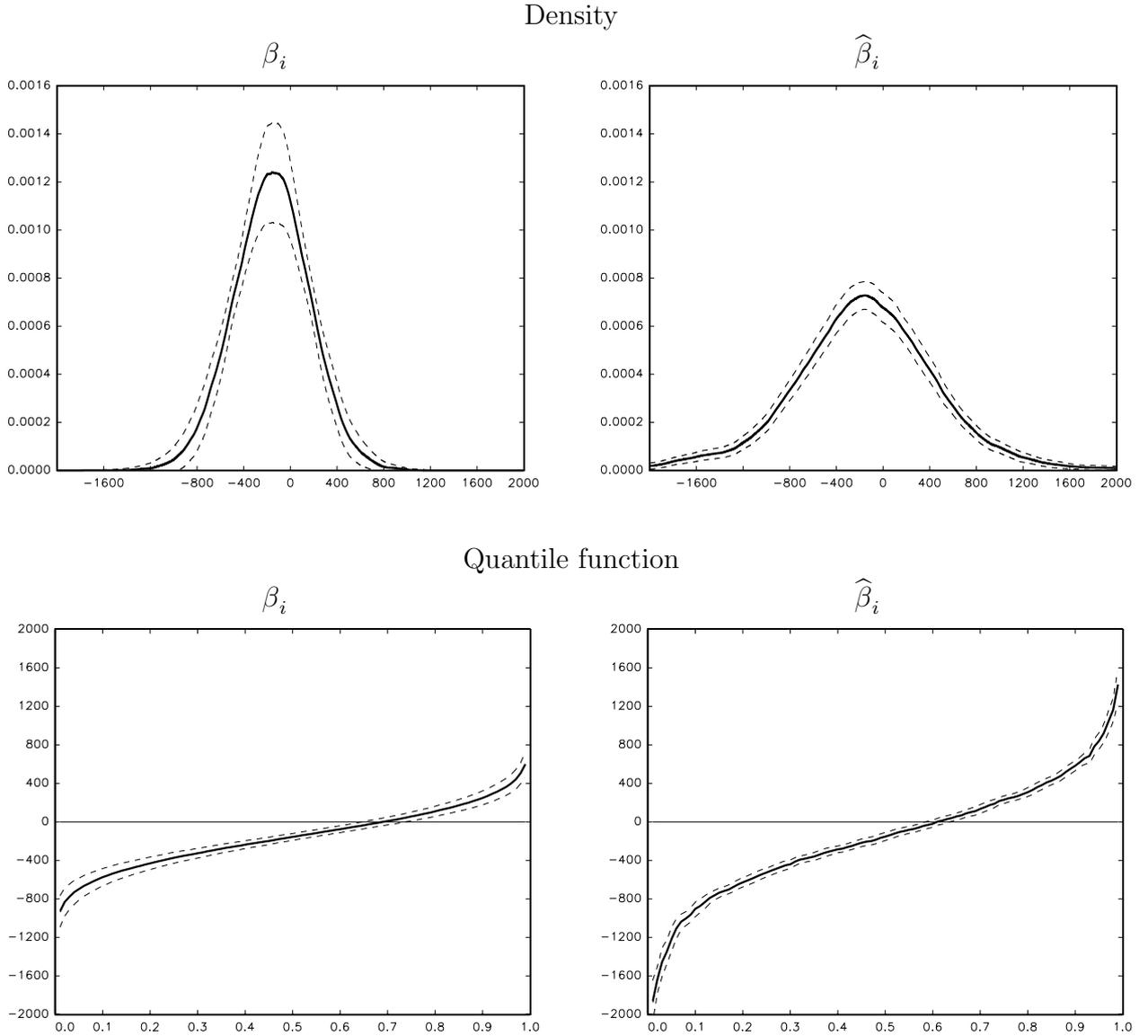
The previous results have been derived under the assumption that the smoking status is strictly exogenous. We now relax the strict exogeneity assumption and assume that smoking is predetermined in model (60), that is:

$$\mathbf{E}(v_{il} | \alpha_i, \beta_i, s_{il}, s_{i,\ell-1}, \dots) = 0. \quad (61)$$

Condition (61) is less restrictive than the strict exogeneity condition (Assumption 1). In particular, (61) could hold in contexts where mothers react to an unexpected birth outcome by changing their smoking behavior.

