Functional Differencing*

Stéphane Bonhomme

New York University
and CEMFI, Madrid

May 2010

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. In addition, and in contrast with common parameters, we emphasize a problem of ill-posedness in the estimation of average marginal effects. Several examples and a small-scale simulation exercise complete the paper.

JEL CODE: C23.
KEYWORDS: Panel data, incidental parameters, inverse problems.

*I thank Manuel Arellano, Alan Bester, Marine Carrasco, Gary Chamberlain, Bryan Graham, Jin Hahn, Chris Hansen, Jim Heckman, Joel Horowitz, 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, and University of Toronto for comments. 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 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

[1] Honoré 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), and linear models with variance dynamics (Meghir and Windmeijer, 2000).

[2] See Arellano and Hahn (2006) for a survey of the bias correction literature in panel data. Recent references include Hahn and Newey (2004), Carro (2007), and Arellano and Bonhomme (2009a).

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, we may apply it to models where differencing strategies are not yet known. We illustrate this point with a censored random coefficients model.[4]

The problem of finding moment restrictions in this way may have no solution. In particular, this will happen when the range of the model spans the whole space. We refer to such models as *surjective*. We show that non-surjectivity holds generally in random coefficients models, nonlinear regression models with independent additive errors, and censored regression models with normal errors, as soon as T is strictly greater than the dimension of the vector of individual effects. We conjecture that models with continuous dependent variables that satisfy the latter condition will generally be non-surjective. In contrast, static binary

[3]See Shen (1997), Ai and Chen (2003), and the recent paper by Chen and Pouzo (2009) for a very general setting that can deal with non-smooth residuals and non-compact sieve spaces. Hu and Schennach (2008) use a sieve maximum likelihood approach in a nonlinear measurement error model where the distribution of the unknown true regressor given the observed error-ridden regressor is left unrestricted. Bester and Hansen (2007) propose to adopt a similar approach in panel data models.

[4]As a possible empirical application of random coefficients models with censoring, one can mention earnings dynamics models with individual-specific slopes (Hause, 1980, Guvenen, 2009), in the presence of top or bottom-coded data.

choice models are generally surjective, with the important exception of the logit. In those models, our approach will not be informative about common parameters.

To describe our approach to construct moment functions for common parameters, we start with the special case where the data and unobserved heterogeneity distributions have known finite supports. Then, the range of the model is 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 finite support case is interesting, as our approach then results in a finite number of conditional moment restrictions on common parameters. We characterize the optimal instruments (Chamberlain, 1987) in this context. Moreover, we show that our differencing strategy achieves the semiparametric information bound of the panel data model. Indeed, the efficient score for the model can be viewed as a particular functional differencing projection.

However, in models where the information bound is zero and the non-surjectivity condition is not satisfied, our approach is not informative about common parameters.[5] In this case, it is important to exploit the fact that the probabilities of individual effects belong to the unit simplex. We outline an extension of our approach that uses this additional information.

When supports are infinite, the matrix of conditional probabilities becomes a linear integral operator, whose range may be infinite-dimensional. We build on the recent econometric literature on inverse problems (Carrasco, Florens and Renault, 2008, Carrasco and Florens, 2009) and endow the spaces of functions with scalar products, making 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 a

[5]Some important recent work has pointed out that, in models where the information bound is zero, the parameters of interest may be 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 dependent variables models.

nonparametric estimate of the outcome distribution function is not needed to estimate common parameters. To study the asymptotic properties of the generalized method-of-moments (GMM) estimator of common parameters, we impose regularity conditions that ensure operator compactness. Under identification and regularity conditions that we provide, the GMM estimator is root-N consistent and asymptotically normal. Thus, *via* our approach, common parameters are estimated at a parametric rate in nonlinear models where the conditional distribution of individual effects is left unrestricted.

Lastly, the framework introduced in this paper is also useful to estimate average marginal effects, which are expectations of structural functions relative to the unknown distribution of individual effects. Interesting policy parameters can often be written in this form. For given common parameters, estimating average marginal effects amounts to estimating a linear functional of the distribution function of individual effects. It is well-known that the problem of nonparametrically recovering that distribution from an empirical estimate of the outcome distribution is *ill-posed* in general, in the sense that small sampling errors in the latter possibly translate into large errors in the distribution to be estimated.[6] This problem may also affect the estimation of linear functionals (Goldenshluger and Pereverzev, 2003, Severini and Tripathi, 2007).

To deal with ill-posedness, we use a regularization approach, which trades off an increase in bias (as the regularized solution is approximate) for a decrease in variance (as the regularized problem is well-posed). The average marginal effects estimates that we construct are large-N consistent and asymptotically normal, although their rate of convergence is less than root-N in general. A similar approach can be used to estimate the asymptotic variance of common parameter estimates.

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. The asymptotic properties of estimators of common parameters and average marginal effects are studied in Sections 5 and 6, respectively. Section 7 presents a small-scale numerical illustration. Lastly, Section 8 concludes.

[6] Ill-posedness has generated a large amount of work in applied mathematics, statistics, and more recently in econometrics, mostly in the context of nonparametric instrumental variables models. Recent references on ill-posed inverse problems in econometrics include Darolles *et al.* (2009), Newey and Powell (2003), Hall and Horowitz (2005), Horowitz and Lee (2007), Blundell *et al.* (2007), and Gagliardini and Scaillet (2008).

2 Incidental parameters

In this section, we present the model and outline our approach *via* examples.

2.1 Likelihood models with fixed effects

Let $(y_{it}, x_{it}')'$, $i = 1, ..., N$ and $t = 1, ..., T$ be the set of observations of an endogenous variable y_{it} and a vector of strictly exogenous variables x_{it}, that we assume 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 $y_i = (y_{i1}, ..., y_{iT})$ conditioned on $x_i = (x_{i1}', ..., x_{iT}')$ and a vector of individual-specific parameters α_i is given by $f_{y|x,\alpha;\theta_0}(.|x_i, \alpha_i)$, where $f_{y|x,\alpha;\theta}$ is a known function given $\theta \in \Theta$. The individual effects α_i are i.i.d. draws from an unrestricted conditional distribution $f_{\alpha|x}$ supported on $\mathcal{A}$. The population distribution function of y_i given $x_i = x$ is thus given by:

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

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.[7] Conditional on the effects, however, the model is fully parametric. In addition, the model may incorporate dynamics such as:

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

where $y^{(t)} = (y_t, y_{t-1}, ...)$, in which case x will contain strictly exogenous regressors and initial conditions.

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" $\alpha_1, ..., \alpha_N$, thus leading to fixed-T consistent estimators of common parameters. Our approach is general, and covers all semiparametric likelihood models of the form (1). To facilitate exposition, we will illustrate how the approach works in three panel data models where standard maximum likelihood fails.

[7] In 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.

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

$$y_i = a\left(x_i, \theta_0\right) + B\left(x_i, \theta_0\right)\alpha_i + v_i, \tag{2}$$

where a $(T \times 1)$ and B $(T \times \dim \alpha_i)$ 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.[8]

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

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

where $\Sigma\left(.,.\right)$ 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:

$$f_{y|x,\alpha;\theta}\left(y|x,\alpha\right) = \left(2\pi\right)^{-\frac{T}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(y - a - B\alpha\right)' \Sigma^{-1}\left(y - a - B\alpha\right)\right], \tag{3}$$

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

Example 2: censored random coefficients model. In our second example, latent outcomes follow a normal random coefficients model:

$$y_i^* = a\left(x_i, \theta_0\right) + B\left(x_i, \theta_0\right)\alpha_i + v_i, \tag{4}$$

where $v_i | x_i, \alpha_i \sim N\left[0, \Sigma\left(x_i, \theta_0\right)\right]$. The difference with Example 1 is that $y_{it} = \max\left(y_{it}^*, c_t\right)$ is observed, where we assume that the censoring thresholds $c_1, ..., c_T$ are known to the researcher. In particular, note that (3) still holds in the censored model, for any y such that $y_t > c_t$ for all $t \in \{1, ..., T\}$.

In the model with a single heterogeneous intercept: $y_{it} = \max\left(x_{it}'\beta_0 + \alpha_i + v_{it}, c_t\right)$, 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.[9]

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

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

Example 3: Static binary choice model. Our third 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 any $t \neq s$. Let $F_t(x_{it}, \alpha_i, \theta) = \Pr(y_{it} = 1 | x_{it}, \alpha_i, \theta)$. In this case, $f_{y|x,\alpha;\theta}$ is a conditional probability mass function that satisfies:

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

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).[10]

2.2 Moment restrictions

The methods used to solve the incidental parameter problem in the three examples outlined above 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:

$$[L_{\theta,x} g](y) = \int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y|x,\alpha) g(\alpha) \, d\alpha. \tag{5}$$

The linear integral operator $L_{\theta,x}$ represents the parametric part of the panel data model. It is a central object in this paper.[11] 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 $\varphi(.,x,\theta)$ such that, for any function $g(\alpha)$:

$$\int_{\mathcal{Y}} \varphi(y,x,\theta) [L_{\theta,x} g](y) dy = 0, \tag{6}$$

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

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

where the integral is taken over the support of the data. Equation (6) 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:

$$\begin{aligned}\mathbb{E}\Big[\varphi(y_i, x_i, \theta_0) \,\Big|\, x_i = x\Big] &= \int_{\mathcal{Y}} \varphi(y, x, \theta_0)\, f_{y|x}(y|x)\, dy \\ &= \int_{\mathcal{Y}} \varphi(y, x, \theta_0) \left[L_{\theta_0,x} f_{\alpha|x}\right](y) dy = 0. \end{aligned} \tag{7}$$

As the conditional moment restrictions (7) 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 (6). 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 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 three main examples.

Example 1 (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 $W\Sigma^{-\frac{1}{2}}B = 0$. Note also the identity:

$$\begin{aligned}(y - a - B\alpha)'\Sigma^{-1}(y - a - B\alpha) &= (y - a - B\alpha)'\Sigma^{-\frac{1}{2}}Q\Sigma^{-\frac{1}{2}}(y - a - B\alpha) \\ &\quad + (y - a)'\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y - a).\end{aligned}$$

We have, for any function $g(\alpha)$, and denoting $q = \dim \alpha_i$:[12]

$$\begin{aligned}[L_{\theta,x}g](y) &= \int_{\mathbb{R}^q} f_{y|x,\alpha;\theta}(y|x,\alpha)\, g(\alpha)\, d\alpha \\ &= (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \left\{\int_{\mathbb{R}^q} \exp\left[-\frac{1}{2}(y - a - B\alpha)'\Sigma^{-\frac{1}{2}}Q\Sigma^{-\frac{1}{2}}(y - a - B\alpha)\right] g(\alpha)\, d\alpha\right\} \\ &\quad \times \left\{\exp\left[-\frac{1}{2}(y - a)'\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y - a)\right]\right\}.\end{aligned}$$

So, any function in the range of $L_{\theta,x}$ can be written as:

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

[12]When errors (independent of individual effects) are non-normal, the range of the operator $L_{\theta,x}$ can be conveniently characterized using the Fourier transformation, as we show in the supplementary appendix to this paper.

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

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

$$\int_{\mathbb{R}^T} \varphi(y)\, h\left(Q\Sigma^{-\frac{1}{2}}y\right) \exp\left[-\frac{1}{2}(y-a)'\,\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y-a)\right] dy = 0, \tag{8}$$

then φ will satisfy (6). 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:

$$\mathbb{E}\left[\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y_i - a)\,\middle|\, x_i\right] = 0. \tag{9}$$

Restrictions (9) 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 (9) achieves the semiparametric information bound, using a suitable sample counterpart for the matrix Σ.

Note that in the version of model (2) that imposes normality our approach yields additional moment restrictions. As an example, we also have:

$$\mathbb{E}\left[\left(\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}(y_i-a)(y_i-a)'\,\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}\right) - \Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}\,\middle|\, x_i\right] = 0.$$

Lastly, note that for this approach to have content we need that there exists *some* non-zero function that is orthogonal to the range of $L_{\theta,x}$. This will require the range not to be dense in the whole space of functions of y, according to a certain topology to be defined below. We will refer to this condition as *non-surjectivity*. In the present case, non-surjectivity is satisfied provided $\text{rank}\,(Q) < T$, hence in particular when $T > \dim \alpha_i$.

Example 2 (cont.) In the censored regression model, any function in the range of $L_{\theta,x}$ will satisfy, for some function h and for any $y > c$:[13]

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

[13]By $y > c$ we denote that $y_t > c_t$ for each $t \in \{1, ..., T\}$.

As an interesting special case, let us start with a simple censored regression model with heterogeneous intercept: $y_{it} = \max(x_{it}'\beta_0 + \alpha_i + v_{it}, 0)$, $T = 2$, and v_{it} i.i.d. $N(0, \sigma_0^2)$. Let also $\theta_0 = (\beta_0, \sigma_0^2)$. Then any element in the range of $L_{\theta,x}$ satisfies, for $y_1 > 0, y_2 > 0$:

$$[L_{\theta,x}g](y) = h(\overline{y}, x) \exp\left[-\frac{1}{4\sigma^2}(\Delta y - \Delta x'\beta)^2\right], \tag{10}$$

where $\overline{y} = (y_1 + y_2)/2$, $\Delta y = y_2 - y_1$, and $\Delta x = x_2 - x_1$.

So, any function φ orthogonal to the functions given by (10), 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:

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

and the following function supported on that rectangle:

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

It is easy to see that φ will satisfy (6) if:

$$\int_c^d \varphi_2(\nu) \exp\left[-\frac{1}{4\sigma^2}(\nu - \Delta x'\beta)^2\right] d\nu = 0. \tag{11}$$

In particular, (11) 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).[14] 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 (11) holds. This strategy will also deliver restrictions on σ_0^2.[15]

A similar approach can be used to derive restrictions on θ_0 in the more general random coefficients model with censoring (4). To see how, let us assume for simplicity that $B(x, \theta)$ has full-column rank q, for all θ, almost surely in x. Let V be a $T \times q$ matrix such that $Q = VV'$ and $V'V = I_q$. Let also U be a $T \times (T - q)$ matrix such that $W = UU'$, and $U'U = I_{T-q}$. Lastly, let $(\mu, \nu) = \left(V'\Sigma^{-\frac{1}{2}}y, U'\Sigma^{-\frac{1}{2}}y\right)$.

[14] In 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.

[15] As 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 (11).

Then, let us consider a region in $\mathbb{R}^T$ of the form:

$$\left\{y \in \mathbb{R}^T, \quad (\mu, \nu) \in R_1 \times R_2\right\} \subset \left\{y \in \mathbb{R}^T, \quad y_1 > c_1, ..., y_T > c_T\right\},$$

where R_1 and R_2 are subsets of $\mathbb{R}^q$ and $\mathbb{R}^{T-q}$, respectively. Finally, let us define the following function supported on that Cartesian product:

$$\varphi(y) = \varphi_2(\nu) \mathbf{1}\{\mu \in R_1\} \mathbf{1}\{\nu \in R_2\}.$$

Then, (6) will hold if φ_2 and R_2 are chosen such that:

$$\int_{R_2} \varphi_2(\nu) \exp\left[-\frac{1}{2}\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)\right] d\nu = 0. \tag{12}$$

In particular, if R_2 is chosen such that $\left\{\nu - U'\Sigma^{-\frac{1}{2}}a, \quad \nu \in R_2\right\}$ is symmetric around zero, then:

$$\mathbb{E}\left[U'\Sigma^{-\frac{1}{2}}(y_i - a) \mathbf{1}\left\{V'\Sigma^{-\frac{1}{2}}y_i \in R_1\right\} \mathbf{1}\left\{U'\Sigma^{-\frac{1}{2}}y_i \in R_2\right\} \Big| \, x_i\right] = 0. \tag{13}$$

Restrictions (13) are valid under non-normality, if the distribution of $U'\Sigma^{-\frac{1}{2}}v_i$ is symmetric around the origin.

In this example and the previous one, the range of $L_{\theta,x}$ is a strict subset of the space of T-variate functions, provided $\dim \alpha_i < T$. The above discussion suggests that, in likelihood models with continuous or censored outcomes that satisfy this condition, there exist some functions (in effect, a continuum of functions) that are orthogonal to the range of $L_{\theta,x}$. This paper proposes a systematic way to construct those functions, thus providing moment restrictions on common parameters.

Example 3 (cont.) In static binary choice panel data models, our approach consists in finding a $2^T \times 1$ vector $\left\{\varphi(y, x, \theta), y \in \{0,1\}^T\right\}$, such that:

$$\sum_{y \in \{0,1\}^T} \varphi(y, x, \theta) \Pr(y|x, \alpha, \theta) = 0, \quad x, \alpha - \text{a.s.} \tag{14}$$

that is, such that:

$$\sum_{y \in \{0,1\}^T} \varphi(y, x, \theta) \prod_{t=1}^{T} F_t^{y_t} (1 - F_t)^{1-y_t} = 0, \quad x, \alpha - \text{a.s.} \tag{15}$$

It can be shown that finding a non-zero $\{\varphi(y, x, \theta)\}$ that satisfies (14) is equivalent to all 2^T products of distinct F's being linearly dependent: $F_1^{k_1} \times ... \times F_T^{k_T}$, $(k_1, ..., k_T) \in \{0,1\}^T$ (see the supplementary appendix).

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 $F_t = \Phi\left(x_{it}'\theta + \alpha_i\right)$, with Φ the standard normal cdf. This situation contrasts with Examples 1 and 2, where a condition of non-surjectivity was guaranteed when $T > \dim \alpha_i$.

In contrast, when errors are logistic the situation is very different. In this case, $F_t = \Lambda\left(x_{it}'\theta + \alpha_i\right)$, where $\Lambda(u) = e^u/\left(1+e^u\right)$ is the standard logistic cdf. We show in the supplementary appendix that (15) is equivalent to:

$$\sum_{y\in\{0,1\}^T} \mathbf{1}\left\{\sum_{t=1}^T y_t = s\right\} \varphi(y,x,\theta) e^{\sum_{t=1}^T y_t x_t'\theta} = 0, \qquad \text{for all } s \in \{0,1,...,T\}. \tag{16}$$

This system of equations has non-trivial solutions as soon as $T \geq 2$. For example, if $T = 2$, (16) implies that: $\varphi_{00} = \varphi_{11} = 0$, and:

$$\varphi_{10} e^{x_{i1}'\theta} + \varphi_{01} e^{x_{i2}'\theta} = 0, \tag{17}$$

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:

$$\mathbb{E}\left(e^{[x_{i2}-x_{i1}]'\theta_0} y_{i1}\left[1-y_{i2}\right] - \left[1-y_{i1}\right] y_{i2} \middle| \, x_i\right) = 0, \tag{18}$$

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

Interestingly, Chamberlain (1987)'s optimal unconditional moment restrictions in (18) are:

$$\mathbb{E}\left[\left(x_{i2}-x_{i1}\right)\frac{1}{e^{[x_{i2}-x_{i1}]'\theta_0}+1}\left(e^{[x_{i2}-x_{i1}]'\theta_0} y_{i1}\left[1-y_{i2}\right] - \left[1-y_{i1}\right] y_{i2}\right)\right] = 0. \tag{19}$$

This coincides exactly with the score equations of the conditional maximum likelihood estimator based on the sufficient statistic $y_{i1} + y_{i2}$ (compare with Arellano, 2003).

In non-logistic binary choice models, the information bound for θ_0 is zero (Chamberlain, 2010). The present discussion suggests that those models are surjective, implying that our approach will not yield informative restrictions on θ_0.[16]

[16]This result is related to Johnson (2004) who shows that, in discrete choice panel data models with compactly supported covariates, common parameters are unidentified unless equation (14) holds for some $\varphi \neq 0$ for at least some value of the covariates, and that when (14) 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 The finite-dimensional case

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.

