

Functional Differencing*

Stéphane Bonhomme

CEMFI, Madrid

Revised version: October 2011

Abstract

In nonlinear panel data models, the incidental parameter problem remains a challenge to econometricians. Available solutions are often based on ingenious, model-specific methods. In this paper, we propose a systematic approach to construct moment restrictions on common parameters that are free from the individual fixed effects. This is done by an orthogonal projection that differences out the unknown distribution function of individual effects. Our method applies generally in likelihood models with continuous dependent variables where a condition of non-surjectivity holds. The resulting method-of-moments estimators are root- N consistent (for fixed T) and asymptotically normal, under regularity conditions that we spell out. Several examples and a small-scale simulation exercise complete the paper.

JEL CODE: C23.

KEYWORDS: Panel data, incidental parameters, inverse problems.

*I thank the Co-Editor and four anonymous referees for very useful comments. I also thank Manuel Arellano, Alan Bester, Marine Carrasco, Gary Chamberlain, Bryan Graham, Jin Hahn, Chris Hansen, Jim Heckman, Joel Horowitz, Yingyao Hu, Konrad Menzel, Ulrich Müller, Jim Powell, Elie Tamer, Harald Uhlig, and seminar participants at Berkeley, Boston University, Brown University, CEMFI, University of Chicago, Chicago Booth, Harvard University, Université de Montréal, New York University, Northwestern University, Toulouse School of Economics, University College London, and University of Toronto. Support from the European Research Council / ERC grant agreement n^o263107 is gratefully acknowledged. All errors are mine.

1 Introduction

A large amount of empirical work has demonstrated the usefulness of panel data to control for unobserved individual heterogeneity. In applications, a common approach is to specify a model that contains a finite-dimensional vector of parameters that are common across individuals, and one or various unit-specific parameters (“fixed effects”) that may reflect heterogeneity in ability, preferences, or technology.

Since the important paper by Neyman and Scott (1948), it is known that a maximum likelihood approach that treats the individual fixed effects as parameters to be estimated may provide inconsistent estimates of common parameters. This “incidental parameter” problem arises because the number of fixed effects grows with the sample size, violating a condition for consistency of maximum likelihood.

For several decades, econometricians and statisticians have proposed various solutions to the incidental parameter problem (see Lancaster, 2000). In the spirit of linear models, where differencing out the individual effects yields moment restrictions on common parameters alone, ingenious methods have been proposed to difference out the fixed effects in various nonlinear panel data models. A celebrated example is the conditional maximum likelihood approach of Rasch (1961) and Andersen (1970) in the static logit model.¹

In a likelihood context, one reaction to the problem is to try to isolate a component in the likelihood that does not depend on the individual effects. This can be done when the likelihood factors, as in the Poisson counts model with reparameterized effects. In general, however, exact separation is not possible. Cox and Reid (1987) proposed an approximate separation procedure, a Bayesian variant of which was applied to panel data models by Lancaster (2002). Estimators based on this idea are first-order unbiased as T increases, although they are not fixed- T consistent in general.²

Another reaction to the incidental parameter problem is to impose some structure on the distribution of unobserved heterogeneity, thereby following a (correlated) random-effects approach (e.g., Chamberlain, 1984). Parametric specifications are popular in applied work. More general semiparametric approaches based on sieve and penalized sieve estimators are

¹Honoré and Kyriazidou (2000) provide a dynamic generalization of this insight. Other nonlinear models where a modified differencing approach has been applied are censored regression models with fixed effects (Honoré, 1992, 1993, Hu, 2002), sample selection models (Kyriazidou, 1997, 2001), multiple-spell duration models (Chamberlain, 1985, Horowitz and Lee, 2004), and linear models with variance dynamics (Meghir and Windmeijer, 2000).

²See Arellano and Hahn (2006) for a survey of the bias correction literature in panel data.

now available.³ In panel data applications, however, the presence of conditioning regressors and initial conditions may complicate the practical implementation of sieve-based methods.

In this paper, we propose a systematic approach to difference out the individual effects and provide restrictions on common parameters alone. We adopt a likelihood setup where T is fixed and, following a “fixed-effects” perspective, the conditional distribution of individual effects given exogenous regressors and initial conditions is left unrestricted.

For a given value of common parameters, the panel data model maps the unknown distribution function of individual effects to the distribution function of the data. The main idea is to search for functions that belong to the orthogonal complement of the range (or image) of that mapping. By construction, such functions have zero expectation, and provide moment restrictions on common parameters. Our approach thus transforms the difficult problem of removing the “incidental” individual effects into a well-defined mathematical problem: constructing functions that belong to the orthogonal complement of a set of functions.

To illustrate the idea, we consider three examples where ingenious ways of differencing out the individual effects have been proposed: the random coefficients model of Chamberlain (1992), the censored regression model of Honoré (1992), and the static logit model. In all three examples, our systematic search for functions that are orthogonal to the range of the model mapping delivers the proposed methods as special cases.

Moreover, as our approach is general, it may be applied to models where differencing strategies are not yet known. A simple example is a censored random coefficients model, for which we derive new analytical moment restrictions on the parameters. Another possible application is a model of a nested CES production function with heterogeneous elasticity of substitution between inputs (e.g., Duffy *et al.*, 2004). In this nonlinear regression model, our approach allows to numerically construct moment restrictions on structural technology parameters. More generally, our approach is well-suited to deal with continuous outcomes models with unobserved heterogeneity, potential applications including structural models of firm investment (Cooper and Haltiwanger, 2006), or multiple-spell duration models in the absence of proportionality (Van den Berg, 2001).

In analogy with standard differencing methods, moment functions may be constructed using an orthogonal projection. To describe our approach, we start with the special case where the distributions of the data and the unobserved heterogeneity have known finite supports.

³See Chen (2007) for a survey of sieve methods in econometrics, and Hu and Schennach (2008) and Bester and Hansen (2007) for applications to latent variables models.

This case is useful for expositional purposes, as its treatment requires only elementary linear algebra. Indeed, the range of the model is just the finite-dimensional vector space spanned by the columns of a (parameter-dependent) matrix of conditional probabilities. Elements that belong to the orthogonal complement of the range can then be constructed using a “within” projection matrix. In effect, this projection differences out the unknown vector of probabilities of individual effects. We refer to this procedure as functional differencing.

The moment restrictions obtained in this way may be uninformative about the parameter of interest, however. In particular, this will happen when the range of the model spans the whole space. A condition of *non-surjectivity* is thus necessary for our approach to yield useful restrictions. In the finite support case, a sufficient condition for non-surjectivity is that the support of the outcomes be richer than that of the individual effects. As an example, non-surjectivity generally fails in binary choice models, the static logit model being an important exception.

When all variables in the model (including covariates) have known finite support, the model is parametric. We show that the moment restrictions based on functional differencing achieve the Cramer-Rao bound. This result is interesting, given that our projection-based approach makes no use of the fact that the probabilities of individual effects belong to the unit simplex.⁴

When supports are infinite, the matrix of conditional probabilities becomes a linear integral operator. Building on the recent econometric literature on inverse problems (Carrasco, Florens and Renault, 2008, Carrasco and Florens, 2009), we endow the spaces of functions with scalar products and make them Hilbert spaces. This construction allows us to define a “within” projection operator that projects functions of the dependent variables onto the orthogonal complement of the range of the model operator. Evaluated at the distribution function of the data, this projection yields a set of restrictions on common parameters alone. Although the within projection operator is generally not available in closed form, it can be approximated using a simple discretization strategy.

As in the finite support case, the functional differencing restrictions can be equivalently written as a system of moment restrictions, conditional on regressors. This means that

⁴In models with zero information bound, however, the restrictions from functional differencing may be completely uninformative. As an important special case, recent work has studied models where the parameter of interest is partially identified. See Honoré and Tamer (2006), who compute the identified sets for the autoregressive parameter in a dynamic probit model, and Chernozhukov *et al.* (2009), who estimate bounds on marginal effects in binary choice.

a nonparametric estimate of the outcome distribution function is not needed to estimate common parameters. In infinite dimensions, one difficulty is that general sufficient conditions for non-surjectivity are no longer easy to find. We show that T being strictly greater than the dimension of the vector of individual effects is sufficient for non-surjectivity to hold in random coefficients models and censored regression models with normal errors. In nonlinear regression models with independent additive errors, in contrast, non-surjectivity requires further restrictions on the range of the regression function.

We estimate common parameters using generalized method-of-moments (GMM), based on a finite number of unconditional moment restrictions obtained from functional differencing. In practice, the choice of moment functions is important. One possibility is to combine a large number of restrictions, generated using a dictionary of functions. Another possibility is based on the insight that, as in the finite support case, there exists a finite subset of unconditional moment restrictions that achieves the information bound of the panel data model. This is because, given a suitable choice for the Hilbert space topology, the efficient score of the model may be interpreted as resulting from a particular functional differencing projection. Although efficient estimation of common parameters seems difficult in general, this result can be used as a guide to select moment restrictions in practice.

To study the asymptotic properties of the GMM estimator of common parameters, we impose regularity conditions that ensure that the model operator is *compact*. This convenient assumption allows us to work with the singular value decomposition of the projection operator. Under identification and regularity conditions that we provide, the GMM estimator is root- N consistent and asymptotically normal. Thus, in our approach, common parameters are estimated at a parametric rate in nonlinear models where the conditional distribution of individual effects is left unrestricted. Note, however, that root- N consistent estimation of common parameters does not imply root- N consistent estimation of average marginal effects. Estimation of the latter is addressed in a companion paper (Bonhomme, 2011).

The rest of the paper is as follows. In Section 2, we describe the model and our general approach. In Section 3, we present the differencing approach in the case where the data and individual effects have known finite supports. The finite support assumption is relaxed in Section 4. Asymptotic properties are studied in Section 5. Section 6 presents a small-scale numerical illustration. Lastly, Section 7 concludes.

2 Incidental parameters

In this section, we present the model and outline our approach in a few examples.

2.1 Likelihood models with fixed effects

Let $(y_{it}, x'_{it})'$, $i = 1, \dots, N$ and $t = 1, \dots, T$ be the set of observations of an endogenous variable y_{it} and a vector of strictly exogenous variables x_{it} . Let also α_i denote a vector of individual-specific unobserved effects. We assume that $y_i = (y_{i1}, \dots, y_{iT})$, $x_i = (x'_{i1}, \dots, x'_{iT})$, and α_i are jointly i.i.d. across individuals. The population contains an infinite number of individual units (large N), observed in a finite number of time periods (fixed T).

The distribution function of y_i conditioned on x_i and α_i is given by $f_{y|x,\alpha}(\cdot|x_i, \alpha_i; \theta_0)$, where $f_{y|x,\alpha}(\cdot|\cdot, \cdot; \theta)$ is a known function given $\theta \in \Theta$. The conditional distribution of α_i given x_i is denoted as $f_{\alpha|x}$. In addition, we let \mathcal{A} be a set that contains the support of $f_{\alpha|x}(\cdot|x)$ for all values of x . The population distribution function of y_i given $x_i = x$ is thus given by:

$$f_{y|x}(y|x; \theta_0, f_{\alpha|x}) = \int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta_0) f_{\alpha|x}(\alpha|x) d\alpha. \quad (1)$$

The model that we consider is semiparametric, because the distribution of the individual effects is not restricted. In particular, we do not restrict the dependence between α_i and x_i , thus following a “fixed-effects” approach.⁵ Conditional on the effects, however, the model is fully parametric. In addition, the model may incorporate dynamics such as:

$$f_{y|x,\alpha}(y|x, \alpha; \theta_0) = \prod_{t=1}^T f_{y_t|y^{(t-1)},x,\alpha}(y_t|y^{(t-1)}, x, \alpha; \theta_0),$$

where $y^{(t)} = (y_t, y_{t-1}, \dots)$, in which case x will contain strictly exogenous regressors and initial conditions. When y_t is a vector of random variables (e.g., outcome and state variables in a dynamic economic model), this representation allows for general predeterminedness and feed-back effects, as long as the researcher is willing to specify the feed-back process parametrically.

Since Neyman and Scott (1948), it is known that the maximum likelihood estimator of θ_0 is generally inconsistent for fixed T . Our aim is to provide restrictions on θ_0 that are free from the “incidental parameters” $\alpha_1, \dots, \alpha_N$, thus leading to fixed- T consistent estimators of common parameters. Our approach is general, and covers all semiparametric likelihood

⁵In common with much of the econometric literature, here we denote as “fixed-effects” an approach which assumes that individual effects and regressors are random draws from an unrestricted distribution.

models of the form (1). To facilitate exposition, we will illustrate how the approach works in various panel data models where standard maximum likelihood fails.

Example 1A: Chamberlain’s random coefficients model. As a first illustrative example, let us consider the model:

$$y_i = a(x_i, \theta_0) + B(x_i, \theta_0) \alpha_i + v_i, \quad (2)$$

where a ($T \times 1$) and B ($T \times \dim \alpha$) are known given x and θ . Chamberlain (1992) considers a version of (2) where errors are mean independent of regressors and effects. He proposes a quasi-differencing strategy that removes the fixed effects α_i , and provides restrictions on common parameters alone.⁶

In addition, here we assume that errors are normally distributed:

$$v_i | x_i, \alpha_i \sim N[0, \Sigma(x_i, \theta_0)], \quad (3)$$

where $\Sigma(\cdot, \cdot)$ is known. This framework includes as special cases linear models with individual-specific intercepts, models with interactive fixed effects, and dynamic autoregressive models.

In this model, the conditional density of the data is given by:

$$f_{y|x,\alpha}(y|x, \alpha; \theta) = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (y - a - B\alpha)' \Sigma^{-1} (y - a - B\alpha) \right], \quad (4)$$

where we have suppressed the reference to (x, θ) for conciseness.

Example 1B: censored random coefficients model. As a variation of the first example, let *latent* outcomes y_i^* follow the normal random coefficients model (2)-(3), with y_i^* in place of y_i . The difference with Example 1A is that $y_{it} = \max(y_{it}^*, c_t)$ is observed, where we assume that the censoring thresholds c_1, \dots, c_T are known to the researcher. In particular, note that (4) still holds in the censored model, for every y such that $y_t > c_t$ for all $t \in \{1, \dots, T\}$.

In the model with a single heterogeneous intercept: $y_{it} = \max(x'_{it} \beta_0 + \alpha_i + v_{it}, c_t)$, Honoré (1992) has derived restrictions on β_0 under the assumption that errors are i.i.d. (though not necessarily normally distributed). To our knowledge, no solution has been proposed to deal with censored models with general random coefficients.⁷

⁶Moreover, Chamberlain (1992) points out that joint estimation of θ_0 and $\alpha_1, \dots, \alpha_N$ will result in an inconsistent estimator for θ_0 when $B(x, \theta)$ depends on θ . This emphasizes the presence of an incidental parameter problem in this model.

⁷Note that the random coefficients framework covers as a special case censored regression models with lagged (latent) dependent variables as considered in Hu (2002).

Example 2: Static binary choice model. Our second example is a static panel data model with a binary dependent variable, where $y_{it} \in \{0, 1\}$ and y_{is} are independent given individual effects and regressors for all $t \neq s$. Let $F_t(x_{it}, \alpha_i, \theta) = \Pr(y_{it} = 1 | x_{it}, \alpha_i; \theta)$. In this case, $f_{y|x, \alpha}(\cdot | x, \alpha; \theta)$ is a conditional probability mass function that satisfies:

$$f_{y|x, \alpha}(y|x, \alpha; \theta) = \prod_{t=1}^T F_t(x, \alpha, \theta)^{y_t} (1 - F_t(x, \alpha, \theta))^{1-y_t}.$$

When errors are logistic, the conditional maximum likelihood estimator based on the sufficient statistic $y_{i1} + y_{i2}$ (for $T = 2$) is root- N consistent for θ_0 (Andersen, 1970). However, when errors are not logistic the semiparametric information bound for θ_0 is zero and there exists no root- N consistent estimator, although θ_0 may still be point-identified (Chamberlain, 2010).⁸

Example 3: Nonlinear regression. Our third example is a nonlinear regression model with individual-specific effects:

$$y_i = m(x_i, \alpha_i, \theta_0) + v_i, \quad (5)$$

where $m(\cdot, \cdot, \cdot)$ is a known $T \times 1$ function. In addition, v_i are assumed normal given by (3).

Nonlinear specifications of this form arise in production functions applications. An example could be the following heterogeneous constant elasticity of substitution (CES) production function:

$$\log y_{it} = \lambda \log \ell_{it} + (1 - \lambda) \log [\gamma h_{it}^{\sigma_i} + (1 - \gamma) k_{it}^{\sigma_i}]^{1/\sigma_i} + \eta_i + v_{it}, \quad (6)$$

which allows for different degrees of complementarity between low-skill labor (ℓ_{it}), high-skill labor (h_{it}), and capital equipment (k_{it}). In this case, θ includes the structural parameters (λ, γ) as well as the variance of v_{it} , and $\alpha_i = (\eta_i, \sigma_i)'$.

In model (5) the conditional density of y_i given (x_i, α_i) is:

$$f_{y|x, \alpha}(y|x, \alpha; \theta) = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (y - m(\alpha))' \Sigma^{-1} (y - m(\alpha)) \right], \quad (7)$$

where we have suppressed the dependence on (x, θ) . Due to the nonlinearity, it does not seem possible to difference out α_i in a straightforward way. Our approach will allow to construct moment restrictions on θ_0 that do not depend on the individual-specific effects.

⁸Estimators that converge at a less-than-parametric rate have been proposed by Manski (1987) and more recently Hoderlein and White (2009).

2.2 Moment restrictions

The methods used to solve the incidental parameter problem in Examples 1 and 2 are *a priori* not obvious, and require the researcher to show considerable ingenuity. Moreover, once a solution has been discovered in one specific model, it is not always clear how to generalize the approach to even closely related models. The comparison between static logit and static probit models illustrates this difficulty.

