A Correlated Random Coefficient Panel Model with Time-Varying Endogeneity

Louise Laage*

Job Market Paper

This version : January 1, 2019

Most recent version : click here

Abstract

This paper studies a class of linear panel models with random coefficients. We do not restrict the joint distribution of the time-invariant unobserved heterogeneity and the covariates. We investigate identification of the average partial effect (APE) when fixed-effect techniques cannot be used to control for the correlation between the regressors and the time-varying disturbances. Relying on control variables, we develop a constructive two-step identification argument. The first step identifies nonparametrically the conditional expectation of the disturbances given the regressors and the control variables, and the second step uses "between-group" variations, correcting for endogeneity, to identify the APE. We propose a natural semiparametric estimator of the APE, show its $\sqrt{n}$ asymptotic normality and compute its asymptotic variance. The estimator is computationally easy to implement, and Monte Carlo simulations show favorable finite sample properties. Control variables arise in various economic and econometric models, and we provide variations of our argument to obtain identification in some applications. As an empirical illustration, we estimate the average elasticity of intertemporal substitution in a labor supply model with random coefficients.

*Department of Economics, Yale University, 30 Hillhouse Avenue, New Haven, CT 06511. louise.laage@yale.edu.
I am grateful to my advisors Donald W. K. Andrews, Xiaohong Chen and especially Yuichi Kitamura for their guidance and support. I also thank Anna Bykhovskaya, Philip A. Haile, Yu Jung Hwang, John Eric Humphries, Rosa Matzkin, Patrick Moran, Peter C. B. Phillips, Pedro Sant'Anna, Masayuki Sawada, Edward Vytlacil as well as participants at the Yale econometrics seminar for helpful conversations and comments on this project. All errors are mine.

1 Introduction

This paper considers a random coefficient panel model whose outcome equation is

$$y_{it} = x_{it}' \mu_i + \alpha_i + \epsilon_{it}, \qquad i \leqslant n,\ t \leqslant T, \tag{1}$$

where the impact of the vector x_{it} of covariates on a scalar dependent variable y_{it} is linear in μ_i, vector-valued time-invariant unobserved heterogeneity. The scalar ϵ_{it} is a time-varying disturbance. We assume that the number of periods T is fixed and the number of units n is large. We adopt a *fixed effect approach*, that is, we do not impose assumptions on the joint distribution of $(\mu_i, (x_{it})_{t \leqslant T})$. This positions (1) in the class of *correlated random coefficient* (CRC) models.

In this class of models, one parameter of interest is the average partial effect (APE). The APE is defined in Wooldridge (2005b) as $\mathbb{E}_q[\partial \mathbb{E}(y_t|x_t, q_t)/\partial x|_{x_t=\bar{x}}]$, where the outer expectation is over the vector of unobservables $q_t = (\mu, \alpha, \epsilon_t)$. It is an average of the partial effect of x_t on y_t over the distribution of the unobserved heterogeneity q_t. In the correlated random coefficient model (1), the average partial effect is equal to $\mathbb{E}(\mu)$. A discussion of the APE in CRC models can be found in Graham and Powell (2012). As they highlight, identifying the average partial effect is not a straightforward task because regressors and unobserved heterogeneity are potentially correlated. For instance, a linear least squares regression computed with a single cross-section will not consistently estimate $\mathbb{E}(\mu)$.

Various versions of (1) are studied in Mundlak (1978), Chamberlain (1992), and more recently in Arellano and Bonhomme (2012) and Graham and Powell (2012). Chamberlain (1992) identifies the average partial effect when some of the coefficients are known to be deterministic, hence constant across individuals, and the number of periods is strictly greater than the number of regressors. His identification method is divided in two steps. The first step exploits within-variation to identify the common parameters, and in a second step, a between-group regression uses these values to obtain the APE. This identification argument requires a strict exogeneity condition on the regressors. In model (1), the strict exogeneity condition can be written $\mathbb{E}(\epsilon_{it}|x_{i1}, ..., x_{iT}) = 0$[1]and we argue that this condition is restrictive. First, note that under the strict exogeneity condition, the covariates $(x_{it})_{t \leqslant T}$ can be correlated with the vector of unobservables $(\alpha_i, \mu_i', (\epsilon_{it})_{t \leqslant T})$ through their correlation with (α_i, μ_i'). This correlation can be controlled for by a fixed effect transformation because (α_i, μ_i') is time-invariant. Loosely speaking, the endogeneity of the model can be "captured by a fixed effect". However, in many cases such an assumption does not hold. One such case is the presence of time-varying omitted variables correlated with the regressors x_{it} and appearing linearly in (1).

[1]In a model more general than (1), where the partial derivative $\partial \mathbb{E}(y|x, q)/\partial x$ depends on the time-varying disturbance, Graham and Powell (2012) allows for a correlation between this disturbance and the regressor x_{it}, but a marginal stationarity condition must be satisfied.

Consider for example an assessment of the impact of maternal time input on an outcome variable measuring child cognitive development, such as test score. This effect may be heterogeneous due to variation in unobserved skills of the mother. Moreover, the time allocation of the mother may be correlated with time-varying unobservables, e.g., stress from the workplace, that also affect child development. In this example, the regressor is correlated with the time-varying disturbance and the strict exogeneity condition does not hold. This paper seeks to relax the strict exogeneity condition and to allow for what we call "time-varying endogeneity".

When instruments satisfying an orthogonality condition are available, one might be tempted to estimate the average effect using a fixed-effect instrumental variables estimator or a first-difference instrumental variables estimator (see, e.g, Wooldridge, 2010). But due to the randomness of the unobserved effect and its potential correlation with the regressors and the instrument, this estimator will be asymptotically biased. Counterfactual computations based on this estimator will be biased as well. The core idea of this paper is to use the control function approach (CFA) to control for endogeneity. More specifically, we assume the existence of control variables such that, conditional on the control variables at all time periods, the time-varying disturbance at time t is conditionally mean independent of the regressors at all time periods. Under this assumption, we show identification of the average partial effect using a two-step approach. We provide a consistent estimator.

Section 2 lays out the two steps of the identification argument. The first step exploits the time variation of the regressors for each individual and "differences away" the individual unobserved heterogeneity in order to nonparametrically identify the conditional expectation of the disturbances given regressors and control variables. The second step corrects for endogeneity using the nonparametric function identified in the first step and uses "between-group" variation to identify the average effect. Because this procedure suffers from the curse of dimensionality, we adapt the identification argument to the case where some of the endogenous regressors are known to have constant coefficients. We also adapt it to the case where the number of periods is larger than the minimum number required for identification.

In Section 3, we illustrate the usefulness of this identification argument. It is valid in a model with time-varying omitted variables, and we discuss a model of a heterogeneous production function as an application. Additionally, our approach to identification relies on the existence of control variables but does not use their exact specification. This information is in fact not part of the model under study. We argue that thanks to this flexibility, the two-step argument can be applied to variations of model (1) and is not limited to endogeneity due to time-varying omitted variables. A similar argument allows us to identify the average effect in models where the current value of the regressor is affected by the outcome in the previous period. Thus, we relax the strict exogeneity condition maintained in the literature. We also adapt our two-step approach to panel models with

random coefficients where sample selection is a source of endogeneity.

The identification argument is constructive, and its structure suggests a natural multi-step estimator. We define the estimator in Section 4 and study its asymptotic properties in Sections 5 through 7. The steps are as follows. We construct first estimates of the control variables, then we estimate nonparametrically the conditional expectation of the time-varying disturbance given the regressors using nonparametric sieve regressions. Finally, we estimate the average effect with a sample analog of the "between-group" regression. The derivation of the asymptotic properties of our estimator is challenging due to the presence of nonparametric regression estimators using nonparametrically estimated regressors. This relates to a broad literature on estimation with generated covariates in which to our knowledge, there are no results on the asymptotic distribution of sample moments depending on nonparametric two-step sieve estimators. We therefore prove that the estimator is consistent and asymptotically normally distributed. This estimator is computationally easy to implement as it uses closed-form expressions and does not require optimization. Monte Carlo simulations show favorable finite sample properties in Section 8.

Section 9 turns to an empirical illustration. Using the Panel Study of Income Dynamics, we estimate the average elasticity of intertemporal substitution in a labor supply model with random coefficients.

We review here the literature this paper is connected to. Linear panel models with random coefficients, such as (1), are sometimes refered to as models with individual-specific slope or variable-coefficient models. They are surveyed for example in Wooldridge (2010) and Hsiao (2014). Wooldridge (2005a) shows that consistency of the standard fixed-effects estimator in these models require the random coefficients to be mean independent of the detrended regressors. Important recent results on correlated random coefficient panel models are in Arellano and Bonhomme (2012) and Graham and Powell (2012). Both papers build upon Chamberlain (1992), which studies efficiency bounds in semiparametric models. As an example, Chamberlain (1992) derives the semiparametric variance bound of the APE in a correlated random coefficient model and provides an efficient estimator. Arellano and Bonhomme (2012) investigate a model very similar to (1) under strict exogeneity. They obtain identification of the variance and of the distribution of the unobserved effect by leveraging information on the time dependence of the time-varying disturbances. They require that the number of periods be strictly greater than the number of regressors (including a constant if any), an assumption that we maintain. On the other hand, Graham and Powell (2012) focus on identification of the APE when T is exactly equal to the number of regressors. In this case, identification is irregular and the method developped in Chamberlain (1992) cannot be applied. They develop an alternative identification argument and construct an estimator. Another example of a fixed effect approach in panel data is Evdokimov (2010) which studies identification

and estimation of a model where the outcome variable depends nonparametrically on regressors and scalar time-invariant unobserved heterogeneity. An alternative to the fixed-effect approach is the correlated random effect approach which imposes restrictions on the conditional distribution of the unobserved heterogeneity given the regressor. Examples of this approach in panel data are, among others, Altonji and Matzkin (2005), Bester and Hansen (2009), Arellano and Bonhomme (2016) and Graham, Hahn, Poirier, and Powell (2018).

As stated earlier, papers studying the APE correlated random coefficients panel models do not allow for time-varying endogeneity. To the best of our knowledge, we are the first to prove identification of the APE in CRC models with time-varying endogeneity. However the use of exclusion restrictions or of the control function approach (see, e.g, Newey, Powell, and Vella, 1999, Blundell and Powell, 2003) in models with random coefficients is not new. Cross section models include Wooldridge (1997), Wooldridge (2003) and Heckman and Vytlacil (1998). They impose an exclusion restriction on the random coefficient and homogeneity conditions on the impact of the instruments on the regressors, and identify the average treatment effect. More recently, Masten and Torgovitsky (2016) specify a nonseparable dependence between the regressors and control variables. An analogous approach in a panel model is employed in Murtazashvili and Wooldridge (2016) which studies a random coefficient model with endogenous regressors and endogenous switching. Additionally, Murtazashvili and Wooldridge (2008) show that the fixed-effect instrumental variables estimator is consistent only under a similar set of assumptions. Exploiting the panel aspect of the data to "difference away" the time-invariant unobserved heterogeneity allows us to avoid imposing such restrictions on the joint distribution of the unobserved heterogeneity, the regressors, and the instruments.

2 Model and Identification

The following section sets up the model and the control function approach. Section 2.2 lays out the identification argument, imposing an invertibility condition which is then studied in more details in Section 2.3. Finally in Section 2.4, we show how to improve upon the identification method in some cases where more is known about the data generating process.

2.1 Model

We restate the model we consider in this paper. For a sample of units indexed by i, for $i \leqslant n$, the outcome variable in period t, for $t \leqslant T$, is given by

$$y_{it} = x_{it}' \mu_i + \alpha_i + \epsilon_{it},$$

where $x_{it} \in \mathbb{R}^{d_x}$ is a vector of observed variables, ϵ_{it} is a time-varying disturbance, and μ_i is a time-invariant vector which represents individual unobserved heterogeneity. We consider the case where T is fixed and n large, and we assume $T \geqslant d_x + 2$. Denoting by $y_i = (y_{i1}, .., y_{iT})'$ the vector of outcomes of unit i, $X_i = (x_{i1}, .., x_{iT})'$ the matrix of regressors, and $\epsilon_i = (\epsilon_{i1}, .., \epsilon_{iT})'$ the vector of error terms, we can rewrite (1) as $y_i = X_i\mu_i + \alpha_i 1_T + \epsilon_i$ where 1_T is the vector of size T composed of ones.

The parameters of interest are the average effects $\mathbb{E}(\mu_i)$ and $\mathbb{E}(\alpha_i)$. Note that (1) is a particular case of the model $y_{it} = \bar{x}_{it}'\beta_i + \epsilon_{it}$ with $\bar{x}_{it} = (x_{it}', 1)'$ and $\beta_i = (\mu_i', \alpha_i)'$. Arellano and Bonhomme (2012) study this general model when some of the regressors are known to have homogeneous impact. One of their objects of interest is the average effect, and their identifying assumption is the strict exogeneity condition $\mathbb{E}(\epsilon_{it}|\bar{X}_i, \beta_i) = 0$. If there is an intercept in the vector of regressors, such a condition does not restrict the joint distribution of (X_i, α_i) hence does not preclude endogeneity. That is, the conditional mean of the additive unobserved error term $\alpha_i + \epsilon_{it}$ given the regressors is not required to be uniformly zero. However, this correlation cannot be due to the presence of time-varying omitted variables, as argued in Section 1. We seek to relax this strict exogeneity condition. We will assume the availability of instrumental variables $z_{it} \in \mathbb{R}^{d_z}$ satisfying the following assumption, where we write Z_i for the individual matrix of instruments and $x_{it,d}$ for each scalar-valued regressor so that $x_{it} = (x_{it,d})_{d \leqslant d_x}$.

Assumption 2.1.

1. *$(X_i, Z_i, \epsilon_i, \mu_i, \alpha_i)$ is i.i.d. accross i, and (1) holds with $\mathbb{E}(\epsilon_{it}) = 0$. For all $d \leqslant d_x$, $x_{it,d}$ is a continuous random variable,*
2. *For each $t \leqslant T$, there exists an identified function C_t such that, defining $v_{it} = C_t(x_{it}, z_{it}) \in \mathbb{R}^{d_v}$,*
$$\mathbb{E}(\epsilon_{it} \,|\, x_{i1}, ..x_{iT}, v_{i1}, ..v_{iT}) = f_t(v_{i1}, ..v_{iT}).$$

Assumption 2.1 (2) is a control function approach (CFA) assumption and v_{it} is a control variable. Besides its panel aspect, this assumption is similar to the condition imposed in Newey et al. (1999)[2]. Define $V_i = (v_{i1}', ..v_{iT}')'$. If for all $t \leqslant T$, a cross section control function assumption is satisfied, that is, $\mathbb{E}(\epsilon_{it}|x_{it}, v_{it}) = h_t(v_{it})$, and if $(x_{it}, \epsilon_{it}, v_{it})$ is i.i.d. over both i and t, then Assumption 2.1 (2) is satisfied with $f_t(V_i) = h_t(v_{it})$. Note that we normalized $\mathbb{E}(\epsilon_{it}) = 0$ so $\mathbb{E}(f_t(V_i)) = 0$ for all $t \leqslant T$. This is without loss of generality since a constant is not separately identifiable from $\mathbb{E}(\alpha_i)$.

Control variables satisfying Assumption 2.1 (2) are typically provided by a first-step selection equation. We will provide a few examples of such selection equations in Section 3. We mention here

[2] A slight difference is that they impose $\mathbb{E}(\epsilon_i|z_i, v_i) = f(v_i)$ in a cross section regression, without endogenous variables in the conditioning set. Their definition of the control variable as $v_i = x_i - \mathbb{E}(x_i|z_i)$ implies $\mathbb{E}(\epsilon_i \,|\, x_i, v_i) = f(v_i)$, which is a cross section version of the condition we impose.

a particular case of primary interest that will be our baseline model for the estimation. Consider

$$\begin{aligned} y_{it} &= x_{it}^{1\prime}\mu_i^1 + x_{it}^{2\prime}\mu_i^2 + \epsilon_{it}, \\ x_{it}^2 &= \mathbb{E}(x_{it}^2|x_{it}^1, z_{it}) + v_{it}, \quad \mathbb{E}(\epsilon_i|V_i, X_i) = f(V_i), \end{aligned} \tag{2}$$

where $x_{it}^1 \in \mathbb{R}^{d_1}$ is the vector of exogenous regressor and $x_{it}^2 \in \mathbb{R}^{d_2}$ can be endogenous. We can rewrite (2) as (1) taking d_x to be $d_1 + d_2$, $x_{it} = (x_{it}^{1\prime}, x_{it}^{2\prime})'$, $\mu_i = (\mu_i^{1\prime}, \mu_i^{2\prime})'$ and $v_{it} \in \mathbb{R}^{d_2}$ as defined. Then Assumption 2.1 (2) is satisfied with $C_t(x, z) = x^2 - \mathbb{E}(x_{it}^2|x_{it}^1 = x^1, z_{it} = z)$. The control variables in this model is the residual of the regression of x_{it}^2 on (x_{it}^1, z_{it}). Note that another choice of control variable studied in Blundell and Powell (2003) and Imbens and Newey (2009) is $v_{it} = F_{x^2|x^1,z}(x_{it}^2|x_{it}^2, z_{it})$, and in that case $v_{it} \in \mathbb{R}$.

Note 2.1. Unlike this particular case, Model (1) and our general definition of the control variables as $v_{it} = C(x_{it}, z_{it})$ in Assumption 2.1 (2) do not make explicit which of the regressors are endogenous. Nor does the condition $\mathbb{E}(\epsilon_{it} \,|\, X_i, V_i) = f_t(V_i)$. Our general identification results will not distinguish endogenous from exogenous variables as this information is embedded in the specific definition of the control variables, which is determined outside of the model. This deliberate lack of precision is offset by a gain in flexibility and we will present in Section 3 a variety of specifications to which our identification argument applies.

Time differencing: Define $u_{it} = \epsilon_{it} - f_t(V_i)$ and $u_i = (u_{i1}, .., u_{iT})'$. By Assumption 2.1, $\mathbb{E}(u_i|X_i, V_i) = 0$. We also write $f(V_i) = (f_1(V_i), .., f_T(V_i))'$, where V_i is of dimension $(T\, d_v)$. The vector of primitives for each unit i is $W_i = (X_i, Z_i, V_i, \mu_i, u_i)$.

The model (1) can be rewritten

$$y_{it} = x_{it}'\,\mu_i + \alpha_i + f_t(V_i) + u_{it}.$$

The extra term $f_t(V_i)$ captures the "time-varying endogeneity": it is an unobserved time-varying random variable correlated with the regressors. The random variable $\alpha_i + f_t(V_i)$ is composed of two unobserved elements. First, α_i, which is time invariant but varies accross individuals. Second, $f_t(V_i)$, which varies with both i and $t \leqslant T$. Thus for any ν_i random variable such that $\mathbb{E}(\nu_i) = 0$, the model with α_i and $(f(V_i))_{t\leqslant T}$ is observationally equivalent to a model where α_i is replaced with $\alpha_i + \nu_i - \mathbb{E}(\nu_i|V_i)$ and $(f(V_i))_{t\leqslant T}$ is replaced with $(f_t(V_i) + \mathbb{E}(\nu_i|V_i))_{t\leqslant T}$. Thus, $f(V_i)$ and $\mathbb{E}(\alpha_i|V_i)$ are not separately identifiable. However,we normalized $\mathbb{E}(f(V_i)) = 0$, which should intuitively guarantee that $f(V_i)$ and $\mathbb{E}(\alpha_i)$ are separately identifiable. Our identification procedure will indeed prove this fact. First, we exploit the time invariance of α_i and take time differences to eliminate this term. We will then later obtain identification of $\mathbb{E}(\alpha_i)$ using $\mathbb{E}(f(V_i)) = 0$.

From now on, we therefore look at a first-differencing transformation of the model, that is, for $t \leqslant T-1$,

$$\begin{aligned} y_{i\,t+1} - y_{it} &= [x_{i\,t+1} - x_{it}]'\mu_i + f_{t+1}(V_i) - f_t(V_i) + u_{i\,t+1} - u_{it}, \\ \dot{y}_{it} &= \dot{x}_{it}'\mu_i + g_t(V_i) + \dot{u}_{it}, \end{aligned} \tag{3}$$

with $\dot{y}_{it} = y_{i\,t+1} - y_{it}$, $\dot{x}_{it} = x_{i\,t+1} - x_{it}$, $g_t(V_i) = f_{t+1}(V_i) - f_t(V_i)$, and $\dot{u}_{it} = u_{i\,t+1} - u_{it}$.

We write the model in vector form, defining $\dot{X}_i = (\dot{x}_{i1}, .., \dot{x}_{i\,T-1})'$ a $(T-1) \times d_x$ matrices, and the $(T-1) \times 1$ vectors $\dot{y}_i = (\dot{y}_{i1}, .., \dot{y}_{i\,T-1})'$, $g(V_i) = (g_1(V_i), .., g_{T-1}(V_i))'$ and $\dot{u}_i = (\dot{u}_{i1}, .., \dot{u}_{i\,T-1})'$. By assumption, $\mathbb{E}(\dot{u}_i|X_i, V_i) = 0$. Equation (3) can be rewritten as

$$\dot{y}_i = \dot{X}_i\mu_i + g(V_i) + \dot{u}_i. \tag{4}$$

Note 2.2. Assumption 2.1 (2) does not require the regressors to be strictly exogenous. Indeed, the conditional expectation $\mathbb{E}(\epsilon_{it}|X_i, V_i)$ is allowed to depend on each of the v_{is}, $s \leqslant T$. In Section 3, we take adantage of this property of Assumption 2.1 (2) and obtain identification in a class of models similar to (1) without contemporaneous endogeneity but where the strict exogeneity condition does not hold and only sequential exogeneity is imposed.

2.2 Identification

2.2.1 Two-step identification

For identification, we need to define two matrices. First, the $(T-1) \times (T-1)$ matrix M_i defined as $M_i = I_{T-1} - \dot{X}_i(\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i'$ if $\dot{X}_i$ is of full rank or $M_i = I - \dot{X}_i\dot{X}_i^+$ if not, where $\dot{X}_i^+$ is the Moore Penrose inverse (implying $\dot{X}_i\dot{X}_i^+\dot{X}_i = \dot{X}_i$). And second, defined only if $\dot{X}_i$ has full column rank, the $d_x \times (T-1)$ matrix $Q_i = (\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i'$. These matrices are such that $M_i\dot{X}_i\mu_i = 0$, and $Q_i\dot{X}_i\mu_i = \mu_i$. Our two-step approach will exploit these properties and first use within-group variations with M_i to identify the function g. Then g will be plugged in in a between-group regression using Q_i to identify $\mathbb{E}(\mu_i)$.

Before explaining the details of the identification method, we mention here some of the limitations of using the matrix Q_i. Whether or not $\dot{X}_i'\dot{X}_i$ is invertible, M_i is an orthogonal projection matrix. This implies that the norm of M_i is bounded with probability 1. However, the norm of the matrix $q(X) = (\dot{X}'\dot{X})^{-1}\dot{X}$ is not uniformly bounded as a function of X as can be seen from the fact that $||q(X)||$ goes to infinity as $\det(\dot{X}'\dot{X}) \to 0$. In particular, this implies that the norm of $q(X)$ is not bounded when X lies in a compact subset of the space of matrices of size $T \times k_x$. However the second step of the identification argument multiplies the vector of outcome variables by $Q_i = q(X_i)$. Taking any expectation to identify the average effect will require that $\mathbb{E}(||Q_i\dot{u}_i||) < \infty$. Given the

properties of Q_i that we highlighted, this is a strong condition. This issue is discussed in details in Graham and Powell (2012). A second issue they raise is that in their model the semiparametric variance bound which depends on Q_i might not be bounded as well, implying that $\mathbb{E}(\mu_i)$ is not regularly identified. We acknowledge these issues and proceed in the following way. In this section, we focus on the identification properties of our model and we assume that the needed moments are finite. However, since regular identification requires assumptions which might not hold in the data, we point out that $\mathbb{E}(\mu_i|\det(\dot{X}_i'\dot{X}_i) > \delta)$ is also identified under standard assumptions and we suggest an estimator for this parameter, the asymptotic properties of which we will study in more details using these standard assumptions. Note that this is also the object estimated by Arellano and Bonhomme (2012).

Assumption 2.2. *$\mathbb{E}(||\dot{u}_i||) < \infty$, $\mathbb{E}(||Q_i\dot{u}_i||) < \infty$, $\mathbb{E}(||Q_i g(V_i)||) < \infty$, and $\mathbb{E}(||(\mu_i', \alpha_i)||) < \infty$.*

Note that Assumption 2.2 implicitly imposes a rank condition on the regressors, since for $\mathbb{E}(||Q_i\dot{u}_i||)$ to exist, Q_i must be well defined with probability one. That is, $\dot{X}_i$ must be of full column rank with probability one. With expectations now properly defined, since M_i and Q_i are function of $\dot{X}_i$, we use Assumption 2.1 (2) to obtain

$$M_i\dot{y}_i = M_i g(V_i) + M_i\dot{u}_i, \qquad \mathbb{E}(M_i\dot{u}_i|X_i, V_i) = 0, \tag{5}$$

$$Q_i\dot{y}_i = \mu_i + Q_i g(V_i) + Q_i\dot{u}_i, \qquad \mathbb{E}(Q_i\dot{u}_i|X_i, V_i) = 0. \tag{6}$$

The role of these matrices is the following. Equation (5) is a within-group transformation that allows to separate g from μ_i and to identify it, while equation (6) isolates μ_i from $\dot{X}_i$ and uses the knowledge of g to identify $\mathbb{E}(\mu)$ by taking population average.

2.2.2 Identification of $g_t(.)$

We start by identifying g_t. This function itself is not an object of interest in the model, but the procedure developed here to identify the average partial effect requires its identification as a first step. Note that (5) gives

$$\mathbb{E}(M_i\dot{y}_i|X_i, V_i) = M_i g(V_i), \tag{7}$$

where $M_i = I_{T-1} - \dot{X}_i(\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i'$ is an observed $(T-1)\times(T-1)$ matrix function of the matrix X_i, which is in the conditioning set of the conditional expectation. For V a given value of V_i, $g(V)$ is a $(T-1)\times 1$ vector and our goal is to recover the function g using the conditional expectation. However, M_i is singular because it is a projection matrix projecting onto the space orthogonal to the columns of $\dot{X}_i$. It is therefore not possible to identify $g(V_i)$ directly using (7).

Instead of using $\mathbb{E}(M\,\dot{y}|X, V)$, we focus on $k(V) = \mathbb{E}(M\,\dot{y}|V)$ which satisfies

$$\mathbb{E}(M\dot{y}|V) = \mathbb{E}(\,M|V)\,g(V) = \mathcal{M}(V)g(V), \tag{8}$$

where we write $\mathcal{M}(V) = \mathbb{E}(M_i | V_i = V) = \mathbb{E}(M(\dot{X}_i) | V_i = V) = \mathbb{E}(I - \dot{X}_i(\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i' | V_i = V)$. If $\mathcal{M}(V)$ is invertible for a given value V of V_i, then (8) gives a closed form expression for $g(V)$. This suggests the following invertibility condition to obtain identification of the whole function,

Assumption 2.3. *The matrix $\mathcal{M}(V_i)$ is invertible, $\mathbb{P}_V$ a.s.*

Note that this assumption is a condition solely on observables, and can therefore be tested using available data. Under Assumptions 2.1, 2.2 and 2.3, we obtain

$$g(V_i) = \mathcal{M}(V_i)^{-1}\mathbb{E}(M_i\dot{y}_i|V_i), \ \mathbb{P}_V \text{ a.s.} \tag{9}$$

Intuitively, Assumption 2.3 precludes $g(V_i)$ from being of the form $\dot{X}_i\beta_i$ and distorting a proper identification of $\mathbb{E}(\mu_i)$. Indeed, $g(V_i) = \dot{X}_i\beta_i \Rightarrow M_i g(V_i) = 0 \Rightarrow \mathcal{M}(V_i)g(V_i) = 0 \Rightarrow g(V_i) = 0$ by invertibility of $\mathcal{M}(V_i)$, which cannot hold if $\beta_i \neq 0$. This means that the term $\dot{X}_i\mu_i$ is separately identifiable from $g(V_i)$ by Assumption 2.3.

Note 2.3. If we did not use time differences, defining $a(V_i) = \mathbb{E}(\alpha_i|V_i)$, $\tilde{M}_i = I_T - X_i(X_i'X_i)^{-1}X_i'$ and $\tilde{\mathcal{M}}(V) = \mathbb{E}(\tilde{M}_i|V_i = V)$, then (8) would become $\mathbb{E}(\tilde{M}_i y_i|V_i) = \tilde{\mathcal{M}}(V_i)[f(V_i)+a(V_i)]+\mathbb{E}(\tilde{M}_i[\alpha_i - a(V_i)])$, where the second term on the RHS is guaranteed to be 0. Hence one cannot use the above explained method to identify $f + a$ and must exploit the time invariance of α_i.

2.2.3 Identification of average effects

Average partial effect $\mathbb{E}(\mu)$

Under Assumption 2.1, 2.2 and 2.3, $g_t(.)$ is identified for $t \leqslant T-1$ and the matrix Q_i is well-defined with probability 1. Equation (6) implies

$$\mu_i = Q_i\dot{y}_i - Q_i g(V_i) - Q_i\dot{u}_i, \tag{10}$$

where by the law of iterated expectations and Assumption 2.2, $\mathbb{E}(Q_i\dot{u}_i) = 0$. This implies

$$\mathbb{E}(\mu) = \mathbb{E}(Q_i\dot{y}_i - Q_i g(V_i)), \tag{11}$$

which identifies $\mathbb{E}(\mu)$ because all elements on the right hand side are observed.

Result 2.1. *Under Assumptions 2.1, 2.2 and 2.3, the average effect $\mathbb{E}(\mu)$ is identified.*

As mentioned in Section 2.2.1, one might be worried that the conditions $\mathbb{E}(||Q_i\dot{u}_i||) < \infty$ and $\mathbb{E}(||Q_i g(V_i)||) < \infty$ of Assumption 2.2 do not hold. In this case, we propose an alternative object of interest which is identified under standard conditions. We define $\delta_i = \mathbb{1}(\det(\dot{X}_i'\dot{X}_i) > \delta_0)$ and $Q_i^\delta = \delta_i Q_i$. Then

$$\mathbb{E}(\mu|\delta) = \mathbb{P}(\det(\dot{X}_i'\dot{X}_i) > \delta_0)^{-1}\mathbb{E}(\delta_i Q_i\dot{y}_i - \delta_i Q_i g(V_i)) = \mathbb{P}(\det(\dot{X}_i'\dot{X}_i) > \delta_0)^{-1}\mathbb{E}(Q_i^\delta\dot{y}_i - Q_i^\delta g(V_i)), \tag{12}$$

which identifies $\mathbb{E}(\mu|\delta)$ because all the terms on the rightmost side of (12) are identified. The required conditions are $\mathbb{E}(||Q_i^\delta \dot{u}_i||) < \infty$ and $\mathbb{E}(||Q_i^\delta g(V_i)||) < \infty$, which one can show are satisfied in particular if X_i has bounded support, $\mathbb{E}(||\dot{u}_i||) < \infty$ and $\mathbb{E}(||g(V_i)||) < \infty$.

It remains to identify $\mathbb{E}(\alpha_i)$, which we obtain using the variables in period 1. We multiply (10) with x_{i1} and substract y_{i1},

$$y_{i1} - x_{i1}'\,\mu_i \;=\; y_{i1} \;-\; x_{i1}'\,[Q_i\dot{y}_i \;-\; Q_i g(V_i) \;-\; Q_i\dot{u}_i]\,.$$

The model gives $y_{i1} - x_{i1}'\,\mu_i = \alpha_i + \epsilon_{it}$ where $\mathbb{E}(\epsilon_{it}) \;=\; 0$. Combining the two, we obtain

$$\mathbb{E}(\alpha_i) = \mathbb{E}(y_{i1} \;-\; x_{i1}'\,[Q_i\dot{y}_i \;-\; Q_i g(V_i)]),$$

where the right hand side is observable. This identified $\mathbb{E}(\alpha_i)$. That is, even if only the difference $g_t = f_{t+1} - f_t$ is identified and not f_t for each t, the normalization $\mathbb{E}(f_t(V_i)) = 0$ identifies $\mathbb{E}(\alpha_i)$.

Average effect of an exogenous intervention

Consider a policy intervention that changes x_{it} for each unit i in a given period t. The average effect of this exterior intervention is an object of interest to analyze such policies and Blundell and Powell (2003) studies its identifiability in different models when the change in covariates is exogenous, i.e, independent of the unobservable error terms. The unobservables in the CRC model we sudy in this paper are $(\mu_i, \alpha_i, (\epsilon_{it})_{t\leqslant T})$, and an exogenous shift can be a variation Δ_i independent of $(\alpha_i, \mu_i, (\epsilon_{it})_{t\leqslant T})$ in which case the average impact of the policy is $\mathbb{E}(\mu_i)\mathbb{E}(\Delta_i)$. However it might be of interest to consider policy interventions where the variation Δ is correlated with x_{it}, hence correlated with (μ_i, α_i) while exogenous in the sense that it is independent of $(\epsilon_{it})_{t\leqslant T}$. For example, consider an exogenous intervention that shifts x_{it} to $l(x_{it})$. The average outcome after this intervention is $\mathbb{E}(l(x_{it})'\mu_i + \alpha_i + \epsilon_{it})$. It depends on the joint distribution of (μ_i, x_{it}) where μ_i is unobservable and could potentially be challenging to obtain since we left this joint distribution unrestricted. However the second step of the proof of identification derived Equation (10), which expresses μ_i as a function of the primitives. Once more it can be plugged in to obtain average effects. The change in expected outcome is

$$\begin{aligned}
\mathbb{E}(l(x_{it})'\mu_i + \alpha_i + \epsilon_{it} - [x_{it}'\mu_i + \alpha_i + \epsilon_{it}]) \;&=\; \mathbb{E}([l(x_{it}) - x_{it}]'\mu_i), \\
&=\; \mathbb{E}([l(x_{it}) - x_{it}]'[Q_i y_i - \; Q_i g(V_i) - Q_i\dot{u}_i]), \\
&=\; \mathbb{E}([l(x_{it}) - x_{it}]'[Q_i y_i - \; Q_i g(V_i)]),
\end{aligned}$$

where the second equality holds by exogeneity of the change in regressors. All elements in the last expectation are identified, thus identifying the average change in outcome.