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. 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:

$$f_{y|x} = L_{\theta_0,x} f_{\alpha|x}, \tag{20}$$

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

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

and:

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

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:

$$W_{\theta,x} = I_{N_y} - L_{\theta,x} L_{\theta,x}^\dagger, \tag{21}$$

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 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 $L^{\dagger}_{\theta,x}$ is a generalized inverse, $W_{\theta,x}$ is such that:

$$W_{\theta,x}L_{\theta,x} = L_{\theta,x} - L_{\theta,x}L^{\dagger}_{\theta,x}L_{\theta,x} = 0,$$

and:

$$L'_{\theta,x}W_{\theta,x} = 0.$$

For any given vector $h \in \mathbb{R}^{N_y}$, the vector $\varphi_{\theta,x} = W_{\theta,x}h \in \mathbb{R}^{N_y}$ is thus orthogonal to the columns of $L_{\theta,x}$. 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 $\varphi\left(\underline{y}_s, x, \theta\right)$ denotes the sth element of $\varphi_{\theta,x}$, where $\underline{y}_s$ $(s = 1, ..., N_y)$ index the points of support of y_i, it follows that:

$$\mathbb{E}\left[\varphi\left(y_i, x_i, \theta_0\right) \,\middle|\, x_i\right] = 0. \tag{22}$$

To interpret our approach, note that the moment restrictions are obtained by left-multiplying (20) 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:

$$W_{\theta_0,x}f_{y|x} \;\; = \;\; 0. \tag{23}$$

The functional differencing restrictions (23) 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 $W_{\theta,x}h = h - L_{\theta,x}L^{\dagger}_{\theta,x}h$ is the least-squares residual in the linear regression of a vector $h \in \mathbb{R}^{N_y}$ on the columns of the matrix $L_{\theta,x}$. 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 (22) can be obtained as residuals in a linear regression. Bajari *et al.* (2009) use a related idea in a game-theoretic context.

Identification. In the finite support case, the functional differencing restrictions can be equivalently written as a system of N_y conditional moment restrictions. To see why, let $\tau\left(y_i\right)$

denote the index in $\{1, ..., N_y\}$ such that $y_i = \underline{y}_{\tau(y_i)}$. Let also $\omega_{\theta,x}[s_1, s_2]$ denote the (s_1, s_2)th element of the matrix $W_{\theta,x}$. Then, (23) can equivalently be written as:[17]

$$\mathbb{E}\left(\omega_{\theta_0,x_i}[s, \tau(y_i)] \,\middle|\, x_i\right) = 0, \quad s \in \{1, ..., N_y\}. \tag{24}$$

The next result provides necessary and sufficient conditions for θ_0 to be point-identified from the moment restrictions (24).

Proposition 1 *i) Suppose that θ_0 is point-identified from (24). Then, for all $\theta \neq \theta_0$ in Θ, $\text{rank}(L_{\theta,x}) < N_y$ with positive probability in x.*

ii) If, for any $\theta \neq \theta_0$, $f_{y|x}$ does not belong to the range of $L_{\theta,x}$ with positive probability in x, then θ_0 is point-identified from (24).

Part *i*) shows that a condition of *non-surjectivity* is necessary for θ_0 to be point-identified from the functional differencing restrictions. The moment restrictions (24) are uninformative about θ_0 when the rows of $L_{\theta,x}$ are linearly independent, i.e., when $\text{rank}(L_{\theta,x}) = N_y$. In this case, the range of $L_{\theta,x}$ is the full space $\mathbb{R}^{N_y}$. For example, if $L_{\theta,x}$ is square and non-singular then the Moore-Penrose inverse coincides will the standard matrix inverse, and $W_{\theta,x} = I_{N_y} - L_{\theta,x}L_{\theta,x}^{-1} = 0$. Note that non-surjectivity is automatically satisfied when $N_y > N_\alpha$.

Part *ii*) in the proposition provides a sufficient condition for identification. In practice, it may be easier to verify local identification using a rank condition. To state the condition, we define the following $N_y \times \dim\theta$ Jacobian matrix:

$$G(x) = \left[\left(\mathbb{E}\left(W_{\theta_0,x_i}\frac{\partial L_{\theta_0,x_i}}{\partial\theta_k}L^{\dagger}_{\theta_0,x_i}[., \tau(y_i)] \,\middle|\, x_i = x\right)\right)_k\right],$$

where $L^{\dagger}_{\theta_0,x_i}[., s]$ denotes the sth column of $L^{\dagger}_{\theta_0,x_i}$, and where:

$$\frac{\partial L_{\theta_0,x_i}}{\partial\theta_k} = \left[\left(\frac{\partial \Pr\left(y_i = \underline{y}_s | x_i, \alpha_i = \underline{\alpha}_n; \theta_0\right)}{\partial\theta_k}\right)_{s,n}\right]$$

is an $N_y \times N_\alpha$ matrix of derivatives.

Proposition 2 *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 $G(x)\,a = 0$ almost surely in x implies $a = 0$, then θ_0 is locally point-identified from (24).*

[17] Note that (24) can be obtained by taking $\varphi_{\theta,x} = W_{\theta,x}e_s$, where $\{e_s\}$ is the canonical basis in $\mathbb{R}^{N_y}$.

Example 3 (cont.) The non-surjectivity condition is satisfied in static binary choice models provided that $N_\alpha < 2^T$. When α_i has more than 2^T points of support the condition will not be satisfied in general. Note that θ_0 may be point-identified even if non-surjectivity does not hold. An example is a static probit model with $N_\alpha \geq 2^T$, when one of the regressors has unbounded support.

Also, θ_0 may fail to be point-identified 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, as the rank condition of Proposition 2 is violated.

3.2 Method-of-moments estimation

Motivated by (24), we propose to estimate θ_0 using the following generalized method-of-moments (GMM) estimator, which relies on a set of $R \geq 1$ unconditional moment restrictions:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \upsilon_{r_1,r_2} \widehat{\mathbb{E}} \left[\varphi_{r_1} \left(y_i, x_i, \theta \right) \right] \widehat{\mathbb{E}} \left[\varphi_{r_2} \left(y_i, x_i, \theta \right) \right], \tag{25}$$

where

$$\varphi_r \left(y_i, x_i, \theta \right) = \omega_{\theta, x_i} \left[s_r, \tau \left(y_i \right) \right] \zeta_r \left(x_i \right), \tag{26}$$

and where $\widehat{\mathbb{E}} \left(z_i \right) = \frac{1}{N} \sum_{i=1}^{N} z_i$ denotes an empirical mean, $\zeta_1, ..., \zeta_R$ are functions of covariates, $s_1, ..., s_R$ are indexes in $\{1, ..., N_y\}$, and $\Upsilon = \left[\upsilon_{r_1,r_2} \right]_{(r_1,r_2) \in \{1,...,R\}^2}$ is a symmetric weighting matrix.

Under standard identification and regularity conditions (e.g., Theorems 2.6 and 3.2 in Newey and McFadden, 1994), $\widehat{\theta}$ is root-N consistent and asymptotically normal for θ_0. The asymptotic results derived in Section 5 cover finite supports as a special case, so we refer the reader to that section for the expression of the asymptotic variance of $\widehat{\theta}$.

Note that the standard regularity conditions in GMM require the moment functions to be continuous in θ. In the present case, in addition to the continuity of $\theta \mapsto f_{y|x,\alpha;\theta} \left(y | x, \alpha \right)$, this requires that the rank of $L_{\theta,x}$ be constant on the parameter space Θ, almost surely in x.[18] Rank constancy is intuitively necessary to ensure the continuity of a projection matrix,

[18] Indeed, Corollary 3.5 in Stewart (1977) then shows that $\theta \mapsto L_{\theta,x}^{\dagger}$ is continuous on Θ, a.s in x, implying the continuity of the within projection matrix $W_{\theta,x}$.

as variations in the rank of $L_{\theta,x}$ induce jumps in its number of non-zero eigenvalues. In particular, rank constancy will be satisfied if $L_{\theta,x}$ has almost surely full column rank.

Efficiency. We now derive Chamberlain (1987)'s optimal instruments 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.

Let $\kappa_{\theta_0,x} = \mathbb{E}\left(U_{\theta_0,x_i}\left[\tau\left(y_i\right),.\right]' U_{\theta_0,x_i}\left[\tau\left(y_i\right),.\right]|x_i = x\right)$, where $U_{\theta_0,x_i}\left[s,.\right]$ denotes the sth row of U_{θ_0,x_i}. The optimal instruments derived in Appendix A yield the optimal unconditional moment restrictions:

$$\mathbb{E}\left(U_{\theta_0,x_i}\left[\tau\left(y_i\right),.\right]\kappa^{-1}_{\theta_0,x_i}\mathbb{E}\left[U'_{\theta_0,x_i}\frac{\partial L_{\theta_0,x_i}}{\partial\theta_k}L^{\dagger}_{\theta_0,x_i}\left[.,\tau\left(y_i\right)\right]\,\middle|\, x_i\right]\right) = 0, \quad k = 1,...,\dim\theta. \quad (27)$$

The optimal instruments in (27) are infeasible. To construct feasible counterparts and estimate θ_0 efficiently on the basis of the functional differencing restrictions (24), one can follow the approach in Newey (1990) and replace the unknown conditional expectations by series estimators.

It is interesting to know whether the functional differencing approach is efficient. The next result shows that estimating θ_0 using the optimal functional differencing restrictions attains the information bound of the panel data model. We prove the result for finitely supported regressors.

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), $\theta \mapsto f_{y|x,\alpha;\theta}\left(y|x,\alpha\right)$ is continuously differentiable in a neighborhood of θ_0. Then the semiparametric information bound for θ_0 in the panel data model coincides with the GMM bound based on the moment restrictions (27).*

From the theory of semiparametric efficiency bounds (e.g., Bickel *et al.*, 1993), the efficient score is obtained as a residual in the linear regression of the score relative to θ on the nonparametric tangent space. In the present case, and provided $f_{\alpha|x} > 0$ (excluding corner cases),[19] the tangent space is equal to the set of score vectors that belong to the range of the matrix $L_{\theta,x}$, scaled by the inverse of $f_{y|x}$ (see Appendix A). As a result, the efficient score is obtained as a particular functional differencing projection.

[19]Note that when the probability distribution of individual effects is not identified, this assumption requires that at least one pseudo-true $f_{\alpha|x}$ be interior.

As a special case, when $W_{\theta,x} = 0$ then the GMM information bound based on the functional differencing restrictions is zero. We thus have the following result.

Corollary 1 *Suppose in addition to the assumptions of Theorem 1 that the range of $L_{\theta,x}$ coincides with $\mathbb{R}^{N_y}$, for all $\theta \in \Theta$ and x. Then the semiparametric information bound for θ_0 is equal to zero.*

In particular, Proposition 1 implies that, in models where the non-surjectivity condition does not hold, there exists no root-N consistent estimator of θ_0 (Chamberlain, 1987).

3.3 Extension: exploiting inequality constraints

The functional differencing approach does not use the fact that the unknown probabilities of individual effects lie in the unit interval. In surjective models– where the information bound for θ_0 is zero– exploiting this additional information will be 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).

To outline an extension that exploits those additional constraints, let $\mathcal{S}$ denote the unit simplex in $\mathbb{R}^{N_\alpha}$, and let us define the following projection, for any given $h \in \mathbb{R}^{N_y}$:

$$Q^+_{\theta,x}(h) = \underset{\widetilde{h} \in L_{\theta,x}(\mathcal{S})}{\operatorname{argmin}} \quad \left(\widetilde{h} - h\right)' \left(\widetilde{h} - h\right). \tag{28}$$

Let us also define the *constrained* within projection as $W^+_{\theta,x} = I_{N_y} - Q^+_{\theta,x}$. To see the link with the within projection matrix introduced above, note that $W_{\theta,x}$ is obtained if we replace $L_{\theta,x}(\mathcal{S})$ in (28) by the range $L_{\theta,x}\left(\mathbb{R}^{N_\alpha}\right)$ of the matrix $L_{\theta,x}$.

The following result provides a characterization of *all* the restrictions implied by the panel data model.

Theorem 2 *The two following statements are equivalent.*

i) There exists some $f_{\alpha|x} \in \mathcal{S}$ such that, almost surely in x, $f_{y|x} = L_{\theta_0,x} f_{\alpha|x}$.

ii) $W^+_{\theta_0,x}\left(f_{y|x}\right) = 0$ almost surely in x.

Theorem 2 shows that the restrictions implied by the panel data model are equivalent to a set of constrained functional differencing restrictions. In particular, in models where

θ_0 is partially identified, part ii) in the theorem characterizes the sharp identified region for common parameters. Using those constrained restrictions for estimation and inference has intuitive appeal, as they may be informative about θ_0 in cases where the within projection matrix $W_{\theta,x}$ is zero.

However, using those restrictions is not direct as, because of the constraints, $W^+_{\theta_0,x}$ is not a matrix but a nonlinear function. In particular, it does not seem possible to write the constrained functional differencing restrictions as a set of conditional moment restrictions. A proper treatment of the difficulties that arise in this extension is left to future work.

4 The infinite-dimensional case

In this section and the next, we provide a generalization of the functional differencing approach to the case where α_i and (possibly) y_i have infinite support, in which case $L_{\theta,x}$ becomes a linear integral operator.

4.1 Linear operators

Let $\mathcal{A} \subset \mathbb{R}^q$ and $\mathcal{Y} \subset \mathbb{R}^T$ denote the supports of α_i and y_i, respectively, where q is the dimension of the vector of individual effects and T is the number of time periods. Let also $\mathcal{X}$ denote the support of x_i. The analysis in this subsection is conditional on 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:[20]

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

The spaces of functions $\mathcal{G}_\alpha$ 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 $\|g\| = \langle g, g \rangle^{\frac{1}{2}}$.

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 (5). The operator $L_{\theta,x}$ can

[20] Note that π_α and π_y may depend on x, which is kept fixed in this subsection, although we omit the x subscript for conciseness.

be understood as an infinite-dimensional analog of the matrix of conditional probabilities introduced in the previous section.

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. Similarly as in Carrasco and Florens (2009), we assume that π_α and π_y are chosen such that the operator $L_{\theta,x}$ is *compact*, as stated in the following assumption.

Assumption 1 *The two following statements hold true:*

i)

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

ii)

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

Part *i)* in Assumption 1 restricts the distribution of individual effects. For example, if $\pi_\alpha = 1$ then $f_{\alpha|x}$ must be square integrable with respect to the Lebesgue measure. For this, $f_{\alpha|x}$ being bounded is sufficient, as in this case:

$$\int_{\mathcal{A}} f_{\alpha|x}(\alpha|x)^2 d\alpha \;\leq\; \left(\sup_{\mathcal{A}} f_{\alpha|x}\right) \underbrace{\int_{\mathcal{A}} f_{\alpha|x}(\alpha|x) d\alpha}_{=1} \;< \infty.$$

Note that $\int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y|x,\alpha)^2/\pi_\alpha(\alpha) d\alpha$ 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 1 when $\mathcal{A} = \mathbb{R}^q$, $\pi_\alpha = 1$, and π_y is integrable, while it does not hold when $\pi_y = 1$. Moreover, *ii)* will automatically hold if $\pi_\alpha = 1$, π_y is integrable, and $f_{y|x,\alpha;\theta}(y|x,\alpha)$ is bounded by a constant $C_{\theta,x}$, provided the support $\mathcal{A}$ of individual effects is bounded in $\mathbb{R}^q$ (given x).

Part *ii)* in Assumption 1 ensures that $L_{\theta,x} g \in \mathcal{G}_y$ for any function $g \in \mathcal{G}_\alpha$, and that the operator $L_{\theta,x} : \mathcal{G}_\alpha \to \mathcal{G}_y$ is Hilbert-Schmidt, hence compact (Theorem 2.32 in Carrasco *et al.*, 2008). Compactness will not be needed to derive the functional differencing moment restrictions below. However, the compactness assumption will be useful for deriving the asymptotic theory of our estimator, as it will allow us to rely on the singular value decomposition of the operator $L_{\theta,x}$.

4.2 Moment restrictions

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

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

and let $\overline{\mathcal{R}\left(L_{\theta,x}\right)}$ denote its closure in $\mathcal{G}_y$, according to the Hilbert space topology.

Next, let us define the "within" projection operator as:

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

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

$$\text{Proj}_{\pi_y}\left[h \middle| \overline{\mathcal{R}\left(L_{\theta,x}\right)} \right] = \underset{\widetilde{h} \in \overline{\mathcal{R}\left(L_{\theta,x}\right)}}{\text{argmin}} \int_{\mathcal{Y}} \left(\widetilde{h}\left(y\right) - h\left(y\right) \right)^2 \pi_y\left(y\right) dy. \tag{29}$$

It follows from the theory of linear operators on Hilbert spaces[21] that the linear operator $W_{\theta,x}$ has domain $\mathcal{G}_y$, 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 $L_{\theta,x}$. 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 the previous section.

The next theorem provides the key restrictions on θ_0.

Theorem 3 *Let Assumption 1 hold. Then the two following equivalent conditions are satisfied:*

$$W_{\theta_0,x} f_{y|x} = 0, \quad \textit{or, equivalently} \tag{30}$$

$$\mathbb{E}\left(\pi_y(y_i) \left[W_{\theta_0,x_i} h \right](y_i) \middle| \, x_i \right) = 0, \quad \textit{for all } h \in \mathcal{G}_y. \tag{31}$$

Theorem 3 provides a set of restrictions on θ_0. As in the finite-dimensional case described in Section 3, those restrictions are obtained using a functional differencing approach that differences out the distribution of individual effects. Moreover, through (31), those restrictions are equivalently written as a set of conditional moment restrictions, leading the way to estimation. Note that the distribution function $f_{y|x}$ enters (31) only *via* the expectation.[22]

[21] We refer to Kress (1989) and Engl, Hanke, and Neubauer (2000) for proofs of the statements in this section and additional background.

[22] Note also that, although the present paper assumes that θ is finite-dimensional, the restrictions of Theorem 3 are still valid if θ is infinite-dimensional.

Singular value decomposition. To provide some insight about the within projection operator, it is convenient to work with the singular value decomposition of the compact operator $L_{\theta,x}$:

$$L_{\theta,x}g = \sum_j \phi_j \lambda_j \left\langle \psi_j, g \right\rangle, \text{ for all } g \in \mathcal{G}_\alpha, \tag{32}$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 ... > 0$ is a sequence of positive real numbers, $\psi_1, \psi_2, ...$ is an orthonormal sequence in $\mathcal{G}_\alpha$, and $\phi_1, \phi_2, ...$ is an orthonormal sequence in $\mathcal{G}_y$. The sum in (32) ranges from 1 to the (possibly infinite) rank of $L_{\theta,x}$. Note that, when $L_{\theta,x}$ is singular, $\{\psi_j\}$ and $\{\phi_j\}$ do not generally span $\mathcal{G}_\alpha$ and $\mathcal{G}_y$, respectively. Note also that λ_j, ψ_j, and ϕ_j, $j = 1, 2, ...$, depend on θ and x, which are kept fixed in the present discussion.

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

$$W_{\theta,x}h \;\; = \;\; h - \sum_j \phi_j \left\langle \phi_j, h \right\rangle, \text{ for all } h \in \mathcal{G}_y. \tag{33}$$

Note that the singular values and functions, and thus the within projection operator $W_{\theta,x}$, do not depend on the distribution of the data. Although they are generally not available in closed form, they can be approximated using a simple discretization strategy that we will describe in Section 7.

It is interesting to contrast the expression in (33) with that of the Moore-Penrose inverse of $L_{\theta,x}$, which is defined on a domain $\mathcal{D} \subset \mathcal{G}_y$ by:

$$L^\dagger_{\theta,x}h \;\; = \;\; \sum_j \psi_j \frac{1}{\lambda_j} \left\langle \phi_j, h \right\rangle, \text{ for all } h \in \mathcal{D}. \tag{34}$$

In the infinite-dimensional case, the Moore-Penrose inverse $L^\dagger_{\theta,x}$ 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, $h \mapsto L^\dagger_{\theta,x}h$ is not continuous, and $L^\dagger_{\theta,x}h$ is very sensitive to any noise in h possibly arising in estimation. 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 Subsection 3.1, $L^\dagger_{\theta,x}h$ 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.

Non-surjectivity. The following proposition establishes that non-surjectivity is necessary for θ_0 to be point-identified from the functional differencing restrictions (30).