To present our approach, it is convenient to introduce the following linear mapping which, for given values of θ and x , maps any function $g(\alpha)$ to the function $[L_{\theta,x}g](y)$ given by:

$$[L_{\theta,x}g](y) = \int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta) g(\alpha) d\alpha. \quad (8)$$

The linear integral operator $L_{\theta,x}$ represents the parametric part of the panel data model. It is a central object in this paper.⁹ In particular, as (1) can be written as: $L_{\theta_0,x}f_{\alpha|x} = f_{y|x}$, the operator $L_{\theta_0,x}$ may be understood as mapping the distribution function of individual effects to that of the data. We defer the precise mathematical definition and properties of $L_{\theta,x}$ until Section 4.

Suppose now that we have found a function $\varphi(\cdot, x, \theta)$ such that, for every function $g(\alpha)$:

$$\int_{\mathcal{Y}} \varphi(y, x, \theta) [L_{\theta,x}g](y) dy = 0, \quad (9)$$

where the integral is taken over the support of the data. Equation (9) means that φ is orthogonal (with respect to the L^2 scalar product) to the *range*– or *image*– of the operator $L_{\theta,x}$. It then follows that:

$$\begin{aligned} \mathbb{E} \left[\varphi(y_i, x_i, \theta_0) \mid x_i = x \right] &= \int_{\mathcal{Y}} \varphi(y, x, \theta_0) f_{y|x}(y|x) dy \\ &= \int_{\mathcal{Y}} \varphi(y, x, \theta_0) [L_{\theta_0,x}f_{\alpha|x}](y) dy = 0. \end{aligned} \quad (10)$$

As the conditional moment restrictions (10) do not involve the individual effects, this discussion suggests that one can difference out the “incidental” individual effects provided we solve the well-defined mathematical problem of finding some functions φ that satisfy (9). When the solutions to this problem are not available in closed form, the functional differencing approach will compute them numerically using projection methods. The next

⁹We follow the notation in Carrasco, Florens and Renault (2008), who provide an excellent overview of linear operators and their applications to econometrics.

section presents our approach to solve this mathematical problem, starting with the finite support case.

In the rest of this section, we illustrate this idea in our examples. Additional details may be found in the supplementary appendix to this paper.¹⁰

Example 1A (cont.) We introduce some notation. Denoting as $\left[\Sigma^{-\frac{1}{2}}B\right]^\dagger$ the Moore-Penrose generalized inverse of the matrix $\Sigma^{-\frac{1}{2}}B$ we define $Q = \Sigma^{-\frac{1}{2}}B \left[\Sigma^{-\frac{1}{2}}B\right]^\dagger$, and $W = I_T - Q$, where I_T is the $T \times T$ identity matrix. Note that Q and W are orthogonal projectors, and that $W\Sigma^{-\frac{1}{2}}B = 0$. It follows from standard arguments that every function in the range of $L_{\theta,x}$ can be written as:

$$[L_{\theta,x}g](y) = h\left(Q\Sigma^{-\frac{1}{2}}y\right) \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y-a)\right], \quad (11)$$

for some function $h : \mathbb{R}^T \rightarrow \mathbb{R}$. As an example, in the special case where $T = 2$, B is a vector of ones, and $\Sigma = \sigma^2 I_2$, $h\left(Q\Sigma^{-\frac{1}{2}}y\right)$ is a function of the individual mean $\bar{y} = (y_1 + y_2)/2$. If $\text{rank}(Q) < T$, the range of $L_{\theta,x}$ is thus strictly included in the space of T -variate functions.

In particular, if we find a function φ such that, for *any* function h :

$$\int_{\mathbb{R}^T} \varphi(y) h\left(Q\Sigma^{-\frac{1}{2}}y\right) \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y-a)\right] dy = 0, \quad (12)$$

then φ will satisfy (9). Finding moment restrictions on θ_0 thus amounts to solving the mathematical problem of constructing such a function φ .

As Q and W are orthogonal to each other, it is easy to see that if we define $\varphi(y) = \Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y-a)$ then φ is orthogonal to all functions in the range of $L_{\theta,x}$. This implies:

$$\mathbb{E}\left[\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y_i - a) \mid x_i\right] = 0. \quad (13)$$

Restrictions (13) remain valid when errors are non-normal, provided $\mathbb{E}(v_i|x_i, \alpha_i) = 0$. Under this assumption, Chamberlain (1992) shows that basing the estimation of θ_0 on the generalized within-group conditional moment restrictions (13) achieves the semiparametric information bound, using a suitable sample counterpart for the matrix Σ .¹¹

Example 1B (cont.) Similarly, in the censored regression model, any function in the range of $L_{\theta,x}$ satisfies (11), for some function h and for all y such that $y_t > c_t$ for all t . For

¹⁰Available at: <http://www.cemfi.es/~bonhomme/>

¹¹Note that in the version of model (2) that imposes normality our approach yields additional moment restrictions. See the supplementary appendix for a detailed discussion.

example, in the special case: $y_{it} = \max(x'_{it}\beta_0 + \alpha_i + v_{it}, 0)$, $T = 2$, with v_{it} i.i.d. $N(0, \sigma_0^2)$, then any element in the range of $L_{\theta,x}$ satisfies, for $y_1 > 0, y_2 > 0$:

$$[L_{\theta,x}g](y) = h(\bar{y}, x) \exp \left[-\frac{1}{4\sigma^2} (\Delta y - \Delta x' \beta)^2 \right], \quad (14)$$

where $\bar{y} = (y_1 + y_2)/2$, $\Delta y = y_2 - y_1$, $\Delta x = x_2 - x_1$, and where $\theta = (\beta, \sigma^2)$.

So, every function φ orthogonal to the functions given by (14), and with support strictly included in the positive orthant, will provide moment conditions on θ_0 . Consider for example a rectangle included in the positive orthant:

$$\left\{ (y_1, y_2), \quad (\bar{y}, \Delta y) \in [a, b] \times [c, d] \right\} \subset \left\{ (y_1, y_2), \quad y_1 > 0, y_2 > 0 \right\},$$

and the following function supported on that rectangle:

$$\varphi(y_1, y_2) = \varphi_2(\Delta y) \mathbf{1}\{\bar{y} \in [a, b]\} \mathbf{1}\{\Delta y \in [c, d]\}.$$

It is easy to see that φ will satisfy (9) if:

$$\int_c^d \varphi_2(\nu) \exp \left[-\frac{1}{4\sigma^2} (\nu - \Delta x' \beta)^2 \right] d\nu = 0. \quad (15)$$

In particular, (15) will be satisfied for $\varphi_2(\nu) = \text{sign}(\nu - \Delta x' \beta)$ and $\varphi_2(\nu) = \nu - \Delta x' \beta$ for example, provided c and d are taken symmetric around $\Delta x' \beta$. Taking the union of all such rectangles in the positive orthant, we obtain restrictions on β_0 that were first derived in Honoré (1992).¹² Those restrictions remain valid in the absence of normality, provided (v_{i1}, v_{i2}) are i.i.d. When assuming normality, our approach suggests additional restrictions which can be obtained by constructing other functions φ_2 such that (15) holds. This strategy will also deliver restrictions on σ_0^2 .¹³ In the supplementary appendix, we show that a similar approach can be used to derive restrictions on θ_0 in the general random coefficients models with censoring.

In Examples 1A and 1B, the range of $L_{\theta,x}$ is a strict subset of the space of T -variate functions, provided $\text{rank}(Q) < T$, hence in particular when $T > \dim \alpha$. This condition, which we refer to as *non-surjectivity*, ensures that the set of non-zero functions that are

¹²In the censored regression model, Honoré's restrictions are slightly different. This is because he uses observations that are partly censored: $(y_1 = 0, y_2 > 0)$ and $(y_1 > 0, y_2 = 0)$, while in the present discussion we focus only on fully uncensored observations.

¹³As an example, it can be shown that, when c and d are taken symmetric around $\Delta x' \beta$, $\varphi_2(\nu) = (\nu - \Delta x' \beta)^2 - 2\sigma^2$ satisfies (15).

orthogonal to the range of $L_{\theta,x}$ is not empty. This paper proposes a systematic way to construct functions in this set, thus providing moment restrictions on common parameters.

Example 2 (cont.) In static binary choice panel data models, our approach consists in finding a $2^T \times 1$ vector $\{\varphi(y, x, \theta), y \in \{0, 1\}^T\}$ such that, almost surely:

$$\sum_{y \in \{0,1\}^T} \varphi(y, x, \theta) \Pr(y_i = y | x, \alpha; \theta) = 0, \quad (16)$$

that is, denoting $F_t = F_t(x, \alpha, \theta)$ for conciseness:

$$\sum_{y \in \{0,1\}^T} \varphi(y, x, \theta) \prod_{t=1}^T F_t^{y_t} (1 - F_t)^{1-y_t} = 0. \quad (17)$$

It can be shown that finding a non-zero $\{\varphi(y, x, \theta)\}$ that satisfies (16) is equivalent to all 2^T products of distinct F 's being linearly dependent: $F_1^{k_1} \times \dots \times F_T^{k_T}$, $(k_1, \dots, k_T) \in \{0, 1\}^T$. F_t being a nonlinear function of individual effects, finding such a φ is often impossible. The reason is that the range of the mapping $L_{\theta,x}$ is likely to span the whole space of vectors in $\{0, 1\}^T$. An example is the static probit model, where $F_t = \Phi(x'_{it}\theta + \alpha_i)$, with Φ the standard normal cdf. This situation contrasts with Examples 1A and 1B, where a condition of non-surjectivity was guaranteed when $T > \dim \alpha$.

In contrast, when errors are logistic the situation is very different. In this case, $F_t = \Lambda(x'_{it}\theta + \alpha_i)$, where $\Lambda(u) = e^u / (1 + e^u)$ is the standard logistic cdf. We show in the supplementary appendix that (17) is equivalent to:

$$\sum_{y \in \{0,1\}^T} \mathbf{1} \left\{ \sum_{t=1}^T y_t = s \right\} \varphi(y, x, \theta) e^{\sum_{t=1}^T y_t x'_t \theta} = 0, \quad \text{for all } s \in \{0, 1, \dots, T\}. \quad (18)$$

This system of equations has non-trivial solutions as soon as $T \geq 2$. For example, if $T = 2$, (18) implies that: $\varphi_{00} = \varphi_{11} = 0$, and:

$$\varphi_{10} e^{x'_{i1}\theta} + \varphi_{01} e^{x'_{i2}\theta} = 0, \quad (19)$$

where with some abuse of notation we have denoted: $\varphi_{y_1 y_2} \equiv \varphi((y_1, y_2)', x, \theta)$. This yields the following conditional moment restriction on θ_0 :

$$\mathbb{E} \left(e^{[x_{i2} - x_{i1}]' \theta_0} y_{i1} [1 - y_{i2}] - [1 - y_{i1}] y_{i2} \mid x_i \right) = 0, \quad (20)$$

which point-identifies θ_0 provided $x_{i2} - x_{i1}$ is not identically zero.

Interestingly, Chamberlain (1987)'s optimal unconditional moment restrictions in (20) are:

$$\mathbb{E} \left[(x_{i2} - x_{i1}) \frac{1}{e^{[x_{i2} - x_{i1}]' \theta_0} + 1} \left(e^{[x_{i2} - x_{i1}]' \theta_0} y_{i1} [1 - y_{i2}] - [1 - y_{i1}] y_{i2} \right) \right] = 0. \quad (21)$$

This coincides exactly with the score equations of the conditional maximum likelihood estimator based on the sufficient statistic $y_{i1} + y_{i2}$.

This discussion closely relates to Chamberlain (2010)'s analysis of identification and information in static binary choice panel data models. Chamberlain shows that, in non-logistic binary choice models, the information bound for θ_0 is zero. The present discussion suggests that those models are surjective, implying that our approach will not yield informative restrictions on θ_0 .¹⁴

Example 3 (cont.) Lastly, in the nonlinear regression model (5) the orthogonal complement of the range of $L_{\theta, x}$ comprises the functions φ such that:

$$\int_{\mathbb{R}^T} \varphi(y) \exp \left[-\frac{1}{2} (y - m(\alpha))' \Sigma^{-1} (y - m(\alpha)) \right] dy = 0, \quad \text{for all } \alpha. \quad (22)$$

Unlike in Examples 1A and 1B, (22) does not seem to have simple analytical solutions in general. In this case, the projection approach of functional differencing will allow to numerically construct functions φ that satisfy (22). As a result, this will deliver conditional moment restrictions on θ_0 .

3 Projection: finite supports

In this section, we present our differencing approach in the special case where the distributions of the data and individual effects have known finite supports. This special case is useful for expositional purposes, as its treatment requires only elementary linear algebra. The next section will consider the general case where supports may be infinite-dimensional.

¹⁴This result is also related to Johnson (2004) who shows that, in discrete choice panel data models with compactly supported covariates, common parameters are unidentified unless equation (16) holds for some $\varphi \neq 0$ for at least some value of the covariates, and that when (16) does not hold for any value of x the information bound for θ_0 is zero. Buchinsky, Hahn and Kim (2008) build on Johnson's results to provide a procedure to test whether the information bound for θ_0 is zero.

3.1 Functional differencing

When y_i and α_i have known finite supports, the linear restrictions (1) simply map the probabilities of α_i to those of y_i , for a given value of x_i .¹⁵ Specifically, let N_y be the number of points of support of y_i , and let N_α be the number of points of support of α_i . Equation (1) can be equivalently written as:

$$f_{y|x} = L_{\theta_0, x} f_{\alpha|x}, \quad (23)$$

where $f_{y|x}$ is the $N_y \times 1$ vector of marginal probabilities of y_i (for a given value $x_i = x$), $f_{\alpha|x}$ is the $N_\alpha \times 1$ vector of marginal probabilities of α_i , and $L_{\theta, x}$ is the matrix of conditional probabilities of y_i given α_i (for given values of x and θ).

Denoting as $\underline{y}_1, \dots, \underline{y}_{N_y}$ and $\underline{\alpha}_1, \dots, \underline{\alpha}_{N_\alpha}$ the points of support of y_i and α_i , respectively, we thus have:

$$f_{y|x} = \begin{bmatrix} \Pr(y_i = \underline{y}_1 | x_i = x) \\ \vdots \\ \Pr(y_i = \underline{y}_{N_y} | x_i = x) \end{bmatrix}, \quad f_{\alpha|x} = \begin{bmatrix} \Pr(\alpha_i = \underline{\alpha}_1 | x_i = x) \\ \vdots \\ \Pr(\alpha_i = \underline{\alpha}_{N_\alpha} | x_i = x) \end{bmatrix},$$

and:

$$L_{\theta, x} = \begin{bmatrix} \Pr(y_i = \underline{y}_1 | x_i = x, \alpha_i = \underline{\alpha}_1; \theta) & \dots & \Pr(y_i = \underline{y}_1 | x_i = x, \alpha_i = \underline{\alpha}_{N_\alpha}; \theta) \\ \vdots & \vdots & \vdots \\ \Pr(y_i = \underline{y}_{N_y} | x_i = x, \alpha_i = \underline{\alpha}_1; \theta) & \dots & \Pr(y_i = \underline{y}_{N_y} | x_i = x, \alpha_i = \underline{\alpha}_{N_\alpha}; \theta) \end{bmatrix},$$

where, for clarity, we have omitted the reference to $(\theta_0, f_{\alpha|x})$ in $f_{y|x}$.

When supports are finite, the range of the matrix $L_{\theta, x}$ is the finite-dimensional vector space spanned by its columns. To construct vectors φ in \mathbb{R}^{N_y} that are orthogonal to the range of $L_{\theta, x}$ one can use the following ‘‘within’’ projection matrix:

$$W_{\theta, x} = I_{N_y} - L_{\theta, x} L_{\theta, x}^\dagger, \quad x - a.s. \quad (24)$$

where I_{N_y} denotes the $N_y \times N_y$ identity matrix, and $L_{\theta, x}^\dagger$ is the Moore-Penrose generalized inverse of $L_{\theta, x}$.

The $N_y \times N_y$ projection matrix $W_{\theta, x}$ satisfies our purpose, as it projects vectors of \mathbb{R}^{N_y} onto the orthogonal complement of the range of the matrix $L_{\theta, x}$. In particular, because $L_{\theta, x}^\dagger$

¹⁵The assumption that the points of support of α_i are known is restrictive. One possibility would be to treat the points of support $\underline{\alpha}_1, \dots, \underline{\alpha}_{N_\alpha}$ as *common* parameters to be estimated jointly with θ_0 , using the conditional moment restrictions from functional differencing.

is a generalized inverse, $W_{\theta,x}$ is such that:

$$W_{\theta,x}L_{\theta,x} = L_{\theta,x} - L_{\theta,x}L_{\theta,x}^\dagger L_{\theta,x} = 0, \quad x - a.s.$$

For every given vector $h \in \mathbb{R}^{N_y}$, the vector $W_{\theta,x}h \in \mathbb{R}^{N_y}$ is thus orthogonal to the columns of $L_{\theta,x}$.¹⁶ Moreover, *any* vector that is orthogonal to the columns of $L_{\theta,x}$ is of the form $W_{\theta,x}h$, for some h . So, if $[W_{\theta,x}h](\underline{y}_s)$ denotes the s th element of $W_{\theta,x}h$, where \underline{y}_s ($s = 1, \dots, N_y$) index the points of support of y_i , it follows that:

$$\mathbb{E} \left([W_{\theta_0,x_i}h](y_i) \middle| x_i \right) = 0, \quad \text{for all } h \in \mathbb{R}^{N_y}. \quad (25)$$

Using the canonical basis of \mathbb{R}^{N_y} as h -vectors in (25), this approach delivers N_y (possibly redundant) conditional moment restrictions on θ_0 . We will estimate θ_0 using generalized method-of-moments (GMM), based on a set of $R \geq \dim \theta$ unconditional moment restrictions. Interestingly, although the unknown probability vector $f_{\alpha|x}$ is conditional on covariates, θ_0 may be estimated using a standard finite-dimensional GMM approach.

