Functional Differencing

Stéphane Bonhomme

New York University

This draft: February 2010
PRELIMINARY AND INCOMPLETE

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 parameter problem, integral operators, inverse problems.

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 to allow one or various parameters to be unit-specific, in order to 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 parameters 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 linear panel data models, differencing out the individual effects yields moment restrictions on common parameters alone. In various nonlinear models, ingenious methods have been proposed to difference out the fixed effects. 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. 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. 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. Then, 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 are orthogonal to the range (or image) of that mapping. By construction, such functions will be orthogonal to the distribution function of the data, providing 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 are orthogonal to 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 (1992a), the censored regression model of Honoré (1992), and the static logit model. In all three examples, our systematic characterization of the range of the model delivers the proposed methods as special cases. Moreover, as our approach is general, we may use it in models where differencing strategies are not yet known. We illustrate this point with a censored random coefficients model.[4]

In a given nonlinear model, there may exist no solution to the problem at hand. 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 choice models are generally surjective, with the important exception of the logit. In those models,

[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.

our approach will not be informative about common parameters.

To describe our approach to construct moment functions, 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 matrix of conditional probabilities that depends on common parameters. Elements that belong to the orthogonal 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, when the columns of the matrix that defines the model are linearly independent, our differencing strategy achieves the semiparametric information bound of the panel data model. This is not true more generally, however. Additional restrictions can be obtained by imposing that the probabilities of individual effects lie in the unit interval. Exploiting those additional constraints will be essential in panel data models with discrete dependent variables that do not satisfy the non-surjectivity condition.[5]

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 distributions with scalar products, making them Hilbert spaces. Moreover, we impose regularity conditions that ensure operator compactness. This construction allows us to define a "within" projection operator that projects functions of the dependent variables onto the orthogonal 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 computed in a convenient basis of functions.

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 those models, 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. Under identification and regularity conditions that we provide, functional differencing estimates of common parameters will be 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. In addition, we show how to use the functional differencing moment restrictions to test parametric random-effects specifications, which are popular in applied work. This provides an analog of the Hausman specification test (Hausman, 1978) in a nonlinear setting.

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. This situation contrasts with the estimation of common parameters. Unlike the (generalized) inverse of the model operator, the within projection operator is continuous in its functional argument, so ill-posedness does *not* arise in the estimation of common parameters.

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

[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).

Section 4, where we introduce some elements of Hilbert space theory and define the within operator. Estimation of common parameters and average marginal effects is presented in Sections 5 and 6, respectively. We then provide the asymptotic theory of our estimators in Section 7. Section 8 presents a small-scale numerical illustration. Lastly, Section 9 concludes.

2 Incidental parameters

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

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), and 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 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 $f_{\alpha|x}$. 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. 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.

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

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

where a $(T \times 1)$ and B $(T \times \dim \alpha_i)$ are known given x and θ. Chamberlain (1992a) 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.[7]

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

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

where $\Sigma(.,.)$ 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:

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

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:

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

where $v_i | x_i, \alpha_i \sim N[0, \Sigma(x_i, \theta_0)]$. The difference with Example 1 is that only $y_{it} = \max(y_{it}^*, c_t)$ is observed, where for simplicity 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 $t \in \{1, ..., T\}$.

When the model includes a single heterogeneous intercept: $y_{it} = \max(x_{it}'\beta_0 + \alpha_i + v_{it}, c_t)$, Honoré (1992) has derived restrictions on β_0. His approach does not require errors to be

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

normally distributed, though it relies on an i.i.d. assumption. To our knowledge, no solution has been proposed to deal with censored models with general random coefficients.[8]

Example 3: Static binary choice model. Our third example is a static panel data model with a binary dependent variable. That is, we assume that $y_{it} \in \{0, 1\}$ 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, 1992b).

2.2 Orthogonality

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.

Our approach relies on the representation (1), understood as a mapping that, for a given value of θ, relates the distribution function of individual effects to that of the data. We will denote this mapping as $L_{\theta,x}$, and write:

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

The mapping $L_{\theta,x}$ is an integral operator which maps functions $g(\alpha)$ to functions $[L_{\theta,x} g](y)$. In particular, (1) is equivalent to: $L_{\theta_0,x} f_{\alpha|x} = f_{y|x}$. We differ the precise mathematical definition and properties of $L_{\theta,x}$ until Section 4.

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

To derive restrictions on common parameters alone, we proceed as follows. We start by characterizing the *range* of the model operator, that is the set of functions $[L_{\theta,x}g](y)$ that can be obtained by using all functions $g(\alpha)$:

$$\mathcal{R}(L_{\theta,x}) = \{L_{\theta,x}g, \text{ for all functions } g\}.$$

Then, we look for functions that are *orthogonal* to $\mathcal{R}(L_{\theta,x})$. Assuming that we have found a function $\varphi(.,x,\theta)$ which is orthogonal to any function of the form $L_{\theta_0,x}g$, it then necessarily follows that $\varphi(.,x,\theta_0)$ is orthogonal to $L_{\theta_0,x}f_{\alpha|x} = f_{y|x}$. So, by orthogonality, θ_0 satisfies the following conditional moment restriction (denoting as $\mathcal{Y}$ the support of y_i):

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

Thus, differencing out the "incidental" individual effects amounts to solving a well-defined mathematical problem: finding some functions that are orthogonal to the space of functions $\mathcal{R}(L_{\theta,x})$. When the solutions to this problem are not available in closed form, we will show how to compute them numerically. 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 the general approach outlined here in our three main examples.

Example 1 (cont.) We introduce some useful 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 $W = I_T - Q$. Note that Q and W are orthogonal projectors, and that $W\Sigma^{-\frac{1}{2}}B = 0$. Note also the identity:

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

Denoting as $\mathcal{A}$ the support of α_i we have, for any function $g(\alpha)$:

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

So, if we find a function φ that is orthogonal to:

$$h\left(Q\Sigma^{-\frac{1}{2}}y\right)\exp\left[-\frac{1}{2}\left(y-a\right)'\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}\left(y-a\right)\right] \tag{5}$$

for *any* function h, then φ will be orthogonal to the range $\mathcal{R}\left(L_{\theta,x}\right)$. 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 themselves, it is easy to see that if we define $\varphi\left(y\right) = \Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}\left(y-a\right)$ then φ is orthogonal to all functions in $\mathcal{R}\left(L_{\theta,x}\right)$. This implies:

$$\mathbb{E}\left[\Sigma^{-\frac{1}{2}}W\Sigma^{-\frac{1}{2}}\left(y_i-a\right)|x_i\right]=0. \tag{6}$$

In a version of model (2) that only assumes $\mathbb{E}\left(v_i|x_i,\alpha_i\right)=0$, Chamberlain (1992a) shows that basing the estimation of θ_0 on the *generalized* within-group conditional moment restrictions (6) 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:

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

Lastly, note that for this approach to have content we need that there exists *some* non-zero function that is orthogonal to $\mathcal{R}\left(L_{\theta,x}\right)$. This will require the range not to be dense in the whole space of functions, according to a certain topology to be defined below. We will refer to this condition as *non-surjectivity*. In the present case, non-surjectivity will be satisfied provided that $\operatorname{rank}Q<T$, 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 $y>c$:[9]

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

As an interesting special case, let us start with a simple censored regression model with heterogeneous intercept: $y_{it}=\max\left(x_{it}'\beta_0+\alpha_i+v_{it},0\right)$, $T=2$, 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 $y_1>0, y_2>0$:

$$\left[L_{\theta,x}g\right]\left(y\right) \;=\; h\left(\overline{y},x\right)\exp\left[-\frac{1}{4\sigma^2}\left(\Delta y-\Delta x'\beta\right)^2\right], \tag{7}$$

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

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 (7), 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:

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

and the following function supported on that rectangle:

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

Orthogonality will hold provided that:

$$\int_c^d \varphi_2(\nu) \exp\left[-\frac{1}{4\sigma^2}(\nu - \Delta x'\beta)^2\right] d\nu = 0. \tag{8}$$

In particular, (8) 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).[10] Note that, as shown by Honoré, those restrictions do not depend on the normality assumption, and will be satisfied when the errors v_{i1} and v_{i2} are i.i.d. However, when assuming normality, our approach suggests additional restrictions which can be obtained by constructing other functions φ_2 (possibly dependent on x) such that (8) holds. This strategy will also deliver restrictions on σ_0^2.[11]

This 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 us define V 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 us denote $(\mu, \nu) = \left(V'\Sigma^{-\frac{1}{2}}y, U'\Sigma^{-\frac{1}{2}}y\right)$.

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

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

[10] 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.

[11] 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 (8).

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:

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

Orthogonality will hold if φ_2 and R_2 are chosen such that:

$$\int_{R_2} \varphi_2(\nu) \exp\left[-\frac{1}{2}\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)'\left(\nu - U'\Sigma^{-\frac{1}{2}}a\right)\right] d\nu = 0. \tag{9}$$

This example and the previous one suggest that, in a likelihood model with continuous or censored dependent variables, it may be possible to derive many (in effect, a continuum of) restrictions on common parameters. This paper proposes a systematic way to generate those restrictions by constructing *all* the functions that are orthogonal to the range of $L_{\theta,x}$.

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

$$\sum_{y\in\{0,1\}^T} \varphi(y,x,\theta)\Pr(y|x,\alpha,\theta) = 0, \quad x,\alpha - \text{a.s.} \tag{10}$$

that is, such that:

$$\sum_{y\in\{0,1\}^T} \varphi(y,x,\theta)\prod_{t=1}^{T} F_t^{y_t}(1-F_t)^{1-y_t} = 0, \quad x,\alpha - \text{a.s.} \tag{11}$$

It can be shown that finding a $\{\varphi(y,x,\theta)\}$ that satisfies (10) 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 Appendix C). F_t being a nonlinear function of individual effects, finding such a φ is often impossible. The reason is that the range of the mapping $L_{\theta,x}$ is likely to span the whole space of vectors in $\{0,1\}^T$. An example is the static probit model, where $F_t = \Phi(x_{it}'\theta + \alpha_i)$, with Φ the standard normal cdf. This situation contrasts with Examples 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, $F_t = \Lambda(x_{it}'\theta + \alpha_i)$, where $\Lambda(u) = e^u/(1+e^u)$ is the standard logistic cdf. We show in Appendix C that (11) is equivalent to:

$$\sum_{y\in\{0,1\}^T} \mathbf{1}\left\{\sum_{t=1}^{T} y_t = s\right\} \varphi(y,x,\theta) e^{\sum_{t=1}^{T} y_t x_t'\theta} = 0, \quad \text{for all } s \in \{0,1,...,T\}. \tag{12}$$

This system of equations has non-trivial solutions as soon as $T \geq 2$. For example, if $T = 2$, (12) implies that: $\varphi_{00} = \varphi_{11} = 0$, and:

$$\varphi_{10} e^{x_{i1}'\theta} + \varphi_{01} e^{x_{i2}'\theta} = 0, \tag{13}$$

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:

$$\mathbb{E}\left(e^{[x_{i2} - x_{i1}]'\theta_0} y_{i1} \left[1 - y_{i2}\right] - \left[1 - y_{i1}\right] y_{i2} | x_i \right) = 0, \tag{14}$$

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

In non-logistic binary choice models, the information bound for θ_0 is zero (Chamberlain, 1992b). The present discussion suggests that those models are surjective, implying that our approach will not yield informative restrictions on θ_0. This result is related to Johnson (2004) who shows that, in discrete choice panel data models, common parameters are unidentified unless equation (10) holds for at least some value of the covariates, and that when (10) does not hold for any value of x the information bound for θ_0 is zero.[12]

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:

$$f_{y|x} = L_{\theta_0, x} f_{\alpha|x}, \tag{15}$$

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).

[12] 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.

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:

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

and:

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

When supports are finite, the range of the matrix $L_{\theta,x}$ is a 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:

$$W_{\theta,x} = I_{N_y} - L_{\theta,x} L_{\theta,x}^{\dagger}, \tag{16}$$

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 $N_y \times N_y$ matrix $W_{\theta,x}$ is simply the orthogonal projection matrix on the null-space of $L'_{\theta,x}$. As such it is symmetric and idempotent. In particular, because $L_{\theta,x}^{\dagger}$ is a generalized inverse, $W_{\theta,x}$ is such that:

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

and:

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

The within projection matrix satisfies our purpose, as it projects vectors of $\mathbb{R}^{N_y}$ onto the orthogonal of the range of the matrix $L_{\theta,x}$. So, given any vector $h \in \mathbb{R}^{N_y}$, the vector $\varphi_{\theta,x} = W_{\theta,x} h \in \mathbb{R}^{N_y}$ is orthogonal to the columns of $L_{\theta,x}$. Moreover, *any* element that is orthogonal to the columns of $L_{\theta,x}$ is of the form $W_{\theta,x} h$, for some h.

Let us denote as $\varphi\left(\underline{y}_s, x, \theta\right)$ the sth element of $\varphi_{\theta,x}$, where $\underline{y}_s, s = 1, ..., N_y$, index the points of support of y_i. Then, as $\varphi_{\theta,x}$ is orthogonal to the columns of $L_{\theta,x}$, it follows that:

$$\mathbb{E}\left[\varphi\left(y_i, x_i, \theta_0\right) | x_i\right] = 0. \tag{17}$$

To interpret our approach, note that the moment restrictions are obtained by left-multiplying (15) 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:

$$W_{\theta_0,x} f_{y|x} \quad = \quad 0. \tag{18}$$

The functional differencing restrictions (18) are thus obtained by differencing out the probability distribution function of individual effects, yielding a set of restrictions on θ_0 alone. This is reminiscent of first-differencing and within-group approaches commonly used in linear panel data models.

As a second interpretation, notice that $W_{\theta,x} h = h - L_{\theta,x} L_{\theta,x}^{\dagger} h$ is the least-squares residual in the linear regression of a vector $h \in \mathbb{R}^{N_y}$ on the columns of the matrix $L_{\theta,x}$. 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 (17) can be obtained as residuals in a linear regression. Bajari *et al.* (2009) use a related idea in a game-theoretic context.

Lastly, note that the moment restrictions (17) are uninformative about θ_0 when the rows of $L_{\theta,x}$ are linearly independent (i.e., when $\text{Rank}\left(L_{\theta,x}\right) = N_y$), as in this case the null-space of $L'_{\theta,x}$ is zero. For example, if $L_{\theta,x}$ is square and non-singular then the Moore-Penrose inverse coincides will the standard matrix inverse, and $W_{\theta,x} = I_{N_y} - L_{\theta,x} L_{\theta,x}^{-1} = 0$. Thus, our differencing approach requires a condition of *non-surjectivity* to be satisfied. In the finite support case, this condition is automatically satisfied when $N_y > N_\alpha$.

Example 3 (cont.) For example, the non-surjectivity condition will be satisfied in the static binary choice model 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, an exception being when errors are logistic.