Proposition 3 *Let Assumption 1 hold, and suppose that θ_0 is globally identified from (30). Then, for all $\theta \neq \theta_0$ in Θ, $\overline{\mathcal{R}(L_{\theta,x})} \neq \mathcal{G}_y$ with positive probability in x.*

As in the finite support case, non-surjectivity is necessary for θ_0 to be point-identified from the functional differencing restrictions, but it is not sufficient. In analogy with simultaneous equations models, non-surjectivity may be understood as an *order* condition for identification. As a complement, the asymptotic analysis in the next section will highlight *rank* conditions, that will ensure that θ_0 is locally point-identified.

As anticipated in Section 2, it can be formally shown that the non-surjectivity condition is satisfied in the random coefficients model (Example 1) and the censored coefficients model (Example 2) with normal errors, provided that $T > \dim \alpha_i$.[23] Non-surjectivity is not satisfied in the static probit model (although it is satisfied in the static logit model).

In the supplementary appendix to this paper, we study non-surjectivity in two additional models: a random coefficients model and a nonlinear regression model with independent additive errors (possibly non-normal). In those examples, we derive closed-form restrictions on θ_0 that involve the characteristic function of time-varying errors. For those restrictions to be informative, $T > \dim \alpha_i$ is sufficient in the random coefficients model, though not in the nonlinear regression model. In this case, non-surjectivity requires that the image of the regression function be non-dense in $\mathbb{R}^T$. When $T > \dim \alpha_i$, this rules out space-filling mappings such as Peano curves (surjective mappings from $\mathbb{R}$ onto $\mathbb{R}^2$).

Those various examples lead us to conjecture that, in models where the dependent variables are continuous, and provided that T be strictly larger than the number of individual effects, the non-surjectivity condition should generally be satisfied.

[23] When $T = \dim \alpha_i$ and B is non-singular, $Q = I_T$ and $W = 0$ in equation (8), and the non-surjectivity condition is *not* satisfied in Example 1.

4.3 Method-of-moments estimation

Let $\{y_i, x_i\}_{i=1,...,N}$ be a random sample. Motivated by the conditional moment restrictions (31) of Theorem 3, we propose to estimate θ_0 by minimizing a GMM criterion of the form (25), with moment functions:

$$\varphi_r(y_i, x_i, \theta) = \pi_y(y_i) [W_{\theta,x_i} h_r](y_i) \zeta_r(x_i), \tag{35}$$

where $h_1, ..., h_R$ are elements of $\mathcal{G}_y$, and $\zeta_1, ..., \zeta_R$ are functions of covariates x_i.

Under regularity conditions given in the next section (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. In the supplementary appendix to this paper, we use this insight to construct a nonlinear analog of the Hausman (1978) specification test for parametric random-effects models.

Turning to the choice of functions h_r and ζ_r in (35), one approach is to choose a finite family h_r, $r = 1, ..., R$, 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 7 we have set $h_r(y) = \phi(y - \mu_r)$, with ϕ the standard normal pdf, and $\mu_1, ..., \mu_R$ 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 so as to maximize efficiency, for example using suitable empirical counterparts of Chamberlain (1987)'s optimal instruments, given a choice of functions h_r.

Optimal moment restrictions. A different approach to the choice of functions h_r is to derive the *optimal* combination of the moment restrictions that functional differencing delivers. When supports are finite, (27) provides the optimal unconditional moment restrictions. When supports are infinite, one can follow the approach in Carrasco and Florens (2000) to construct a finite-dimensional set of optimal instrument functions $h_k^{\text{opt}} \in \mathcal{G}_y$, $k \in \{1, ..., \dim \theta\}$. The expression of the instrument functions given in Appendix A is similar to (27), with U_{θ_0,x_i}, L_{θ_0,x_i}, and κ_{θ_0,x_i} being linear operators. Then, estimation of θ_0 may be

based on the following $\dim\theta$ unconditional moment restrictions:

$$\mathbb{E}\Big(\pi_y(y_i)\left[U_{\theta_0,x_i}h_k^{\text{opt}}\right](y_i)\Big)=0,\quad k=1,...,\dim\theta.$$

To construct feasible counterparts of the optimal instrument functions, the covariance operator κ_{θ_0,x_i} must be *regularized* (Carrasco and Florens, 2000). In addition, a regularization of $L^{\dagger}_{\theta_0,x_i}f_{y|x}$ (which appears in the expression of h_k^{opt} given in the appendix) is also needed. We will use a regularization approach to estimate average marginal effects in Section 6.

5 Asymptotic properties

In this section, we study the asymptotic properties of the estimator of common parameters, which we write in a more compact form as:[24]

$$\widehat{\theta}=\underset{\theta}{\text{argmin}}\quad\widehat{\mathbb{E}}\left[\varphi\left(y_i,x_i,\theta\right)'\right]\Upsilon\widehat{\mathbb{E}}\left[\varphi\left(y_i,x_i,\theta\right)\right],$$

where the moment functions are given by:

$$\varphi\left(y_i,x_i,\theta\right)\;=\;\left[\begin{array}{c}\pi_y(y_i)\left[W_{\theta,x_i}h_1\right](y_i)\,\zeta_1\left(x_i\right)\\ ...\\ \pi_y(y_i)\left[W_{\theta,x_i}h_R\right](y_i)\,\zeta_R\left(x_i\right)\end{array}\right].\tag{36}$$

To state the assumptions we will need some additional notation. We denote the *norm* of a bounded operator as $\|L\|=\max_{\|h\|\leq1}\|Lh\|/\|h\|$. We also denote the *adjoint* of an operator $L:\mathcal{G}_1\to\mathcal{G}_2$ 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.

5.1 Consistency

We make the following assumptions that ensure the consistency of $\widehat{\theta}$ as N tends to infinity. 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 *The following statements hold true.*

i) Θ *is compact.*

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

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

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

iv) Almost surely in x*:*

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

Moreover, for any $r = 1, ..., R$*:*

v) For any j*:*

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

vi) Almost surely in x*:*

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

vii)

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

viii)

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

ix)

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

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.

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. A sufficient condition for $L_{\theta,x}$ to have constant rank is that it is injective, i.e. that $g = 0$ is the only solution to $L_{\theta,x}g = 0$.[25] 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.[26]

[25] As an example, rank constancy is also satisfied in the static logit model: when $T = 2$, the null-space of $L^*_{\theta,x}$ has dimension 1, so $\text{rank}(L_{\theta,x}) = 2^T - 1 = 3$ irrespective of (θ, x).

[26] This is because the function $\theta \mapsto \lambda_{j,\theta,x}$ is continuous on Θ. See Theorem 15.17 in Kress (1989).

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. We will invoke source conditions when studying the asymptotic properties of average marginal effects, but we do not need them for the estimation of common parameters.

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 1) with known error variance. In addition, in Section 7 we provide numerical evidence supporting uniform Fourier convergence in the two simple models that we use as illustrations.

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:

$$\Sigma(\theta) = \mathbb{E}\left[\varphi(y_i, x_i, \theta)\, \varphi(y_i, x_i, \theta)'\right]. \tag{37}$$

This property will be useful to derive the asymptotic distribution of $\widehat{\theta}$.[27] This implies that there is no need to regularize the estimates of the moment functions.

Finally, Conditions *viii*) and *ix*) are moment existence conditions.

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

Theorem 4 *Let Assumptions 1 and 2 hold. Then* $\widehat{\theta} \xrightarrow{p} \theta_0$.

5.2 Asymptotic normality

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

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

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

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

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

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

For any $r = 1,...,R$:

iii) Almost surely in x:

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

iv)

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

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

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

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

vi) As N tends to infinity:

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

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

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

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

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

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

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 $\widehat{\theta}$.

Theorem 5 *Let the assumptions of Theorem 4 be satisfied and let Assumption 3 hold. Then:*

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, \left(G'\Upsilon G\right)^{-1} G'\Upsilon\Sigma\left(\theta_0\right)\Upsilon G\left(G'\Upsilon G\right)^{-1}\right]. \tag{38}$$

Remark 1. The proof of Theorem 5 does not require the empirical moment functions $\theta \mapsto \widehat{\mathbb{E}}\left[\varphi\left(y_i, x_i, \theta\right)\right]$ to be continuous. In practice, as outlined in Section 7 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 $\widehat{\theta}$, we need to compute consistent estimates of Σ and G. The outer product Σ is readily estimated as:

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

In contrast, the Jacobian matrix G involves the unbounded Moore-Penrose inverse $L^{\dagger}_{\theta_0,x_i}$. Interestingly, the matrix G can be interpreted as an average marginal effect. The analysis in the next section will show that the problem of estimating G, and hence the *variance* of the common parameter estimates $\widehat{\theta}$, may be *ill-posed*. A simple truncated estimate that relies on the singular value decomposition (34) is:

$$\widehat{G} = \left[\left(-\widehat{\mathbb{E}}\left(\sum_{j=1}^{J}\pi_y\left(y_i\right)\phi_{j,\widehat{\theta},x_i}\left(y_i\right)\frac{1}{\lambda_{j,\widehat{\theta},x_i}}\left\langle\frac{\partial L^{*}_{\widehat{\theta},x_i}}{\partial\theta}W_{\widehat{\theta},x_i}h_r, \psi_{j,\widehat{\theta},x_i}\right\rangle\zeta_r\left(x_i\right)\right)\right)_{r,k}\right]. \tag{39}$$

Below we provide conditions (on the rate of convergence of J as a function of the sample size) under which $\widehat{G}$ is consistent for G.[28]

6 Average marginal effects

In this section, we study average marginal effects, or policy parameters, of the form:

$$M = \mathbb{E}\left[m\left(x_i, \alpha_i\right)\right],$$

where $m()$ is a known function. We focus on scalar marginal effects to simplify the notation, although our approach could easily be extended to vector-valued $m()$. Average marginal effects are often of interest in applications. Examples include the average effect of a covariate on a conditional mean, or moments of individual fixed effects.

6.1 Identification

Let us denote $M = \mathbb{E}\left[M\left(x_i\right)\right]$, where

$$M(x) = \int_{\mathcal{A}} m\left(x, \alpha\right) f_{\alpha|x}(\alpha|x) d\alpha.$$

In the following we assume that θ_0 is point-identified. Moreover, we suppose that m/π_α belongs to $\mathcal{G}_\alpha$, so that $M(x)$ is well-defined.

The distinctive feature of average marginal effects is that they involve the unknown distribution of individual effects. This distribution may be not point-identified for fixed T. The next result gives a condition for $M(x)$ to be identified.

Proposition 4 *Suppose that Assumption 1 holds. Suppose also that θ_0 is point-identified. Then $M(x)$ is point-identified if:*

$$\frac{m}{\pi_\alpha} \in \overline{\mathcal{R}\left(L^*_{\theta_0,x}\right)}, \tag{40}$$

where $L^*_{\theta_0,x}$ is the adjoint (or transpose) of $L_{\theta_0,x}$. Moreover, in this case:

$$M(x) = \left\langle \frac{m}{\pi_\alpha}, L^\dagger_{\theta_0,x} f_{y|x} \right\rangle. \tag{41}$$

Proposition 4 gives a sufficient condition for $M(x)$ to be point-identified. The reason why this condition is not necessary is that (40) does not take into account that the distribution

[28] Given 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.

function of individual effects $f_{\alpha|x}$ is non-negative. Note that (40) holds obviously when $L_{\theta_0,x}$ is *injective*, i.e. when the unique solution to $L_{\theta_0,x}g = 0$ is $g = 0$. Intuitively, in that case the distribution of individual effects can be uniquely recovered from the data, so any marginal effect is point-identified. In non-injective models, average marginal effects may be partially identified, as happens in models with binary dependent variables (e.g., Chernozhukov *et al.*, 2009).

Examples 1 and 2 (cont.) In Chamberlain (1992)'s random coefficients model with normal errors, and in the censored random coefficients model with normal errors, a necessary and sufficient condition for $L_{\theta,x}$ to be injective is that $\text{rank}\,(B) = \dim \alpha_i$ (see the supplementary appendix). When this happens almost surely in x, any average marginal effect is point-identified.

Example 3 (cont.) In the static logit model, the only $M(x)$ that are identified by Proposition 4 are averages of the form $M(x) = \mathbb{E}\left[h\left(y_i\right)|x_i = x\right]$, where $h \in \mathcal{G}_y$.[29]

6.2 Estimation

We now propose a method to estimate an average marginal effect $M = \mathbb{E}\left[M\left(x_i\right)\right]$. For this we suppose that the identification conditions of Proposition 4 are satisfied.

Combining (41) the singular value decomposition of the Moore-Penrose generalized inverse $L^{\dagger}_{\theta_0,x}$ given by equation (34), we obtain:

$$\begin{aligned} M &= \mathbb{E}\left[\sum_j \left\langle \phi_{j,x_i}, f_{y|x}\right\rangle \frac{1}{\lambda_{j,x_i}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right] \\ &= \mathbb{E}\left[\sum_j \pi_y\left(y_i\right)\phi_{j,x_i}\left(y_i\right) \frac{1}{\lambda_{j,x_i}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right], \end{aligned} \tag{42}$$

where singular values and functions also depend on θ_0, which we omit for conciseness.

The presence of $1/\lambda_{j,x_i}$ in (42) implies that, unless the function m is sufficiently smooth (in a sense to made precise below), replacing the expectation by an empirical mean in (42) delivers an inconsistent estimate of M as the sample size tends to infinity. Building on the literature on ill-posed inverse problems, we propose to estimate M by the following

[29] The reason is that, in this case, $\mathcal{R}\left(L^*_{\theta_0,x}\right)$ is finite-dimensional, so it is closed in $\mathcal{G}_\alpha$. So, (40) holds if and only if $\frac{m}{\pi_\alpha} \in \mathcal{R}\left(L^*_{\theta_0,x}\right)$.

regularized estimate:

$$\widehat{M}_{\delta_N} = \widehat{\mathbb{E}}\left[\sum_j q_{j,x_i}(\delta_N)\,\pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right], \tag{43}$$

where δ_N tend to zero as N tends to infinity at a rate to be specified, and where q_{j,x_i} is a regularization scheme that decreases the contribution of those terms for which λ_{j,x_i} is small, hence $1/\lambda_{j,x_i}$ is large, in the sum. Here also, all singular values and singular functions are computed at θ_0. In practice, θ_0 is replaced by a root-N consistent estimate $\widehat{\theta}$. As the final rate of convergence in Theorem 6 below is slower than root-N, this will not affect the asymptotic distribution of $\widehat{M}_{\delta_N}$.

Specifically, we choose $q_{j,x}$ such that, almost surely in x, there exists a constant $a_x > 0$ such that:

$$\begin{cases} |q_{j,x}(\delta)| \leq a_x \frac{\lambda_{j,x}^2}{\delta}, \\ \lim_{\delta\to 0} q_{j,x}(\delta) = 1. \end{cases} \tag{44}$$

An important example of regularization scheme is *Tikhonov* regularization, in which case

$$q_{j,x}(\delta) = \frac{\lambda_{j,x}^2}{\lambda_{j,x}^2 + \delta}. \tag{45}$$

Other popular regularization schemes that satisfy (44) are spectral cut-off and Landweber-Fridman. See Section 3.3 in Carrasco *et al.* (2008).

Let:

$$m_{i,\delta_N} = \sum_j q_{j,x_i}(\delta_N)\,\pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle. \tag{46}$$

We have:

$$\widehat{M}_{\delta_N} - M = \underbrace{\left[\widehat{\mathbb{E}}\left(m_{i,\delta_N}\right) - \mathbb{E}\left(m_{i,\delta_N}\right)\right]}_{A_N} + \underbrace{\left[\mathbb{E}\left(m_{i,\delta_N}\right) - M\right]}_{B_N}.$$

The term B_N is responsible for the asymptotic bias of $\widehat{M}_{\delta_N}$, while A_N is related to its asymptotic variance. To derive the asymptotic properties of $\widehat{M}_{\delta_N}$ we make the following assumptions.

Assumption 4 *The following conditions hold.*

i) There exist $\beta_y \geq 0$ and $\beta_m \geq 0$ such that $\beta_y + \beta_m > 1$ and, a.s. in x:

$$C_y(x) \equiv \sum_j \frac{\left\langle \phi_{j,x}, f_{y|x}\right\rangle^2}{\lambda_{j,x}^{2\beta_y}} < \infty, \tag{47}$$

$$C_m(x) \equiv \sum_j \frac{\left\langle \psi_{j,x}, \frac{m}{\pi_\alpha}\right\rangle^2}{\lambda_{j,x}^{2\beta_m}} < \infty. \tag{48}$$

ii)

$$\sqrt{N}\delta_N \mathbb{E}\left[\left(\sup_j \left|\lambda_{j,x_i}^{\beta_y+\beta_m-1}\left(q_{j,x_i}\left(\delta_N\right)-1\right)\right|\right) C_y\left(x_i\right)^{\frac{1}{2}} C_m\left(x_i\right)^{\frac{1}{2}}\right] \overset{N\to\infty}{\to} 0. \tag{49}$$

iii)

$$\mathbb{E}\left[\left(\sup_j \left|\frac{\delta_N q_{j,x_i}\left(\delta_N\right)}{\lambda_{j,x_i}}\right|^2\right) \sup_{y\in\mathcal{Y}}\left(f_{y|x}\left(y|x_i\right)\pi_y\left(y\right)\right)\left\|\frac{m}{\pi_\alpha}\right\|^2\right] < \infty.$$

iv)

$$\mathbb{E}\left[\left(\sup_j \left|\frac{\delta_N q_{j,x_i}\left(\delta_N\right)}{\lambda_{j,x_i}}\right|^2\right) \left\|f_{y|x}\right\|^2\left\|\frac{m}{\pi_\alpha}\right\|^2\right] < \infty.$$

v) As N tends to infinity:

$$\sqrt{N}\delta_N\left[\widehat{\mathbb{E}}\left(m_{i,\delta_N}\right)-\mathbb{E}\left(m_{i,\delta_N}\right)\right] \overset{d}{\to} N\left[0,\Sigma_M\right],$$

where

$$\Sigma_M = \lim_{N\to\infty}\mathrm{Var}\left[\delta_N\cdot m_{i,\delta_N}\right] < \infty.$$

Part *i)* in Assumption 4 imposes smoothness conditions on the distribution of the data and the form of the marginal effect, requiring that $f_{y|x}$ and m/π_α belong to regularity spaces. Source conditions like (47) are routinely assumed in the literature on ill-posed inverse problems. In econometrics, several variants of this assumption have already been applied.[30] Note that (47) and (48) are automatically satisfied for $\beta_y = 0$ and $\beta_m = 0$, respectively.[31]

In a given model, (47) may substantially restrict the class of data distributions. For example, in the classical nonparametric deconvolution model $y_i = \alpha_i + v_i$, the source condition (47) with $\beta_y = 1$ will require the distribution of v_i to be *less smooth* than that of α_i (Carrasco and Florens, 2009). Notice that, in contrast, smoothness restrictions were not needed when considering common parameters.

Part *ii)* guarantees that the bias term B_N tends fast enough to zero when N tends to infinity. To assess the rate of convergence of B_N, the *total* degree of smoothness $\beta_y + \beta_m$ is key. To see why, consider for example the case where Tikhonov regularization (45) is used. Then:

$$\sup_j \quad \left|\lambda_{j,x}^{\beta_y+\beta_m-1}\left(q_{j,x}\left(\delta_N\right)-1\right)\right| = O\left(\delta_N^{\gamma}\right), \tag{50}$$

[30] See for example Carrasco *et al.* (2008), and Darolles, Florens and Renault (2009). Related assumptions have been made in Blundell *et al.* (2007), and in Hall and Horowitz (2005).

[31] This is because $f_{y|x} \in \mathcal{G}_y$, and $\frac{m}{\pi_\alpha} \in \mathcal{G}_\alpha$.

with $\gamma = \min\left(1, \frac{\beta_y+\beta_m-1}{2}\right)$. By comparison, when using spectral cut-off or Landweber-Fridman regularization, (50) holds with $\gamma = \frac{\beta_y+\beta_m-1}{2}$. See Proposition 3.11 in Carrasco *et al.* (2008).

Parts *iii*), *iv*) and *v*) ensure that A_N satisfies a central limit theorem. In particular, Conditions *iii*) and *iv*) guarantee that $\text{Var}\left[\delta_N \cdot m_{i,\delta_N}\right]$ is finite. Note that, by (44), $\left|\frac{\delta_N q_{j,x}(\delta_N)}{\lambda_{j,x}}\right| \leq a_x \lambda_{j,x}$ is almost surely bounded, as the operator $L_{\theta,x}$ is compact. Condition *v*) requires additional moments to be finite, in order for a Liapunov central limit theorem to be applicable.