2.3 Invertibility of $\mathcal{M}(V)$

We now provide conditions satisfying Assumption 2.3 under which the matrix $\mathcal{M}(V)$ is nonsingular almost surely in V. We first state a set of high level conditions and prove that they satisfy Assumption 2.3. We will then explore on a case-by-case basis situations in which these high-level conditions are satisfied. We also provide some extensions. We use the notation Int() to refer to the interior of a set, $\mathcal{S}_{W|\bar{V}}$ refers to the support[3] of the random variable W conditional on the variable V taking the value $\bar{V}$. $Rank(A)$ refers to the column rank of a matrix A, and $p_{W|V}(.|.)$ is the conditional density of a random variable W conditional on the variable V.

We will write, for two random variables A and B, $\mathcal{S}_A$ the support of A and $\mathcal{S}_{A|b}$ the support of the A conditional on $B = b$.

2.3.1 High-level Condition

Assumption 2.4. *The following holds almost surely in V.*

1. $\text{Int}\left(\mathcal{S}_{\dot{X}|V}\right) \neq \varnothing$,
2. *There exists a basis $e = (e_1, .., e_{T-1})$ of $\mathbb{R}^{T-1}$ such that for each $t \leqslant T-1$, there exists $\dot{X}^{(t)} \in Int\left(\mathcal{S}_{\dot{X}|V}\right)$ such that $p_{\dot{X}|V}(\dot{X}^{(t)}|V) > 0$, $Rank(\dot{X}^{(t)}) = d_x$ and $\dot{X}^{(t)\prime} e_t = 0$.*

Result 2.2. *Under Assumption (2.4), $\mathcal{M}(V)$ is nonsingular almost surely in V.*

Proof. Recall that $\mathcal{M}(V) = \mathbb{E}(\, M \,|V) = \mathbb{E}(\, M(X) \,|V)$ where we define $M(X) = I_{T-1} - \dot{X}(\dot{X}'\dot{X})^{-1}\dot{X}'$ is an orthogonal projection matrix. $M(X)$ projects onto the space orthogonal to the k_x columns of $\dot{X}$, where each column k corresponds to the $T-1$ values of the scalar r.v $\dot{x}_{i.,k}$. By the properties of orthogonal projection matrices, $M(X) = M(X)' = M(X)^2 = M(X)'M(X)$. This implies that $\mathcal{M}(V) = \mathbb{E}(\, M(X)'M(X) \,|V)$. Thus, for a given $V \in \mathcal{S}_V$,

$$\begin{aligned}
\mathcal{M}(V) \notin GL_{T-1}(\mathbb{R}) &\Leftrightarrow \exists c \in \mathbb{R}^{T-1},\ c \neq 0,\ , \mathcal{M}(V)\, c = 0 \\
&\Rightarrow \exists c \in \mathbb{R}^{T-1},\ c \neq 0,\ , c'\, \mathcal{M}(V)\, c = 0 \\
&\Rightarrow \exists c \in \mathbb{R}^{T-1},\ c \neq 0,\ , \mathbb{E}(\, c'\, M(X)'M(X) c \,|\, V) = 0 \\
&\Rightarrow \exists c \in \mathbb{R}^{T-1},\ c \neq 0,\ , \mathbb{E}(\, ||M(X)\, c||^2 \,|\, V) = 0,
\end{aligned}$$

and since $||M(X)\, c||^2$ is a positive function, this implies that $||M(X)\, c|| = 0$ with probability 1 on the support of $\dot{X}$ conditional on V, i.e $\mathbb{P}_{\dot{X}|V}$−a.s. That is,

$$M(X)\, c = 0,\ \mathbb{P}_{X|V}\text{-a.s}\,.$$

[3]The support of a continuous r.v Z with density p_Z is defined as the closure of the set where p_Z takes nonzero values.

This result is very useful here, it implies that if a sum of orthogonal projections of a given vector x is zero, then each of the orthogonal projections of x is zero. This allows us to exploit variations of X conditional on V, focusing on the kernels of the matrices $M(X)$ to show that the intersections of all of them as X varies is the trivial set $\{0\}$.

Consider a draw of $V \in \mathcal{S}_V$ satisfying Assumption 2.4 (1) and (2). $\text{Int}\left(\mathcal{S}_{X|V}\right) \neq \varnothing$ and there exists a basis $e = (e_1, .., e_{T-1})$ and $\dot{X}^{(t)} \in \text{Int}\left(\mathcal{S}_{\dot{X}|V}\right)$ such that for each $t \leqslant T-1$, $p_{\dot{X}|V}(\dot{X}^{(t)}|V) > 0$, $Rank(\dot{X}^{(t)}) = d_x$ and $\dot{X}^{(t)\prime} e_t = 0$.

$\mathcal{S}_{X|V}$ is a subset of $\mathbb{R}^{(T-1)\times d_x}$, the continuity arguments will therefore be in $\mathbb{R}^{(T-1)\times d_x}$. Fix $t \leqslant T-1$. $\dot{X}^{(t)}$ is of full column rank k_x implies that $\det(\dot{X}^{(t)\prime}\dot{X}^{(t)}) \neq 0$. The determinant function being continuous, as well as the density $p_{\dot{X}|V}(.|V)$, this implies that there exists an open ball $\mathcal{B}_t \subset \text{Int}\left(\mathcal{S}_{\dot{X}|V}\right)$ such that **(1)** $\dot{X}^{(t)} \in \mathcal{B}_t$, **(2)** $\forall \dot{X} \in \mathcal{B}_t$, $p_{\dot{X}|V}(\dot{X}|V) > 0$, and **(3)** $\forall \dot{X} \in \mathcal{B}_t, Rank(\dot{X}) = k_x$.

Take $c \in \mathbb{R}^{T-1}$ such that $\mathcal{M}(V)c = 0$. Then we know that $M(X)\,c = 0$, $\mathbb{P}_{\dot{X}|V}$-a.s. Since the density $p_{\dot{X}|V}(.|V)$ is strictly positive on $\mathcal{B}_t$, $M(X)\,c = 0$, for all $\dot{X}$ in $\mathcal{B}_t$ except on a set of measure 0. Additionally, since for all $\dot{X}$ in $\mathcal{B}_t$, $\dot{X}$ is of full rank, then $M(\dot{X})\,c$ is a continuous function of $\dot{X}$. Those two facts imply that $M(X)\,c$ is uniformly 0 on $\mathcal{B}_t$, and in particular, $M(X^{(t)})\,c = 0$. Moreover $\dot{X}^{(t)\prime} e_t = 0$ implies $M(X^{(t)})e_t = M(X^{(t)})'e_t = e_t$. Thus,

$$\begin{aligned}
\forall t \leqslant T-1,\ M(X^{(t)})\,c = 0 &\Rightarrow \forall t \leqslant T-1,\ e_t'\,M(X^{(t)})\,c = 0,\\
&\Rightarrow \forall t \leqslant T-1,\ e_t'\,c = c_t = 0,\\
&\Rightarrow c = 0.
\end{aligned}$$

Hence $\mathcal{M}(V)$ is invertible and this holds almost surely in V. □

Note 2.4. In this note, we discuss the conditions imposed in Assumption 2.4.

1. If e is the canonical basis of $\mathbb{R}^{T-1}$, that is, $e_1 = (1, 0, ..., 0)'$, ... and $e_{T-1} = (0, ..., 0, 1)'$, then $\dot{X}'d_t = 0$ is equivalent to $x_{it+1} = x_{it+1} = \bar{x}$. In this case, the terms $x_{it}'\mu_i$ and $x_{it+1}'\mu_i$ are equal, $y_{it+1} - y_{it} = \epsilon_{it+1} - \epsilon_{it}$ and

$$\mathbb{E}(y_{it+1} - y_{it}|x_{it} = x_{it+1} = \bar{x}, V_i = V) = f_{t+1}(V) - f_t(V) = g_t(V).$$

   The condition $p_{\dot{X}|V}(\dot{X}^{(t)}|V) > 0$ is similar to a condition in Evdokimov (2010), which studies a nonparametric panel model with scalar time-invariant unobserved heterogeneity and additively separable time-varying disturbance. The difference, due to the linear structure of the model we study here, is that we require this condition to hold only for a finite number of points ($T-1$, exactly) and we do not use these values of $\bar{x}$ to obtain identification, we simply need their existence.

2. We impose the full rank condition on the draws $\dot{X}^{(t)}$ to be able to use the continuity of the function $M(.)$. Indeed $M(.)$ is not continuous at the points $X \in \mathcal{S}_X$ such that X is not of full rank because the Moore Penrose inverse function is not a continuous function. This condition might look strong and unintuitive. It might seem unnecessary as well, since the invertibility of $\mathcal{M}(V)$ yields identification of g and not of μ; an explicit full rank condition on the regressors should therefore not be needed. However, note that linear dependence between the regressors is assumed away in Assumption 2.2 to obtain identification of $\mathbb{E}(\mu)$ in the second step.

3. The existence of a draw $\dot{X}^{(t)}$ such that $\dot{X}^{(t)}$ is of full column rank d_x and $\dot{X}^{(t)} e_t = 0$ implicitly requires the number of columns of $\dot{X}^{(t)}$ to be lower than $T-2$, i.e $d_x \leqslant T-2$ as we imposed. Moreover, note that if we do not consider a first-differencing transformation of the model and use a vector form of (1), then the column $(1, .., 1)'$ is always included in the matrix M_i. This would imply that the matrix $\mathcal{M}(V_i)$ is singular.

4. The statement here is about invertibility of $\mathcal{M}(V)$, almost everywhere on the support of V. In Section 4, we will construct an estimator, and to study its asymptotic properties we will assume that $\mathcal{M}(V)$ is invertible for all $V \in \mathcal{S}_V$ to ensure that $\lambda_{\min}(\mathcal{M}(V))$ is bounded away from zero uniformly in V.

2.3.2 Examples and Variations of the Conditions

Examples

We focus here on the model (2) when there are no exogenous regressors, i.e, $d_1 = 0$. We write $x_{it} = b_t(z_{it}) + v_{it}$. We consider first the scalar case, where x_{it} and z_{it} are real. Write $\dot{X}_i = \dot{B}(Z_i) + \dot{V}_i$, so $\dot{X}_i$ is a column vector of size $T-1$.

Assumption 2.5.

1. *Almost surely in V,* $\text{Int}\left(\mathcal{S}_{Z|V}\right) \neq \varnothing$,
2. *For all $t \leqslant T$, b_t is a continuously differentiable function,*
3. *Almost surely in V, there exists $\bar{Z} \in \text{Int}\left(\mathcal{S}_{Z|V}\right)$ such that $p_{Z|V}(\bar{Z}|V) > 0$, $\bar{\dot{X}} = \dot{B}(\bar{Z}) + \dot{V} \neq 0$, and for some $t \leqslant T-1$, $\bar{\dot{X}}_t \neq 0$ and $db_t(\bar{z}_{it})/dz_t \neq 0$.*

Result 2.3. *Under Assumption 2.5, $\mathcal{M}(V)$ is nonsingular $\mathbb{P}_V$ a.s.*

Proof. We define for a draw of V and the corresponding $\bar{Z}$ defined in Condition (3) of Assumption 2.5, an open ball $\mathcal{B}$ around $\bar{\dot{X}}$ such that for all $\dot{X} \in \mathcal{B}$, $\dot{X} \neq 0$ and $p_{\dot{X}|V}(\dot{X}|V) > 0$. The function $M(.)$ which maps X to the orthogonal projection matrix projecting onto the space orthogonal to the space generated by the columns of $\dot{X}$, is continuous on $\mathcal{B}$ by the same argument as in the proof of Result 2.2. As in this proof, for c such that $\mathcal{M}(V)c = 0$, we have $||M_i c|| = 0$ for all $\dot{X} \in \mathcal{B}$.

For this same draw of V and the corresponding $\bar{Z}$, there exists $t \leqslant T-1$ such that $\mathrm{d}b_t(\bar{z}_{it})/dz_t \neq 0$. For δ small enough, $\underline{Z}_\delta = (\bar{z}_{i1}, .., \bar{z}_{it-1}, \bar{z}_{it} + \delta, \bar{z}_{it+1}, .., \bar{z}_{iT})' \in \mathcal{B}$. Define $\underline{\dot{X}}_\delta = \dot{B}(\underline{Z}_\delta) + V$. All components of $\underline{\dot{X}}_\delta$ are the same as those of $\bar{\dot{X}}$ but one. Therefore there exists δ such that $\underline{\dot{X}}_\delta$ and $\bar{\dot{X}}$ are not collinear and $\underline{\dot{X}}_\delta \neq 0$. However $M(\underline{\dot{X}})c = 0$ implies that c and $\underline{\dot{X}}$ are collinear, and $M(\bar{\dot{X}})c = 0$ also implies that c and $\bar{\dot{X}}$ are collinear, which would imply, if $c \neq 0$, that $\underline{\dot{X}}$ and $\bar{\dot{X}}$ are collinear. Therefore c must be 0. This implies that $\mathcal{M}(V)$ is nonsingular. □

We now give conditions in the case where x_{it} is vector valued, $x_{it} \in \mathbb{R}^{d_x}$ and $z \in \mathbb{R}^{d_z}$, but where b_t is assumed to be a linear function, i.e,

$$x_{it} = A_t z_{it} + v_{it}, \qquad \mathbb{E}(v_{it}|z_{it}) = 0, \tag{13}$$

with A_t of size $d_x \times d_z$. Taking the basis e to be the canonical basis of $\mathbb{R}^{T-1}$, Condition (2) of Assumption 2.4 implies that almost surely in V there exists $X^{(t)} \in \mathcal{S}_{X|V}$ such that $e_t'\dot{X}^{(t)} = 0$ for all $t \leqslant T$, which is equivalent to $\dot{x}_{it}^{(t)} = 0$. In the case of the linear control function approach in (13), it imposes for almost all draws of V the existence of $Z^{(t)} \in \mathcal{S}_{Z|V}$ such that $A_{t+1}z_{it+1}^{(t)} - A_t z_{it}^{(t)} + \dot{v}_{it} = 0$, for all $t \leqslant T-1$, . This condition will be satisfied if A_t (or A_{t+1}) has full row rank, implying $d_z \geqslant d_x$, and if $\mathrm{Int}\left(\mathcal{S}_{z_t|(V,z_{t+1})}\right)$ (or $\mathrm{Int}\left(\mathcal{S}_{z_{t+1}|(V,z_t)}\right)$) is rich enough. A stronger version of these arguments is the following assumption.

Assumption 2.6.

1. *$\dot{X}$ has full rank a.s,*
2. *The matrices $(A_t)_{t\leqslant T}$ have full row rank,*
3. *Almost surely in V, for all $t \leqslant T$, $\mathrm{Int}\left(\mathcal{S}_{z_{t+1}|(V,z_t)}\right) = \mathbb{R}^{k_z}$.*

Result 2.4. *If (13) holds and Assumption 2.6 is satisfied, then Assumpion 2.4 holds.*

Note that Condition (3) of Assumption 2.6 does not allow the instrument to be time-invariant. However identification can still be obtained for a constant instrument and for example one of our applications in Section 3, studying sequential exogeneity, uses x_1 as an instrument for all time periods. If $z_{it} = z$ does not vary over time in (13), one can see that the condition $e_t'\dot{X}^{(t)} = 0$ becomes $(A_{t+1} - A_t)z + \dot{v}_{it} = 0$: $A_{t+1} - A_t$ must be of full rank, i.e, A_{t+1} has to be "different enough" from A_t. The same intuition applies to our sequential exogeneity example. On the other hand if $A_{t+1} = A_t$, what is needed is that $z_{it+1} - z_{it}$ has rich enough support conditional on V: z must vary sufficiently over time.

Deterministic relation between regressors

The support condition in Assumption 2.4 does not allow any regressor to be a function of other regressors. We now show that if some regressors do depend deterministically on others, it is possible to rewrite the model and obtain sufficient conditions guaranteeing invertibility of the matrix.

We write $x_{it} = (x_{it}^1, .., x_{it}^s, x_{it}^{s+1}, .., x_{it}^{d_x})$, where the first s components of x_{it} do not have functional dependence, while there are $d_x - s$ functions $(l_k)_{s+1 \leqslant k \leqslant d_x}$ such that for $s+1 \leqslant k \leqslant d_x$, $x_{it}^k = l_k(x_{it}^1, .., x_{it}^s)$. We define $\dot{X}^s = (\dot{x}_i^1, .., \dot{x}_i^s)$ the collection of time differences for the first s components of x_{it}. With this new setting we can now rewrite Assumption 2.4.

Assumption 2.7. *The following holds almost surely in V.*

1. $\text{Int}\left(\mathcal{S}_{\dot{X}^s|V}\right) \neq \varnothing$,
2. *For all $s+1 \leqslant k \leqslant d_x$, l_k is a continuous function,*
3. *For all $t \leqslant T-1$, there exists $\dot{X}^{(t)}$ such that $(\dot{X}^{(t)})^s \in \text{Int}\left(\mathcal{S}_{\dot{X}^s|V}\right)$, $p_{\dot{X}|V}(\dot{X}^{(t)}|V) > 0$, $Rank(\dot{X}^{(t)}) = d_x$ and $\dot{X}^{(t)\prime} e_t = 0$.*

Note that the support condition is on $\dot{X}^s$ while the orthogonality condition is on the whole collection of columns of $\dot{X}$. Assuming full rank implies that the l_k functions cannot be linear. The proof of invertibility of $\mathcal{M}$ under Assumption 2.7 follows the same steps as under Assumption 2.4, using now continuity of the functions $(l_k)_{s+1 \leqslant k \leqslant d_x}$.

Discrete distributions

Because of the support conditions they include, the various sets of assumptions we suggested do not handle the case where the conditional distribution of X given V is discrete. This implies that if the control variable comes from a selection equation $x_{it} = c(z_{it}, v_{it})$, then z_{it} cannot be a discrete random variable. It is however possible to extend the previous framework to obtain invertibility of $\mathcal{M}(V)$ when z is a discrete random variable. We point out this compatibility here, which extends to the overall identification argument, but throughout the rest of the paper we will assume for convenience that x, z and v are continuously distributed. We assume here that $x_{it} = c(z_{it}, v_{it})$, and z is a discrete random vector that takes finitely many values. More precisely, we assume that the vector Z_i takes $N(V)$ values with positive probability, conditional on $V_i = V$.

For each value $Z_{(N)}$, $N \leqslant N(V)$, we denote by $\dot{X}_{(N)}$ and $M_{(N)}$ the corresponding matrix of regressors and projection matrix. Using the fact that each $M_{(N)}$ is an orthogonal projection matrix, we look at the singularity condition on $\mathcal{M}(V)$. For a given $V \in \mathcal{S}_V$,

$$\begin{aligned} \mathcal{M}(V) \notin GL_{T-1}(\mathbb{R}) &\Leftrightarrow \exists c \in \mathbb{R}^{T-1}, \mathcal{M}(V)\, c = 0 \\ &\Rightarrow \exists c \in \mathbb{R}^{T-1},\, c'\, \mathcal{M}(V)\, c = 0 \end{aligned}$$

$$\Rightarrow \exists c \in \mathbb{R}^{T-1}, \sum_{N \leqslant N(V)} c' M'_{(N)} M_{(N)} c = 0$$
$$\Rightarrow \exists c \in \mathbb{R}^{T-1}, \sum_{N \leqslant N(V)} ||M_{(N)} c||^2 = 0 \Rightarrow \forall N \leqslant N(V),\ M_{(N)} c = 0.$$

An assumption yielding invertibility of $\mathcal{M}(V)$, $\mathbb{P}_V$ − a.s, is the following.

Assumption 2.8. *Almost surely in V,*

For each $t \leqslant T-1$, there exists $N_t[V] \leqslant N(V)$ such that $\dot{X}'_{(N_t[V])} d_t = 0$.

Result 2.5. *If Assumption 2.8 holds, then $\mathcal{M}(V)$ is nonsingular $\mathbb{P}_V$ a.s.*

The proof of this result follows as in the proof of Result 2.2, without the continuity arguments.

Moreover, we can use the argument of Result 2.3 if x_{it} is scalar: indeed in this case, what is required for invertibility of $\mathcal{M}(V)$ is that there are two draws $\dot{X}_{N1}$ and $\dot{X}_{N2}$ from the conditional distribution of $\dot{X}$ given V, such that the intersection of the null spaces of M_{N1} and M_{N2}, the corresponding projection matrices, is the trivial set $\{0\}$. That is, if x_{it} is scalar, $\mathcal{M}(V)$ is invertible if conditional on V, there are two draws $\dot{X}_{N1}$ and $\dot{X}_{N2}$ that are not collinear.

2.4 Extensions

2.4.1 Combining random coefficients with common parameters

If it is known to the researcher that the random coefficients associated to some covariates $l_{it} \in \mathbb{R}^{d_l}$ have a degenerate distribution, we propose a different procedure. Consider the model

$$y_{it} = l'_{it} b + x'_{it} \mu_i + \epsilon_{it}, \tag{14}$$

where $x_{it} = (x^1_{it}, x^2_{it}) \in \mathbb{R}^{d_x}$ where as in Model (2), $x^1_{it} \in \mathbb{R}^{d_1}$ are exogenous regressors and $x^2_{it} \in \mathbb{R}^{d_2}$ are allowed to be endogenous, and we also write $l_{it} = (l^1_{it}, l^2_{it}) \in \mathbb{R}^{k_m}$ where $l^1_{it} \in \mathbb{R}^{d_{l1}}$ is exogenous, $l^2_{it} \in \mathbb{R}^{d_{l2}}$ is endogenous.

In what follows we explain how to proceed when l_{it} is either exogenous or endogenous. In the case where the control variables are the residuals of the regression of x^2_{it}, these extensions are useful for two reasons. First, if all coefficients are assumed heterogeneous, T is required to be greater than or equal to $d_x + d_l + 2$, which means that the vector of control variables will be of dimension at least $(d_{l2} + d_2)(d_x + d_l + 2)$. V is an argument of the function g which will be nonparametrically estimated. A high dimension of V is undesirable because of the curse of dimensionality. However if l_{it} is known to have homogeneous impact, T needs to be higher than $(d_x + 2)$ which is a less restrictive requirement. Moreover, the nonparametric regressions necessary to construct $g(V)$ will only be conditional on the vector of control variables constructed for x^2_t as our extensions do not require the construction of control variables for l^2_t. We will show that the dimensions of conditioning

sets in that case does not have to exceed $d_2(d_x + 2)$ (which is reached if T is taken to be exactly $d_x + 2$).

We modify Assumption 2.1 (2) and impose $\mathbb{E}(\epsilon_{it}|Z_i^L, X_i, V_i) = f_t(V_i)$, where as before, V_i is an identified function of the regressors X_i and the instruments Z_i, and Z_i^L is composed of L_i^1 and instruments for L_i^2. The matrices M_i and Q_i are the same objects and depend on $\dot{X}_i$.

The within-group operation gives

$$M_i\dot{y}_i = M_i\dot{L}_i b + M_i g(V_i) + M_i\dot{u}_i, \text{ with } \mathbb{E}(M_i\dot{u}_i|Z_i^L, X_i, V_i) = 0, \tag{15}$$

which implies

$$\begin{aligned} \mathbb{E}(M_i\dot{y}_i|V_i) &= E(M_i\dot{L}_i|V_i)\, b + \mathcal{M}(V_i)g(V_i) \\ \Rightarrow M_i\,\mathcal{M}(V_i)^{-1}\mathbb{E}(M_i\dot{y}_i|V_i) &= M_i\,\mathcal{M}(V_i)^{-1}E(M_i\dot{L}_i|V_i)\, b + M_i g(V_i), \end{aligned}$$

where the first equality holds by the law of iterated expectations since M_i is function of X_i.

The left multiplication by $M_i\mathcal{M}(V_i)^{-1}$ creates the term $M_i g(V_i)$, which also appears in (15). This suggests a modification of the procedure developed in Robinson (1988) for the identification of b. Indeed, defining $\Delta\dot{y}_i = \dot{y}_i - \mathcal{M}(V_i)^{-1}\mathbb{E}(M_i\dot{y}_i|V_i)$ and $\Delta\dot{L}_i = \dot{L}_i - \mathcal{M}(V_i)^{-1}E(M_i\dot{L}_i|V_i)$, we obtain

$$M_i\,\Delta\dot{y}_i = M_i\,\Delta\dot{L}_i b + M_i\dot{u}_i.$$

Since $\mathbb{E}(\dot{Z}_i^L\, M_i\dot{u}_i) = \mathbb{E}(\dot{Z}_i^L\, M_i\mathbb{E}(\dot{u}_i|Z_i^L, X_i, V_i)) = 0$, (note that the multiplication is simplified by $M_i = M_i' = M_i^2$) we obtain

$$b = \mathbb{E}(\dot{L}_i^{W\prime}\, M_i\,\Delta\dot{L}_i)^{-1}\mathbb{E}(\dot{Z}_i^{L\prime}\, M_i\Delta\dot{y}_i), \tag{16}$$

under the assumption that $\mathbb{E}(\dot{Z}_i^{L\prime}\, M_i\,\Delta\dot{L}_i)$ is nonsingular.

Once b is identified, then identification of g and $\mathbb{E}(\mu_i)$ will be obtained by applying the results of Sections 2.2.2 and 2.2.3 to $y_{it} - l_{it}'b$.

If $d_x = 1$, that is, the researcher is interested in relaxing the homogeneity assumption for one endogenous regressor, then it is required that $T \geqslant 3$. If the other regressors are exogenous, v_{it} will be scalar. Taking $T = 3$, then the dimension of the conditioning set for the identification of δ is 3, independently of the number of regressors in l_{it}.

2.4.2 Case where $T > d_x + 2$

The identification method developed here is constructive and suggests natural estimators. However a first step requires the identification of g, function of a Tk_2 dimensional vector. If T is too large, identification would still hold but estimation of conditional expectations, conditioning on a

set of variables of high dimension, has undesirable properties by the curse of dimensionality. For identification, it is required that $T \geqslant k_x + 2$, so if $T > k_x + 2$, one can select $k_x + 2$ time periods among the T available and obtain identification. However, it is possible[4] to use the T time periods without increasing the dimension of the conditioning set.

We assume here that $T > k_x + 2$, and denote $\mathcal{T}$ the set of subsets of $\{1, .., T\}$ of cardinality $k_x + 2$. The cardinality of $\mathcal{T}$ is $\binom{T}{k_x+2}$. Consider $\tau \in \mathcal{T}$, a subset of $k_x + 2$ time periods that we write $\tau = (t_1, ..., t_{k_x+2})$ where $t_1, < \cdots < t_{k_x+2}$. We write with a superscript τ the vectors that are defined using only the time periods in τ. For instance, $V_i^\tau = (v_{it_1}, .., v_{it_{k_x+2}})$. Then the model implies

$$y_i^\tau = X_i^\tau \mu_i + \epsilon_i^\tau.$$

Now we modify the control function assumption.

Assumption 2.9. *There exist a set of functions* $(h_t^\tau)_{\substack{t\in\tau \\ \tau\in\mathcal{T}}}$ *and identified functions* $(C_t)_{t\leqslant T}$ *such that, defining* $v_{it} = C_t(x_{it}, z_{it}) \in \mathbb{R}^{d_v}$,

$$\forall\, \tau \in \mathcal{T},\ \forall\, t \in \tau,\ \mathbb{E}(\epsilon_{it} \,|\, X_i^\tau, V_i^\tau) \,=\, h_t^\tau(V_i^\tau).$$

Indeed the assumption that we used in the main model is $\mathbb{E}(\epsilon_{it} \,|\, X_i, V_i) \,=\, f_t(V_i)$, which does not imply Assumption 2.9. If Assumption 2.1 (2) holds, then by the law of iterated expectations $\mathbb{E}(\epsilon_{it} \,|\, X_i^\tau, V_i^\tau) \,=\, \mathbb{E}(f_t(V_i) \,|\, X_i^\tau, V_i^\tau)$ which is not necessarily a function of V_i^τ only. However, the independence assumptions we make in all our applications directly satisfy Assumption 2.9.

For a given τ in $\mathcal{T}$, changing the definition of g_t to $g_t^\tau = h_{t+1}^\tau - h_t^\tau$, identification of the vector of functions g^τ follows from the same first step provided that $\mathcal{M}^\tau(V^\tau)$ is invertible. In the main model, identification of $\mathbb{E}(\mu_i)$ follows from (10), which would become $\mathbb{E}(\mu_i) \,=\, \mathbb{E}(Q_i^\tau \dot{y}_i^\tau \,-\, Q_i^\tau g^\tau(V_i^\tau))$. But since this holds for all subset τ, we can also write

$$\mathbb{E}(\mu_i) \,=\, \frac{1}{\binom{T}{k_x+2}} \sum_{\tau\in\mathcal{T}} \mathbb{E}(Q_i^\tau \dot{y}_i^\tau \,-\, Q_i^\tau g^\tau(V_i^\tau)).$$

3 Applications and Variations of the model

In this section, we propose direct applications of the model where we suggest control variables and give conditions under which the identification assumptions, Assumptions 2.1, 2.2 and 2.3, hold. We also describe some models, different from the main model studied in the identification section, but where, using the appropriate control variables, the two-step approach also provides identification results under some conditions.

[4] I thank Donald Andrews for this suggestion.

3.1 Omitted variables

As we explained in the introduction, the model studied in this paper fits particularly well a framework in which the endogeneity of x_{it}^2 is caused by time-varying omitted variables. Indeed, we introduce an unobserved variable r_{it}, and consider the following model

$$y_{it} = x_{it}^{1\prime}\alpha_i^1 + x_{it}^{2\prime}\alpha_i^2 + r_{it} + u_{it}. \tag{17}$$

Equation (17) is a special case of (1), where $\epsilon_{it} = r_{it} + u_{it}$. Assume there exists an instrument z_{it} and two functions g_t and h_t such that $x_{it}^2 = g_t(x_{it}^1, z_{it}) + h_t(r_{it}, \nu_{it})$, with $\mathbb{E}(h_t(r_{it}, \nu_{it})|x_{it}^1, z_{it}) = 0$, and take the control variable to be $v_{it} = h_t(r_{it}, \nu_{it})$. The control variable is identified as the residual of the regression of x^2 on x^1 and z. We also assume the strict exogeneity condition $\mathbb{E}(u_{it}|\nu_i, R_i, X_i^1, Z_i) = 0$, as well as $\mathbb{E}(r_{it}|V_i, X_i^1, Z_i) = f_t(V_i)$, which will be satisfied in particular if for all $t \leqslant T$, $(r_{it}, \nu_{it}) \perp\!\!\!\perp (x_{is}^1, z_{is})_{s \leqslant T}$. This is a possible formalization of x^2 being endogenous because of omitted variables. Under these assumptions,

$$\mathbb{E}(\epsilon_{it}|V_i, X_i) = \mathbb{E}(\mathbb{E}(\epsilon_{it}|V_i, X_i^1, Z_i)|V_i, X_i) = f_t(V_i) + \mathbb{E}(\mathbb{E}(u_{it}|V_i, X_i^1, Z_i)|V_i, X_i) = f_t(V_i),$$

which satisfies Assumption 2.1.

In this model the endogeneity is caused by an omitted variable: if r_{it} is not constant over time, the endogeneity cannot be controlled for using a fixed-effect transformation. In a later part of this section, we provide an example in which the strict exogeneity condition imposed in Arellano and Bonhomme (2012) is relaxed.

3.2 Heterogeneous production function

Consider a decision variable x_{it} chosen by an agent and an outcome variable y_{it} realized after the choice of x_{it}, given by

$$y_{it} = x_{it}'\mu_i + \epsilon_{it}.$$

Such a production function can be used to model an education outcome, where x_{it} is any type of parental investment, and the random coefficients represent an heterogeneity in the returns to investment, at the child level. But y_{it} can also be a firm or farm output, with x_{it} being capital, labor and/or land inputs.

The use of a triangular system in such a model is suggested in Imbens and Newey (2009) and we follow this example here, using for each time period the decision problem to obtain a selection equation. An important difference is that we assume that the agent does not know (μ_i, ϵ_{it}) at the time of the decision. Instead, she has information about it contained in $\eta_{it} \in \mathbb{R}$, scalar random variable. Writing $y_{it} = q(x_{it}, \mu_i, \epsilon_{it})$, and $C_t(x, z)$ a cost function with z cost shifters, the choice

of x_{it} maximizes an expected profit,

$$x_{it} \;=\; \underset{x}{\operatorname{argmax}}\, \mathbb{E}\left(q(x, \mu_i, \epsilon_{it}) - C_t(x, z_{it}) \,|\, z_{it}, \eta_{it}\right). \tag{18}$$

This implies the existence of a function H_t such that $x_{it} = H_t(z_{it}, \eta_{it})$. We assume that for all $t \leqslant T$, $H_t(z_{it}, \eta)$ is strictly monotonic in η with probability 1, η_t is continuously distributed and its CDF is strictly increasing. We also assume that $(\eta_{it}, \epsilon_{it})_{t\leqslant T} \perp\!\!\!\perp (z_{it})_{t\leqslant T}$ and defining $v_{it} = F_{x_t|z_t}(x_{it}|z_{it}) = F_{\eta_t}(\eta_{it})$, these assumptions imply that ϵ_{it} is independent of Z_i conditional on V_i. Therefore,

$$\mathbb{E}(\epsilon_{it}|V_i, X_i) \;=\; \mathbb{E}\left(\mathbb{E}(\epsilon_{it}|V_i, Z_i)|V_i, Z_i\right) = \mathbb{E}(\mathbb{E}(\epsilon_{it}|V_i, Z_i)|X_i, V_i)) \;=:\; f_t\left(V_i\right).$$

This proves that the model studied here satisfies the control function assumption, Assumption 2.1. If in addition the invertibility assumption, Assumption 2.3, holds, the identification results obtained in the previous section apply, and the average returns to input are identified. Note however that the level of generality of this result is restricted. First, η_{it}, which could be the productivity in the standard model, is scalar, while there is more than one scalar unknown to the agent. And second, we need to impose $(\eta_{it}, \epsilon_{it})_{t\leqslant T} \perp\!\!\!\perp (z_{it})_{t\leqslant T}$, while one might reasonably want to allow z_{it+1} to be correlated with $(\eta_{it}, \epsilon_{it})$.