3.2 Estimation

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 us denote as $\tau\left(y_i\right)$ the index in $\{1, ..., N_y\}$ such that $y_i = \underline{y}_{\tau(y_i)}$. Let also $\omega_{\theta,x}\left[s_1, s_2\right]$ denote the (s_1, s_2)th element of the matrix $W_{\theta,x}$. Then, (18) is equivalent to:

$$\mathbb{E}\left(\omega_{\theta_0,x_i}\left[s, \tau\left(y_i\right)\right] | x_i\right) \quad = \quad 0, \quad s \in \{1, ..., N_y\}. \tag{19}$$

This motivates estimating θ_0 using the following generalized method-of-moments (GMM) estimator, which relies on a set of $R \geq 1$ unconditional moment restrictions based on (19):

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \upsilon_{r_1,r_2} \widehat{\mathbb{E}} \left[\varphi_{r_1} (y_i, x_i, \theta) \right] \widehat{\mathbb{E}} \left[\varphi_{r_2} (y_i, x_i, \theta) \right], \tag{20}$$

where

$$\varphi_r (y_i, x_i, \theta) = \omega_{\theta, x_i} [s_r, \tau (y_i)] \zeta_r (x_i), \tag{21}$$

and where $\widehat{\mathbb{E}} (z_i) = \frac{1}{N} \sum_{i=1}^{N} z_i$ denotes an empirical mean, $\zeta_1, ..., \zeta_R$ are functions of covariates, $s_1, ..., s_R$ are indexes in $\{1, ..., N_y\}$, and $\Upsilon = [\upsilon_{r_1,r_2}]_{(r_1,r_2) \in \{1,...,R\}^2}$ is a symmetric weight 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 7 cover finite supports as a special case, so we refer the reader to that section for the expression of the asymptotic variance.

Note that non-surjectivity is clearly necessary for θ_0 to be point-identified, as when the model is surjective $W_{\theta,x}$ is identically zero. However, it is not sufficient. From Lemma 2.3 in Newey and McFadden (1994), a sufficient condition is that $f_{y|x}$ does not belong to the range of $L_{\theta,x}$ for *any* $\theta \neq \theta_0$, with positive probability in x. In practice, local point-identification may be verified by checking that a rank condition is satisfied at θ_0.

Lastly, note that the standard regularity conditions in GMM require the moment functions to be continuous in θ. In the present case, in addition to imposing smoothness conditions on $\theta \mapsto f_{y|x,\alpha;\theta} (y|x, \alpha)$, this requires that the rank of $L_{\theta,x}$ be constant on the parameter space Θ, almost surely in x. Then, Corollary 3.5 in Stewart (1977) shows that $\theta \mapsto L^{\dagger}_{\theta,x}$ is continuous on Θ, a.s in x, implying the continuity of the within projection matrix $W_{\theta,x}$. Rank constancy is intuitively necessary to ensure the continuity of a projection matrix, as variations in the rank of $L_{\theta,x}$ induce jumps in its number of non-zero eigenvalues. In particular, rank constancy will be satisfied if $L_{\theta,x}$ has almost surely full column rank.

Optimal instruments. Following Chamberlain (1987), it is possible to derive the optimal instruments for this GMM estimation problem. For this purpose, it is useful to introduce a $N_y \times (N_y - \operatorname{rank} L_{\theta,x})$ 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. As a convention, we denote the rows of a matrix A as $A[s, .]$ and its columns as $A[., s]$.

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

$$\mathbb{E}\left(U_{\theta_0,x_i}\left[\tau\left(y_i\right),.\right]\kappa_{\theta_0,x_i}^{-1}\mathbb{E}\left[U'_{\theta_0,x_i}\frac{\partial L_{\theta_0,x_i}}{\partial\theta_k}L^{\dagger}_{\theta_0,x_i}\left[.,\tau\left(y_i\right)\right]\Big|x_i\right]\right) = 0, \quad k = 1,...,\dim\theta. \tag{22}$$

The optimal instruments in (22) are infeasible. To construct feasible counterparts and estimate θ_0 efficiently on the basis of the functional differencing restrictions (19), one can follow the approach of Newey (1990) and replace the unknown conditional expectations by series estimators.

Example 3 (cont.). Consider the static logit model with $T = 2$. Chamberlain's optimal unconditional moment restrictions in (14) are:

$$\mathbb{E}\left[\left(x_{i2} - x_{i1}\right)\frac{1}{e^{\left[x_{i2}-x_{i1}\right]'\theta_0}+1}\left(e^{\left[x_{i2}-x_{i1}\right]'\theta_0}y_{i1}\left[1-y_{i2}\right] - \left[1-y_{i1}\right]y_{i2}\right)\right] = 0. \tag{23}$$

This coincides exactly with the score equations of the conditional maximum likelihood estimator (CMLE) based on the sufficient statistic $y_{i1} + y_{i2}$ (compare with Arellano, 2003).[13]

3.3 Efficiency

As the way we have derived the above restrictions on θ_0 may seem arbitrary, it is of interest to know whether the conditional moment restrictions derived using the functional differencing approach lead to efficient estimation of θ_0. The next result gives the semiparametric information bound when the range of $L_{\theta,x}$ spans the whole space, that is when the model is surjective.

Proposition 1 *Assume 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.*

This result was first derived in Johnson (2004), and an intuition for the result was provided in Buchinsky *et al.* (2008). For completeness, we provide a simple sketch of the proof in Appendix A. In particular, Proposition 1 implies that, in surjective models, there exists no root-N consistent estimator of θ_0 (Chamberlain, 1987). In those models, the within projection matrix is equal to zero, and our differencing approach has no identification power.

[13] See also Buchinsky *et al.* (2008). Interestingly, this optimality property of the CMLE is not limited to the case $T = 2$. See Appendix C for the case $T = 3$. However, this result does not necessarily imply that the CMLE is semiparametrically efficient, as we argue below.

The next result complements Proposition 1 by showing that, in the special case where the columns of $L_{\theta,x}$ are linearly independent, basing the estimation of θ_0 on the functional differencing moment restrictions (19) will achieve the semiparametric information bound of the panel data model. We prove the result in Appendix A, for the case where covariates x_i have finite support.

Theorem 1 *Assume that the supports of y_i, α_i and x_i are known and finite. Suppose also that:*

i) $\theta \mapsto f_{y|x,\alpha;\theta}(y|x,\alpha)$ is continuous on Θ, for all y, α, x.

ii) $L_{\theta,x}$ has full-column rank, for all θ, x.

Then the semiparametric information bound of the panel data model coincides with the bound associated with the functional differencing restrictions (19).

When the columns of $L_{\theta,x}$ are not linearly independent, it may be that the functional differencing approach is inefficient. In Section 5 we shall see that this situation arises in a simple dynamic logit model. Nevertheless, Theorem 1 suggests that the way our approach removes the "incidental" individual effects exploits most of the information of the panel data model, at least when columns are independent.

Extension: exploiting inequality constraints. The reason why functional differencing may be inefficient is that it does not enforce the fact that the unknown probabilities of individual effects lie in the unit interval. Exploiting those additional constraints in estimation may be helpful to improve finite-sample precision, even if the conditions of Theorem 1 are satisfied.

Moreover, in models where the information bound for θ_0 is zero, exploiting those restrictions will be essential. Examples that fall into this category are set identified models with discrete dependent variables, 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, 1992b).

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

$$Q^{+}_{\theta,x}(h) = \underset{\widetilde{h} \in L_{\theta,x}(\mathcal{S}_{\mathcal{A}})}{\operatorname{argmin}} \left\| \widetilde{h} - h \right\|, \tag{24}$$

where $\|.\|$ denotes the Euclidean norm. Let us also define the *constrained* within projection as $W^+_{\theta,x} = I_{N_y} - Q^+_{\theta,x}$.

It is easy to see that, by construction:

$$W^+_{\theta_0,x}\left(f_{y|x}\right) \;=\; 0. \tag{25}$$

Using the constrained functional differencing restrictions (25) for estimation and inference has intuitive appeal. Indeed, it may be that (25) is informative about θ_0 while the within projection matrix $W_{\theta,x}$ is zero. In particular, in models where θ_0 partially identified, the constrained restrictions (25) may still characterize useful bounds on common parameters.

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 (25) as a set of conditional moment restrictions. A proper treatment of the difficulties that arise in this extension is left to future work.

4 Linear operators and within transformations

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.

4.1 Linear operators

When α_i has infinite support, $L_{\theta,x}$ becomes a linear integral operator.[14] We will make assumptions that ensure that the operator $L_{\theta,x}$ is compact, hence allowing to replicate the analysis of the finite-dimensional case in a more general setting.

Formally, let $\mathcal{X}$ be the support of x_i, and let $x \in \mathcal{X}$. Let $\theta \in \Theta$ be a given value of the common parameters. Let $\mathcal{A} \subset \mathbb{R}^q$ and $\mathcal{Y} \subset \mathbb{R}^T$ be 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 $\mathcal{G}_\alpha$ and $\mathcal{G}_y$ be two spaces of functions with domains $\mathcal{A}$ and $\mathcal{Y}$, respectively.

We define $L_{\theta,x}$ as the integral operator that maps $g \in \mathcal{G}_\alpha$ to $L_{\theta,x}g \in \mathcal{G}_y$ such that, for all $y \in \mathcal{Y}$:

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

[14] See Carrasco, Florens and Renault (2008) for an excellent overview of linear operators and their applications to econometrics.

The operator $L_{\theta,x}$ can be understood as an infinite-dimensional analog of the matrix of conditional probabilities considered in the previous section.

Next, we define $\mathcal{G}_\alpha$ and $\mathcal{G}_y$ as the spaces of square integrable functions with respect to two positive functions $\pi_\alpha > 0$ and $\pi_y > 0$, respectively:[15]

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

Then, $\mathcal{G}_\alpha$ 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 $\|g\| = \langle g, g \rangle^{\frac{1}{2}}$.

The functions π_α and π_y are selected in order to ensure compactness of the operator $L_{\theta,x}$, as stated in the following assumption (see Carrasco and Florens, 2009, for a similar setup).

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)

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

Assumption 1 restricts the distribution of individual effects. For example, if $f_{\alpha|x}$ is assumed square integrable with respect to the Lebesgue measure, we can choose $\pi_\alpha = 1$.[16] Then, π_y needs to be chosen such that *ii)* is satisfied. If y_i and α_i have bounded support, one can choose $\pi_\alpha = 1$ and $\pi_y = 1$.

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). An alternative, which does not require to assume compactness, would be to define $\mathcal{G}_\alpha$ and $\mathcal{G}_y$ as L^1 spaces of integrable functions. This is the approach pursued in Hu

[15] 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.

[16] A sufficient condition for square-integrability is that $f_{\alpha|x}$ be bounded, as in this case:

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

and Schennach (2008), who use general results from the spectral theory of linear operators. In our case, the compactness assumption will be useful for estimation, as it will allow us to compute the singular value decomposition of the operator $L_{\theta,x}$, and thus to compute explicit moment functions for θ_0.[17]

4.2 The within projection operator

We now generalize the concepts of the transpose of a matrix, the Moore-Penrose inverse, and the within projection matrix, to the infinite-dimensional case. We refer to Kress (1989) and Engl, Hanke, and Neubauer (2000) for proofs of the statements in this section and additional background on the theory of compact linear operators on Hilbert spaces.

First, let $\mathcal{N}(L_{\theta,x})$ denote the null-space of the operator $L_{\theta,x}$:

$$\mathcal{N}(L_{\theta,x}) = \{g \in \mathcal{G}_\alpha, L_{\theta,x}g = 0\},$$

and let $\mathcal{R}(L_{\theta,x})$ denote its range:

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

We say that $L_{\theta,x}$ is *injective* if and only if:

$$\mathcal{N}(L_{\theta,x}) = \{0\}. \tag{26}$$

Examples of non-injective panel data models are models with discrete dependent variables. Operator injectivity has recently received some attention in econometrics (see Hu and Schennach, 2008, and references therein). It will play an important role in the discussion of marginal effects in Section 6.

In addition, we will say that $L_{\theta,x}$ is *surjective* if:

$$\overline{\mathcal{R}(L_{\theta,x})} = \mathcal{G}_y, \tag{27}$$

where $\overline{A}$ denotes the closure of A in $\mathcal{G}_y$. As in the finite-dimensional case, non-surjectivity of $L_{\theta,x}$ will be an essential requirement of the functional differencing approach.

Next, we define the *adjoint* of the operator $L_{\theta,x}$, denoted as $L^*_{\theta,x}$, which is the unique operator that maps $\mathcal{G}_y$ onto $\mathcal{G}_\alpha$ such that, for all $(g,h) \in \mathcal{G}_\alpha \times \mathcal{G}_y$:

$$\langle L_{\theta,x}g, h\rangle = \left\langle g, L^*_{\theta,x}h\right\rangle.$$

[17] In contrast, in Hu and Schennach (2008) operator theory is used for identification only, while estimation is done using sieve maximum likelihood.

It follows from this definition that:

$$\left[L_{\theta,x}^{*}h\right](\alpha) = \int_{\mathcal{Y}} f_{y|x,\alpha;\theta}(y|x,\alpha)h(y)\frac{\pi_y(y)}{\pi_\alpha(\alpha)}dy.$$

As $L_{\theta,x}$ is compact, $L_{\theta,x}^{*}$ is also compact, and is simply the operator analog of the transpose of a matrix. An important remark is that the orthogonal of $\overline{\mathcal{R}(L_{\theta,x})}$, relative to the scalar product $\langle .,. \rangle$, is $\mathcal{N}\left(L_{\theta,x}^{*}\right)$. So, non-surjectivity of $L_{\theta,x}$ is equivalent to $\mathcal{N}\left(L_{\theta,x}^{*}\right) \neq \{0\}$.

We can now define the Moore-Penrose generalized inverse of the operator $L_{\theta,x}$, and the associated within operator. The *Moore-Penrose inverse* $L_{\theta,x}^{\dagger}$ is defined by the following limit:

$$L_{\theta,x}^{\dagger} = \lim_{\delta \overset{>}{\to} 0} \left(L_{\theta,x}^{*}L_{\theta,x} + \delta I_\alpha\right)^{-1} L_{\theta,x}^{*},$$

where I_α is the identity operator on $\mathcal{G}_\alpha$, i.e. $I_\alpha g = g$ for all $g \in \mathcal{G}_\alpha$. The domain of $L_{\theta,x}^{\dagger}$ is strictly included in $\mathcal{G}_y$ in general, unless the range of $L_{\theta,x}$ is closed.[18] This means that $L_{\theta,x}^{\dagger}h$ will generally not be defined for all $h \in \mathcal{G}_y$. Note that the range $\mathcal{R}(L_{\theta,x})$ is closed when y_i has finite support. Also, $L_{\theta,x}^{\dagger}$ satisfies the generalized inverse property:

$$L_{\theta,x}L_{\theta,x}^{\dagger}L_{\theta,x} = L_{\theta,x}.$$

The *within operator* $W_{\theta,x}$ is then defined as:

$$W_{\theta,x} = \lim_{\delta \overset{>}{\to} 0} \quad I_y - L_{\theta,x}\left(L_{\theta,x}^{*}L_{\theta,x} + \delta I_\alpha\right)^{-1} L_{\theta,x}^{*},$$

where I_y denotes the identity operator on $\mathcal{G}_y$. The operator $W_{\theta,x}$ has domain $\mathcal{D}(W_{\theta,x}) = \mathcal{G}_y$, and it is bounded, hence continuous (see Theorem 2.16 in Carrasco *et al.*, 2008). Note that for any $h \in \mathcal{D}\left(L_{\theta,x}^{\dagger}\right)$ we have: $W_{\theta,x}h = \left(I_y - L_{\theta,x}L_{\theta,x}^{\dagger}\right)h$. The within operator is simply the continuous extension of $\left(I_y - L_{\theta,x}L_{\theta,x}^{\dagger}\right)$ on $\mathcal{G}_y$, with $W_{\theta,x}h = 0$ for all $h \in \overline{\mathcal{R}(L_{\theta,x})}$. In addition, it can be shown that $W_{\theta,x}$ is the orthogonal projector on $\mathcal{N}\left(L_{\theta,x}^{*}\right)$. In particular, $W_{\theta,x}$ is self-adjoint and idempotent, i.e. $W_{\theta,x}^{*} = W_{\theta,x}$, and $W_{\theta,x}^{2} = W_{\theta,x}$.