To interpret our approach, note that the moment restrictions are obtained by left-multiplying (23) by the within projection matrix $W_{\theta_0,x}$, yielding $W_{\theta_0,x}f_{y|x} = W_{\theta_0,x}L_{\theta_0,x}f_{\alpha|x}$, and thus:

$$W_{\theta_0,x}f_{y|x} = 0, \quad x - a.s. \quad (26)$$

The functional differencing restrictions (26) are thus obtained by differencing out the probability distribution function of individual effects, yielding a set of restrictions on θ_0 alone. This is reminiscent of first-differencing and within-group approaches commonly used in linear panel data models.

As a second interpretation, notice that $W_{\theta,x}h = h - L_{\theta,x}L_{\theta,x}^\dagger h$ is the least-squares residual in the linear regression of a vector $h \in \mathbb{R}^{N_y}$ on the columns of the matrix $L_{\theta,x}$. As an example, in the special case where $L_{\theta,x}$ has full-column rank $W_{\theta,x}h$ is simply $h - L_{\theta,x} (L_{\theta,x}'L_{\theta,x})^{-1} L_{\theta,x}'h$. By construction, this residual is orthogonal to the columns of $L_{\theta,x}$. In particular, at the true value θ_0 , $W_{\theta_0,x}h$ is orthogonal to $L_{\theta_0,x}f_{\alpha|x} = f_{y|x}$. This means that the moment functions in (25) can be obtained as residuals in a linear regression. Bajari *et al.* (2009) use a related idea in a game-theoretic context.

¹⁶To see this, note that the projection matrix $W_{\theta,x}$ is symmetric, so $(W_{\theta,x}h)'L_{\theta,x} = h'W_{\theta,x}L_{\theta,x} = 0$, x -a.s.

3.2 Identification and information

Depending on the form of the panel data model, common parameters θ_0 may fail to be point-identified from the conditional moment restrictions (25). In particular this will happen if, for some $\theta \neq \theta_0$ in Θ , the matrix $L_{\theta,x}$ has full-row rank N_y almost surely in x . Indeed, in that case the range of $L_{\theta,x}$ is the full space \mathbb{R}^{N_y} , so $W_{\theta,x}$ is identically zero. As a consequence, (25) is satisfied at a θ -value different from the truth.¹⁷ Hence, a condition of *non-surjectivity* is necessary for θ_0 to be point-identified from the functional differencing restrictions.

Non-surjectivity holds in static binary choice models (Example 2) provided $N_\alpha < 2^T$. When α_i has more than 2^T points of support, however, the condition generally fails. When this happens, θ_0 is not point-identified from the functional differencing restrictions (25). Note that, under support conditions on covariates, θ_0 is point-identified in Example 2 (Manski, 1987), irrespective of N_α . This means that our approach may be uninformative about θ_0 in models where it is actually identified.

In addition, θ_0 may fail to be point-identified from (25) even though the non-surjectivity condition holds. Hence, non-surjectivity is not a sufficient condition for identification. To provide an example, let us consider a static logit model with two periods and a scalar time-invariant regressor $x_{i1} = x_{i2}$. In this model, as we argued above, the non-surjectivity condition is satisfied irrespective of N_α . However, the slope coefficient (θ_0 in $x'_{it}\theta_0$) is not identified from (25).

A general sufficient condition for point-identification is that, for every $\theta \neq \theta_0$ in Θ , $f_{y|x}$ does not belong to the range of $L_{\theta,x}$ with positive probability in x . In practice, this condition may be difficult to check and it may be easier to verify *local* identification using a rank condition on the following $N_y \times K$ Jacobian matrix (where $K = \dim \theta$):

$$G(x) = \left[-W_{\theta_0,x} \frac{\partial L_{\theta_0,x}}{\partial \theta_1} L_{\theta_0,x}^\dagger f_{y|x}, \dots, -W_{\theta_0,x} \frac{\partial L_{\theta_0,x}}{\partial \theta_K} L_{\theta_0,x}^\dagger f_{y|x} \right],$$

where $\frac{\partial L_{\theta_0,x_i}}{\partial \theta_k}$ is an $N_y \times N_\alpha$ matrix of derivatives with (s, n) th element $\frac{\partial \Pr(y_i = y_s | x_i, \alpha_i = \alpha_n; \theta_0)}{\partial \theta_k}$.

The following result is shown in Appendix A.

Proposition 1 *Let us assume that $L_{\theta,x}$ has constant rank in a neighborhood of θ_0 , and that $\theta \mapsto L_{\theta,x}$ is continuously differentiable at θ_0 , almost surely in x . If $G(x)c = 0$ almost surely in x implies $c = 0$, then θ_0 is locally point-identified from (25).*

¹⁷As an example, if $L_{\theta,x}$ is square and non-singular then the Moore-Penrose inverse coincides with the standard matrix inverse, and $W_{\theta,x} = I_{N_y} - L_{\theta,x}L_{\theta,x}^{-1} = 0$.

The result is derived under the assumption that $\text{rank}(L_{\theta,x})$ is locally constant. This type of condition is important to ensure that $W_{\theta,x}$ be continuous (and differentiable) around θ_0 .¹⁸ Rank constancy is intuitively necessary to ensure the continuity of a projection matrix, as variations in the rank of $L_{\theta,x}$ induce jumps in its number of non-zero eigenvalues. A sufficient condition for $L_{\theta,x}$ to have constant rank is that it has full-column rank N_α . The constant rank assumption is also satisfied in non-injective panel data models, such as static logit.¹⁹

To illustrate Proposition 1, consider the case where x_i has finite support \mathcal{X} . Then, a necessary condition for the result to hold is:

$$\sum_{x \in \mathcal{X}} (N_y - \text{rank}(L_{\theta_0,x})) \geq \dim \theta.$$

Note that this condition may fail to hold even though the model is non-surjective (i.e., even if $\text{rank}(L_{\theta_0,x}) < N_y$).

Information. Let \mathcal{S} denote the unit simplex in \mathbb{R}^{N_α} . The structure of the panel data model implies that there exists an $f_{\alpha|x} \in \mathcal{S}$ such that (23) holds. The following result shows that, although the functional differencing restrictions do not exploit the fact that $f_{\alpha|x}$ belongs to \mathcal{S} , they achieve the information bound for θ_0 . We prove the result for finitely supported regressors, in which case the model is parametric. A more general discussion of efficiency will be given in Section 4.

Theorem 1 *Let us assume that regressors x_i have finite support, that $f_{y|x} > 0$ and $f_{\alpha|x} > 0$, and that, for all (y, x, α) , $\theta \mapsto f_{y|x,\alpha}(y|x, \alpha; \theta)$ is continuously differentiable in a neighborhood of θ_0 . Then the information bound for θ_0 in the panel data model coincides with the GMM bound based on the conditional moment restrictions (25).*

In particular, Theorem 1 implies that the bound for θ_0 is zero when $L_{\theta_0,x}$ is surjective, as in this case (25) is not informative about common parameters. Although imposing that $f_{\alpha|x}$ is non-negative and sums to one may make a difference in finite sample, it does not improve on the bound.²⁰

¹⁸See e.g. Corollary 3.5 in Stewart (1977).

¹⁹For example, with $T = 2$ the null-space of $L'_{\theta_0,x}$ has dimension 1, so $\text{rank}(L_{\theta_0,x}) = 2^T - 1 = 3$, irrespective of (θ, x) .

²⁰To see intuitively why *equality* constraints on $f_{\alpha|x}$ do not affect the bound, note that summing the right and left-hand sides of (23) implies: $\sum_{n=1}^{N_\alpha} f_{\alpha|x}(\alpha_n|x) = 1$. The reason why *inequality* (non-negativity) constraints do not matter is that it is assumed that $f_{\alpha|x} > 0$ is interior.

In models with zero information bound, however, exploiting the constraints on $f_{\alpha|x}$ is often essential. Examples that fall into this category are set identified models with discrete outcomes, such as the dynamic probit model considered in Honoré and Tamer (2006). This situation may also happen when θ_0 is point-identified, but not estimable at a root- N rate, an example being a static probit model with an unbounded regressor (Chamberlain, 2010).

Let $L_{\theta_0,x}(\mathcal{S})$ denote the image by $L_{\theta_0,x}$ of the unit simplex. In order to incorporate the constraints on $f_{\alpha|x}$ in estimation, a natural approach would be to exploit the fact that, at true parameter values, $f_{y|x}$ coincides with its projection on $L_{\theta_0,x}(\mathcal{S})$. Using a *constrained* projection approach has intuitive appeal, especially in cases where the linear projection matrix $W_{\theta_0,x}$ is zero.

Minimizing the distance between data probabilities and their projection on the convex set $L_{\theta_0,x}(\mathcal{S})$ takes the form of a quadratic programming problem. Chernozhukov *et al.* (2009) have recently used this strategy in discrete choice panel data models. Note however that, in this minimum-distance approach, estimates of the conditional probabilities $f_{y|x}$ are needed, raising a curse of dimensionality as the dimension of x_i increases. In contrast, the conditional moment restrictions (25) obtained by functional differencing are immune to this problem. In the rest of this paper, we do *not* exploit the fact that $f_{\alpha|x}$ is a distribution function, although we think this is an important task for future work.

4 Projection: general case

When α_i or y_i have infinite support, $L_{\theta_0,x}$ becomes a linear integral operator. In this section we describe our framework in this general case, derive the moment restrictions, and discuss identification and efficiency.

4.1 Moment restrictions

We start with some notation. Let \mathcal{X} denote the support of x_i . Let also $\mathcal{A} \subset \mathbb{R}^q$ (respectively, $\mathcal{Y} \subset \mathbb{R}^T$) be a set that contains the support of $f_{\alpha|x}(\cdot|x)$ (resp. $f_{y|x}(\cdot|x)$) for all values $x \in \mathcal{X}$, where q is the dimension of the vector of individual effects and T is the number of time periods. We start the analysis by taking a value $x \in \mathcal{X}$.

Given two positive functions $\pi_\alpha > 0$ and $\pi_y > 0$, we define two spaces of square integrable

functions with domains \mathcal{A} and \mathcal{Y} , respectively, as:²¹

$$\begin{aligned}\mathcal{G}_\alpha &= \left\{ g : \mathcal{A} \rightarrow \mathbb{R}, \int_{\mathcal{A}} g(\alpha)^2 \pi_\alpha(\alpha) d\alpha < \infty \right\}, \\ \mathcal{G}_y &= \left\{ h : \mathcal{Y} \rightarrow \mathbb{R}, \int_{\mathcal{Y}} h(y)^2 \pi_y(y) dy < \infty \right\}.\end{aligned}$$

The functions π_α and π_y are specified by the researcher, and they will be important to derive projection-based moment restrictions on common parameters.

The spaces of functions \mathcal{G}_α and \mathcal{G}_y are *Hilbert spaces*, endowed with two scalar products that with some abuse of notation we denote similarly: $\langle g_1, g_2 \rangle = \int_{\mathcal{A}} g_1(\alpha) g_2(\alpha) \pi_\alpha(\alpha) d\alpha$, and $\langle h_1, h_2 \rangle = \int_{\mathcal{Y}} h_1(y) h_2(y) \pi_y(y) dy$, respectively. The associated norms are denoted as $\|g\| = \langle g, g \rangle^{\frac{1}{2}}$ and $\|h\| = \langle h, h \rangle^{\frac{1}{2}}$.

For a given value of common parameters $\theta \in \Theta$, we define $L_{\theta,x}$ as the integral operator that maps $g \in \mathcal{G}_\alpha$ to $L_{\theta,x}g \in \mathcal{G}_y$, where $L_{\theta,x}g$ is given by (8). The operator $L_{\theta,x}$ may be understood as an infinite-dimensional analog of a matrix. Indeed, when \mathcal{A} and \mathcal{Y} are finite and π_α and π_y are chosen as discrete uniform probability distributions, the operator $L_{\theta,x}$ is just the matrix of conditional probabilities introduced in the previous section.

Let us denote the *range* of the operator $L_{\theta,x}$ as:

$$\mathcal{R}(L_{\theta,x}) = \left\{ L_{\theta,x}g \in \mathcal{G}_y, \quad g \in \mathcal{G}_\alpha \right\},$$

and let $\overline{\mathcal{R}(L_{\theta,x})}$ denote its closure in \mathcal{G}_y (according to the Hilbert space topology). We define the “within” projection operator as:

$$W_{\theta,x}h = h - \text{Proj}_{\pi_y} \left[h \mid \overline{\mathcal{R}(L_{\theta,x})} \right], \quad \text{for all } h \in \mathcal{G}_y,$$

where the orthogonal projection of h onto $\overline{\mathcal{R}(L_{\theta,x})}$ satisfies:

$$\text{Proj}_{\pi_y} \left[h \mid \overline{\mathcal{R}(L_{\theta,x})} \right] = \underset{\tilde{h} \in \overline{\mathcal{R}(L_{\theta,x})}}{\text{argmin}} \int_{\mathcal{Y}} \left(\tilde{h}(y) - h(y) \right)^2 \pi_y(y) dy. \quad (27)$$

As $\overline{\mathcal{R}(L_{\theta,x})}$ is a closed linear subspace of \mathcal{G}_y , the projection operator $W_{\theta,x}$ is well-defined on \mathcal{G}_y , is continuous in its functional argument (i.e., $h \mapsto W_{\theta,x}h$ is continuous), and projects functions in \mathcal{G}_y onto the orthogonal complement of the range of $L_{\theta,x}$. Note that $W_{\theta,x}$ depends on π_y , although for conciseness we leave that dependence implicit. In the special case where the supports \mathcal{Y} and \mathcal{A} are finite and π_y is discrete uniform, $W_{\theta,x}$ coincides with the within projection matrix of Section 3.

The next result provides the key restrictions on θ_0 .

²¹Note that π_α and π_y may depend on x , although we omit the x subscript for conciseness.

Theorem 2 *Let us assume that $f_{y|x} \in \mathcal{G}_y$, x -a.s. Then the two following equivalent conditions are satisfied:*

$$W_{\theta_0,x} f_{y|x} = 0, \quad x - a.s. \quad \text{or, equivalently} \quad (28)$$

$$\mathbb{E} \left(\pi_y(y_i) [W_{\theta_0,x_i} h](y_i) \middle| x_i \right) = 0, \quad \text{for all } h \in \mathcal{G}_y. \quad (29)$$

Theorem 2 provides a set of restrictions on θ_0 . As in the finite-dimensional case described in Section 3, these restrictions are obtained using a functional differencing approach that differences out the distribution of individual effects.²² Note that the required condition on $f_{y|x}$ is automatically satisfied if $\pi_y f_{y|x}$ is almost surely bounded by a constant, as in this case:

$$\int_{\mathcal{Y}} f_{y|x}(y|x)^2 \pi_y(y) dy \leq \left(\sup_{\mathcal{Y}} \pi_y(y) f_{y|x}(y|x) \right) \underbrace{\int_{\mathcal{Y}} f_{y|x}(y|x) dy}_{=1} < \infty, \quad x - a.s.$$

The functional differencing restrictions are equivalently written as a set of conditional moment restrictions. Importantly, the distribution function $f_{y|x}$ enters (29) only through the expectation. We will base estimation on a finite set of unconditional moment restrictions implied by (29), and study the properties of the GMM estimator in Section 5.

4.2 Non-surjectivity

As in the finite support case, the conditional moment restrictions given by (29) may fail to point-identify θ_0 . A necessary condition for point-identification of θ_0 based on (29) is that $L_{\theta,x}$ be non-surjective in the following sense.

Definition 1 (*non-surjectivity*) *The operator $L_{\theta,x}$ is non-surjective if, for all $\theta \neq \theta_0$ in Θ , $\overline{\mathcal{R}(L_{\theta,x})} \neq \mathcal{G}_y$ with positive probability in x .*

The non-surjectivity condition is related to the size of the support of the dependent variables. For example, non-surjectivity is generally not satisfied in discrete choice models when the support of individual effects is unrestricted, the logit model being an important exception. By comparison, non-surjectivity holds in the random coefficients and censored random coefficients models with normal errors (Examples 1A and 1B), provided $T > \dim \alpha$.²³

²²Note that, although the present paper assumes that θ is finite-dimensional, the restrictions of Theorem 2 are still valid if θ is infinite-dimensional.

²³When $T = \dim \alpha$ and B is non-singular, $Q = I_T$ and $W = 0$ in equation (12), and the non-surjectivity condition is *not* satisfied in Example 1A.

In contrast with the finite support case, however, non-surjectivity may fail to hold even though the support of outcomes is richer than the support of individual effects. Indeed, it is known that there exist surjective mappings from \mathcal{G}_α onto \mathcal{G}_y , even when $T > \dim \alpha$.²⁴ When dimensions are infinite, non-surjectivity imposes a condition on the structure of the nonlinear panel data model, in addition to a condition on the respective supports of outcomes and individual effects.²⁵

In the supplementary appendix to this paper, we study non-surjectivity in the nonlinear regression model (5) of Example 3 with independent additive errors (possibly non-normal). There we show that $T > \dim \alpha$ is *not* sufficient for non-surjectivity to hold. To guarantee non-surjectivity, a condition must be imposed on the image of the regression function in \mathbb{R}^T , which rules out space-filling mappings such as Peano curves (surjective mappings from \mathbb{R} onto \mathbb{R}^2).²⁶

Lastly note that, as in the finite support case, non-surjectivity is not sufficient for θ_0 to be identified from the functional differencing restrictions in general. In analogy with simultaneous equations models, non-surjectivity may be understood as an *order* condition for identification.²⁷ Sufficient conditions for *local* point-identification similar to Proposition 1 may be obtained. As stating rank conditions requires additional notation and assumptions, we will discuss them in Section 5, when analyzing the properties of GMM estimators based on a finite number of unconditional moment restrictions.

4.3 Information bound

It is interesting to relate the functional differencing approach to the efficient score and information bound of the panel data model. To proceed, let:

$$f_{y|x}(y|x; \theta_0, \eta_0) = \int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta_0) f_{\alpha|x}(\alpha|x; \eta_0) d\alpha$$

be a parametric submodel, where the parametric model for the individual effects depends on a scalar parameter η_0 .

²⁴This is because the Hilbert spaces \mathcal{G}_α and \mathcal{G}_y admit countable orthogonal bases. See for example Goldberg and Kruse (1962).