Note that a direct application of the identification results in Imbens and Newey (2009) would require the instruments to satisfy the condition $z_{it} \perp\!\!\!\perp (\mu_i, \epsilon_{it}, \eta_{it})$, while we do not impose any conditions on the joint distribution of (μ_i, x_i, z_i). In a model of heterogenous production function, finding instruments independent of the unobserved heterogeneity μ_i might be challenging. In this case and if the panel is long enough, our two-step approach would guarantee identification.

3.3 Sample selection

Now consider a panel model with random coefficients and sample selection. Loosely speaking, if the selection is correlated with the disturbance of the main equation, an endogeneity problem arises, since the regressors of the selected individuals will be correlated with the disturbance as well. Das, Newey, and Vella (2003) study a nonparametric model of sample selection in a cross sectional setting, specifying a selection equation which provides them with a control variable.

Our selection equation will be similar, so some of our arguments closely follow theirs, but our model differs in some respects: it includes random coefficients and the outcome equation depends linearly on the regressors. Moreover, we consider a panel model while theirs is a cross-section. This is non negligible as we will for instance consider selection in all periods in addition to selection in each period separately.

The selection model we consider is

$$y_{it}^{*} \;=\; x_{it}'\,\mu_i + \alpha_i + \epsilon_{it},$$

$$d_{it} \;=\; \mathbb{1}\left(\eta_{it} \leqslant C_t(x_{it}, z_{it})\right), \tag{19}$$
$$y_{it} \;=\; d_{it}\, y^*_{it},$$

where z_{it} is an instrument. Let $d_i \;=\; (d_{it})_{t\leqslant T}$, and write $d_i = 1$ to denote the event that $d_{it} = 1$ for all $t \leqslant T$. Also, let $p_{it} \;=\; \mathbb{E}\left(d_{it} | \, x^1_{it}, z_{it}\right) \;=\; \mathbb{P}\left(\eta_{it} \leqslant C_t(x_{it}, z_{it})\right)$, $P_i \;=\; (p_{it})_{t\leqslant T}$, and assume that for each t there is a function f_t such that for all $t \leqslant T$,

$$\mathbb{E}\left(\epsilon_{it} | d_i = 1, X_i, P_i\right) \;=\; f_t(P_i). \tag{20}$$

Note that this assumption is satisfied in particular if $(\epsilon_{is}, \eta_{is})_{s\leqslant T} \perp\!\!\!\perp (X_i, Z_i)$ and if the cdf of η_t, F_t, is strictly increasing. Indeed, in this case defining $\nu_{it} \;=\; F_{\eta,t}(\eta_{it})$, $\nu_i \;=\; (\nu_{it})_{t\leqslant T}$, $p_{it} \;=\; F_{\eta,t}(C_t(x_{it}, z_{it}))$, and $d_{it} \;=\; \mathbb{1}\left(\nu_{it} \leqslant p_{it}\right)$, then

$$\mathbb{E}\left(\epsilon_{it} | d_i = 1, X_i, P_i\right) \;=\; \mathbb{E}\left(\mathbb{E}\left(\epsilon_{it} | \nu_i, X_i, Z_i\right) | d_i = 1, X_i, P_i\right) = \mathbb{E}\left(\mathbb{E}\left(\epsilon_{it} | \nu_i\right) | \nu_i \leqslant P_i\right) \;:=\; f_t(P_i),$$

which is as desired (where by an abuse of notation we write the inequality $\nu_i \leqslant P_i$ to denote the inequality component by component). The joint distribution of $(\epsilon_{it}, \eta_{)}$ is unrestricted under these assumptions. The conditional expectation has a form similar to the control function assumption we maintained in the identification section on the main model, where the control variable is now p_{it} and is identified through a cross sectional regression of d_{it}, for each period t. We write $u_{it} \;=\; \epsilon_{it} - f_t(P_i)$. The identification argument in this selection model will have a two-step structure similar to that of the main model. The important difference is that all the conditional expectations are evaluated for the subsample such that $d_i = 1$, that is, the subsample of individuals who are selected in all periods. To be more precise, define $\tilde{x}_{it} \;=\; d_{i\,t+1} x_{i\,t+1} - d_{it} x_{it}$, $\tilde{g}_t(P_i) \;=\; d_{i\,t+1} f_{t+1}(P_i) - d_{it} f_t(P_i)$, and similarly, $\tilde{u}_{it}$ and the matrices and vectors $\tilde{X}_i$, $\tilde{u}_i$, $\tilde{g}(P_i)$, $\tilde{M}_i$ and $\tilde{Q}_i$. Note that for the subsample such that $d_i \;=\; 1$, we have $\tilde{X}_i = \dot{X}_i$. Hence,

$$\begin{aligned}\mathbb{E}\left(\tilde{M}_i \dot{y}_i \,|\, (d_i \;=\; 1), P_i\right) \;&=\; \mathbb{E}\left(\tilde{M}_i \tilde{X}_i \mu_i + \tilde{M}_i \tilde{g}(P_i) + \tilde{M}_i \tilde{u}_i \,|\, (d_i \;=\; 1), P_i\right) \\ &=\; \mathbb{E}\left(M_i \dot{X}_i \mu_i + M_i g(P_i) + M_i \dot{u}_i \,|\, (d_i \;=\; 1), P_i\right) \;=\; \mathcal{M}(P_i) g(P_i),\end{aligned}$$

where we define $\mathcal{M}(P_i) = \mathbb{E}\left(M_i \tilde{y}_i \,|\, (d_i \;=\; 1), P_i\right)$. This is the first step equation. The second step equation will be given by

$$\mathbb{E}\left(\tilde{Q}_i \tilde{y}_i - \; \tilde{Q}_i \tilde{g}(P_i) | d_i = 1\right) \;=\; \mathbb{E}\left(\mu_i | d_i = 1\right).$$

Assumption 3.1. *$\mathbb{E}\left(\epsilon_t | d = 1, X, P\right) \;=\; f_t(P)$, and $\mathcal{M}(P)$ is invertible almost surely in P.*

Result 3.1. *Under Assumption 3.1, $\mathbb{E}\left(\mu | d = 1\right)$ is identified.*

The identified object is the average effect conditional on selection, $\mathbb{E}(\mu|d = 1)$ which is in general different from $\mathbb{E}(\mu)$ unless $\mu \perp\!\!\!\perp (\eta, X, Z)$. Comparing an estimated value of $\mathbb{E}(\mu|d = 1)$ with an estimate derived from a model without random coefficients would allow one to test whether or not there is heterogeneity. The type of counterfactual that one can compute with this object would describe the effect of policies which impact the intensive margin, not the extensive margin, i.e, which do not affect whether individuals are selected. If $T > d_x + 2$, we recommend using the procedure described in Section 2.4.2 and computing the average effect conditional on being selected in a subset of time periods. Averaging over all subsets identifies a conditional average effect under some additional conditions. This avoids using only the subsample of individuals for whom $d_{it} = 1$ for all $t \leqslant T$, which can be quite small if T is large (and still fixed).

Note that the model allows for regressors s_{it} to be suject to selection as well, that is, regressors not observed for the population such that $d_{it} = 0$ (for example, the wage variable is not defined for unemployed individuals). As long as these regressors are not arguments of the function C_t and the condition $\mathbb{E}\left(\epsilon_t|d = 1, X, S, P\right) = f_t(P)$ holds, the identification argument remains valid.

This sample selection model can also be adapted to the case where some of the regressors are endogenous. If these regressors are not multiplied by random coefficients, the argument of Section 2.4.1 can be applied. If they are accompanied by random coefficients, on the other hand, we suggest using the control function approach on the endogenous regressors, the control variables being for instance the residuals of the regression of the endogenous regressors on the exogenous regressors and instruments. The identification method developed above would then use a vector of control variables which include these residuals in addition to the propensity scores.

One last case worth mentioning, to which our two-step approach can be adapted, is when some regressors are endogenous and subject to selection. We briefly explain how to construct the control variables. The model is

$$\begin{aligned} y_{it} &= d_{it}\, y_{it}^*, \quad \text{with} \quad y_{it}^* = x_{it}^{1\,\prime}\, \mu_i^1 + x_{it}^{2\,\prime}\, \mu_i^2 + \epsilon_{it}, \\ x_{it}^2 &= d_{it}\, x_{it}^{2*}, \quad \text{with} \quad x_{it}^{2*} = \pi_t^2(x_{it}^1, z_{it}^1) + v_{it}, \\ d_{it} &= \mathbb{1}\left(\nu_{it} \leqslant p_t(x_{it}^1, z_{it}^1, z_{it}^2)\right) = \mathbb{1}\left(\nu_{it} \leqslant p_{it}\right), \end{aligned} \tag{21}$$

with $\nu_{it} \sim \mathcal{U}[0;1]$. Assume that $(\epsilon_{is}, \nu_{is}, v_{is})_{s\leqslant T} \perp\!\!\!\perp (x_{is}^1, z_{is}^1, z_{is}^2)_{s\leqslant T}$. In this model, x^2 is the endogenous regressor. Identification of p_{it} holds by $\mathbb{E}\left(d_{it}|\, x_{it}^1, z_{it}^1, z_{it}^2\right) = p_{it}$. Moreover

$$\mathbb{E}\left(v_{it}|d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2\right) = \mathbb{E}\left(\mathbb{E}\left(v_{it}|\nu_{it}\right)|(\nu_{it} \leqslant p_{it}), x_{it}^1, z_{it}^1, z_{it}^2\right) := \phi_t(p_{it}),$$

implying $\mathbb{E}\left(x_{it}^2|d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2\right) = \pi_t^2(x_{it}^1, z_{it}^1) + \phi_t(p_{it})$. This gives

$$x_{it}^2 - \mathbb{E}\left(x_{it}^2|d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2\right) = v_{it} - \phi_t(p_{it}) := \bar{v}_{it}, \tag{22}$$

where $\bar{v}_{it}$ is identified. That is, the residuals for individuals selected in the sample are also control variables. The corresponding estimator will not need generated covariates. We define again $\nu_i = (\nu_{is})_{s \leqslant T}$, and similarly d_i, P_i, $\bar{V}_i$, $X_i = (X_i^1, X_i^2)$ and $\phi(P_i) = (\phi_t(p_{it}))_{t \leqslant T}$. Note that the function $(V_i, P_i) \mapsto (\bar{V}_i, P_i)$ is one-to-one. Therefore,

$$\begin{aligned}
\mathbb{E}\left(\epsilon_{it} | d_i = 1, X_i, \bar{V}_i, P_i\right) &= \mathbb{E}\left(\mathbb{E}\left(\epsilon_{it} | \nu_i, V_i, X_i^1, Z_i^1, Z_i^2\right) | (\nu_i \leqslant P_i), X_i, V_i, P_i\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\epsilon_{it} | \nu_i, V_i\right) | (\nu_i \leqslant P_i), X_i, V_i, P_i\right) := F_t(V_i, P_i), \\
&= F_t(\bar{V}_i + \phi(P_i), P_i) := f_t(\bar{V}_i, P_i).
\end{aligned}$$

This conditional expectation is as required by Assumption 2.1 to use our two-step method, where the control variables are $(\bar{V}_i, P_i)$.

This double use of the control function approach is already suggested in Das et al. (2003). We presented here a slight modification such that the identification requires two steps instead of three. This allows us to use our formula for the asymptotic variance provided that π^2 is not an object of interest. Indeed, increasing the number of steps changes the asymptotic variance matrix.

3.4 Relaxing a strict exogeneity condition

Instead of focusing on contemporaneous endogeneity, that is relaxing restrictions on the joint dependence of (ϵ_{it}, x_{it}), one can use our framework to relax restrictions on the joint dependence of $(\epsilon_{it}, x_{it+1})$. This corresponds to relaxing the strict exogeneity condition imposed in Arellano and Bonhomme (2012), while keeping the contemporaneous exogeneity.

The model is

$$y_{it} = x_{it}' \mu_i + \alpha_i + \epsilon_{it}, \tag{23}$$

where $\mathbb{E}(\epsilon_{it} | x_{it}) = 0$ but where the strict exogeneity condition $\mathbb{E}(\epsilon_{it} | X_i) = 0$ fails to hold because there is a feedback effect. For instance, x_{it+1} can be impacted by ϵ_{it}. As in the main model, the idea is to look for an identified vector V_i such that $\mathbb{E}(\epsilon_{it} \,|\, x_{i1}, ..x_{iT}, v_{i1}, ..v_{iT}) = f_t(v_{i1}, ..v_{iT}) = f_t(V_i)$.

We consider the case where x_{it} is a Markov process. We write $x_{i\,t+1} = m_t(x_{it}) + \eta_{i\,t+1}$, $1 \leqslant t \leqslant T-1$, where η_{it} are i.i.d over time. One could alternatively consider $x_{i\,t+1} = m_{t+1}(x_{it}, \eta_{i\,t+1})$ with η scalar and m_t strictly monotonic in η and use the control variable suggested in Imbens and Newey (2009). We also assume that $(\epsilon_{it}, \eta_{i\,t+1}, .., \eta_{iT}) \perp\!\!\!\perp (x_{i1}, \eta_{i2}, .., \eta_{it})$, which implies that $(\epsilon_{it}, \eta_{i\,t+1}, .., \eta_{iT}) \perp\!\!\!\perp (x_{i1}, .., x_{it})$. That is, the innovations giving the evolution of x after time t and ϵ_t are independent of past values of x. However the joint distribution of $(\epsilon_{it}, \eta_{i\,t+1}, .., \eta_{iT})$ is not restricted, which corresponds to relaxing the above-mentioned strict exogeneity. Then, for all t less than T,

$$\mathbb{E}(\epsilon_{it} \,|\, x_{i1}, .., x_{iT}, \eta_{i2}, .., \eta_{iT}) = \mathbb{E}(\epsilon_{it} | x_{i1}, \eta_{i2}, .., \eta_{iT}) = \mathbb{E}(\epsilon_{it} | \eta_{i\,t+1}, .., \eta_{iT})$$

$$:= f_t(\eta_{i2}, .., \eta_{iT}) := f_t(V_i),$$

where the first equality holds by the Markov structure of x, and the second equality holds by the independence assumption on the error terms. We define $V_i = (\eta_{i2}, .., \eta_{iT})$ to be the control variable. Note that the η_{it} are all identified as the residuals of reduced form regressions. Conditional mean independence is sufficient, that is, $\mathbb{E}(\epsilon_{it}|x_{i1}, \eta_{i2}, .., \eta_{iT})$ being independent of x_{i1}.

Defining as previously $M_i = I - \dot{X}_i(\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i'$, $\mathcal{M}(V_i) = \mathbb{E}(M_i|V_i)$, $u_{it} = \epsilon_{it} - f_t(V_i)$ and $g_t(V_i) = f_{t+1}(V_i) - f_t(V_i)$, (23) together with the independence assumptions guarantees, as in the main model,

$$\mathbb{E}(M_i\dot{y}_{it}|V_i) = \mathcal{M}(V_i)g(V_i), \text{ and } \mathbb{E}(Q_i\dot{y}_i) = \mathbb{E}(\mu_i) + \mathbb{E}(Q_ig(V_i). \tag{24}$$

A two-step procedure, as in the main model, requires $\mathcal{M}(V)$ to be nonsingular: a first step identifies the vector of functions g and a second step identifies the average effect.

The method applies as long as $\mathcal{M}(V)$ is nonsingular. By definition, $V_i = (\eta_{i2}, ..\eta_{iT})$ while M_i is constructed using the vector of variables X_i: therefore the expectation of M_i conditional on V_i is an expectation over x_{i1} only, which is the instrument here. The invertibility of $\mathcal{M}(V)$ might not seem intuitive. To show that this invertibility condition can actually hold, let us look more closely at the case where x_t is a scalar AR(1) process, that is, $x_{i\,t+1} = \rho x_{it} + \eta_{i\,t+1}$ with $\rho \neq 1$ and $(\epsilon_{it}, \eta_{i\,t+1}, .., \eta_{iT}) \perp\!\!\!\perp (x_{i1}, \eta_{i2}, .., \eta_{it})$. By definition, $M_i = I - \dot{X}_i\dot{X}_i'/(\dot{X}_i'\dot{X}_i)$. Moreover,

$$x_{it} = \rho^{t-1}x_{i1} + \sum_{s=2}^{t}\rho^{t-s}\eta_{is} \Rightarrow \quad x_{i\,t+1} - x_{it} = \rho^{t-1}(\rho-1)x_{i1} + (\rho-1)\sum_{s=2}^{t}\rho^{t-s}\eta_{is} + \eta_{i\,t+1},$$

therefore defining the two vectors $C_1 = [\rho - 1](1, \rho, .., \rho^{T-2})' \in \mathbb{R}^{T-1}$ and $C_2(V) = (\eta_{i2}, [\rho-1]\eta_{i2} + \eta_{i3}, .., [\rho-1]\sum_{s=2}^{T}\rho^{T-s}\eta_{is} + \eta_{iT})' \in \mathbb{R}^{T-1}$, we can write

$$\dot{x}_i = x_{i1}C_1 + C_2(V_i). \tag{25}$$

For a given value of $\bar{V} \in \mathcal{S}_V$, we proved in Section 2.3 that

$$\begin{aligned}
\mathcal{M}(\bar{V}) \notin GL_{T-1}(\mathbb{R}) &\Leftrightarrow \exists\, a(\bar{V}) \in \mathbb{R}^{T-1},\ a(\bar{V}) \neq 0,\ \mathcal{M}(\bar{V})\, a(\bar{V}) = 0, \\
&\Rightarrow \exists a(\bar{V}),\ M\, a(\bar{V}) = 0\ \ \mathbb{P}_{\dot{X}|V=\bar{V}} a.s, \\
&\Rightarrow \exists a(\bar{V}),\ a(\bar{V}) \text{ is collinear to } \dot{x}\ \ \mathbb{P}_{\dot{X}|V=\bar{V}} \text{ a.s.}
\end{aligned}$$

The draws of $\dot{x}$ from $\mathbb{P}_{\dot{X}|V=\bar{V}}$, as can be seen in (25), differ only in the value of x_1: these draws are the sum of two vectors, $C_2(\bar{V})$ which is fixed since the draws are conditional on $V = \bar{V}$, and x_1C_1 which is proportional to the constant vector C_1. Note that $\mathcal{S}_{x_1|V=\bar{V}} = \mathcal{S}_{x_1}$ since $x_1 \perp\!\!\!\perp V$. If there are at least two nonzero points $\underline{x_1}$ and $\overline{x_1}$ in $\mathcal{S}_{x_1}$ such that $a(\bar{V})$ is collinear to $\underline{x_1}C_1 + C_2(\bar{V})$ and to $\overline{x_1}C_1 + C_2(\bar{V})$, then since $a(\bar{V}) \neq 0$ and $C_1 \neq 0$, this implies that either C_1 and $C_2(\bar{V})$ are

proportional, or $C_2(\bar{V}) = 0$. Note that $C_2(\bar{V}) = 0$ implies $\bar{V} = 0$, and one can show that if C_1 and $C_2(\bar{V})$ are proportional, this implies that $\bar{V} \in \left\{ a. \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} | a \in \mathbb{R} \right\} =: \mathcal{D}$. Hence, the existence of such $\underline{x_1}$ and $\overline{x_1}$ implies $\bar{V} \in \mathcal{D}$, and $\mathcal{D}$ is a subset of $\mathbb{R}^{T-1}$, which has $\mathbb{P}_V$ measure 0 if V_i is continuously distributed on $\mathbb{R}^{T-1}$. We summarize the arguments in the following assumption and result.

Assumption 3.2.

1. *(23) holds and for all* $t \leqslant T$, $x_{i\,t+1} = \rho x_{it} + \eta_{i\,t+1}$ *with* $\rho \neq 1$, $(\epsilon_{it}, \eta_{i\,t+1}, .., \eta_{iT}) \perp\!\!\!\perp (x_{i1}, \eta_{i2}, .., \eta_{it})$, *and* $(X_i, \mu_i, \alpha_i, \epsilon_i, \eta_i)$ *is i.i.d,*
2. *Either* x_{i1} *has a discrete distribution with support including two nonzero points, or* x_{i1} *is continuously distributed and* $\text{Int}(\mathcal{S}_{x_1}) \neq \varnothing$. *Moreover,* $(\eta_{i2}, .., \eta_{iT})$ *is continuously distributed on* $\mathbb{R}^{T-1}$, *and* f_t *is continuous.*

Result 3.2. *Under Assumption 3.2,* $\mathbb{E}(\mu_i', \alpha_i)$ *is identified.*

Proof. Under Assumption 3.2, for a given value $\bar{V}$, if $\mathcal{M}(\bar{V})$ is not invertible, there exist two nonzero draws of x_1, $\underline{x_1}$ and $\overline{x_1}$, with positive density in the continuously distributed case, or probability in the discretely distributed case, such that $\underline{x_1} C_1 + C_2(\bar{V})$ and $\overline{x_1} C_1 + C_2(\bar{V})$ are proportional. That is, if $\mathcal{M}(\bar{V})$ is not invertible then $\bar{V} \in \mathcal{D}$: the function g is identified over $\mathcal{S}_V \backslash \mathcal{D}$. However, g is continuous and the support of V is dense in $\mathbb{R}^{T-1}$, which allows us to identify g over $\mathcal{S}_V$. Since g is identified, the second identification step described in Section 2 allows for identification of $\mathbb{E}(\mu_i)$ and $\mathbb{E}(\alpha_i)$. □

This result is an example of the case where identification does not require the instrument to vary over time, but where the impact of $z_{it} = x_{i1}$ on x_{it} is not the same for each time period. Indeed, one can write $x_{i2} = m_2(x_{i1}) + \eta_{i2}$, $x_{i3} = m_3(m_2(x_{i1}) + \eta_{i2}) + \eta_{i3}$, ... For this reason, we imposed the condition $\rho \neq 1$ in the case $x_{i\,t+1} = \rho x_{it} + \eta_{i\,t+1}$.

4 Estimators

In this section, we explain how to construct an estimator of the average partial effect and provide implementation details in a particular case.

At each step of the argument, the proofs were constructive. To construct an estimator of the APE, our suggestion is therefore to follow the structure of the identification proof replacing population moments with their sample analogs. We assumed that the control variables are given by $v_{it} = C_t(x_{it}, z_{it})$ where C_t is identified: for $\hat{C}_t$ an estimator of this function, either parametric or nonparametric depending on the form of the control variables, v_{it} is estimated with $\hat{v}_{it} = \hat{C}_t(x_{it}, z_{it})$. The conditional expectation functions $\mathcal{M}(V) = \mathbb{E}(M_i | V_i = V)$ and $k(V) = \mathbb{E}(M_i \dot{y}_i | V_i = V)$ are estimated nonparametrically using the generated values of V as regressors and the function

$g = \mathcal{M}^{-1}k$ is estimated plugging in the estimators $\hat{\mathcal{M}}(V)$ and $\hat{k}(V)$ in this formula. The estimator for $\mathbb{E}(\mu|\delta)$ will be a sample analog of Equation (12), plugging in the estimator of g. As we highlighted in Section 2.2.1, the condition $\mathbb{E}(||Q_i||^2) < \infty$ may not hold in the data so out of caution we estimate $\mathbb{E}(\mu|\delta)$, where we defined $\delta_i = \mathbb{1}(\det(\dot{X}_i'\dot{X}_i) > \delta_0)$. Allowing for $\delta_0 \to 0$, as in Graham and Powell (2012), to estimate $\mathbb{E}(\mu)$ without imposing $\mathbb{E}(||Q_i||^2) < \infty$ is the subject of further research.

It is clear that the asymptotic properties will depend on the definition of the control variables. In the remaining part of the paper, we focus on a model with random coefficients on both strictly exogenous and endogenous regressors. The control variables in this model are the residuals of the nonparametric regression of the endogenous regressors on the exogenous regressors and the instruments. We consider this model to be the leading example in the class of models satisfying (1) and Assumption 2.1. Indeed as explained in Section 3.1 it can describe a situation in which regressors and disturbances covary due to the presence of omitted variables. More specifically, the model is

$$\begin{aligned} y_{it} &= x_{it}^{1\prime}\mu_i^1 + x_{it}^{2\prime}\mu_i^2 + \alpha_i + \epsilon_{it}, \\ x_{it}^2 &= b_t(x_{it}^1, z_{it}) + v_{it}, \qquad \mathbb{E}(v_{it}|x_{it}^1, z_{it}) = 0, \end{aligned} \tag{26}$$

where $x_{it}^1 \in \mathbb{R}^{d_1}$, $x_{it}^2 \in \mathbb{R}^{d_2}$, $z_{it} \in \mathbb{R}^{d_z}$, and where Assumption 2.1 holds. Here the regressors x^1 are exogenous while x^2 can be endogenous. Note that $v_{it} = x_{it}^2 - \mathbb{E}(x_{it}^2|x_{it}^1, z_{it})$. The estimators we define for our asymptotic analysis are as described above, where the estimator $\hat{v}_{it}$ is the residual from the nonparametric regression estimation of x_{it}^2, and all estimators of the nonparametric regressions will be series estimators. We provide details on implementation in what follows.

4.1 Series estimator of V_i

The vector of control variables is $V_i = (v_{i1}', .., v_{iT}')' \in \mathbb{R}^{Td_2}$, where $v_{it} = x_{it}^2 - \mathbb{E}(x_{it}^2|x_{it}^1, z_{it})$. We write $\xi_{it} = (x_{it}^1, z_{it})$. Consider the $L \times 1$ vector of approximating functions $r^L(\xi_t) = (r_{1L}(\xi_t), .., r_{LL}(\xi_t))'$. We define the series estimators of the regression function $\mathbb{E}(x_{it}^2|\xi_{it} = \xi_t) = b_t(\xi_t)$ to be $\hat{\beta}_t' r^L(\xi_t)$ where $\hat{\beta}_t$ is $L \times d_2$, and

$$\hat{\beta}_t = (R_t R_t')^{-1} \sum_i r^L(\xi_{it}) x_{it}^{2\prime} = (R_t R_t')^{-1} R_t X_t^{2\prime}, \tag{27}$$

where $R_t = (r^L(\xi_{1t}), .., r^L(\xi_{nt}))$ is $L \times n$ and $X_t^2 = (x_{1t}^2, .., x_{nt}^2)$ is $d_2 \times n$.

The control variables are defined as the residuals of the regression of x_t^2. Later, we will assume that the support $\mathcal{S}_V$ of V is bounded. However, the values obtained using the estimated residuals might not be in $\mathcal{S}_V$: it will be convenient for the asymptotic analysis to introduce a transformation τ of the generated variables to ensure that their transformed values lie in $\mathcal{S}_V$. Specifically, we assume that the support of v_t is of the form $\times_{d=1}^{d_2} [\underline{v}_{td}, \bar{v}_{td}]$ so that the support of V is $\mathcal{S}_V =$

$\times_{d\leqslant d_2,\, t\leqslant T}[\underline{v}_{td}, \bar{v}_{td}]$. We define τ such that for $V = (v_1', .., v_T')' \in \mathbb{R}^{Td_2}$, then $\tau(V) \in \mathbb{R}^{Td_2}$ and the $(d_2(t-1)+d)$th component of $\tau(V)$ satisfies

$$\tau(V)_{(t-1)d_2+d} = \begin{cases} v_{t,d}, & \text{if } v_{t,d} \in [\underline{v}_{td}; \bar{v}_{td}], \\ \underline{v}_{td}, & \text{if } v_{t,d} \leqslant \underline{v}_{td}, \\ \bar{v}_{td}, & \text{if } v_{t,d} \geqslant \bar{v}_{td}, \end{cases}$$

where $v_{t,d}$ is the d^{th} component of v_t.

Write $r_{it} = r^L(\xi_{it})$ and $\hat{b}_{it} = \hat{\beta}_t' r_{it}$. We define $b_{it} = b_t(\xi_{it})$ and $\hat{b}_t = \hat{\beta}_t' r^L$. We also define the residuals $\tilde{v}_{it} = x_{it}^2 - \hat{b}_{it}$, and $\tilde{V}_i = (\tilde{v}_{i1}', .., \tilde{v}_{iT}')'$. Our estimator for V_i will be $\hat{V}_i = \tau(\tilde{V}_i)$: $\tau(\tilde{V}_i)$ is the projection of $\tilde{V}_i$ onto $\mathcal{S}_V$ such that if $\tilde{V}_i$ lies outside of $\mathcal{S}_V$, the function $\tau(\tilde{V})$ is the point on the boundaries of the support that is the closest to V. Note that for all draws of V_i, $\tau(V_i) = V_i$ and $||\hat{V}_i - V_i|| \leqslant ||\tilde{V}_i - V_i||$.

4.2 Series estimator of $\mathcal{M}$ and k

For a real-valued random variable w_i, we propose a series estimator using the generated $\hat{V}_i$ as regressors for the conditional expectation $h^W(V) = \mathbb{E}(w_i | V_i = V)$.

Let $p^K(V) = (p_{1K}(V), .., p_{KK}(V))'$ denote a $K \times 1$ vector of approximating functions. An estimator of $h^W(V)$ is $p^K(V)'\hat{\pi}^W$ where $\hat{\pi}^W$ is a vector of size K given by

$$\hat{\pi}^W = (\hat{P}\hat{P}')^{-1} \sum_i p^K(\hat{V}_i) W_i' = (\hat{P}\hat{P}')^{-1}\hat{P}W,$$

where $\hat{P} = (p^K(\hat{V}_1), .., p^K(\hat{V}_n))$ is $K \times n$ and $W = (w_1, .., w_n)'$ is a vector of size n.

Using this general definition, we construct component by component estimators $\hat{\mathcal{M}}$ and $\hat{k}$ for the matrix and vector valued functions $\mathcal{M}$ and k. We obtain $p^K(V)'\hat{\pi}^{M,st}$ an estimator of the (s,t) component of the matrix $\mathcal{M}$, taking w_i to be $(M_i)_{s,t}$. Similarly, an estimator of the sth component of k will be $p^K(V)'\hat{\pi}^{k,s}$, choosing $w_i = (M_i \dot{y}_i)_s$.

4.3 Construction of $\hat{g}$ and $\hat{\mu}$

Under Assumptions 2.1 and 2.3, we have $g(V) = \mathcal{M}(V)^{-1} k(V)$. A straightforward estimator of g is $\hat{g}(V) = \hat{\mathcal{M}}(V)^{-1}\hat{k}(V)$.

The closed-form expression (12) suggests the use of a sample average to estimate $\mathbb{E}(\mu|\delta)$, plugging in the nonparametric estimator of g evaluated at the generated values. The estimator is

$$\hat{\mu} = \frac{\sum_{i=1}^n \delta_i Q_i [\dot{y}_i - \hat{g}(\hat{V}_i)]}{\sum_{i=1}^n \delta_i} = \frac{\sum_{i=1}^n Q_i^{\bar{\delta}} [\dot{y}_i - \hat{g}(\hat{V}_i)]}{\sum_{i=1}^n \delta_i}.$$

In the next three sections, we study the asymptotic properties of our estimator. Section 5 establishes the rates of convergence of the nonparametric two-step sieve estimators $\hat{k}$ and $\hat{\mathcal{M}}$,

and of $\hat{g}$. Section 6 uses this analysis and states consistency of $\hat{\mu}$. Section 7 focuses on proving asymptotic normality and deriving the asymptotic variance.

5 Convergence rates of the nonparametric two-step estimators

The multi-step estimation procedure only uses closed form expressions: its ease of implementation comes with a layered asymptotic analysis, as each step needs to be analyzed one by one. This type of asymptotic analysis is the subject of the literature on semiparametric estimation with generated covariates. A review of this literature can be found in Mammen, Rothe, Schienle, et al. (2012) and Mammen, Rothe, and Schienle (2016). Among important recent contributions, Hahn and Ridder (2013) derives a general formula of the asymptotic variance of such estimators. However they do not provide results on how to obtain asymptotic normality for given classes of nonparametric estimators. Mammen et al. (2016) establish the asymptotic normality of GMM type estimators depending on a nonparametric nuisance parameter, which is estimated with a nonparametric two-step estimator where the second step is a kernel regression. Hahn, Liao, and Ridder (2018) focuses on sieve estimators.

We introduce some notations. For a vector $a \in \mathbb{R}^p$, $||a||$ is its Euclidean norm. We also denote by $||.||_F$ the Frobenius norm (the canonical norm) in the space of matrices $\mathcal{M}_p(\mathbb{R})$, and $||.||_2$ the matrix norm induced by $||.||$ on $\mathbb{R}^p$ (the spectral norm). We recall that for a given matrix $A \in \mathcal{M}_p(\mathbb{R})$, $||A||_F = \left(\sum_{i,j\leqslant p} a_{ij}^2\right)^{1/2} = \text{tr}(A'A)^{1/2}$. To avoid tedious notations, we will regularly omit the subscript F, $||A||$ without index implies that the norm considered is the Frobenius norm. The index will be displayed when clarity requires it. We define $\lambda_{\min}(A)$ to be the smallest eigenvalue of the matrix A (when it has one), similarly $\lambda_{\max}(A)$, as well as $\lambda_1(A) \leqslant ... \leqslant \lambda_p(A)$ all the eigenvalues ranked by increasing order (when they exist).

We will use the following results. First, for all $A \in \mathcal{M}_p(\mathbb{R})$, $||A||_2 \leqslant ||A||_F$. This inequality also holds for nonsquare matrices. Also, for A a symmetric matrix, $||A||_2 = |\lambda_{\max}(A)|$ and $||A||_F^2 = \sum_{i=1}^p \lambda_i(A)^2$. By definition of $||.||_2$, $||A\,a|| \leqslant ||A||_2\,||a||$.

We also write, for g a vector of functions of $x \in \mathcal{S}_x \subset R^k$, $||g||_\infty = \sup_{\mathcal{S}_x} ||g(.)||$. For $l = (l_1, .., l_k) \in \mathbb{N}^k$, we define $|l| = \sum_{j=1}^k l_j$, and the partial derivative $\partial^l g(x) = \partial^{|l|} g(x)/\partial^{l_1} x_1 ... \partial^{l_k}$. We will use the norm $|g|_d = \max_{|l|\leqslant d} \sup_{x\in\mathcal{S}_x} ||\partial^l g(x)||$. We denote by $\partial g(x)$ the Jacobian matrix $(\partial g(x)/\partial x_1, ..., \partial g(x)/\partial x_k)$. In what follows, T will denote the triangular inequality, M the Markov inequality, CS indicates the use of the Cauchy Schwarz inequality, LLN the weak law of large numbers, C a generic constant (whose value can change from one line to another), and we follow Imbens and Newey (2009) in denoting with CM (for Conditional Markov) the result that if $\mathbb{E}\left(|a_n|\,\middle|\,b_n\right) = O_{\mathbb{P}}(r_n)$ then $|a_n| = O_{\mathbb{P}}(r_n)$. For a sequence $(c_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$, the notation $c_n \to 0$ should

be understood as $c_n \to_{n\to\infty} 0$.