Singular value decompositions. Given that we are working with compact operators, it is convenient to introduce their singular value decomposition (SVD, see Theorem 15.16 in Kress, 1989):

$$L_{\theta,x}g = \sum_j \phi_j \lambda_j \left\langle \psi_j, g \right\rangle, \text{ for all } g \in \mathcal{G}_\alpha, \tag{28}$$

[18] This is because $L_{\theta,x}^{\dagger}$ has domain: $\mathcal{D}\left(L_{\theta,x}^{\dagger}\right) = \mathcal{R}(L_{\theta,x}) + \mathcal{R}(L_{\theta,x})^{\perp}$.

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 (28) ranges from 1 to the (possibly infinite) rank of $L_{\theta,x}$. Note that $\{\psi_j\}$ does not form a basis of $\mathcal{G}_\alpha$, unless $L_{\theta,x}$ is injective. Likewise, $\{\phi_j\}$ does not form a basis of $\mathcal{G}_y$, except when $L_{\theta,x}$ is surjective. 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.

With this representation, it is easily verified that the Moore-Penrose inverse of $L_{\theta,x}$ is given by (see Theorem 2.8 in Engl *et al.*, 2000):

$$L_{\theta,x}^{\dagger} h \;=\; \sum_j \psi_j \frac{1}{\lambda_j} \left\langle \phi_j, h \right\rangle, \text{ for all } h \in \mathcal{D}\left(L_{\theta,x}^{\dagger}\right), \tag{29}$$

and the within operator is given by:

$$W_{\theta,x} h \;=\; h - \sum_j \phi_j \left\langle \phi_j, h \right\rangle, \text{ for all } h \in \mathcal{G}_y. \tag{30}$$

Note that the singular values and functions, and thus the within 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 computed in a suitable basis of functions. In Section 8 we will use a numerical approximation method due to Nashed and Wahba (1974) for this purpose. Working with the proposed approximation will result in an expression for the moment functions that is similar to the one we derived in the finite support case.

In the infinite-dimensional case, $L_{\theta,x}^{\dagger}$ is not bounded (hence not continuous) in general. The reason is that, when $\mathcal{R}\left(L_{\theta,x}\right)$ is not closed, 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, $L_{\theta,x}^{\dagger} h$ is not continuous in h, and it is very sensitive to any noise in h possibly arising in estimation. In contrast, $W_{\theta,x} h$ is always continuous in h.

A finite-dimensional intuition for this result is as follows. In the least-squares interpretation of Subsection 3.1, $L_{\theta,x}^{\dagger} h$ and $W_{\theta,x} h$ are 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 good prediction does not require to accurately estimate the contributions of the various regressors separately.

5 Restrictions on common parameters

In this section, we derive moment restrictions on θ_0. Then, we propose method-of-moments estimators of common parameters, and introduce a specification test for parametric random-effects models.

5.1 Functional differencing restrictions

The next theorem provides the key restrictions on θ_0.

Theorem 2 *Let Assumption 1 hold. Then the two following equivalent conditions are satisfied:*

$$W_{\theta_0,x} f_{y|x} = 0, \quad \textit{or, equivalently} \tag{31}$$

$$\mathbb{E}\left(\pi_y(y_i)\left[W_{\theta_0,x_i} h\right](y_i) \middle| x_i\right) = 0, \quad \textit{for all } h \in \mathcal{G}_y. \tag{32}$$

Theorem 2 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 (32), 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 (32) only *via* the expectation.

Efficiency. It is of interest to know whether the equivalent restrictions (31) and (32) exhaust all the information contained in the panel data model. The next theorem shows that this is *not* the case.

Theorem 3 *Let Assumption 1 hold. Then the two following statements are equivalent.*

i) There exists $f_{\alpha|x} \in \mathcal{G}_\alpha$ *such that* $f_{\alpha|x} \geq 0$, $\int_{\mathcal{A}} f_{\alpha|x}(\alpha|x)\, d\alpha = 1$, *and* $L_{\theta_0,x} f_{\alpha|x} = f_{y|x}$.

ii) $W_{\theta_0,x} f_{y|x} = 0$, $f_{y|x} \in \mathcal{D}\left(L^\dagger_{\theta_0,x}\right)$, *and there exists* $g \in \mathcal{G}_\alpha$ *such that*

$$L^\dagger_{\theta_0,x} f_{y|x} + \left(I_\alpha - L^\dagger_{\theta_0,x} L_{\theta_0,x}\right) g \geq 0. \tag{33}$$

Theorem 3 shows that the restrictions implied by the panel data model are equivalent to either of the two equivalent conditions (31) and (32), plus two extra conditions. The first requires that $f_{y|x}$ belongs to the domain of $L^\dagger_{\theta_0,x}$. This is an existence condition, which

imposes that a solution to: $f_{y|x} = L_{\theta_0,x} f_{\alpha|x}$ exists (see Definition 3.2 in Carrasco *et al.*, 2008). Using the singular value decomposition (29), this condition can be written as:

$$\sum_j \frac{1}{\lambda_j^2} \left\langle \phi_j, f_{y|x} \right\rangle^2 < \infty, \tag{34}$$

where λ_j and ϕ_j depend on (θ_0, x). Equation (34) imposes conditions on the respective rates of convergence to zero of λ_j and $\left\langle \phi_j, f_{y|x} \right\rangle$, respectively, and requires $f_{y|x}$ to belong to a certain smoothness class. The second extra condition not exploited by (31) is a set of inequality constraints, which comes from the fact that the distribution function of individual effects $f_{\alpha|x}$ needs to be non-negative.[19]

We will not exploit non-negativity constraints (33) to estimate θ_0. As the following example illustrates, our approach may thus result in inefficient estimates.

Example 4: Dynamic logit model. Consider the simple dynamic logit model

$$\begin{aligned} y_{it} &= \mathbf{1}\{\theta_0 y_{i,t-1} + \alpha_i + v_{it} \geq 0\}, \quad t = 2, ..., T, \\ y_{i1} &= \mathbf{1}\{\alpha_i + v_{i1} \geq 0\}, \end{aligned}$$

with v_{it} i.i.d. logistic, and where we take $T = 3$. This setting has been considered in Hahn (2001). It follows from some simple calculations reported in Appendix C that the functional differencing restrictions can be written as:[20]

$$\mathbb{E}\left[e^{\theta_0} y_{i1} \left(1 - y_{i2}\right) y_{i3} - \left(1 - y_{i1}\right) y_{i2} y_{i3} \right] = 0. \tag{35}$$

We note that (35) is the first-order condition of the conditional maximum likelihood estimator (CMLE), based on the density of (y_{i1}, y_{i2}) given the sufficient statistic $(y_{i1} + y_{i2}, y_{i3})$. Interestingly, Hahn (2001) shows that the CMLE does not achieve the information bound for θ_0 in this model. This provides an example of a (non-injective) panel data model where the functional differencing approach is inefficient. In this situation, efficient estimation of θ_0 should exploit the non-negativity constraints on $f_{\alpha|x}$.[21]

[19] Note that $I_\alpha - L^\dagger_{\theta_0,x} L_{\theta_0,x}$ is the orthogonal projection operator on $\mathcal{N}(L_{\theta_0,x})$. So, (33) simply means that there exists a non-negative solution $g \in \mathcal{G}_\alpha$ to the problem $f_{y|x} = L_{\theta_0,x} g$.

[20] See also Buchinsky *et al.* (2008).

[21] When supports are infinite, an operator analog of the constrained projection function $W^+_{\theta,x}$ that we introduced in Section 3 can be constructed using the techniques employed in Section 5.4 in Engl *et al.* (2000). Then, $W^+_{\theta,x}$ becomes a *nonlinear* operator that can be computed by means of projected Landweber iterations, as explained in Eicke (1992) and Sabharwal and Potter (1998).

Non-surjectivity. From Theorem 2, θ_0 satisfies a continuum of restrictions given by (31). For θ_0 to be point-identified from those restrictions, we need the equation: $W_{\theta,x} f_{y|x} = 0$ to have a unique solution θ_0 that is an interior point of Θ. Though this condition guarantees global identification of θ_0, it is little explicit. The following proposition states that non-surjectivity is necessary for θ_0 to be point-identified.

Proposition 2 *Let Assumption 1 hold, and suppose that θ_0 is globally identified from the functional differencing restrictions (31). Then, for all $\theta \neq \theta_0$ in Θ, the two following equivalent conditions hold with positive probability in x:*

$$\mathcal{N}\left(L^*_{\theta,x}\right) \neq \{0\}, \tag{36}$$

$$\overline{\mathcal{R}\left(L_{\theta,x}\right)} \neq \mathcal{G}_y. \tag{37}$$

Non-surjectivity of $L_{\theta,x}$ implies that the functional differencing restrictions given by equation (32) have (potentially) some identification content for θ_0. If on the contrary $L_{\theta,x}$ is surjective for all θ in a neighborhood of θ_0, then those restrictions are completely uninformative about θ_0, because $W_{\theta,x}$ is identically zero.

Note that, 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 Section 7 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)[22] and the censored coefficients model (Example 2) with normal errors, provided that $T > \dim \alpha_i$. Non-surjectivity is not satisfied in static probit model (though it is satisfied in the static logit model).

In appendix C, 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

[22]When $T = \dim \alpha_i$ and B is non-singular, $Q = I_T$ and $W = 0$ in equation (5), and the non-surjectivity condition is *not* satisfied.

$\mathbb{R}^T$. 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.

5.2 Method of moments

Let $\{y_i, x_i\}_{i=1,...,N}$ be a random sample. Motivated by the conditional moment restrictions (32) of Theorem 2, we propose to estimate θ_0 by minimizing a GMM criterion of the form (20), with moment functions:

$$\varphi_r(y_i, x_i, \theta) = \pi_y(y_i) [W_{\theta,x_i} h_r](y_i) \zeta_r(x_i), \tag{38}$$

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 Section 7 (which include point-identification of θ_0), $\widehat{\theta}$ will be root-N consistent and asymptotically normal. The main reason for this is the boundedness of the within projection operator.

Turning to the choice of functions h_r and ζ_r in (38), 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 8 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 matrix Υ so as to maximize efficiency, for example using suitable empirical counterparts of Chamberlain's (1987) 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, (22) provides the optimal unconditional moment restrictions. When supports are infinite, one can follow the approach of Carrasco and Florens (2000) to construct a finite-dimensional set of 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 (22), 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:

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

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 the optimal instrument functions given in the appendix) is also needed.

5.3 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:

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

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 (38). The statistic S is simply an empirical counterpart of the functional differencing moment restrictions, evaluated at the random-effects MLE.

We show in Appendix A that, under the null, and under regularity conditions given in Section 7 and standard regularity assumptions on the MLE:

$$\sqrt{N}S \xrightarrow{d} \mathcal{N}(0, V_S),$$

where the expression of V_S is provided in the appendix.

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:

$$NS'\widehat{V}_S^{-1}S \xrightarrow{d} \chi^2_R, \tag{39}$$

where $\widehat{V}_S$ is a consistent estimator of V_S. Thus, (39) 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.

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

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

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 3 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 function of individual effects $f_{\alpha|x}$ is non-negative. Note that (40) holds obviously when $L_{\theta_0,x}$ is *injective*. 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's (1992a) 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 $\operatorname{rank} B = \dim \alpha_i$ (see Appendix C). In this case, any average marginal effect is point-identified.

Example 3 (cont.) In the static logit model, the only $M(x)$ that are identified by Proposition 3 are averages of the form $M(x) = \mathbb{E}\left[h\left(y_i\right)|x_i = x\right]$, where $h \in \mathcal{G}_y$. 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 $\frac{m}{\pi_\alpha} \in \mathcal{R}\left(L^*_{\theta_0,x}\right)$.

6.2 Estimation

We now explain how to estimate an average marginal effect $M = \mathbb{E}\left[M\left(x_i\right)\right]$. We distinguish two different cases. To proceed, note that the singular value decomposition (SVD) of the Moore-Penrose generalized inverse $L^\dagger_{\theta_0,x}$ given by equation (29) implies the following SVD for the adjoint of $L^\dagger_{\theta_0,x}$, for g in its domain:

$$\left(L^\dagger_{\theta_0,x}\right)^* g \;=\; \sum_j \phi_j \frac{1}{\lambda_j} \left\langle \psi_j, g \right\rangle, \tag{42}$$

where λ_j, ϕ_j, and ψ_j depend on (x, θ_0).

Just as $L^\dagger_{\theta_0,x}$, its adjoint is not defined everywhere, and is not bounded. The following condition is key to assess the precision of marginal effects estimates.

Condition 1 *The function m/π_α belongs to the domain of $\left(L^\dagger_{\theta_0,x}\right)^*$, that is:*

$$\sum_j \frac{1}{\lambda_j^2} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle^2 < \infty. \tag{43}$$

Condition 1 requires that $\left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle$ tends fast enough to zero as j tends to infinity relative to λ_j. 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 Appendix C.

More generally, it can be shown that Condition 1 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 $m(x,\alpha) = \mathbb{E}\left[\pi_y(y_i)\, h(y_i) \,|x,\alpha\right]$. Hence, the set of marginal effects that satisfy Condition 1 is very special, as it corresponds to mean effects of the form: $M(x) = \mathbb{E}\left[\pi_y(y_i)\, h(y_i) \,|x\right]$. As intuition suggests, estimating averages of those marginal effects across x's will not be difficult in general.[23]

The importance of Condition 1 is shown by the next proposition.

Proposition 4 *Assume that the identification condition (40) is satisfied, and let Condition 1 hold. Then*

$$M(x) = \mathbb{E}\left(\pi_y(y_i) \left[\left(L^\dagger_{\theta_0,x_i}\right)^* \frac{m}{\pi_\alpha}\right](y_i) \quad \Big| x_i = x\right). \tag{44}$$

Proposition 4 suggests to estimate M by:

$$\widehat{M} = \widehat{\mathbb{E}}\left(\pi_y(y_i) \left[\left(L^\dagger_{\widehat{\theta},x_i}\right)^* \frac{m}{\pi_\alpha}\right](y_i)\right), \tag{45}$$

where $\widehat{\theta}$ is root-N consistent for θ_0, and where $\widehat{\mathbb{E}}(z_i) = \frac{1}{N}\sum_{i=1}^N z_i$ is an empirical mean. When Condition 1 holds, $\widehat{M}$ will be, under standard regularity conditions, a root-N consistent estimator of M. When the condition does *not* hold, M can be consistently estimated using a data-dependent regularization scheme, as we now explain.

Regularization. When Condition 1 does not hold, the empirical average in (45) will not be large-N consistent for M. Our solution involves replacing the operator $\left(L^\dagger_{\theta_0,x}\right)^*$ by a suitably chosen approximation, whose domain contains m/π_α. This type of approximation is referred to as a *regularization* in the literature on inverse problems.

[23]Note that, in injective models: $\overline{\mathcal{R}\left(L^*_{\theta_0,x}\right)} = \mathcal{N}\left(L_{\theta_0,x}\right)^\perp = \mathcal{G}_\alpha$. So, when $L_{\theta_0,x}$ is injective the set of marginal effects that satisfy Condition 1 is dense in $\mathcal{G}_\alpha$. More generally, this set is always dense in the set of identified marginal effects, as shown by Proposition 3.