²⁵This discussion is reminiscent of the problem of verifying injectivity (or completeness) conditions in instrumental variables models (Newey and Powell, 2003). Showing non-surjectivity is actually equivalent to verifying the *non*-injectivity of the adjoint (or transpose) of the operator $L_{\theta,x}$.

²⁶In addition, the discretization strategy outlined in Section 6 may be used to provide numerical evidence on non-surjectivity. See the supplementary appendix for an illustration.

²⁷Note that although non-surjectivity is necessary for θ_0 to be point-identified from the restrictions (29), θ_0 may be point-identified in the panel data model even though $L_{\theta,x}$ is surjective. An example is a binary choice model with an unbounded covariate, which we already mentioned in Section 3.

We will assume that $\inf_{\alpha \in \mathcal{A}} f_{\alpha|x}(\alpha|x; \eta_0) > 0$, and work under regularity conditions that ensure that second-order moments of scores taken at true values are finite. Statements of the required regularity conditions (mean-square differentiability) may be found in the literature on semiparametric information bounds, such as Bickel *et al.* (1993), and more specifically Chamberlain (2010) and Hahn (1997, 2001) in a panel data context.

Let $L^2(f_{y|x})$ denote the space of square-integrable functions with respect to $f_{y|x}$. The *nonparametric tangent space* is defined as the $L^2(f_{y|x})$ -closure of the linear span of the η -scores across all parametric submodels. The efficient score for the k -th component of θ_0 (where $k \in \{1, \dots, \dim \theta\}$) is then constructed as a residual in the population projection of the θ_k -score on the nonparametric tangent space.

To derive the efficient score, let \mathcal{R}_y denote the set of zero-mean functions in $\mathcal{R}(L_{\theta_0,x})$. We show in Appendix B that the nonparametric tangent space of the model is the $L^2(f_{y|x})$ -closure of $\frac{1}{f_{y|x}} \cdot \mathcal{R}_y$. In particular, the zero-mean condition ensures that the tangent space is a set of score functions. The characterization of the nonparametric tangent space in terms of the range of the operator $L_{\theta_0,x}$ is important, as it will provide the link with functional differencing.

It follows from the characterization of the tangent space that the efficient score corresponds to the following $\dim \theta$ efficient moment restrictions on θ_0 :

$$\mathbb{E} \left[\frac{1}{f_{y|x}(y_i|x_i)} \left(\frac{\partial}{\partial \theta_k} \Big|_{\theta_0} [L_{\theta_0,x} f_{\alpha|x}(\cdot|x_i)](y_i) - h^{(k)}(y_i, x_i) \right) \right] = 0, \quad k = 1, \dots, \dim \theta, \quad (30)$$

where:

$$h^{(k)}(\cdot, x) = \operatorname{argmin}_{h \in \overline{\mathcal{R}_y}} \int_{\mathcal{Y}} \left(\frac{\partial}{\partial \theta_k} \Big|_{\theta_0} [L_{\theta_0,x} f_{\alpha|x}(\cdot|x)](y) - h(y) \right)^2 \frac{1}{f_{y|x}(y|x)} dy, \quad x - a.s. \quad (31)$$

To interpret the efficient moment restrictions in the light of the functional differencing approach, it is convenient to consider the following choice for the weight functions:

$$\pi_\alpha = \frac{1}{f_{\alpha|x}}, \quad \pi_y = \frac{1}{f_{y|x}}. \quad (32)$$

Note that, for this specific choice of weight functions, $h^{(k)}$ given by (31) may be interpreted as the orthogonal projection of $\frac{\partial L_{\theta_0,x} f_{\alpha|x}}{\partial \theta_k}$ on $\overline{\mathcal{R}_y}$, the closure of \mathcal{R}_y in \mathcal{G}_y . Moreover we show in the appendix that, for similar reasons as in the finite support case, the zero-mean condition on the elements of the tangent space is irrelevant for the purpose of computing the efficient score. As a result, $h^{(k)}$ is also the orthogonal projection of $\frac{\partial L_{\theta_0,x} f_{\alpha|x}}{\partial \theta_k}$ on $\overline{\mathcal{R}(L_{\theta_0,x})}$.

This key observation implies that the efficient moment restrictions (30) coincide with the following $\dim \theta$ restrictions from functional differencing:

$$\mathbb{E} \left(\pi_y(y_i) \left[W_{\theta_0, x_i} \frac{\partial L_{\theta_0, x_i} f_{\alpha|x}}{\partial \theta_k} \right] (y_i) \right) = 0, \quad k = 1, \dots, \dim \theta, \quad (33)$$

where π_y is given by (32).

This discussion thus shows that, for the specific choice of Hilbert space norms implied by (32), the efficient set of moment restrictions on θ_0 may be achieved using an equivalent set of restrictions from functional differencing. When supports \mathcal{A} and \mathcal{Y} are finite, all norms on \mathcal{G}_α and \mathcal{G}_y are equivalent, so the functional differencing approach is efficient for *any* choice of weight functions. This result is consistent with the finite support case that we considered in Theorem 1.

When supports are infinite, however, the equivalence of norms no longer holds and the choice of weight functions becomes essential. As a consequence, taking the efficient moment restrictions to the data is challenging.²⁸ Although efficient estimation of θ_0 seems a difficult task in general, in the next section we will use the above theoretical arguments as a guide to choose moment functions in practice.

5 Estimation

In this section we study the properties of a GMM estimator of θ_0 based on a finite set of unconditional moment restrictions obtained from (29).

5.1 Method-of-moments

Let $\{y_i, x_i\}_{i=1, \dots, N}$ be an i.i.d. sample. Motivated by the conditional moment restrictions (29) of Theorem 2, we propose to estimate θ_0 by minimizing the GMM criterion:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \quad \widehat{\mathbb{E}} [\varphi(y_i, x_i, \theta)'] \Upsilon \widehat{\mathbb{E}} [\varphi(y_i, x_i, \theta)], \quad (34)$$

where the moment functions are given by:

$$\varphi(y_i, x_i, \theta) = \begin{bmatrix} \pi_y(y_i) [W_{\theta, x_i} h_1](y_i) \zeta_{s_1}(x_i) \\ \vdots \\ \pi_y(y_i) [W_{\theta, x_i} h_R](y_i) \zeta_{s_R}(x_i) \end{bmatrix}. \quad (35)$$

²⁸A first difficulty, common with other efficient estimation problems, is that $f_{\alpha|x}$ and $f_{y|x}$ are unknown and must be replaced by empirical counterparts. A second difficulty is that, when π_α and π_y are chosen according to (32), $L_{\theta, x}$ may fail to satisfy the compactness assumption 1 *ii*) below.

In (34)-(35), $\widehat{\mathbb{E}}(z_i) = \frac{1}{N} \sum_{i=1}^N z_i$ denotes an empirical mean, h_1, \dots, h_R are elements of \mathcal{G}_y , ζ_1, \dots, ζ_M are functions of covariates, s_1, \dots, s_R are indexes in $\{1, \dots, M\}$, and $\Upsilon = \left[(v_{r_1, r_2})_{r_1, r_2} \right]$ is a symmetric $R \times R$ weighting matrix.

Under regularity conditions given below (which include point-identification of θ_0), $\widehat{\theta}$ will be root- N consistent and asymptotically normal. The main reason for this result is the boundedness of the within projection operator. Importantly, $\widehat{\theta}$ is consistent irrespective of the form of the distribution of individual effects.²⁹

Turning to the choice of functions h_r and ζ_{s_r} in (35), one approach is to choose a finite family h_r , $r = 1, \dots, R$, that covers (in some sense) \mathcal{G}_y . A possibility is to take orthogonal polynomials on \mathbb{R}^T (e.g., section 6.12 in Judd, 1998). As a closely related option, one may choose $\{h_r\}$ as a “flexible” family of densities, such as normal mixtures. In the simulation experiments reported in Section 6 we have set $h_r(y) = \phi(y - \mu_r)$, with ϕ the pdf of the normal distribution with zero mean and covariance matrix I_T , and μ_1, \dots, μ_R elements of \mathbb{R}^T .

In the presence of covariates, one could let the coefficients of the orthogonal polynomials—or of the chosen “flexible” family of densities—depend on x_i in some way, e.g. letting μ_r in $\phi(y - \mu_r)$ depend linearly on x_i . In addition, one may also want to choose the functions ζ_r and the weighting matrix Υ based on efficiency considerations.

A second approach to the choice of moment functions builds on the efficiency discussion of Section 4.3. Indeed, the form of the efficient moment restrictions (33) suggests to base estimation on the following *feasible* set of $\dim \theta$ restrictions:

$$\mathbb{E} \left(\pi_y(y_i) \left[W_{\theta_0, x_i} \frac{\partial L_{\tilde{\theta}, x_i} \tilde{g}_{\alpha, x_i}}{\partial \theta_k} \right] (y_i) \right) = 0, \quad k = 1, \dots, \dim \theta, \quad (36)$$

for given $(\tilde{\theta}, \tilde{g}_{\alpha, x}) \in \Theta \times \mathcal{G}_\alpha$.

Note that (36) is a valid set of moment restrictions on θ_0 for *any* choice of $\tilde{\theta}$, $\tilde{g}_{\alpha, x}$, π_α , and π_y . In addition, it follows from the discussion in Section 4 that the choice $(\tilde{\theta}, \tilde{g}_{\alpha, x}) = (\theta_0, f_{\alpha|x})$ achieves asymptotic efficiency, provided the weight functions are taken according to (32).³⁰ In the simulation exercise of Section 6, we will compare the performance of the two above approaches to choose the moment functions.

²⁹In the supplementary appendix we use this insight to construct a nonlinear analog of the Hausman (1978) specification test for parametric random-effects models.

³⁰In practice, a parametric random-effects model (based on a parametric model for $f_{\alpha|x}$) may be fitted to the data to choose $\tilde{\theta}$ and $\tilde{g}_{\alpha, x}$, and one may take $\pi_\alpha = 1/\tilde{g}_{\alpha, x}$, and $\pi_y = 1/L_{\tilde{\theta}, x} \tilde{g}_{\alpha, x}$. Although the resulting GMM estimator of θ_0 will not be efficient in general if the random-effects model is misspecified, it will be consistent under the conditions of Theorem 3 below.

5.2 Compactness

To derive the asymptotic properties of the GMM estimator $\widehat{\theta}$, we will work with the following regularity conditions (as in Carrasco and Florens, 2009).

Assumption 1 *The two following statements hold true, almost surely in x :*

i)

$$\int_{\mathcal{A}} f_{\alpha|x}(\alpha|x)^2 \pi_{\alpha}(\alpha) d\alpha < \infty,$$

ii)

$$\int_{\mathcal{Y}} \int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta)^2 \frac{\pi_y(y)}{\pi_{\alpha}(\alpha)} dy d\alpha < \infty, \quad \text{for all } \theta \in \Theta.$$

Part *i)* in Assumption 1 restricts the distribution of individual effects to be square integrable with respect to π_{α} . Note that $\int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta)^2 / \pi_{\alpha}(\alpha) d\alpha$ being bounded by a constant independent of y , and π_y being integrable, are sufficient conditions for part *ii)*. In particular, it can be shown that *ii)* holds in Example 1A when $\mathcal{A} = \mathbb{R}^q$, $\pi_{\alpha} = 1$, and π_y is integrable, while it does not hold in that example when $\pi_y = 1$. Moreover, *ii)* will automatically hold when \mathcal{A} is bounded in \mathbb{R}^q , provided $\pi_y(\cdot) f_{y|x,\alpha}(\cdot|x, \cdot; \theta)$ and $1/\pi_{\alpha}(\cdot)$ are bounded on $\mathcal{Y} \times \mathcal{A}$ and \mathcal{A} , respectively.

Part *ii)* in Assumption 1 ensures that $L_{\theta,x}g \in \mathcal{G}_y$ for all functions $g \in \mathcal{G}_{\alpha}$, and that the operator $L_{\theta,x} : \mathcal{G}_{\alpha} \rightarrow \mathcal{G}_y$ is Hilbert-Schmidt, hence compact. Compact operators are convenient to work with, as they can be approximated by finite-dimensional sums (see, e.g., Carrasco *et al.*, 2008).

The spectrum of the compact operator $L_{\theta,x}$ consists of a finite or countable set of singular values, which (when infinite) accumulates to zero.³¹ It will be convenient to work with the *singular value decomposition* of $L_{\theta,x}$:

$$L_{\theta,x}g = \sum_j \phi_j \lambda_j \langle \psi_j, g \rangle, \quad \text{for all } g \in \mathcal{G}_{\alpha}, \quad (37)$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \dots > 0$ is a sequence of positive real numbers, ψ_1, ψ_2, \dots is an orthonormal sequence in \mathcal{G}_{α} , and ϕ_1, ϕ_2, \dots is an orthonormal sequence in \mathcal{G}_y . The sum in (37) ranges from 1 to the (possibly infinite) rank of $L_{\theta,x}$. Note that, when $L_{\theta,x}$ is singular, $\{\psi_j\}$ and $\{\phi_j\}$ (corresponding to non-zero singular values) do not generally span \mathcal{G}_{α} and \mathcal{G}_y , respectively.

³¹We refer to Kress (1999) and Engl, Hanke, and Neubauer (2000) for results on compact linear operators in Hilbert spaces.

Note also that λ_j , ψ_j , and ϕ_j , $j = 1, 2, \dots$, depend on θ and x , although here the subscripts are omitted for clarity.

In this representation, the within projection operator can be written as:

$$W_{\theta,x}h = h - \sum_j \phi_j \langle \phi_j, h \rangle, \text{ for all } h \in \mathcal{G}_y. \quad (38)$$

Note that the singular values of the projector $W_{\theta,x}$ are zeros and ones. Moreover, note that $W_{\theta,x}$ does not depend on the distribution of the data. Although the singular functions ϕ_j are generally not available in closed form, they can be approximated using a simple discretization strategy that we will describe in Section 6.

It is interesting to contrast the expression in (38) with that of the Moore-Penrose inverse of $L_{\theta,x}$, which is defined on a domain $\mathcal{D} \subset \mathcal{G}_y$ by:³²

$$L_{\theta,x}^\dagger h = \sum_j \psi_j \frac{1}{\lambda_j} \langle \phi_j, h \rangle, \text{ for all } h \in \mathcal{D}. \quad (39)$$

In the infinite-dimensional case, the Moore-Penrose inverse $L_{\theta,x}^\dagger$ is not bounded in general. The reason is that, when the range of $L_{\theta,x}$ is not closed in \mathcal{G}_y , the singular values λ_j of the compact operator $L_{\theta,x}$ tend to zero as j tends to infinity (e.g., Engl *et al.*, 2000, p. 37). Hence, $h \mapsto L_{\theta,x}^\dagger h$ is not continuous, and $L_{\theta,x}^\dagger h$ is very sensitive to any noise in h possibly arising in estimation, reflecting an *ill-posedness* problem. In contrast, the within projection operator is always bounded, and thus continuous in its functional argument.

A finite-dimensional intuition for this result is as follows. Following the least-squares interpretation of Section 3, $L_{\theta,x}^\dagger h$ and $W_{\theta,x}h$ may be understood as the least-squares coefficients and residuals, respectively, in the linear regression of h on the columns of $L_{\theta,x}$. Now, when y_i and α_i have large supports, the columns of $L_{\theta,x}$ tend to be close to collinear. This will typically affect the precision of the coefficient estimates. However, the fitted values and predicted residuals will not be sensitive to the multicollinearity problem, as accurate (in-sample) prediction does not require to precisely estimate the contributions of the various regressors separately.

5.3 Asymptotic properties

To derive the asymptotic distribution of $\hat{\theta}$ we will need some additional notation. We denote the *norm* of a bounded operator as $\|L\| = \max_{\|h\| \leq 1} \|Lh\|$. Let \mathcal{G}_1 and \mathcal{G}_2 be two Hilbert

³²The domain \mathcal{D} is the linear span of $\mathcal{R}(L_{\theta,x})$ and its orthogonal complement in \mathcal{G}_y . It is thus a strict subspace of \mathcal{G}_y , unless $L_{\theta,x}$ has closed range.

spaces, and let $L : \mathcal{G}_1 \rightarrow \mathcal{G}_2$ be a bounded linear operator. We denote the *adjoint* of L as L^* , which is the unique linear operator that maps \mathcal{G}_2 onto \mathcal{G}_1 such that $\langle Lg, h \rangle = \langle g, L^*h \rangle$ for all $(g, h) \in \mathcal{G}_1 \times \mathcal{G}_2$. The adjoint operator may be interpreted as an infinite-dimensional analog of the matrix transpose.

5.3.1 Consistency

We make the following assumptions that ensure the consistency of $\widehat{\theta}$ as N tends to infinity.³³ For clarity, we now indicate with a subscript that $\lambda_{j,\theta,x}$, $\psi_{j,\theta,x}$ and $\phi_{j,\theta,x}$ depend on (θ, x) .

Assumption 2

i) Θ is compact.

ii) $\mathbb{E}(\pi_y(y_i) [W_{\theta,x_i} h_r](y_i) \zeta_{s_r}(x_i)) = 0$, $r = 1, \dots, R$, has a unique solution θ_0 that is an interior point of Θ .

iii) The function $\theta \mapsto f_{y|x,\alpha}(y|x, \alpha; \theta)$ is continuous on Θ , almost surely in y, x, α .

iv) Almost surely in x :

$$\sup_{\theta \in \Theta} \int_{\mathcal{Y}} \int_{\mathcal{A}} f_{y|x,\alpha}(y|x, \alpha; \theta)^2 \frac{\pi_y(y)}{\pi_\alpha(\alpha)} dy d\alpha < \infty.$$

Moreover, for every $r = 1, \dots, R$:

v) For every j :

$$\mathbb{E} \left[\left(\frac{1}{\inf_{\theta \in \Theta} \lambda_{j,\theta,x_i}^2} \right) \|f_{y|x}\| \|h_r\| |\zeta_{s_r}(x_i)| \right] < \infty.$$

vi) Almost surely in x :

$$\sup_{\theta \in \Theta} \left(\sum_{j>J} \langle \phi_{j,\theta,x}, f_{y|x} \rangle^2 \right) \xrightarrow{J \rightarrow \infty} 0.$$

vii)

$$\mathbb{E} \left[\sup_{y \in \mathcal{Y}} (\pi_y(y) f_{y|x}(y|x_i)) \|h_r\|^2 \zeta_{s_r}(x_i)^2 \right] < \infty.$$

viii)

$$\mathbb{E} \left[\|f_{y|x}\|^2 \|h_r\|^2 \zeta_{s_r}(x_i)^2 \right] < \infty.$$

³³Note that the weighting matrix Υ is assumed known. It can be replaced by a consistent estimate, with no change in the proof.