The proof of consistency of $\hat{\mu}$ is split in several steps. We start by deriving the convergence rates of the step estimators $\hat{V}$, $\mathcal{M}$, $\hat{k}$ and $\hat{g}$.

5.1 Sample mean square error for estimator of the control variables

The generated covariates $\hat{V}_i$ are constructed as the estimated residuals of T regressions. Newey et al. (1999) uses similar two step series estimators, and we use some of their results for each of the T nonparametric estimations.

Assumption 5.1.

1. *(X_i, Z_i) is i.i.d, $\mathrm{Var}(x_t^2|\xi_t)$ is bounded, and ξ_t is continuously distributed,*
2. *There exists a $L\times L$ nonsingular matrix Γ_{1t} such that for $R^L(\xi_t) = \Gamma_{1t} r^L(\xi_t)$, then $\mathbb{E}(R^L(\xi_t)R^L(\xi_t)')$ has smallest eigenvalue bounded away from zero uniformly in L,*
3. *There exist γ_1 and β_t^L such that $\sup_{\mathcal{S}_\xi} ||b_t(\xi_t) - \beta_t^{L\prime} r^L(\xi_t)|| \leqslant CL^{-\gamma_1}$,*
4. *For $a_1(L)$ such that $\sup_{\mathcal{S}_\xi} ||R^L(\xi_t)|| \leqslant a_1(L)$, $\sqrt{L/n}\; a_1(L) \to_{n\to\infty} 0$.*

Under Assumption 5.1, Lemma A1 of Newey et al. (1999) shows that $\frac{1}{n}\sum_{i=1}^n || v_{it} - \tilde{v}_{it} ||^2 = O_{\mathbb{P}}\left(L/n + L^{-2\gamma_1}\right) = O_{\mathbb{P}}(\Delta_n^2)$, and $\max_i || v_{it} - \tilde{v}_{it} || = O_{\mathbb{P}}(a_1(L)\Delta_n)$, where $\Delta_n = \sqrt{L/n} + L^{-\gamma_1}$. Note that the same result implies

$$\sup_{\mathcal{S}_{\xi t}} ||b_t - \hat{b}_t|| = O_{\mathbb{P}}(a_1(L)\Delta_n) \tag{28}$$

Define $\tilde{V}_i = (\tilde{v}_{i1}, ..., \tilde{v}_{iT})$. Since $||\hat{V}_i - V_i|| \leqslant ||\tilde{V}_i - V_i||$, the cited result implies

$$\frac{1}{n}\sum_{i=1}^n || V_i - \hat{V}_i ||^2 = O_{\mathbb{P}}\left(L/n + L^{-2\gamma_1}\right) = O_{\mathbb{P}}(\Delta_n^2). \tag{29}$$

$$\max_i || V_i - \hat{V}_i || = O_{\mathbb{P}}(a_1(L)\Delta_n). \tag{30}$$

When b_t is continuously differentiable up to order p, writing $d_\xi = d_1 + d_z$, then Assumption 5.1 (3) holds with $\gamma_1 = p/d_\xi$ for different choices of sieve basis.

Conditions satisfying Assumption 5.1 (2) typically require the support of ξ_t to be bounded and the density of ξ_t to be bounded away from 0 on its support. This restriction is not desirable. Indeed, in applications where the density of the regressors goes to 0 at the boundaries, regressors will be trimmed to consider only a subset of $\mathcal{S}_\xi$ where the density is bounded away from 0. However, we are interested here in the identification and estimation of the average effect $\mathbb{E}(\mu)$, which are useful when a researcher is interested in counterfactuals involving average effects. Trimming on regressors to estimate a conditional effect is contrary to this goal, and the trimming is arbitrary.

We therefore provide a set of conditions allowing the density of the regressor to go to 0 at the boundary of its support when the support is bounded. We follow Imbens and Newey (2009) which

develops an argument of Andrews (1991) (Example II) in assuming a polynomial lower bound on the rate of decrease of the density.

We assume that $\mathcal{S}_{\xi t} \subset \mathbb{R}^{d_\xi}$ is of the form $\times_{d=1}^{d_\xi} [\underline{\xi}_{td}; \bar{\xi}_{td}]$ and defining $f_{\xi t}$ the density of ξ_t, assume that for all $\xi_t \in \mathcal{S}_{\xi t}$, $f_{\xi t}(\xi_t) \geqslant \Pi_{d=1}^{d_\xi} (\xi_{dt} - \underline{\xi}_{dt})^\alpha (\bar{\xi}_{dt} - \xi_{dt})^\alpha$. Under these conditions, a slight modification of Lemma S.3 of Imbens and Newey (2009) shows that if r^L is a polynomial basis of functions, there exists a nonsingular $L \times L$ matrix $\tilde{\Gamma}_{\xi t}$ such that for $\tilde{r}^L(\xi_t) = \tilde{\Gamma}_{\xi t} r^L(\xi_t)$, $\mathbb{E}(\tilde{r}^L(\xi_t)\tilde{r}^L(\xi_t)') = I_L$, implying that Assumption 5.1 (2) holds. They also obtain $\sup_{\mathcal{S}_\xi} ||\tilde{r}^L(\xi)|| \leqslant a_1(L)$ with $a_1(L) = CL^{\alpha+1}$. In this case, the set of conditions in Assumption 5.1 can be modified as follows.

Assumption 5.2.

1. *(X_i, Z_i) is i.i.d,* $\mathrm{Var}(x_t^2|\xi_t)$ *is bounded, and ξ_t is continuously distributed,*
2. *$r^L(.)$ is the power series basis, and* $\forall\, \xi_t \in \mathcal{S}_{\xi t}$, $f_\xi(\xi_t) \geqslant \Pi_{d=1}^{d_\xi} (\xi_{dt} - \underline{\xi}_{dt})^{\alpha_1} (\bar{\xi}_{dt} - \xi_{dt})^{\alpha_1}$,
3. *There exists γ_1 and β_t^L such that* $\sup_{\mathcal{S}_\xi} ||b_t(\xi_t) - \beta_t^{L\prime} r^L(\xi_t)|| \leqslant CL^{-\gamma_1}$,
4. *For* $a_1(L) = L^{\alpha_1+1}$, $\sqrt{L/n}\; a_1(L) \to_{n\to\infty} 0$.

Under Assumption 5.2, (29) and (30) hold. Allowing for unbounded support is a desirable extension as well, and is possible using a method similar to Chen, Hong, and Tamer (2005) and Chen, Hong, Tarozzi, et al. (2008).

5.2 Convergence rates of two-step series estimators

The estimator of the conditional expectation $\mathbb{E}(w_i|V_i = V) = h^W(V)$ is constructed using the generated control variables. This two-step strucure where each step is nonparametric is very similar to the estimators studied in Newey et al. (1999), aside from the panel aspect. However, they impose an orthogonality condition which will not hold for our choices of w_i. More specifically, our model is

$$w = h^W(V) + e^W,\ \mathbb{E}(e^W|V) = 0,$$

and Newey et al. (1999) imposes additionally that $\mathbb{E}(e^W|X^1, V, Z) = 0$, that is, e^W is conditionally mean-independent of the variables used in the first step. This assumption has implications for the convergence rate of the two-step estimator and, as will be obvious in a later part of the paper, on the asymptotic variance of a linear functional of this estimator. This condition does not hold for our choices of w_i, as w_i is either a component of the matrix $M = I - \dot{X}(\dot{X}'\dot{X})^{-1}\dot{X}'$ or of the vector $M\dot{y}$, and

$$\mathbb{E}(M|X^1, X^2, Z) = M \neq \mathbb{E}(M|V) = \mathcal{M}(V),$$
$$\mathbb{E}(M\dot{y}|X^1, X^2, Z) = Mg(V) + M\mathbb{E}(\dot{u}|X^1, X^2, Z) \neq \mathbb{E}(M\dot{y}|V) = k(V) = \mathcal{M}(V)\, g(V).$$

This difference has been documented for instance in Hahn and Ridder (2013) and Mammen et al. (2016). To account for the extra term $\mathbb{E}(e^W|X^1,V,Z) = \mathbb{E}(e^W|X^1,X^2,V)$, we write

$$w = h^W(V) + e^W,\ \mathbb{E}(e^W|V) = 0, w = h^W(V) + \rho^W(X,Z) + e^{W*},\ \mathbb{E}(e^{W*}|X^1,X^2,Z) = 0. \quad (31)$$

Note that $\rho^W(X,Z) = \mathbb{E}(w|X^1,X^2,Z) - \mathbb{E}(w|V)$. As in Section 4.2, we first state our results under generic assumptions then show that they are satisfied when the regressors have bounded support and their joint density goes to 0 at the boundary of the support.

Assumption 5.3.

1. *e_i^{W*} is i.i.d, V is continuously distributed, $\mathbb{E}((e_i^{W*})^2 \mid X^1,X^2,Z)$ is bounded, $\rho^W(.)$ is bounded over the support of (X^1,X^2,Z),*
2. *h^W is Lipschitz on $\mathcal{S}_V$,*
3. *There exists a $K\times K$ nonsingular matrix Γ_2, such that for $P^K(V) = \Gamma_2 p^K(V)$, then $\mathbb{E}(P^K(V)P^K(V)')$ has smallest eigenvalue bounded away from zero uniformly in K,*
4. *There exists γ_2 and π^K such that $\sup_{\mathcal{S}_V} |h^W(V) - p^K(V)'\pi_W^K| \leqslant CK^{-\gamma_2}$,*
5. *For $\sup_{\mathcal{S}_V} ||P^K(V)|| \leqslant b_1(K)$ and $\sup_{\mathcal{S}_V} ||\partial P^K(V)/\partial V|| \leqslant b_2(K)$, $\sqrt{K}\ b_2(K)\Delta_n \to_{n\to\infty} 0$ and $\sqrt{K/n}\, b_1(K) \to_{n\to\infty} 0$.*

Result 5.1. *Under Assumption 5.1 and 5.3,*

$$\int \left|\hat{h}^W(V) - h^W(V)\right|^2 dF(V) \ = \ O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2), \quad (32)$$

$$\sup_{V\in\mathcal{S}_V} |\hat{h}^W(V) - h^W(V)| \ = \ O_{\mathbb{P}}\left(b_1(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2}\right). \quad (33)$$

Proof. The proof follows the steps of the proof of Theorem 12 of Imbens and Newey (2009) (IN09 thereafter), but we adapt some of their claims to our model where $\mathbb{E}(e_i^W|X_i^1,X_i^2,Z_i) \neq 0$.

Define $p_i = p^K(V_i)$, $P = (p_1, .., p_n)$, $Q = PP'/n$, $\hat{p}_i = p^K(\hat{V}_i)$, $\hat{Q} = \hat{P}\hat{P}'/n$, $\rho_i^W = \rho^W(X_i^1,X_i^2,Z_i)$, as well as the vectors $e^W = (e_1^W, .., e_n^W)'$, $\rho^W = (\rho_1^W, .., \rho_n^W)'$ and $e^{W*} = (e_1^{W*}, .., e_n^{W*})'$. Note that $e^W = \rho^W + e^{W*}$. Because the series estimator is unchanged by a linear transformation of the basis of functions, we can assume that $p_i(V_i) = P^K(V_i)$. As argued in Newey (1997), we can assume without loss of generality that under Assumption 5.3 $\mathbb{E}(p_i^K p_i^{K\prime}) = I_K$. By construction, $\hat{V}_i \in \mathcal{S}_V$, and under Assumption 5.1, (29) holds. Therefore, as in Lemma S.5 of Imbens and Newey (2009), we have

$$||Q - I_K|| = O_{\mathbb{P}}(b_1(K)\sqrt{K/n}), \quad (34)$$

$$||P'e^W/n|| = O_{\mathbb{P}}(\sqrt{K/n}), \quad (35)$$

$$||\hat{P} - P||^2/n = O_{\mathbb{P}}(b_2(K)^2\Delta_n^2), \quad (36)$$

$$||\hat{Q} - Q|| = O_{\mathbb{P}}(b_2(K)^2\Delta_n^2 + \sqrt{K}b_2(K)\Delta_n). \quad (37)$$

Hence by Assumption 5.3 (5), $||\hat{Q} - I_K|| = o_{\mathbb{P}}(1)$ and as in Lemma S.6 of Imbens and Newey (2009), with probability going to 1, $\lambda_{\min}(\hat{Q}) \geqslant C$ and $\lambda_{\min}(Q) \geqslant C$.

We now show how the rate of convergence is impacted by the conditional mean dependence of e^W on (X, Z) by deriving the rate of $||\hat{\pi}^W - \pi_W^K||$, where we recall $\hat{\pi}^W = \hat{Q}^{-1}\hat{P}W$. We define $H^W = (h^W(V_1), .., h^W(V_n))'$, $\hat{H}^W = (h^W(\hat{V}_1), .., h^W(\hat{V}_n))'$, $\tilde{\pi}^W = \hat{Q}^{-1}\hat{P}\hat{H}^W/n$, $\bar{\pi}^W = \hat{Q}^{-1}\hat{P}H^W/n$. We decompose

$$||\hat{\pi}^W - \pi_W^K|| \leqslant \underbrace{||\hat{\pi}^W - \bar{\pi}^W||}_{(A)} + \underbrace{||\bar{\pi}^W - \tilde{\pi}^W||}_{(B)} + \underbrace{||\tilde{\pi}^W - \pi_W^K||}_{(C)}.$$

The first term can in turn be decomposed as

$$(A) = \hat{Q}^{-1}\hat{P}\,[\rho^W/n \,+\, e^{W*}/n].$$

Since (X_i, Z_i, e_i^{W*}) are i.i.d, we have $\mathbb{E}(e_i^{W*}|X_1, Z_1, .., X_n, Z_n) = 0$, $\mathbb{E}((e_i^{W*})^2|X_1, Z_1, .., X_n, Z_n) = \mathbb{E}((e_i^{W*})^2|X_i, Z_i) \leqslant C$, and $\mathbb{E}(e_i^{W*}e_j^{W*}|X_1, Z_1, .., X_n, Z_n) = 0$. This gives

$$\begin{aligned}\mathbb{E}(||\hat{Q}^{1/2}\,\hat{Q}^{-1}\hat{P}e^{W*}/n||^2|X_1, Z_1, .., X_n, Z_n) &= \operatorname{tr}(\hat{Q}^{-1/2}\hat{P}\,\mathbb{E}(e^{W*}e^{W*\prime}|X_1, Z_1, .., X_n, Z_n)\,\hat{P}'\hat{Q}^{-1/2})/n^2,\\ &\leqslant C\operatorname{tr}(\hat{P}'(PP')^{-1}\hat{P})/n \leqslant CK/n,\end{aligned}$$

where the last inequality holds by $\hat{P}'\hat{Q}^{-1}\hat{P}$ being an orthogonal projection matrix. This implies by M that $\hat{Q}^{1/2}\,\hat{Q}^{-1}\hat{P}e^{W*}/n = O_{\mathbb{P}}(\sqrt{K/n})$, and by $\lambda_{\min}(Q) \geqslant C$ w. p. a 1, that

$$\hat{Q}^{-1}\hat{P}e^{W*}/n = O_{\mathbb{P}}(\sqrt{K/n}).$$

This rate is the same as Lemma S.7 (i) of IN09, which holds since e^{W*} is by definition conditionally mean-independent of the regressors generating V. As for the second term appearing in (A), we write

$$\hat{Q}^{-1}\hat{P}\rho^W/n = \hat{Q}^{-1}P\rho^W/n + \hat{Q}^{-1}(P - \hat{P})\rho^W/n.$$

Since $\mathbb{E}(\rho_i^W|V_i) = 0$ and (ρ_i^W, V_i) is i.i.d, we know that as in (35), $||P\rho^W/n||^2 = O_{\mathbb{P}}(K/n)$. Therefore, by $\lambda_{\min}(Q) \geqslant C$ w. p. a 1, $||\hat{Q}^{-1}P\rho^W/n|| = O_{\mathbb{P}}(\sqrt{K/n})$.

Moreover, $(P - \hat{P})\rho^W/n = \frac{1}{n}\sum_{i=1}^n(\hat{p}_i^K - p_i^K)\rho_i^W$ and

$$\begin{aligned}\frac{1}{n}||\sum_{i=1}^n(\hat{p}_i^K - p_i^K)\rho_i^W|| &\leqslant \frac{1}{n}\sum_{i=1}^n||(\hat{p}_i^K - p_i^K)\rho_i^W|| \leqslant C\left(\frac{1}{n}\sum_{i=1}^n||(\hat{p}_i^K - p_i^K)||^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n|\rho_i^W|^2\right)^{1/2},\\ &\leqslant Cb_2(K)\left(\frac{1}{n}\sum_{i=1}^n||(\hat{V}_i - V_i)||^2\right)^{1/2}||\rho||_\infty \;=\; O_{\mathbb{P}}(b_2(K)\Delta_n).\end{aligned}$$

This implies by $\lambda_{\min}(Q) \geqslant C$ w. p. a 1 that $\hat{Q}^{-1}(P - \hat{P})\rho^W/n = O_{\mathbb{P}}(b_2(K)\Delta_n)$. This gives a convergence rate for (A), $||\hat{\pi}^W - \bar{\pi}^W||^2 = O_{\mathbb{P}}(K/n + K/n + b_2(K)^2\Delta_n^2) = O_{\mathbb{P}}(K/n + b_2(K)^2\Delta_n^2)$.

By Lemma S.7 (ii) and (iii) of IN09, $(B) = O_{\mathbb{P}}(\Delta_n)$ since $h^W(.)$ is Lipschitz, and $(C) = O_{\mathbb{P}}(K^{-\gamma_2})$ using Assumption 5.3 (4). This implies that

$$||\hat{\pi}^W - \pi_W^K||^2 = O_{\mathbb{P}}(K/n + b_2(K)^2\Delta_n^2 + \Delta_n^2 + K^{-2\gamma_2}) = O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + b_2(K)^2\Delta_n^2),$$

which differs from the rate $(K/n + K^{-2\gamma_2} + \Delta_n^2)$ obtained in IN09. Clearly, the structure of the proof showed that the extra term $b_2(K)\Delta_n$ comes from the correlation between e^W and $\hat{P} - P$, which is nonzero because $\hat{P}$ is constructed using the estimated $\hat{V}$ which themselves depend on the covariates (X^1, X^2, Z). Deriving a rate of convergence requires linearizing the term $\hat{P} - P$. If it were true that $\mathbb{E}(e_i^W | X_i^1, X_i^2, Z_i) = 0$, then $\mathbb{E}((\hat{P} - P)\nu) = 0$ would hold, which would yield a simpler convergence rate.

We can now conclude that

$$\begin{aligned}
\int \left|\hat{h}^W(V) - h^W(V)\right|^2 dF_V(V) &\leqslant \int \left|\hat{h}^W(V) - p^K(V)'\pi_W^K\right|^2 dF_V(V) + \int \left|p^K(V)'\pi_W^K - h^W(V)\right|^2 dF_V(V) \\
&\leqslant (\hat{\pi}^W - \pi_W^K)'[\int p^K(V)p^K(V)'dF_V(V)](\hat{\pi}^W - \pi_W^K) + CK^{-2\gamma_2} \\
&= O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2),
\end{aligned}$$

where the last line holds by the normalization $\mathbb{E}(Q) = I_K$. Moreover

$$\begin{aligned}
\sup_{V\in\mathcal{S}_V} |\hat{h}^W(V) - h^W(V)| &\leqslant \sup_{V\in\mathcal{S}_V} \left|\hat{h}^W(V) - p^K(V)'\pi_W^K\right| + O_{\mathbb{P}}(K^{-\gamma_2}) \\
&\leqslant O_{\mathbb{P}}\left(b_1(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2}\right).
\end{aligned}$$

□

As highlighted in the proof, the rate we obtained is different from the rate derived in NPV99 because of the absence of conditional mean independence condition. The following corollary, stating the rate of convergence of the first order partial derivative of the two step nonparametric estimator, will be used in the proof of asymptotic normality.

Corollary 5.1. *Under Assumption 5.1 and 5.3, and assuming moreover that* $\sup_{\mathcal{S}_V} |\partial h^W(V) - \pi_W^{K\prime}\partial p^K(V)| \leqslant CK^{-\gamma_2}$, *then* $\sup_{V\in\mathcal{S}_V} |\partial\hat{h}^W(V) - \partial h^W(V)| = O_{\mathbb{P}}\left(b_2(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2}\right)$.

As we mentioned earlier, Assumption 5.3 (4) is often shown to hold for regressors with bounded support and density bounded away from 0 on their support. For reasons stated earlier, we prefer avoiding any trimming on covariates. We therefore give a set of conditions allowing the density of the regressors to go to zero on the boundaries of the support. Recall that the support of V is $\mathcal{S}_V = \times_{d\leqslant d_2, t\leqslant T}[\underline{v}_{td}, \bar{v}_{td}]$.

Assumption 5.4.

1. e_i^{W*} *is i.i.d,* V *is continuously distributed,* $\mathbb{E}((e^{W*})^2 \mid X^1, X^2, Z)$ *is bounded,* $\rho^W(.)$ *is bounded over the support of* (X^1, X^2, Z)*,*
2. h^W *is Lipschitz on* $\mathcal{S}_V$*,*
3. $p^K(.)$ *is the power series basis, and forall* $V \in \mathcal{S}_V$*,* $f_V(V) \geqslant \Pi_{d \leqslant k_2, t \leqslant T} (v_{t,d} - \underline{v}_{td})^{\alpha_2} (\bar{v}_{td} - v_{t,d})^{\alpha_2}$*,*
4. *There exists* γ_2 *and* π_W^K *such that* $\sup_{\mathcal{S}_V} |h^W(V) - \pi_W^{K\prime} p^K(V)| \leqslant CK^{-\gamma_2}$*,*
5. $K^{\alpha_2+7/2}\Delta_n \to_{n\to\infty} 0$ *and* $K^{\alpha_2+3/2}/\sqrt{n} \to_{n\to\infty} 0$*.*

Result 5.2. *Under Assumptions 5.2 and 5.4 , (32) and (33) hold.*

As mentioned earlier, $b_1(K) = K^{\alpha_2+1}$ and $b_2(K) = K^{\alpha_2+3}$, therefore Assumption 5.4 (5) implies Assumption 5.3 (5), and together with Assumption 5.4 (3), implies Assumption 5.3 (3).

For these results to apply to our choice of $w_i = (M_i)_{s,t}$ and $w_i = (M_i \dot{y}_i)_t$ for $1 \leqslant s, t \leqslant T-1$, we adapt Assumption 5.4 to the model primitives.

Assumption 5.5.

1. V *is continuously distributed,* $\mathbb{E}(\dot{u}|X^1, X^2, Z)$ *and* $\mathrm{Var}(||\dot{u}|| \,|X^1, X^2, Z)$ *are bounded,*
2. $\mathcal{M}$ *and* k *are Lipschitz,*
3. $p^K(.)$ *is the power series basis, and for all* $V \in \mathcal{S}_V$*,* $f_V(V) \geqslant \Pi_{d \leqslant k_2, t \leqslant T} (v_{t,d} - \underline{v}_{td})^{\alpha_2} (\bar{v}_{td} - v_{t,d})^{\alpha_2}$*,*
4. *There exists* γ_2 *and, for* $1 \leqslant s, t \leqslant T-1$*,* $\pi_{M,st}^K$ *and* $\pi_{k,t}^K$ *such that* $\sup_{\mathcal{S}_V} |\mathcal{M}_{st}(V) - \pi_{M,st}^{K\prime} p^K(V)| \leqslant CK^{-\gamma_2}$ *and* $\sup_{\mathcal{S}_V} |k_t(V) - \pi_{k,t}^{K\prime} p^K(V)| \leqslant CK^{-\gamma_2}$*,*
5. $K^{\alpha_2+7/2}\Delta_n \to_{n\to\infty} 0$ *and* $K^{\alpha_2+3/2}/\sqrt{n} \to_{n\to\infty} 0$*.*

Under Assumptions 2.1, 5.2 and 5.5, the convergence rate of $\hat{\mathcal{M}}$ and $\hat{k}$ in sup norm and mean square norms are therefore given by (32) and (33).

5.3 Convergence rate for $\hat{g}$

We defined $\hat{g}(V) = \hat{\mathcal{M}}(V)^{-1} \hat{k}(V)$. The rate of convergence of $\hat{g}(.)$ will be obtained using continuity arguments. We will assume the following set of conditions.

Assumption 5.6. $\mathcal{M}$ *and* g *are continuous on* $\mathcal{S}_V$*,* $\mathcal{S}_V$ *is a compact set and* $\mathcal{M}(V)$ *is invertible for all values* $V \in \mathcal{S}_V$*.*

This implies that $k = \mathcal{M}\, g$ is continuous as well and that $||m||_\infty$, $||\mathcal{M}||_\infty$ and $||g||_\infty$ exist. Note that the continuity assumption somewhat overlaps with Assumption 5.3 (4) as the existence of a linear approximation relies on smoothness assumptions. Moreover, we assume that $\mathcal{M}(V)$ is

invertible for all values in the support while Assumption 2.3 requires the matrix to be invertible only $\mathbb{P}_V$ a.s. I prove the following results.

Result 5.3. *Under Assumption 5.6, the function $V \in \mathcal{S}_V \mapsto \lambda_{\min}(\mathcal{M}(V))$ is continuous.*

Proof. For all V, $\mathcal{M}(V)$ is symmetric. Using the Weilandt-Hoffman inequality for the 2-Schatten norm, for two values V and V',

$$\sum_{i=1}^{T-1} \left| \lambda_i(\mathcal{M}(V)) - \lambda_i(\mathcal{M}(V')) \right|^2 \leqslant ||\mathcal{M}(V) - \mathcal{M}(V')||_F^2,$$

where we index the eigenvalues $(\lambda_i)_{i=1}^{T-1}$ by increasing order. This implies

$$\left| \lambda_{\min}(\mathcal{M}(V)) - \lambda_{\min}(\mathcal{M}(V')) \right| \leqslant ||\mathcal{M}(V) - \mathcal{M}(V')||_F^{1/2}. \tag{38}$$

Since $\mathcal{M}(.)$ is a continuous function, this concludes the argument. □

The Lipschitz inequality (38) will be used in the proof for the convergence rates.

Result 5.4. *Under Assumption 5.6, there exists $c > 0$, such that for all $V \in \mathcal{S}_V$, $\lambda_{\min}(\mathcal{M}(V)) \geqslant c$.*

Proof. Under Assumption 5.6, $\mathcal{M}(V)$ is nonsingular for all $V \in \mathcal{S}_V$. This implies that for all $V \in \mathcal{S}_V$, $\lambda_{\min}(\mathcal{M}(V)) > 0$. Since $\mathcal{S}_V$ is a compact set, using Result 5.3, the function $V \mapsto \lambda_{\min}(\mathcal{M}(V))$ is continuous, therefore it has a minimum and reaches it. That minimum cannot be 0, hence $\exists\, c > 0,\ \forall\, V \in \mathcal{S}_V,\ \lambda_{\min}(\mathcal{M}(V)) \geqslant c$. □

Result 5.5. *Under Assumptions 5.2, 5.5 and 5.6, and assuming $b_1(K)^2(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2) \to_{n\to\infty} 0$,*

$$\begin{aligned} \int ||\hat{g}(V) - g(V)||^2 \, dF(V) &= O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2), \\ ||\hat{g}(V) - g(V)||_\infty &= O_{\mathbb{P}}(b_1(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2}). \end{aligned}$$

Proof. Under Assumptions 5.1 and 5.3, writing Γ_n^2 and γ_n respectively the mean square and sup norm rates of convergence, we know that $\int ||\hat{m}(V) - m(V)||^2 \, dF(V) = O_{\mathbb{P}}(\Gamma_n^2)$, $\sup_{V\in\mathcal{S}_V} ||\hat{m}(V) - m(V)|| = O_{\mathbb{P}}(\gamma_n)$, $\int \left|\left|\hat{\mathcal{M}}(V) - \mathcal{M}(V)\right|\right|_F^2 dF(V) = O_{\mathbb{P}}(\Gamma_n^2)$, and $\sup_{V\in\mathcal{S}_V} \left|\left|\hat{\mathcal{M}}(V) - \mathcal{M}(V)\right|\right|_F = O_{\mathbb{P}}(\gamma_n)$,

since the Frobenius norm and the Euclidean norm are square roots of the sum of squared elements, and this rate was obtained for each element of $\mathcal{M}(V)$ and $m(V)$.

We write $\mathbb{1}_n = \mathbb{1}\left(\min_{V\in\mathcal{S}_V} \lambda_{\min}(\hat{\mathcal{M}}(V)) > \frac{c}{2}\right)$. Using (38), we have

$$\lambda_{\min}(\hat{\mathcal{M}}(V)) > \lambda_{\min}(\mathcal{M}(V)) - \left|\left| ||\hat{\mathcal{M}}(V) - \mathcal{M}(V)||_F \right|\right|_\infty,$$

$$\Rightarrow \min_{V \in \mathcal{S}_V} \lambda_{\min}(\hat{\mathcal{M}}(V)) \;>\; c - \Big|\Big| \, ||\hat{\mathcal{M}}(V) - \mathcal{M}(V)||_F \Big|\Big|_\infty,$$

where the last implication uses Result 5.4. Hence $\mathbb{1}_n \geqslant \mathbb{1}\left(\Big|\Big| \, ||\hat{\mathcal{M}}(V) - \mathcal{M}(V)||_F \Big|\Big|_\infty \leqslant c/2\right)$. By Assumptions 5.1 and 5.3, $\gamma_n \to 0$ which implies $\mathbb{1}_n \;=\; 1$ w. p. a 1.

To obtain the sup norm rate, we write,

$$\begin{aligned}
\forall\, V \in \mathcal{S}_V,\; g(V) - \hat{g}(V) \;&=\; \mathcal{M}(V)^{-1}\, m(V) - \hat{\mathcal{M}}(V)^{-1} \hat{m}(V) \\
&=\; \mathcal{M}(V)^{-1} \left[\, \hat{\mathcal{M}}(V) - \mathcal{M}(V) \,\right] \hat{\mathcal{M}}(V)^{-1}\, m(V) \;+\; \hat{\mathcal{M}}(V)^{-1} \left[\, m(V) - \hat{m}(V) \,\right],
\end{aligned}$$

which gives, using T, the norm inequality and definition of induced norm,

$$\mathbb{1}_n ||\hat{g}(V) - g(V)||_\infty \;\leqslant\; \mathbb{1}_n \, c \, \frac{c}{2} \, \Big|\Big| \, ||\hat{\mathcal{M}}(V) - \mathcal{M}(V)||_F \, \Big|\Big|_\infty ||m||_\infty \;+\; \mathbb{1}_n \, \frac{c}{2} \, ||m - \hat{m}||_\infty.$$

This implies that $||\hat{g}(V) - g(V)||_\infty = O_{\mathbb{P}}(\gamma_n)$. To obtain the mean square error rate, we write

$$\begin{aligned}
&\mathbb{1}_n \int ||\hat{g}(V) - g(V)||^2 dF(V) \\
&\quad=\; \mathbb{1}_n \int \left|\left| \mathcal{M}(V)^{-1} \left[\, \hat{\mathcal{M}}(V) - \mathcal{M}(V) \,\right] \hat{\mathcal{M}}(V)^{-1}\, m(V) \;+\; \hat{\mathcal{M}}(V)^{-1} \left[\, m(V) - \hat{m}(V) \,\right] \right|\right|^2 dF(V) \\
&\quad\leqslant\; \mathbb{1}_n \; c \; \frac{c}{2} \; ||m||_\infty \int \left|\left| \hat{\mathcal{M}}(V) - \mathcal{M}(V) \right|\right|_F^2 dF(V) \;+\; \mathbb{1}_n \, \frac{c}{2} \int ||\hat{m}(V) - m(V)||^2 \, dF(V),
\end{aligned}$$

which implies $\int ||\hat{g}(V) - g(V)||^2 \, dF(V) \;=\; O_{\mathbb{P}}(\Gamma_n^2)$. □

6 Consistency of $\hat{\mu}$

Recall that, writing $\delta_i = \mathbb{1}(\det(\dot{X}_i' \dot{X}_i) > \delta_0)$ and $Q_i^\delta = \delta_i Q_i$, we defined the estimator for $\mathbb{E}(\mu|\delta)$ to be

$$\hat{\mu} \;=\; \frac{\sum_{i=1}^n Q_i^\delta \, [\dot{y}_i - \, \hat{g}(\hat{V}_i)]}{\sum_{i=1}^n \delta_i}.$$

Assumption 6.1. *Assume* $\mathbb{E}(||Q^\delta||) < \infty$ *and* $\mathbb{E}(||Q^\delta \, \dot{y}||) < \infty$.

Result 6.1. *Suppose Assumptions 5.1, 5.3, 5.6, and 6.1 hold, with* $\gamma_n \to_{n\to\infty} 0$ *and* $a_1(L)\Delta_n \to_{n\to\infty} 0$. *If additionally* g *is continuously differentiable on* $\mathcal{S}_V$, *then* $\hat{\mu} \to_{\mathbb{P}} \mathbb{E}(\mu|\delta)$.

Proof. To prove consistency, we need to show that $\frac{1}{n}\sum_{i=1}^n Q_i^\delta \, \hat{g}(\hat{V}_i) \to_{\mathbb{P}} \mathbb{E}(Qg(V)\delta)$. Then by the LLN, $\frac{1}{n}\sum_{i=1}^n \delta_i \to_{\mathbb{P}} \mathbb{P}(\det(\dot{X}_i' \dot{X}_i) > \delta)$, and by Assumption 6.1 and the LLN, $\frac{1}{n}\sum_{i=1}^n Q_i^\delta \, \dot{y}_i \to_{\mathbb{P}} \mathbb{E}(Q \, \dot{y} \, \delta)$, consistency would follow from Equation (12).