We consider a simple version of the model without exogenous time-varying regressors. Including time-varying regressors reduces the possibilities of point identification of effects of interest, requiring to restrict the correlation between individual effects and regressors.

Figure 1: Density and quantile estimates of the smoking effect



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

Taking differences between child ℓ and child $m < \ell$ we have:

$$y_{i\ell} - y_{im} = \beta_i [s_{i\ell} - s_{im}] + v_{i\ell} - v_{im}. \quad (62)$$

It turns out that interesting average effects are point identified in this framework under the predeterminedness condition (61). To see why, remark that, for $k = 0, 1$:

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

where we have used that, because of (61), both $v_{i\ell}$ and v_{im} are mean independent of s_{im} .

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

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

Hence, the following average effects are identified:

$$\mathbf{E}(\beta_i | s_{im} = k, s_{i\ell} = 1 - k) = (1 - 2k) \frac{\mathbf{E}(y_{i\ell} - y_{im} | s_{im} = k)}{\Pr(s_{i\ell} = 1 - k | s_{im} = k)}. \quad (63)$$

Table 4: Average smoking effects under predeterminedness

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

Note: Estimates of the mean of β_i in model (60) without exogenous regressors, for various smoking sequences. For example, (0, 1, .) refers to mothers who did not smoke during the pregnancy of their first child, and smoked while pregnant of their second child. Estimates in column 1 are computed under predeterminedness of the smoking status, while estimates in column 3 are computed under strict exogeneity. Standard errors are clustered at the mother level.

We report empirical estimates of (63) in Table 4, for various values of m , ℓ and k . In the same table (column 3), we report the estimates calculated under strict exogeneity.⁴³ Table

⁴³That is, computing the mean of $\hat{\beta}_i$ on the various sequences of smoking statuses.

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

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

Another interesting result from Table 4 is that, though quantitatively distinct, the results obtained under predeterminedness and strict exogeneity of smoking behavior are qualitatively similar. For example, under strict exogeneity the mean effect is -189 g for mothers who quitted smoking between the first and second child, while it is -117 g for mothers who started to smoke during the second pregnancy, the difference being significant at 5%. Indeed, none of the effects obtained under strict exogeneity is statistically different from the one obtained under predeterminedness (for a given smoking sequence) at the 5% level.⁴⁴ This suggests that the strict exogeneity assumption is not unreasonable on these data.

7 Conclusion

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

In many microeconomic applications, it is of interest to estimate the distributions of individual-specific effects. We have provided fixed- T identification results for variances and

⁴⁴The only significant difference at the 10% level is the one for the sequence (1, 0, .). In addition, a joint Wald test of significance has a p-value of .60.

more generally distributions of random coefficients and time-varying errors, in linear panel data models with strictly exogenous regressors. Distributional characteristics of individual effects (other than the mean) are not identified under the assumptions of unrestricted intertemporal distribution of the errors and unrestricted distribution of the effects conditioned on the regressors. In our results we have exploited the identifying content of limited time dependence of time varying errors.

In addition, we have proposed fixed- T consistent estimators of variances and densities. Density estimators rely on a conditional nonparametric deconvolution approach. We have not studied the asymptotic properties of the density estimators. When covariates are discrete, proving consistency and deriving rates of convergence is a simple extension of the existing literature. When covariates are continuous, however, the extension is not immediate and is an interesting topic for future work.

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

References

- [1] Aaronson, D., L. Barrow, and W. Sander (2007): "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics*, 25, 95-135.
- [2] Abrevaya, J. (2006): "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," *Journal of Applied Econometrics*, vol. 21(4), 489-519.
- [3] Arellano, M., and O. Bover (1995): "Another Look at the Instrumental-Variable Estimation of Error-Components Models," *Journal of Econometrics*, 68, 29-51.
- [4] 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.