Specifically, we consider the following regularized version of $\left(L^{\dagger}_{\theta_0,x}\right)^*$ (see Definition 3.9 in Carrasco *et al.*, 2008). Given any $\delta > 0$, it is defined by the following SVD:

$$\sum_j q_j(\delta)\phi_j \frac{1}{\lambda_j} \langle \psi_j, . \rangle , \tag{46}$$

where for some constant $a > 0$:

$$\begin{cases} |q_j(\delta)| & \leq & a\frac{\lambda_j^2}{\delta}, \\ \lim\limits_{\delta \to 0} \quad q_j(\delta) & = & 1. \end{cases} \tag{47}$$

The first condition in (47) ensures that m/π_α belongs to the domain of the regularized version of $\left(L^{\dagger}_{\theta_0,x}\right)^*$. Technically, the presence of $q_j(\delta)$ decreases the contribution of those terms for which λ_j is small, hence $1/\lambda_j$ is large, in the sum (46). The second condition implies that the regularized operator tends to the operator $\left(L^{\dagger}_{\theta_0,x}\right)^*$ as the regularization parameter δ tends to zero.

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

$$q_j(\delta) = \frac{\lambda_j^2}{\lambda_j^2 + \delta}. \tag{48}$$

Other popular regularization schemes that satisfy (47) are spectral cut-off and Landweber-Fridman (see Section 3.3 in Carrasco *et al.*, 2008).

Working with the regularization scheme $q_j(\delta)$ allows us to define the following regularized version of $M(x)$, which is well-defined for any $\delta > 0$ whether or not Condition 1 holds:

$$M_\delta(x) \quad = \quad \sum_j q_j(\delta) \langle \phi_j, f_{y|x} \rangle \frac{1}{\lambda_j} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle . \tag{49}$$

In estimation, we will let $\delta = \delta_N$ tend to zero as N tends to infinity at a rate to be specified. We propose to estimate M by an empirical analog of (49) averaged over x_i:

$$\widehat{M}_{\delta_N} \quad = \quad \widehat{\mathbb{E}} \left[\sum_j q_j \left(\delta_N\right) \pi_y(y_i) \phi_j(y_i) \frac{1}{\lambda_j} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle \right] . \tag{50}$$

Note that λ_j, ψ_j and ϕ_j all depend on x_i, although the subscript is implicit. They also depend on common parameter estimates $\widehat{\theta}$. We derive the asymptotic rate of convergence of $\widehat{M}_{\delta_N}$ in the next section.

7 Asymptotic properties

In this section, we study the asymptotic properties of the estimators of common parameters and average marginal effects that we have introduced in Sections 5 and 6.

7.1 Common parameter estimates

We start by studying the properties of $\widehat{\theta}$, which we write in a more compact form as:[24]

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

where the moment functions are given by:

$$\varphi\left(y_i, x_i, \theta\right) \quad = \quad \begin{pmatrix} \pi_y(y_i)\left[W_{\theta,x_i} h_1\right](y_i)\,\zeta_1\left(x_i\right) \\ \cdots \\ \pi_y(y_i)\left[W_{\theta,x_i} h_R\right](y_i)\,\zeta_R\left(x_i\right) \end{pmatrix}. \tag{51}$$

We make the following assumptions that ensure the consistency of $\widehat{\theta}$ as N tend 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.

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

iv) Almost surely in x*:*

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

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

v) For any j*:*

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

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)

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

viii)

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

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

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 in Section 5, this condition may fail when the non-surjectivity condition does not hold.

Conditions *iii*) and *iv*) impose that the conditional distribution of the data given α_i vary continuously with θ. This will imply that the mapping $\theta \mapsto L_{\theta,x}$ is continuous on Θ with respect to the operator norm,[25] almost surely in x.

In the consistency proof given in Appendix B we find it useful to introduce a modified version of the within operator, which is defined as

$$W_{\theta,x}^{(\mu)} = I_y - L_{\theta,x}\left(L_{\theta,x}^* L_{\theta,x} + \mu I_\alpha\right)^{-1} L_{\theta,x}^*. \tag{52}$$

As a consequence of *iii*) and *iv*), for any fixed $\mu > 0$, the mapping $\theta \mapsto W_{\theta,x}^{(\mu)}$ is continuous. Then, Conditions *v*) and *vi*) ensure that the convergence of $W_{\theta,x}^{(\mu)} f_{y|x}$ to $W_{\theta,x} f_{y|x}$ as μ tends to zero is uniform on Θ.

Condition *v*) requires that $\lambda_{j,\theta,x}$ be bounded from below. This requires $\operatorname{rank} L_{\theta,x}$, when finite, to be constant on Θ. A sufficient condition for $L_{\theta,x}$ to have constant rank is that it is injective.[26] 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.[27]

Condition *vi*) requires that $\sum_{j>J} \left\langle \phi_{j,\theta,x}, f_{y|x} \right\rangle^2$ 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 $\left\langle \phi_{j,\theta,x}, f_{y|x} \right\rangle$ 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

[25] The norm of a bounded operator L is defined as: $\|L\| = \max_{\|h\|\leq 1} \frac{\|Lh\|}{\|h\|}$.

[26] Rank constancy is also (locally) satisfied in the static and dynamic logit models considered above. In Example 3, $\dim \mathcal{N}\left(L_{\theta,x}^*\right) = 1$, so $\operatorname{rank} L_{\theta,x} = 2^T - 1 = 3$ irrespective of (θ, x). In Example 4, $\operatorname{rank} L_\theta = 2^T - 2 = 6$ when $\theta \neq 0$, and $\operatorname{rank} L_\theta = 2^T - 4 = 4$ when $\theta = 0$ (see Appendix C).

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

$f_{y|x}$ 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 Appendix C, we analytically verify uniform Fourier convergence in Chamberlain's (1992a) random coefficients model (Example 1) with known error variance. In addition, in Section 8 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:

$$\Sigma(\theta) = \mathbb{E}\left[\varphi(y_i, x_i, \theta)\,\varphi(y_i, x_i, \theta)'\right]. \tag{53}$$

This property will be useful to derive the asymptotic distribution of $\widehat{\theta}$.[28] This implies that there is no need to regularize the estimates of the moment functions, and contrasts with the need to regularize marginal effects estimates.

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

We then can state the following consistency result.

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

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

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

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

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

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

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

[28] 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.

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)

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

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

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

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

vi) As N tends to infinity:

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

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

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

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

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 $\frac{\partial L_{\theta,x}}{\partial\theta_k}$ associated with the kernel $\frac{\partial f_{y|x,\alpha;\theta}}{\partial\theta_k}$, for any $k\in\{1,...,\dim\theta\}$.

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, $\Sigma\left(\theta\right)$ 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:*

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0,\left(G'\Upsilon G\right)^{-1}G'\Upsilon\Sigma\left(\theta_0\right)\Upsilon G\left(G'\Upsilon G\right)^{-1}\right]. \tag{54}$$

Importantly, 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, working with the following slightly modified version of the within operator will typically ensure that the objective function varies smoothly with θ:

$$W_{\theta,x} \approx I_y - \sum_{j=1}^{J}\left\langle \phi_{j,\theta,x},.\right\rangle \phi_{j,\theta,x}, \tag{55}$$

where $J \geq 1$ is some integer. In order for this modification not to affect the asymptotic distribution of $\sqrt{N}\left(\widehat{\theta}-\theta_0\right)$, $J = J_N$ needs to tend to infinity fast enough as N tends to infinity.[29] When implementing this approach, we found it convenient to set J such that very small singular values (possibly leading to numerical errors due to finite precision) are discarded. In Section 8 we will provide evidence that the estimate $\widehat{\theta}$ is little sensitive to the choice of J.

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, to estimate the Hessian term G, a *regularization* is needed. The reason is that G involves the Moore-Penrose inverse $L^{\dagger}_{\theta_0,x_i}$, so G is analogous to an average marginal effect. A simple truncated estimate can be obtained as

$$\widehat{G} = \left[-\widehat{\mathbb{E}}\left(\sum_{j=1}^{J}\pi_y\left(y_i\right)\phi_{j,\widehat{\theta},x_i}\left(y_i\right)\frac{1}{\lambda_{j,\widehat{\theta},x_i}}\left\langle \frac{\partial L^{*}_{\widehat{\theta},x_i}}{\partial\theta}W_{\widehat{\theta},x_i}h_r,\psi_{j,\widehat{\theta},x_i}\right\rangle \zeta_r\left(x_i\right)\right)\right]_{r,k}. \tag{56}$$

[29] Technically, we need to choose J_N such that the bias of the moment function scaled by root-N tends to zero as N tends to infinity.

In practice, when using a finite number of singular functions in (55) to ensure that the GMM objective function is regular, $\widehat{G}$ can simply be estimated as the (numerical) derivatives of the empirical moment functions at the optimum. The next subsection details the asymptotic properties of estimates of average marginal effects, in particular providing conditions under which $\widehat{G}$ is consistent for G.

7.2 Marginal effects estimates

As explained in Subsection 6.2, we distinguish two cases. First, let us assume that Condition 1 holds, and let us suppose for simplicity that θ_0 is known. Then the asymptotic properties of marginal effects estimates are standard. To see why, define

$$m_i(\theta_0) = \pi_y(y_i)\left[\left(L^{\dagger}_{\theta_0,x_i}\right)^* \frac{m}{\pi_\alpha}\right](y_i),$$

and write the average marginal effect estimate as $\widehat{M} = \widehat{\mathbb{E}}[m_i(\theta_0)]$. The variance of $m_i(\theta_0)$ being finite is enough to apply a central limit theorem, and to prove that $\widehat{M}$ is root-N consistent and asymptotically normal.[30]

When Condition 1 does not hold we consider the regularized estimator $\widehat{M}_{\delta_N}$ as defined by (50). Let:

$$m_{i,\delta_N} = \sum_j q_j(\delta_N)\,\pi_y(y_i)\phi_j(y_i)\frac{1}{\lambda_j}\left\langle \psi_j, \frac{m}{\pi_\alpha}\right\rangle. \tag{57}$$

In (57) all singular values and singular functions are computed at (θ_0, x_i). 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 does not affect the asymptotic distribution of $\widehat{M}_{\delta_N}$. For conciseness, we drop the (θ_0, x_i) subscript in the rest of this subsection.

We have:

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

[30] Root-N consistency may also hold when θ_0 is not known, under the assumptions needed for two-step GMM estimation. In particular, it can be shown that under the identification conditions of Proposition 3:

$$\frac{\partial}{\partial\theta_k}\Big|_{\theta_0}\mathbb{E}\left[m_i(\theta)\right] = -\left\langle\left(L^{\dagger}_{\theta_0,x_i}\right)^* \frac{m}{\pi_\alpha}, \frac{\partial L_{\theta_0,x}}{\partial\theta_k}L^{\dagger}_{\theta_0,x}f_{y|x}\right\rangle$$

is well-defined. Estimating this term– which appears in the asymptotic variance of $\widehat{\mathbb{E}}\left[m_i\left(\widehat{\theta}\right)\right]$– will require the use of regularization.

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 exists $\beta > 1$ such that

$$C_\beta(x_i) \equiv \sum_j \frac{\left\langle \phi_j, f_{y|x} \right\rangle^2}{\lambda_j^{2\beta}} < \infty. \tag{58}$$

ii)

$$\sqrt{N}\delta_N \mathbb{E}\left[\left(\sup_j \left|\lambda_j^{\beta-1}\left(q_j(\delta_N)-1\right)\right|\right) C_\beta(x_i)^{\frac{1}{2}} \left\|\frac{m}{\pi_\alpha}\right\|\right] \overset{N\to\infty}{\to} 0. \tag{59}$$

iii)

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

iv)

$$\mathbb{E}\left[\left(\sup_j \left|\frac{\delta_N q_j(\delta_N)}{\lambda_j}\right|^2\right) \left\|f_{y|x}\right\|^2 \left\|\frac{m}{\pi_\alpha}\right\|^2\right] < \infty.$$

v) As N tends to infinity:

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

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, requiring that $f_{y|x}$ belongs to a regularity space (see Definition 3.4 in Carrasco *et al.*, 2008). Source conditions like (58) are routinely assumed in the ill-posed inverse problems literature. In econometrics, several variants of this assumption have already been applied.[31]

Recall that smoothness restrictions were not needed for estimating common parameters. In a given model, (58) may substantially restrict the class of data distributions. For example, in the classical nonparametric deconvolution model $y_i = \alpha_i + v_i$, Condition *i*) with $\beta = 1$ will require the distribution of v_i to be *less smooth* than that of α_i (Carrasco and Florens, 2009).

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

Part *ii*) guarantees that the bias term B_N is small when N tends to infinity. Consider for example the case where Tikhonov regularization (48) is used. Then:

$$\sup_j \quad \left|\lambda_j^{\beta-1}\left(q_j\left(\delta_N\right)-1\right)\right| = O\left(\delta_N^{\gamma}\right), \tag{60}$$

with $\gamma = \min\left(1, \frac{\beta-1}{2}\right)$. By comparison, when using spectral cut-off or Landweber-Fridman regularization, (60) holds with $\gamma = \frac{\beta-1}{2}$, for any smoothness parameter β. 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 $\mathrm{Var}\left[\delta_N \cdot m_{i,\delta_N}\right]$ is finite, in close analogy to Theorem 4. Note that, by (47), $\left|\frac{\delta_N q_j(\delta_N)}{\lambda_j}\right| \leq a\lambda_j$ is 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 $N^{\frac{\gamma}{2+2\gamma}}$.[32]

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

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

$$\sqrt{N}\delta_N\left[\widehat{M}_{\delta_N} - M\right] \xrightarrow{d} N\left[0, \Sigma_M\right]. \tag{61}$$

To conclude this section, note that the asymptotic variance of $\widehat{M}_{\delta_N}$ can simply be estimated as:

$$\widehat{\mathrm{Var}}\left(\widehat{M}_{\delta_N}\right) = \frac{\widehat{\mathrm{Var}}\left[m_{i,\delta_N}\right]}{N}, \tag{62}$$

[32]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 obtains better rates of convergence when β in (58) is large, i.e. when the distribution of the data is very smooth.

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

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

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.

8 Numerical illustration

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

8.1 Practical implementation

To implement our method in practice, we approximate the within projection operator using a method due to Nashed and Wahba (1974).[33] The approximation method works well in our context, as it uses the parametric probability model of y_i given (x_i, α_i) to generate natural bases of functions. Singular values and singular functions are then computed in those bases.[34] We present the approach in some detail in Appendix D, where we also explain how we compute estimates of average marginal effects.

In practice, the method leads to approximating moment functions for common parameters as follows. First, we sample N_y values $\underline{y}_s$ from π_y, and N_α values $\underline{\alpha}_n$ from a user-specified density $\overline{\pi}$ whose support contains $\mathcal{A}$. Then, we define the $N_y \times 1$ and $N_\alpha \times 1$ vectors (for a given $y \in \mathcal{Y}$):

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

and the $N_y \times N_\alpha$ matrix:

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

[33] The approach 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.

[34] GAUSS codes implementing the approach are available from the author.

We approximate the moment functions in (38) as:

$$\varphi_r(y_i, x_i, \theta) \approx \pi_y(y_i) \left[h_r(y_i) - \left(\underline{f}^{(y_i)}_{\theta,x_i} \right)' \underline{L}^{\dagger}_{\theta,x_i} \underline{h}_r \right] \zeta_r(x_i). \tag{63}$$

So, approximating the moment functions in this way yields an expression that is similar to the one that we derived in the finite support case.

As N_y and N_α tend to infinity, the right-hand side in (63) converges almost surely to the true moment function (see Appendix D). 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.[35]

When the dimensions of the matrix $\underline{L}_{\theta,x}$ are large, the numerical computation of the Moore-Penrose generalized inverse may be affected by errors due to finite machine precision. For this reason, we compute a modified generalized inverse that uses only J eigenvalues.[36] This amounts to approximating the modified version of the within operator given by the right-hand side of (55). The simulation evidence below suggests that taking any J in a reasonable range leads to very similar results.