The compactness assumption *i*) is standard. Condition *ii*) requires θ_0 to be point-identified from the moment restrictions. In particular, as argued above, this condition fails when the non-surjectivity condition does not hold. Note that θ_0 being an interior point of Θ is not needed for consistency, though it will be used to show asymptotic normality. Condition *iii*) imposes that the conditional distribution of the data given α_i varies continuously with θ . Together with the uniform boundedness condition *iv*), this implies that the mapping $\theta \mapsto L_{\theta,x}$ is continuous on Θ with respect to the operator norm, almost surely in x .

Conditions *v*) and *vi*) guarantee that the population objective function is continuous in θ . Condition *v*) requires that $\lambda_{j,\theta,x}$ be bounded from below. This requires $\text{rank}(L_{\theta,x})$, when finite, to be constant on Θ . When the rank of $L_{\theta,x}$ is infinite, it will always be the case that $\inf_{\theta \in \Theta} \lambda_{j,\theta,x} > 0$, a.s. in x .³⁴ Condition *vi*) requires that $\sum_{j>J} \langle \phi_{j,\theta,x}, f_{y|x} \rangle^2$ tends to zero as J tends to infinity, uniformly on Θ . Note that the convergence to zero at each θ value is ensured by the fact that $f_{y|x} \in \mathcal{G}_y$. Condition *vi*) imposes the stronger requirement that the convergence be uniform, thus restricting the behavior of Fourier coefficients $\langle \phi_{j,\theta,x}, f_{y|x} \rangle$ across θ parameters. For this reason, we refer to Condition *vi*) as *uniform Fourier convergence*.

Note that uniform Fourier convergence holds trivially when $L_{\theta,x}$ has finite rank. When the rank is infinite, the rate of convergence to zero of Fourier coefficients is allowed to be arbitrarily slow. This shows that Condition *vi*) does not restrict the distribution of the data $f_{y|x}$ to belong to a certain smoothness class, unlike the *source* conditions often considered in the literature on ill-posed inverse problems. Condition *vi*) seems new in the literature. In the supplementary appendix to this paper, we analytically verify uniform Fourier convergence in Chamberlain (1992)'s random coefficients model (Example 1A) with known error variance. In addition, we provide numerical evidence supporting uniform Fourier convergence in the two simple models that we use as illustrations in Section 6.

Condition *vii*) is useful to show the uniform convergence of the sample moment functions to the population ones. This condition is stronger than actually needed for consistency. However, it guarantees that the following variance-covariance matrix is well-defined:

$$\Sigma(\theta) = \mathbb{E} [\varphi(y_i, x_i, \theta) \varphi(y_i, x_i, \theta)'] . \quad (40)$$

This property will be useful to derive the asymptotic distribution of $\hat{\theta}$.³⁵ This implies that there is no need to regularize the estimates of the moment functions. Finally, *viii*) is a

³⁴This is because the function $\theta \mapsto \lambda_{j,\theta,x}$ is continuous on Θ , x -a.s. See Theorem 15.17 in Kress (1999).

³⁵In particular, *vii*) requires that $\pi_y f_{y|x}$ be bounded on the support \mathcal{Y} , x -a.s. See Carrasco and Florens (2009) for a related assumption.

moment existence condition.

We then can state the following consistency result, which is proved in Appendix C.

Theorem 3 *Let Assumptions 1 and 2 hold. Then $\widehat{\theta} \xrightarrow{p} \theta_0$.*

5.3.2 Asymptotic normality

We now state assumptions that ensure that $\widehat{\theta}$ is a root- N consistent, asymptotically normal estimator of θ_0 .

Assumption 3 *There exists a neighborhood \mathcal{V} of θ_0 such that:*

i) The function $\theta \mapsto f_{y|x,\alpha}(y|x, \alpha; \theta)$ is continuously differentiable on \mathcal{V} , almost surely in y, x, α .

ii) Almost surely in x and for $(k, \ell) \in \{1, \dots, \dim \theta\}^2$:

$$\sup_{\theta \in \mathcal{V}} \int_{\mathcal{Y}} \int_{\mathcal{A}} \left| \frac{\partial f_{y|x,\alpha}(y|x, \alpha; \theta)}{\partial \theta_k} \frac{\partial f_{y|x,\alpha}(y|x, \alpha; \theta)}{\partial \theta_\ell} \right| \frac{\pi_y(y)}{\pi_\alpha(\alpha)} dy d\alpha < \infty.$$

For every $r = 1, \dots, R$:

iii) Almost surely in x :

$$\sup_{\theta \in \mathcal{V}} \left(\sum_{j>J} \langle \phi_{j,\theta,x}, h_r \rangle^2 \right) \xrightarrow{J \rightarrow \infty} 0.$$

iv)

$$\mathbb{E} \left[\left(\sup_{\theta \in \mathcal{V}} \left\| \frac{\partial L_{\theta, x_i}}{\partial \theta_k} \right\| \right) \|h_r\| \left\| L_{\theta_0, x_i}^\dagger f_{y|x} \right\| |\zeta_{s_r}(x_i)| \right] < \infty, \quad k = 1, \dots, \dim \theta.$$

v) The $R \times \dim \theta$ matrix:

$$G = \left[\left(-\mathbb{E} \left(\left\langle \frac{\partial L_{\theta_0, x_i}^*}{\partial \theta_k} W_{\theta_0, x_i} h_r, L_{\theta_0, x_i}^\dagger f_{y|x} \right\rangle \zeta_{s_r}(x_i) \right) \right)_{r,k} \right]$$

is such that $G' \Upsilon G$ is nonsingular.

vi) As N tends to infinity:

$$\sqrt{N} \widehat{\mathbb{E}} [\varphi(y_i, x_i, \theta_0)] \xrightarrow{d} N[0, \Sigma(\theta_0)].$$

Moreover, for all $\theta \in \mathcal{V}$:

$$\sqrt{N} \left(\widehat{\mathbb{E}} [\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0)] - \mathbb{E} [\varphi(y_i, x_i, \theta)] \right) \xrightarrow{d} N[0, \Sigma(\theta, \theta_0)],$$

where $\Sigma(\theta, \theta_0) = \text{Var} [\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0)]$.

Conditions *i)* and *ii)* impose regularity restrictions on the conditional density $f_{y|x,\alpha}$ as a function of common parameters. In particular, they allow us to define a bounded integral operator $\frac{\partial L_{\theta,x}}{\partial \theta_k} : \mathcal{G}_\alpha \rightarrow \mathcal{G}_y$ (for $k = 1, \dots, \dim \theta$) as:

$$\left[\frac{\partial L_{\theta,x}}{\partial \theta_k} g \right] (y) = \int_{\mathcal{A}} \frac{\partial f_{y|x,\alpha}(y|x, \alpha; \theta)}{\partial \theta_k} g(\alpha) d\alpha, \quad \text{for all } g \in \mathcal{G}_\alpha.$$

Condition *iii)* is similar in spirit to Condition *v)* of Assumption 2. Indeed, as $h \in \mathcal{G}_y$ the partial sums of squared Fourier coefficients converge to zero at each θ . Condition *iii)* requires this convergence to be uniform, here in a local neighborhood around θ_0 . Together with $\lambda_{j,\theta,x}$ being bounded from below, this guarantees that the mapping $\theta \mapsto W_{\theta,x} h_r$ is continuous on \mathcal{V} , almost surely in x .

Condition *iv)* requires some moments to be finite. This will ensure the differentiability of the population objective function at θ_0 . Then, Condition *v)* is a familiar condition on the non-singularity of the Jacobian matrix. G having full-column rank can be understood as a *rank condition* for local point-identification of θ_0 .

The two parts in Condition *vi)* will be satisfied if one can apply a central limit theorem to the empirical moment functions. As, by Assumption 2, $\Sigma(\theta)$ is finite for all $\theta \in \mathcal{V}$, and given that data are i.i.d, the conditions of application of the Lindeberg-Levy central limit theorem are satisfied if $\Sigma(\theta) \neq 0$. In particular, this requires the model to be non-surjective.

We now can state the next result, which proves the root- N consistency and asymptotic normality of $\hat{\theta}$.

Theorem 4 *Let the assumptions of Theorem 3 be satisfied and let Assumption 3 hold. Then:*

$$\sqrt{N} (\hat{\theta} - \theta_0) \xrightarrow{d} N \left[0, (G' \Upsilon G)^{-1} G' \Upsilon \Sigma(\theta_0) \Upsilon G (G' \Upsilon G)^{-1} \right]. \quad (41)$$

Remark 1. The proof of Theorem 4 does not require the empirical moment functions $\theta \mapsto \hat{\mathbb{E}}[\varphi(y_i, x_i, \theta)]$ to be continuous. In practice, as outlined in Section 6 below, we will work with a discretized version of the operator $W_{\theta,x}$ associated with continuous moment functions.

Remark 2. In order to estimate the asymptotic variance of $\hat{\theta}$, we need to compute consistent estimates of Σ and G . The outer product Σ is readily estimated as:

$$\hat{\Sigma} = \hat{\mathbb{E}} \left[\varphi(y_i, x_i, \hat{\theta}) \varphi(y_i, x_i, \hat{\theta})' \right].$$

In contrast, the Jacobian matrix G involves the unbounded Moore-Penrose inverse $L_{\theta_0, x_i}^\dagger$. Interestingly, the matrix G can be interpreted as an average marginal effect. As a consequence, the problem of estimating G , and hence the *variance* of the common parameter estimates $\widehat{\theta}$, may be *ill-posed* (as in Bonhomme, 2011). A simple truncated estimate that relies on the singular value decomposition (39) is:

$$\widehat{G} = \left[\left(-\widehat{\mathbb{E}} \left(\sum_{j=1}^J \pi_y(y_i) \phi_{j, \widehat{\theta}, x_i}(y_i) \frac{1}{\lambda_{j, \widehat{\theta}, x_i}} \left\langle \frac{\partial L_{\widehat{\theta}, x_i}^*}{\partial \theta_k} W_{\widehat{\theta}, x_i} h_r, \psi_{j, \widehat{\theta}, x_i} \right\rangle \zeta_{s_r}(x_i) \right) \right) \right]_{r,k}, \quad (42)$$

where $J = J_N$ tends to infinity at a suitable rate so that \widehat{G} is consistent for G .³⁶

6 Numerical illustration

In this last section of the paper, we illustrate the functional differencing approach in two simple models. We start by discussing implementation issues.

6.1 Practical implementation

To implement our method in practice, we approximate the within projection operator using a discretization approach.³⁷ The approximation method uses the parametric probability model of y_i given (x_i, α_i) to generate natural bases of functions.

Specifically, we start by sampling N_y values \underline{y}_s from π_y , and approximate the projection of $h \in \mathcal{G}_y$ on $\overline{\mathcal{R}(L_{\theta, x})}$ by its projection on the span of $\{\mu_s\}$, where:

$$\mu_s(y) = \int_{\mathcal{A}} \frac{1}{\pi_\alpha(\alpha)} f_{y|x, \alpha}(y|x, \alpha; \theta) f_{y|x, \alpha}(\underline{y}_s|x, \alpha; \theta) d\alpha, \quad s = 1, \dots, N_y. \quad (43)$$

This projection takes an explicit form, which we approximate by simulation using the N_y draws \underline{y}_s . We approximate the integral in (43) using importance sampling, drawing N_α values $\underline{\alpha}_n$ from a user-specified density $\bar{\pi}$ whose support contains \mathcal{A} .³⁸

This yields the following approximation to the moment functions (see Appendix D):

$$\varphi_r(y_i, x_i, \theta) \approx \pi_y(y_i) \left(h_r(y_i) - \left(\underline{f}_{\theta, x_i}^{(y_i)} \right)' \underline{L}_{\theta, x_i}^\dagger \underline{h}_r \right) \zeta_{s_r}(x_i), \quad (44)$$

³⁶Given that $\widehat{\theta}$ has a stable asymptotic distribution, the subsampling approach of Politis *et al.* (1999) provides a non-analytical alternative to conduct inference on θ_0 .

³⁷This approach is sometimes called *least-squares collocation* (e.g., Chapter 17 in Kress, 1999, and Carrasco and Florens, 2009). Here we assume that π_y is a distribution function, although this may easily be relaxed. Matlab codes implementing the approach are available at: <http://www.cemfi.es/~bonhomme/>

³⁸The choice of the importance sampling weight function $\bar{\pi}$ may affect the quality of the numerical approximation. In the special case where the support \mathcal{A} is known and finite, $\bar{\pi}$ may be chosen as a discrete uniform distribution on \mathcal{A} .

where

$$\underline{h}_r = \left[\left(h_r \left(\underline{y}_s \right) \right)_s \right], \quad \underline{f}_{\theta,x}^{(y)} = \left[\left(\frac{1}{\sqrt{\pi_\alpha(\underline{\alpha}_n)} \bar{\pi}(\underline{\alpha}_n)} f_{y|x,\alpha}(y|x, \underline{\alpha}_n; \theta) \right)_n \right]$$

are $N_y \times 1$ and $N_\alpha \times 1$ vectors, respectively, and where:

$$\underline{L}_{\theta,x} = \left[\left(\frac{1}{\sqrt{\pi_\alpha(\underline{\alpha}_n)} \bar{\pi}(\underline{\alpha}_n)} f_{y|x,\alpha} \left(\underline{y}_s | x, \underline{\alpha}_n; \theta \right) \right)_{s,n} \right]$$

is an $N_y \times N_\alpha$ matrix. So, approximating the moment functions in this way yields an expression that is similar to the one that we obtained in the finite support case.

Note that, as the operator L_{θ,x_i} is parametric, i.e. known for given θ and x_i , we are not limited in the precision of the approximation. This means (at least conceptually) that we can choose unrestrictedly large values for N_y and N_α . In practice, however, matrix dimensions are limited by memory requirements and computation time. As a consequence, a computational curse of dimensionality may arise when T , $\dim \alpha$, or the dimension of the covariates vector are moderately large.

For this reason, it is interesting to assess the effect of approximation error. Under suitable regularity conditions (Geweke, 1989) the simulation-based approximation to $\hat{\theta}$ has an error of order $O_p \left(N_\alpha^{-\frac{1}{2}} \right) + O_p \left(N_y^{-\frac{1}{2}} \right)$. As a result, the discretization will not affect the asymptotic distribution of $\hat{\theta}$ if N_α and N_y tend to infinity faster than N . In contrast, the asymptotic distribution will generally differ in an asymptotic where the size of the discretization grows at the same rate as the sample size.

Lastly, an additional numerical issue arises when the dimensions of the matrix $\underline{L}_{\theta,x_i}$ are large. This is because, due to finite machine precision, the computation of singular vectors corresponding to small singular values may be affected by numerical errors. For this reason, we compute a modified generalized inverse that uses only $J \geq 1$ eigenvalues.³⁹ The simulation evidence displayed in the supplementary appendix suggests that taking any J in a reasonable range leads to very similar results.

6.2 Simulation evidence

The first model we consider is a tobit model with fixed effects:

$$y_{it}^* = \alpha_i + v_{it}, \quad t = 1, 2, \tag{45}$$

³⁹This modification is easily implemented using the singular value decomposition: $\underline{L}_{\theta,x_i} = \underline{\Phi} \cdot \underline{\Lambda} \cdot \underline{\Psi}'$, the J -modified Moore-Penrose inverse being equal to $\underline{\Psi}[:, 1 : J] \left(\underline{\Lambda}[1 : J, 1 : J]^{-1} \right) \underline{\Phi}[:, 1 : J]'$, where $A[1 : J, 1 : J]$ and $A[:, 1 : J]$ denote self-explanatory selections of a matrix A .

where the distribution of v_{it} given α_i is i.i.d normal $(0, \sigma^2)$, where σ is the common parameter of interest. In addition, y_{it}^* is observed only when $y_{it}^* \geq c_t$, where the thresholds c_t are known. To generate the data, we take α_i to be standard normal and $c_t = 0$ (50% censoring).

The second model is a simple version of Chamberlain (1992)'s random coefficients model:

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

where v_{i1} and v_{i2} are independent standard normal. We are interested in the common parameter θ . In the simulations we take α_i to be normal with mean 1 and unitary variance.

In the tobit model we let π_y be the density of a homogeneous tobit model with underlying normal innovations $(0, 3)$. In the random coefficients model we let π_y be a normal density $(1, 3)$. In both cases π_α is set to one, the weighting matrix Υ is chosen to be the identity, and we let $\bar{\pi}$ be uniform on $[-5, 5]$. Lastly, we set $h_r(y) = \phi(y - \mu_r)$, where ϕ is the standard bivariate normal pdf and μ_r belongs to either of three increasing sets containing 9, 25, and 49 points, respectively.⁴⁰

The first three rows of the two panels in Table 1 report the mean and standard deviation of $\hat{\sigma}$ and $\hat{\theta}$ across 1000 simulations, for two sample sizes: $N = 100$ and $N = 500$.⁴¹ We see that the functional differencing estimates present small biases. However, the difference in standard deviations between the three specifications suggests that the choice of moment functions may have important consequences on the estimator's properties.⁴²

In the third rows of the two panels, we show the results for an infeasible estimator of θ_0 based on the optimal moment restriction (33). We see that this estimator has smaller variance than the GMM estimators based on various moment functions, even though it is based on a single restriction. In order to assess the effect of using different (feasible) moment restrictions in practice, we experimented with different choices for $(\tilde{\theta}, \tilde{g}_\alpha)$ when using the strategy proposed in Section 4, and found similar results (not reported). Exploring the performance of feasible counterparts to (33) is an important task for future work.