We decompose

$$\frac{1}{n}\sum_{i=1}^n Q_i^\delta \, \hat{g}(\hat{V}_i) \;-\; \mathbb{E}(Qg(V)\delta) \;=\; \frac{1}{n}\sum_{i=1}^n Q_i^\delta \, [\hat{g}(\hat{V}_i) \;-\; g(V_i)] \;+\; \frac{1}{n}\sum_{i=1}^n Q_i^\delta \, g(V_i) \;-\; \mathbb{E}(Qg(V)\delta)$$

$$:= A_n + B_n.$$

We have

$$\begin{aligned}
||A_n|| &= \Big|\Big| \frac{1}{n}\sum_{i=1}^{n} Q_i^\delta \left[\hat{g}(\hat{V}_i) - g(\hat{V}_i) \right] + \frac{1}{n}\sum_{i=1}^{n} Q_i^\delta \left[g(\hat{V}_i) - g(V_i) \right] \Big|\Big| \\
&\leqslant ||\hat{g} - g||_\infty \frac{1}{n}\sum_{i=1}^{n} ||Q_i^\delta|| + C \max_i ||\hat{V}_i - V_i|| \frac{1}{n}\sum_{i=1}^{n} ||Q_i^\delta||_2 \\
&= O_{\mathbb{P}}(\gamma_n + a_1(L)\Delta_n) \frac{1}{n}\sum_{i=1}^{n} ||Q_i^\delta||_2,
\end{aligned}$$

where the first term in the inequality follows from $\hat{V}_i \in \mathcal{S}_V$ by design, and the second sum term follows from g being continuously differentiable on a compact set, hence Lipschitz continuous on this set. The last equality follows from Equation (30) and Result (5.5). By Assumption 6.1, $\gamma_n \to 0$ and $a_1(L)\Delta_n \to 0$ as n goes to infinity, and we obtain $||A_n|| = o_{\mathbb{P}}(1)$. □

7 Asymptotic normality

In this section we derive the asymptotic normality of $\hat{\mu}$. We carry out the analysis in several steps. First, we modify the trimming function. We then express the dependence of our estimator on a sample moment that depends on the nonparametric two-step sieve estimators. We explain how to linearize it. We turn to the general case of a linear functional of nonparametric two-step sieve estimator, derive the asymptotic distribution, then apply the obtained results to the linearized part of our estimator. Finally we prove that the linearization is valid, and derive asymptotic normality of $\hat{\mu}$.

7.1 Trimming

We will require the supports of V and $(\xi_t)_{t\leqslant T}$ to be bounded. Before getting into the details of the proof of asymptotic normality, we introduce a slight modification of the estimator $\hat{V}_i$ of the control variables V.

Recall that we defined $\hat{V}_i = \tau(\tilde{V}_i)$, where $\tilde{V}_i = (\tilde{v}_{it})_{t\leqslant T}$ is the vector of residual from the sieve regression of x_{it}^2 on $\xi_{it} = (x_{it}^1, z_{it})$ and τ projects onto $\mathcal{S}_V = \times_{d\leqslant d_2,\, t\leqslant T} [\underline{v}_{td}; \bar{v}_{td}]$. The proof of asymptotic normality will use the smoothness properties of τ and will require it to be twice differentiable, which is not the case when τ is as defined in Section 4.1. We change the definition of τ. Define $\varsigma > 0$, and $\tau_\varsigma : x \in \mathbb{R} \mapsto \varsigma(e^{-x^2/(2\varsigma^2) + x/\varsigma} - 1)$. Note that $\lim_{x\to-\infty} \tau_\varsigma(x) = -\varsigma$, $\lim_{x\to+\infty} \tau_\varsigma(x) = -\varsigma$ and we also have $\tau_\varsigma(0) = 0$, $\tau_\varsigma'(0) = 1$ and $\tau_\varsigma''(0) = 0$. For $V \in \mathbb{R}^{Td_2}$, the

$(d_2(t-1)+d)^{\text{th}}$ component of $\tau(V)$ is given by

$$\tau(V)_{(t-1)k_2+d} = \begin{cases} v_{t,d}, & \text{if } v_{t,d} \in [\underline{v}_{td}; \bar{v}_{td}], \\ \underline{v}_{td} + \tau_\varsigma(v_{t,d} - \underline{v}_{td}), & \text{if } v_{t,d} \leqslant \underline{v}_{td}, \\ \bar{v}_{td} - \tau_\varsigma(\bar{v}_{td} - v_{t,d}), & \text{if } v_{t,d} \geqslant \bar{v}_{td}, \end{cases}$$

and define as before $\hat{V}_i = \tau(\tilde{V}_i)$. The support of τ is $\mathbb{R}^{Td_2}$ and we now have $\hat{V}_i \in \mathcal{S}_V^\varsigma = \times_{d \leqslant d_2,\, t \leqslant T} [\underline{v}_{td} - \varsigma; \bar{v}_{td} + \varsigma]$. We will refer to $\mathcal{S}_V^\varsigma$ as the "extended support".

Each component of τ is a twice differentiable function of V, implying that τ itself is twice continuously differentiable. Moreover for all $V \in \mathcal{S}_V$, $\partial\tau/\partial V = I_{Tk_2}$ which will imply that the derivative of the function m composed with τ evaluated at V, $m(\tau(V))$, is equal to the derivative of $m(V)$ whenever $V \in \mathcal{S}_V$. On the extended support, that is for all $V \in \mathcal{S}_V^\varsigma$, $|\partial\tau/\partial V| \leqslant C$ and $|\partial^2\tau/\partial V^2| \leqslant C$ for some constant C.

It will also be convenient to use extensions of the various regression functions used at different places in our proofs. For a function $m : \mathcal{S}_V \to \mathbb{R}^k$ (for any given k) such that m is twice continuously differentiable on $\mathcal{S}_V$, we define $m^\varsigma : \mathcal{S}_V^\varsigma \to \mathbb{R}^k$ an extension of m, twice continuously differentiable. That is, for all V in $\mathcal{S}_V$, $m^\varsigma(V) = m(V)$, and m^ς must be twice continuously differentiable on the extended support $\mathcal{S}_V^\varsigma$. Note that if there exists a sequence of functions $(m_n)_{n\in\mathbb{N}}$ converging uniformly to m^ς on the extended support $\mathcal{S}_V^\varsigma$, the sequence of restrictions of $(m_n)_{n\in\mathbb{N}}$ on $\mathcal{S}_V$ converges uniformly to m. We previously used, for g a function of the variable V, the norm $|g|_d = \max_{|l|\leqslant d} \sup_{V\in\mathcal{S}_V} ||\partial^l g(.)||$. A corresponding norm for the extended functions will change the supremum to a supremum over the extended support, i.e., $|g|_d^\varsigma = \max_{|l|\leqslant d} \sup_{V\in\mathcal{S}_V^\varsigma} ||\partial^l g(.)||$.

Note that as was the case with our previous definition of τ, $||\hat{V}_i - V_i|| \leqslant ||\tilde{V}_i - V_i||$. This guaranteees that our results on the sup-norm convergence rates of the nonparametric two-step estimators $\hat{\mathcal{M}}$ and $\hat{k}$, and their derivatives, remain valid. This simply requires some changes provided some changes are made to the definition of the vector of basis functions $p^K(.)$ and to the approximation condition (4) of Assumption 5.3. First, $p^K(.)$ is defined on the extended support, and the bounds $b_1(K)$ and $b_2(K)$ are also defined as bounds on the sup norm over the extended support. Second, the approximation condition must be imposed on the extended functions $\mathcal{M}^\varsigma$ and k^ς. Under these modified conditions, the rate of convergence of $\hat{g}$ is unchanged and consistency of $\hat{\mu}$ holds.

7.2 Linearization

We study the asymptotic normality of $\sqrt{n}(\hat{\mu} - \mathbb{E}(\mu))$, where $\hat{\mu}$ is a sample average of a function of the nonparametric estimate $\hat{g}$ evaluated at the generated values $\hat{V}$. It is a semiparametric estimator where the infinite dimensional nuisance parameter is estimated with generated covariates. General

conditions for asymptotic normality of semiparametric estimators are given for instance in Newey (1994b), Chen, Linton, and Van Keilegom (2003), Ai and Chen (2003) and Ichimura and Lee (2010). Many other references exist. However, the conditions given in these papers are "high-level" conditions and cannot be directly applied to estimators using generated covariates. Mammen et al. (2016) studies the asymptotic normality of a general class of semiparametric GMM estimators with generated covariates. Our estimator of the APE $\hat{\mu}$ belongs to this class of estimators, although of a simpler form since it has a closed-form expression. The infinite dimensional nuisance parameter in Mammen et al. (2016) is a conditional expectation which they estimate with a local polynomial estimator and they do not specify an estimator for the generated covariates. To the best of our knowledge, there are no such results for two-step series estimators. We will therefore develop the results that are needed for asymptotic normality of our specific estimator.

The estimator of the average effect $\mathbb{E}(\mu|\delta)$ was defined as

$$\hat{\mu} \;=\; \frac{\sum_{i=1}^n \delta_i Q_i \left[\dot{y}_i - \hat{g}(\hat{V}_i)\right]}{\sum_{i=1}^n \delta_i} \;=\; \frac{\sum_{i=1}^n Q_i^\delta \left[\dot{y}_i - \hat{g}(\hat{V}_i)\right]}{\sum_{i=1}^n \delta_i} = \frac{1}{\sum_{i=1}^n \delta_i/n}\,\hat{\mu}^\delta,$$

where $\hat{\mu}^\delta = \sum_{i=1}^n Q_i^\delta \left[\dot{y}_i - \hat{g}(\hat{V}_i)\right]/n$. To obtain the asymptotic distribution of $\hat{\mu} - \mathbb{E}(\mu|\delta)$, we will first study $\hat{\mu}^\delta - \mathbb{E}(\mu_i\delta_i)$. We write $\mathcal{G} = ((b_t)_{t\leqslant T}, k, \mathcal{M})$ for a vector of generic functions with $b_t : \mathcal{S}_{\xi t} \mapsto \mathbb{R}^{d_2}$, $k : \mathcal{S}_V^\varsigma \mapsto \mathbb{R}^{T-1}$ and $\mathcal{M} : \mathcal{S}_V^\varsigma \mapsto \mathcal{M}_{T-1}(\mathbb{R})$. For clarity we choose to write $\mathcal{G}_0 = ((b_{0t})_{t\leqslant T}, k_0, \mathcal{M}_0)$, for the **true values** of these functions, that is, for the nonparametric primitives of the model. Note that the functions we consider here are functions on the extended support. We dropped the exponant ς and will display it to avoid confusion whenever necessary. We decompose

$$\begin{aligned}
\sqrt{n}(\hat{\mu}^\delta - \mathbb{E}(\mu_i\delta_i)) \;&=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_i^\delta \left[\dot{y}_i - \;\hat{g}(\hat{V}_i)\right] - \mathbb{E}(\mu_i\delta_i),\\
&= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n [\delta_i\mu_i - \mathbb{E}(\mu_i\delta_i)] + \sum_{i=1}^n Q_i^\delta \dot{u}_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n [Q_i^\delta g(V_i) - \mathbb{E}(Q^\delta g(V))] - \sum_{i=1}^n [Q_i^\delta \hat{g}(\hat{V}_i) - \mathbb{E}(Q^\delta g(V))]\right],\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \big[\delta_i\mu_i - \mathbb{E}(\mu_i\delta_i)\big] + \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_i^\delta \dot{u}_i \;-\; \sqrt{n}\big[\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0)\big], \qquad (39)
\end{aligned}$$

where we define

$$\begin{aligned}
\chi(W_i, \mathcal{G}) \;&=\; Q_i^\delta\, \mathcal{M}\left(\tau\left[(x_{it}^2 - b_t(\xi_{it}))_{t\leqslant T}\right]\right)^{-1}\, k\left(\tau\left[(x_{it}^2 - b_t(\xi_{it}))_{t\leqslant T}\right]\right),\\
\mathcal{X}(\mathcal{G}) \;&=\; \mathbb{E}(\chi(W_i, \mathcal{G})) - \mathbb{E}(\chi(W_i, \mathcal{G}_0)),\\
\mathcal{X}_n(\mathcal{G}) \;&=\; \frac{1}{n}\sum_{i=1}^n [\chi(W_i, \mathcal{G}) - \mathbb{E}(\chi(W_i, \mathcal{G}_0))],
\end{aligned}$$

and where τ is as defined in Section 7.1. The function τ ensures that the argument of $\mathcal{M}$ and k lies in $\mathcal{S}_V^\varsigma$. We recall that $W_i = (X_i, u_i, V_i, Z_i, \mu_i, \alpha_i)$ stands for the whole vector of primitive variables. Note that $\mathcal{X}(\mathcal{G}_0) = 0$. We write the variables as column vectors, e.g $V_i = (V_{i1}', ..., V_{iT}')$.

The decomposition as well as our choice of arguments in $\mathcal{X}$ make explicit the dependence of our estimator on the functions $(b_t)_{t\leqslant T}$. Indeed, the generated covariates are defined as the vector of $v_{it} = x_{it}^2 - b_t(\xi_{it})$, and $(b_t)_{t\leqslant T}$ are nuisance parameters. The use of generated covariates in place of the true value of the variables has a twofold impact on semiparametric estimators such as $\hat{\mu}$. First the nuisance parameter $\mathcal{M}$ and k are estimated using the generated values. Second the estimators $\hat{\mathcal{M}}$ and $\hat{k}$ are evaluated at the generated values when plugged in in the sample average that defines $\hat{\mu}$. The dependence of $\mathcal{X}$ on $(b_t)_{t\leqslant T}$ highlights the latter aspect.

The two first terms in Equation (39) are normalized sums of i.i.d random variables. Their asymptotic normality can be established by a standard CLT argument. We focus on the last term, $\sqrt{n}\big[\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0)\big]$, which has a standard form except for its dependence on a composition of the infinite dimensional nuisance parameters. Arguments yielding asymptotic normality typically require a fast enough rate of convergence of the estimators of the nuisance parameters, a stochastic equicontinuity condition that imposes restrictions on the class of functions the parameters can lie in, and an asymptotic normality condition for linear functionals of the nonparametric estimates of the nuisance parameters. Specifically, we define a class of continuous functions $\mathcal{H}$ endowed with a pseudometric $||.||_{\mathcal{H}}$ such that $\mathcal{G} \in \mathcal{H}$. As we just described, we require the following set of conditions for our asymptotic analysis. Define $\mathcal{H}_\delta = \{\mathcal{G} \in \mathcal{H} : ||\mathcal{G} - \mathcal{G}_0|| \leqslant \delta\}$.

Assumption 7.1.

1. *For all* $\delta_n = o(1)$, $\sup_{||\mathcal{G}-\mathcal{G}_0||_{\mathcal{H}}\leqslant\delta_n} ||\mathcal{X}_n(\mathcal{G}) - \mathcal{X}(\mathcal{G}) - \mathcal{X}_n(\mathcal{G}^0)|| = o_{\mathbb{P}}(n^{-1/2})$.
2. *The pathwise derivative of* $\mathcal{X}$ *at* $\mathcal{G}_0$ *evaluated at* $\mathcal{G}-\mathcal{G}_0$, $\mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G}-\mathcal{G}_0]$, *exists in all directions* $[\mathcal{G} - \mathcal{G}_0]$, *and for all* $\mathcal{G} \in \mathcal{H}_{\delta_n}$ *with* $\delta_n = o(1)$, $||\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G} - \mathcal{G}_0]\,|| \leqslant c||\mathcal{G} - \mathcal{G}_0||_{\mathcal{H}}^2$, *for some constant* $c \geqslant 0$,
3. $||\hat{\mathcal{G}} - \mathcal{G}_0||_{\mathcal{H}} = o_{\mathbb{P}}(n^{-1/4})$.

Result 7.1. *Under Assumption 7.1,*

$$\sqrt{n}\big[\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0)\big] \;=\; \sqrt{n}\,\mathcal{X}^{(G)}(\mathcal{G}_0)[\hat{\mathcal{G}} - \mathcal{G}_0] + o_{\mathbb{P}}(1).$$

Proof. This proof is a special case of Theorem 2 in Chen et al. (2003). By Assumption 7.1 (3), there exists $\delta_n = o(1)$ such that $\mathbb{P}(||\hat{\mathcal{G}}-\mathcal{G}_0|| > \delta_n) \to 0$. Take $\mathbb{1}_n = \mathbb{1}(||\hat{\mathcal{G}}-\mathcal{G}_0|| \leqslant \delta_n)$. $\mathbb{1}_n||\mathcal{X}_n(\hat{\mathcal{G}})-\mathcal{X}_n(\mathcal{G}_0)-\mathcal{X}^{(G)}(\mathcal{G}_0)[\hat{\mathcal{G}} - \mathcal{G}_0]|| \leqslant \mathbb{1}_n||\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}(\mathcal{G}) - \mathcal{X}_n(\mathcal{G}_0)|| + \mathbb{1}_n||\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\hat{\mathcal{G}} - \mathcal{G}_0]|| = o_{\mathbb{P}}(n^{-1/2})$ by Assumption 7.1. Hence $||\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\hat{\mathcal{G}} - \mathcal{G}_0]|| = o_{\mathbb{P}}(n^{-1/2})$. □

The definition of $\mathcal{H}$ and in particular of $||.||_{\mathcal{H}}$ is not straightforward here. The choice of $\mathcal{H}$ will be driven by the stochastic equicontinuity condition, that is, Condition (1) of Assumption 7.1, following Chen et al. (2003) and the choice of $||.||_{\mathcal{H}}$ will be driven by Condition (2). We examine this in more details in Section 7.4.2.

Under Assumption 7.1, the asymptotic distribution of $\hat{\mu}$ depends on the asymptotic behavior of $\sqrt{n}\,\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0]$, where we write $\mathcal{X}_0^{(G)}$ for the pathwise derivative of $\mathcal{X}$ at $\mathcal{G}_0$. It is a linear functional of a nonparametric estimator, an object that has been widely studied (see, e.g, Newey (1994a) for kernel estimators and Newey (1997) for series estimators). However the pathwise derivative is evaluated at the constructed two-step nonparametric estimators and its asymptotic distribution cannot be obtained directly. Hahn et al. (2018) derives asymptotic normality results for nonlinear functionals of two-step nonparametric sieves estimators when the sieve estimators are from a general class of nonlinear sieve regression estimators. They however do not specify a formula for the asymptotic variance, arguing that it might not exist. This is not an issue in our case and we derive using a different type of proof the asymptotic normality and asymptotic variance of our linear sieve estimators.

The structure of our asymptotic analysis is as follows: we first derive the asymptotic distribution of $\sqrt{n}\,\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0]$ by studying the general case of a linear functional of the two-step sieve estimator of a nonparametric regression function and obtain its asymptotic variance. We then specify our choice of $\mathcal{H}$ and $||.||_{\mathcal{H}}$, show that Assumption 7.1 holds for this choice. Using Result 7.1, we obtain the asymptotic distribution of the standardized $\hat{\mu}$.

We therefore focus on the linearized term. The pathwise derivative applied to the estimators can be decomposed as the sum of $T+2$ partial pathwise derivatives applied to each nonparametrically estimated function. We define $\chi_0^{(k)}(W_i)[\tilde{k}]$, $(\chi_0^{(bt)}(W_i)[\tilde{b}_t])_{t\leqslant T}$ and $\chi_0^{(M)}(W_i)[\tilde{\mathcal{M}}]$) to be the partial pathwise derivatives of χ respectively with respect to k, b_t and $\mathcal{M}$ at the true value $\mathcal{G}_0$, evaluated respectively at $\tilde{k}$, $\tilde{b}_t$ and $\tilde{\mathcal{M}}$. We have

$$\begin{aligned}
\chi_0^{(k)}(W_i)[\tilde{k}] &:= \chi^{(k)}(W_i, \mathcal{G}_0)[\tilde{k}] \;=\; Q_i^{\delta}\mathcal{M}_0(V_i)^{-1}\tilde{k}(V_i), \\
\chi_0^{(bt)}(W_i)[\tilde{b}_t] &:= \chi^{(2t)}(W_i, \mathcal{G}_0)[\tilde{b}_t] \;=\; -Q_i^{\delta}\frac{\partial g}{\partial v_t}(V_i)\tilde{b}_t(\xi_{it}) \\
\chi_0^{(M)}(W_i)[\tilde{\mathcal{M}}] &:= \chi^{(M)}(W_i, \mathcal{G}_0)[\tilde{\mathcal{M}}] \;=\; -Q_i^{\delta}\mathcal{M}_0(V_i)^{-1}\,\tilde{\mathcal{M}}(V_i)\,\mathcal{M}_0(V_i)^{-1}k_0(V_i) \\
&\;=\; -Q_i^{\delta}\mathcal{M}_0(V_i)^{-1}\,\tilde{\mathcal{M}}(V_i)g(V_i) = \; -[g(V_i)' \otimes (Q_i^{\delta}\mathcal{M}_0(V_i)^{-1})]\;\mathrm{Vec}(\tilde{\mathcal{M}}(V_i)),
\end{aligned}$$

where v_t denotes the tth component of V, and where $\frac{\partial g}{\partial V}(V_i)$ is a Jacobian matrix of size $(T-1)\times Td_2$. Note that the function τ does not appear in the above formula, nor does any of its partial order derivatives. This is because when evaluated at the true value of V, by design τ simplifies to the identity function on $\mathcal{S}_V$ while its Jacobian simplifies to the identity matrix.

We define $\mathcal{X}_0^{(k)}[\tilde{k}]$ the partial pathwise derivative of $\mathcal{X}$ with respect to k at $\mathcal{G}_0$ and evaluated at $\tilde{k}$, and similarly $\mathcal{X}_0^{(M)}[\tilde{\mathcal{M}}]$ and $(\mathcal{X}_0^{(bt)}[\tilde{b}_t])_{t\leqslant T}$. Assuming one can interchange expectation and differentiation, we follow Mammen et al (2016) and write

$$\mathcal{X}_0^{(G)}[\mathcal{G} - \mathcal{G}_0] \;= \mathcal{X}_0^{(k)}[k - k_0] + \mathcal{X}_0^{(M)}[\mathcal{M} - \mathcal{M}_0] + \sum_{t=1}^{T}\mathcal{X}_0^{(bt)}[b_t - b_{0,t}],$$

$$= \int_V \left[\lambda_{\mathcal{M}}(v) \;\; \text{Vec}\left((\mathcal{M}-\mathcal{M}_0)(v)\right)\right] dF_V(v) + \int_V \left[\lambda_k(v)\,(k-k_0)(v)\right] dF_V(v)$$
$$+ \quad \sum_{t=1}^{T} \int_{\xi_t} \lambda_{bt}(\xi_t)\,(b_t - b_{0,t})(\xi_t) dF_{\xi_t}(\xi_t), \tag{40}$$

where the functions $\lambda_.$ are defined using the partial pathwise derivatives as

$$\begin{aligned}
\lambda_{\mathcal{M}}(v) &= -\mathbb{E}\left(g(V_i)' \otimes (Q_i^\delta \mathcal{M}_0(V_i)^{-1}) \,|\, V_i = v\right) = -g(v)' \otimes [\mathbb{E}(Q_i^\delta | V_i = v)\mathcal{M}_0(v)^{-1}], \\
\lambda_k(v) &= \mathbb{E}(Q_i^\delta \mathcal{M}_0(V_i)^{-1} \,|\, V_i = v) = \mathbb{E}(Q_i^\delta \,|\, V_i = v)\mathcal{M}_0(v)^{-1}, \\
\lambda_{bt}(\xi_t) &= -\mathbb{E}\left(Q_i^\delta \frac{\partial g(V_i)}{\partial v_t} \,|\, \xi_{it} = \xi_t\right).
\end{aligned}$$

We point out the fact that even if $\mathcal{X}_0^{(k)}$ and $\mathcal{X}_0^{(M)}$ are evaluated at functions defined on the extended support $\mathcal{S}_V^\varsigma$, they are unaffected by the values of these functions when evaluated at $V \notin \mathcal{S}_V$. Therefore it will not be necessary to consider extended functions. The following section studies the general case of a linear function applied to a (non extended) nonparametric two-step sieve estimator.

7.3 Linear application of a nonparametric two-step sieve estimator

We consider the model (31) described in Section 5.2. The object of interest in this section is the value of a linear function $a(.)$ evaluated at h^W where $h^W(.) = \mathbb{E}(W|V = .)$. We use the nonparametric two-step sieve estimator $\hat{h}^W$. The estimator of $a(h^W)$ will be $a(\hat{h}^W)$, and the purpose of this section is to write, under general conditions on the random variables W, V, e^{W*}, the functions h^W and ρ^W, the term $\sqrt{n}(a(h^W) - a(\hat{h}^W))$ as $\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{i,n}^W + o_{\mathbb{P}}(1)$. Treating this general case first will allow us to apply the derived results to $\mathcal{X}_0^{(k)}[\hat{k} - k_0]$ and $\mathcal{X}_0^{(M)}[\hat{\mathcal{M}} - \mathcal{M}_0]$.

To obtain our asymptotic normality result, we will use Lemma 2 of Newey et al. (1999). However as mentioned earlier, an essential orthogonal condition they assumed, namely the independence of the conditional mean of $w - \mathbb{E}(w|V)$ on (X, Z), does not hold in our model. For this reason we obtain an extra term in $s_{i,n}^W$, which depends on ρ^W. Note that V is composed of T generated covariates coming from T different cross-section regressions. We treat the case where $w \in \mathbb{R}$, $V \in \mathbb{R}^{Td_2}$, $a(h) \in \mathbb{R}^{d_a}$. We define $b_{it} = b_t(\xi_{it})$, $b_t = (b_{1t}, ..., b_{nt})$, the vector $h_\varsigma^W = (h^W(V_1), .., h^W(V_n)) = (h_\varsigma^W(V_1, .., h_\varsigma^W(V_n))$, for $h_\varsigma^W(.)$ the functional extension of $h^W(.)$, and the vector $\hat{h}_\varsigma^W = (h_\varsigma^W(\hat{V}_1), .., h_\varsigma^W(\hat{V}_n))$. Define $\rho_i^W = \rho^W(X_i, Z_i)$ and $\vec{v}_t = (v_{1t}, .., v_{nt})$. Finally we define the following matrices,

$$\begin{aligned}
\overline{H}_t^W &= \frac{1}{n}\sum_{i=1}^n \hat{p}_i \left(\partial h_\varsigma^W(V_i)/\partial v_t \otimes r_{it}'\right) & A &= (a(p_{1K}), .., a(p_{KK})), \\
&= \frac{1}{n}\sum_{i=1}^n \hat{p}_i \left(\partial h^W(V_i)/\partial v_t \otimes r_{it}'\right), & \overline{\mathrm{d}P}_t^W &= \frac{1}{n}\sum_{i=1}^n \rho_i^W \left(\frac{\partial p^K(V_i)}{\partial v_t} \otimes r_{it}'\right),
\end{aligned}$$

$$H_t^W = \mathbb{E}[p_i \left(\partial h_\varsigma^W(V_i)/\partial v_t \otimes r_{it}'\right)] \qquad \mathrm{d}P_t^W = \mathbb{E}(\rho_i^W \left(\frac{\partial p^K(V_i)}{\partial v_t} \otimes r_{it}'\right)),$$
$$= \mathbb{E}[p_i \left(\partial h^W(V_i)/\partial v_t \otimes r_{it}'\right)],$$

where v_t is the t^{th} generated covariate, and the equalities on the first two matrices hold by definition of h^ς as a convenient functional extension of h. We point out that this section, and our asymptotic normality analysis in general, will use series estimators without specifying the basis of approximating functions. We showed consistency under two sets of conditions, one using a general sieve basis, and one using power series and allowing the density of the regressors to go to zero on the boundaries of their support. Unlike the latter assumptions however, the assumptions we impose for asymptotic normality use rates on the sup norm of the functions as well as of their derivatives when the suprema are defined over the extended support. This rules out a direct application of the results allowing the density of the regressors to go to zero on the boundaries of their support. Indeed these results use a linear transformation of the power series and provide rates on the bounds of the norms of the vector of functions which are valid on the support, not on the extended support. Computing rates on the extended support is beyond the scope of this paper. We therefore remain silent on the choice of the basis. Recall that by $||\tau(v_1) - \tau(v_2)|| \leqslant ||v_1 - v_2||$, under Assumption 5.1 we have $\frac{1}{n}\sum_{i=1}^n ||v_i - \hat{v}_i||^2 = O_{\mathbb{P}}\left(L/n + L^{-2\gamma_1}\right) = O_{\mathbb{P}}(\Delta_n^2)$.

Assumption 7.2.

1. *The data W_i is i.i.d,*
2. $||a(g)|| \leqslant C|g|_0$,
3. *h^W is twice continuously differentiable and the first and second derivatives are bounded, ρ^W is bounded,*
4. *There exists γ_1 and β_t^L such that for all $t \leqslant T$, $\sup_{\mathcal{S}_{\xi_t}} ||b_t(\xi_t) - \beta_t^{L\prime} r^L(x_t^1, z_t)|| \leqslant CL^{-\gamma_1}$. There exists γ_2, π_W^K such that $\sup_{\mathcal{S}_V^\varsigma} ||h_\varsigma^W(v) - p^K(v)'\pi_W^K|| \leqslant CK^{-\gamma_2}$,*
5. *For all $t \leqslant T$, there exists Γ_{1t}, a $L \times L$ nonsingular matrix such that for $R_t^L(\xi_t) = \Gamma_{1t} r_t^L(\xi_t)$, $\mathbb{E}(R_t^L(\xi_t)R_t^L(\xi_t)')$ has smallest eigenvalue bounded away from zero uniformly in L. Similarly, there exists Γ_2, a $K \times K$ nonsingular matrix such that for $P^K(V) = \Gamma_2 p^K(V)$, $\mathbb{E}(P^K(V)P^K(V)')$ has smallest eigenvalue bounded away from zero uniformly in K,*
6. *$||A||$ is bounded,*
7. *For $|R_t^L(\xi_t)|_0 \leqslant a_1(L)$, $|P^K(V)|_0^\varsigma \leqslant b_1(K)$, $|P^K(V)|_1^\varsigma \leqslant b_2(K)$, $|P^K(V)|_2^\varsigma \leqslant b_3(K)$, we have $\sqrt{n}K^{-\gamma_2} = o(1)$, $\max(\sqrt{K}, \sqrt{L}b_2(K))\, a_1(L)\sqrt{L/n} = o(1)$, $b_2(K)[\sqrt{L/n} + L^{-\gamma_1}][K + L] = o(1)$,*
   $b_2(K)\sqrt{n}L^{-\gamma_1} = o(1)$, $b_3(K)[L/\sqrt{n} + \sqrt{n}L^{-2\gamma_1}] = o(1)$, $b_2(K)^2\sqrt{K}[\sqrt{L/n} + L^{-\gamma_1}] = o(1)$,
8. *For all ξ_t, $\mathbb{E}(||v_{it}||^2|\xi_{it} = \xi_t) \leqslant C$ and for all (X, Z), $\mathrm{Var}(e_i^{W*}|X_i = X, Z_i = Z) \leqslant C$.*

As stated above, we do not specify the sieve basis but for simplification in the computations,

we assume that $b_1(K) \leqslant b_2(K) \leqslant b_3(K)$.

Lemma 7.1. *Under Assumptions 5.1 and 7.2,*

$$\sqrt{n}[a(\hat{h}^W) - a(h^W)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A\, \mathbb{E}(p_i p_i')^{-1} \left[p_i e_i^W + \sum_{t=1}^{T} (H_t^W - dP_t^W)(I_{k_2} \otimes \mathbb{E}(r_{it} r_{it}')^{-1}) (v_{it} \otimes r_{it}) \right] + o_{\mathbb{P}}(1).$$

Proof. As argued in the proof of Result 5.1, under Assumption 7.2 (5), we can impose without loss of generality the normalization $\mathbb{E}(p_i p_i') = I_K$ and $\mathbb{E}(r_i r_i') = I_L$.

We define $\Delta_{\partial P} = b_2(K)\sqrt{L/n}$, $\Delta_Q = b_2(K)^2\Delta_n^2 + \sqrt{K} b_2(K)\Delta_n + b_1(K)\sqrt{K/n}$, $\Delta_{Q1} = a_1(L)\sqrt{L/n}$ and $\Delta_H = b_2(K)\Delta_n\sqrt{L} + a_1(L)\sqrt{K/n}$. Recall that $\Delta_n^2 = L/n + L^{-2\gamma_1}$. Under Assumption 7.1 (3), $\sqrt{n}K^{-\gamma_2} = o(1)$ and

$$\begin{aligned}
\sqrt{K}\Delta_Q &= O(\sqrt{K}[\sqrt{K} b_2(K)\Delta_n + b_1(K)\sqrt{K/n}]) = O(K b_2(K)[\sqrt{L/n} + L^{-\gamma_1}]) = o(1), \\
\Delta_{\partial P}\sqrt{L} &= b_2(K)L/\sqrt{n} = o(1), \qquad b_2(K)\Delta_{Q1}\sqrt{L} = b_2(K)a_1(L)L/\sqrt{n} = o(1), \\
b_2(K)\Delta_n &= O(b_2(K)\sqrt{n}L^{-\gamma_1} + \Delta_{\partial P}) = o(1), \\
\sqrt{L}\Delta_H &= O(b_2(K)L(\sqrt{L/n} + L^{-\gamma_1}) + \max(\sqrt{K}, \sqrt{L}b_2(K))\, a_1(L)\sqrt{L/n}) = o(1), \\
\Delta_Q b_2(K) &= O\left(b_2(K)^2\sqrt{K}[L^{-\gamma_1} + \sqrt{L/n}] + b_2(K)^2\sqrt{K/n}\right) = o(1), \\
\sqrt{n} b_3(K)\Delta_n^2 &= b_3(K)[L/\sqrt{n} + \sqrt{n}L^{-2\gamma_1}] = o(1).
\end{aligned}$$

These results also imply $\Delta_{\partial P} = o(1)$, $\Delta_Q = o(1)$, $\Delta_{Q1} = o(1)$ and $\Delta_H = o(1)$. The convergence of all these rates to zero as n grows will be used in several steps of the proof.