8.2 Simulation evidence

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

$$y^*_{it} = \alpha_i + v_{it}, \quad t = 1, 2, \tag{64}$$

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's (1992a) random coefficients model:

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

[35] 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.

[36] 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.

where v_{i1} and v_{i2} are i.i.d 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 an 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 used in the numerical computation of the within operator, i.e. J in (55). 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 7 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.[37] 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.

[37] 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.

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.[38] Lastly, we have used $J = 12$ singular values to compute the within 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's (1992a) 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 $\widetilde{\theta}$ 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 (50), 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.

[38] 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 $\{(0,0),(0,1),(0,-1),(0,2),(0,-2),(0,3),(0,-3),...,(-3,-3)\}$.

This is consistent with ill-posedness affecting the estimation of the mean individual effect.[39]

Interestingly, there is also evidence of ill-posedness in the estimation of $\mathbb{E}(\alpha_i)$ in the random coefficients model. In this case, and given our choice of functions π_α and π_y, the identity function $m_0(\alpha) = \alpha$ does *not* satisfy Condition 1 (actually, $m_0 \notin \mathcal{G}_\alpha$). 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]

9 Conclusion

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

The approach consists in finding functions that are orthogonal to the range of the model operator. When supports are finite, this can be done using a simple "within" projection matrix, which differences out the unknown probabilities of individual effects. When supports are infinite, we have shown how to use a linear projection operator for the same purpose. This 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. We have used the moment restrictions obtained from functional differencing to construct an analog of the Hausman specification test of random versus fixed effects in a nonlinear setting. We have also studied estimation of average marginal effects and found that, in contrast with common parameters,

[39] We also applied the method outlined in Section 7 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.

[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 Condition 1. 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.

a problem of ill-posedness arises in this case.

This paper raises a number of open questions. First, in infinite dimensions, the orthogonal 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. (1991): "Moment Testing with non-ML Estimators," *mimeo*.

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

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

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

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

[8] 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*.

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

[10] 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.

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

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

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

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

[15] 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.

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

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

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

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

[20] Chamberlain, G. (1992b): "Binary Response Models for Panel Data: Identification and Information", to appear in *Econometrica*.

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

[22] 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.

[23] 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.

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

[25] Eicke, B. (1992): "Iteration Methods for Convexly Constrained Ill-Posed Problems in Hilbert Spaces," *Numerical Functional Analysis and Optimization*, 13, 413–429.

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

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

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

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

[30] Hahn, J. (2001): "The Information Bound Of A Dynamic Panel Logit Model With Fixed Effects," *Econometric Theory*, 17, 913–932.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[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] Nashed, M.Z., and G. Wahba (1974): "Convergence Rates of Approximate Least Squares Solutions of Linear Integral and Operator Equations of the First Kind", *Mathematics of Computation*, 28, 69–80.

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

[53] Newey, W., 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.

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

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

[56] Sabharwal, A., and L. C. Potter (1998): "Convexly Constrained Linear Inverse Problems: Iterative Least-Squares and Regularization," *IEEE Transactions on Signal Processing*, 46(9), 2345–2352.

[57] 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.

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

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

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

APPENDIX

A Proofs

Optimal moment restrictions (finite support). To proceed, let $r = \operatorname{rank} L_{\theta,x}$. For simplicity, we assume that r does not depend on x. Let also $U_{\theta,x}$ be $N_y \times (N_y - r)$ such that $W_{\theta,x} = U_{\theta,x}U'_{\theta,x}$ and $U'_{\theta,x}U_{\theta,x} = I_{N_y-r}$. Then the restrictions from functional differencing can be written as:

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

where $\varphi\left(y_i, x_i, \theta\right) = U_{\theta,x_i}\left[\tau\left(y_i\right), .\right]'$ is $(N_y - r) \times 1$. This is a finite set of $N_y - r$ conditional moment restrictions. The next proposition gives the optimal instruments in this case. Note that, in the special case where there are no exogenous regressors, the optimal GMM estimator associated with the optimal instruments (A2) coincides with (A5) below.

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

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

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

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:

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

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

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

The conclusion then follows from Chamberlain (1987).

■

Sketch of the proof of Proposition 1. Following the standard semiparametric efficiency bounds setup (e.g., Bickel *et al.*, 1993), consider a regular parametric submodel $f_{\alpha|x;\eta}$ for the unknown probability function of individual effects indexed by a scalar parameter η, which coincides with the true $f_{\alpha|x}$ when $\eta = 0$. The partial score with respect to η is, at $\underline{y}_s$:

$$\frac{\partial}{\partial \eta}\Big|_{\eta=0} \ln\left(\sum_{n=1}^{N_\alpha} f_{y|x,\alpha;\theta_0}\left(\underline{y}_s|x,\underline{\alpha}_n\right) f_{\alpha|x;\eta}\left(\underline{\alpha}_n|x\right)\right) = \frac{\sum_{n=1}^{N_\alpha} f_{y|x,\alpha;\theta_0}\left(\underline{y}_s|x,\underline{\alpha}_n\right) \frac{\partial}{\partial \eta}\Big|_{\eta=0} f_{\alpha|x;\eta}\left(\underline{\alpha}_n|x\right)}{\sum_{n=1}^{N_\alpha} f_{y|x,\alpha;\theta_0}\left(\underline{y}_s|x,\underline{\alpha}_n\right) f_{\alpha|x}\left(\underline{\alpha}_n|x\right)}.$$

The nonparametric tangent space is the span of such scores (which is finite-dimensional, hence closed in $\mathbb{R}^{N_y}$).

Now,

$$a_s \equiv \sum_{n=1}^{N_\alpha} f_{y|x,\alpha;\theta_0}\left(\underline{y}_s|x,\underline{\alpha}_n\right) \frac{\partial}{\partial \eta}\Big|_{\eta=0} f_{\alpha|x;\eta}\left(\underline{\alpha}_n|x\right) = L_{\theta_0,x}\left[s,.\right] \frac{\partial}{\partial \eta}\Big|_{\eta=0} f_{\alpha|x;\eta}$$

may take any value, subject to the restriction $\sum_{s=1}^{N_y} a_s = 0$. This is because $L_{\theta_0,x}$ is surjective, and $\frac{\partial}{\partial \eta}\Big|_{\eta=0} f_{\alpha|x;\eta}$ is unrestricted, apart from the fact that its elements need to sum to zero. This suggests that the nonparametric tangent set is the full set of scores in $\mathbb{R}^{N_y}$, hence that θ_0 has zero information.

Proof of Theorem 1. To simplify the notation we assume x away in the proof. Let N_y and N_α denote the number of points of supports of y_i and α_i, respectively. In this case, the information bound for θ_0 is achieved by the minimum-distance (MD) estimator based on $L_{\theta_0} g = f_y$, subject to the restrictions $\sum_n g_n = 1$, and $g_n \geq 0$ for all $n \in \{1, ..., N_\alpha\}$.

First, note that $L_{\theta_0} g = f_y$ implies that $\sum_n g_n = 1$. This is because:

$$\begin{aligned} 1 = \sum_s f_y\left(\underline{y}_s\right) &= \sum_s \sum_n f_{y|\alpha;\theta_0}\left(\underline{y}_s|\underline{\alpha}_n\right) g_n \\ &= \sum_n \left(\sum_s f_{y|\alpha;\theta_0}\left(\underline{y}_s|\underline{\alpha}_n\right)\right) g_n = \sum_n g_n. \end{aligned}$$

This shows that the following MD estimator achieves the information bound for θ_0:

$$\widehat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \left\{ \underset{g,g_n\geq 0}{\operatorname{argmin}} \left(\widehat{f}_y - L_\theta g\right)' \Upsilon_{\varepsilon_N} \left(\widehat{f}_y - L_\theta g\right) \right\}, \tag{A3}$$

where $\widehat{f}_y$ is a nonparametric estimate of f_y, and the weighting matrix is chosen as:

$$\Upsilon_{\varepsilon_N} = \left(\operatorname{Var}\left(\widehat{f}_y\right) + \varepsilon_N * I_{N_y}\right)^{-1},$$

where $\varepsilon_N > 0$ tends to zero as N tends to infinity. For technical reasons we take $\varepsilon_N = N^{-\frac{1}{4}}$.

If the information bound for θ_0 were zero, then the information bound associated with (31) would also be zero. Let us thus focus on the case where the bound is not zero. In particular, $\widehat{\theta}$ is root-N consistent for θ_0.

Next, consider, for given $\theta \in \Theta$:

$$\widehat{g}_\theta = \underset{g\in\mathbb{R}^{N_\alpha}}{\operatorname{argmin}} \left(\widehat{f}_y - L_\theta g\right)' \Upsilon_{\varepsilon_N} \left(\widehat{f}_y - L_\theta g\right),$$

where the non-negativity constraints on g are not imposed. Note that:

$$\widehat{g}_\theta = \left[\Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta\right]^\dagger \Upsilon_{\varepsilon_N}^{\frac{1}{2}} \widehat{f}_y.$$

As L_θ is injective, and as $\theta \mapsto L_\theta$ is continuous on Θ we have:

$$\begin{aligned}
\widehat{g}_{\widehat{\theta}} &= \left[L_{\widehat{\theta}}' \Upsilon_{\varepsilon_N} L_{\widehat{\theta}}\right]^{-1} L_{\widehat{\theta}}' \Upsilon_{\varepsilon_N} \widehat{f}_y \\
&= \left[L_{\theta_0}' \Upsilon_{\varepsilon_N} L_{\theta_0}\right]^{-1} L_{\theta_0}' \Upsilon_{\varepsilon_N} f_y + o_p(1) \\
&= \left[L_{\theta_0}' \Upsilon_{\varepsilon_N} L_{\theta_0}\right]^{-1} L_{\theta_0}' \Upsilon_{\varepsilon_N} L_{\theta_0} f_\alpha + o_p(1) \\
&= f_\alpha + o_p(1),
\end{aligned}$$

where in the second equality we have used that ε_N tends to zero more slowly than $N^{-\frac{1}{2}}$.

As $f_\alpha > 0$, it follows that $\widehat{g}_{\widehat{\theta}} > 0$ with probability approaching one. This shows that, with probability approaching one, $\widehat{\theta}$ coincides with:

$$\operatorname*{argmin}_{\theta\in\Theta} \left(\widehat{f}_y - L_\theta \left[\Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta\right]^\dagger \Upsilon_{\varepsilon_N}^{\frac{1}{2}} \widehat{f}_y\right)' \Upsilon_{\varepsilon_N} \left(\widehat{f}_y - L_\theta \left[\Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta\right]^\dagger \Upsilon_{\varepsilon_N}^{\frac{1}{2}} \widehat{f}_y\right). \tag{A4}$$

To see that (A4) coincides with the optimal MD estimator based on the functional differencing restrictions (31), let U_θ be $N_y \times (N_y - N_\alpha)$ such that $W_\theta = U_\theta U_\theta'$ and $U_\theta' U_\theta = I_{N_y - N_\alpha}$. Notice that $W_\theta f_y = 0$ is equivalent to $U_\theta' f_y = 0$.

As $\Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta$ is injective we can apply a standard result on partitioned matrices and obtain:

$$I_{N_y} - \Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta \left[\Upsilon_{\varepsilon_N}^{\frac{1}{2}} L_\theta\right]^\dagger = U_\theta \left(U_\theta' \Upsilon_{\varepsilon_N}^{-1} U_\theta\right)^{-1} U_\theta'.$$

Note that $U_\theta' \Upsilon_{\varepsilon_N}^{-1} U_\theta$ can be replaced by $U_\theta' \operatorname{Var}\left(\widehat{f}_y\right) U_\theta$ with no effect on the first-order asymptotic properties of the estimator. So, with probability approaching one, $\widehat{\theta}$ coincides with the following MD estimator:

$$\widetilde{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \quad \widehat{f}_y' \left[U_\theta \left(U_\theta' \operatorname{Var}\left(\widehat{f}_y\right) U_\theta\right)^{-1} U_\theta'\right] \widehat{f}_y. \tag{A5}$$

This implies that $\widehat{\theta}$ and $\widetilde{\theta}$ are asymptotically equivalent.[41] As $\widetilde{\theta}$ coincides with the optimal MD estimator based on the functional differencing restrictions (31), the conclusion follows.

Proof of Theorem 2. First note that $W_{\theta_0,x} f_{y|x} = W_{\theta_0,x} L_{\theta_0,x} f_{\alpha|x} = 0$, with probability one. Hence (31). To show that (31) and (32) are equivalent, note that:

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

[41] For this, note that, as $\mathbf{1}\left\{\widehat{\theta} = \widetilde{\theta}\right\} \xrightarrow{p} 1$ and: $\mathbf{1}\left\{\widehat{\theta} = \widetilde{\theta}\right\} \sqrt{N}\left(\widetilde{\theta} - \theta_0\right) = \mathbf{1}\left\{\widehat{\theta} = \widetilde{\theta}\right\} \sqrt{N}\left(\widehat{\theta} - \theta_0\right)$, it follows that: $\sqrt{N}\left(\widetilde{\theta} - \theta_0\right) = (1 + o_p(1)) \sqrt{N}\left(\widehat{\theta} - \theta_0\right)$.

Proof of Theorem 3. We start with a lemma.

Lemma A1 *Let $h \in \mathcal{D}\left(L^{\dagger}_{\theta_0,x}\right)$. A necessary and sufficient condition for $L_{\theta_0,x}g = h$ to have a solution is that $W_{\theta_0,x}h = 0$, in which case the general solution is: $L^{\dagger}_{\theta_0,x}h + \left(I_\alpha - L^{\dagger}_{\theta_0,x}L_{\theta_0,x}\right)\widetilde{g}$, for some $\widetilde{g} \in \mathcal{G}_\alpha$.*

Proof. Clearly, if $L_{\theta_0,x}g = h$ then $W_{\theta_0,x}h = 0$. Conversely, suppose that $W_{\theta_0,x}h = 0$. From Theorem 2.5 in Engl *et al.* (2000, p. 34), the set of solutions is:

$$L^{\dagger}_{\theta_0,x}h + \mathcal{N}\left(L_{\theta_0,x}\right).$$

Now, from Proposition 2.3 in Engl *et al.* (2000, p. 33), $\left(I_\alpha - L^{\dagger}_{\theta_0,x}L_{\theta_0,x}\right)$ is the orthogonal projector onto $\mathcal{N}\left(L_{\theta_0,x}\right)$. Hence: $\mathcal{N}\left(L_{\theta_0,x}\right) = \mathcal{R}\left(I_\alpha - L^{\dagger}_{\theta_0,x}L_{\theta_0,x}\right)$.

This ends the proof. ■

Suppose that $i)$ holds. Then $f_{\alpha|x}$ satisfies $L_{\theta_0,x}f_{\alpha|x} = f_{y|x}$. So, $f_{y|x} \in \mathcal{R}\left(L_{\theta_0,x}\right) \subset \mathcal{D}\left(L^{\dagger}_{\theta_0,x}\right)$, and $W_{\theta_0,x}f_{y|x} = 0$. So, using Lemma A1, there exists $g \in \mathcal{G}_\alpha$ such that $f_{\alpha|x} = L^{\dagger}_{\theta_0,x}f_{y|x} + \left(I_\alpha - L^{\dagger}_{\theta_0,x}L_{\theta_0,x}\right)g$, and $f_{\alpha|x} \geq 0$ a.s. Hence $ii)$.