⁴⁰Those three sets are $\{(0, 0), (0, 1), (0, -1), \dots, (-1, -1)\}$, $\{(0, 0), (0, 1), (0, -1), (0, 2), (0, -2), \dots, (-2, -2)\}$, and $\{(0, 0), (0, 1), (0, -1), (0, 2), (0, -2), (0, 3), (0, -3), \dots, (-3, -3)\}$.

⁴¹In the discretization, we have taken $N_y = 1000$ and $N_\alpha = 100$, and used Halton's quasi-random sequences to generate $\{y_s\}$ and $\{\alpha_n\}$. We used $J = 10$ singular values in the computation.

⁴²Note that estimates of asymptotic standard errors are not reported in Table 1. In our experience, the simple truncated estimate that we described at the end of Section 5 gave good results. For example, we found the average of standard errors estimates across simulations to be .078, .063, and .059 in the tobit model for $N = 500$ (the Monte Carlo standard deviations being .079, .065, and .061 in Table 1).

Lastly, it is interesting to compare the Monte Carlo standard deviations of GMM estimates to the information bounds of the models. In the supplementary appendix, we show how to adapt our discretization strategy to numerically compute the bounds. The last two rows of the panels in Table 1 show the implied theoretical values for the standard deviations. We see that not knowing $f_{\alpha|x}$ entails a loss of information. This is reflected in the fact that the infeasible random-effects estimator of θ_0 based on the true f_{α} has small variance in all designs. In addition, while the GMM estimators based on various moment functions have a much higher variance in some cases, the standard deviations of the infeasible estimator based on (33) are close to the bound.⁴³

7 Conclusion

Dealing with the incidental parameter problem in nonlinear panel data models remains a challenge to econometricians. Available solutions are often based on ingenious, model-specific methods. In a likelihood setup, we have proposed a systematic approach to construct moment restrictions on common parameters that are free from the “incidental” individual effects.

The approach consists in finding functions that are orthogonal to the range of the model operator. When supports are finite, this can be done using a simple “within” projection matrix, which differences out the unknown probabilities of individual effects. When supports are infinite, we have shown how to use a linear projection operator for the same purpose.

Our approach yields conditional moment restrictions on common parameters alone which may be informative when a condition of non-surjectivity holds. The resulting method-of-moments estimators are root- N consistent (for fixed T) and asymptotically normal, under suitable regularity conditions. This situation contrasts with the estimation of average marginal effects, which is potentially subject to a problem of ill-posedness (Bonhomme, 2011).

This paper raises a number of open questions. A first issue is implementation: the functional differencing approach requires various inputs from the researcher, such as the number and choice of moment functions to use in practice. In infinite dimensions, this choice may very much affect the finite-sample performance of the estimator. We have introduced an infeasible just-identified GMM estimator that is asymptotically efficient under suitable

⁴³The second panel in the table also reports the performance of Chamberlain (1992)’s estimator: $\tilde{\theta} = \widehat{\mathbb{E}}(y_{i2}) / \widehat{\mathbb{E}}(y_{i1})$. This estimator does not require knowledge of the distribution of α_i . The results suggest that it is more precise than the GMM estimators based on a combination of various moment restrictions, but less precise than the infeasible estimator based on (33).

conditions. Constructing feasible counterparts is an important task for the future.

A second avenue for future research is the treatment of partially identified models. In those models, it is essential to exploit the non-negativity constraints implied by the panel data model. It seems interesting to study the possibility of extending the functional differencing approach in order to incorporate these restrictions.

Lastly, a maintained assumption in this paper is that, while the distribution of individual effects given regressors is unspecified, the conditional distribution of the data given the effects is parametric. It may be important to relax the parametric assumption. For example, Hu and Schennach (2008) and Hu and Shum (2009) prove general identification results in models with latent variables under conditional independence restrictions. In panel data models with continuous dependent variables, the functional differencing approach generates a continuum of identifying restrictions on common parameters. In linear models, this allows to relax the parametric setting, provided some restrictions are imposed on the dynamics of time-varying errors (Arellano and Bonhomme, 2011). The framework introduced in this paper should be useful to extend those results to nonlinear panel data models.

References

- [1] Andersen, E.B. (1970): “Asymptotic Properties of Conditional Maximum Likelihood Estimators,” *Journal of the Royal Statistical Society B*, 32, 283-301.
- [2] Arellano, M., and S. Bonhomme (2011): “Identifying Distributional Characteristics in Random Coefficients Panel Data Models,” to appear in the *Review of Economic Studies*.
- [3] 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.
- [4] Bajari, P., J. Hahn, H. Hong, and G. Ridder (2009): “A Note on Semiparametric Estimation of Finite Mixtures of Discrete Choice Models with Application to Game Theoretic Models,” unpublished manuscript.
- [5] Bester, A., and C. Hansen (2007): “Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity,” unpublished manuscript.
- [6] Bickel, P.J., C.A.J. Klassen, Y. Ritov, and J.A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press. Baltimore and London.
- [7] Bonhomme, S. (2011): “Panel Data, Inverse Problems, and the Estimation of Policy Parameters”, unpublished manuscript.
- [8] Buchinsky, M., J. Hahn, and K.I. Kim (2008): “Semiparametric Information Bound of Dynamic Discrete Choice Models,” unpublished manuscript.
- [9] Carrasco, M., and J. P. Florens (2009): “Spectral Methods for Deconvolving a Density,” to appear in *Econometric Theory*.
- [10] Carrasco, M., J. P. Florens, and E. Renault (2008): “Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization,” *Handbook of Econometrics*, J.J. Heckman and E.E. Leamer (eds), vol. 6, North Holland.
- [11] Chamberlain, G. (1984): “Panel Data”, in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Elsevier Science.

- [12] Chamberlain, G. (1985): “Heterogeneity, Omitted Variable Bias, and Duration Dependence”, in: J.J. Heckman and B. Singer, eds. *Longitudinal analysis of labor market data*, Cambridge University Press, Cambridge, 3–38.
- [13] Chamberlain, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions”, *Journal of Econometrics*, 34, 305–334.
- [14] Chamberlain, G. (1992): “Efficiency Bounds for Semiparametric Regression”, *Econometrica*, 60, 567–596.
- [15] Chamberlain, G. (2010): “Binary Response Models for Panel Data: Identification and Information”, *Econometrica*, 78, 159–168.
- [16] Chen, X., (2007): “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in J. Heckman and E. Leamer eds. *Handbook of Econometrics*, vol. 6B, Chapter 76, 5549–5632. New York: Elsevier Science.
- [17] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2009): “Identification and Estimation of Marginal Effects in Nonlinear Panel Models,” CeMMAP working papers CWP05/09, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.
- [18] Cooper, R. W., and J. C. Haltiwanger (2006): “On the Nature of Capital Adjustment Costs,” *Review of Economic Studies*, 73(3), 611–633.
- [19] Cox, D. R. and N. Reid (1987): “Parameter Orthogonality and Approximate Conditional Inference” (with discussion), *Journal of the Royal Statistical Society, Series B*, 49, 1–39.
- [20] Duffy, J., C. Papageorgiou, and F. Perez-Sebastian (2004): “Capital-Skill Complementarity? Evidence from a Panel of Countries,” *Review of Economics and Statistics*, 86(1), 327–344.
- [21] Engl, H.W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems*, Kluwer Academic Publishers.
- [22] Geweke, J. (1989): “Bayesian Inference in Econometric Models Using Monte Carlo Integration”, *Econometrica*, 57, 1317–1339.

- [23] Goldberg, S., and A.H. Kruse (1962): “The Existence of Compact Linear Maps Between Banach Spaces”, *Proc. Amer. Math. Soc.*, 13, 808–811.
- [24] Hahn, J. (1997): “A Note on the Efficient Semiparametric Estimation of some Exponential Panel Models”, *Econometric Theory*, 13, 583–588.
- [25] Hahn, J. (2001): “The Information Bound of a Dynamic Panel Logit Model with Fixed Effects,” *Econometric Theory*, 17, 913–932.
- [26] Hausman, J. A. (1978): “Specification Tests in Econometrics,” *Econometrica*, 46, 1251–1272.
- [27] Hoderlein, S., and H. White (2009): “Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects”, unpublished manuscript.
- [28] Honoré, B. (1992): “Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects,” *Econometrica*, 60, 533–565.
- [29] Honoré, B. (1993): “Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variable,” *Journal of Econometrics*, 59, 35–61.
- [30] Honoré, B. and E. Kyriazidou (2000): “Panel Data Discrete Choice Models with Lagged Dependent Variables,” *Econometrica*, 68, 839–874.
- [31] Honoré, B., and E. Tamer (2006): “Bounds on Parameters in Dynamic discrete-Choice Models,” *Econometrica*, 74(3), 611-629.
- [32] Horowitz, J. L., and S. Lee (2004): “Semiparametric Estimation of a Panel Data Proportional Hazards Model with Fixed Effects”, *Journal of Econometrics*, 119(1), 155–198.
- [33] Hu, L. (2002): “Estimation of a Censored Dynamic Panel Data Model,” *Econometrica*, 70(6), 2499-2517.
- [34] Hu, Y., and S.M. Schennach (2008): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76(1), 195-216.
- [35] Hu, Y., and M. Shum (2009): “Nonparametric Identification of Dynamic Models with Unobserved State Variables,” unpublished manuscript.

- [36] Johnson, E.G. (2004): “Identification in Discrete Choice Models with Fixed Effects,” Working paper, Bureau of Labor Statistics.
- [37] Judd, K. (1998): *Numerical Methods in Economics*, MIT Press. Cambridge, London.
- [38] Kress, R. (1999): *Linear Integral Equations*, Springer.
- [39] Kyriazidou, E. (1997): “Estimation of a Panel Data Sample Selection Model,” *Econometrica*, 65, 1335–1364.
- [40] Kyriazidou, E. (2001): “Estimation of Dynamic Panel Data Sample Selection Models,” *Review of Economic Studies*, 68, 543–572.
- [41] Lancaster, T. (2000): “The Incidental Parameter Problem Since 1948,” *Journal of Econometrics*, 95, 391–413.
- [42] Lancaster, T. (2002): “Orthogonal Parameters and Panel Data”, *Review of Economic Studies*, 69, 647–666.
- [43] Lindsay, B.G. (1983): “Efficiency of the Conditional Score in a Mixture Setting”, *Annals of Statistics*, 11(2), 486–497.
- [44] Manski, C. (1987): “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data,” *Econometrica*, 55(2), 357–362.
- [45] Meghir, C., and F. Windmeijer (2000): “Moment Conditions for Dynamic Panel Data Models with Multiplicative Individual Effects in the Conditional Variance”, *Annales d’Economie et de Statistique*, 55-56, 317–330.
- [46] Newey, W.K., and D. McFadden (1994): “Large Sample Estimation and Hypothesis Testing,” in R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics* vol 4: 2111-245. Amsterdam: Elsevier Science.
- [47] Newey, W., and J. Powell (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.
- [48] Neyman, J. and E. L. Scott (1948): “Consistent Estimates Based on Partially Consistent Observations”, *Econometrica*, 16, 1–32.

- [49] Politis, N., J. Romano, and M. Wolf (1999): *Subsampling*, Springer Verlag, New York.
- [50] Rasch, G. (1961): “On the General Laws and the Meaning of Measurement in Psychology”, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, University of California Press, Berkeley and Los Angeles.
- [51] Stewart, G.W. (1977): “On the Perturbation of Pseudo-Inverses, Projections, and Linear Least Squares Problems,” *SIAM Review*, 19, 634-666.
- [52] Van den Berg, G. J. (2001): “Duration Models: Specification, Identification, and Multiple Durations,” in *Handbook of Econometrics*, Volume V, ed. by J. J. Heckman and E. Leamer. Amsterdam: North Holland.
- [53] Yoshida, K. (1971): *Functional Analysis*. Springer Verlag. New York.

APPENDIX

A Proofs (Sections 3 and 4)

Proof of Proposition 1. Under the proposition's assumptions, by a result in Stewart (1977, p. 653), $\theta \mapsto W_{\theta,x}$ is differentiable at θ_0 , a.s. in x , with derivatives:

$$-W_{\theta_0,x} \frac{\partial L_{\theta_0,x}}{\partial \theta_k} L_{\theta_0,x}^\dagger - \left(W_{\theta_0,x} \frac{\partial L_{\theta_0,x}}{\partial \theta_k} L_{\theta_0,x}^\dagger \right)', \quad k = 1, \dots, \dim \theta.$$

Noting that $W_{\theta_0,x} f_{y|x} = 0$, it follows that $\theta \mapsto W_{\theta,x} f_{y|x}$ is differentiable at θ_0 , a.s. in x , with Jacobian matrix $G(x)$. Therefore, $G(x)c = 0$ having $c = 0$ as only solution is the rank condition for local point-identification of θ_0 in (25).

Proof of Theorem 1. As a preliminary, we shall first derive Chamberlain (1987)'s optimal instruments and unconditional moment restrictions for the GMM estimation problem. For this purpose, it is useful to introduce an $N_y \times (N_y - \text{rank}(L_{\theta,x}))$ matrix $U_{\theta,x}$ with orthogonal columns such that $U_{\theta,x} U_{\theta,x}' = W_{\theta,x}$. Working with this matrix allows to remove redundant restrictions. We will denote as $\tau(y_i)$ the index in $\{1, \dots, N_y\}$ such that $y_i = \underline{y}_{\tau(y_i)}$. Lastly, $A[\tau, \cdot]$ and $A[\cdot, \tau]$ will denote the τ th row or column, respectively, of a matrix A .

Lemma A1 *Assume that $\text{rank}(L_{\theta,x})$ is constant in θ in a neighborhood \mathcal{V} of θ_0 , a.s. in x , and that $\theta \mapsto f_{y|x,\alpha}(y|x, \alpha; \theta)$ is continuously differentiable on \mathcal{V} , a.s. Lastly, assume that $\kappa_{\theta_0,x_i} = \mathbb{E}(U_{\theta_0,x_i}[\tau(y_i), \cdot]' U_{\theta_0,x_i}[\tau(y_i), \cdot] | x_i)$ is a.s. non-singular.*

Then the optimal unconditional moment conditions corresponding to (25) are given by:

$$\mathbb{E} \left(U_{\theta_0,x_i}[\tau(y_i), \cdot]' \kappa_{\theta_0,x_i}^{-1} \mathbb{E} \left[U_{\theta_0,x_i}' \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^\dagger [\cdot, \tau(y_i)] \mid x_i \right] \right) = 0, \quad k = 1, \dots, \dim \theta. \quad (\text{A1})$$

Proof.

As the rank of $L_{\theta,x}$ is independent of θ and $L_{\theta,x}$ is continuous, $\theta \mapsto W_{\theta,x}$ is continuous in a neighborhood of θ_0 (Stewart, 1977, Theorem 4.1), and so is $\theta \mapsto U_{\theta,x}$. We have:

$$\begin{aligned} \mathbb{E}[\varphi(y_i, x_i, \theta) | x_i] - \mathbb{E}[\varphi(y_i, x_i, \theta_0) | x_i] &= U_{\theta,x_i}' f_{y|x} - U_{\theta_0,x_i}' f_{y|x} \\ &= U_{\theta,x_i}' L_{\theta_0,x_i} L_{\theta_0,x_i}^\dagger f_{y|x} - U_{\theta_0,x_i}' L_{\theta_0,x_i} L_{\theta_0,x_i}^\dagger f_{y|x} \\ &= U_{\theta,x_i}' L_{\theta_0,x_i} L_{\theta_0,x_i}^\dagger f_{y|x} \\ &= -U_{\theta,x_i}' (L_{\theta,x_i} - L_{\theta_0,x_i}) L_{\theta_0,x_i}^\dagger f_{y|x}, \end{aligned}$$

where we have used that $f_{y|x} = L_{\theta_0,x} L_{\theta_0,x}^\dagger f_{y|x}$, and that

$$U_{\theta,x}' L_{\theta,x} = U_{\theta,x}' U_{\theta,x} U_{\theta,x}' L_{\theta,x} = U_{\theta,x}' W_{\theta,x} L_{\theta,x} = 0.$$

As $U_{\theta,x}$ is continuous and as $\theta \mapsto f_{y|x,\alpha}(y|x, \alpha; \theta)$ is continuously differentiable in a neighborhood of θ_0 , it follows that the moment functions are differentiable at θ_0 with derivatives:

$$\begin{aligned} \frac{\partial}{\partial \theta_k} \Big|_{\theta_0} \mathbb{E}[\varphi(y_i, x_i, \theta) | x_i] &= -U_{\theta_0,x_i}' \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^\dagger f_{y|x} \\ &= -\mathbb{E} \left[U_{\theta_0,x_i}' \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^\dagger [\cdot, \tau(y_i)] \mid x_i \right]. \end{aligned}$$

The conclusion then follows from Chamberlain (1987).

■

We shall now prove Theorem 1. We assume x away in the proof. Allowing for finitely supported x 's complicates the notation, but leaves the substance of the proof unchanged.

To proceed, let $\eta \mapsto f_\alpha(\cdot; \eta)$ be a model for the distribution of α_i , which depends on a scalar parameter η , and coincides with the true f_α when $\eta = \eta_0$. The model tangent space consists of the linear span of:

$$\frac{\partial}{\partial \eta} \Big|_{\eta_0} \ln \left(\sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) f_\alpha(\underline{\alpha}_n; \eta) \right) = \frac{\sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) \frac{\partial f_\alpha(\underline{\alpha}_n; \eta_0)}{\partial \eta}}{\sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) f_\alpha(\underline{\alpha}_n)}.$$

As the only restriction on $\frac{\partial f_\alpha(\underline{\alpha}_n; \eta_0)}{\partial \eta}$ is that $\sum_{n=1}^{N_\alpha} \frac{\partial f_\alpha(\underline{\alpha}_n; \eta_0)}{\partial \eta} = 0$, the tangent space coincides with:

$$\left\{ \frac{1}{f_y(y)} \sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) v_n, \quad \sum_{n=1}^{N_\alpha} v_n = 0 \right\}.$$

So, as $\sum_{s=1}^{N_y} f_{y|\alpha}(\underline{y}_s | \alpha; \theta_0) = 1$ for all α , the tangent space also coincides with:

$$\left\{ \mathbb{S}(y) = \frac{1}{f_y(y)} \sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) v_n, \quad \sum_{s=1}^{N_y} \mathbb{S}(\underline{y}_s) = 0 \right\}.$$

Given that the scores are defined over the finite support of y_i , it is convenient to work with $N_y \times 1$ vectors: $\mathbb{S} = \left(\mathbb{S}(\underline{y}_1), \dots, \mathbb{S}(\underline{y}_{N_y}) \right)'$. Using this convention, the tangent space is the set of score vectors (which sum to zero) that belong to the range of $D_f^{-1} L_{\theta_0}$, where D_f is the $N_y \times N_y$ matrix with s th diagonal element $f_y(\underline{y}_s)$.