From the results stated in the proof of Result 5.1, we obtain $||Q_1 - I_L|| = O_{\mathbb{P}}(\Delta_{Q1}) = o_{\mathbb{P}}(1)$ and $||\hat{Q} - I_K|| = O_{\mathbb{P}}(\Delta_Q) = o_{\mathbb{P}}(1)$. This implies, as argued in NPV99, that the eigenvalues of $\hat{Q}$ are bounded away from 0 w. p. a 1, therefore $||B\hat{Q}^{-1}|| \leqslant ||B|| O_{\mathbb{P}}(1)$ and $||B\hat{Q}^{-1/2}|| \leqslant ||B|| O_{\mathbb{P}}(1)$ for any matrix B. Using

$$\begin{aligned}
||\overline{H}_t^W - H_t^W|| \leqslant & ||\frac{1}{n}\sum_{i=1}^{n} (\hat{p}_i - p_i)(\partial h_\varsigma^W(V_i)/\partial v_t \otimes r_{it}')|| \\
& + ||\frac{1}{n}\sum_{i=1}^{n} p_i\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}') - \mathbb{E}[p_i\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}')]||,
\end{aligned}$$

with $||\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - p_i)(\partial h_\varsigma^W(V_i)/\partial v_t \otimes r_{it}')|| \leqslant \frac{1}{n}||\hat{P} - P||\, \sup_{\mathcal{S}_V^\varsigma} ||\partial h_\varsigma^W||\, \mathrm{tr}(RR')^{1/2} = O_{\mathbb{P}}(b_2(K)\Delta_n\sqrt{L})$, and

$$\begin{aligned}
&\mathbb{E}(||\frac{1}{n}\sum_{i=1}^{n} p_i\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}') - \mathbb{E}[p_i\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}')]||^2) \\
&\quad \leqslant \mathbb{E}(\mathrm{tr}\left(\frac{1}{n^2}\sum_{i=1}^{n} p_i\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}')\, (\partial h^W(V_i)/\partial v_t \otimes r_{it}')' p_i'\right)) \leqslant C\frac{1}{n}\mathbb{E}(\mathrm{tr}(p_i r_i' r_i p_i')) \leqslant C a_1(L)^2 K/n,
\end{aligned}$$

we obtain $||\overline{H}_t^W - H_t^W|| = O_{\mathbb{P}}(\Delta_H)$. By Assumption 7.2 (7), $||\overline{H}_t^W - H_t^W|| = o_{\mathbb{P}}(1)$. Moreover, since $\rho^W(.)$ is a bounded function and $\mathbb{E}(r'_{it}r_{it}) = \mathbb{E}(tr(I_L)) = L$, we have for all $t \leqslant T$,

$$\mathbb{E}(||\overline{\mathrm{d}P}_t - \mathrm{d}P_t||^2) \leqslant \frac{C}{n^2}\sum_{i=1}^{n}\mathbb{E}(\mathrm{tr}\left(\frac{\partial p^K(V_i)}{\partial v_t}\left(\frac{\partial p^K(V_i)}{\partial v_t}\right)' r'_{it}r_{it}\right)) \leqslant Cb_2(K)^2L/n,$$

which implies by M that $||\overline{\mathrm{d}P}_t - \mathrm{d}P_t|| = O_{\mathbb{P}}(\Delta_{\partial P}) = o_{\mathbb{P}}(1)$.

Since $a(.)$ is linear, $a(\hat{h}^W) = A\hat{\pi}_W$. Using Assumption 7.2 (2),

$$||a(p^K(.)'\pi_W^K) - a(h^W(.))|| \leqslant \sup_{\mathcal{S}_V}||p^K(.)'\pi_W^K - g^W(.)|| \leqslant C\sup_{\mathcal{S}_V^\varsigma}||p^K(.)'\pi_W^K - g_\varsigma^W(.)|| \leqslant CK^{-\gamma_2},$$

which implies using Assumptions 7.2 (7) that

$$\begin{aligned}\sqrt{n}[a(\hat{h}^W) - a(h^W)] &= \sqrt{n}FA[\hat{\pi}_W - \pi_W^K] + o_{\mathbb{P}}(1),\\ &= A\hat{Q}^{-1}\hat{P}(W - \hat{P}'\pi_W^K)/\sqrt{n} + o_{\mathbb{P}}(1).\end{aligned}$$

Since $||A\hat{Q}^{-1}\hat{P}(\hat{h}_\varsigma^W - \hat{P}'\pi_W^K)|| \leqslant ||A\hat{Q}^{-1}\hat{P}||||\hat{h}_\varsigma^W - \hat{P}'\pi_W^K|| \leqslant \sqrt{n}||A\hat{Q}^{-1/2}||\sqrt{n}\sup_{\mathcal{S}_V^\varsigma}||p^K(.)'\pi_W^K - h^W(.)||$, $A\hat{Q}^{-1}\hat{P}(\hat{h}_\varsigma^W - \hat{P}'\pi_W^K)/\sqrt{n} = o_{\mathbb{P}}(1)$ by $||A||$ bounded. Therefore we obtain as in NPV99,

$$\sqrt{n}(a(\hat{h}^W) - a(h^W)) = \underbrace{A\hat{Q}^{-1}\hat{P}e^W/\sqrt{n}}_{\text{(B)}} + \underbrace{A\hat{Q}^{-1}\hat{P}(h_\varsigma^W - \hat{h}_\varsigma^W)/\sqrt{n}}_{\text{(C)}} + o_{\mathbb{P}}(1). \tag{41}$$

We first focus on the term (B), which we decompose

$$\text{(B)} = AP'e^W/\sqrt{n} + \underbrace{A(\hat{Q}^{-1} - I)Pe^W/\sqrt{n}}_{(B1)} + \underbrace{A\hat{Q}^{-1}(\hat{P} - P)e^{W*}/\sqrt{n}}_{(B2)} + \underbrace{A\hat{Q}^{-1}(\hat{P} - P)\rho^W/\sqrt{n}}_{(B3)}.$$

Define $\vec{v} = (v_1, .., v_n)$. Since $\mathbb{E}(||Pe^W/\sqrt{n}||^2) = 1/n\,\mathrm{tr}[\mathbb{E}(Pe^W(e^W)'P')] = 1/n\,\mathrm{tr}[\mathbb{E}(P\,\mathbb{E}(e^W(e^W)'|\vec{v})\,P')] \leqslant C\,\mathrm{tr}[I_k] = O(K)$ by Assumption 7.2 (8), we have by the Markov inequality $||Pe^W/\sqrt{n}|| = O_{\mathbb{P}}(K^{1/2})$. Therefore,

$$\begin{aligned}&||(B1)|| \leqslant ||A\hat{Q}^{-1}||\,||I_K - \hat{Q}||\,||P\epsilon/\sqrt{n}|| = ||A||O_{\mathbb{P}}(1)\,||I_K - \hat{Q}||\,||P\epsilon/\sqrt{n}|| = O_{\mathbb{P}}(\Delta_Q K^{1/2})\\ &\quad\Rightarrow (B1) = o_{\mathbb{P}}(1),\end{aligned}$$

under Assumption 7.2 (8).

We now look at the extra terms (B2) and (B3), where we decomposed e^W as $\rho^W + e^{W*}$ since e^W itself is not conditionally mean independent of $(\hat{P} - P)$, while e^{W*} is. Indeed, since $\mathbb{E}(e^{W*}|\vec{x}, \vec{z}) = 0$,

$$\begin{aligned}\mathbb{E}(||(\hat{P} - P)e^{W*}/\sqrt{n}||^2|\vec{x}, \vec{z}) &= 1/n\,\mathrm{tr}[\mathbb{E}((\hat{P} - P)e^{W*}(e^{W*})'(\hat{P} - P)'|\vec{x}, \vec{z})]\\ &= 1/n\,\mathrm{tr}[(\hat{P} - P)\,\mathbb{E}(e^{W*}(e^{W*})'|\vec{x}, \vec{z})\,(\hat{P} - P)'] \leqslant C/n||\hat{P} - P||^2.\end{aligned}$$

By the proof of Result 5.1, $||\hat{P}-P||^2/n = O_{\mathbb{P}}(b_2(K)^2\Delta_n^2)$ (the difference is that now $b_2(K)$ is defined as the sup rate over the extended support $\mathcal{S}_V^S$) hence by CM, $||(\hat{P}-P)e^{W*}/\sqrt{n}|| = O_{\mathbb{P}}(b_2(K)\Delta_n)$.

$$||A\hat{Q}^{-1}(\hat{P}-P)e^{W*}/\sqrt{n}|| \leqslant ||A\hat{Q}^{-1}||\,||(\hat{P}-P)e^{W*}||/\sqrt{n} = ||A||O_{\mathbb{P}}(1)||(\hat{P}-P)e^{W*}||/\sqrt{n}$$
$$\Rightarrow (B2) = O_{\mathbb{P}}(b_2(K)\Delta_n) = o_{\mathbb{P}}(1).$$

We now focus on (B3). We have $(\hat{P}-P)\rho^W/\sqrt{n} = 1/\sqrt{n}\sum_{i=1}^n(\hat{p}_i^K - p_i^K)\rho_i^W$ and a second order Taylor expansion gives

$$||(\hat{p}_i^K - p_i^K) - \frac{\partial p^K(\tau(V_i))}{\partial V}\frac{\partial \tau(V_i)}{\partial V}(\tilde{V}_i - V_i)|| \leqslant Cb_3(K)||\tilde{V}_i - V_i||^2,$$

which can be rewritten as $||(\hat{p}_i^K - p_i^K) - \frac{\partial p^K(V_i)}{\partial v}(\tilde{V}_i - V_i)|| \leqslant Cb_3(K)||\tilde{V}_i - V_i||^2$, since $V_i \in \mathcal{S}_V$ and we chose τ so that its Jacobian matrix is the identity matrix on $\mathcal{S}_V$. Hence

$$||A\hat{Q}^{-1}\frac{1}{\sqrt{n}}[(\hat{P}-P)\rho^W - \sum_{i=1}^n \frac{\partial p^K(V_i)}{\partial V}(\tilde{V}_i - V_i)\rho_i^W]\,|| \leqslant CO_{\mathbb{P}}(1)b_3(K)\sum_{i=1}^n ||\tilde{V}_i - V_i||^2/\sqrt{n}$$
$$= O_{\mathbb{P}}(\sqrt{n}b_3(K)\Delta_n^2) = o_{\mathbb{P}}(1),$$

by Assumption 7.2 (7). Therefore $(B3) = A\hat{Q}^{-1}\sum_{i=1}^n \frac{\partial p^K(V_i)}{\partial v}(\tilde{V}_i - V_i)\rho_i^W/\sqrt{n} + o_{\mathbb{P}}(1)$. We can decompose

$$\begin{aligned}(-1)(\tilde{v}_{it} - v_{it}) &= \hat{\beta}_t' r_{it} - b_{it} = (Q_{1t}^{-1}R_t[X_t^{2\prime} - R_t'\beta_t^L]/n)'r_{it} + \beta_t^{L\prime}r_{it} - b_{it},\\ &= [(Q_{1t}^{-1}R_t\vec{v}_t'/n)'r_{it}] + [(Q_{1t}^{-1}R_t[(b_t)' - R_t'\beta_t^L]/n)'r_{it}] + [\beta_t^{L\prime}r_{it} - b_{it}],\end{aligned}$$

and then apply this decomposition to $(B.3)$, $(B.3) = -[(B3.1) + (B3.2) + (B3.3)] + o_{\mathbb{P}}(1)$, where

$$\begin{aligned}(B3.1) &= \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_t}\left[Q_{1t}^{-1}R_t[(b_t)' - R_t'\beta_t^L]/n\right]' r_{it}/\sqrt{n}\\ (B3.2) &= \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_t}[\beta_t^{L\prime}r_{it} - b_{it}]/\sqrt{n}\\ (B3.3) &= \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_t}\left[Q_{1t}^{-1}R_t\vec{v}_t'/n\right]' r_{it}/\sqrt{n}\end{aligned}$$

The first term in this expression of $(B.3)$ can be rewritten

$$\begin{aligned}(B3.1) &= \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_t}\left[Q_{1t}^{-1}R_t[(b_t)' - R_t'\beta_t^L]/n\right]' r_{it}/\sqrt{n}\\ &= \sum_{t=1}^T A\hat{Q}^{-1}\frac{1}{n}\sum_{i=1}^n \rho_i^W \left(\frac{\partial p^K(V_i)}{\partial v_t}\otimes r_{it}'\right)\mathrm{Vec}(Q_{1t}^{-1}R_t[(b_t)' - R_t'\beta_t^L])/\sqrt{n}\\ &= \sum_{t=1}^T A\hat{Q}^{-1}\overline{\mathrm{d}P}_t\,(I_{k_2}\otimes Q_{1t}^{-1}R_t)\,\mathrm{Vec}((b_t)' - R_t'\beta_t^L)/\sqrt{n},\end{aligned}$$

where $||\operatorname{Vec}((b_t)' - R_t'\beta_t^L)|| \leqslant \sqrt{n}\,\sup_{\mathcal{S}_{\xi_t}} ||b_t(.) - \beta_t^{L\prime} r^L(.)|| \leqslant \sqrt{n} L^{-\gamma_1}$. Defining, for $d \leqslant d_2$, the matrix $\overline{\mathrm{d}P}_{td} = \frac{1}{n}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_{td}} r_{it}'$ where v_{td} is the d^{th} component of v_t, then $\overline{\mathrm{d}P}_t = (\overline{\mathrm{d}P}_{t1}, .., \overline{\mathrm{d}P}_{tk_2})$, and we can write

$$||A\hat{Q}^{-1}\overline{\mathrm{d}P}_t(I_{d_2} \otimes Q_{1t}^{-1}R_t)||^2 = \sum_{d=1}^{d_2} ||A\hat{Q}^{-1}\overline{\mathrm{d}P}_{td}Q_{1t}^{-1}R_t)||^2 = \sum_{d=1}^{d_2} \operatorname{tr}(A\hat{Q}^{-1}\overline{\mathrm{d}P}_{td}Q_{1t}^{-1}R_tR_t'Q_{1t}^{-1}\overline{\mathrm{d}P}_{td}'\hat{Q}^{-1}A')$$
$$= n\sum_{d=1}^{d_2} \operatorname{tr}(A\hat{Q}^{-1}\overline{\mathrm{d}P}_{td}Q_{1t}^{-1}\overline{\mathrm{d}P}_{td}'\hat{Q}^{-1}A'),$$

and $\overline{\mathrm{d}P}_{td}Q_t^{-1}\overline{\mathrm{d}P}_{td}' = \left(\frac{1}{n}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_{td}} r_{it}'\right)\left(\frac{1}{n}\sum_{i=1}^n r_{it}r_{it}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_{td}} r_{it}'\right)'$.

By $R_t'(R_tR_t')^{-1}R_t$ being an orthogonal projection matrix,

$$\overline{\mathrm{d}P}_{td}Q_t^{-1}\overline{\mathrm{d}P}_{td}' \leqslant \frac{1}{n}\sum_{i=1}^n (\rho_i^W)^2 \frac{\partial p^K(V_i)}{\partial v_{td}}\left(\frac{\partial p^K(V_i)}{\partial v_{td}}\right)',$$

implying $||\overline{\mathrm{d}P}_{td}Q_t^{-1/2}|| \leqslant b_2(K)$ and

$$\operatorname{tr}(A\hat{Q}^{-1}\overline{\mathrm{d}P}_{td}Q_t^{-1}\overline{\mathrm{d}P}_{td}'\hat{Q}^{-1}A') \leqslant \frac{1}{n}\sum_{i=1}^n (\rho_i^W)^2 \operatorname{tr}(A\hat{Q}^{-1}\frac{\partial p^K(V_i)}{\partial v_{td}}\left(\frac{\partial p^K(V_i)}{\partial v_{td}}\right)'\hat{Q}^{-1}A')$$
$$\leqslant \frac{1}{n}||A\hat{Q}^{-1}||^2 \sum_{i=1}^n (\rho_i^W)^2 ||\frac{\partial p^K(V_i)}{\partial v_{td}}||^2 = O_{\mathbb{P}}(1)\, b_2(K)^2,$$

since $\rho^W(.)$ is bounded. Hence $||A\hat{Q}^{-1}\overline{\mathrm{d}P}_t(I_{k_2} \otimes Q_{1t}^{-1}R_t)||^2 = O_{\mathbb{P}}(nb_2(K)^2)$ and we obtain by Assumption 7.2 (7),

$$||(B3.1)|| = O_{\mathbb{P}}(\sqrt{n}b_2(K)\sqrt{n}L^{-\gamma_1}/\sqrt{n}) = O_{\mathbb{P}}(b_2(K)\sqrt{n}L^{-\gamma_1}) = o_{\mathbb{P}}(1).$$

Focusing now on the second term in the expression of $(B3)$,

$$||(B3.2)|| = \Big|\Big|\sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \rho_i^W \frac{\partial p^K(V_i)}{\partial v_t}[\beta_t^{L\prime} r_{it} - b_{it}]\Big|\Big|/\sqrt{n} \leqslant ||A\hat{Q}^{-1}||nb_2(K)\, CL^{-\gamma_1}/\sqrt{n}$$
$$= O_{\mathbb{P}}(1)b_2(K)\sqrt{n}L^{-\gamma_1} = o_{\mathbb{P}}(1),$$

again by Assumption 7.2 (7). This implies that $(B3) = -(B3.3) + o_{\mathbb{P}}(1)$, with

$$(B3.3) = \sum_{t=1}^T A\hat{Q}^{-1}\overline{\mathrm{d}P}_t(I_{k_2} \otimes Q_{1t}^{-1})\operatorname{Vec}(R_t\vec{v}_t')/\sqrt{n}.$$

First, by $\mathbb{E}(||v_{it}||^2|\xi_{it} = \xi_t)$ bounded, $||R_t\vec{v}_t'/\sqrt{n}|| = O_{\mathbb{P}}(\sqrt{L})$, which gives

$$||\sum_{t=1}^T A\hat{Q}^{-1}\overline{\mathrm{d}P}_t[I_{k_2} \otimes Q_{1t}^{-1} - I_{k_2}L]\operatorname{Vec}(R_t\vec{v}_t')/\sqrt{n}||,$$
$$\leqslant ||A\hat{Q}^{-1}|| \sum_{t=1}^T ||\overline{\mathrm{d}P}_t(I_{k_2} \otimes Q_{1t}^{-1/2})||\, ||(I_{k_2} \otimes Q_{1t}^{-1/2}||\, ||I_{k_2} \otimes (I_L - Q_{1t})||\, ||R_t\vec{v}_t'/\sqrt{n}||,$$

$$\leqslant O_{\mathbb{P}}(1)Cb_2(K)O_{\mathbb{P}}(1)\Delta_{Q1}\sqrt{L} = O_{\mathbb{P}}(b_2(K)\Delta_{Q1}\sqrt{L}) = o_{\mathbb{P}}(1),$$

by Assumption 7.2 (7). Similarly

$$||\sum_{t=1}^{T} A\hat{Q}^{-1}(\overline{\mathrm{d}P}_t - \mathrm{d}P_t)\,\mathrm{Vec}(R_t\vec{v}_t')/\sqrt{n}|| \leqslant C||A\hat{Q}^{-1}||\,||\overline{\mathrm{d}P}_t - \mathrm{d}P_t||\,||R_t\vec{v}_t'/\sqrt{n}|| = O_{\mathbb{P}}(\Delta_{\partial P}\sqrt{L}) = o_{\mathbb{P}}(1).$$

Finally, we write $\mathrm{d}P_{td} = \mathbb{E}\left(\rho_i^W \partial p^K(V_i)/\partial v_{td} r_{it}'\right)$, as well as v_{itd} the d^{th} component of v_{it} and $\vec{v}_{td} = (v_{1td}, .., v_{ntd})'$. Then

$$\begin{aligned}
\mathbb{E}(||\mathrm{d}P_{td}R_t\vec{v}_{td}'/\sqrt{n}||^2) &= \mathrm{tr}\left(\mathrm{d}P_{td}\mathbb{E}\left(R_t\mathbb{E}(\vec{v}_{td}\vec{v}_{td}'|\vec{x}^1,\vec{z})R_t'\right)\mathrm{d}P_{td}'\right)/n \leqslant C\,\mathrm{tr}\left(\mathrm{d}P_{td}\mathbb{E}\left(R_tR_t'\right)\mathrm{d}P_{td}'\right)/n \\
&\leqslant C\,\mathrm{d}P_{td}\mathrm{d}P_{td}' = C\mathbb{E}\left(h_i^W\partial p^K(V_i)/\partial v_{td}r_{it}'\right)\mathbb{E}(r_{it}r_{it}')^{-1}\mathbb{E}\left(\rho_i^W r_{it}\partial p^K(V_i)'/\partial v_{td}\right) \\
&\leqslant C\mathbb{E}((\rho_i^W)^2\partial p^K(V_i)/\partial v_{td}(\partial p^K(V_i)/\partial v_{td})') \leqslant Cb_2(K)^2,
\end{aligned}$$

where the second to last inequality follows from taking the orthogonal projection matrix argument to the limit.This implies that $||\mathrm{d}P_{td}R_t\vec{v}_{td}'/\sqrt{n}|| = O_{\mathbb{P}}(b_2(K))$, and

$$\begin{aligned}
||\sum_{t=1}^{T} A(\hat{Q}^{-1} - I_{k_2})\mathrm{d}P_t\,\mathrm{Vec}(R_t\vec{v}_t')/\sqrt{n}|| &\leqslant \sum_{t=1}^{T}\sum_{d=1}^{d_2} ||A\hat{Q}^{-1}(I_{k_2} - \hat{Q})\mathrm{d}P_{td}R_t\vec{v}_{td}')/\sqrt{n}|| \\
&\leqslant ||A\hat{Q}^{-1}||\,O_{\mathbb{P}}(\Delta_Q)O_{\mathbb{P}}(b_2(K)) = O_{\mathbb{P}}(\Delta_Q b_2(K)) = o_{\mathbb{P}}(1),
\end{aligned}$$

by Assumption 7.2 (7). We can now write

$$(B3.3) = \frac{1}{\sqrt{n}}A\sum_{t=1}^{T}\mathrm{d}P_t\,\mathrm{Vec}(R_t\vec{v}_t') + o_{\mathbb{P}}(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A\sum_{t=1}^{T}\mathrm{d}P_t v_{it}\otimes r_{it} \;+\; o_{\mathbb{P}}(1),$$

where the term appearing in the sum over n, were the weight matrices not normalized, would become $\sum_{t\leqslant T} A\mathbb{E}(p_ip_i')^{-1}\mathrm{d}P_t\left(I_{k_2}\otimes\mathbb{E}(r_{it}r_{it}')^{-1}\right)v_{it}\otimes r_{it}$. Adding all terms appearing in (B), one obtains,

$$(\mathrm{B}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A\left[p_ie_i^W \;-\; \sum_{t=1}^{T}\mathrm{d}P_t\left(v_{it}\otimes r_{it}\right)\right] + o_{\mathbb{P}}(1).$$

The remaining term in the expression of $\sqrt{n}(a(\hat{h}^W) - a(h^W))$ is $(C) = A\hat{Q}^{-1}\hat{P}(h_\varsigma^W - \hat{h}_\varsigma^W)/\sqrt{n}$. This term is similar to the second term in equation $(A.16)$ of NPV99, p598, where the regression function is becomes h_ς^W. Since $v\mapsto h_\varsigma^W(\tau(v))$ is by composition twice continuously differentiable and has bounded second order derivative on the extended support, one obtains using NPV99

$$(\mathrm{C}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A\sum_{t=1}^{T} H_t^W(v_{it}\otimes r_{it}) + o_{\mathbb{P}}(1),$$

adapting to the fact that h_ς^W is here function of T generated covariates instead of one and using $\sqrt{n}L^{-\gamma_1}$, $\sqrt{n}b_1(K)\Delta_n^2$, $\sqrt{L}\Delta_{Q1}$, $\sqrt{K}\Delta_Q$ and $\sqrt{L}\Delta_H$ converge to zero as n goes to infinity. Note that, absent the normalization of the weight matrices, the term summed over n in the previous equation would become $A\,\mathbb{E}(p_ip_i')^{-1}\sum_{t=1}^{T} H_t^W(I_{k_2}\otimes\mathbb{E}(r_{it}r_{it}')^{-1})(v_{it}\otimes r_{it})$. □

Compared to the linearization obtained in Newey et al. (1999), there is an extra term $\mathrm{d}P_t^W$ which would be zero if $\rho^W(X,Z) = 0$. This refers to the absence of the above mentioned orthogonality condition. Another difference is the sumation over t of the H_t^W and $\mathrm{d}P_t^W$ terms: this is due to vector of control variables being composed of T estimated residuals.

Assumption 7.2 (6) is a condition on the functional a applied to the elements of the approximating basis. The functionals applied to our estimators are derived from the linearization of $\mathcal{X}$, which is the sum of linear functionals applied to the conditional expectations $\mathcal{M}$ and k. These functionals take the form of an expectation $a(h^W) = \int \lambda_a(v) h^W(v) dF_V(v)$. This corresponds to the mean square continuity condition of Newey (1997) under which he obtains $\sqrt{n}$ asymptotic normality of a linear functional of a linear sieve estimator. We exploit properties implied by this specification of a but instead obtain an intermediary result on the mean square convergence of the term $\mathbb{E}(p_i p_i')^{-1}\left[p_i e_i^W + \sum_{t=1}^T (H_t^W - \mathrm{d}P_t^W)\left(I_{k_2} \otimes \mathbb{E}(r_{it} r_{it}')^{-1}\right)(v_{it} \otimes r_{it})\right]$. This term is the sum of three elements, and we give separate results for each element. These results will then be used in the following section to show that Condition (6) of Assumption 7.2 holds and to obtain the total asymptotic variance matrix of $\sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0]$.

Assumption 7.3.

1. *There exists a function* $\lambda_a : \mathbb{R}^{Td_2} \mapsto \mathbb{R}^{d_a}$ *such that* $a(h^W) = \int \lambda_a(v) h^W(v) dF_V(v)$,
2. *There exists* ι_a^K *such that* $|\lambda_a(V) - \iota_a^K p^K(V)|_1 = O(K^{-\gamma_3})$, *there exist* ι_{aht}^L *and* $\iota_{a\rho t}^L$ *such that* $\mathbb{E}\left(||\mathbb{E}\left[\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right] - \iota_{aht}^L r^L(\xi_t)||^2\right) \to 0$ *and* $\mathbb{E}\left(||\mathbb{E}\left[\rho_i^W \frac{\partial \lambda_a(V)}{\partial v_{td}}|\xi_t\right] - \iota_{a\rho t}^L r^L(\xi_t)||^2\right) \to 0$ *as* $L \to \infty$,
3. $b_2(K)K^{-\gamma_3} = o(1)$,
4. *For all* V, $\mathrm{Var}(e_i^W|V) \leqslant C$.

Note that Condition (2) is a stronger condition than $\mathbb{E}(||\lambda_a(V) - \iota_a^K p^K(V)||^2) \to 0$, which would usually be assumed to obtain $\sqrt{n}$ asymptotic normality. Indeed, Assumption 7.3 (2) is a sup norm rate condition on both λ_a and $\partial\lambda_a/\partial V$. The $\mathrm{d}P_t^W$ term we obtained in Lemma 7.1 includes derivatives of the vector of basis functions. Loosely speaking, left multiplication of $\mathrm{d}P_t^W$ by the matrix A, which is the matrix of expectations of λ_a multiplied by the functions, will yield an approximation of the derivative of λ_a under Condition (2), which will appear in the asymptotic variance of our estimator when applied to specific functionals.

This mathematical intuition will be made explicit in the proof and requires the definition of the functions $\tilde{\lambda}_a(v) = A\,\mathbb{E}(p_i p_i')^{-1} p^K(v)$, $\tilde{\lambda}_{a,td}^{\partial}(\xi_t) = A\,\mathbb{E}(p_i p_i')^{-1} H_{td}^W \mathbb{E}(r_{it} r_{it}')^{-1} r^L(\xi_t)$ and ${}^{\partial}\tilde{\lambda}_{a,td}(\xi_t) = A\,\mathbb{E}(p_i p_i')^{-1} \mathrm{d}P_{td}^W \mathbb{E}(r_{it} r_{it}')^{-1} r^L(\xi_t)$.

Lemma 7.2. *Under Assumption 7.2 and 7.3, as $K, L \to \infty$, $\mathbb{E}(||e_i^W(\tilde{\lambda}_a(V) - \lambda_a(V))||^2) \to 0$, $\mathbb{E}\left(||v_{i,td}(\tilde{\lambda}^{\partial}_{a,td}(\xi_t) - \mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right))||^2\right) \to 0$, and $\mathbb{E}(||v_{i,td}(^{\partial}\tilde{\lambda}_{a,td} - \mathbb{E}\left[\rho_i^W \frac{\partial \lambda_a(V)}{\partial v_{td}}|\xi_t\right])||^2) \to 0$.*

Proof. By Assumption 7.2 (5), we can assume wlog that $\mathbb{E}(p_i p_i') = I_K$ and $\mathbb{E}(r_{it} r_{it}') = I_L$. Note that $\tilde{\lambda}_a(v) = A\, p^K(v) = \mathbb{E}\left(\lambda_a(V) p^K(V)'\right) p^K(V)$ is the mean square projection of λ_a on the functional space spanned by p^K. As in Newey (1997), Theorem 3, this implies that $\mathbb{E}(||e_i^W(\tilde{\lambda}_a(V) - \lambda_a(V))||^2) \leqslant C\mathbb{E}(||\pi_a^K p^K(V) - \lambda_a(V)||^2) \to 0$, using Assumption 7.3(4).

Following NPV99, noticing that $\tilde{\lambda}^{\partial}_{a,td}(\xi_t) = A\, H^W_{td} r^L(\xi_t) = \mathbb{E}\left(\tilde{\lambda}_a(V)\frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)'\right) r^L(\xi_t)$, and since $\mathbb{E}\left([\tilde{\lambda}_a(V) - \lambda_a(V)]\frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)'\right) r^L(\xi_t)$ is the mean square projection of $\mathbb{E}(\tilde{\lambda}_a(V) - \lambda_a(V)|\xi_t)$ on the functional space spanned by r^L, by properties of projection we have

$$\begin{aligned}\mathbb{E}\left(\left|\left|\tilde{\lambda}^{\partial}_{a,td}(\xi_t) - \mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)'\right) r^L(\xi_t)\right|\right|^2\right) \\ \leqslant \mathbb{E}\left(||[\tilde{\lambda}_a(V) - \lambda_a(V)]\frac{\partial h^W(V)}{\partial v_{td}}||^2\right) \leqslant C\mathbb{E}\left(||[\tilde{\lambda}_a(V) - \lambda_a(V)]||^2\right) \to 0,\end{aligned}$$

where the last inequality holds by Assumption 7.2 (3).

As for the $\tilde{\lambda_a}$, since $\mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)'\right) r^L(\xi_t)$ is the mean square projection of $\mathbb{E}\left[\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right]$, then

$$\begin{aligned}\mathbb{E}\left(\left|\left|\mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)'\right) r^L(\xi_t) - \mathbb{E}\left[\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right]\right|\right|^2\right) \\ \leqslant \mathbb{E}(||\iota^L_{aht} r^L(\xi_t) - \mathbb{E}\left[\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right]||^2) \to 0.\end{aligned}$$

This implies $\mathbb{E}\left(||\tilde{\lambda}^{\partial}_{a,td}(\xi_t) - \mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right)||^2\right) \to 0$, and, by Assumption 7.2 (8),

$$\mathbb{E}\left(||v_{i,td}(\tilde{\lambda}^{\partial}_{a,td}(\xi_t) - \mathbb{E}\left(\lambda_a(V)\frac{\partial h^W(V)}{\partial v_{td}}|\xi_t\right))||^2\right) \to 0.$$

For the third result, we need

$$\begin{aligned}\mathbb{E}(||\frac{\partial \tilde{\lambda}_a(V_i)}{\partial v_{td}} - \frac{\partial \lambda_a(V_i)}{\partial v_{td}}||^2) \leqslant & 2\mathbb{E}(||\mathbb{E}(\lambda_a(V)p^K(V)')\frac{\partial p^K(V_i)}{\partial v_{td}} - \iota_a^K \frac{\partial p^K(V_i)}{\partial v_{td}}||^2) \\ & + 2\mathbb{E}(||\iota_a^K \frac{\partial p^K(V_i)}{\partial v_{td}} - \frac{\partial \lambda_a(V_i)}{\partial v_{td}}||^2),\end{aligned}$$

where the second term in the sum converges to 0 by Assumption 7.3 (2). The first term is

$$\begin{aligned}\mathbb{E}\left(\left|\left|\mathbb{E}([\lambda_a(V) - \iota_a^K p^K(V)]p^K(V)')\frac{\partial p^K(V_i)}{\partial v_{td}}\right|\right|^2\right) &\leqslant b_2(K)^2\mathbb{E}(||[\lambda_a(V) - \iota_a^K p^K(V)||] \\ &= O(b_2(K)^2 K^{-2\gamma_3}) \to 0,\end{aligned}$$

which implies that $\mathbb{E}(||\frac{\partial \tilde{\lambda}_a(V_i)}{\partial v_{td}} - \frac{\partial \lambda_a(V_i)}{\partial v_{td}}||^2) \to 0$.

Since ${}^{\partial}\tilde{\lambda}_{a,td}(\xi_t) = A\,\mathrm{d}P^W_{td}\,r^L(\xi_t) = \mathbb{E}(A\rho^W_i\left(\frac{\partial p^K(v_i)}{\partial v_{td}}\otimes r'_{it}\right))r^L(\xi_t)$, we have by property of MSE projection,

$$\mathbb{E}\left(\left|\left|{}^{\partial}\tilde{\lambda}_{a,td}(\xi_t) - \mathbb{E}\left(\rho^W_i\frac{\partial\lambda_a(V_i)}{\partial v_{td}}\otimes r^L(\xi_t)'\right)r^L(\xi_t)\right|\right|^2\right) \leqslant \mathbb{E}\left(|\rho^W_i|\,\left|\left|\frac{\partial\tilde{\lambda}_a(V_i)}{\partial v_{td}} - \frac{\partial\lambda_a(V_i)}{\partial v_{td}}\right|\right|^2\right) \to 0$$

by Assumption 7.2 (3) and the result obtained above. Moreover,

$$\begin{aligned}\mathbb{E}\left(\left|\left|\mathbb{E}\left(\rho^W_i\frac{\partial\lambda_a(V_i)}{\partial v_{td}}\otimes r^L(\xi_t)'\right)r^L(\xi_t) - \mathbb{E}\left[\rho^W_i\frac{\partial\lambda_a(V)}{\partial v_{td}}|\xi_t\right]\right|\right|^2\right)& \\ \leqslant \mathbb{E}(||\iota^L_{a\rho t}r^L(\xi_t) - \mathbb{E}\left[\rho^W_i\frac{\partial\lambda_a(V)}{\partial v_{td}}|\xi_t\right]||^2) \to 0,&\end{aligned}$$

which implies $\mathbb{E}(||{}^{\partial}\tilde{\lambda}_{a,td} - \mathbb{E}\left[\rho^W_i\frac{\partial\lambda_a(V)}{\partial v_{td}}|\xi_t\right]||^2) \to 0$, and $\mathbb{E}(||v_{i,td}({}^{\partial}\tilde{\lambda}_{a,td} - \mathbb{E}\left[\rho^W_i\frac{\partial\lambda_a(V)}{\partial v_{td}}|\xi_t\right])||^2) \to 0$. □

7.4 Application to the model

7.4.1 Asymptotics of the linear part

We use the results derived in the previous section to obtain an asymptotic equivalent of $\sqrt{n}\mathcal{X}^G_0[\hat{\mathcal{G}} - \mathcal{G}_0]$. Indeed, (40) expresses $\mathcal{X}^G_0[\mathcal{G} - \mathcal{G}_0]$ as a sum of linear functionals applied to the components of $\mathcal{G} = ((b_t)_{t\leqslant T}, k, \mathcal{M})$, and we estimate both k and $\mathcal{M}$ with two-step sieve estimators.