Suppose $ii)$. Then let $f_{\alpha|x} \equiv L^{\dagger}_{\theta_0,x}f_{y|x} + \left(I_\alpha - L^{\dagger}_{\theta_0,x}L_{\theta_0,x}\right)g \geq 0$ with probability one. By Lemma A1, $W_{\theta_0,x}f_{y|x} = 0$ implies that $L_{\theta_0,x}f_{\alpha|x} = f_{y|x}$. Moreover:

$$\int_{\mathcal{Y}} \left| \int_{\mathcal{A}} f_{y|x,\alpha;\theta_0}(y|x,\alpha) f_{\alpha|x}(\alpha|x) d\alpha \right| dy = \int_{\mathcal{Y}} f_{y|x}(y|x) dy = 1 < \infty.$$

So we can apply the Fubini theorem and obtain:

$$\begin{aligned} 1 = \int_{\mathcal{Y}} f_{y|x}(y|x) dy &= \int_{\mathcal{A}} \left[\int_{\mathcal{Y}} f_{y|x,\alpha;\theta_0}(y|x,\alpha) dy \right] f_{\alpha|x}(\alpha|x) d\alpha \\ &= \int_{\mathcal{A}} f_{\alpha|x}(\alpha|x) d\alpha. \end{aligned}$$

This implies $i)$ and ends the proof.

Proof of Proposition 2. Assume that θ_0 is globally identified from (31), and suppose that $\mathcal{N}\left(L^{*}_{\theta,x}\right) = \{0\}$ with probability one for some $\theta \neq \theta_0$ in Θ.

Then, as $W_{\theta,x}$ is the orthogonal projector on $\mathcal{N}\left(L^{*}_{\theta,x}\right)$, it follows that $W_{\theta,x} = 0$. So $W_{\theta,x}f_{y|x} = 0$, contradicting the fact that θ_0 is globally identified. Hence (36). The equivalence with (37) comes from standard results on linear operators in Hilbert spaces.

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

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

where $\{\nu_j\}$ is *any* orthonormal family in $\mathcal{G}_y$, $\{\xi_{j,\theta,x}\}$ is any orthonormal basis of $\mathcal{N}\left(L^{*}_{\theta,x}\right)$, 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 (31) is equivalent to

$$\begin{aligned} \mathbb{E}\left[\pi_y(y_i)\omega(y, y_i, x_i, \theta_0) | x\right] &= \left[U_{\theta_0,x}^* f_{y|x}\right](y) \\ &= 0, \quad \text{for all } y \in \mathcal{Y}. \end{aligned} \tag{A6}$$

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:

$$h_k^{\text{opt}} = \kappa_{\theta_0,x_i}^{-1} U_{\theta_0,x_i}^* \frac{\partial L_{\theta_0,x_i}}{\partial \theta_k} L_{\theta_0,x_i}^{\dagger} f_{y|x}, \quad k = 1, ..., \dim \theta. \tag{A7}$$

In this expression, $\frac{\partial L_{\theta_0,x_i}}{\partial \theta_k}$ is an operator with kernel $\frac{\partial f_{y|x,\alpha;\theta}}{\partial \theta_k}$. Regularity conditions that ensure that the population moment functions are differentiable, and that this operator is well-defined, are given in Section 7. The operator $\kappa_x : \mathcal{G}_y \to \mathcal{G}_y$ is a non-singular *covariance operator* (Carrasco and Florens, 2000) given by:

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

Hausman specification test. 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:

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

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):

$$\sqrt{N} S \xrightarrow{d} \mathcal{N}(0, V_S),$$

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

where:

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

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 (56) with $\widetilde{\theta}$ in place of $\widehat{\theta}$.

Proof of Proposition 3. First, note that $\left(I_\alpha - L_{\theta,x}^\dagger L_{\theta,x}\right)$ is the orthogonal projection operator on $\mathcal{N}\left(L_{\theta,x}\right)$. So, $\frac{m}{\pi_\alpha} \in \mathcal{N}\left(L_{\theta_0,x}\right)^\perp = \overline{\mathcal{R}\left(L_{\theta_0,x}^*\right)}$ if and only if:

$$\left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right)\frac{m}{\pi_\alpha} = 0.$$

Suppose that θ_0 is point-identified. 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, as $L_{\theta_0,x}\left(f_\alpha - g_\alpha\right) = 0$, it follows from Lemma A1 that there exists $g \in \mathcal{G}_\alpha$ such that $g_\alpha - f_\alpha = \left(I_\alpha - L_{\theta_0,x}^\dagger L_{\theta_0,x}\right) g$.

Now, note that: $M(x) = \left\langle \frac{m}{\pi_\alpha}, f_\alpha \right\rangle$. We have:

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

provided that (40) holds, 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, noticing that $f_{y|x} = L_{\theta_0,x} L_{\theta_0,x}^\dagger f_{y|x}$, we have:

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

This ends the proof.

Proof of Proposition 4. From the fact that $\left(L_{\theta_0,x}^\dagger\right)^* \frac{m}{\pi_\alpha} \in \mathcal{G}_y$ we have:

$$\begin{aligned}
M(x) &= \left\langle \frac{m}{\pi_\alpha}, L_{\theta_0,x}^\dagger f_{y|x} \right\rangle \\
&= \left\langle \left(L_{\theta_0,x}^\dagger\right)^* \frac{m}{\pi_\alpha}, f_{y|x} \right\rangle \\
&= \mathbb{E}\left(\pi_y(y_i)\left[\left(L_{\theta_0,x}^\dagger\right)^* \frac{m}{\pi_\alpha}\right](y_i) \quad \middle| x_i = x\right).
\end{aligned}$$

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. 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.

Thus, by (52), 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:*

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

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 $J \geq 1$:

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

So, using the Cauchy-Schwartz inequality:

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

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

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

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

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

We will show:

$$\sup_{\theta \in \Theta} \quad \mathbb{E}\left(\left[\widehat{\mathbb{E}}(\varphi_r) - \mathbb{E}(\varphi_r) \right]^2 \right) \overset{N \to \infty}{\to} 0. \tag{B1}$$

For this, we will show two lemmas.

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

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

Proof.

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

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*

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

Proof. We have:

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

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 (B1) 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:

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

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

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

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 $J \geq 1$:

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

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 $W_{\theta,x}^{(\mu)} h_r$ tends to $W_{\theta,x} h_r$ as μ tends to zero, uniformly on $\mathcal{V}$.

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

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

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

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

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

As in the proof of Lemma B3 we have:

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

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:

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

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 $\zeta_r\left(x_i\right)^2$) 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* $\operatorname{Var}\left[\delta_N \cdot m_{i,\delta_N}\right] < \infty$.

Proof. We have:

$$\begin{aligned}
\operatorname{Var}\left[\delta_N \cdot m_{i,\delta_N}\right] &= \operatorname{Var}\left(\sum_j \delta_N q_j\left(\delta_N\right) \pi_y(y_i) \phi_j(y_i) \frac{1}{\lambda_j} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle\right) \\
&= \mathbb{E}\left[\operatorname{Var}\left(\sum_j \delta_N q_j\left(\delta_N\right) \pi_y(y_i) \phi_j(y_i) \frac{1}{\lambda_j} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle \Big| x_i\right)\right] \\
&\quad + \operatorname{Var}\left(\sum_j \delta_N q_j\left(\delta_N\right) \left\langle \phi_j, f_{y|x} \right\rangle \frac{1}{\lambda_j} \left\langle \psi_j, \frac{m}{\pi_\alpha} \right\rangle\right).
\end{aligned}$$

Starting with the second term in the sum:

$$\text{Var}\left(\sum_j \delta_N q_j(\delta_N)\langle\phi_j, f_{y|x}\rangle \frac{1}{\lambda_j}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\right) \leq \mathbb{E}\left(\sup_j\left|\frac{\delta_N q_j(\delta_N)}{\lambda_j}\right|^2 \|f_{y|x}\|^2 \left\|\frac{m}{\pi_\alpha}\right\|^2\right)$$

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

As for the first term in the sum, define: $K\frac{m}{\pi_\alpha} \equiv \sum_j \delta_N q_j(\delta_N)\frac{1}{\lambda_j}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\phi_j$. We have:

$$\begin{aligned}\text{Var}\left(\sum_j \delta_N q_j(\delta_N)\pi_y(y_i)\phi_j(y_i)\frac{1}{\lambda_j}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\Big|x_i\right) &= \text{Var}\left(\pi_y(y_i)\left[K\frac{m}{\pi_\alpha}\right](y_i)\Big|x_i\right)\\ &\leq \int_{\mathcal{Y}}\pi_y(y)^2\left(\left[K\frac{m}{\pi_\alpha}\right](y)\right)^2 f_{y|x}(y|x_i)\,dy\\ &\leq \sup_{y\in\mathcal{Y}}\left(f_{y|x}(y|x_i)\pi_y(y)\right)\left\|K\frac{m}{\pi_\alpha}\right\|^2. \qquad \text{(B2)}\end{aligned}$$

Noticing that $\|K\|^2 \leq \sup_j\left|\frac{\delta_N q_j(\delta_N)}{\lambda_j}\right|^2$ by Cauchy-Schwartz inequality, the expectation of (B2) 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:

$$\begin{aligned}B_N &= \mathbb{E}_{y_i,x_i}\left[\sum_j \lambda_j^{\beta-1}(q_j(\delta_N)-1)\pi_y(y_i)\phi_j(y_i)\frac{1}{\lambda_j^\beta}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\right]\\ &= \mathbb{E}_{x_i}\left[\sum_j \lambda_j^{\beta-1}(q_j(\delta_N)-1)\langle\phi_j, f_{y|x}(.|x_i)\rangle\frac{1}{\lambda_j^\beta}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\right].\end{aligned}$$

From (58) and the Cauchy-Schwartz inequality we have:

$$\left|\sum_j \langle\phi_j, f_{y|x}(.|x_i)\rangle\frac{1}{\lambda_j^\beta}\left\langle\psi_j, \frac{m}{\pi_\alpha}\right\rangle\right| \leq C_\beta(x_i)^{\frac{1}{2}}\left\|\frac{m}{\pi_\alpha}\right\|.$$

Hence:

$$|B_N| \leq \mathbb{E}_{x_i}\left[\left(\sup_j\left|\lambda_j^{\beta-1}(q_j(\delta_N)-1)\right|\right)C_\beta(x_i)^{\frac{1}{2}}\left\|\frac{m}{\pi_\alpha}\right\|\right].$$

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

C Examples

Operator injectivity in the random coefficients model (normal errors). Here we show that $\operatorname{rank} B = q$ (where $q = \dim \alpha_i$) is necessary and sufficient for $L_{\theta,x}$ to be injective in Example 1. To prove the result we take $\pi_\alpha = 1$, 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:

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

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 $z \in \mathbb{R}^q$:

$$\int_{\mathcal{A}} \exp\left[-\frac{1}{2} (z - \alpha)' B' \Sigma^{-1} B (z - \alpha) \right] g(\alpha)\, d\alpha = 0. \tag{C1}$$

As $\mathcal{G}_\alpha = L^2(\mathbb{R}^q)$, we can take L^2-Fourier transforms in (C1) 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 $r = \dim \mathcal{N}(B)$. Let $\widetilde{V}$ be a $q \times (q-r)$ matrix such that $\widetilde{V}\widetilde{V}' = B^{\dagger}B$ and $\widetilde{V}'\widetilde{V} = I_{q-r}$, and let $\widetilde{U}$ be a $q \times r$ matrix such that $\widetilde{U}\widetilde{U}' = I_q - B^{\dagger}B$ and $\widetilde{U}'\widetilde{U} = I_r$. Let $\widetilde{g}_1 \in L^2(\mathbb{R}^{q-r})$ and $\widetilde{g}_2 \in L^1(\mathbb{R}^r) \cap L^2(\mathbb{R}^r)$ such that $\widetilde{g}_1 \neq 0$, $\widetilde{g}_2 \neq 0$, and $\int_{\mathbb{R}^r} \widetilde{g}_2(\nu)\, d\nu = 0$. 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 $C = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}}$:

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

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 $\operatorname{rank} B = q$, 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 $\psi : \mathbb{R}^q \to \mathbb{R}^q$ 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 $\psi \in \mathcal{H}$:

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

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 C1 *The left singular functions of the operator $L_{\theta,x} : \mathcal{G}_\alpha \to \mathcal{G}_y$ are given by:*

$$\phi_j(y) = C(\theta) H_j\left(V'\Sigma^{-\frac{1}{2}}y\right) \exp\left[-\frac{1}{2}(y-a)'\Sigma^{-\frac{1}{2}}UU'\Sigma^{-\frac{1}{2}}(y-a)\right], \tag{C2}$$

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:

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

Let $A \propto B$ denote 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:

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

So, from the usual decomposition of quadratic forms:

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

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:

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

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, (C2) follows. Lastly, as $\|\phi_j\|=1$ the proportionality constant $C(\theta)$ satisfies:

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

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 C1 *The following condition is satisfied for any $h \in \mathcal{G}_y$, a.s. in x:*

$$\sup_{\theta\in\Theta}\left(\sum_{j>J}\left\langle \phi_j, h\right\rangle^2\right) \overset{J\to\infty}{\to} 0. \tag{C3}$$

Proof.

We start by checking Condition (C3) 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:

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

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:

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

where we have factored π_y as in the proof of Proposition C1, 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 $\mathcal{H}$:

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

In addition:

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

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 $J \geq J_1$:

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

Therefore:

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

and the corollary is proved.

■

Average marginal effects in the random coefficients model (normal errors). As before we take $\pi_\alpha=1$, and $\pi_y(y)=\exp\left[-\frac{1}{2}\eta y'\Sigma^{-1}y\right]$, where $\eta>0$. Let us assume that $L_{\theta_0,x}$ is injective. Suppose that Condition 1 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 (5), and using the same notation as in the proof of Proposition C1, 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 $\int_{\mathbb{R}^q}H(\mu)^2\exp\left[-\frac{1}{2}\eta\mu'\mu\right]d\mu<\infty$.

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 $\tau\in\mathbb{R}^q$:

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

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 Condition 1:[43]

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

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

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. To show the result we take $\pi_\alpha = 1$.

Proposition C2 *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:

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

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:

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

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:

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

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

[44] Condition 1 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 $\mu \mapsto \exp\left[-\frac{1}{4}\eta\mu'\mu\right]$.

[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.

■

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 $f_{v_1|x_1}\left(y_1 - x_1'\theta - \alpha\right) = 0$ 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 $g \in \mathcal{G}_\alpha \cap L^1\left(\mathbb{R}^q\right)$. Note that, in this case, $L_{\theta,x}g \in \mathcal{G}_y \cap L^1\left(\mathbb{R}^T\right)$. Moreover, $\mathcal{G}_\alpha \cap L^1\left(\mathbb{R}^q\right)$ is dense in $\mathcal{G}_\alpha$,[46] and $\mathcal{G}_y \cap L^1\left(\mathbb{R}^T\right)$ is dense in $\mathcal{G}_y$.

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

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

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

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

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:

$$\left[\mathcal{F}\left[L_{\theta,x}g\right]\right](\xi + W\chi|x)\,\Psi_{v|x;\theta}\left(\xi|x\right) \quad = \quad e^{\sqrt{-1}\chi' W a}\left[\mathcal{F}\left[L_{\theta,x}g\right]\right](\xi|x)\,\Psi_{v|x;\theta}\left(\xi + W\chi|x\right), \quad (\xi,\chi) \in \mathbb{R}^{2T}. \tag{C4}$$

Equation (C4) suggests that the non-surjectivity condition is satisfied unless $W = 0$, that is $\mathcal{N}\left(B'\right) = \{0\}$. In addition, evaluating (C4) at $\theta = \theta_0$ and $g = f_{\alpha|x}$ yields:

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

that is:

$$\mathbb{E}\left[e^{\sqrt{-1}(\xi + W\chi)' y_i}\Psi_{v|x;\theta_0}\left(\xi|x_i\right) - e^{\sqrt{-1}\chi' W a}e^{\sqrt{-1}\xi' y_i}\Psi_{v|x;\theta_0}\left(\xi + W\chi|x_i\right)\Big|x_i\right] \quad = \quad 0, \quad (\xi,\chi) \in \mathbb{R}^{2T}. \tag{C5}$$

Equation (C5) shows that θ_0 satisfies a continuum of conditional moment restrictions, which are informative when $\mathcal{N}\left(B'\right) \neq \{0\}$. Moreover, in this model those restrictions are analytical.