Under those conditions, the mean squared error (MSE) of the marginal effects estimator satisfies:

$$\mathbb{E}\left[\left(\widehat{M}_{\delta_N} - M\right)^2\right] \;=\; O\left(\frac{1}{N\delta_N^2}\right) + O\left(\delta_N^{2\gamma}\right),$$

where the first term on the right-hand side accounts for the variance of the estimator, while the second term accounts for the squared bias. The usual trade-off arises, as a smaller regularization parameter δ_N decreases the bias, but increases the variance at the same time. The rate of convergence of the estimator is thus always slower than the one obtained for $\delta_N = N^{-\frac{1}{2+2\gamma}}$, where the rate of convergence in terms of root-MSE is $N^{\frac{\gamma}{2+2\gamma}}$.

Since when using Tikhonov regularization γ is always lower than 1, the rate of convergence is thus slower than $N^{\frac{1}{4}}$, irrespective of the degree of smoothness of $f_{y|x}$. Using spectral cut-off or Landweber-Fridman instead, one may obtain better rates of convergence when $\beta_y + \beta_m$ is large, i.e. when the distribution of the data $f_{y|x}$– or the function m/π_α– is very smooth.

Finally, the next result gives the asymptotic distribution of $\widehat{M}_{\delta_N}$.

Theorem 6 *Let Assumptions 1 and 4 hold. Then:*

$$\sqrt{N}\delta_N\left(\widehat{M}_{\delta_N} - M\right) \xrightarrow{d} N\left[0, \Sigma_M\right]. \tag{51}$$

Remark 1. Note that the asymptotic variance of $\widehat{M}_{\delta_N}$ can simply be estimated as:

$$\widehat{\text{Var}}\left(\widehat{M}_{\delta_N}\right) = \frac{\widehat{\text{Var}}\left[m_{i,\delta_N}\right]}{N}, \tag{52}$$

where $\widehat{\text{Var}}\left[m_{i,\delta_N}\right]$ denotes the sample variance of m_{i,δ_N}. Note also that one can use the MSE calculations to choose δ_N in practice, as a minimizer of $\widehat{\text{Var}}\left(\widehat{M}_{\delta_N}\right) + \widehat{\text{Bias}}^2$, where

$$\widehat{\text{Bias}} = \widehat{\mathbb{E}}\left[\sum_j \left(q_{j,x_i}\left(\delta_N\right) - 1\right)\left\langle \psi_{j,x_i}, f_{\alpha|x}\right\rangle \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right].$$

In practice, $f_{\alpha|x}$ is unknown, and can be replaced by an estimate of $L_{\theta_0}^{\dagger} f_{y|x}$, possibly regularized. See Carrasco and Florens (2009) and Gagliardini and Scaillet (2008) for related approaches to choose the regularization parameter.

Remark 2. It is interesting to provide conditions under which M can be estimated at a root-N rate. This will happen when (48) holds with $\beta_m = 1$:

$$C(x) \equiv \sum_j \frac{\left\langle \psi_{j,x}, \frac{m}{\pi_\alpha} \right\rangle^2}{\lambda_{j,x}^2} < \infty, \quad x - \text{a.s.} \tag{53}$$

We show in Appendix B that, when (53) holds and under some additional regularity conditions, we can take $q_{j,x} = 1$ in (43) and obtain a root-N consistent and asymptotically normal estimator $\widehat{M}$ of M.

Condition (53) requires that $\left\langle \psi_{j,x}, \frac{m}{\pi_\alpha} \right\rangle$ tends fast enough to zero as j tends to infinity relative to $\lambda_{j,x}$. In the random coefficients model of Example 1, this condition requires the Fourier transform of m/π_α to decay fast enough to zero, as we argue in the supplementary appendix to this paper.

More generally, it can be shown that (53) holds if and only if $\frac{m}{\pi_\alpha} \in \mathcal{R}\left(L_{\theta_0,x}^*\right)$, that is if there exists a function $h \in \mathcal{G}_y$ such that $m(x,\alpha) = \mathbb{E}\left[\pi_y(y_i)\, h(y_i) \,|x, \alpha\right]$. This corresponds to mean effects of the form: $M(x) = \mathbb{E}\left[\pi_y(y_i)\, h(y_i) \,|x\right]$.[32]

7 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.

7.1 Practical implementation

To implement our method in practice, we approximate the within projection operator using a discretization approach.[33] The approximation method uses the parametric probability

[32] Note that, in injective models: $\overline{\mathcal{R}\left(L_{\theta_0,x}^*\right)} = \mathcal{G}_\alpha$. So, when $L_{\theta_0,x}$ is injective the set of marginal effects that satisfy (53) is dense in $\mathcal{G}_\alpha$. More generally, this set is always dense in the set of identified marginal effects, as shown by Proposition 4.

[33] This approach is sometimes referred to as *least-squares collocation*. It is presented in Kress (1989, Chapter 17) and Engl *et al.* (2000, Section 3.3). In the econometric literature, Carrasco and Florens (2009) have applied this approach to a nonparametric deconvolution model. GAUSS codes implementing the approach are available from the author.

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, replace the integral in (29) by a sum over $s = 1, ..., N_y$, and look for a solution in the range of $L_{\theta,x}$ of the form:

$$\text{Proj}_{\pi_y}\left[h\middle|\overline{\mathcal{R}\left(L_{\theta,x}\right)}\right] \approx L_{\theta,x}\widetilde{g}, \quad \text{where} \quad \widetilde{g}\left(\alpha\right) = \sum_{s=1}^{N_y} \frac{1}{\pi_\alpha\left(\alpha\right)} f_{y|x,\alpha;\theta}\left(\underline{y}_s|x,\alpha\right) g_s.$$

To motivate using this particular family of functions, note that:

$$\left\{\alpha \mapsto \frac{f_{y|x,\alpha;\theta}\left(y|x,\alpha\right)}{\pi_\alpha\left(\alpha\right)}, \quad y \in \mathcal{Y}\right\}$$

spans the orthogonal complement of the null-space of $L_{\theta,x}$.[34] Imposing that $\widetilde{g}$ belongs to that particular subspace of $\mathcal{G}_\alpha$ is without loss of generality.

Then, we solve for $\left(g_1, ..., g_{N_y}\right)$ in:

$$\min_{\left(g_1,...,g_{N_y}\right)\in\mathbb{R}^{N_y}} \quad \sum_{s=1}^{N_y}\left(\sum_{s'=1}^{N_y}\left[\int_{\mathcal{A}} \frac{f_{y|x,\alpha;\theta}\left(\underline{y}_s|x,\alpha\right) f_{y|x,\alpha;\theta}\left(\underline{y}_{s'}|x,\alpha\right)}{\pi_\alpha\left(\alpha\right)} d\alpha\right] g_{s'} - h\left(\underline{y}_s\right)\right)^2.$$

This yields:

$$\begin{aligned}\left[L_{\theta,x}\widetilde{g}\right]\left(y\right) \quad = \quad & \left[\left(\int_{\mathcal{A}} \frac{f_{y|x,\alpha;\theta}\left(y|x,\alpha\right) f_{y|x,\alpha;\theta}\left(\underline{y}_s|x,\alpha\right)}{\pi_\alpha\left(\alpha\right)} d\alpha\right)_s\right]' \\ & \times \left[\left(\int_{\mathcal{A}} \frac{f_{y|x,\alpha;\theta}\left(\underline{y}_s|x,\alpha\right) f_{y|x,\alpha;\theta}\left(\underline{y}_{s'}|x,\alpha\right)}{\pi_\alpha\left(\alpha\right)} d\alpha\right)_{s,s'}\right]^\dagger \left[\left(h\left(\underline{y}_{s'}\right)\right)_{s'}\right].\end{aligned} \tag{54}$$

Lastly, the moment functions are approximated as:

$$\varphi_r\left(y_i, x_i, \theta\right) \approx \pi_y\left(y_i\right)\left(h_r\left(y_i\right) - \left[L_{\theta,x}\widetilde{g}_r\right]\left(y_i\right)\right)\zeta_r\left(x_i\right), \tag{55}$$

where $L_{\theta,x}\widetilde{g}_r$ is given by (54) with h_r in place of h. As N_y tends to infinity, the right-hand side in (55) converges almost surely to the true moment function, provided $\{\underline{y}_s\}$ becomes dense in $\mathcal{Y}$ as N_y increases (see Engl *et al.*, 2000, p. 67-68).

In practice, we approximate the integrals in (54) using importance sampling (Geweke, 1989), drawing N_α values $\underline{\alpha}_n$ from a user-specified density $\overline{\pi}$ whose support contains $\mathcal{A}$. It

[34] This family spans the range of the adjoint (or transpose) operator $L^*_{\theta,x}$, which is also the orthogonal complement of the null-space of $L_{\theta,x}$.

is easy to see that using this approach yields the following approximation to the moment functions:

$$\varphi_r(y_i, x_i, \theta) \approx \pi_y(y_i) \left(h_r(y_i) - \left(\underline{f}_{\theta,x_i}^{(y_i)} \right)' \underline{L}_{\theta,x_i}^{\dagger} \underline{h}_r \right) \zeta_r(x_i), \tag{56}$$

where

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

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

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

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, one may want to assess the effect of approximation error. In our context, this could be done along the lines of Carrasco and Florens (2009), who work in an asymptotic where the size of the discretization grows at the same rate as the sample size.

When the dimensions of the matrix $\underline{L}_{\theta,x_i}$ are large, the numerical computation of the Moore-Penrose generalized inverse $\underline{L}_{\theta,x_i}^{\dagger}$ may be affected by errors due to finite machine precision. For this reason, we compute a modified generalized inverse that uses only $J \geq 1$ eigenvalues.[35] The simulation evidence below suggests that taking any J in a reasonable range leads to very similar results.

Lastly, a similar approach delivers an approximation to the average marginal effects estimate $\widehat{M}_{\delta_N}$ in (43) as:

$$\widehat{M}_{\delta_N} \approx \widehat{\mathbb{E}} \left[\sum_{j=1}^{J} q_{j,x_i}(\delta_N)\, \pi_y(y_i) \frac{N_y}{\underline{\lambda}_{j,x_i}^2} \left(\underline{f}_{\widehat{\theta},x_i}^{(y_i)} \right)' \underline{\psi}_{j,x_i} \underline{\psi}_{j,x_i}' \underline{m} \right],$$

where

$$\underline{m} = \left[\left(\frac{1}{\sqrt{\overline{\pi}(\underline{\alpha}_n)\,\pi_\alpha(\underline{\alpha}_n)}} m(\underline{\alpha}_n) \right)_n \right]$$

is an $N_\alpha \times 1$ vector, and where $\underline{\lambda}_{j,x_i}$ and $\underline{\psi}_{j,x_i}$ are the singular values and right singular vectors, respectively, of the matrix $\underline{L}_{\widehat{\theta},x_i}$.

[35] This 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.

7.2 Simulation evidence

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

$$y_{it}^{*} = \alpha_i + v_{it}, \quad t = 1, 2, \tag{57}$$

where the distribution of v_{it} given α_i is i.i.d normal $(0, \sigma^2)$. In addition, y_{it}^{*} is observed only when $y_{it}^{*} \geq c_t$, where the thresholds c_t are known. Our interest will center on the common parameter σ and the average marginal effect $\mathbb{E}(\alpha_i)$. 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 and the mean of α_i. In the simulations we take α_i to be normal with mean 1 and unitary variance.

Common parameters. In the two upper panels of Figure 1 we show the mean of $\widehat{\sigma}$ and $\widehat{\theta}$, as well as asymptotic 95%-confidence intervals, across 1000 simulations, for a sample size $N = 100$. 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 models π_α is set to one, and we set $h_r(y) = \phi(y - \mu_r)$, where ϕ is the standard normal pdf and where μ_r takes 49 different values in $\mathbb{R}^2$:

$$\{(0,0), (0,1), (0,-1), (0,2), (0,-2), (0,3), (0,-3), ..., (-3,-3)\}.$$

The weighting matrix Υ is chosen to be the identity. In addition, we let $\overline{\pi}$ be uniform on $[-5, 5]$. Moreover, we take $N_y = 500$ and $N_\alpha = 50$, and we use Halton's quasi-random sequences to generate $\{\underline{y}_s\}$ and $\{\underline{\alpha}_n\}$, in view of their superior convergence properties relative to standard Monte-Carlo methods (see Chapter 9 in Judd, 1998).

On the x-axis of the figure we report the number of singular values J used in the numerical computation of the discretized version of the within projection operator. We see that the results quickly stabilize around the true value ($\sigma_0 = 1$ and $\theta_0 = 1$, respectively). This result is consistent with the absence of ill-posedness in the estimation of common parameters.

Next, we provide some numerical evidence on uniform Fourier convergence in the two models. In Section 5 we assumed uniform Fourier convergence to show root-N consistency and asymptotic normality of common parameter estimates. In Figure 2 we report the sum $\sum_{j>J} \langle \phi_{j,\theta}, f_y \rangle^2$, for various J and for common parameters (θ and σ) in a grid of values ranging between .5 and 1.5.[36] Figure 2 shows that the Fourier coefficients tend quickly to zero, and there is visual evidence that the convergence is uniform over the set of parameters that we have considered. This provides numerical support for uniform Fourier convergence in those two models.

Returning to common parameter estimates, Table 1 reports the mean and standard deviation of $\widehat{\sigma}$ and $\widehat{\theta}$ across 1000 simulations, for two sample sizes: $N = 100$ and $N = 500$. We report the results for three choices of functions h_r, taking μ_r as an element of either of three increasing sets containing 9, 25, and 49 points, respectively.[37] Lastly, we have used $J = 12$ singular values in our discretization of the within projection operator.

To provide a benchmark, we also report in the table the maximum likelihood estimates of σ and θ. Note that those estimates require knowledge of the true distribution of α_i. In addition, for the random coefficients model we report Chamberlain (1992)'s GMM estimator: $\widetilde{\theta} = \widehat{\mathbb{E}}(y_{i2}) / \widehat{\mathbb{E}}(y_{i1})$. This last estimator does not require knowledge of the distribution of α_i.

Table 1 shows that functional differencing estimates behave well, with moderate biases. However, comparison with the infeasible random-effects estimator shows that the loss of efficiency relative to maximum likelihood is large. In the tobit model for $N = 100$, the standard deviation of the best functional differencing estimate ($R = 49$) is 60% higher than the one of the infeasible MLE.

The results for the random coefficients model (lower part of the table) suggest that our choice of moment functions is not optimal, and that there exist potential efficiency gains within the functional differencing framework. Indeed, when $N = 100$ the standard deviation of the simple GMM estimator $\widetilde{\theta}$ is 30% *lower* than the one of the best functional differencing estimate. As we have seen in Section 2, the mean restrictions that motivate $\widetilde{\theta}$

[36] In our experiments, we observed that estimates of singular vectors associated with very small singular values were affected by numerical error. In Chamberlain' s model, the sum $\sum_{j=1}^{J} \langle \phi_{j,\theta}, f_y \rangle^2$ increased steadily with J and seemed to reach a plateau after a few singular values, yet the sum jumped after the 19th singular value (and actually became $>> \|f_y\|^2$). For this reason, we discarded the singular values $\lambda_{j,\theta}, j \geq 19$ in the sum. For the tobit model, this phenomenon occurred after the 14th singular value, and we proceeded similarly.

[37] Those three sets are $\{(0,0),(0,1),(0,-1)...,(-1,-1)\}$, $\{(0,0),(0,1),(0,-1),(0,2),(0,-2),...,(-2,-2)\}$, and $\{(0,0),(0,1),(0,-1),(0,2),(0,-2),(0,3),(0,-3),...,(-3,-3)\}$.

are strictly contained in the full set of restrictions that the functional differencing approach can potentially exploit. Exploring those efficiency gains is an important avenue for future research.

Average marginal effects. We then report in the lower panels of Figure 1 the mean and 95%-confidence intervals of the functional differencing estimates of $\mathbb{E}(\alpha_i)$ in the two models. We take $q_j = \mathbf{1}\{j \leq J\}$ in (43), and report the number J of singular values used in the computation on the x-axis. This amounts to using a truncated singular value decomposition as regularization scheme.

In sharp contrast with common parameters (upper part of the figure) the variance of the estimates increases rapidly as one uses a larger number of singular values in the computation. This is consistent with ill-posedness affecting the estimation of the mean individual effect.[38] There is also evidence of ill-posedness in the estimation of $\mathbb{E}(\alpha_i)$ in the random coefficients model.[39] Still, the increase in variance with the number of singular values is less dramatic than for the tobit model, suggesting that ill-posedness is less severe in the random coefficients model.

Lastly, we report in Table 2 the estimates of the unweighted mean of α_i, and of the weighted mean $\mathbb{E}[\alpha_i \phi(\alpha_i)] / \mathbb{E}[\phi(\alpha_i)]$, where ϕ is the standard normal pdf. The results show strong evidence of ill-posedness when estimating the unweighted mean, while the estimates of the weighted mean behave better. Intuitively, weighting the mean of α_i by the normal density acts as an implicit regularization.[40]

8 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.

[38] We also applied the method outlined in Section 6 to choose the regularization parameter. The minimization of the approximate MSE worked well, implying that keeping between 2 and 3 singular values is optimal to estimate the mean of α_i in the tobit model.

[39] This is because, given our choice of functions π_α and π_y, the identity function $m_0(\alpha) = \alpha$ does *not* satisfy (53).

[40] In Chamberlain's model and given our choice of weighting function π_y, it can be shown that $m_1(\alpha) = \alpha\phi(\alpha)$ and $m_2(\alpha) = \phi(\alpha)$ satisfy (53). This explains why ill-posedness does not affect the estimation of the weighted mean of α_i in this model, as evidenced by Table 2.

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, for which we have emphasized a problem of ill-posedness.

This paper raises a number of open questions. First, in infinite dimensions, the orthogonal complement of the range of the model operator is often infinite-dimensional. The preliminary simulation evidence that we have presented suggests that using one set of moment functions or another in estimation may very much affect finite-sample precision. Direct implementation of the optimal instruments is likely to be difficult, because of the necessary regularizations involved. It is thus of interest to suggest alternative moment functions to use in practice.

A second avenue for future work 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. With this aim, we have outlined a constrained functional differencing approach that yields additional restrictions on common parameters. It seems promising to develop this insight, particularly to deal with partially identified marginal effects in general models.

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) prove general identification results in models with latent variables under conditional independence restrictions. Hu and Shum (2009) discuss the nonparametric identification of Markovian dynamic models with unobserved states. 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 that some restrictions are imposed on the dynamics of time-varying errors (Arellano and Bonhomme, 2009b). The framework introduced in this paper should be useful to extend those results to nonlinear panel data models.

References

[1] Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

[2] Andersen, E.B. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society B*, 32, 283-301.

[3] Arellano, M. (2003): "Discrete Choices with Panel Data," *Investigaciones Económicas*, XXVII, 423–458.

[4] Arellano, M., and S. Bonhomme (2009a): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

[5] Arellano, M., and S. Bonhomme (2009b): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *mimeo*.

[6] 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.

[7] 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," *mimeo*.

[8] Bester, A., and C. Hansen (2007): "Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity," *mimeo*.

[9] 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.

[10] Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 7, 1613-1669.

[11] Buchinsky, M., J. Hahn, and K.I. Kim (2008): "Semiparametric Information Bound of Dynamic Discrete Choice Models," *mimeo*.

[12] Carrasco, M., and J. P. Florens (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16, 797-834.

[13] Carrasco, M., and J. P. Florens (2009): "Spectral Methods for Deconvolving a Density," to appear in *Econometric Theory*.

[14] 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.

[15] Carro, J. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *Journal of Econometrics*, 127, 503-528.

[16] Chamberlain, G. (1984): "Panel Data", in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Elsevier Science.

[17] Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, 34, 305–334.

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

[19] Chamberlain, G. (2010): "Binary Response Models for Panel Data: Identification and Information", *Econometrica*, 78, 159–168.

[20] Chen, X., and D. Pouzo (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152, 46–60.

[21] 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.

[22] 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.