We will first apply Lemma 7.1 for these estimators, choosing w_i to be either a component of $M_i\dot{y}_i$ or of M_i. By analogy with the general model, to apply Lemma 7.1 to $\mathcal{X}^M_0[\hat{\mathcal{M}}]$ we define the following objects

$$\begin{aligned}e^M_i &= M_i - \mathbb{E}(M_i|V_i) = M_i - \mathcal{M}_0(V_i),\\ e^{M*}_i &= M_i - \mathbb{E}(M_i|X^1_i, X^2_i, Z_i) = 0,\\ \rho^M(X^1_i, X^2_i, Z_i) &= \mathbb{E}(M_i|X^1_i, X^2_i, Z_i) - \mathbb{E}(M_i|V_i) = M_i - \mathcal{M}_0(V_i),\end{aligned}$$

and similarly for $\mathcal{X}^k_0[\hat{k}]$, we define

$$\begin{aligned}e^k_i &= M_i\dot{y}_i - \mathbb{E}(M_i\dot{y}_i|V_i) = [M_i - \mathcal{M}_0(V_i)]g(V_i) + M_i\dot{u}_i,\\ e^{k*}_i &= M_i\dot{y}_i - \mathbb{E}(M_i\dot{y}_i|X^1_i, X^2_i, Z_i) = M_i[\dot{u}_i - \mathbb{E}(\dot{u}_i|X^1_i, X^2_i, Z_i)],\\ \rho^k(X^1_i, X^2_i, Z_i) &= \mathbb{E}(M_i\dot{y}_i|X^1_i, X^2_i, Z_i) - \mathbb{E}(M_i\dot{y}_i|V_i) = [M_i - \mathcal{M}_0(V_i)]g(V_i) + M_i\mathbb{E}(\dot{u}_i|X^1_i, X^2_i, Z_i).\end{aligned}$$

It will be convenient to assume $\mathbb{E}(\dot{u}_i|X^1_i, X^2_i, Z_i) = 0$. We define now the analogs of the matrices A, $\mathrm{d}P^W_t$ and H^W_t. Write for a given v, $\lambda^j_M(v)$ the j^{th} column of the matrix $\lambda_M(v)$. We will need the following matrix

$$\Lambda^M = \int_v \left[\,\lambda^1_M(v)p^K_1(v),\ \ldots,\lambda^1_M(v)p^K_K(v),\ \ldots,\lambda^{(T-1)^2}_M(v)p^K_1(v),\ \ldots,\lambda^{(T-1)^2}_M(v)p^K_K(v)\right]dF_V(v),$$

which is of dimension $d_x \times K(T-1)^2$. Similarly, we define the matrix Λ^k. We will interchangeably index the columns of λ_M as λ_M^d with $d \leqslant (T-1)^2$ and as λ_M^{st} with $1 \leqslant s,t \leqslant T-1$. We will also need

$$H_t^M = \mathbb{E}\left[\frac{\partial \operatorname{Vec}(\mathcal{M}_0(V_i))}{\partial v_t} \otimes p_i \otimes r'_{it}\right],\ H_t^k = \mathbb{E}\left[\frac{\partial k_0(V_i)}{\partial v_t} \otimes p_i \otimes r'_{it}\right],$$
$$\mathrm{d}P_t^M = \mathbb{E}\left[\operatorname{Vec}(\rho_i^M) \otimes \frac{\partial p^K(V_i)}{\partial v_t} \otimes r'_{it}\right],\ \mathrm{d}P_t^k = \mathbb{E}\left[\rho_i^k \otimes \frac{\partial p^K(V_i)}{\partial v_t} \otimes r'_{it}\right].$$

The regression functions b_{0t} are estimated nonparametrically and the asymptotic distribution of functionals of such objects is studied in Newey (1997). For those, writing for a given ξ_t, $\lambda_{bt}^j(\xi_t)$ the j^{th} column of the matrix $\lambda_{bt}(\xi_t)$, we define for each t,

$$\Lambda^{bt} = \int_{\xi_t} \Big[\lambda_{bt}^1(\xi_t) r_1^L(\xi_t),\ \ldots,\lambda_{bt}^1(\xi_t) r_L^L(\xi_t),\ \ldots,\lambda_{bt}^{T-1}(\xi_t) r_1^L(\xi_t),\ \ldots,\lambda_{bt}^{T-1}(\xi_t) r_L^L(\xi_t)\Big] dF_{\xi_t}(\xi_t).$$

We now state the assumptions required to apply Lemma 7.1 on the functionals $\mathcal{X}_0^M$ and $\mathcal{X}_0^k$ applied respectively to $\hat{\mathcal{M}}$ and $\hat{k}$ where mentioned in the discussion before Assumption 7.2, we do not specify the basis of approximating functions.

Assumption 7.4.

1. *$W_i = (X_i, Z_i, V_i, \mu_i, u_i)$ are i.i.d, $\mathcal{S}_{\xi t}$ for all $t \leqslant T$ and $\mathcal{S}_V$ are bounded,*
2. *$\mathcal{M}_0$ and g are twice continuously differentiable and bounded, and their first and second order derivatives are bounded,*
3. *$\mathbb{E}(||Q_i^\delta||) < \infty$, and for all (X,Z), $\mathbb{E}(||\dot{u}_i|| \,|X_i^1, X_i^2, Z_i) < C$,*
4. *There exists γ_1 and β_t^L such that for all $t \leqslant T$, $\sup_{\mathcal{S}_{\xi_t}} ||b_{0t}(\xi_t) - \beta_t^{L\prime} r^L(\xi_t)|| \leqslant CL^{-\gamma_1}$. There exists γ_2, π_M^K and π_k^K such that $\sup_{\mathcal{S}_V^\varsigma} ||\mathcal{M}_0^\varsigma(v) - p^K(v)'\pi_M^K|| \leqslant CK^{-\gamma_2}$ and $\sup_{\mathcal{S}_V^\varsigma} ||k_0^\varsigma(v) - p^K(v)'\pi_k^K|| \leqslant CK^{-\gamma_2}$,*
5. *For all $t \leqslant T$, there exists a $L \times L$ nonsingular matrix Γ_{1t} such that for $R_t^L(\xi_t) = \Gamma_{1t} r_t^L(\xi_t)$, $\mathbb{E}(R_t^L(\xi_t) R_t^L(\xi_t)')$ has smallest eigenvalue bounded away from zero uniformly in L. Similarly, there exists a $K \times K$ nonsingular matrix Γ_2 such that for $P^K(V) = \Gamma_2 p^K(V)$, $\mathbb{E}(P^K(V)P^K(V)')$ has smallest eigenvalue bounded away from zero uniformly in K,*
6. *$||\Lambda^M||$, $||\Lambda^k||$, and $||\Lambda^{bt}||$ are bounded,*
7. *For $|R_t^L(\xi_t)|_0 \leqslant a_1(L)$, $|P^K(V)|_0^\varsigma \leqslant b_1(K)$, $|P^K(V)|_1^\varsigma \leqslant b_2(K)$, $|P^K(V)|_2^\varsigma \leqslant b_3(K)$, we have $\sqrt{n}K^{-\gamma_2} = o(1)$, $\max(\sqrt{K}, \sqrt{L}b_2(K))\, a_1(L)\sqrt{L/n} = o(1)$, $b_2(K)[\sqrt{L/n} + L^{-\gamma_1}][K+L] = o(1)$,*
   $b_2(K)\sqrt{n}L^{-\gamma_1} = o(1)$, $b_3(K)[L/\sqrt{n} + \sqrt{n}L^{-2\gamma_1}] = o(1)$, $b_2(K)^2\sqrt{K}[\sqrt{L/n} + L^{-\gamma_1}] = o(1)$,
8. *For all ξ_t, $\mathbb{E}(||v_{it}||^2|\xi_t) \leqslant C$ and for all (X,Z), $\mathbb{E}(||\dot{u}_i - \mathbb{E}(\dot{u}_i|X_i,Z_i)||^2|X_i,Z_i) \leqslant C$,*

Applying Lemma 7.1, we obtain the following linearization.

Result 7.2. *Under Assumptions 2.1, 2.3, 5.6 and 7.4,*

$$\begin{aligned}\sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0] \;=\; & o_{\mathbb{P}}(1) \\ & + \;\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda^M(I_{(T-1)^2}\otimes\Theta)\;\mathrm{Vec}(e_i^M)\otimes p_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda^k(I_{T-1}\otimes\Theta)\,e_i^k\otimes p_i \qquad (42) \\ & + \;\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big[\Lambda^M(I_{(T-1)^2}\otimes\Theta)(H_t^M - dP_t^M) + \Lambda^k(I_{T-1}\otimes\Theta)(H_t^k - dP_t^k) + \Lambda^{bt}\Big](I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it},\end{aligned}$$

where we define $\Theta = \mathbb{E}(p_i p_i')$ *and* $\Theta_1 = \mathbb{E}(r_i r_i')$.

Proof. We first focus on the functional $\mathcal{X}_0^{(2t)}[b_t] = \int_\xi \lambda_{bt}(\xi_{it})b_t(\xi_{it})dF_{\xi_t}(\xi_t)$. Newey (1997) shows in the proof of Theorem 2 (equation (A.7) p164 and the subsequent text) that if $||\mathcal{X}_0^{(bt)}[b_t]|| \leqslant C|b_t|_0$, $\sqrt{n}L^{-\gamma_1} \to 0$, $\Delta_{Q1} = a_1(L)\sqrt{L/n} \to 0$, and $||\Lambda^{bt}||$ is bounded, then $\sqrt{n}\mathcal{X}_0^{(bt)}(\hat{b}_t - b_{0t}) = \Lambda^{bt}\sum_{i=1}^n r_{it}\otimes v_{it}/\sqrt{n} + o_{\mathbb{P}}(1)$. For all $t \leqslant T$, under Assumption 7.4 (2) and (3), $||\mathcal{X}_0^{(bt)}[b_t]|| = ||\mathbb{E}(Q_i\frac{\partial g}{\partial v_t}(V_i)b_t(\xi_{it}))|| \leqslant C|b_t|_0$. The other required conditions hold by Assumption 7.4.

We now check that the conditions of Assumption 7.2 hold for the functionals $\mathcal{X}_0^{(k)}$ and $\mathcal{X}_0^{(M)}$ applied to the two-step estimators $\hat{k}$ and $\hat{\mathcal{M}}$. Under Assumption 5.6 we have by Result 5.4 that $\lambda_{\min}(\mathcal{M}(V)) \geqslant C$. Together with Assumption 7.4 (2) and (3), this guarantees that $||\mathcal{X}_0^{(k)}[\tilde{k}]|| = ||\mathbb{E}(Q_i\mathcal{M}_0(V_i)^{-1}\tilde{k}(V_i))|| \leqslant C|\tilde{k}|_0$ and $||\mathcal{X}_0^{(M)}[\tilde{\mathcal{M}}]|| = ||\mathbb{E}(Q_i\mathcal{M}_0(V_i)^{-1}\tilde{\mathcal{M}}(V_i)\mathrm{d}f(V_i)|| \leqslant C|\tilde{\mathcal{M}}|_0$. Moreover $\rho^M(X_i^1, X_i^2, Z_i) = M_i - \mathcal{M}_0(V_i)$ where $M_i = I - \dot{X}_i(\dot{X}_i'\dot{X}_i)^{-1}\dot{X}_i'$ if $\dot{X}_i$ is of full rank, or $M_i = I - \dot{X}_i\dot{X}_i^+$ if not, with $\dot{X}_i^+$ is the Moore Penrose inverse. In either case, $||M_i||_2 \leqslant 1$ implying $||M_i||_F \leqslant C$, as well as $\mathcal{M}_0(V_i) = \mathbb{E}(M_i|V_i)$, ensuring that ρ^M is a bounded function for the norm considered here. By the same argument and Assumption 7.4 (2) and (3), $\rho^k(X_i^1, X_i^2, Z_i) = [M_i - \mathcal{M}_0(V_i)]g(V_i) + M_i\mathbb{E}(\dot{u}_i|X_i^1, X_i^2, Z_i)$ is uniformly bounded. Hence Assumption 7.2 (2) holds for each functional.

By Assumption 2.1 and 2.3, $k_0(V) = \mathcal{M}_0^{-1}(V)g(V)$ therefore by Assumption 7.4 (2), k is twice continuously differentiable, implying that Assumption 7.2 (3) holds for each functional. Also, $\mathbb{E}(||e^{M*}||^2|X, Z) = 0$ for all (X, Z), and by Assumption 7.4 (8), for all (X, Z), $\mathbb{E}(||e^{k*}||^2|X, Z) = \mathbb{E}(||M[\dot{u} - \mathbb{E}(\dot{u}|X, Z)]||^2|X, Z) \leqslant C$, ensuring that Assumption 7.2 (8) holds for each functional. The other conditions of Assumption 7.2 are more direct consequences of Assumption 7.4. □

Assumption 7.4 (6) requires $||\Lambda^M||$, $||\Lambda^k||$, and $||\Lambda^{bt}||$ to be bounded : this condition is not a direct restriction on primitives of the model. We apply Lemma 7.2 to obtain such restrictions. This lemma will also be useful to obtain the asymptotic variance matrix of our estimator in the next section. Recall that $\sqrt{n}(\hat{\mu}^\delta - \mathbb{E}(\mu_i\delta_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n[\delta_i\mu_i - \mathbb{E}(\mu\delta)] + \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_i^\delta\dot{u}_i - \sqrt{n}\big[\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0)\big]$. Under Assumption 7.4, we can write $\frac{1}{\sqrt{n}}\sum_{i=1}^n[\delta_i\mu_i - \mathbb{E}(\mu\delta)] + \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_i^\delta\dot{u}_i - \sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0]$ as $\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{i,n} + o_{\mathbb{P}}(1)$ where

$$s_{i,n} = [\delta_i\mu_i - \mathbb{E}(\mu\delta)] + Q_i^\delta\dot{u}_i - \Lambda^M(I_{(T-1)^2}\otimes\Theta)\;\mathrm{Vec}(e_i^M)\otimes p_i - \Lambda^k(I_{T-1}\otimes\Theta)\,e_i^k\otimes p_i$$

$$+\sum_{t=1}^{T}\left[\Lambda^M(I_{(T-1)^2}\otimes\Theta)(\mathrm{d}P_t^M-H_t^M)+\Lambda^k(I_{T-1}\otimes\Theta)(\mathrm{d}P_t^k-G_t^k)-\Lambda^{bt}\right](I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it}.$$

We also define $\Omega=\mathrm{Var}(s_{i,n})$.

We define the following objects, $\tilde{Q}_i^\delta = Q_i^\delta - \mathbb{E}(Q_i^\delta|V_i)\mathcal{M}_0(V_i)^{-1}M_i$, and $\Omega_0 = \mathrm{Var}\Big([\delta_i\mu_i - \mathbb{E}(\mu\delta)] + \tilde{Q}_i^\delta\dot{u}_i + \sum_{t=1}^T \mathbb{E}\left(\tilde{Q}_i^\delta\frac{\partial g(V_i)}{\partial v_t}|\,\xi_{it}\right)v_{it}\Big)$.

Assumption 7.5.

1. *For each function $\lambda_a(.)$, column of $\lambda_{\mathcal{M}}(.)$ and $\lambda_k(.)$, $\lambda_a(.)$ is continuously differentiable and there exists ι_a^K such that $|\lambda_a(V)-\iota_a^K p^K(V)|_1 = O(K^{-\gamma_3})$. Moreover, there exists ι_{aht}^L and $\iota_{a\rho t}^L$ such that $\mathbb{E}\left(||\mathbb{E}\left[\lambda_a(V)\frac{\partial h^a(V)}{\partial v_{td}}|\xi_t\right]-\iota_{aht}^L r^L(\xi_t)||^2\right)\to 0$ and $\mathbb{E}\left(||\mathbb{E}\left[\rho_i^W\frac{\partial\lambda_a(V)}{\partial v_{td}}|\xi_t\right]-\iota_{a\rho t}^L r^L(\xi_t)||^2\right)\to 0$ as $L\to\infty$.*
   For all $t\leqslant T$, b_t is continuous and there exists ι_{2t}^L such that $\mathbb{E}\left(||\lambda_{bt}(\xi_t)-\iota_{bt}^L r^L(\xi_t)||^2\right)\to 0$ as $L\to\infty$,
2. *$b_2(K)K^{-\gamma_3}=o(1)$,*
3. *$\mathbb{E}(Q_i^\delta||^2)<\infty$, $\mathbb{E}(||\mu_i||^2)<\infty$, and there exists $C>0$ such that $\Omega_0\geqslant CI_{d_x}$,*
4. *$\mathbb{E}(\dot{u}_i|X_i^1,X_i^2,Z_i)=0$.*

Assumption 7.5 (4) is imposed to simplify computations. It amounts to strengthening the control function assumption, that is, Assumption 2.1 (2). The condition $\Omega_0\geqslant CI_{d_x}$ holds if for instance $\mathrm{Var}(\mu_i|X_i,Z_i,u_i,V_i)\geqslant CI_{d_x}$ for some $C>0$, or if a similar condition holds on the conditional variance of $\dot{u}_i$, as is typically assumed. We now state the result giving the asymptotic variance of the estimator, as well as the boundedness condition needed in the previous result. The boundedness of the two last matrices is added for later results on asymptotic normality.

Result 7.3. *Under Assumptions 2.1, 2.3, 5.6, 7.4 and 7.5, $\Omega^{-1/2}\to_{n\to\infty}\Omega_0^{-1/2}$. Moreover, $||\Lambda^M||$, $||\Lambda^k||$, $||\Lambda^{bt}||$, $||\Lambda^M(I_{(T-1)^2}\otimes\Theta)(H_t^M-dP_t^M)||$, and $||\Lambda^k(I_{T-1}\otimes\Theta)(H_t^k-dP_t^k)||$ are bounded.*

Proof. Under Assumptions 7.5, Conditions (1), (2) and (3) of Assumption 7.3 holds for $w_i=(M_i\dot{y}_i)_t$ and $w_i=(M_i)_{st}$ associated respectively with λ_k^t and $\lambda_M^{s,t}$. We showed that Assumption 7.4 implied that ρ^M and ρ^k are bounded, as well as $\mathbb{E}(||e^{M*}||^2|X,Z)$ and $\mathbb{E}(||e^{k*}||^2|X,Z)$: this implies that for all V, $\mathrm{Var}(e_i^{\mathcal{M}}|V)\leqslant C$ and $\mathrm{Var}(e_i^k|V)\leqslant C$. Condition (4) of Assumption 7.3 is also satisfied for our choices of w_i, hence we can apply Lemma 7.2.

We now use Equation (42) to construct the asymptotic variance. Define

$$s_i=[\delta_i\mu_i-\mathbb{E}(\mu\delta)]+Q_i^\delta\dot{u}_i-\lambda_{\mathcal{M}}(V_i)\,\mathrm{Vec}(e_i^M)-\lambda_k(V_i)e_i^k$$
$$+\sum_{t=1}^{T}\left(\mathbb{E}\left[\frac{\partial\lambda_{\mathcal{M}}(V)}{\partial v_t}\,\mathrm{Vec}(\rho_i^{\mathcal{M}})|\xi_{it}\right]-\mathbb{E}\left[\lambda_{\mathcal{M}}(V)\frac{\partial\mathcal{M}_0(V)}{\partial v_t}|\xi_t\right]\right.$$

$$+\mathbb{E}\left[\frac{\partial \lambda_k(V)}{\partial v_t}\rho_i^k|\xi_{it}\right] - \mathbb{E}\left[\lambda_k(V)\frac{\partial k_0(V)}{\partial v_t}|\xi_t\right] - \lambda_{bt}(\xi_{it})\Big)\; v_{it},$$

where, by a convenient abuse of notation, we denote with $\frac{\partial \lambda_{\mathcal{M}}(V)}{\partial v_t}\,\mathrm{Vec}(\rho_i^{\mathcal{M}})$ the sum $\sum_{j\leqslant(T-1)^2}\rho_{i,j}^{\mathcal{M}}\frac{\partial \lambda_{\mathcal{M}}^j(V)}{\partial v_t}$ with $\rho_{i,j}^{\mathcal{M}}$ the j^{th} component of the vector $\mathrm{Vec}(\rho^M(X_i^1, X_i^2, Z_i))$, and similarly for λ_k. We will, in a later step of this proof, simplify the formula for s_i.

For a constant vector $c \in \mathbb{R}^{d_x}$, $|c'[\mathbb{E}(s_{in}s_{in}') - \mathbb{E}(s_i s_i')]c| \leqslant \mathbb{E}([s_{in}'c - s_i'c]^2) + 2\mathbb{E}([s_i'c]^2)^{1/2}\mathbb{E}([s_{in}'c - s_i'c]^2)^{1/2}$. We conveniently decompose the difference $s_{i,n} - s_i$ as

$$\begin{aligned}
s_{i,n} - s_i \;=\; & \lambda_{\mathcal{M}}(V_i)\,\mathrm{Vec}(e_i^M) - \Lambda^M(I_{(T-1)^2}\otimes\Theta)\;\mathrm{Vec}(e_i^M)\otimes p_i \\
& + \lambda_k(V_i)e_i^k - \Lambda^k(I_{T-1}\otimes\Theta)\,e_i^k\otimes p_i \\
& + \sum_{t=1}^{T}\mathbb{E}\left[\lambda_{\mathcal{M}}(V)\frac{\partial \mathcal{M}_0(V)}{\partial v_t}|\xi_t\right] v_{it} - \Lambda^M(I_{(T-1)^2}\otimes\Theta)H_t^M(I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it} \\
& + \sum_{t=1}^{T}\mathbb{E}\left[\lambda_k(V)\frac{\partial k_0(V)}{\partial v_t}|\xi_t\right] v_{it} - \Lambda^k(I_{T-1}\otimes\Theta)H_t^k(I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it} \\
& + \sum_{t=1}^{T}\Lambda^M(I_{(T-1)^2}\otimes\Theta)\mathrm{d}P_t^M(I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it} - \mathbb{E}\left[\frac{\partial \lambda_{\mathcal{M}}(V)}{\partial v_t}\,\mathrm{Vec}(\rho_i^{\mathcal{M}})|\xi_{it}\right] v_{it} \\
& + \sum_{t=1}^{T}\Lambda^k(I_{T-1}\otimes\Theta)\mathrm{d}P_t^k(I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it} - \mathbb{E}\left[\frac{\partial \lambda_k(V)}{\partial v_t}\rho_i^k|\xi_{it}\right] \\
& + \lambda_{bt}(\xi_{it})v_{it} - \Lambda^{bt}(I_{k_2}\otimes\Theta_1)\,v_{it}\otimes r_{it},
\end{aligned}$$

where each line in this sum is of one of the three types of elements analyzed in Lemma 7.2, except for the last line. $\mathbb{E}(||s_{i,n} - s_i||^2)$ is bounded by the sum of the expected squared norms of the elements of each line. To show that it converges to zero as n goes to infinity, we use the fact that Assumption 7.3 holds for each λ_a, where λ_a is a column of either $\lambda_{\mathcal{M}}$ or λ_k. By Assumption 7.5 (1) and Assumption 7.4 (8), the expected squared norm of the term in the last line also converges to 0 as $n \to \infty$.

These arguments imply that $\mathbb{E}(||s_{i,n} - s_i||^2) \to 0$. By the proof of Result 7.2 and Assumption 7.5 (1), the functions multiplying the residuals appearing in the definition of s_i are all bounded. Together with Assumption 7.5 (3), this guarantees $\mathbb{E}([s_i'c]^2) < \infty$. Hence, $|c'[\mathbb{E}(s_{in}s_{in}') - \mathbb{E}(s_i s_i')]c| \to 0$, and $\mathbb{E}(s_{in}s_{in}') - \mathbb{E}(s_i s_i') \to 0$. That is, $\Omega \to \mathrm{Var}(s_i)$.

We can now simplify the formula for s_i using the primitives of the model. Indeed, note that

$$\begin{aligned}
\lambda_{\mathcal{M}}(V_i)\,\mathrm{Vec}(e_i^M) + \lambda_k(V_i)e_i^k \;&=\; \mathbb{E}(Q_i^\delta|V_i)\mathcal{M}_0(V_i)^{-1}M_i\dot{u}_i, \\
\lambda_{\mathcal{M}}(V)\frac{\partial \mathcal{M}_0(V)}{\partial v_t} + \lambda_k(V)\frac{\partial k_0(V)}{\partial v_t} \;&=\; \mathbb{E}(Q_i^\delta|V_i)\,\frac{\partial g(V_i)}{\partial v_t}, \\
\frac{\partial \lambda_{\mathcal{M}}(V)}{\partial v_t}\,\mathrm{Vec}(\rho_i^{\mathcal{M}}) + \frac{\partial \lambda_k(V)}{\partial v_t}\rho_i^k \;&=\; -\mathbb{E}(Q_i^\delta|V_i)\mathcal{M}_0(V_i)^{-1}(M_i - \mathcal{M}_0(V_i))\frac{\partial g(V_i)}{\partial v_t},
\end{aligned}$$

and since $\lambda_{bt}(\xi_{it}) = -\mathbb{E}\left(Q_i^\delta \frac{\partial g(V_i)}{\partial v_t} \,|\, \xi_{it} = \xi_t\right)$, we obtain

$$\begin{aligned} s_i =& [\delta_i\mu_i - \mathbb{E}(\mu\delta)] + Q_i^\delta \dot{u}_i - \mathbb{E}(Q_i^\delta|V_i)\mathcal{M}_0(V_i)^{-1}M_i\dot{u}_i \\ &+ \sum_{t=1}^{T} \mathbb{E}\left(\left[Q_i^\delta - \mathbb{E}(Q_i^\delta|V_i)\mathcal{M}_0(V_i)^{-1}M_i\right]\frac{\partial g(V_i)}{\partial v_t}|\xi_{it}\right) v_{it}, \\ =& [\delta_i\mu_i - \mathbb{E}(\mu\delta)] + \tilde{Q}_i^\delta \dot{u}_i + \sum_{t=1}^{T} \mathbb{E}\left(\tilde{Q}_i^\delta \frac{\partial g(V_i)}{\partial v_t}|\,\xi_{it}\right) v_{it}. \end{aligned}$$

By our definition of Ω_0, we have $\mathrm{Var}(s_i) = \Omega_0$. This implies $\Omega \to_{n\to\infty} \Omega_0$, and by $\Omega_0 \geqslant CI_{d_x}$, we obtain $\Omega^{-1/2} \to_{n\to\infty} \Omega_0^{-1/2} \leqslant C^{-1/2}I_{k_x}$.

We again use Lemma 7.2 to prove that $||\Lambda^M||$, $||\Lambda^k||$, and $||\Lambda^{bt}||$ are bounded. Indeed, since wlog we can assume $\Theta_1 = I_L$, we have $||\Lambda^{bt}||^2 = \mathrm{tr}(\Lambda^{bt}\Lambda^{bt\prime}) = \mathrm{tr}(\Lambda^{bt}(I_{k_2}\otimes\Theta_1)\Lambda^{bt\prime})$. Using the notation of Lemma 7.2 with $\tilde{\lambda}_{bt}^j(\xi) = \mathbb{E}(\lambda_{bt}^j(\xi_t)r^L(\xi_t)')\, r^l(\xi)$, this gives $||\Lambda^{bt}||^2 = \mathrm{tr}\left(\sum_{j\leqslant T-1}\mathbb{E}\left[\tilde{\lambda}_{bt}^j(\xi_t)\tilde{\lambda}_{bt}^j(\xi_t)'\right]\right)$. However, by Lemma 7.2, we know that $\mathbb{E}(||\tilde{\lambda}_{bt}^j(\xi_t) - \lambda_{bt}^j(\xi_t)||^2) \to 0$, under Assumption 7.5. The same reasoning we used for $\mathbb{E}(s_{in}s_{in}') - \mathbb{E}(s_is_i')$ applies, and since $\lambda_{bt}^j(.)$ is a bounded function, we obtain $\mathbb{E}\left(\tilde{\lambda}_{bt}^j(\xi_t)\tilde{\lambda}_{bt}^j(\xi_t)'\right) \to_{n\to\infty} \mathbb{E}\left(\lambda_{bt}^j(\xi_t)\lambda_{bt}^j(\xi_t)'\right)$.

Therefore $||\Lambda^{bt}||^2 \to_{n\to\infty} \mathrm{tr}\left(\Omega_0^{-1/2}\sum_{j\leqslant T-1}\mathbb{E}\left(\lambda_{bt}^j(\xi_t)\lambda_{bt}^j(\xi_t)'\right)\Omega_0^{-1/2}\right) \leqslant C$. Hence $||\Lambda^{bt}||^2$ is bounded. The same arguments applied to the functions $\lambda_{\mathcal{M}}$, λ_k, as well as to

$$\mathbb{E}\left[\frac{\partial\lambda_{\mathcal{M}}(V)}{\partial v_t}\mathrm{Vec}(\rho_i^{\mathcal{M}})|\xi_{it}\right],\ \mathbb{E}\left[\lambda_{\mathcal{M}}(V)\frac{\partial\mathcal{M}_0(V)}{\partial v_t}|\xi_t\right],\ \mathbb{E}\left[\frac{\partial\lambda_k(V)}{\partial v_t}\rho_i^k|\xi_{it}\right] \text{ and } \mathbb{E}\left[\lambda_k(V)\frac{\partial k_0(V)}{\partial v_t}|\xi_t\right],$$

would imply that $||\Lambda^M||$, $||\Lambda^k||$, $||\Lambda^M(I_{(T-1)^2}\otimes\Theta)(H_t^M - \mathrm{d}P_t^M)||$, and $||\Lambda^k(I_{T-1}\otimes\Theta)(H_t^k - \mathrm{d}P_t^k)||$ are bounded. □

We now know that under Assumption 7.5, Assumption 7.4 (6) holds. We will write Assumption 7.4' for Assumption 7.4 without its condition (6): our previous results imply that under Assumption 7.4' and Assumption 7.5, Equation (42) on $\sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0]$ holds.

7.4.2 Asymptotic distribution of $\hat{\mu}$

We now assemble the arguments of Section 7.2 and 7.4.1. We showed that $\frac{1}{\sqrt{n}}\sum_{i=1}^n[\delta_i\mu_i - \mathbb{E}(\mu_i\delta_i)] + \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_i^\delta\dot{u}_i - \sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0] = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{i,n} + o_{\mathbb{P}}(1)$. Recall that Result 7.1, that is, $\sqrt{n}\big[\mathcal{X}_n(\hat{\mathcal{G}}) - \mathcal{X}_n(\mathcal{G}_0)\big] = \sqrt{n}\mathcal{X}_0^{(G)}[\hat{\mathcal{G}} - \mathcal{G}_0] + o_{\mathbb{P}}(1)$ will guarantee that $\sqrt{n}(\hat{\mu}^\delta - \mathbb{E}(\mu_i\delta_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{i,n} + o_{\mathbb{P}}(1)$.

Let us show that Assumption 7.1 and Result 7.1 hold. We restate here Assumption 7.1 for $\mathcal{G} = ((b_t)_{t\leqslant T}, k, \mathcal{M})$.

Assumption 7.1.

1. *For all $\delta_n = o(1)$, $\sup_{||\mathcal{G}-\mathcal{G}_0||_{\mathcal{H}}\leqslant\delta_n} ||\mathcal{X}_n(\mathcal{G}) - \mathcal{X}(\mathcal{G}) - \mathcal{X}_n(\mathcal{G}^0)|| = o_{\mathbb{P}}(n^{-1/2})$.*

2. *The pathwise derivative of* $\mathcal{X}$ *at* $\mathcal{G}_0$ *evaluated at* $\mathcal{G}-\mathcal{G}_0$, $\mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G}-\mathcal{G}_0]$, *exists in all directions* $[\mathcal{G}-\mathcal{G}_0]$, *and for all* $\mathcal{G}\in\mathcal{H}_{\delta_n}$ *with* $\delta_n = o(1)$, $||\mathcal{X}(\mathcal{G})-\mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G}-\mathcal{G}_0]\,|| \leqslant C||\mathcal{G}-\mathcal{G}_0||^2_{\mathcal{H}}$, *for some constant* $C\geqslant 0$,
3. $||\hat{\mathcal{G}}-\mathcal{G}_0||_{\mathcal{H}} = o_{\mathbb{P}}(n^{-1/4})$.