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

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

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

$$y_i = m\left(x_i, \alpha_i, \theta_0\right) + v_i, \quad i = 1, ..., N, \tag{C6}$$

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, (C6) may be used to model nonlinear production functions with heterogeneous technology parameters. We define $\mathcal{Y} = \mathbb{R}^T$, $\mathcal{A} = \mathbb{R}^q$, and we take π_α and π_y such that Assumption 1 holds.

We make the following assumption.

Assumption C1 *For $\theta \in \Theta$ and with probability one:*

$$\overline{\left\{m\left(x, \alpha, \theta\right), \alpha \in \mathbb{R}^q\right\}} \underset{\neq}{\subset} \mathbb{R}^T, \tag{C7}$$

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

Assumption C1 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 C1 will fail to hold, however, when $T = \dim \alpha_i$ and m is one-to-one.

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

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

Taking Fourier transforms we obtain:

$$\left[\mathcal{F}\left[L_{\theta,x} g\right]\right](\xi) \quad = \quad \left(\int_{\mathcal{A}} e^{\sqrt{-1}\xi' m(x,\alpha,\theta)} g\left(\alpha\right) d\alpha\right) \cdot \Psi_{v|x;\theta}\left(\xi|x\right), \quad \xi \in \mathbb{R}^T. \tag{C8}$$

We have the next result.

Proposition C3 *Let Assumption C1 hold, and assume that $\Psi_{v|x;\theta}$ does not vanish on $\mathbb{R}^T$. Then, for any $\mu \notin \overline{\left\{m\left(x, \alpha, \theta\right), \alpha \in \mathbb{R}^q\right\}}$ and any $g \in \mathcal{G}_\alpha \cap L^1\left(\mathbb{R}^q\right)$:*

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}^T} e^{-\frac{1}{2}\varepsilon\xi'\xi} e^{-\sqrt{-1}\xi'\mu} \left(\frac{\left[\mathcal{F}\left[L_{\theta,x} g\right]\right](\xi)}{\Psi_{v|x;\theta}\left(\xi|x\right)}\right) d\xi = 0. \tag{C9}$$

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

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

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 $m()$:

$$\inf_{\alpha \in \mathbb{R}^q} |\mu - m\left(x, \alpha, \theta\right)|^2 \geq \eta > 0.$$

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 (C8) 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 C3 provides a set of restrictions on $L_{\theta,x}g$, which is non-empty when Assumption C1 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 (C9) 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 C1 holds.

Static probit model. To see why finding a set $\{\varphi_y\}$ that satisfies (10) 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, (10) 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:

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

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

Static logit model. We first prove (12). We have:

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

So, as $e^{s\alpha}$, $s = 0, ..., T$, are linearly independent, (12) follows.

Next, we characterize the optimal linear combination of functional differencing restrictions when $T = 3$. In this case, the elements $\varphi \in \mathcal{N}\left(L^*_{\theta,x}\right)$ are characterized by the following equations, obtained from (12): $\varphi_{000} = \varphi_{111} = 0$, and

$$\begin{aligned} \varphi_{100}e^{x'_{i1}\theta} + \varphi_{010}e^{x'_{i2}\theta} + \varphi_{001}e^{x'_{i3}\theta} &= 0, \\ \varphi_{110}e^{(x_{i1}+x_{i2})'\theta} + \varphi_{101}e^{(x_{i1}+x_{i3})'\theta} + \varphi_{011}e^{(x_{i2}+x_{i3})'\theta} &= 0. \end{aligned}$$

This shows that $\dim \mathcal{N}\left(L^*_{\theta,x}\right) = 4$ (a.s in x), and the full set of conditional moment restrictions from functional differencing can be written as:

$$\begin{aligned} \mathbb{E}\left[y_{i1}\left(1-y_{i2}\right)\left(1-y_{i3}\right)e^{x'_{i2}\theta_0} - \left(1-y_{i1}\right)y_{i2}\left(1-y_{i3}\right)e^{x'_{i1}\theta_0}|x_i\right] &= 0, \\ \mathbb{E}\left[y_{i1}\left(1-y_{i2}\right)\left(1-y_{i3}\right)e^{x'_{i3}\theta_0} - \left(1-y_{i1}\right)\left(1-y_{i2}\right)y_{i3}e^{x'_{i1}\theta_0}|x_i\right] &= 0, \\ \mathbb{E}\left[y_{i1}y_{i2}\left(1-y_{i3}\right)e^{x'_{i3}\theta_0} - y_{i1}\left(1-y_{i2}\right)y_{i3}e^{x'_{i2}\theta_0}|x_i\right] &= 0, \\ \mathbb{E}\left[y_{i1}y_{i2}\left(1-y_{i3}\right)e^{x'_{i3}\theta_0} - \left(1-y_{i1}\right)y_{i2}y_{i3}e^{x'_{i1}\theta_0}|x_i\right] &= 0. \end{aligned}$$

Let $e_j = e^{x'_{i1}\theta_0}$, $a = \frac{1}{e_1+e_2+e_3}$, $b = \frac{1}{e_1e_2+e_2e_3+e_3e_1}$, and

$$d_{stw} = y_{i1}^s\left(1-y_{i1}\right)^{1-s}y_{i2}^t\left(1-y_{i2}\right)^{1-t}y_{i3}^w\left(1-y_{i3}\right)^{1-w}.$$

Computing Chamberlain's (1987) optimal unconditional moments for this set of conditional moment restrictions yields, after some calculation:

$$\begin{aligned} \mathbb{E}\Big[&ad_{100}\left(\left(x_2-x_1\right)e_2 + \left(x_3-x_1\right)e_3\right) + ad_{010}\left(-\left(x_2-x_1\right)\left(e_2+e_3\right) + \left(x_3-x_1\right)e_3\right) \\ &+ad_{001}\left(\left(x_2-x_1\right)e_2 - \left(x_3-x_1\right)\left(e_1+e_2\right)\right) + bd_{110}\left(\left(x_3-x_2\right)e_1e_3 + \left(x_3-x_1\right)e_2e_3\right) \\ &+bd_{101}\left(-\left(x_3-x_2\right)e_2\left(e_1+e_3\right) + \left(x_3-x_1\right)e_2e_3\right) \\ &+bd_{011}\left(\left(x_3-x_2\right)e_1e_3 - \left(x_3-x_1\right)e_1\left(e_2+e_3\right)\right)\Big] = 0. \end{aligned} \tag{C10}$$

We check that (C10) coincides exactly with the score equation from the CMLE, using $y_{i1} + y_{i2} + y_{i3}$ as a sufficient statistic.

Dynamic logit model. The eight probabilities of y_i given $\alpha_i = \alpha$ are, denoting $X = e^{\alpha}$ and $Y = e^{\alpha+\theta}$:

$$\begin{aligned} \mathrm{P}\left(y_{i1}=0, y_{i2}=0, y_{i3}=0|\alpha\right) &= \left(1+X\right)^{-3} \\ \mathrm{P}\left(y_{i1}=1, y_{i2}=0, y_{i3}=0|\alpha\right) &= X\left(1+X\right)^{-2}\left(1+Y\right)^{-1} \\ \mathrm{P}\left(y_{i1}=0, y_{i2}=1, y_{i3}=0|\alpha\right) &= X\left(1+X\right)^{-2}\left(1+Y\right)^{-1} \\ \mathrm{P}\left(y_{i1}=0, y_{i2}=0, y_{i3}=1|\alpha\right) &= X\left(1+X\right)^{-3} \\ \mathrm{P}\left(y_{i1}=1, y_{i2}=1, y_{i3}=0|\alpha\right) &= XY\left(1+X\right)^{-1}\left(1+Y\right)^{-2} \\ \mathrm{P}\left(y_{i1}=1, y_{i2}=0, y_{i3}=1|\alpha\right) &= X^2\left(1+X\right)^{-2}\left(1+Y\right)^{-1} \\ \mathrm{P}\left(y_{i1}=0, y_{i2}=1, y_{i3}=1|\alpha\right) &= XY\left(1+X\right)^{-2}\left(1+Y\right)^{-1} \\ \mathrm{P}\left(y_{i1}=1, y_{i2}=1, y_{i3}=1|\alpha\right) &= XY^2\left(1+X\right)^{-1}\left(1+Y\right)^{-2}. \end{aligned}$$

From (10) it thus follows that:

$$\begin{aligned} 0 &= (1+X)^{-3}\varphi_{000} + X(1+X)^{-2}(1+Y)^{-1}\varphi_{100} + X(1+X)^{-2}(1+Y)^{-1}\varphi_{010} \\ &\quad + X(1+X)^{-3}\varphi_{001} + XY(1+X)^{-1}(1+Y)^{-2}\varphi_{110} + X^2(1+X)^{-2}(1+Y)^{-1}\varphi_{101} \\ &\quad + XY(1+X)^{-2}(1+Y)^{-1}\varphi_{011} + XY^2(1+X)^{-1}(1+Y)^{-2}\varphi_{111}. \end{aligned}$$

So, multiplying by $(1+X)^3(1+Y)^2$ and setting the coefficients of degree 0 to 5 of the polynomial in e^α to zero, we obtain:

$$\begin{aligned} 0 &= \varphi_{000} \\ 0 &= \varphi_{100} + \varphi_{010} + \varphi_{001} \\ 0 &= (Y-X)(\varphi_{001} + \varphi_{110}) \\ 0 &= (Y-X)(Y\varphi_{001} + X\varphi_{110}) \\ 0 &= X\varphi_{110} + X\varphi_{101} + Y\varphi_{011} \\ 0 &= \varphi_{111}. \end{aligned}$$

So, if $\theta \neq 0$ then $Y \neq X$ and we obtain $\varphi_{000} = \varphi_{001} = \varphi_{110} = \varphi_{111} = 0$ and:

$$\begin{aligned} \varphi_{100} + \varphi_{010} &= 0 \\ \varphi_{101} + e^\theta \varphi_{011} &= 0. \end{aligned}$$

In this case $\dim \mathcal{N}(L_\theta^*) = 2$. Hence the functional differencing restrictions:

$$\mathbb{E}\left[y_{i1}(1-y_{i2})y_{i3}e^{\theta_0} - (1-y_{i1})y_{i2}y_{i3}\right] = 0 \tag{C11}$$

$$\mathbb{E}\left[y_{i1}(1-y_{i2})(1-y_{i3}) - (1-y_{i1})y_{i2}(1-y_{i3})\right] = 0. \tag{C12}$$

Moreover, as the moment functions in (C11) and (C12) are orthogonal, it is equivalent to base the estimation of θ_0 on (C11) only.

If $\theta = 0$ then $\varphi_{000} = \varphi_{111} = 0$ and:

$$\begin{aligned} \varphi_{100} + \varphi_{010} + \varphi_{001} &= 0 \\ \varphi_{110} + \varphi_{101} + \varphi_{011} &= 0. \end{aligned}$$

In this case $\dim \mathcal{N}(L_\theta^*) = 4$.

D Computation

In this section of the appendix we explain how we implement our approach in practice.

Discretization. Estimating common parameters and average marginal effects requires computing the singular values and singular functions of $L_{\theta,x}$: $\{\lambda_j\}$, $\{\phi_j\}$, and $\{\psi_j\}$.

Let us first consider the right singular functions $\{\psi_j\}$. From (28) we have (for all j):

$$L_{\theta,x}^* L_{\theta,x} \psi_j = \lambda_j^2 \psi_j.$$

Using the expressions for $L_{\theta,x}$ and $L_{\theta,x}^*$ we also have, for any $\alpha \in \mathcal{A}$:

$$\begin{aligned} \left[L_{\theta,x}^* L_{\theta,x} \psi_j\right](\alpha) &= \int_{\mathcal{Y}} f(y|x,\alpha)\left[L_{\theta,x}\psi_j\right](y)\frac{\pi_y(y)}{\pi_\alpha(\alpha)}dy \\ &= \int_{\mathcal{Y}}\int_{\mathcal{A}} f(y|x,\alpha) f(y|x,a)\psi_j(a)\frac{\pi_y(y)}{\pi_\alpha(\alpha)}dyda, \end{aligned}$$

where for clarity we have denoted $f \equiv f_{y|x,\alpha;\theta}$.

Following the least squares approach of Nashed and Wahba (1974), we sample N_y values from π_y and replace the integral with respect to y by an empirical mean:

$$\left[L^*_{\theta,x}L_{\theta,x}\psi_j\right](\alpha) \approx \frac{1}{N_y}\sum_{s=1}^{N_y}\int_{\mathcal{A}} f\left(\underline{y}_s|x,\alpha\right) f\left(\underline{y}_s|x,a\right)\psi_j(a)\frac{1}{\pi_\alpha(\alpha)}da. \tag{D1}$$

Equating the right-hand side of (D1) with $\lambda_j^2\psi_j(\alpha)$, we see that the solutions are of the form:

$$\widetilde{\psi}_j(\alpha) = \sum_{s=1}^{N_y} a_{js}\frac{1}{\pi_\alpha(\alpha)} f\left(\underline{y}_s|x,\alpha\right), \tag{D2}$$

for some a_{js} that depend on x and θ.

This key observation shows that one can use $\left\{f\left(\underline{y}_s|x,\alpha\right)/\pi_\alpha\right\}_s$ as an approximate generating family of $\overline{\mathcal{R}\left(L^*_{\theta,x}\right)}$, hence as a natural basis for the singular functions $\{\psi_j\}$. Replacing ψ_j by $\widetilde{\psi}_j$ in (D1) and equating with $\widetilde{\lambda}_j^2\widetilde{\psi}_j(\alpha)$, we obtain that $\{a_{js}\}$ and $\{\widetilde{\lambda}_j\}$ satisfy:

$$\frac{1}{\pi_\alpha(\alpha)}\frac{1}{N_y}\sum_{s=1}^{N_y}\left(\sum_{s'=1}^{N_y} a_{js'}\int_{\mathcal{A}}\frac{1}{\pi_\alpha(a)} f\left(\underline{y}_s|x,a\right) f\left(\underline{y}_{s'}|x,a\right)da\right) f\left(\underline{y}_s|x,\alpha\right)$$
$$= \widetilde{\lambda}_j^2\sum_{s=1}^{N_y} a_{js}\frac{1}{\pi_\alpha(\alpha)} f\left(\underline{y}_s|x,\alpha\right).$$

This will be satisfied if:

$$\frac{1}{N_y}\sum_{s'=1}^{N_y} a_{js'}\int_{\mathcal{A}}\frac{1}{\pi_\alpha(a)} f\left(\underline{y}_s|x,a\right) f\left(\underline{y}_{s'}|x,a\right)da = \widetilde{\lambda}_j^2 a_{js}, \quad s=1,...,N_y.$$

So, $\widetilde{\lambda}_j^2$ and $\{a_{js}\}_s$ can be chosen as the jth eigenvalue and (one of) the jth eigenvector(s), respectively, of:

$$P = \left[\frac{1}{N_y}\int_{\mathcal{A}}\frac{1}{\pi_\alpha(a)} f\left(\underline{y}_s|x,a\right) f\left(\underline{y}_{s'}|x,a\right)da\right]_{(s,s')}. \tag{D3}$$

Then, computing $\widetilde{\psi}_j$ according to (D2), we need to rescale $\widetilde{\psi}_j$ such that it as unitary norm, that is:

$$\widetilde{\psi}_j(\alpha)\Big/\left(\int_{\mathcal{A}}\widetilde{\psi}_j^2(a)\,\pi_\alpha(a)\,da\right)^{\frac{1}{2}}. \tag{D4}$$

When N_y tends to infinity, $\widetilde{\lambda}_j^2$ and the rescaled $\widetilde{\psi}_j$ will converge to the true λ_j^2 and ψ_j, respectively, provided that $\{\underline{y}_s\}$ becomes dense in $\mathcal{Y}$ as N_y increases (see Engl *et al.*, 2000, p. 67-68).