[23] Darolles, S., J.P. Florens, and E. Renault (2009): "Nonparametric Instrumental Regression," *mimeo*. Available at SSRN: http://ssrn.com/abstract=1338775

[24] Engl, H.W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems*, Kluwer Academic Publishers.

[25] Gagliardini, P., and O. Scaillet (2008): "Tikhonov Regularization for Nonparametric Instrumental Variable Estimators," *WP*.

[26] Geweke, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317–1339.

[27] Goldenshluger, A. and S. V. Pereverzev (2003): "On Adaptive Inverse Estimation of Linear Functionals in Hilbert Scales," *Bernoulli*, 9(5), 783–807.

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

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

[30] Hall, P., and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

[31] Hause, J. (1980): "The Fine Structure of Earnings and the On-the-Job Training Hypothesis," *Econometrica*, 48, 1013–1029.

[32] Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

[33] Hoderlein, S., and H. White (2009): "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects", *WP*.

[34] Honoré, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 533–565.

[35] Honoré, B. (1993): "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variable," *Journal of Econometrics*, 59, 35–61.

[36] Honoré, B. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

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

[38] Horowitz, J., and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75(4), 1191–1208.

[39] Hu, L. (2002): "Estimation of a Censored Dynamic Panel Data Model," *Econometrica*, 70(6), 2499-2517.

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

[41] Hu, Y., and M. Shum (2009): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," *mimeo*.

[42] Johnson, E.G. (2004): "Identification in Discrete Choice Models with Fixed Effects," Working paper, Bureau of Labor Statistics.

[43] Judd, K. (1998): *Numerical Methods in Economics*, MIT Press. Cambridge, London.

[44] Kress, R. (1989): *Linear Integral Equations*, Springer.

[45] Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

[46] Kyriazidou, E. (2001): "Estimation of Dynamic Panel Data Sample Selection Models," *Review of Economic Studies*, 68, 543–572.

[47] Lancaster, T. (2000): "The Incidental Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413.

[48] Lancaster, T. (2002): "Orthogonal Parameters and Panel Data", *Review of Economic Studies*, 69, 647–666.

[49] Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), 357–362.

[50] 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.

[51] Newey, W.K. (1990a): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809-837.

[52] 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. Amsterdan: Elsevier Science.

[53] Newey, W., and J. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

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

[55] Politis, N., J. Romano, and M. Wolf (1999): *Subsampling*, Springer Verlag, New York.

[56] Severini, T. A. and Tripathi, G. (2007): "Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," Working Paper, University of Connecticut.

[57] Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

[58] Stewart, G.W. (1977): "On the Perturbation of Pseudo-Inverses, Projections, and Linear Least Squares Problems," *SIAM Review*, 19, 634-666.

APPENDIX

A Proofs

Proof of Proposition 1. Part *i*). Conversely, assume that Rank $(L_{\theta_1,x}) = N_y$ for some $\theta_1 \neq \theta_0$ in Θ, almost surely in x. Then $W_{\theta_1,x} f_{y|x} = 0$ a.s. in x, so (24) holds at θ_1 also. This implies that θ_0 is not identified from (24).

Part *ii*). Conversely, suppose that θ_0 is not identified from (24). Then, there exists a $\theta_1 \neq \theta_0$ in Θ such that $W_{\theta_1,x} f_{y|x} = 0$ a.s. in x. This implies that $f_{y|x}$ belongs to the range of $L_{\theta_1,x}$ with probability one in x.

Proof of Proposition 2. Under the proposition's assumptions, by a result in Stewart (1977, p. 653), $\theta \mapsto W_{\theta,x}^{\dagger}$ is differentiable at θ_0, a.s in x, with derivatives:

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

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, $G(x)\, a = 0$ having $a = 0$ as only solution is the rank condition for local point-identification of θ_0 in (24).

Optimal moment restrictions (finite support). Let $U_{\theta,x}$ be the matrix defined in Subsection 3.2. The restrictions from functional differencing can be written as:

$$\mathbb{E}\left[\varphi(y_i, x_i, \theta_0) \,|x_i\right] = 0, \tag{A1}$$

where $\varphi(y_i, x_i, \theta) = U_{\theta,x_i}\left[\tau(y_i), .\right]'$ is $(N_y - \text{rank}\,(L_{\theta,x_i})) \times 1$. This is a finite set of conditional moment restrictions, for which the optimal instruments are given by the next proposition.

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

Then the optimal instruments corresponding to (A1) are given by:

$$\kappa_{\theta_0,x_i}^{-1} \mathbb{E}\left[U_{\theta_0,x_i}' \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^{\dagger}\left[., \tau(y_i)\right] \,\middle|\, x_i \right], \quad k = 1, ..., \dim \theta. \tag{A2}$$

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:

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

where we have used that $f_{y|x} = L_{\theta_0,x} L_{\theta_0,x_i}^{\dagger} f_{y|x}$, 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 $\theta \mapsto f_{y|x,\alpha;\theta}(y|x,\alpha)$ is continuously differentiable in a neighborhood of θ_0, it follows that the moment functions are differentiable at θ_0 with derivatives:

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

The conclusion then follows from Chamberlain (1987).

■

Proof of 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.

Next, let:

$$f_y(y)=\sum_{n=1}^{N_\alpha}f_{y|\alpha;\theta_0}(y|\underline{\alpha}_n)f_{\alpha|\eta_0}\left(\underline{\alpha}_n\right)$$

be a regular parametric submodel, which includes a parametric model for the individual effects that depends on a scalar parameter η_0. The nonparametric tangent space consists of the (closure of the) linear span of:

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

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

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

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

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

Given that the scores are defined over the finite support of y_i, it is convenient to work with $N_y\times 1$ vectors: $\mathbb{S}=\left(\mathbb{S}\left(\underline{y}_1\right),...,\mathbb{S}\left(\underline{y}_{N_y}\right)\right)'$. Using this convention, the nonparametric 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 sth diagonal element $f_y\left(\underline{y}_s\right)$.

The efficient score is obtained by projecting:

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

on the nonparametric 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 $f_y\left(\underline{y}_s\right)$– are: $\frac{1}{f_y(\underline{y}_s)^2} \times f_y\left(\underline{y}_s\right) = \frac{1}{f_y(\underline{y}_s)}$. 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 A1

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

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

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 A1 we can rewrite the efficient score as:

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

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

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

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

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

So, we finally obtain:

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

Comparing with (27) ends the proof.

Proof of Theorem 2. Suppose *i*). Then, as $f_{\alpha|x} \in \mathcal{S}$, it follows that $Q^+_{\theta_0,x}\left(L_{\theta_0,x}f_{\alpha|x}\right) = L_{\theta_0,x}f_{\alpha|x}$, hence that $Q^+_{\theta_0,x}\left(f_{y|x}\right) = f_{y|x}$. This shows *ii*).

Conversely, suppose that *ii*) holds. Then $f_{y|x} = Q^+_{\theta_0,x}\left(f_{y|x}\right)$. From the definition of the projection $Q^+_{\theta_0,x}$ and the fact that $\mathcal{S}$ is a closed subset of $\mathbb{R}^{N_\alpha}$, there exists a $g \in \mathcal{S}$ such that $Q^+_{\theta_0,x}\left(f_{y|x}\right) = L_{\theta_0,x}g$. Hence $f_{y|x} = L_{\theta_0,x}g$, and *i*) holds with $f_{\alpha|x} = g$.

Proof of Theorem 3. First note that:

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

It follows that $W_{\theta_0,x} f_{y|x} = 0$ with probability one. Hence (30). To show that (30) and (31) are equivalent, note that, as $W_{\theta_0,x}$ is self-adjoint:

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

Proof of Proposition 3. Assume that θ_0 is globally identified from (30), and suppose that $\overline{\mathcal{R}\left(L_{\theta_1,x}\right)} = \mathcal{G}_y$ with probability one for some $\theta_1 \neq \theta_0$ in Θ. Then, as $h \mapsto W_{\theta_1,x} h$ is continuous and $W_{\theta_1,x}$ is zero on $\mathcal{R}\left(L_{\theta_1,x}\right)$, it follows that $W_{\theta_1,x} = 0$, a.s. So $W_{\theta_1,x} f_{y|x} = 0$, contradicting the fact that θ_0 is globally identified.

Optimal moment restrictions (infinite support). Let us define the following linear operator:

$$U_{\theta,x} = \sum_j \left\langle \nu_j, . \right\rangle \xi_{j,\theta,x},$$

where $\{\nu_j\}$ is *any* orthonormal family in $\mathcal{G}_y$, $\{\xi_{j,\theta,x}\}$ is any orthonormal basis of the null-space of the adjoint operator $\mathcal{N}\left(L^*_{\theta,x}\right)$,[41] and the sum ranges from $j = 1$ to the (possibly infinite) dimension of $\mathcal{N}\left(L^*_{\theta,x}\right)$.

By construction, $W_{\theta,x} = U_{\theta,x} U^*_{\theta,x}$. Moreover:

$$\begin{aligned}
W_{\theta_0,x} f_{y|x} = 0 \quad &\Leftrightarrow \quad \sum_j \left\langle \xi_{j,\theta_0,x}, f_{y|x}\right\rangle \xi_{j,\theta_0,x} = 0 \\
&\Leftrightarrow \quad \left\langle \xi_{j,\theta_0,x}, f_{y|x}\right\rangle = 0 \text{ for all } j \\
&\Leftrightarrow \quad \sum_j \left\langle \xi_{j,\theta_0,x}, f_{y|x}\right\rangle \nu_j = 0 \\
&\Leftrightarrow \quad U^*_{\theta_0,x} f_{y|x} = 0.
\end{aligned}$$

Now, this set of restrictions can be equivalently written as a set of conditional moment restrictions indexed by $y \in \mathcal{Y}$. To see this, note that from the Riesz representation theorem[42] for each $\theta \in \Theta$ and $x \in \mathcal{X}$ there exists a set of functions $\{\omega(y, ., x, \theta) \in \mathcal{G}_y, y \in \mathcal{Y}\}$ such that, for any $h \in \mathcal{G}_y$:

$$\left[U^*_{\theta,x} h\right](y) = \int_{\mathcal{Y}} \omega(y, \widetilde{y}, x, \theta)\, h(\widetilde{y})\pi_y(\widetilde{y})d\widetilde{y}.$$

Hence (30) is equivalent to

$$\begin{aligned}
\mathbb{E}\left[\pi_y(y_i)\omega(y, y_i, x_i, \theta_0)\,|x\right] &= \left[U^*_{\theta_0,x} f_{y|x}\right](y) \\
&= 0, \quad \text{for all } y \in \mathcal{Y}. \qquad \text{(A3)}
\end{aligned}$$

[41] The null-space is defined as: $\mathcal{N}\left(L^*_{\theta,x}\right) = \left\{h \in \mathcal{G}_y, \quad L^*_{\theta,x} h = 0\right\}$.

[42] The Riesz representation theorem can be applied here because $U^*_{\theta,x}$ is bounded, see Theorem 2.18 in Carrasco *et al.* (2008).

This shows that θ_0 is characterized as the solution of a set of conditional moment restrictions, which becomes a continuum when $\mathcal{Y}$ is continuous.

The analogy with the finite-dimensional case motivates considering the following instruments:

$$h_k^{\text{opt}} = \kappa_{\theta_0,x_i}^{-1} U_{\theta_0,x_i}^* \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^\dagger f_{y|x}, \quad k = 1, ..., \dim \theta. \tag{A4}$$

In this expression, $\frac{\partial L_{\theta_0,x_i}}{\partial \theta_k}$ is an operator with kernel $\frac{\partial f_{y|x,\alpha;\theta_0}}{\partial \theta_k}$. Regularity conditions that ensure that the population moment functions are differentiable, and that this operator is well-defined, are given in Section 5. The operator $\kappa_{\theta_0,x} : \mathcal{G}_y \to \mathcal{G}_y$ is a non-singular *covariance operator* (Carrasco and Florens, 2000) given by:

$$[\kappa_{\theta_0,x} h](y) = \int_{\mathcal{Y}} \mathbb{E}\Big[\pi_y(y_i)^2 \omega(y, y_i, x_i, \theta_0)\, \omega(\widetilde{y}, y_i, x_i, \theta_0) \,\Big|\, x\Big] h(\widetilde{y}) d\widetilde{y}, \quad \text{for all } h \in \mathcal{G}_y.$$

Proof of Proposition 4. Suppose that θ_0 is point-identified, and that $\frac{m}{\pi_\alpha} \in \overline{\mathcal{R}\left(L_{\theta_0,x}^*\right)}$. Let f_α and g_α be such that $f_{y|x} = L_{\theta_0,x} f_\alpha$ and $f_{y|x} = L_{\theta_0,x} g_\alpha$. Then, from Proposition 2.3 in Engl *et al.* (2000, p. 33), $\left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right)$ is the orthogonal projector onto the null-space of $L_{\theta_0,x}$. So, as $L_{\theta_0,x}(f_\alpha - g_\alpha) = 0$, there exists a $g \in \mathcal{G}_\alpha$ such that $g_\alpha - f_\alpha = \left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right) g$. Moreover, as $\overline{\mathcal{R}\left(L_{\theta_0,x}^*\right)}$ is the orthogonal complement of the null-space of $L_{\theta_0,x}$, we have that $\left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right) \frac{m}{\pi_\alpha} = 0$.

Noting that $M(x) = \left\langle \frac{m}{\pi_\alpha}, f_\alpha \right\rangle$, we thus have:

$$\begin{aligned}
\left\langle \frac{m}{\pi_\alpha}, g_\alpha \right\rangle &= M(x) + \left\langle \frac{m}{\pi_\alpha}, g_\alpha - f_\alpha \right\rangle \\
&= M(x) + \left\langle \frac{m}{\pi_\alpha}, \left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right) g \right\rangle \\
&= M(x) + \left\langle \left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right) \frac{m}{\pi_\alpha}, g \right\rangle = M(x),
\end{aligned}$$

where we have used that $I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}$ is self-adjoint. Hence $M(x)$ is identified.

In particular, as $f_{y|x} = L_{\theta_0,x} L_{\theta_0,x}^\dagger f_{y|x}$, we have:

$$M(x) = \left\langle \frac{m}{\pi_\alpha}, L_{\theta_0,x}^\dagger f_{y|x} \right\rangle.$$

This ends the proof.

B Proofs of asymptotic results

Proof of Theorem 4. 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 $\mu > 0$:

$$W_{\theta,x}^{(\mu)} = I_y - L_{\theta,x}\left(L_{\theta,x}^* L_{\theta,x} + \mu I_\alpha\right)^{-1} L_{\theta,x}^*. \tag{B1}$$

We start with the following result.

Lemma B1 *Let iii), iv), and viii) in Assumption 2 hold. Then, for any r and for $\mu > 0$ given, the function:*

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

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 $\theta_s \overset{s\to\infty}{\to} \theta$, then (e.g., Section 2.2 in Carrasco *et al.*, 2008):

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

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 $j \in \{1,2,...\}$. It thus follows that $\left\|W_{\theta,x}^{(\mu)}\right\| \leq 1$ for any θ, x. So, letting again $\theta_s \overset{s\to\infty}{\to} \theta$ we have:

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

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

■

Lemma B2 *Let v), vi) and viii) in Assumption 2 hold. Then, for any r:*

$$\mathbb{E}\left(\left[W_{\theta,x_i}^{(\mu)} h_r\right](y_i)\,\pi_y(y_i)\,\zeta_r(x_i)\right) \overset{\mu\to 0}{\to} \mathbb{E}\left([W_{\theta,x_i} h_r](y_i)\,\pi_y(y_i)\,\zeta_r(x_i)\right)$$

where the convergence holds uniformly on Θ.

Proof. We have:

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

So, for any $J \geq 1$:

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

So, using the Cauchy-Schwarz inequality:

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

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 B1 and B2 then shows that

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

is continuous on Θ, for any 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

$$\varphi_r=\pi_y\left(y_i\right)\left[W_{\theta,x_i}h_r\right]\left(y_i\right)\zeta_r\left(x_i\right).$$

We will show:

$$\sup_{\theta\in\Theta}\ \mathbb{E}\left(\left[\widehat{\mathbb{E}}\left(\varphi_r\right)-\mathbb{E}\left(\varphi_r\right)\right]^2\right)\overset{N\to\infty}{\to}0. \tag{B2}$$

For this, we will show two lemmas.

Lemma B3 *Let ix) in Assumption 2 hold. Then*

$$\sup_{\theta\in\Theta}\ \mathrm{Var}\left(\mathbb{E}\left(\left[W_{\theta,x_i}h_r\right]\left(y_i\right)\pi_y\left(y_i\right)\zeta_r\left(x_i\right)|x_i\right)\right)<\infty.$$

Proof.

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

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

■

Lemma B4 *Let vii) in Assumption 2 hold. Then*

$$\sup_{\theta\in\Theta}\ \mathbb{E}\left(\mathrm{Var}\left(\left[W_{\theta,x_i}h_r\right]\left(y_i\right)\pi_y\left(y_i\right)\zeta_r\left(x_i\right)|x_i\right)\right)<\infty.$$

Proof. We have, almost surely in x:

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

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

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

■

Finally, combining Lemmas B3 and B4, $\mathrm{Var}(\varphi_r)$ is uniformly bounded. So, the left-hand side in (B2) is bounded by a constant divided by N. This shows convergence in mean squares, which implies convergence in probability.

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

Proof of Theorem 5. 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:

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

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

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

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

Now, as in the proof of Theorem 4 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 any $J \geq 1$:

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

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 $j \in \{1,...,J\}$, and as $\left\langle\phi_{j,\theta,x},h_r\right\rangle^2 \leq \|h_r\|^2$, the first

term tends uniformly to zero as μ tends to zero (for fixed J). This shows that $W^{(\mu)}_{\theta,x}h_r$ tends to $W_{\theta,x}h_r$ as μ tends to zero, uniformly on $\mathcal{V}$.

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

$$\left\langle \frac{\partial L^*_{\tilde{\theta},x}}{\partial\theta_k} W_{\theta,x}h_r, L^\dagger_{\theta_0,x}f_{y|x} \right\rangle \overset{\theta\to\theta_0}{\to} \left\langle \frac{\partial L^*_{\theta_0,x}}{\partial\theta_k} W_{\theta_0,x}h_r, L^\dagger_{\theta_0,x}f_{y|x} \right\rangle.$$

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:

$$\sqrt{N}\left(\widehat{\mathbb{E}}[\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0)] - \mathbb{E}[\varphi(y_i, x_i, \theta)]\right) \overset{d}{\to} N[0, \mathrm{Var}(\varphi(y_i, x_i, \theta) - \varphi(y_i, x_i, \theta_0))].$$

As in the proof of Lemma B3 we have:

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

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 ix) 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 B4 we have, almost surely in x:

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

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 $\zeta_r(x_i)^2$) 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.

Proof of Theorem 6. We start with the following lemma.

Lemma B5 *Let Conditions iii) and iv) in Assumption 4 hold. Then* $\mathrm{Var}[\delta_N \cdot m_{i,\delta_N}] < \infty$.

Proof. We have:

$$\begin{aligned}
\mathrm{Var}[\delta_N \cdot m_{i,\delta_N}] &= \mathrm{Var}\left(\sum_j \delta_N q_{j,x_i}(\delta_N)\,\pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right) \\
&= \mathbb{E}\left[\mathrm{Var}\left(\sum_j \delta_N q_{j,x_i}(\delta_N)\,\pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle \,\middle|\, x_i\right)\right] \\
&\quad + \mathrm{Var}\left(\sum_j \delta_N q_{j,x_i}(\delta_N)\left\langle \phi_{j,x_i}, f_{y|x}\right\rangle \frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right).
\end{aligned}$$

Starting with the second term in the sum:

$$\mathrm{Var}\left(\sum_j \delta_N q_{j,x_i}(\delta_N)\left\langle \phi_{j,x_i}, f_{y|x}\right\rangle \frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right) \leq \mathbb{E}\left(\sup_j\left|\frac{\delta_N q_{j,x_i}(\delta_N)}{\lambda_{j,x_i}}\right|^2 \|f_{y|x}\|^2 \left\|\frac{m}{\pi_\alpha}\right\|^2\right)$$

where we have used the Cauchy-Schwarz inequality. This term is bounded by *iv*).