- [5] 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.
- [6] Beran, R., A. Feuerverger, and P. Hall (1996): “On Nonparametric Estimation of Intercept and Slope Distributions in Random Coefficient Regression,” *Annals of Statistics*, 24(6), 2569–2592.
- [7] Bonhomme, S., and J. M. Robin (2010): “Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics,” *Review of Economic Studies*, 77(2), 491–533.
- [8] Cameron, C., and P.K. Trivedi (2005): *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.
- [9] Carrasco, M., and J. P. Florens (2009): “Spectral Methods for Deconvolving a Density,” to appear in *Econometric Theory*.
- [10] Carroll, R. J., and P. Hall (1988): “Optimal rates of Convergence for Deconvoluting a Density,” *Journal of the American Statistical Association*, 83, 1184-1186.
- [11] Chamberlain, G. (1992): “Efficiency Bounds for Semiparametric Regression”, *Econometrica*, 60, 567–596.
- [12] Chamberlain, G. (1993): “Feedback in Panel Data Models”, unpublished manuscript, Department of Economics, Harvard University.
- [13] Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2009): “Identification and Estimation of Marginal Effects in Nonlinear Panel Models,” to appear in *Econometrica*.
- [14] Demidenko, E. (2004): *Mixed Models. Theory and Applications*, John Wiley & Sons.
- [15] Dobbelaere, S., and J. Mairesse (2007): “Panel Data Estimates of the Production Function and Product and Labor Market Imperfections”, unpublished manuscript.
- [16] Dudley (2002): *Real Analysis and Probability*, Cambridge University Press.
- [17] Evans, W. N., and J. S. Ringel (1999): “Can Higher Cigarette Taxes Improve Birth Outcomes?,” *Journal of Public Economics*, 72, 135-154.

- [18] Evdokimov, K. (2009): “Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity,” unpublished manuscript.
- [19] Evdokimov, K., and H. White (2010): “An Extension of a Lemma of Kotlarski,” unpublished working paper.
- [20] Fan, J. Q. (1991): “On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems,” *Annals of statistics*, 19, 1257–1272.
- [21] Graham, B.S., and J.L. Powell (2008): “Identification and Estimation of Irregular Correlated Random Coefficient Models,” unpublished manuscript.
- [22] Guvenen, F. (2009): “An Empirical Investigation of Labor Income Processes,” *Review of Economic Dynamics*, 12, 58-79.
- [23] 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.
- [24] Hoderlein, S., J. Klemelä, and E. Mammen (2010): “Analyzing the Random Coefficient Model Nonparametrically,” *Econometric Theory*, 26, 804–837.
- [25] Holtz-Eakin, D., W. Newey, and H. Rosen (1988): “Estimating Vector Autoregressions with Panel Data”, *Econometrica*, 56, 1371–1395.
- [26] Horowitz, J. L., and M. Markatou (1996): “Semiparametric Estimation of Regression Models for Panel Data”, *Review of Economic Studies*, 63, 145–168.
- [27] Hu, Y., and S.M. Schennach (2008): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76(1), 195-216.
- [28] Kotlarski, I. (1967): “On Characterizing the Gamma and Normal Distribution,” *Pacific Journal of Mathematics*, 20, 69-76.
- [29] Li, T., and Q. Vuong (1998): “Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators,” *Journal of Multivariate Analysis*, 65, 139–165.
- [30] Lillard, L., and Y. Weiss (1979): “Components of Variation in Panel Earnings Data: American Scientists, 1960-70,” *Econometrica*, Vol.47, 437-454.