The efficient score is obtained by projecting:

$$\frac{\partial}{\partial \theta_k} \Big|_{\theta_0} \ln \left(\sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta) f_\alpha(\underline{\alpha}_n) \right) = \frac{\sum_{n=1}^{N_\alpha} \frac{\partial f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0)}{\partial \theta_k} f_\alpha(\underline{\alpha}_n)}{\sum_{n=1}^{N_\alpha} f_{y|\alpha}(y|\underline{\alpha}_n; \theta_0) f_\alpha(\underline{\alpha}_n)}, \quad k = 1, \dots, \dim \theta,$$

on the tangent space, and taking the residual.

In vector form, we thus obtain the efficient score as the residual in the population regression of $D_f^{-1} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha$ on the columns of $D_f^{-1} L_{\theta_0}$. The coefficient in that regression is:

$$\left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger D_f^{-\frac{1}{2}} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha,$$

because the least squares weights—taking into account the fact that each \underline{y}_s is weighted by $f_y(\underline{y}_s)$ —are: $\frac{1}{f_y(\underline{y}_s)^2} \times f_y(\underline{y}_s) = \frac{1}{f_y(\underline{y}_s)}$. So the efficient score with respect to θ_k is:

$$\begin{aligned} \mathbb{S}_k^* &= D_f^{-1} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha - D_f^{-1} L_{\theta_0} \left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger D_f^{-\frac{1}{2}} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha \\ &= D_f^{-\frac{1}{2}} \left[I_{N_y} - D_f^{-\frac{1}{2}} L_{\theta_0} \left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger \right] D_f^{-\frac{1}{2}} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha. \end{aligned}$$

To make the link with functional differencing, we state the following result.

Lemma A2

$$D_f^{-\frac{1}{2}} L_{\theta_0} \left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger + D_f^{\frac{1}{2}} U_{\theta_0} \left(D_f^{\frac{1}{2}} U_{\theta_0} \right)^\dagger = I_{N_y},$$

where U_{θ_0} is the $N_y \times (N_y - \text{rank}(L_{\theta_0,x}))$ matrix with orthogonal columns such that $U_{\theta_0,x} U'_{\theta_0,x} = W_{\theta_0,x}$.

Proof. Note that $D_f^{-\frac{1}{2}} L_{\theta_0} \left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger$ is the orthogonal projector on the range of $D_f^{-\frac{1}{2}} L_{\theta_0}$. Likewise, $D_f^{\frac{1}{2}} U_{\theta_0} \left(D_f^{\frac{1}{2}} U_{\theta_0} \right)^\dagger$ is the orthogonal projector on the range of $D_f^{\frac{1}{2}} U_{\theta_0}$, which is the orthogonal complement of the range of $D_f^{-\frac{1}{2}} L_{\theta_0}$. ■

From Lemma A2 we can rewrite the efficient score as:

$$\begin{aligned} \mathbb{S}_k^* &= D_f^{-\frac{1}{2}} \left[I_{N_y} - D_f^{-\frac{1}{2}} L_{\theta_0} \left(D_f^{-\frac{1}{2}} L_{\theta_0} \right)^\dagger \right] D_f^{-\frac{1}{2}} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha \\ &= D_f^{-\frac{1}{2}} \left[D_f^{\frac{1}{2}} U_{\theta_0} \left(D_f^{\frac{1}{2}} U_{\theta_0} \right)^\dagger \right] D_f^{-\frac{1}{2}} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha \\ &= U_{\theta_0} (U'_{\theta_0} D_f U_{\theta_0})^{-1} U'_{\theta_0} \frac{\partial L_{\theta_0}}{\partial \theta_k} f_\alpha, \end{aligned}$$

where $U'_{\theta_0} D_f U_{\theta_0}$ is non-singular as U_{θ_0} has full-column rank.

Lastly, taking derivatives in the identity $L_{\theta_0} L_{\theta_0}^\dagger L_{\theta_0} = L_{\theta_0}$ and left-multiplying by U'_θ we obtain:

$$U'_\theta \frac{\partial L_{\theta_0}}{\partial \theta_k} L_{\theta_0}^\dagger L_{\theta_0} = U'_\theta \frac{\partial L_{\theta_0}}{\partial \theta_k},$$

where we have used that $U'_\theta L_{\theta_0} = 0$.

So, we finally obtain:

$$\mathbb{S}_k^* = U_{\theta_0} (U'_{\theta_0} D_f U_{\theta_0})^{-1} U'_{\theta_0} \frac{\partial L_{\theta_0}}{\partial \theta_k} L_{\theta_0}^\dagger f_y.$$

Comparing with (A1) ends the proof.

Proof of Theorem 2. First note that:

$$\text{Proj}_{\pi_y} \left[f_{y|x} \mid \overline{\mathcal{R}(L_{\theta_0,x})} \right] = L_{\theta_0,x} f_{\alpha|x} = f_{y|x}.$$

It follows that $W_{\theta_0,x} f_{y|x} = 0$, x -a.s. Hence (28). To show that (28) and (29) are equivalent, note that, as $W_{\theta_0,x}$ is self-adjoint (or symmetric, e.g., Yoshida, 1971, p. 83):

$$\begin{aligned} W_{\theta_0,x} f_{y|x} = 0 &\Leftrightarrow \langle h, W_{\theta_0,x} f_{y|x} \rangle = 0 \text{ for all } h \in \mathcal{G}_y \\ &\Leftrightarrow \langle W_{\theta_0,x} h, f_{y|x} \rangle = 0 \text{ for all } h \in \mathcal{G}_y \\ &\Leftrightarrow \left[\int_{\mathcal{Y}} [W_{\theta_0,x} h](y) f_{y|x}(y|x) \pi_y(y) dy = 0 \text{ for all } h \in \mathcal{G}_y \right] \\ &\Leftrightarrow \left[\mathbb{E} \left(\pi_y(y_i) [W_{\theta_0,x_i} h](y_i) \mid x_i = x \right) = 0 \text{ for all } h \in \mathcal{G}_y \right]. \end{aligned}$$

B Information bound (Section 4)

Information bound : characterization. We start by noting that the analysis may be conditioned on an $x \in \mathcal{X}$ (as in Hahn, 1997). The nonparametric tangent space of the model consists of the $L^2(f_{y|x})$ -closure of the linear span of:

$$\frac{\partial}{\partial \eta} \Big|_{\eta_0} \ln ([L_{\theta_0, x} f_{\alpha|x}(\cdot|x; \eta)](y)) = \frac{[L_{\theta_0, x} \frac{\partial f_{\alpha|x}(\cdot|x; \eta_0)}{\partial \eta}](y)}{[L_{\theta_0, x} f_{\alpha|x}(\cdot|x)](y)}.$$

Note that $\frac{\partial \ln f_{\alpha|x}(\cdot|x; \eta_0)}{\partial \eta}$ is not restricted beyond the fact that it has zero expectation and is square integrable with respect to $f_{\alpha|x}(\cdot|x; \eta_0) = f_{\alpha|x}(\cdot|x)$. This is because $\inf_{\alpha \in \mathcal{A}} f_{\alpha|x}(\alpha|x) > 0$, and thus $f_{\alpha|x}(\cdot|x)$ is *interior*. This guarantees that the tangent space is a linear vector space, as opposed to a convex cone (see Lindsay, 1983). This assumption (which implies that individual effects must have compact support) may be stronger than necessary. For example, in exponential family models it is enough to assume that $f_{\alpha|x}(\cdot|x)$ has non-empty interior in order not to restrict the tangent space (Hahn, 1997).

It follows from this discussion that, letting:

$$\mathcal{D}_\alpha = \left\{ v(\cdot, x), \frac{v(\cdot, x)}{f_{\alpha|x}} \in L^2(f_{\alpha|x}), \mathbb{E} \left(\frac{v(\alpha_i, x_i)}{f_{\alpha|x}(\alpha_i|x_i)} \Big| x_i = x \right) = 0 \right\},$$

the tangent space is the $L^2(f_{y|x})$ -closure of the linear span of $\left\{ \frac{L_{\theta_0, x} v(\cdot, x)}{L_{\theta_0, x} f_{\alpha|x}}, v(\cdot, x) \in \mathcal{D}_\alpha \right\}$.

Now, for all v such that $\frac{v(\cdot, x)}{f_{\alpha|x}} \in L^2(f_{\alpha|x})$ we have almost surely:⁴⁴

$$\begin{aligned} \int_{\mathcal{Y}} [L_{\theta_0, x} v(\cdot, x)](y) dy &= \int_{\mathcal{Y}} \left(\int_{\mathcal{A}} f_{y|x, \alpha}(y|x, \alpha; \theta_0) v(\alpha, x) d\alpha \right) dy \\ &= \int_{\mathcal{A}} \left(\int_{\mathcal{Y}} f_{y|x, \alpha}(y|x, \alpha; \theta_0) dy \right) v(\alpha, x) d\alpha \\ &= \int_{\mathcal{A}} v(\alpha, x) d\alpha. \end{aligned}$$

It thus follows that the tangent space is the $L^2(f_{y|x})$ -closure of the linear span of:

$$\frac{1}{f_{y|x}} \cdot \left\{ h(\cdot, x) \in \mathcal{R}(L_{\theta_0, x}), \mathbb{E} \left(\frac{h(y_i, x_i)}{f_{y|x}(y_i|x_i)} \Big| x_i = x \right) = 0 \right\} = \frac{1}{f_{y|x}} \cdot \mathcal{R}_y.$$

Denoting as \mathbb{S}_k^* the efficient score for θ_k we thus have:

$$\mathbb{S}_k^*(y, x; \theta_0) = \frac{\partial}{\partial \theta_k} \Big|_{\theta_0} \ln ([L_{\theta, x} f_{\alpha|x}](y)) - m^{(k)}(y, x), \quad k = 1, \dots, \dim \theta,$$

where:

$$m^{(k)}(\cdot, x) = \underset{m \in \text{closure} \left(\frac{1}{f_{y|x}} \cdot \mathcal{R}_y \right)}{\text{argmin}} \mathbb{E} \left[\left(\frac{\partial}{\partial \theta_k} \Big|_{\theta_0} \ln ([L_{\theta, x_i} f_{\alpha|x}](y_i)) - m(y_i) \right)^2 \Big| x_i = x \right], \quad x - a.s.$$

The equivalence with (30)-(31) follows by taking $h^{(k)}(y, x) = f_{y|x}(y|x) \cdot m^{(k)}(y, x)$.

⁴⁴Note that, by the Cauchy-Schwarz inequality:

$$\left(\int_{\mathcal{A}} |v(\alpha, x)| d\alpha \right)^2 \leq \int_{\mathcal{A}} \left(\frac{v(\alpha, x)}{f_{\alpha|x}(\alpha|x)} \right)^2 f_{\alpha|x}(\alpha|x) d\alpha < \infty, \quad x - a.s.$$

Information bound: adding-up constraint. To show that $h^{(k)}$ is the orthogonal projection of $\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k}$ on $\overline{\mathcal{R}(L_{\theta_0, x})}$, it is sufficient to show that:

$$\tilde{h}^{(k)} \equiv \text{Proj}_{\pi_y} \left[\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k} \Big| \overline{\mathcal{R}(L_{\theta_0, x})} \right] \in \overline{\mathcal{R}_y}. \quad (\text{B2})$$

From the above characterization of \mathcal{R}_y , it is thus sufficient to show that:

$$\int_{\mathcal{Y}} \tilde{h}^{(k)}(y, x) dy = 0, \quad x - a.s.$$

Or, equivalently, given the choice of weight functions (32), that:

$$\left\langle \tilde{h}^{(k)}(\cdot, x), f_{y|x}(\cdot|x) \right\rangle = 0, \quad x - a.s. \quad (\text{B3})$$

Now, as $f_{y|x, \alpha}(\cdot|x, \alpha; \theta_0)$ is a conditional distribution function, we have:

$$\left\langle \frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k}, f_{y|x}(\cdot|x) \right\rangle = \int_{\mathcal{Y}} \left[\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k} \right] (y) dy = 0, \quad x - a.s.$$

In addition, as $f_{y|x} \in \overline{\mathcal{R}(L_{\theta_0, x})}$, we have that $\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k} - \tilde{h}^{(k)}$ is \mathcal{G}_y -orthogonal to $f_{y|x}$. Combining the results, it then follows that $\tilde{h}^{(k)} = \frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k} - \left(\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k} - \tilde{h}^{(k)} \right)$ is also \mathcal{G}_y -orthogonal to $f_{y|x}$.

This shows (B3). Hence, $h^{(k)} = \tilde{h}^{(k)}$ coincides with the orthogonal projection of $\frac{\partial L_{\theta_0, x} f_{\alpha|x}}{\partial \theta_k}$ on $\overline{\mathcal{R}(L_{\theta_0, x})}$.

C Asymptotic results (Section 5)

Proof of Theorem 3. We verify the conditions of Theorem 2.1 in Newey and McFadden (1994). First, note that observations are i.i.d., and that the global identification condition holds, with Θ compact. The rest of the proof consists of two steps.

Step 1 consists in showing that the population objective function is continuous on the parameter space. Let, for $\mu > 0$:

$$W_{\theta, x}^{(\mu)} = I_y - L_{\theta, x} (L_{\theta, x}^* L_{\theta, x} + \mu I_{\alpha})^{-1} L_{\theta, x}^*. \quad (\text{C4})$$

We start with the following result.

Lemma C3 *Let iii), iv), and viii) in Assumption 2 hold. Then, for every r and for $\mu > 0$ given, the function:*

$$\theta \mapsto \mathbb{E} \left(\left[W_{\theta, x_i}^{(\mu)} h_r \right] (y_i) \pi_y(y_i) \zeta_{s_r}(x_i) \right)$$

is continuous on Θ .

Proof. Conditions iii) and iv) imply that the mapping $\theta \mapsto L_{\theta, x}$ is continuous on Θ with respect to the operator norm, x -a.s. This statement follows from the fact that, if $\theta_s \xrightarrow{s \rightarrow \infty} \theta$, then (e.g., Section 2.2 in Carrasco *et al.*, 2008):

$$\|L_{\theta_s, x} - L_{\theta, x}\|^2 \leq \sup_{\theta \in \Theta} \int_{\mathcal{Y}} \int_{\mathcal{A}} [f_{y|x, \alpha}(y|x, \alpha; \theta_s) - f_{y|x, \alpha}(y|x, \alpha; \theta)]^2 \frac{\pi_y(y)}{\pi_{\alpha}(\alpha)} dy d\alpha,$$

which tends to zero by iii), iv), and an application of Lebesgue's dominated convergence theorem.

So, the mapping $\theta \mapsto W_{\theta,x}^{(\mu)}$ is also continuous on Θ with respect to the operator norm, a.s. in x . Now, note that the singular values of $W_{\theta,x}^{(\mu)}$ are either equal to 1 or to some $-\frac{\mu}{\mu + \lambda_{j,\theta,x}^2}$, for $j \in \{1, 2, \dots\}$. It thus follows that $\|W_{\theta,x}^{(\mu)}\| \leq 1$ for all θ, x . So, letting again $\theta_s \xrightarrow{s \rightarrow \infty} \theta$ we have:

$$\begin{aligned} \left| \mathbb{E} \left(\left[\left(W_{\theta_s, x_i}^{(\mu)} - W_{\theta, x_i}^{(\mu)} \right) h_r \right] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \right) \right| &= \left| \mathbb{E} \left(\left\langle \left(W_{\theta_s, x_i}^{(\mu)} - W_{\theta, x_i}^{(\mu)} \right) h_r, f_{y|x} \right\rangle \zeta_{s_r} (x_i) \right) \right| \\ &\leq \mathbb{E} \left(\left\| \left(W_{\theta_s, x_i}^{(\mu)} - W_{\theta, x_i}^{(\mu)} \right) h_r \right\| \|f_{y|x}\| |\zeta_{s_r} (x_i)| \right). \end{aligned}$$

The term within the expectation tends to zero by continuity of $\theta \mapsto W_{\theta,x}^{(\mu)}$. Moreover, it is dominated by $2 \|h_r\| \|f_{y|x}\| |\zeta_{s_r} (x_i)|$, which has finite expectation by *viii*) and the Cauchy-Schwarz inequality. The conclusion follows from the dominated convergence theorem.

■

Lemma C4 *Let $v)$, $vi)$ and $viii)$ in Assumption 2 hold. Then, for every r :*

$$\mathbb{E} \left(\left[W_{\theta, x_i}^{(\mu)} h_r \right] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \right) \xrightarrow{\mu \rightarrow 0} \mathbb{E} \left([W_{\theta, x_i} h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \right)$$

where the convergence holds uniformly on Θ .