As for the first term in the sum, define:

$$K_x \frac{m}{\pi_\alpha} \equiv \sum_j \delta_N q_{j,x}(\delta_N) \frac{1}{\lambda_{j,x}} \left\langle \psi_{j,x}, \frac{m}{\pi_\alpha} \right\rangle \phi_{j,x}.$$

We have, almost surely in x:

$$\begin{aligned} \mathrm{Var}\left(\sum_j \delta_N q_{j,x_i}(\delta_N) \pi_y(y_i) \phi_{j,x_i}(y_i) \frac{1}{\lambda_{j,x_i}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha} \right\rangle \Big| x \right) &= \mathrm{Var}\left(\pi_y(y_i) \left[K_x \frac{m}{\pi_\alpha} \right](y_i) \Big| x \right) \\ &\leq \int_{\mathcal{Y}} \pi_y(y)^2 \left(\left[K_x \frac{m}{\pi_\alpha} \right](y) \right)^2 f_{y|x}(y|x)\, dy \\ &\leq \sup_{y \in \mathcal{Y}} \left(f_{y|x}(y|x) \pi_y(y) \right) \left\| K_x \frac{m}{\pi_\alpha} \right\|^2. \quad \text{(B3)} \end{aligned}$$

Notice that, by the Cauchy-Schwarz inequality:

$$\|K_x\|^2 \leq \sup_j \left| \frac{\delta_N q_{j,x}(\delta_N)}{\lambda_{j,x}} \right|^2, \quad x - a.s.$$

Therefore, the expectation of (B3) is bounded by *iii*).
This ends the proof.
■

From part *v*) in Assumption 4, we have:

$$\sqrt{N} \delta_N A_N \xrightarrow{d} N[0, \Sigma_M].$$

So, from the Mann-Wald theorem we only need to verify that

$$\sqrt{N} \delta_N B_N \xrightarrow{p} 0.$$

Now, we have:

$$\begin{aligned} B_N &= \mathbb{E}\left[\sum_j (q_{j,x_i}(\delta_N) - 1) \pi_y(y_i) \phi_{j,x_i}(y_i) \frac{1}{\lambda_{j,x_i}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha} \right\rangle \right] \\ &= \mathbb{E}\left[\sum_j \lambda_{j,x_i}^{\beta_y + \beta_m - 1} (q_{j,x_i}(\delta_N) - 1) \frac{1}{\lambda_{j,x_i}^{\beta_y}} \langle \phi_{j,x_i}, f_{y|x} \rangle \frac{1}{\lambda_{j,x_i}^{\beta_m}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha} \right\rangle \right]. \end{aligned}$$

From (47), (48), and the Cauchy-Schwarz inequality we have, almost surely:

$$\left| \sum_j \frac{1}{\lambda_{j,x_i}^{\beta_y}} \langle \phi_{j,x_i}, f_{y|x} \rangle \frac{1}{\lambda_{j,x_i}^{\beta_m}} \left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha} \right\rangle \right| \leq C_y(x_i)^{\frac{1}{2}} C_m(x_i)^{\frac{1}{2}}.$$

Hence:

$$|B_N| \leq \mathbb{E}\left[\left(\sup_j \left| \lambda_{j,x_i}^{\beta_y + \beta_m - 1} (q_{j,x_i}(\delta_N) - 1) \right| \right) C_y(x_i)^{\frac{1}{2}} C_m(x_i)^{\frac{1}{2}} \right].$$

The conclusion follows from part *ii*) in Assumption 4.

Root-N consistent estimation. Suppose that (53) holds, and take $q_{j,x} = 1$. Let us define:

$$\widehat{M} \;=\; \widehat{\mathbb{E}}\left[\sum_j \pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right].$$

We have, using the Cauchy-Schwarz inequality:

$$\mathrm{Var}\left(\sum_j \left\langle \phi_{j,x_i}, f_{y|x}\right\rangle \frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right) \;\leq\; \mathbb{E}\left(\left\|f_{y|x}\right\|^2 C\left(x_i\right)\right), \tag{B4}$$

where $C\left(x_i\right)$ is given by (53).

Let us define:

$$K_x\frac{m}{\pi_\alpha} \equiv \sum_j \frac{1}{\lambda_{j,x}}\left\langle \psi_{j,x}, \frac{m}{\pi_\alpha}\right\rangle \phi_{j,x}.$$

We have, almost surely in x:

$$\begin{aligned}
\mathrm{Var}\left(\sum_j \pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle \Big|\, x\right) \;&=\; \mathrm{Var}\left(\pi_y(y_i)\left[K_x\frac{m}{\pi_\alpha}\right](y_i)\,\Big|\, x\right) \\
&\leq\; \int_{\mathcal{Y}} \pi_y(y)^2\left(\left[K_x\frac{m}{\pi_\alpha}\right](y)\right)^2 f_{y|x}\left(y|x\right)dy \\
&\leq\; \sup_{y\in\mathcal{Y}}\left(f_{y|x}\left(y|x\right)\pi_y\left(y\right)\right)\left\|K_x\frac{m}{\pi_\alpha}\right\|^2 \\
&\leq\; \sup_{y\in\mathcal{Y}}\left(f_{y|x}\left(y|x\right)\pi_y\left(y\right)\right)C\left(x\right). \qquad \text{(B5)}
\end{aligned}$$

This implies that, if the right-hand side of (B4) and the expectation of the right-hand side of (B5) are finite:

$$\mathrm{Var}\left(\sum_j \pi_y(y_i)\phi_{j,x_i}(y_i)\frac{1}{\lambda_{j,x_i}}\left\langle \psi_{j,x_i}, \frac{m}{\pi_\alpha}\right\rangle\right) \;<\; \infty.$$

Because the bias term is zero in this case, root-N consistency and asymptotic normality of $\widehat{M}$ follow from standard arguments.

Table 1: Common parameter estimates ($T = 2$)

Tobit model: σ (true=1)

	$N = 100$		$N = 500$	
	Mean	Std	Mean	Std
Grid ($R = 9$)	1.021	.175	.998	.085
Grid ($R = 25$)	1.022	.169	.994	.071
Grid ($R = 49$)	1.011	.146	.994	.065
Infeasible REML	.996	.090	.997	.043

Chamberlain's model: θ (true=1)

	$N = 100$		$N = 500$	
	Mean	Std	Mean	Std
Grid ($R = 9$)	1.040	.286	1.022	.152
Grid ($R = 25$)	1.028	.221	1.011	.101
Grid ($R = 49$)	1.024	.191	1.009	.084
Infeasible REML	1.000	.125	.997	.054
GMM	1.006	.146	.999	.062

Note: Mean and standard deviations of $\widehat{\sigma}$ and $\widehat{\theta}$ across 1000 simulations. "Grid (R)" refers to using $\phi(. - \mu_r)$, $r = 1, ..., R$, to construct moment functions, where the set of values μ_r is indicated in the text. "Infeasible REML" is the infeasible random-effects maximum likelihood estimate, which assumes knowledge of f_α. "GMM" is Chamberlain (1992)'s estimator of θ.

Table 2: Average marginal effects estimates ($T = 2$)

Tobit model

	$N = 100$		$N = 500$		$N = 100$		$N = 500$	
	Unweighted mean (true=0)				Weighted mean (true=0)			
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$J = 2$	-.046	.238	-0.042	.108	.014	.021	.016	.009
$J = 5$	-.029	.431	-.016	.200	-.024	.275	-.021	.128
$J = 8$	.139	12.10	-.106	5.69	.147	1.047	.017	.436

Chamberlain's model

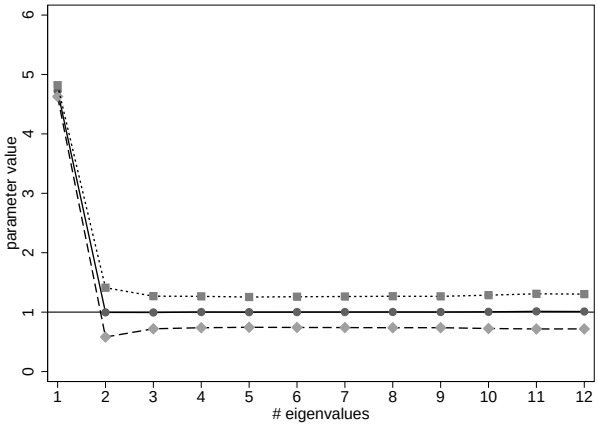
	$N = 100$		$N = 500$		$N = 100$		$N = 500$	
	Unweighted mean (true=1)				Weighted mean (true=.5)			
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$J = 2$	1.117	.227	1.116	.102	.485	.094	.478	.040
$J = 5$	.938	.374	.939	.170	.514	.140	.514	.060
$J = 8$	1.071	1.212	1.037	.534	.510	.134	.510	.058

Note: Mean and standard deviations of the estimates of the unweighted mean $\mathbb{E}(\alpha_i)$ and the weighted mean $\mathbb{E}[\alpha_i \phi(\alpha_i)] / \mathbb{E}[\phi(\alpha_i)]$ across 1000 simulations. J refers to the number of singular values used in estimation.

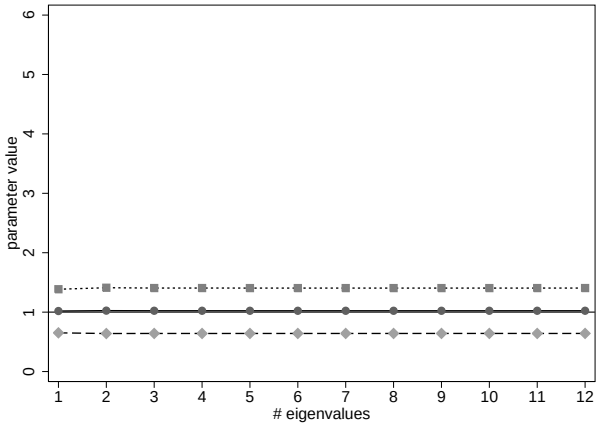
Figure 1: Parameter estimates ($N = 100, T = 2$)

Common parameters

Tobit model: σ (true=1)

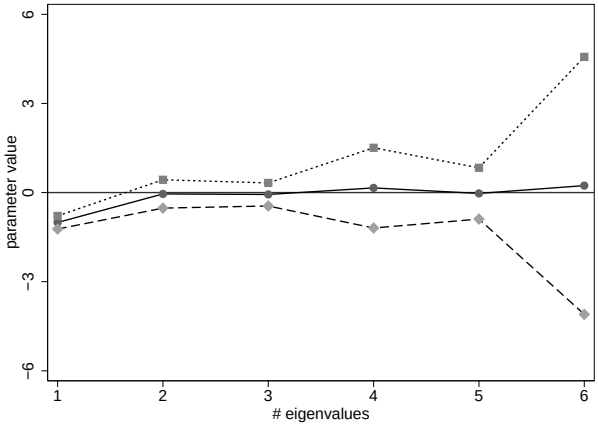


Chamberlain's model: θ (true=1)

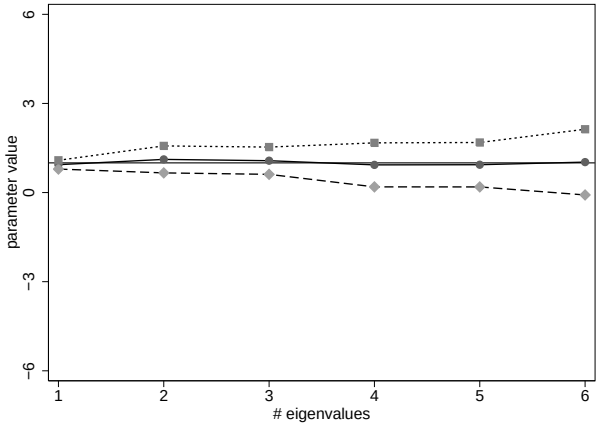


Average marginal effects

Tobit model: $\mathbb{E}(\alpha_i)$ (true=0)

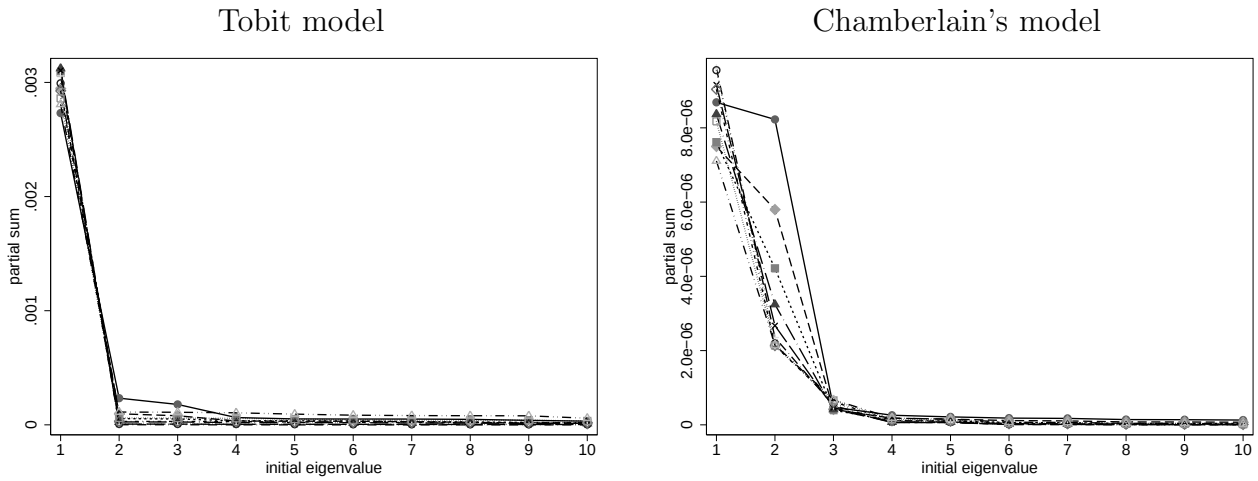


Chamberlain's model: $\mathbb{E}(\alpha_i)$ (true=1)



Note: On the x-axis we report the number of singular values used in estimation, while the y-axis shows parameter estimates. The functions used to construct moment functions are $\phi(. - \mu_r)$, $r = 1, ..., 49$, where the set of values for μ_r is indicated in the text (upper panels). The solid and discontinuous lines show the mean estimate and the 95%-confidence interval, respectively. The thin solid line indicates the true parameter value.

Figure 2: Uniform Fourier convergence ($T = 2$)



Note: We report the quantity $\sum_{j>J} \langle \phi_{j,\theta}, f_y \rangle^2$*, where* $(J+1)$ *is shown on the x-axis. The various curves correspond to different parameters* θ *(*σ *on the left panel), which belong to a grid* $\{.5, .6, ..., 1.5\}$.

Supplement to "Functional Differencing": Supplementary Appendix

By Stéphane Bonhomme

This supplementary appendix contains results on the examples not covered in the main text. It also presents a specification test that may be viewed as a nonlinear analog of the Hausman (1978) test of parametric random-effects models.

A Examples

Operator injectivity in the random coefficients model (normal errors). Here we show that $\text{rank}\,(B) = q$ (where $q = \dim \alpha_i$) is necessary and sufficient for $L_{\theta,x}$ to be injective in Example 1. We prove the result for $\pi_\alpha = 1$, so that $\mathcal{G}_\alpha = L^2(\mathbb{R}^q)$.

Let $g \in \mathcal{G}_\alpha$ such that $L_{\theta,x} g = 0$, that is:

$$(2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \left\{ \int_{\mathcal{A}} \exp\left[-\frac{1}{2} (y - a - B\alpha)' \Sigma^{-\frac{1}{2}} Q \Sigma^{-\frac{1}{2}} (y - a - B\alpha) \right] g(\alpha)\, d\alpha \right\} \times \left\{ \exp\left[-\frac{1}{2} (y-a)' \Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}} (y-a) \right] \right\} = 0.$$

This implies that:

$$\int_{\mathcal{A}} \exp\left[-\frac{1}{2} (y - a - B\alpha)' \Sigma^{-\frac{1}{2}} Q \Sigma^{-\frac{1}{2}} (y - a - B\alpha) \right] g(\alpha)\, d\alpha = 0.$$

Using the properties of Q this is equivalent to:

$$\int_{\mathcal{A}} \exp\left[-\frac{1}{2} \left(\left(\Sigma^{-\frac{1}{2}} B \right)^{\dagger} \Sigma^{-\frac{1}{2}} (y-a) - \alpha \right)' B' \Sigma^{-1} B \left(\left(\Sigma^{-\frac{1}{2}} B \right)^{\dagger} \Sigma^{-\frac{1}{2}} (y-a) - \alpha \right) \right] g(\alpha)\, d\alpha = 0.$$

Now, if B has full-column rank, then $\left(\Sigma^{-\frac{1}{2}} B \right)^{\dagger} \Sigma^{-\frac{1}{2}}$ is surjective. So we have, for all $z \in \mathbb{R}^q$:

$$\int_{\mathcal{A}} \exp\left[-\frac{1}{2} (z - \alpha)' B' \Sigma^{-1} B (z - \alpha) \right] g(\alpha)\, d\alpha = 0. \tag{A1}$$

As $\mathcal{G}_\alpha = L^2(\mathbb{R}^q)$, we can take L^2-Fourier transforms in (A1) and obtain, using that $B'\Sigma^{-1}B$ is non-singular:

$$[\mathcal{F}g](\tau)\, e^{-\frac{1}{2}\tau' \left(B'\Sigma^{-1}B \right)^{-1} \tau} = 0, \quad \tau \in \mathbb{R}^q,$$

where $\mathcal{F}$ is the L^2-Fourier transform operator (Yoshida, 1971, p. 154). This implies that $\mathcal{F}g = 0$, hence that $g = 0$. This shows that $L_{\theta,x}$ is injective.

Conversely, when B does not have full-column rank, let $r = \text{rank}\,(B)$. Let $\widetilde{V}$ be a $q \times r$ matrix such that $\widetilde{V}\widetilde{V}' = B^{\dagger}B$ and $\widetilde{V}'\widetilde{V} = I_r$, and let $\widetilde{U}$ be a $q \times (q-r)$ matrix such that $\widetilde{U}\widetilde{U}' = I_q - B^{\dagger}B$ and $\widetilde{U}'\widetilde{U} = I_{q-r}$. Let $\widetilde{g}_1 \in L^2(\mathbb{R}^r)$ and $\widetilde{g}_2 \in L^1(\mathbb{R}^{q-r}) \cap L^2(\mathbb{R}^{q-r})$ such that $\widetilde{g}_1 \neq 0$, $\widetilde{g}_2 \neq 0$, and $\int_{\mathbb{R}^{q-r}} \widetilde{g}_2(\nu)\, d\nu = 0$. Lastly, let $g(\alpha) = \widetilde{g}_1\left(\widetilde{V}'\alpha \right) \widetilde{g}_2\left(\widetilde{U}'\alpha \right)$. Note that $g \in \mathcal{G}_\alpha$ by construction.

Then, noting that $B = BB^{\dagger}B = B\widetilde{V}\widetilde{V}'$ we have, letting $C = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}}$:

$$\begin{aligned}
[L_{\theta,x} g](\alpha) &= C \int_{\mathcal{A}} \exp\left[-\frac{1}{2}(y - a - B\alpha)' \Sigma^{-1} (y - a - B\alpha)\right] g(\alpha)\, d\alpha \\
&= C \int_{\mathcal{A}} \exp\left[-\frac{1}{2}\left(y - a - B\widetilde{V}\widetilde{V}'\alpha\right)' \Sigma^{-1} \left(y - a - B\widetilde{V}\widetilde{V}'\alpha\right)\right] \widetilde{g}_1\left(\widetilde{V}'\alpha\right) \widetilde{g}_2\left(\widetilde{U}'\alpha\right) d\alpha \\
&= C \int_{\mathbb{R}^r} \exp\left[-\frac{1}{2}\left(y - a - B\widetilde{V}\mu\right)' \Sigma^{-1} \left(y - a - B\widetilde{V}\mu\right)\right] \widetilde{g}_1(\mu)\, d\mu \int_{\mathbb{R}^{q-r}} \widetilde{g}_2(\nu)\, d\nu \\
&= 0,
\end{aligned}$$

where we have used the change in variables $(\mu, \nu) = \left(\widetilde{V}'\alpha, \widetilde{U}'\alpha\right)$.