Once λ_j and ψ_j have been computed, we obtain the left singular vector ϕ_j from:

$$\phi_j(y) = \frac{1}{\lambda_j}\left[L_{\theta,x}\psi_j\right](y)$$
$$= \frac{1}{\lambda_j}\int_{\mathcal{A}} f(y|x,\alpha)\,\psi_j(\alpha)\,d\alpha.$$

Replacing ψ_j by $\widetilde{\psi}_j$ and λ_j by $\widetilde{\lambda}_j$, we thus obtain that $\phi_j(y) \approx \widetilde{\phi}_j(y)$, where

$$\widetilde{\phi}_j(y) = \frac{1}{\widetilde{\lambda}_j} \sum_{s=1}^{N_y} a_{js} \int_{\mathcal{A}} \frac{1}{\pi_\alpha(\alpha)} f(y|x,\alpha) f\left(\underline{y}_s|x,\alpha\right) d\alpha. \tag{D5}$$

So, $\left\{\int_{\mathcal{A}} \frac{1}{\pi_\alpha(\alpha)} f(y|x,\alpha) f\left(\underline{y}_s|x,\alpha\right) d\alpha\right\}_s$ is a natural basis for the singular functions $\{\phi_j\}$. Here also, $\widetilde{\phi}_j$ needs to be rescaled so that it has unitary norm.

In sum, the singular values and functions are computed in four steps.

Algorithm

- *In a first step, for any j and given values of x, θ we compute the jth eigenvalue $\widetilde{\lambda}_j^2$ and the (arbitrarily normalized) jth eigenvector $\{a_{js}\}_s$ of the matrix P given by (D3).*
- *In a second and third steps, we compute $\widetilde{\psi}_j$ and $\widetilde{\phi}_j$ using (D2) and (D5).*
- *Finally, in a fourth step $\widetilde{\psi}_j$ and $\widetilde{\phi}_j$ are rescaled so that $\left\|\widetilde{\psi}_j\right\| = \left\|\widetilde{\phi}_j\right\| = 1$.*

In practice, one can discretize the integrals in (D2), (D3), and (D5). For example, drawing $\underline{\alpha}_n$, $n = 1, ..., N_\alpha$, from a density $\overline{\pi}$ whose support contains $\mathcal{A}$ we can replace the matrix P in (D3) by:

$$\widetilde{P} = \left[\frac{1}{N_y}\frac{1}{N_\alpha}\sum_{n=1}^{N_\alpha} \frac{1}{\pi_\alpha(\underline{\alpha}_n)}\frac{1}{\overline{\pi}(\underline{\alpha}_n)} f\left(\underline{y}_s|x,\underline{\alpha}_n\right) f\left(\underline{y}_{s'}|x,\underline{\alpha}_n\right)\right]_{(s,s')}. \tag{D6}$$

Common parameters. To estimate common parameters, we need to compute $[W_{\theta,x_i} h_r](y_i)$, for all functions h_r and all observations $i = 1, ..., N$. We propose to proceed as follows.

For each value of covariates $x_i = x$, we compute the first J singular values and vectors $\widetilde{\lambda}_j$, $\widetilde{\psi}_j$, and $\widetilde{\phi}_j$ as explained above. In practice, J will be chosen large enough such that λ_j is very close to zero for $j > J$. For example, one could set a threshold ε that depends on machine precision, and discard all singular values that are below ε. Then we set:

$$[W_{\theta,x} h_r](y) \approx h_r(y) - \sum_{j=1}^{J} \widetilde{\phi}_j(y) \left\langle \widetilde{\phi}_j, h_r \right\rangle, \tag{D7}$$

where the integral $\left\langle \widetilde{\phi}_j, h_r \right\rangle = \int_{\mathcal{Y}} \widetilde{\phi}_j(y) h_r(y) \pi_y(y) dy$ can be discretized using $\{\underline{y}_s\}$ as points of support.

When discretizing the integrals, the expression of the approximate moment functions coincides with (63). To see this, in the notation of Section 8 we let $\underline{L} = \underline{\Phi} \cdot \underline{\Lambda} \cdot \underline{\Psi}'$ be the SVD of $\underline{L}$, where we remove the θ, x subscript for conciseness. Then, discretizing the integral in (D5) we obtain, up to a multiplicative constant:

$$\widetilde{\phi}_j(y) \propto \left(\underline{f}^{(y)}\right)' \underline{L}' a_j.$$

Now, as $\widetilde{P} \propto \underline{L} \cdot \underline{L}'$, the vector $a_j = \{a_{js}\}_s$ coincides with the jth column of $\underline{\Phi}$, which we denote as $\underline{\phi}_j$. So:

$$\begin{aligned} \widetilde{\phi}_j(y) &\propto \left(\underline{f}^{(y)}\right)' \underline{L}' \underline{\phi}_j \\ &\propto \left(\underline{f}^{(y)}\right)' \underline{\psi}_j. \end{aligned}$$

The squared norm of $\widetilde{\phi}_j$ is approximated as:

$$\frac{1}{N_y}\sum_{s=1}^{N_y}\widetilde{\phi}_j\left(\underline{y}_s\right)^2 = \frac{1}{N_y}\underline{\psi}_j'\underline{L}'\underline{L}\underline{\psi}_j = \underline{\lambda}_j^2\left(\frac{\underline{\phi}_j'\underline{\phi}_j}{N_y}\right) = \frac{\underline{\lambda}_j^2}{N_y},$$

where $\underline{\lambda}_j$ is the jth diagonal element of $\underline{\Lambda}$.

This means that, after rescaling:

$$\widetilde{\phi}_j(y) \approx \frac{\sqrt{N_y}}{\underline{\lambda}_j}\left(\underline{f}^{(y)}\right)'\underline{\psi}_j.$$

In particular:

$$\begin{aligned}\left\langle\widetilde{\phi}_j, h_r\right\rangle &= \int_{\mathcal{Y}}\widetilde{\phi}_j(y)\,h_r(y)\,\pi_y(y)\,dy \\ &\approx \frac{1}{N_y}\sum_{s=1}^{N_y}\widetilde{\phi}_j\left(\underline{y}_s\right)h_r\left(\underline{y}_s\right) \\ &\approx \frac{1}{N_y}\left(\frac{\sqrt{N_y}}{\underline{\lambda}_j}\underline{L}\underline{\psi}_j\right)'\underline{h}_r = \frac{1}{\sqrt{N_y}}\underline{\phi}_j'\underline{h}_r.\end{aligned}$$

Finally, $W_{\theta,x}h_r$ is approximated as:

$$\begin{aligned}\left[W_{\theta,x}h_r\right](y) &\approx h_r(y) - \sum_{j=1}^{J}\frac{\sqrt{N_y}}{\underline{\lambda}_j}\left(\underline{f}^{(y)}\right)'\underline{\psi}_j\left\langle\widetilde{\phi}_j, h_r\right\rangle \\ &\approx h_r(y) - \sum_{j=1}^{J}\frac{\sqrt{N_y}}{\underline{\lambda}_j}\left(\underline{f}^{(y)}\right)'\underline{\psi}_j\frac{1}{\sqrt{N_y}}\underline{\phi}_j'\underline{h}_r \\ &= h_r(y) - \left(\underline{f}^{(y)}\right)'\left(\sum_{j=1}^{J}\underline{\psi}_j\frac{1}{\underline{\lambda}_j}\underline{\phi}_j'\right)\underline{h}_r.\end{aligned}$$

This coincides with (63), using the J-modified Moore-Penrose inverse of $\underline{L}$.

Average marginal effects. Turning to marginal effects estimates, we approximate $\widehat{M}_{\delta_N}$ in (50) by:

$$\widehat{M}_{\delta_N} \approx \widehat{\mathbb{E}}\left[\sum_{j=1}^{J}q_j(\delta_N)\,\pi_y(y_i)\widetilde{\phi}_j(y_i)\frac{1}{\widetilde{\lambda}_j}\left\langle\widetilde{\psi}_j, \frac{m}{\pi_\alpha}\right\rangle\right], \tag{D8}$$

and discretize the integral $\left\langle\widetilde{\psi}_j, \frac{m}{\pi_\alpha}\right\rangle = \int_{\mathcal{A}}\widetilde{\psi}_j(\alpha)\,m(\alpha)\,d\alpha$ on the support $\{\underline{\alpha}_n\}$.

Denoting as $\underline{m}$ the $N_\alpha \times 1$ vector with elements $\left[\frac{1}{\sqrt{\overline{\pi}(\underline{\alpha}_n)\pi_\alpha(\underline{\alpha}_n)}}m(\underline{\alpha}_n)\right]_n$ we obtain, after some calculation:

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

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's (1992a) 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

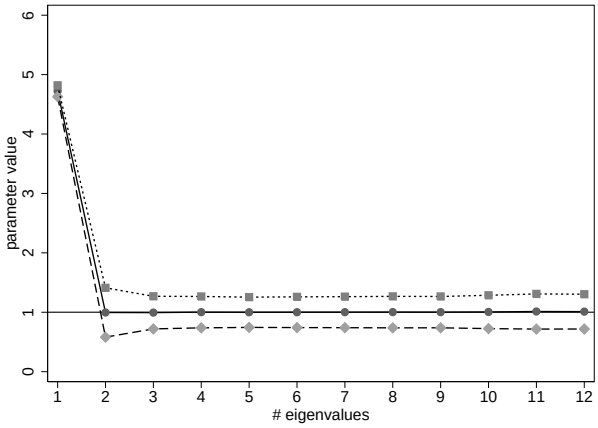
	$N = 100$		$N = 500$		$N = 100$		$N = 500$	
	Unweighted mean (true=1)				Weighted mean (true=.5)			
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$J = 2$	1.117	.227	1.116	.102	.485	.094	.478	.040
$J = 5$	.938	.374	.939	.170	.514	.140	.514	.060
$J = 8$	1.071	1.212	1.037	.534	.510	.134	.510	.058

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.

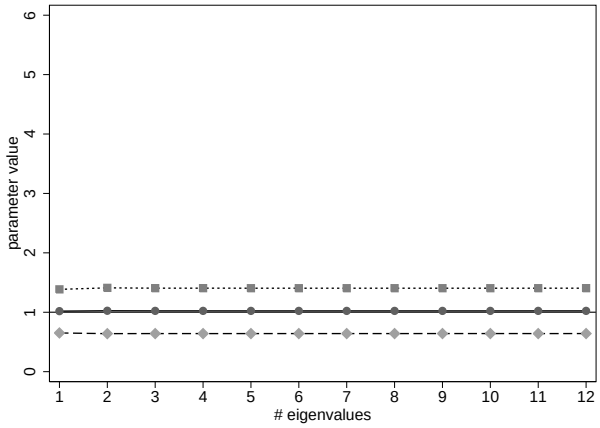
Figure 1: Parameter estimates ($N = 100, T = 2$)

Common parameters

Tobit model: σ (true=1)

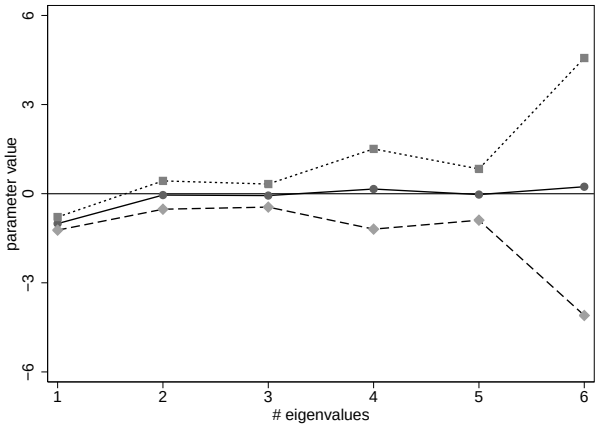


Chamberlain's model: θ (true=1)

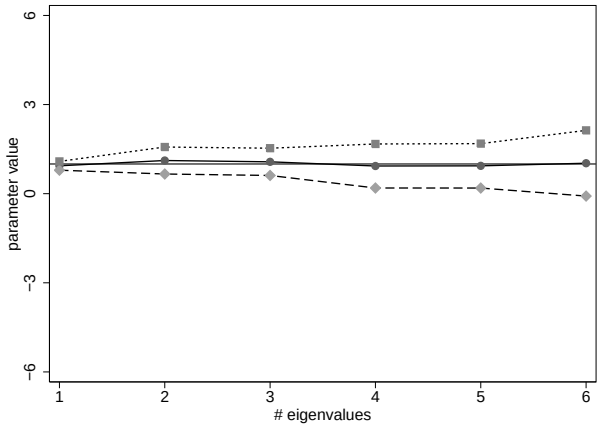


Average marginal effects

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

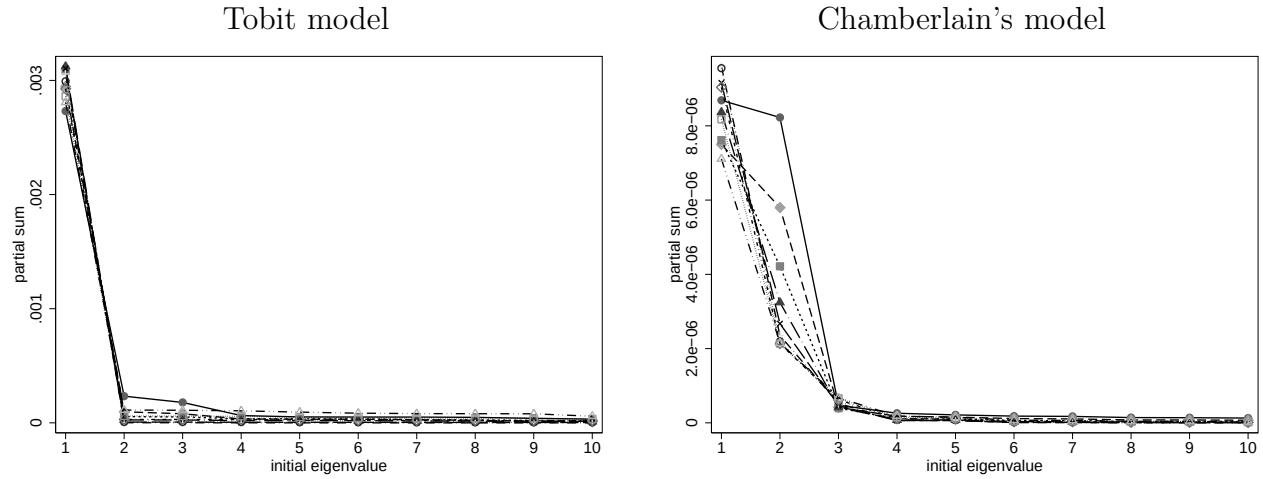


Chamberlain's model: $\mathbb{E}(\alpha_i)$ (true=1)



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 asymptotic 95%-confidence intervals, 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* $\{.5, .6, ..., 1.5\}$.