- [31] MaCurdy, T. (1981): “An Empirical Model of Labor Supply in a Life-Cycle Setting,” *Journal of Political Economy*, 89, 1059-1085.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] Neyman, J. and E. L. Scott (1948): “Consistent Estimates Based on Partially Consistent Observations”, *Econometrica*, 16, 1–32.
- [36] Permutt, T., and J. R. Hebel (1989): “Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight,” *Biometrics*, 45, 619-622.
- [37] Robinson, P. (1987): “Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form,” *Econometrica*, 55, 855-891.
- [38] Rosensweig, M.R., and K.I. Wolpin (1991): “Inequality at Birth : The Scope for Policy Intervention,” *Journal of Econometrics*, 50, 205-228.
- [39] Swamy, P. A. (1970): “Efficient Inference in a Random Coefficient Model,” *Econometrica*, 38, 311–323.
- [40] Székely, G.J., and C.R. Rao (2000): “Identifiability of Distributions of Independent Random Variables by Linear Combinations and Moments,” *Sankhyä*, 62, 193-202.
- [41] Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

APPENDIX

A Proofs

A.1 Proofs

Proposition 1. We have, using Assumption 1:

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

Likewise, again using Assumption 1:

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

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

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

By the law of iterated expectations we obtain:

$$\mathbf{E}(\mathbf{F}_i\widehat{\boldsymbol{\gamma}}_i|\mathbb{S}) = \mathbf{E}(\mathbf{F}_i\boldsymbol{\gamma}_i|\mathbb{S}).$$

Lastly, (6) implies that $\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i|\mathbb{S}) = \mathbf{E}(\boldsymbol{\gamma}_i|\mathbb{S})$, so:

$$\mathbf{Cov}(\mathbf{F}_i, \boldsymbol{\gamma}_i|\mathbb{S}) = \mathbf{E}(\mathbf{F}_i\boldsymbol{\gamma}_i|\mathbb{S}) - \mathbf{E}(\mathbf{F}_i|\mathbb{S})\mathbf{E}(\boldsymbol{\gamma}_i|\mathbb{S}) = \mathbf{E}(\mathbf{F}_i\widehat{\boldsymbol{\gamma}}_i|\mathbb{S}) - \mathbf{E}(\mathbf{F}_i|\mathbb{S})\mathbf{E}(\widehat{\boldsymbol{\gamma}}_i|\mathbb{S}) = \mathbf{Cov}(\mathbf{F}_i, \widehat{\boldsymbol{\gamma}}_i|\mathbb{S}).$$

The conclusion follows.

Corollary 2. Similar to the proof of Proposition 1.

Theorem 1. It follows from (23) and (24) that $\boldsymbol{\omega}_i$ is identified. Hence $\boldsymbol{\Omega}_i$ is identified. So, by (19), $\mathbf{X}_i\mathbf{E}(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i'|\mathbf{W}_i)\mathbf{X}_i'$ is also identified.

For any $i \in \mathbb{S}$, \mathbf{X}_i has full-column rank. So, $\mathbf{E}(\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i'|\mathbf{W}_i, \mathbb{S})$ is identified. This ends the proof.

Proof of (25). In matrix form, (23) becomes:

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

Applying the trace operator yields, in the particular case where errors are i.i.d. independent of \mathbf{W}_i with variance σ^2 :

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

where we have used that $\text{Tr}(\mathbf{Q}_i) = T - q$. Hence:

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

Theorem 2. We need the following elementary lemma.

Lemma A1 Let $g : \mathbf{R}^L \rightarrow \mathbf{R}$ be a twice-differentiable function such that $\frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} = \mathbf{0}$ for all $\mathbf{x} \in \mathbf{R}^L$, $\frac{\partial g(\mathbf{0})}{\partial \mathbf{x}} = \mathbf{0}$, and $g(\mathbf{0}) = 0$. Then $g(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbf{R}^L$.

Proof. As $\frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} \right) = \mathbf{0}$, it follows that $\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$ is a constant, equal to zero as $\frac{\partial g(\mathbf{0})}{\partial \mathbf{x}} = \mathbf{0}$. Hence $g(\mathbf{x})$ is a constant, equal to zero as $g(\mathbf{0}) = 0$. ■

Similarly as in the proof of Theorem 1, it follows from (24) and (41) that $\boldsymbol{\omega}_i(\boldsymbol{\tau})$ is identified for all $\boldsymbol{\tau} \in \mathbf{R}^T$. Hence

$$\frac{\partial^2 \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} = \mathbf{S}_2 \boldsymbol{\omega}_i(\boldsymbol{\tau}), \quad \boldsymbol{\tau} \in \mathbf{R}^T,$$

is identified.