So $L_{\theta,x}$ is not injective. This ends the proof.

Uniform Fourier convergence in the random coefficients model (normal errors). Consider model (2) with normal errors, where in addition we assume that Σ is *known*. We also assume that $\text{rank}(B) = q$, i.e. that $L_{\theta,x}$ is injective.

Let us take $\pi_\alpha = 1$, and $\pi_y(y) = \exp\left[-\frac{1}{2}\eta y' \Sigma^{-1} y\right]$, where $\eta > 0$. Let $Q = \Sigma^{-\frac{1}{2}} B \left[\Sigma^{-\frac{1}{2}} B\right]^{\dagger}$, and define V a $T \times q$ matrix such that $Q = VV'$ and $V'V = I_q$. Let also $W = I_T - Q$, and define U a $T \times (T - q)$ matrix such that $W = UU'$, and $U'U = I_{T-q}$.

Let us define $\mathcal{H}$ the Hilbert space of functions $\psi : \mathbb{R}^q \to \mathbb{R}$ such that:

$$\int_{\mathbb{R}^q} \psi(\mu)^2 \exp\left[-\frac{1}{2}\eta \mu'\mu\right] d\mu < \infty,$$

endowed with its canonical scalar product. Lastly, let $L_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ be the integral operator such that, for all $\psi \in \mathcal{H}$:

$$[L_{\mathcal{H}}\psi](z) = \int_{\mathbb{R}^q} \exp\left[-\frac{1}{4}(z - \mu)'(z - \mu)\right] \times \exp\left[-\frac{1}{2}\eta\mu'\mu\right] \psi(\mu)\, d\mu, \quad \text{for all } z \in \mathbb{R}^q.$$

We note that $L_{\mathcal{H}}$ is Hilbert-Schmidt, so it admits a singular value decomposition, and that $L_{\mathcal{H}}$ is self-adjoint.

We have the following result.

Proposition A1 *The left singular functions of the operator $L_{\theta,x} : \mathcal{G}_\alpha \to \mathcal{G}_y$ are given by:*

$$\phi_j(y) = C(\theta) H_j\left(V'\Sigma^{-\frac{1}{2}} y\right) \exp\left[-\frac{1}{2}(y - a)' \Sigma^{-\frac{1}{2}} UU' \Sigma^{-\frac{1}{2}} (y - a)\right], \tag{A2}$$

where H_j, $j = 1, 2, ...$ are the singular functions of the self-adjoint operator $L_{\mathcal{H}}$, and where $C(\theta)$ is a positive constant, uniformly bounded on Θ provided that $a()$ is continuous in θ and Θ is compact.

Proof.

Let $\mathcal{Y} = \mathbb{R}^T$, and $\mathcal{A} = \mathbb{R}^q$. We have:

$$\begin{aligned}
\left[L_{\theta,x} L^*_{\theta,x} h\right](y) &= \int_{\mathcal{Y}} \int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y|x,\alpha)\, f_{y|x,\alpha;\theta}(\widetilde{y}|x,\alpha)\, \pi_y(\widetilde{y})\, h(\widetilde{y})\, d\alpha d\widetilde{y} \\
&= \int_{\mathcal{Y}} \underbrace{\left\{\int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y|x,\alpha)\, f_{y|x,\alpha;\theta}(\widetilde{y}|x,\alpha)\, d\alpha\right\}}_{k(y,\widetilde{y})} \pi_y(\widetilde{y})\, h(\widetilde{y})\, d\widetilde{y}.
\end{aligned}$$

Moreover:

$$f_{y|x,\alpha;\theta}(y|x,\alpha) \propto \exp\left[-\frac{1}{2}\left(V'\Sigma^{-\frac{1}{2}}(y-a)-V'\Sigma^{-\frac{1}{2}}B\alpha\right)'\left(V'\Sigma^{-\frac{1}{2}}(y-a)-V'\Sigma^{-\frac{1}{2}}B\alpha\right)\right] \times \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right],$$

where $A \propto B$ denotes the fact that A and B are equal up to a multiplicative constant (possibly dependent on θ, x).

Using the change of variables $\beta = V'\Sigma^{-\frac{1}{2}}B\alpha$, and noting that $V'\Sigma^{-\frac{1}{2}}B$ is non-singular, we obtain:

$$\begin{aligned} k(y,\widetilde{y}) &= \int_{\mathcal{A}} f_{v|x;\theta}(y-a-B\alpha)\, f_{v|x;\theta}(\widetilde{y}-a-B\alpha)\, d\alpha \\ &\propto \int_{\mathcal{A}} \exp\Big[-\frac{1}{2}\left(V'\Sigma^{-\frac{1}{2}}(y-a)-\beta\right)'\left(V'\Sigma^{-\frac{1}{2}}(y-a)-\beta\right) \\ &\qquad -\frac{1}{2}\left(V'\Sigma^{-\frac{1}{2}}(\widetilde{y}-a)-\beta\right)'\left(V'\Sigma^{-\frac{1}{2}}(\widetilde{y}-a)-\beta\right)\Big] d\beta \\ &\quad \times \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)-\frac{1}{2}(\widetilde{y}-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(\widetilde{y}-a)\right]. \end{aligned}$$

So, from the usual decomposition of quadratic forms:

$$\begin{aligned} k(y,\widetilde{y}) &\propto \exp\left[-\frac{1}{4}\left(V'\Sigma^{-\frac{1}{2}}(y-\widetilde{y})\right)'\left(V'\Sigma^{-\frac{1}{2}}(y-\widetilde{y})\right)\right] \\ &\quad \times \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)-\frac{1}{2}(\widetilde{y}-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(\widetilde{y}-a)\right]. \end{aligned}$$

As the left singular function ϕ_j belongs to the range of $L_{\theta,x}$, there exists a function h_j such that:

$$\phi_j(y) = h_j\left(V'\Sigma^{-\frac{1}{2}}y\right)\exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right].$$

The function ϕ_j satisfies:

$$\left[L_{\theta,x}L^*_{\theta,x}\phi_j\right](y) \propto \phi_j(y).$$

This is equivalent to:

$$\begin{aligned} h_j\left(V'\Sigma^{-\frac{1}{2}}y\right) &\propto \int_{\mathcal{Y}}\Big\{\exp\left[-\frac{1}{4}\left(V'\Sigma^{-\frac{1}{2}}(y-\widetilde{y})\right)'\left(V'\Sigma^{-\frac{1}{2}}(y-\widetilde{y})\right)\right] \\ &\qquad \times \exp\left[-\frac{1}{2}(\widetilde{y}-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(\widetilde{y}-a)\right]\pi_y(\widetilde{y})\, h_j\left(V'\Sigma^{-\frac{1}{2}}\widetilde{y}\right)\Big\}d\widetilde{y}. \end{aligned}$$

Then, we note that, as $VV'+UU'=I_T$:

$$\begin{aligned} \pi_y(\widetilde{y}) &= \exp\left[-\frac{1}{2}\eta\widetilde{y}'\Sigma^{-1}\widetilde{y}\right] \\ &= \exp\left[-\frac{1}{2}\eta\left(V'\Sigma^{-\frac{1}{2}}\widetilde{y}\right)'V'\Sigma^{-\frac{1}{2}}\widetilde{y}\right]\times\exp\left[-\frac{1}{2}\eta\widetilde{y}'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}\widetilde{y}\right]. \end{aligned}$$

We thus obtain, using the change in variables $(\mu,\nu) = \left(V'\Sigma^{-\frac{1}{2}}\widetilde{y}, U'\Sigma^{-\frac{1}{2}}\widetilde{y}\right)$:

$$h_j\left(V'\Sigma^{-\frac{1}{2}}y\right) \propto \int_{\mathbb{R}^q}\exp\left[-\frac{1}{4}\left(V'\Sigma^{-\frac{1}{2}}y-\mu\right)'\left(V'\Sigma^{-\frac{1}{2}}y-\mu\right)\right]\exp\left[-\frac{1}{2}\eta\mu'\mu\right]h_j(\mu)\,d\mu.$$

So, (A2) follows. Lastly, as $\|\phi_j\| = 1$ the proportionality constant $C(\theta)$ satisfies:

$$\begin{aligned}
\frac{1}{C(\theta)^2} &= \int_{\mathcal{Y}} \left(H_j\left(V'\Sigma^{-\frac{1}{2}}y\right) \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right]\right)^2 \exp\left[-\frac{1}{2}\eta y'\Sigma^{-1}y\right] dy \\
&= |\Sigma|^{\frac{1}{2}} \int_{\mathbb{R}^q} H_j(\mu)^2 \exp\left[-\frac{1}{2}\eta\mu'\mu\right] d\mu \\
&\qquad \times \int_{\mathbb{R}^{T-q}} \exp\left[-\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)\right] \exp\left[-\frac{1}{2}\eta\nu'\nu\right] d\nu \\
&= |\Sigma|^{\frac{1}{2}} \left(\frac{2\pi}{2+\eta}\right)^{\frac{T-q}{2}} \exp\left[-\frac{\eta}{2+\eta}a'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}a\right],
\end{aligned}$$

where we have used that $\|H_j\| = 1$. As $a()$ is continuous in θ and Θ is compact, and as $W = UU'$ is a projector, $a'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}a$ is bounded. So, $C(\theta)$ is uniformly bounded.

The result follows.

■

Using the expression for the left singular functions, we then verify uniform Fourier convergence for model (2).

Corollary A1 *The following condition is satisfied for any $h \in \mathcal{G}_y$, a.s. in x:*

$$\sup_{\theta\in\Theta}\left(\sum_{j>J}\left\langle \phi_j, h\right\rangle^2\right) \overset{J\to\infty}{\to} 0. \tag{A3}$$

Proof.

We start by checking Condition (A3) when h is a polynomial. It is enough to check the result for h of the form $\left(\Sigma^{-\frac{1}{2}}y\right)^{(k)}$, where $y^{(k)} = y_1^{k_1} \times ... \times y_T^{k_T}$. Let $(\mu,\nu) = \left(V'\Sigma^{-\frac{1}{2}}y, U'\Sigma^{-\frac{1}{2}}y\right)$. We have:

$$\left(\Sigma^{-\frac{1}{2}}y\right)^{(k)} = \left(VV'\Sigma^{-\frac{1}{2}}y + UU'\Sigma^{-\frac{1}{2}}y\right)^{(k)} = (V\mu + U\nu)^{(k)}.$$

We note that $(V\mu + U\nu)^{(k)}$ is a polynomial in μ and ν, the coefficients of which are uniformly bounded as U and V are orthogonal matrices. So it is sufficient to check the result for h of the form $\left(V'\Sigma^{-\frac{1}{2}}y\right)^{(m)}\left(U'\Sigma^{-\frac{1}{2}}y\right)^{(\ell)}$.

For such an h, we have:

$$\begin{aligned}
\left\langle \phi_j, h\right\rangle &= C(\theta)\int_{\mathcal{Y}} \left\{ \left(V'\Sigma^{-\frac{1}{2}}y\right)^{(m)}\left(U'\Sigma^{-\frac{1}{2}}y\right)^{(\ell)} H_j\left(V'\Sigma^{-\frac{1}{2}}y\right) \right. \\
&\qquad \left. \times \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right]\pi_y(y)\right\} dy \\
&= C(\theta)|\Sigma|^{\frac{1}{2}}\int_{\mathbb{R}^q} \mu^{(m)} H_j(\mu) \exp\left[-\frac{1}{2}\eta\mu'\mu\right] d\mu \\
&\qquad \times \int_{\mathbb{R}^{T-q}} \nu^{(\ell)} \exp\left[-\frac{1}{2}\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)\right] \exp\left[-\frac{1}{2}\eta\nu'\nu\right] d\nu,
\end{aligned}$$

where we have factored π_y as in the proof of Proposition A1, and where we have used the change in variables $(\mu,\nu) = \left(V'\Sigma^{-\frac{1}{2}}y, U'\Sigma^{-\frac{1}{2}}y\right)$.

Now, as $\mu^{(m)}$ belongs to $\mathcal{H}$:

$$\sum_{j>J}\left(\int_{\mathbb{R}^q}\mu^{(m)}H_j\left(\mu\right)\exp\left[-\frac{1}{2}\eta\mu'\mu\right]d\mu\right)^2 \overset{J\to\infty}{\to} 0.$$

In addition:

$$\left|\int_{\mathbb{R}^{T-q}}\nu^{(\ell)}\exp\left[-\frac{1}{2}\left(\nu-U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu-U'\Sigma^{-\frac{1}{2}}a\right)\right]\exp\left[-\frac{1}{2}\eta\nu'\nu\right]d\nu\right| \leq \int_{\mathbb{R}^{T-q}}|\nu|^{(\ell)}\exp\left[-\frac{1}{2}\eta\nu'\nu\right]d\nu<\infty.$$

This shows uniform Fourier convergence for polynomial h.

Lastly let $h\in\mathcal{G}_y$, and fix $\varepsilon>0$. We start by noting that polynomials are dense in $\mathcal{G}_y$. For example, when $T=1$ the (generalized) Hermite polynomials form an orthogonal basis of the weighted L^2 space $\mathcal{G}_y$. So, there exists a polynomial $\widetilde{h}$ such that: $\left\|h-\widetilde{h}\right\|^2<\frac{\varepsilon}{4}$.

For this $\widetilde{h}$, and by the previous result, there exists a J_1 such that, for all $J\geq J_1$:

$$\sup_{\theta\in\Theta}\quad\sum_{j>J}\left\langle\phi_j,\widetilde{h}\right\rangle^2<\frac{\varepsilon}{4}.$$

Therefore:

$$\begin{aligned}\sup_{\theta\in\Theta}\quad\sum_{j>J}\left\langle\phi_j,h\right\rangle^2 &\leq \sup_{\theta\in\Theta}\quad\sum_{j>J}2\left(\left\langle\phi_j,\widetilde{h}\right\rangle^2+\left\langle\phi_j,h-\widetilde{h}\right\rangle^2\right)\\ &\leq 2\times\sup_{\theta\in\Theta}\quad\sum_{j>J}\left\langle\phi_j,\widetilde{h}\right\rangle^2+2\times\left\|h-\widetilde{h}\right\|^2\\ &< 2\times\frac{\varepsilon}{4}+2\times\frac{\varepsilon}{4}\\ &= \varepsilon,\end{aligned}$$

and the corollary is proved.

■

Average marginal effects in the random coefficients model (normal errors). As before we take $\pi_\alpha=1$, and $\pi_y\left(y\right)=\exp\left[-\frac{1}{2}\eta y'\Sigma^{-1}y\right]$, where $\eta>0$. Let us assume that $L_{\theta_0,x}$ is injective. Suppose that (53) holds, so that there exists $h\in\mathcal{G}_y$ such that $m=L^*_{\theta_0,x}h$ (as $\pi_\alpha=1$). If $L_{\theta_0,x}$ is non-surjective, there are many h that satisfy this equation. Without loss of generality we assume that:

$$h\in\mathcal{N}\left(L^*_{\theta_0,x}\right)^\perp=\overline{\mathcal{R}\left(L_{\theta_0,x}\right)}.$$

So, by (8), and using the same notation as in the proof of Proposition A1, there exists a function H such that:

$$h(y)=H\left(V'\Sigma^{-\frac{1}{2}}y\right)\exp\left(-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right),$$

where H is such that $\int_{\mathbb{R}^q}H\left(\mu\right)^2\exp\left[-\frac{1}{2}\eta\mu'\mu\right]d\mu<\infty$.

It thus follows that:

$$\begin{aligned} m(\alpha) &= \left[L^*_{\theta_0,x}h\right](\alpha) \\ &= \int_{\mathcal{Y}} f_{v|x,\alpha;\theta_0}(y|x,\alpha)\,\pi_y(y)\,h(y)\,dy \\ &\propto \int_{\mathbb{R}^q}\int_{\mathbb{R}^{T-q}} H(\mu)\exp\left[-\frac{1}{2}\left(\mu - V'\Sigma^{-\frac{1}{2}}(a+B\alpha)\right)'\left(\mu - V'\Sigma^{-\frac{1}{2}}(a+B\alpha)\right)\right] \\ &\qquad \times\exp\left[-\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)\right]\exp\left[-\frac{1}{2}\eta\mu'\mu\right]\times\exp\left[-\frac{1}{2}\eta\nu'\nu\right]d\mu d\nu \\ &\propto \int_{\mathbb{R}^q} H(\mu)\exp\left[-\frac{1}{2}\left(\mu - V'\Sigma^{-\frac{1}{2}}(a+B\alpha)\right)'\left(\mu - V'\Sigma^{-\frac{1}{2}}(a+B\alpha)\right)\right]\exp\left[-\frac{1}{2}\eta\mu'\mu\right]d\mu, \end{aligned}$$

where we have used the change in variables $(\mu,\nu) = \left(V'\Sigma^{-\frac{1}{2}}y, U'\Sigma^{-\frac{1}{2}}y\right)$.

Taking Fourier transforms we obtain, for $\tau\in\mathbb{R}^q$:

$$[\mathcal{F}m]\left(B'\Sigma^{-\frac{1}{2}}V\tau\right) \propto \left[\mathcal{F}\widetilde{H}\right](\tau)\,e^{-\sqrt{-1}\tau'V'\Sigma^{-\frac{1}{2}}a}e^{-\frac{1}{2}\tau'\tau}.$$

where $\widetilde{H}(\mu) = H(\mu)\exp\left[-\frac{1}{2}\eta\mu'\mu\right]$, and $\mathcal{F}$ is the L^2-Fourier transform operator. Note that $\mu\mapsto H(\mu)\exp\left[-\frac{1}{4}\eta\mu'\mu\right]$ belongs to $L^2(\mathbb{R}^q)$, so $\mathcal{F}\widetilde{H}$ is well-defined.

As $\mathcal{F}\widetilde{H}$ is square integrable, it follows that, as a consequence of (53):[43]

$$\tau\mapsto[\mathcal{F}m](\tau)\,e^{\frac{1}{2}\tau'\left(B'\Sigma^{-1}B\right)^{-1}\tau}$$

must be square integrable. This imposes restrictions on the rate at which $[\mathcal{F}m](\tau)$ tends to zero as $|\tau|$ tends to infinity.[44]

Operator injectivity in the censored random coefficients model (normal errors). In the censored random coefficients model, we define $\mathcal{A}=\mathbb{R}^q$, and $\mathcal{Y}=\{y\in\mathbb{R}^T, y_t\geq c_t \text{ for all } t\}$. The next proposition shows that $L_{\theta,x}$ is *injective* when v_{it} is normally distributed and B has full-column rank. We show the result for $\pi_\alpha = 1$.

Proposition A2 *Suppose that* $\operatorname{rank}[B(x,\theta)] = q$ *for all* θ, x*-a.s. Then* $L_{\theta,x}:\mathcal{G}_\alpha\to\mathcal{G}_y$ *is injective in Example 2.*

Proof.

In the proof we drop the reference to x to simplify the notation. Let $g\in\mathcal{G}_\alpha$ such that $L_\theta g = 0$. Then, for all $y > c$ (where $y > c$ denotes that $y_t > c_t$ for all t):

$$\int_{\mathcal{A}} f_v(y - a - B\alpha)\,g(\alpha)d\alpha = 0.$$

This implies:

$$\int_{\mathcal{A}} e^{-\frac{1}{2}[y-a-B\alpha]'\Sigma^{-1}[y-a-B\alpha]}g(\alpha)d\alpha = 0,$$

[43] Note that: $B'\Sigma^{-\frac{1}{2}}VV'\Sigma^{-\frac{1}{2}}B = B'\Sigma^{-1}B$.