Proof. We have:

$$\begin{aligned} B &\equiv \mathbb{E} \left(\left[\left(W_{\theta, x_i}^{(\mu)} - W_{\theta, x_i} \right) h_r \right] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \right) \\ &= \mathbb{E} \left(\left\langle \left(W_{\theta, x_i}^{(\mu)} - W_{\theta, x_i} \right) h_r, f_{y|x} \right\rangle \zeta_{s_r} (x_i) \right) \\ &= \mathbb{E} \left(\sum_j \frac{-\mu}{\mu + \lambda_{j,\theta, x_i}^2} \langle \phi_{j,\theta, x_i}, f_{y|x} \rangle \langle \phi_{j,\theta, x_i}, h_r \rangle \zeta_{s_r} (x_i) \right). \end{aligned}$$

So, for every $J \geq 1$:

$$\begin{aligned} |B| &\leq \mu \sum_{j \leq J} \mathbb{E} \left(\frac{1}{\inf_{\theta \in \Theta} \lambda_{j,\theta, x_i}^2} |\langle \phi_{j,\theta, x_i}, f_{y|x} \rangle \langle \phi_{j,\theta, x_i}, h_r \rangle \zeta_{s_r} (x_i)| \right) \\ &\quad + \mathbb{E} \left(\sum_{j > J} |\langle \phi_{j,\theta, x_i}, f_{y|x} \rangle \langle \phi_{j,\theta, x_i}, h_r \rangle \zeta_{s_r} (x_i)| \right). \end{aligned}$$

So, using the Cauchy-Schwarz inequality:

$$\begin{aligned} \sup_{\theta \in \Theta} |B| &\leq \mu \sum_{j \leq J} \mathbb{E} \left(\frac{1}{\inf_{\theta \in \Theta} \lambda_{j,\theta, x_i}^2} \|f_{y|x}\| \|h_r\| |\zeta_{s_r} (x_i)| \right) \\ &\quad + \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\sum_{j > J} \langle \phi_{j,\theta, x_i}, f_{y|x} \rangle^2 \right)^{\frac{1}{2}} \|h_r\| |\zeta_{s_r} (x_i)| \right]. \end{aligned}$$

Fix $\varepsilon > 0$. By *vi)*, *viii)* and the dominated convergence theorem, the second term on the right-hand side tends to zero as J tends to infinity. So there exists a J such that this term is $< \varepsilon/2$.

For that J , take μ small enough such that the first term is $< \varepsilon/2$. Such a μ exists by v). This shows the lemma.

■

Combining Lemmas C3 and C4 then shows that

$$\theta \mapsto \mathbb{E} \left([W_{\theta, x_i} h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \right)$$

is continuous on Θ , for all r . This ends Step 1 of the proof.

Lastly, in Step 2 we show uniform convergence in probability of the sample moment restrictions to the population moment restrictions. To do this, let us denote

$$\varphi_r = \pi_y (y_i) [W_{\theta, x_i} h_r] (y_i) \zeta_{s_r} (x_i).$$

We will show:

$$\sup_{\theta \in \Theta} \mathbb{E} \left(\left[\widehat{\mathbb{E}}(\varphi_r) - \mathbb{E}(\varphi_r) \right]^2 \right) \xrightarrow{N \rightarrow \infty} 0. \quad (\text{C5})$$

For this, we will show two lemmas.

Lemma C5 *Let viii) in Assumption 2 hold. Then*

$$\sup_{\theta \in \Theta} \text{Var} \left(\mathbb{E} \left([W_{\theta, x_i} h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \mid x_i \right) \right) < \infty.$$

Proof.

$$\begin{aligned} \text{Var} \left(\mathbb{E} \left([W_{\theta, x_i} h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \mid x_i \right) \right) &= \text{Var} \left(\langle W_{\theta, x_i} h_r, f_{y|x} \rangle \zeta_{s_r} (x_i) \right) \\ &\leq \mathbb{E} \left(\langle W_{\theta, x_i} h_r, f_{y|x} \rangle^2 \zeta_{s_r} (x_i)^2 \right) \\ &\leq \mathbb{E} \left(\|W_{\theta, x_i} h_r\|^2 \|f_{y|x}\|^2 \zeta_{s_r} (x_i)^2 \right) \\ &\leq \mathbb{E} \left(\|h_r\|^2 \|f_{y|x}\|^2 \zeta_{s_r} (x_i)^2 \right), \end{aligned}$$

where we have used that $\|W_{\theta, x_i}\| \leq 1$. The conclusion follows from viii).

■

Lemma C6 *Let vii) in Assumption 2 hold. Then*

$$\sup_{\theta \in \Theta} \mathbb{E} \left(\text{Var} \left([W_{\theta, x_i} h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) \mid x_i \right) \right) < \infty.$$

Proof. We have, almost surely in x :

$$\begin{aligned} \text{Var} \left([W_{\theta, x} h_r] (y_i) \pi_y (y_i) \mid x \right) &\leq \int_{\mathcal{Y}} \{ [W_{\theta, x} h_r] (y) \pi_y (y) \}^2 f_{y|x} (y|x) dy \\ &\leq \sup_{y \in \mathcal{Y}} \left(\pi_y (y) f_{y|x} (y|x) \right) \int_{\mathcal{Y}} \{ [W_{\theta, x} h_r] (y) \}^2 \pi_y (y) dy \\ &= \sup_{y \in \mathcal{Y}} \left(\pi_y (y) f_{y|x} (y|x) \right) \|W_{\theta, x} h_r\|^2 \\ &\leq \sup_{y \in \mathcal{Y}} \left(\pi_y (y) f_{y|x} (y|x) \right) \|h_r\|^2, \end{aligned}$$

where we have used that $\|W_{\theta,x}\| \leq 1$.

So, by *vii*), $\mathbb{E} \left[\text{Var} ([W_{\theta,x_i} h_r](y_i) \pi_y(y_i) | x_i) \zeta_{s_r}(x_i)^2 \right]$ is uniformly bounded, and the conclusion follows.

■

Finally, combining Lemmas C5 and C6, $\text{Var}(\varphi_r)$ is uniformly bounded. So, the left-hand side in (C5) is bounded by a constant divided by N . This shows convergence in mean squares, which implies convergence in probability.

So the consistency of $\hat{\theta}$ is proved.

Proof of Theorem 4. We verify the conditions of Theorem 7.2 in Newey and McFadden (1994). First, we prove that $\theta \mapsto \mathbb{E}(\varphi(y_i, x_i, \theta))$ is differentiable at θ_0 with derivative G . For this, note that:

$$\begin{aligned} \mathbb{E}(\varphi_r(y_i, x_i, \theta)) - \mathbb{E}(\varphi_r(y_i, x_i, \theta_0)) &= \mathbb{E}(\langle W_{\theta,x_i} h_r, f_{y|x} \rangle \zeta_{s_r}(x_i)) - \mathbb{E}(\langle W_{\theta_0,x_i} h_r, f_{y|x} \rangle \zeta_{s_r}(x_i)) \\ &= \mathbb{E}(\langle (W_{\theta,x_i} - W_{\theta_0,x_i}) h_r, f_{y|x} \rangle \zeta_{s_r}(x_i)) \\ &= \mathbb{E}(\langle (W_{\theta,x_i} - W_{\theta_0,x_i}) h_r, L_{\theta_0,x_i} L_{\theta_0,x_i}^\dagger f_{y|x} \rangle \zeta_{s_r}(x_i)) \\ &= \mathbb{E}(\langle L_{\theta_0,x_i}^* W_{\theta,x_i} h_r, L_{\theta_0,x_i}^\dagger f_{y|x} \rangle \zeta_{s_r}(x_i)) \\ &= -\mathbb{E}(\langle (L_{\theta,x_i} - L_{\theta_0,x_i})^* W_{\theta,x_i} h_r, L_{\theta_0,x_i}^\dagger f_{y|x} \rangle \zeta_{s_r}(x_i)), \end{aligned}$$

where we have used that $f_{y|x} = L_{\theta_0,x_i} L_{\theta_0,x_i}^\dagger f_{y|x}$, and that $L_{\theta_0,x_i}^* W_{\theta_0,x_i} = 0$ for all θ .

By *i*) and *ii*) in Assumption 3 the mapping $\theta \mapsto L_{\theta,x}$ is continuously differentiable on \mathcal{V} , x -a.s. It follows from the mean-value theorem that

$$\mathbb{E}(\varphi_r(y_i, x_i, \theta)) - \mathbb{E}(\varphi_r(y_i, x_i, \theta_0)) = -\mathbb{E} \left(\left\langle \frac{\partial L_{\theta_0,x_i}^*}{\partial \theta'} W_{\theta,x_i} h_r, L_{\theta_0,x_i}^\dagger f_{y|x} \right\rangle \zeta_{s_r}(x_i) \right) (\theta - \theta_0),$$

where $\tilde{\theta}$ lies between θ and θ_0 .

Now, as in the proof of Theorem 3 and using in addition Condition *iii*), the function $\theta \mapsto W_{\theta,x} h_r$ is continuous on \mathcal{V} , a.s. in x . To see this, note that, for every $J \geq 1$:

$$\left\| W_{\theta,x}^{(\mu)} h_r - W_{\theta,x} h_r \right\|^2 \leq \mu^2 \sum_{j=1}^J \frac{1}{\lambda_{j,\theta,x}^4} \langle \phi_{j,\theta,x}, h_r \rangle^2 + \sum_{j>J} \langle \phi_{j,\theta,x}, h_r \rangle^2.$$

The second term on the right-hand side tends uniformly to zero as J tends to infinity by *iii*). Moreover, as $\lambda_{j,\theta,x}$ is bounded from below for $j \in \{1, \dots, J\}$, and as $\langle \phi_{j,\theta,x}, h_r \rangle^2 \leq \|h_r\|^2$, the first term tends uniformly to zero as μ tends to zero (for fixed J). This shows that $W_{\theta,x}^{(\mu)} h_r$ tends to $W_{\theta,x} h_r$ as μ tends to zero, uniformly on \mathcal{V} .

It follows that, for every $k \in \{1, \dots, \dim \theta\}$ and a.s. in x :

$$\left\langle \frac{\partial L_{\tilde{\theta},x}^*}{\partial \theta_k} W_{\theta,x} h_r, L_{\theta_0,x}^\dagger f_{y|x} \right\rangle \xrightarrow{\theta \rightarrow \theta_0} \left\langle \frac{\partial L_{\theta_0,x}^*}{\partial \theta_k} W_{\theta_0,x} h_r, L_{\theta_0,x}^\dagger f_{y|x} \right\rangle.$$

Thus, by *iv*) and the dominated convergence theorem, $\theta \mapsto \mathbb{E}(\varphi(y_i, x_i, \theta))$ is differentiable at θ_0 with derivative G .

Next, by the first part of *vi*) the empirical moment functions tend in distribution to $N[0, \Sigma(\theta_0)]$. The theorem will thus be proved if we can show stochastic equicontinuity. Now, by the second part of *vi*) we have:

$$\sqrt{N} \left(\hat{\mathbb{E}}[\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0)] - \mathbb{E}[\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0)] \right) \xrightarrow{d} N[0, \text{Var}(\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0))].$$

As in the proof of Lemma C5 we have:

$$\text{Var} \left(\mathbb{E} \left([(W_{\theta, x_i} - W_{\theta_0, x_i}) h_r] (y_i) \pi_y (y_i) \zeta_{s_r} (x_i) | x_i \right) \right) \leq \mathbb{E} \left(\|W_{\theta, x_i} h_r - W_{\theta_0, x_i} h_r\|^2 \|f_{y|x}\|^2 \zeta_{s_r} (x_i)^2 \right).$$

The term inside the expectation tends to zero as θ tends to θ_0 , as $\theta \mapsto W_{\theta, x} h_r$ is continuous. Condition *viii*) in Assumption 2 and the dominated convergence theorem thus imply that the between- x variance tends to zero as θ tends to θ_0 .

Lastly, as in the proof of Lemma C6 we have, almost surely in x :

$$\text{Var} \left([W_{\theta, x} h_r - W_{\theta_0, x} h_r] (y_i) \pi_y (y_i) | x \right) \leq \sup_{y \in \mathcal{Y}} \left(\pi_y (y) f_{y|x} (y|x) \right) \|W_{\theta, x} h_r - W_{\theta_0, x} h_r\|^2.$$

The right-hand side in this expression tends to zero as θ tends to θ_0 , again by the continuity of $\theta \mapsto W_{\theta, x} h_r$. Moreover, Condition *vii*) in Assumption 2 shows that this term (multiplied by $\zeta_{s_r} (x_i)^2$) is dominated in expectation, and the dominated convergence theorem concludes that the within- x variance tends to zero as θ tends to θ_0 .

This shows stochastic equicontinuity and ends the proof.

D Discretization strategy (Section 6)

Let \mathcal{M} be the span of $\{\mu_s\}$, and let $a \in \mathcal{G}_y$ such that $\text{Proj}_{\pi_y} [h_r | \overline{\mathcal{R}(L_{\theta, x})}] = L_{\theta, x} L_{\theta, x}^* a$.⁴⁵ We have, for given y, x :

$$\begin{aligned} [L_{\theta, x} L_{\theta, x}^* a] (y) &= \int_{\mathcal{Y}} \int_{\mathcal{A}} f_{y|x, \alpha} (y|x, \alpha; \theta) f_{y|x, \alpha} (\tilde{y}|x, \alpha; \theta) a(\tilde{y}) \frac{\pi_y(\tilde{y})}{\pi_\alpha(\alpha)} d\tilde{y} d\alpha \\ &\approx \frac{1}{N_y} \sum_{s=1}^{N_y} a(\underline{y}_s) \underbrace{\int_{\mathcal{A}} \frac{1}{\pi_\alpha(\alpha)} f_{y|x, \alpha} (y|x, \alpha; \theta) f_{y|x, \alpha} (\underline{y}_s|x, \alpha; \theta) d\alpha}_{=\mu_s(y)}, \end{aligned}$$

where the difference is $O_p(1/\sqrt{N_y})$ under suitable assumptions (e.g. Geweke, 1989).

Let $\tilde{h}_r = \text{Proj}_{\pi_y} [h_r | \overline{\mathcal{R}(L_{\theta, x})}]$. As the orthogonal projection of h_r on \mathcal{M} and the orthogonal projection of \tilde{h}_r on \mathcal{M} coincide, $\tilde{h}_r(y)$ may thus be approximated by $\sum_{s=1}^{N_y} a_s \mu_s(y)$, where $\{a_s\}$ minimizes $\int_{\mathcal{Y}} \left(h_r(y) - \sum_{s=1}^{N_y} a_s \mu_s(y) \right)^2 \pi_y(y) dy$. Solving for $\{a_s\}$ and approximating the integrals by simulation we have:

$$\tilde{h}_r(y) \approx \sum_{s=1}^{N_y} \tilde{a}_s \mu_s(y),$$

where:

$$[(\tilde{a}_s)_s] = \left[\left(\sum_{s=1}^{N_y} \mu_{s_1}(\underline{y}_s) \mu_{s_2}(\underline{y}_s) \right)_{s_1, s_2} \right]^\dagger \left[\left(\sum_{s=1}^{N_y} \mu_{s_1}(\underline{y}_s) h_r(\underline{y}_s) \right)_{s_1} \right]. \quad (\text{D6})$$

Lastly, in order to compute $\mu_s(y)$ we approximate the integrals with respect to α by importance sampling using $\bar{\pi}$ as our proposal density. This yields: $[(\mu_s(y))_s] \approx \frac{1}{N_\alpha} \underline{L}_{\theta, x} \underline{f}_{\theta, x}^{(y)}$, where the

⁴⁵Note that $\mathcal{R}(L_{\theta, x} L_{\theta, x}^*)$ is dense in $\overline{\mathcal{R}(L_{\theta, x})}$.

approximation is $O_p(1/\sqrt{N_\alpha})$. So, using (D6) and the fact that $A^\dagger = (A'A)^\dagger A' = A'(AA')^\dagger$ we have:

$$\begin{aligned}\tilde{h}_r(y) &\approx \left(\underline{f}_{\theta,x}^{(y)}\right)' \underline{L}'_{\theta,x} \left(\underline{L}_{\theta,x} \underline{L}'_{\theta,x} \underline{L}_{\theta,x} \underline{L}'_{\theta,x}\right)^\dagger \underline{L}_{\theta,x} \underline{L}'_{\theta,x} h_r \\ &= \left(\underline{f}_{\theta,x}^{(y)}\right)' \underline{L}_{\theta,x}^\dagger h_r.\end{aligned}$$

This yields the approximation (44).

Table 1: Simulation results ($T = 2$)

Tobit model: σ (true=1)				
	$N = 100$		$N = 500$	
	Mean	Std	Mean	Std
Grid (9 restrictions)	1.001	.143	1.003	.079
Grid (25 restrictions)	1.000	.123	1.001	.065
Grid (49 restrictions)	1.000	.119	1.000	.061
Efficient (1 restriction)	1.014	.113	.990	.047
Infeasible REML	1.001	.092	1.000	.043
Bound (unknown f_α)	-	$\approx .112$	-	$\approx .050$
Bound (known f_α)	-	$\approx .095$	-	$\approx .042$

Chamberlain's model: θ (true=1)				
	$N = 100$		$N = 500$	
	Mean	Std	Mean	Std
Grid (9 restrictions)	1.063	.326	1.014	.155
Grid (25 restrictions)	1.026	.225	1.001	.104
Grid (49 restrictions)	1.016	.188	.998	.085
Efficient (1 restriction)	1.030	.136	.970	.054
Infeasible REML	.999	.083	1.000	.037
Chamberlain	1	.173	1	.078
Bound (unknown f_α)	-	$\approx .112$	-	$\approx .050$
Bound (known f_α)	-	$\approx .089$	-	$\approx .040$

Note: Mean and standard deviation of $\hat{\sigma}$ and $\hat{\theta}$ across 1000 simulations. “Grid (R restrictions)” refers to using $\phi(\cdot - \mu_r)$, $r = 1, \dots, R$, to construct moment functions, where the set of values μ_r is indicated in the text. “Efficient” is the infeasible estimator based on (33). “Infeasible REML” is the infeasible random-effects maximum likelihood estimate, which assumes knowledge of f_α . “Chamberlain” is Chamberlain (1992)’s estimator of θ . Due to non-existence of moments for a ratio of sample means, asymptotic means and standard deviations are reported for that estimator. We checked that the implied quantiles provide a reasonable approximation to the quantiles of the finite-sample distribution. Lastly, the “Bounds” are numerical approximations to the minimal asymptotic standard deviations theoretically attainable.