Note that, because of Assumption 1:

$$\frac{\partial \ln \Psi_{\mathbf{v}_i | \mathbf{W}_i}(\mathbf{0} | \mathbf{W}_i)}{\partial \boldsymbol{\tau}} = \mathbf{E}(\mathbf{v}_i | \mathbf{W}_i) = \mathbf{0}.$$

In addition, because of the definition of a characteristic function:

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

Moreover, by Lemma A1 it follows that the log-characteristic function of errors is uniquely determined by these restrictions.

Next, using (41) we have that:

$$\mathbf{X}_i \left(\frac{\partial^2 \ln \Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i}(\mathbf{X}'_i \boldsymbol{\tau} | \mathbf{W}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'} \right) \mathbf{X}'_i, \quad \boldsymbol{\tau} \in \mathbf{R}^T,$$

is identified. Using that \mathbf{X}_i has full-column rank in \mathbb{S} , this implies that:

$$\frac{\partial^2 \ln \Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}(\boldsymbol{\xi} | \mathbf{W}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'}, \quad \boldsymbol{\xi} \in \mathbf{R}^q,$$

is also identified. Next, noting that $\ln \Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}(\mathbf{0} | \mathbf{W}_i) = 0$, and that:

$$\frac{\partial \ln \Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}(\mathbf{0} | \mathbf{W}_i)}{\partial \boldsymbol{\xi}} = \mathbf{E}(\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S})$$

is identified by Proposition 1, it follows from Lemma A1 that $\ln \Psi_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}$ is identified.

Lastly, distribution functions $f_{\mathbf{v}_i | \mathbf{W}_i}$ and $f_{\boldsymbol{\gamma}_i | \mathbf{W}_i, \mathbb{S}}$ are identified by the uniqueness property of characteristic functions (e.g., Dudley, 2002, p.303).

This ends the proof.

B Mallows' algorithm (2007)

The algorithm. The model is:

$$A_i = B_i + C_i,$$

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

The algorithm is as follows.

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

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

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

Illustration. We here briefly present some simulation results, which suggest that Mallows’ algorithm works well in practice. We calibrate the data generating process (DGP) to the results of Section 6. Formally, the DGP is:

$$\begin{cases} y_{i1} &= \alpha_i + \beta_i + v_{i1} \\ y_{i2} &= \alpha_i + v_{i2} \\ y_{i3} &= \alpha_i + v_{i3}, \end{cases}$$

where β_i , v_{i1} , v_{i2} , and v_{i3} follow independent normal distributions.⁴⁵ The mean of β_i is -150 and its standard deviation is 300 , while $v_{i\ell}$ has zero mean and standard deviation $\sigma \in \{100, 300, 400, 500\}$. For comparison, our empirical estimate of the standard deviation is $\hat{\sigma} \approx 450$.

We apply Mallows’ algorithm to equations (58) and (59).⁴⁶ We also estimate the density of β_i using a nonparametric kernel deconvolution estimator. As an infeasible choice for the truncation parameter T_N , we minimize the mean integrated squared error (MISE) of the density estimate on an equidistant grid of 99 points on $(1/1500, 9/1500)$.

Figure B1 shows the results of 1000 simulations. The column on the left of the figure shows estimates using Mallows’ algorithm, while the column on the right refers to kernel deconvolution estimates. We report the median, and 10% and 90% quantiles across simulations, along with the normal density.

The results show that Mallows’ algorithm performs well, showing no bias and tight confidence bands. By comparison, kernel deconvolution estimates perform worse. This is especially so when σ is large and the signal-to noise ratio in the first-differenced equation is small. This situation arises in the empirical application of Section 6. In this case, the simulation-based approach of Mallows seems to strongly outperform characteristic-function based estimators.