[44] Condition (53) imposes *more* than square integrability, as $\widetilde{H}$ is the product of a function in $L^2(\mathbb{R}^q)$ with the rapidly decaying function $\mu\mapsto\exp\left[-\frac{1}{4}\eta\mu'\mu\right]$.

or equivalently:

$$\int_{\mathcal{A}} e^{-\frac{1}{2}\alpha' B'\Sigma^{-1}B\alpha} e^{(y-a)'\Sigma^{-1}B\alpha} g(\alpha)d\alpha = 0.$$

As $B'\Sigma^{-1}B$ is positive definite, one can differentiate under the integral sign and obtain, for all $y > c$, and all $k = (k_1, ..., k_T) \in \{0, 1, 2, ...\}^T$:

$$\int_{\mathcal{A}} e^{-\frac{1}{2}\alpha' B'\Sigma^{-1}B\alpha} \left[\Sigma^{-1}B\alpha\right]^{\otimes k} e^{(y-a)'\Sigma^{-1}B\alpha} g(\alpha)d\alpha = 0,$$

where

$$y^{\otimes k} = \underbrace{y_1 \otimes ... \otimes y_1}_{k_1 \text{ times}} \otimes ... \otimes \underbrace{y_T \otimes ... \otimes y_T}_{k_T \text{ times}}.$$

For any $0 < \eta < 1/2$ we thus have:

$$\int_{\mathcal{A}} \left(\left[\Sigma^{-1}B\alpha\right]^{\otimes k} e^{-\eta\alpha' B'\Sigma^{-1}B\alpha} \right) e^{-\left(\frac{1}{2}-\eta\right)\alpha' B'\Sigma^{-1}B\alpha} e^{(y-a)'\Sigma^{-1}B\alpha} g(\alpha)d\alpha \quad = \quad 0.$$

As B has full-column rank, $\left\{ \left[\Sigma^{-1}B\alpha\right]^{\otimes k} e^{-\eta\alpha' B'\Sigma^{-1}B\alpha}, k \in \{0, 1, 2, ...\}^T \right\}$ is a complete family in $L^2(\mathbb{R}^q)$.[45] It follows that:

$$e^{-\left(\frac{1}{2}-\eta\right)\alpha' B'\Sigma^{-1}B\alpha} e^{(y-a)'\Sigma^{-1}B\alpha} g(\alpha) \quad = \quad 0, \text{ a.s. in } \alpha, y > c,$$

which implies that $g = 0$. This ends the proof.

■

Lastly, note that, to show injectivity, it is important that the support of v_i be large enough. To see this, consider the simple model (with $T = 1$):

$$y_{i1} = \max\left(x_{i1}'\theta_0 + \alpha_i + v_{i1}, c_1\right),$$

where $\text{Supp}(v_{i1}) = [a, b]$. Clearly, if $\alpha \leq c_1 - x_1'\theta - b$, then $f_{v_1|x_1}(y_1 - x_1'\theta - \alpha) = 0$ for all $y_1 \geq c_1$. So, any function g in $\mathcal{G}_\alpha$ that is zero on $]c_1 - x_1'\theta - b, +\infty[$ belongs to the null-space of the operator $L_{\theta,x}$. Hence $L_{\theta,x}$ is not injective.

Random coefficients model (non-normal errors). Consider model (2), where now the distribution of v_i given x_i and α_i is known given θ (possibly non-normal), and is independent of α_i with zero mean. We let $\mathcal{Y} = \mathbb{R}^T$, $\mathcal{A} = \mathbb{R}^q$, and we take π_α and π_y such that Assumption 1 is satisfied.

We start by obtaining restrictions on $L_{\theta,x}g$ for $g \in \mathcal{G}_\alpha \cap L^1(\mathbb{R}^q)$. Note that, in this case, $L_{\theta,x}g \in \mathcal{G}_y \cap L^1\left(\mathbb{R}^T\right)$. Moreover, $\mathcal{G}_\alpha \cap L^1(\mathbb{R}^q)$ is dense in $\mathcal{G}_\alpha$,[46] and $\mathcal{G}_y \cap L^1\left(\mathbb{R}^T\right)$ is dense in $\mathcal{G}_y$.

[45] This is because polynomials form a complete family in the weighted L^2 space with weighting function $\pi(\alpha) = e^{-\eta\alpha' B'\Sigma^{-1}B\alpha}$. For example, for $q = 1$ the (generalized) Hermite polynomials are dense in that space.

[46] To see this, let $g \in \mathcal{G}_\alpha$ and consider $g_M(\alpha) = \mathbf{1}\{|\alpha| \leq M\}g(\alpha)$. We have:

$$\|g - g_M\|^2 = \int_{|\alpha|>M} g^2(\alpha)\pi_\alpha(\alpha)d\alpha \overset{M\to+\infty}{\to} 0.$$

We have, for any $g \in \mathcal{G}_\alpha$:

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

So, if in addition $g \in L^1(\mathbb{R}^q)$ we can take Fourier transforms and obtain:

$$[\mathcal{F}[L_{\theta,x} g]](\xi) = e^{\sqrt{-1}\xi' a} \cdot [\mathcal{F} g](B'\xi) \cdot \Psi_{v|x;\theta}(\xi|x), \quad \xi \in \mathbb{R}^T,$$

where $\Psi_{v|x;\theta} = \mathcal{F} f_{v|x;\theta}$ is the conditional characteristic function of v_i given x_i.

Denoting $W = I_T - BB^\dagger$, and noting that $B'W = 0$, we obtain:

$$[\mathcal{F}[L_{\theta,x} g]](\xi + W\chi|x)\, \Psi_{v|x;\theta}(\xi|x) = e^{\sqrt{-1}\chi' W a} [\mathcal{F}[L_{\theta,x} g]](\xi|x)\, \Psi_{v|x;\theta}(\xi + W\chi|x), \quad (\xi,\chi) \in \mathbb{R}^{2T}. \tag{A4}$$

Equation (A4) suggests that the non-surjectivity condition is satisfied unless $W = 0$, that is $\text{rank}(B) = T$. In addition, evaluating (A4) at $\theta = \theta_0$ and $g = f_{\alpha|x}$ yields:

$$\Psi_{y|x}(\xi + W\chi|x)\, \Psi_{v|x;\theta_0}(\xi|x) = e^{\sqrt{-1}\chi' W a} \Psi_{y|x}(\xi|x)\, \Psi_{v|x;\theta_0}(\xi + W\chi|x), \quad (\xi,\chi) \in \mathbb{R}^{2T},$$

that is:

$$\mathbb{E}\left[e^{\sqrt{-1}(\xi + W\chi)' y_i} \Psi_{v|x;\theta_0}(\xi|x_i) - e^{\sqrt{-1}\chi' W a} e^{\sqrt{-1}\xi' y_i} \Psi_{v|x;\theta_0}(\xi + W\chi|x_i) \,\middle|\, x_i \right] = 0, \quad (\xi,\chi) \in \mathbb{R}^{2T}. \tag{A5}$$

Equation (A5) shows that θ_0 satisfies a continuum of conditional moment restrictions, which are informative when $\text{rank}(B) < T$. Moreover, in this model those restrictions are analytical.

Nonlinear regression model (non-normal errors). Let us consider the model:

$$y_i = m(x_i, \alpha_i, \theta_0) + v_i, \quad i = 1, ..., N, \tag{A6}$$

where $m()$ is a known $T \times 1$ function. The distribution of v_i given x_i and α_i is known given θ_0, and is independent of α_i. For example, (A6) may be used to model nonlinear production functions with heterogeneous technology parameters. We define $\mathcal{Y} = \mathbb{R}^T$, $\mathcal{A} \subset \mathbb{R}^q$, and we take π_α and π_y such that Assumption 1 holds.

We make the following assumption.

Assumption A1 *For $\theta \in \Theta$ and with probability one:*

$$\overline{\{m(x, \alpha, \theta), \alpha \in \mathcal{A}\}} \underset{\neq}{\subset} \mathbb{R}^T, \tag{A7}$$

where the closure is relative to the Euclidean topology in $\mathbb{R}^T$.

Assumption A1 will typically hold if $T > \dim \alpha_i$, that is when the number of time periods is strictly greater than the number of heterogeneous components. In this case, the assumption rules out space-filling mappings (such as Peano curves) that map surjectively $\mathbb{R}^q$ onto $\mathbb{R}^T$. Assumption A1 will fail to hold, however, when $T = \dim \alpha_i$ and m is one-to-one.

As in the linear case we let $g \in \mathcal{G}_\alpha \cap L^1(\mathcal{A})$ and we derive restrictions on $L_{\theta,x}g$. We have:

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

Taking Fourier transforms we obtain:

$$[\mathcal{F}[L_{\theta,x}g]](\xi) = \left(\int_{\mathcal{A}} e^{\sqrt{-1}\xi' m(x,\alpha,\theta)} g(\alpha)\, d\alpha\right) \cdot \Psi_{v|x;\theta}(\xi|x), \quad \xi \in \mathbb{R}^T. \tag{A8}$$

We have the next result.

Proposition A3 *Let Assumption A1 hold, and assume that $\Psi_{v|x;\theta}$ does not vanish on $\mathbb{R}^T$. Then, for any $\mu \notin \overline{\{m(x,\alpha,\theta), \alpha \in \mathcal{A}\}}$ and any $g \in \mathcal{G}_\alpha \cap L^1(\mathcal{A})$:*

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}^T} e^{-\frac{1}{2}\varepsilon\xi'\xi} e^{-\sqrt{-1}\xi'\mu} \left(\frac{[\mathcal{F}[L_{\theta,x}g]](\xi)}{\Psi_{v|x;\theta}(\xi|x)}\right) d\xi = 0. \tag{A9}$$

Proof. Let $\varepsilon > 0$. We have, using the Fubini theorem:

$$\begin{aligned} A(\varepsilon) &\equiv \int_{\mathbb{R}^T} e^{-\frac{1}{2}\varepsilon\xi'\xi} e^{-\sqrt{-1}\xi'\mu} \left(\int_{\mathcal{A}} e^{\sqrt{-1}\xi' m(x,\alpha,\theta)} g(\alpha)\, d\alpha\right) d\xi \\ &= \int_{\mathcal{A}} \left(\int_{\mathbb{R}^T} e^{-\sqrt{-1}\xi'\mu} e^{\sqrt{-1}\xi' m(x,\alpha,\theta)} e^{-\frac{1}{2}\varepsilon\xi'\xi} d\xi\right) g(\alpha)\, d\alpha \\ &= \int_{\mathcal{A}} \left((2\pi)^{\frac{T}{2}} \varepsilon^{-\frac{T}{2}} e^{-\frac{1}{2\varepsilon}(\mu - m(x,\alpha,\theta))'(\mu - m(x,\alpha,\theta))}\right) g(\alpha)\, d\alpha, \end{aligned}$$

where we have used the expression of the Fourier transform of a Gaussian distribution.

Now, as μ does not belong to the closure of the range of $m()$:

$$\inf_{\alpha \in \mathcal{A}} |\mu - m(x,\alpha,\theta)|^2 \geq \eta > 0.$$

It thus follows that:

$$|A(\varepsilon)| \leq (2\pi)^{\frac{T}{2}} \varepsilon^{-\frac{T}{2}} e^{-\frac{\eta}{2\varepsilon}} \int_{\mathcal{A}} |g(\alpha)|\, d\alpha \underset{\varepsilon \to 0}{\to} 0.$$

Lastly, by (A8) and the fact that $\Psi_{v|x;\theta}$ is non-vanishing:

$$A(\varepsilon) = \int_{\mathbb{R}^T} e^{-\frac{1}{2}\varepsilon\xi'\xi} e^{-\sqrt{-1}\xi'\mu} \left(\frac{[\mathcal{F}[L_{\theta,x}g]](\xi)}{\Psi_{v|x;\theta}(\xi|x)}\right) d\xi.$$

This ends the proof.

■

Proposition A3 provides a set of restrictions on $L_{\theta,x}g$, which is non-empty when Assumption A1 holds. This suggests that $L_{\theta,x}$ is not surjective under that assumption, provided that $\Psi_{v|x;\theta}$ is non-vanishing. This last assumption is commonly made in the nonparametric deconvolution literature (e.g., Carrasco and Florens, 2009).

In addition, the proposition allows us to derive simple restrictions on θ_0. Evaluating (A9) at $\theta = \theta_0$ and $g = f_{\alpha|x}$ we obtain, for any μ outside the closure of the range of $m(x,.;\theta_0)$:

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}^T} e^{-\frac{1}{2}\varepsilon\xi'\xi} e^{-\sqrt{-1}\xi'\mu} \frac{\Psi_{y|x}(\xi|x)}{\Psi_{v|x;\theta_0}(\xi|x)} d\xi = 0.$$

This yields a continuum of restrictions on θ_0 (indexed by μ), when Assumption A1 holds.

Static probit model. To see why finding a non-zero $\{\varphi_y\}$ that satisfies (14) is equivalent to all 2^T products of distinct F's being linearly dependent: $F_1^{k_1} \times ... \times F_T^{k_T}$, $(k_1, ..., k_T) \in \{0,1\}^T$, consider the case $T = 2$. Then, (14) can be written as:

$$\varphi_{00} + (\varphi_{10} - \varphi_{00}) F_1 + (\varphi_{01} - \varphi_{00}) F_2 + (\varphi_{11} - \varphi_{10} - \varphi_{01} + \varphi_{00}) F_1 F_2 = 0,$$

and we have:

$$\begin{pmatrix} \varphi_{00} \\ \varphi_{10} - \varphi_{00} \\ \varphi_{01} - \varphi_{00} \\ \varphi_{11} - \varphi_{10} - \varphi_{01} + \varphi_{00} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_{00} \\ \varphi_{10} \\ \varphi_{01} \\ \varphi_{11} \end{pmatrix}.$$

This triangular structure holds for any $T \geq 2$.

Static logit model. We prove (16). We have:

$$\begin{aligned}
(15) \quad &\Leftrightarrow \quad \sum_{y \in \{0,1\}^T} \varphi_y(x,\theta) \prod_{t=1}^{T} \Lambda\left(x_t'\theta + \alpha\right)^{y_t} \left(1 - \Lambda\left(x_t'\theta + \alpha\right)\right)^{1-y_t} = 0 \\
&\Leftrightarrow \quad \sum_{y \in \{0,1\}^T} \varphi_y(x,\theta) \prod_{t=1}^{T} \left[\frac{e^{x_t'\theta+\alpha}}{1+e^{x_t'\theta+\alpha}}\right]^{y_t} \left[\frac{1}{1+e^{x_t'\theta+\alpha}}\right]^{1-y_t} = 0 \\
&\Leftrightarrow \quad \sum_{y \in \{0,1\}^T} \varphi_y(x,\theta) e^{\sum_{t=1}^{T} y_t (x_t'\theta+\alpha)} = 0 \\
&\Leftrightarrow \quad \sum_{y \in \{0,1\}^T} \varphi_y(x,\theta) e^{\sum_{t=1}^{T} y_t x_t'\theta} e^{\alpha \sum_{t=1}^{T} y_t} = 0.
\end{aligned}$$

So, as $e^{s\alpha}$, $s = 0, ..., T$, are linearly independent, (16) follows.

B Specification test

In applied work, a common approach is to assume a parametric model for the individual effects. Here we show how to use the functional differencing restrictions for the purpose of specification testing.

Let:

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

be a complete parametric specification of the distribution of the data, which includes a parametric model for the individual effects. A popular choice is to let $f_{\alpha|x;\eta_0}(\alpha|x)$ be a Gaussian density, with means and variances that are parsimonious functions of covariates x_i (Chamberlain, 1984).

We wish to test the null hypothesis that $f_{\alpha|x}$ is correctly specified. For this, we consider the random-effects maximum likelihood estimator (MLE) of θ_0, which solves:

$$\widetilde{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\underset{\eta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln \left(\int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y_i|x_i,\alpha) f_{\alpha|x;\eta}(\alpha|x_i)\, d\alpha \right) \right].$$

Then, we define the following statistic:

$$S = \frac{1}{N} \sum_{i=1}^{N} \varphi\left(y_i, x_i, \widetilde{\theta}\right),$$

where $\varphi = (\varphi_1, ..., \varphi_R)'$, with φ_r given by (35). The statistic S is simply an empirical counterpart of the functional differencing moment restrictions, evaluated at the random-effects MLE.

Proposition B1 *Under the null of correct specification, and under regularity conditions given in Section 5 and standard regularity assumptions on the MLE:*

$$\sqrt{N}S \xrightarrow{d} N\left[0, V_S\right],$$

where the expression of V_S is provided in equation (B1) below.

Proof.

Let us denote $\ell_i(\theta, \eta) = \ln\left[\int_{\mathcal{A}} f_{y|x,\alpha;\theta}(y_i|x_i,\alpha) f_{\alpha|x;\eta}(\alpha|x_i)\, d\alpha\right]$, and $L_{\theta\theta} = \mathbb{E}\left[\frac{\partial^2 \ell_i(\theta_0,\eta_0)}{\partial\theta\partial\theta'}\right]$, with a similar notation for the three other components of the Hessian: $L_{\theta\eta}$, $L_{\eta\theta}$, and $L_{\eta\eta}$. Then, under standard regularity conditions and under the null of correct specification:

$$\sqrt{N}\left(\widetilde{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, V_{\widetilde{\theta}}\right],$$

where $V_{\widetilde{\theta}} = \left[L_{\theta\theta} - L_{\theta\eta}L_{\eta\eta}^{-1}L_{\eta\theta}\right]^{-1}$.

Let $\varphi_i(\theta) = \varphi(y_i, x_i, \theta)$. It is easy to show that, under the null, and under the regularity conditions of Theorem 5 and standard regularity assumptions on the MLE (see Arellano, 1991):

$$\sqrt{N}S \xrightarrow{d} N\left[0, V_S\right],$$

where:

$$V_S = \mathbb{E}\left[\left(\varphi_i(\theta_0) - GV_{\widetilde{\theta}}s_i\right)\left(\varphi_i(\theta_0) - GV_{\widetilde{\theta}}s_i\right)'\right], \tag{B1}$$

with $s_i = \frac{\partial \ell_i(\theta_0,\eta_0)}{\partial\theta} - L_{\theta\eta}L_{\eta\eta}^{-1}\frac{\partial \ell_i(\theta_0,\eta_0)}{\partial\eta}$, and $G = \mathbb{E}\left[\frac{\partial\varphi_i(\theta_0)}{\partial\theta'}\right]$.

A consistent estimator of V_S is then obtained as:

$$\widehat{V}_S = \widehat{\mathbb{E}}\left[\left(\varphi_i\left(\widetilde{\theta}\right) - \widehat{G}\widehat{V}_{\widetilde{\theta}}\widehat{s}_i\right)\left(\varphi_i\left(\widetilde{\theta}\right) - \widehat{G}\widehat{V}_{\widetilde{\theta}}\widehat{s}_i\right)'\right],$$

where $\widehat{V}_{\widetilde{\theta}}$ is a consistent estimator of $V_{\widetilde{\theta}}$, $\widehat{s}_i = \frac{\partial \ell_i(\widetilde{\theta},\widetilde{\eta})}{\partial\theta} - \widehat{L}_{\theta\eta}\widehat{L}_{\eta\eta}^{-1}\frac{\partial \ell_i(\widetilde{\theta},\widetilde{\eta})}{\partial\eta}$, with $\widehat{L}_{\theta\eta}$ and $\widehat{L}_{\eta\eta}$ consistent estimators of $L_{\theta\eta}$ and $L_{\eta\eta}$, respectively, and $\widehat{G}$ is given by (39) with $\widetilde{\theta}$ in place of $\widehat{\theta}$.

■

Let us assume that V_S is non-singular. In particular, this requires that the vector of moment functions φ is not identically zero, thus restricting the model to be non-surjective. As N tends to infinity we then have, under the null of correct specification:

$$NS'\widehat{V}_S^{-1}S \xrightarrow{d} \chi^2_R, \tag{B2}$$

where $\widehat{V}_S$ is a consistent estimator of V_S.[47] Thus, (B2) provides a simple way to test the validity of random-effects specifications in non-surjective models. This provides an analog of the Hausman test (Hausman, 1978) in a nonlinear context.

[47] As shown in Sections 5 and 6, the problem of estimating V_S (which involves the Jacobian matrix G) is generally ill-posed. In particular, $\widehat{V}_S$ may not be root-N consistent for V_S.

References

[1] Arellano, M. (1991): "Moment Testing with non-ML Estimators," *mimeo.*

[2] Carrasco, M., and J. P. Florens (2009): "Spectral Methods for Deconvolving a Density," to appear in *Econometric Theory.*

[3] Chamberlain, G. (1984): "Panel Data", in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Elsevier Science.

[4] Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

[5] Yoshida, K. (1971): *Functional Analysis.* Springer Verlag. New York.