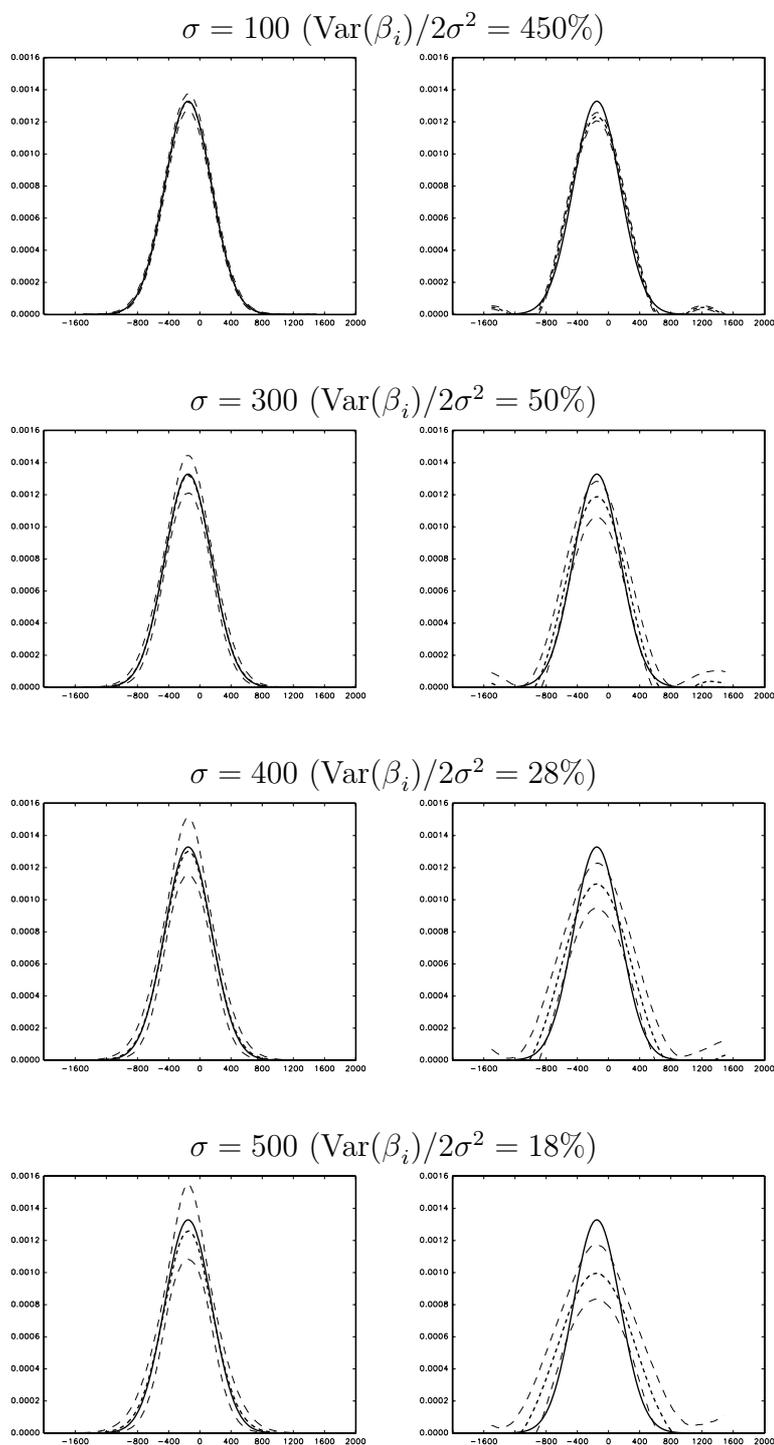
⁴⁵Note that the presence of α_i does not affect estimation.

⁴⁶The density of β_i is estimated using a Gaussian kernel with a rule-of-thumb bandwidth.

Figure B1: Estimates of f_{β_i} on simulated data

Mallows' algorithm

Kernel deconvolution



Note: The DGP is that of Example 2 with parameters roughly chosen to match the empirical results (where $\hat{\sigma} \approx 450$). Thick dashed line is the pointwise median across 1000 simulations, thin dashed lines are the 10%-90% pointwise confidence bands. The thick solid line is the truth. For kernel deconvolution, the truncation parameter T_N minimizes the MISE on a grid of values.