Condition (1) is a stochastic equicontinuity condition. We follow Section 4 in Chen et al. (2003) (CLVK thereafter) in our choice of the space $\mathcal{H}$, as they establish easy-to-check conditions implying stochastic equicontinuity in some spaces. For $\mathcal{S}_W$ a bounded subset of $\mathbb{R}^k$, we define a function $g:\mathcal{S}_W\mapsto\mathbb{R}$, and $\varrho>0$, the norm $||g||_{\infty,\varrho} = |g|_{\lfloor\varrho\rfloor} + \max_{|r|=\lfloor\varrho\rfloor}\sup_{w\neq w'}\frac{|\partial^r g(w)-\partial^r g(w')|}{||w-w'||^{\varrho-\lfloor\varrho\rfloor}}$. We define $\mathcal{C}^{\varrho}_c(\mathcal{S}_W)$ to be the set of continuous functions $g:\mathcal{S}_W\mapsto\mathbb{R}$ such that $||g||_{\infty,\varrho}\leqslant c$. The set $\mathcal{H}^{\varrho}_{2t,c} = \mathcal{C}^{\varrho}_c(\mathcal{S}_{\xi_t})^{k_2}$ will be the class of vector valued functions taking values in $\mathbb{R}^{k_2}$, each component of which lies in $\mathcal{C}^{\varrho}_c(\mathcal{S}_{\xi_t})$. We recall that the generic functions k and $\mathcal{M}$ are defined on the extended support $\mathcal{S}^{\varsigma}_V$. Hence we define $\mathcal{H}^{\varrho}_{\mathcal{M},c,c'} = \mathcal{C}^{\varrho}_c(\mathcal{S}^{\varsigma}_V)^{(T-1)^2}\cap\{g:\forall w\in\mathcal{S}_W,\,\lambda_{\min}(M_{g(w)})>c'\}$, where $M_{g(w)}$ is the matrix formed by the coefficients of $g(w)$, and $\mathcal{H}^{\varrho}_{MY,c} = \mathcal{C}^{\varrho}_c(\mathcal{S}^{\varsigma}_V)^{T-1}$. Finally, for the entire vector of infinite dimensional parameter $\mathcal{G}$, we define the set $\mathcal{H}^{\varrho}_{c,c'} = \left(\times_{t\leqslant T}\mathcal{H}^{\varrho}_{2t,c}\right)\times\mathcal{H}^{\varrho}_{\mathcal{M},c,c'}\times\mathcal{H}^{\varrho}_{MY,c}$ and take $\mathcal{H}$ to be $\mathcal{H}^{\varrho}_{c,c'}$.

Our choice of $||.||_{\mathcal{H}}$ is justified by Condition (2). The functional $\mathcal{X}$ is a function of $\mathcal{M}$, k and $(b_t)_{t\leqslant T}$, it is an expectation over values at which these functions are evaluated, but where $\mathcal{M}$ and k are composed with $(b_t)_{t\leqslant T}$. These compositions imply, as was clear in the computations, that the linearization will involve the first order partial derivatives of $\mathcal{M}_0$ and k_0. It also implies that the difference between $\mathcal{X}(\mathcal{G})$ and $\mathcal{X}(\mathcal{G})-\mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G}-\mathcal{G}_0]$ can be easily controlled by, among other terms, the distance between first order partial derivatives of these functions. A natural norm is therefore $||\mathcal{G}||_{\mathcal{H}} = \sum_{j=1}^{(T-1)^2}|\mathcal{M}_j-\mathcal{M}_{0,j}|^{\varsigma}_1 + \sum_{j=1}^{T-1}|k_j-k_{0,j}|^{\varsigma}_1 + \sum_{t\leqslant T}\sum_{j=1}^{d_2}||b_{t,j}-b_{0,t,j}||_{\infty}$ where by an abuse of notation $\mathcal{M}_j$ is the j^{th} component of $\mathrm{Vec}(\mathcal{M})$. This norm is our choice of norm in the remainder of this section.

Assumption 7.6. $\varrho > \max(Td_2, d_z+d_1)/2$.

Result 7.4. *Defining* $\mathcal{H}=\mathcal{H}^{\varrho}_{c,c'}$ *and* $||.||_{\mathcal{H}}$ *as described, if Assumption 7.5 (3) and Assumption 7.6 hold, then Assumption 7.1 (1) and (2) hold.*

Proof. We start by showing that the stochastic equicontinuity condition, Condition (1), holds. Lemma 1 of CLVK shows that if $(W_i)_{i=1}^n$ is i.i.d, Assumption 7.1 (1) holds if : (A) the class $\mathcal{F} = \{\chi(W,\mathcal{G}) : \mathcal{G}\in\mathcal{H}^{\varrho}_{c,c'}\}$ is $\mathbb{P}$-Donsker, i.e it satisfies $\int_0^{\infty}\sqrt{\log N_{[]}(\epsilon,\mathcal{F},||.||_{L_2(P)})}d\epsilon<\infty$, where $N_{[]}(\epsilon,\mathcal{F},||.||_{L_2(P)})$ is the covering number with bracketing, and if (B) $\chi(.,\mathcal{G})$ is $L_2(P)$ continuous at $\mathcal{G}_0$, that is, $\mathbb{E}(||\chi(W_i,\mathcal{G})-\chi(W_i,\mathcal{G}_0)||^2)\to 0$ as $||\mathcal{G}-\mathcal{G}_0||_{\mathcal{H}}\to 0$. We now check that each of these conditions is satisfied by our assumptions.

Condition (A): We use j, l to index vectors. As in CLVK, it is enough to prove that $\mathcal{F}_l = \{\chi_l(W, \mathcal{G}) : \mathcal{G} \in \mathcal{H}^{\varrho}_{c,c'}\}$ is $\mathbb{P}$-Donsker for each component l of $\chi(.)$. Recall that $\chi(W_i, \mathcal{G}) = Q_i^{\delta}\, \mathcal{M}\left(\tau\left[(x_{it}^2 - b_t(\xi_{it}))_{t\leqslant T}\right]\right)^{-1} k\left(\tau\left[(x_{it}^2 - b_t(\xi_{it}))_{t\leqslant T}\right]\right)$. We examine $\chi(W_i, \mathcal{G}) - \chi(W_i, \mathcal{G}_0)$ and write, by an abuse of notation and only in this proof, $V_i = (x_{it}^2 - b_t(\xi_{it}))_{t\leqslant T}$ and $V_{0,i} = (x_{it}^2 - b_{0,t}(\xi_{it}))_{t\leqslant T}$. Note that $V_0 = \tau(V_0)$. We decompose

$$\begin{aligned}
\chi(W_i, \mathcal{G}) - \chi(W_i, \mathcal{G}_0) &= Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[k(\tau[V]) - k_0(\tau[V])] \\
&+ Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[\mathcal{M}_0(\tau[V]) - \mathcal{M}(\tau[V])]\mathcal{M}_0(\tau[V])^{-1}k_0(V_0) \\
&+ Q_i^{\delta}\mathcal{M}_0(\tau[V])^{-1}[\mathcal{M}_0(V_0) - \mathcal{M}_0(\tau[V])]\mathcal{M}_0(V_0)^{-1}k_0(V_0) \\
&+ Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[k_0(\tau[V]) - k_0(V_0)].
\end{aligned} \tag{43}$$

Since $(\mathcal{G}, \mathcal{G}_0) \in \mathcal{H}^{\varrho}_{c,c'} \times \mathcal{H}^{\varrho}_{c,c'}$, the norms of each functional and its first order derivatives are bounded. Moreover the derivatives of τ are bounded. This implies that $||\mathcal{M}_0(V_0) - \mathcal{M}_0(\tau[V])|| \leqslant c||V_0 - V||$, and the same result holds for k. Hence, using (43), $|\chi_l(W_i, \mathcal{G}) - \chi_l(W_i, \mathcal{G}_0)| \leqslant ||\chi(W_i, \mathcal{G}) - \chi(W_i, \mathcal{G}_0)|| \leqslant C||Q_i^{\delta}||\,(\sum_{j=1}^{(T-1)^2} |\mathcal{M}_j - \mathcal{M}_{0,j}|_0^{\varsigma} + \sum_{j=1}^{T-1} |k_j - k_{0,j}|_0^{\varsigma} + \sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_{\infty})$, where the constant C depends on c and c'. By Assumption 7.6, $\mathbb{E}(||Q_i^{\delta}||^2) < \infty$ which implies by the proof of Theorem 3 of CLVK that

$$N_{[]}(\epsilon, \mathcal{F}, ||.||_{L_2(P)}) \leqslant N\left(\epsilon/c^Q, \mathcal{C}_c^{\varrho}(\mathcal{S}_V^{\varsigma}), ||.||_{\infty}\right)^{(T-1)^2+T-1} \Pi_{t\leqslant T} N\left(\epsilon/c^Q, \mathcal{C}_c^{\varrho}(\mathcal{S}_{\xi_t}), ||.||_{\infty}\right)^{d_2},$$

where $N(\epsilon, \mathcal{C}_c^{\varrho}(\mathcal{S}_W), ||.||_{\infty})$ denotes the covering number of the class $\mathcal{C}_c^{\varrho}(\mathcal{S}_W)$, and $c^Q = 2[(T-1)^2 + T - 1 + Td_2]\,\mathbb{E}(||Q_i^{\delta}||^2)$, is the size of the brackets constructed in CLVK.

It is known that for $\mathcal{S}_W$ a bounded subset of $\mathbb{R}^k$, $\log N(\epsilon, \mathcal{C}_c^{\varrho}(\mathcal{S}_W), ||.||_{\infty}) \leqslant \epsilon^{-k/\varrho}$. By Assumption 7.6, $\varrho > \max(Td_2, d_z + d_1)/2$, which implies that $\mathcal{F}_j$ is $\mathbb{P}$-Donsker. Therefore, Condition (A) is satisfied.

Condition (B) : By $\mathbb{E}(||Q_i^{\delta}||^2) \leqslant C$ and using once more the decomposition given by (43), $\mathbb{E}(||\chi(W_i, \mathcal{G}) - \chi(W_i, \mathcal{G}_0)||^2) \leqslant C||\mathcal{G} - \mathcal{G}_0||^2_{\mathcal{H}}$ which gives the wanted result.

We now show that Assumption 7.1 (2) holds. This condition is on the remainder of the linearization, $||\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G} - \mathcal{G}_0]\,||$. Note that

$$\begin{aligned}
&\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G} - \mathcal{G}_0] \\
&= \quad \mathbb{E}(Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[k(\tau[V]) - k_0(\tau[V])]) - \mathbb{E}(Q_i^{\delta}\mathcal{M}_0(V_0)^{-1}[k(V_0) - k_0(V_0)]) \\
&\quad + \mathbb{E}(Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[k_0(\tau[V]) - k_0(V_0)]) - \mathbb{E}(Q_i^{\delta}\mathcal{M}_0(V_0)^{-1}\frac{\partial k_0}{\partial V_0}(V_0)[V - V_0]) \\
&\quad + \mathbb{E}(Q_i^{\delta}\mathcal{M}(\tau[V])^{-1}[\mathcal{M}_0(\tau[V]) - \mathcal{M}(\tau[V])]\mathcal{M}_0(\tau[V])^{-1}k_0(V_0)) \\
&\qquad\qquad - \mathbb{E}(Q_i^{\delta}\mathcal{M}_0(V_0)^{-1}\,[\mathcal{M}_0(V_0) - \mathcal{M}(V_0)]\mathcal{M}_0(V_0)^{-1}k_0(V_0)) \\
&\quad + \mathbb{E}(Q_i^{\delta}\mathcal{M}_0(\tau[V])^{-1}[\mathcal{M}_0(V_0) - \mathcal{M}_0(\tau[V])]\mathcal{M}_0(V_0)^{-1}k_0(V_0)
\end{aligned}$$

$$- \mathbb{E}([k_0(V_0)'\mathcal{M}_0(V_0)' \otimes (Q_i^\delta \mathcal{M}_0(V_0)^{-1})] \text{ Vec}(\frac{\partial \mathcal{M}_0}{\partial V_0}(V_0))[V_0 - V]).$$

We use this decomposition and we bound each line separately. We show here how to find upper bounds for the first and second lines, both of which will be less than $||\mathcal{G}-\mathcal{G}_0||^2_{\mathcal{H}}$ up to a multiplicative constant. The upper bounds for the third and forth lines of this decomposition can be obtained in a similar fashion. By the triangular inequality, this will give $||\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G} - \mathcal{G}_0]\,|| \leqslant C||\mathcal{G} - \mathcal{G}_0||^2_{\mathcal{H}}$, as desired. First,

$$\begin{aligned}
&\big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}(V)^{-1}[k(\tau[V]) - k_0(V)]) - \mathbb{E}(Q_i^\delta \mathcal{M}_0(V_0)^{-1}[k(V_0) - k_0(V_0)])\big|\big| \\
&\quad = \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}(\tau[V])^{-1}[\mathcal{M}_0(\tau[V]) - \mathcal{M}(\tau[V])]\mathcal{M}_0(\tau[V])^{-1}[k(\tau[V]) - k_0(V)])\big|\big| \\
&\qquad\qquad + \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}_0(\tau[V])^{-1}[\mathcal{M}_0(V_0) - \mathcal{M}_0(\tau[V])]\mathcal{M}_0(V_0)^{-1}[k(\tau[V]) - k_0(\tau[V])])\big|\big| \\
&\qquad\qquad\qquad + \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}_0(V_0)^{-1}[(k-k_0)(\tau[V]) - (k-k_0)(V_0)])\big|\big| \\
&\leqslant C\mathbb{E}(||Q_i^\delta||)\left((\sum_{j=1}^{(T-1)^2} |\mathcal{M}_j - \mathcal{M}_{0,j}|_0^\varsigma)(\sum_{j=1}^{T-1} |k_j - k_{0,j}||_0^\varsigma) \right. \\
&\qquad\qquad + (\sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_t - b_{0,t}||_\infty)(\sum_{j=1}^{T-1} |k_j - k_{0,j}|_0^\varsigma) \\
&\qquad\qquad\qquad \left. + (\sum_{j=1}^{T-1} |k_j - k_{0,j}|_1^\varsigma)(\sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_\infty) \right) \leqslant C||\mathcal{G} - \mathcal{G}_0||^2_{\mathcal{H}}.
\end{aligned}$$

As for the second line of the decomposition of $\mathcal{X}(\mathcal{G}) - \mathcal{X}^{(G)}(\mathcal{G}_0)[\mathcal{G} - \mathcal{G}_0]$, we write

$$\begin{aligned}
&\big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}(\tau[V])^{-1}[k_0(V) - k_0(V_0)]) - \mathbb{E}(Q_i^\delta \mathcal{M}_0(V_0)^{-1}\frac{\partial k_0}{\partial V_0}(V_0)[V - V_0])\big|\big| \\
&\quad = \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}(\tau[V])^{-1}[\mathcal{M}(V_0) - \mathcal{M}(\tau[V])]\mathcal{M}(V_0)^{-1}[k_0(\tau[V]) - k_0(V_0)])\big|\big| \\
&\qquad\qquad + \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}(V_0)^{-1}[\mathcal{M}_0(V_0) - \mathcal{M}(V_0)]\mathcal{M}_0(V_0)^{-1}[k_0(V) - k_0(V_0)])\big|\big| \\
&\qquad\qquad\qquad + \big|\big|\mathbb{E}(Q_i^\delta \mathcal{M}_0(V_0)^{-1}[k_0(\tau[V]) - k_0(V_0) - \frac{\partial k_0}{\partial V_0}(V_0)[V - V_0]])\big|\big| \\
&\leqslant C\mathbb{E}(||Q_i^\delta||)\left((\sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_\infty)^2 \right. \\
&\qquad\qquad + (\sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_\infty)(\sum_{j=1}^{(T-1)^2} |\mathcal{M}_j - \mathcal{M}_{0,j}|_0^\varsigma) \\
&\qquad\qquad\qquad \left. + (\sum_{t\leqslant T}\sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_\infty)^2 \right) \leqslant C||\mathcal{G} - \mathcal{G}_0||^2_{\mathcal{H}},
\end{aligned}$$

where the inequality for the third term in this equation holds by the Jacobian of τ being the identity matrix when evaluated at V_0 (since $V_0 \in \mathcal{S}_V$) and by the second order derivative of τ being bounded. □

We provided a set of conditions satisfying Conditions (1) and (2) of Assumption 7.1 for our choice of $\mathcal{H}$ and $||.||_{\mathcal{H}}$. Condition (3) is a condition on the convergence rate of the estimators $\hat{b}_t$, $\hat{k}$ and $\hat{\mathcal{M}}$. The rate of convergence of $||\hat{b}_{t,j} - b_{0,t,j}||_\infty$ for all (t,j) is given by Equation (28). The rates of convergence of $|\hat{k}_j - k_{0,j}|_1^\varsigma$ and $|\hat{\mathcal{M}}_j - \mathcal{M}_{0,j}|_1^\varsigma$ are given by Corollary 5.1. As explained in Section 7.1, the conditions required to apply this corollary must be adapted to the extended support. Assumption 7.4 already includes most of these conditions as it specifies an approximation rate of $\mathcal{M}_0$ and k_0 over the extended support and defines the rates b_1, b_2 and b_3 as bounds on sup-norms over the extended support. What we add is a slight modification of Condition (4) as follows.

"There exists γ_1 and β_t^L such that for all $t \leqslant T$, $\sup_{\mathcal{S}_{\xi_t}} ||g^{2t}(\xi_t) - \beta_t^{L\prime} r^L(\xi_t)|| \leqslant CL^{-\gamma_1}$. There exists γ_2, $\pi_{M,st}^K$ and $\pi_{k,t}^K$ such that $|\mathcal{M}_0^\varsigma(.)_{st} - p^K(.)'\pi_{M,st}^K|_1^\varsigma \leqslant CK^{-\gamma_2}$ and $|k_0^\varsigma(.)_t - p^K(.)'\pi_{k,t}^K|_1^\varsigma \leqslant CK^{-\gamma_2}$, for all $1 \leqslant s,t \leqslant T-1$".

This modification is a stronger assumption, changing the approximation rate to be over the $|.|_1$ norm instead of the sup norm. We incorporate it in the set of assumptions 7.4'. We can now state the following result.

Result 7.5. *Under Assumptions 2.1, 2.3, 5.6, 7.4', 7.5, 7.6 and 7.7, assuming moreover that $a_1(L)\Delta_n = o(n^{-1/4})$ and $b_2(K)[K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2]^{1/2} = o(n^{-1/4})$, then*

$$[\hat{\mu}^\delta - \mathbb{E}(\mu_i\delta_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{i,n} + o_{\mathbb{P}}(1).$$

Define $\Phi = \mathbb{P}(\det(\dot{X}_i'\dot{X}_i) > \delta_0)^{-1}$ and $\phi_n = \frac{1}{n}\sum_{i=1}^n \delta_i$. The estimator of the average effect $\mathbb{E}(\mu|\delta)$ is $\hat{\mu} = \frac{1}{\sum_{i=1}^n \delta_i/n}\hat{\mu}^\delta/\phi_n$. We decompose

$$\begin{aligned}\sqrt{n}\phi_n[\hat{\mu} - \mathbb{E}(\mu|\delta)] &= \sqrt{n}[\hat{\mu}^\delta - \mathbb{E}(\mu\delta)] + \sqrt{n}\frac{\mathbb{E}(\mu\delta)}{\Phi}[\Phi - \phi_n], \\ &= \sqrt{n}\ (I_{d_x}; \mathbb{E}(\mu|\delta))\left[\begin{pmatrix}\hat{\mu}^\delta \\ \phi_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\mu\delta) \\ \Phi\end{pmatrix}\right], \end{aligned} \tag{44}$$

where $(I_{d_x}; \mathbb{E}(\mu|\delta))$ is a matrix of size $d_x \times (d_x+1)$. Equation (44) gives the asymptotic distribution of $\hat{\mu} - \mathbb{E}(\mu|\delta)$ once the asymptotic distribution distribution of $\sqrt{n}\left[\begin{pmatrix}\hat{\mu}^\delta \\ \Delta_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\mu\delta) \\ \Delta\end{pmatrix}\right]$ is known.

Assumption 7.7. $\mathbb{E}[||\mu_i - \mathbb{E}(\mu)||^4] < +\infty$, $\mathbb{E}[||Q_i^\delta||^4] < \infty$. *Also for all (X,Z), $\mathbb{E}[||\dot{u}_i||^4|X_i = X, Z_i = Z] \leqslant C$, for all ξ_t, $\mathbb{E}[||v_{it}||^4|\xi_{it} = \xi_t] \leqslant C$.*

We define $\Sigma_0 = \mathrm{Var}\left((s_i', \delta_i)'\right)$. Since $\Omega_0 > 0$, then $\Sigma_0 > 0$.

Result 7.6. *Under Assumptions 2.1, 2.3, 5.6, 7.4', 7.5, 7.6 and 7.7, assuming moreover that $a_1(L)\Delta_n = o(n^{-1/4})$ and $b_2(K)[K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2]^{1/2} = o(n^{-1/4})$,*

$$\sqrt{n}\Sigma_0^{-1/2}\left[\begin{pmatrix}\hat{\mu}^\delta \\ \phi_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\mu\delta) \\ \Phi\end{pmatrix}\right] \to^d \mathcal{N}(0, I_{d_x+1}). \tag{45}$$

Proof. By Result 7.5, $\sqrt{n}\left[\begin{pmatrix}\hat{\mu}^{\delta}\\ \Delta_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\mu\delta)\\ \Delta\end{pmatrix}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}s_{i,n}\\ \delta_i - \Delta\end{pmatrix} + o_{\mathbb{P}}(1)$.

Define $\Sigma_n = \mathrm{Var}((s'_{i,n}, \delta_i - \Delta)')$. We obtain the asymptotic distribution in two steps. We prove first that $\sqrt{n}\Sigma_n\left[\begin{pmatrix}\hat{\mu}^{\delta}\\ \Delta_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\mu\delta)\\ \Delta\end{pmatrix}\right] \to^d \mathcal{N}(0, I_{d_x+1})$. We show in a second step that $\Sigma_n \to \Sigma_0$, which will yield the wanted result. We follow Newey et al. (1999) in proving a Lindeberg condition for $c'\Sigma_n(s'_{i,n}, \delta_i - \Delta)'$ for any constant vector $c \in \mathbb{R}^{d_x+1}$ such that $||c|| = 1$. More precisely if for all such c, $\frac{1}{\sqrt{n}}c'\Sigma_n^{-1/2}\sum_{i=1}^{n}(s'_{i,n}, \delta_i - \Phi)' \to^d \mathcal{N}(0,1)$, this first result will be a consequence the Cramér-Wold theorem. Write $S_{i,n} = c'\Sigma_n(s'_{i,n}, \delta_i - \Phi)'$, then $\mathbb{E}(S_{i,n}) = 0$ and $\mathrm{Var}(S_{i,n}) = 1$. Asymptotic normality is a consequence of the CLT, provided that the Lindeberg condition holds for $S_{i,n}$, i.e, for any $\epsilon > 0$, $\mathbb{E}(S_{i,n}^2\mathbb{1}(|S_{i,n}| > \epsilon\sqrt{n})) \to 0$. Note that by ρ^M and ρ^k bounded under Assumption 7.4, $\mathbb{E}(\dot{u}_i|X_i^1, X_i^2, Z_i) = 0$ and $\mathbb{E}[||\dot{u}_i||^4|X_i = X, Z_i = Z] \leqslant C$ for all (X, Z), $\mathbb{E}[||e_i^k||^4|V_i = V] \leqslant C$ and $\mathbb{E}[||\operatorname{Vec}(e_i^{\mathcal{M}})||^4|V_i = V] \leqslant C$. Fix $\epsilon > 0$. We normalize $\Theta = I_K$ and $\Theta_1 = I_L$, and obtain

$$
\begin{aligned}
n\epsilon^2\mathbb{E}(S_{i,n}^2\mathbb{1}(|S_{i,n}| > \epsilon\sqrt{n})) \;&\leqslant\; \mathbb{E}(S_{i,n}^4\mathbb{1}(|S_{i,n}| > \epsilon\sqrt{n})) \;\leqslant\; \mathbb{E}(S_{i,n}^4),\\
&\leqslant C\Big(\mathbb{E}[||\mu_i - \mathbb{E}(\mu)||^4] + \mathbb{E}[||Q_i^{\delta}\dot{u}_i||^4] + ||\Lambda^M||^4\,\mathbb{E}[||\operatorname{Vec}(e_i^M)\otimes p_i||^4] + ||\Lambda^k||^4\,\mathbb{E}[||e_i^k\otimes p_i||^4]\\
&\quad + ||[\Lambda^M(H_t^M - \mathrm{d}P_t^M) + \Lambda^k(H_t^k - \mathrm{d}P_t^k) + \Lambda^{bt}]||^4\,\mathbb{E}[||v_{it}\otimes r_{it}||^4] + \mathbb{E}[||\delta_i - \Delta||^4]\Big).
\end{aligned}
$$

We can bound $\mathbb{E}(||v_{it}\otimes r_{it}||^4) = \mathbb{E}(||r_{it}||^4||v_{it}^4||) \leqslant C\mathbb{E}(||r_{it}||^4)$ by Assumption 7.7, and $\mathbb{E}(||r_{it}||^4) \leqslant a_1(L)^2\operatorname{tr}(\mathbb{E}(r'_{it}r_{it})) = a_1(L)^2L$. Similarly, by Assumption 7.7, $\mathbb{E}(||e_i^k\otimes p_i||^4) = O(b_1(K)^2K)$ and $\mathbb{E}(||\operatorname{Vec}(e_i^M)\otimes p_i||^4) = O(b_1(K)^2K)$. Therefore, by Result 7.3, $n\epsilon^2\mathbb{E}(S_{i,n}^2\mathbb{1}(|S_{i,n}| > \epsilon\sqrt{n})) = O(b_1(K)^2K + a_1(L)^2L)$.

Assumption 7.4 (7) implies $\Delta Q = o(1)$ and $\Delta Q_1 = o(1)$, in turn implying $\sqrt{K/n}\,b_1(K) \to 0$ and $\sqrt{L/n}\,a_1(L) \to 0$. Therefore the condition $\mathbb{E}(S_{i,n}^2\mathbb{1}(|S_{i,n}| > \epsilon\sqrt{n})) \to 0$ holds.

The second step to obtain (45) requires $\Sigma_n \to \Sigma_0$. This is a consequence of the mean squared convergences obtained in the proof of Result 7.3. □

By a delta method argument, we can use (44) with the previous result to obtain

$$\sqrt{n}\phi_n[\hat{\mu} - \mathbb{E}(\mu|\delta)] \to^d \mathcal{N}\left(0, (I_{d_x}; \mathbb{E}(\mu|\delta))\,\Sigma_0\,(I_{d_x}; \mathbb{E}(\mu|\delta))'\right),$$

hence $\sqrt{n}[\hat{\mu} - \mathbb{E}(\mu|\delta)] \to^d \mathcal{N}\left(0, \Phi^{-2}\,(I_{d_x}; \mathbb{E}(\mu|\delta))\,\Sigma_0\,(I_{d_x}; \mathbb{E}(\mu|\delta))'\right)$. We can now state the main result of this section,

Result 7.7. *Under Assumptions 2.1, 2.3, 5.6, 7.4', 7.5, 7.6 and 7.7, assuming moreover that* $a_1(L)\Delta_n = o(n^{-1/4})$ *and* $b_2(K)[K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2]^{1/2} = o(n^{-1/4})$,

$$\sqrt{n}[\hat{\mu} - \mathbb{E}(\mu|\delta)] \to^d \mathcal{N}(0, \Phi^{-2}\Xi),$$

where $\Xi = \Omega_0 + \mathbb{E}((\delta_i - \Phi)s_i)\mathbb{E}(\mu|\delta)' + \mathbb{E}(\mu|\delta)\mathbb{E}((\delta_i - \Phi)s_i)' + (\Phi - \Phi^2)\mathbb{E}(\mu|\delta)\mathbb{E}(\mu|\delta)'$.

8 Monte Carlo simulations

We explore the properties of our multi-step estimator with Monte Carlo simulations when the model is a specific case of the model studied in the asymptotic analysis, Model (26). More specifically, the data generating process we consider is the following.

$$y_{it} = x_{it}^{1\prime}\mu_i^1 + x_{it}^{2\prime}\mu_i^2 + \underbrace{\sin(3v_{it}) + u_{it}}_{=\epsilon_{it}},\ i = 1..n,\ t \leqslant T.$$

where the random coefficients are drawn according to

$$\mu_i = A\nu_i, \text{ with } A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \text{ and } \nu_i^1 \sim \mathcal{U}[0,1],\ \nu_i^2 \sim \mathcal{U}[0,1],\ \nu^1 \perp\!\!\!\perp \nu^2,$$

and the specification for the covariates, instruments and time-varying disturbances are, for all $t \leqslant T$,

$$\begin{aligned} &\tilde{x}_{it}^1 \sim \mathcal{U}[0,1], \qquad \tilde{z}_{it} \sim \mathcal{U}[0,1], \qquad v_{it} \sim \mathcal{U}[-0.5, 0.5], \qquad X_i^1 \perp\!\!\!\perp Z_i, \\ &x_{it}^1 = 5(\mu_i^1)^{1/4}\,\tilde{x}_{it}^1, \qquad z_{it} = 5(\mu_i^2)^{1/4}\,\tilde{z}_{it}, \\ &x_{it}^2 = (x_{it}^1 + z_{it})^{1/2} + v_{it}. \end{aligned}$$

In this design, the control function is $f_t(V_i) = \sin(3v_{it})$, giving $g_t(V_i) = \sin(3v_{it+1}) - \sin(3v_{it})$. As for the random coefficients, the design implies that μ^1 has support $[0,3]$, $\mathbb{E}(\mu_i^1) = 1.5$, $\text{Var}(\mu_i^1) = 5/12$, and that μ^2 has support $[0,4]$, $\mathbb{E}(\mu_i^2) = 2$ and $\text{Var}(\mu_i^2) = 5/6$. The supports of the regressors are for x_t^1, $[0, 6.58]$ and $[-0.5, 4.2]$ for x_t^2. The heterogeneity is quite substantial in this design. This simulation design imposes the random coefficients and the regressors to covary. To ensure that the condition $\mathbb{E}(||Q_i\dot{u}_i||) < \infty$ holds, we imposed z_{it} and x_{it}^1 to depend multiplicatively on μ_i^1 and μ_i^2 raised to the power $1/4$, following an observation made in Graham and Powell (2012). Although there are, to our knowledge, no exact results on when this condition does or does not hold, this choice should intuitively imply finiteness of the expectation.

We show here the results of $R = 1000$ simulations of two different sample sizes, $n = 1000$ and $n = 2000$. Our choice of sieve approximating functions is a third order multivariate B-spline basis for both estimation of the conditional expectation of x^2 conditional on (x^1, z) and estimation of the functions $\mathcal{M}(.) = \mathbb{E}(M_i|V_i = .)$ and $k(.) = \mathbb{E}(M_i\dot{y}_i|V_i = .)$. The conditional expectation of x^2 is used to construct the generated covariates $\hat{V}_i$. Recall that $g(V) = \mathcal{M}^{-1}(V)k(V)$. For each of the simulation draw r, an estimate $\hat{g}^r$ of the function g is computed. We report in Figure 1 the pointwise average of these esimates $\bar{g}(V) = \sum_{r \leqslant R} \hat{g}^r(V)/R$ as well as the 5^{th} and 95^{th} quantiles $g^5(V)$ and $g^{95}(V)$ for each value of V.

For each draw r, the estimators $\hat{\mu}^{1r}$ and $\hat{\mu}^{2r}$ of the average partial effects $\mathbb{E}(\mu_i^1)$ and $\mathbb{E}(\mu_i^2)$ are computed following the second step of our estimation procedure. That is, we plug in the estimator

of the function g in a sample analog of the formula (11). Figures 2 and 3 are smoothed histograms of the obtained estimators of these average effects. For each coefficient, we used the same scale for different sample sizes but did not use the same sample scale for each coefficient. These plots are compatible with the asymptotic normality result of Section 7. It is noticeable that the variance of the estimator of $\mathbb{E}(\mu_i^2)$, which is the average partial effect of the endogenous variable, is larger than the variance of the estimator of $\mathbb{E}(\mu_i^1)$. However, this shows that even in small samples of size 1000, the estimator for the average partial effects performs relatively well and in particular does not seem to be biased.

As an additional exercise, we compare in Figure 4 the distribution of the estimator constructed in this paper to two different estimators of the impact of x^1 and x^2. The first one is the first-difference instrumental variable, $\hat{\mu}^{FDIV}$, as defined in Wooldridge (2010) Section 11.4. This estimator is consistent under an homogeneity assumption. The second estimator is $\hat{\mu}^{CRC}$, an estimator which is consistent under heterogeneity if there is no time-varying endogeneity. More precisely, $\hat{\mu}^{CRC} = \sum_{i=1}^n Q_i \dot{y}_i / n$: it corresponds to the second step of the estimator studied in this paper. It is visible from the figure that because of the biases coming from either heterogeneity or time-varying endogeneity, the true value of the average effect might not be in the confidence intervals of estimators neglecting either of these features.

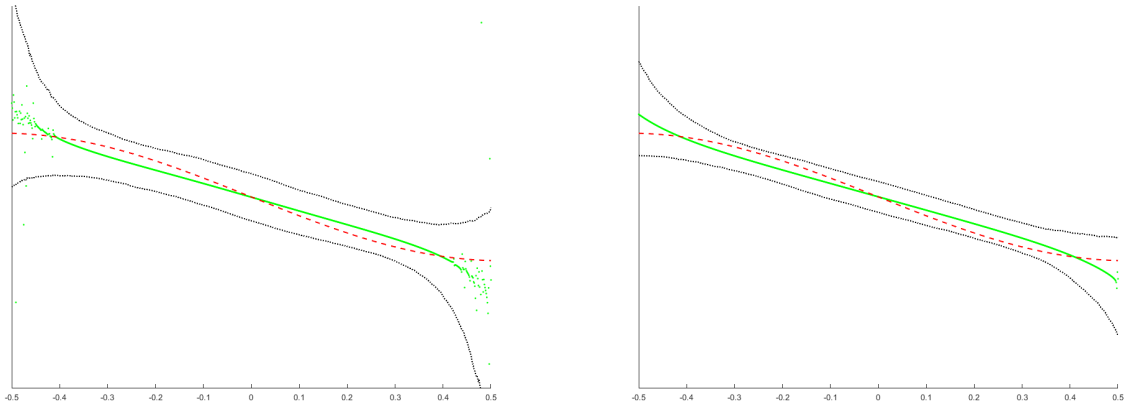

Figure 1: *Estimation of g, $n = 1000$ (left) and $n = 2000$ (right)*
Plot of the true value g (red dashed line), the pointwise average $\bar{g}$ (green line) and the 90 percent MC confidence bands g^5 and g^{95} (black dotted line). These functions are evaluated at $V = (v_1, 0, 0, 0)$ where $v_1 \in [-0.5, 0.5]$ (in this case $g(V) = -\sin(3v_1)$). We note the presence of a boundary effect.

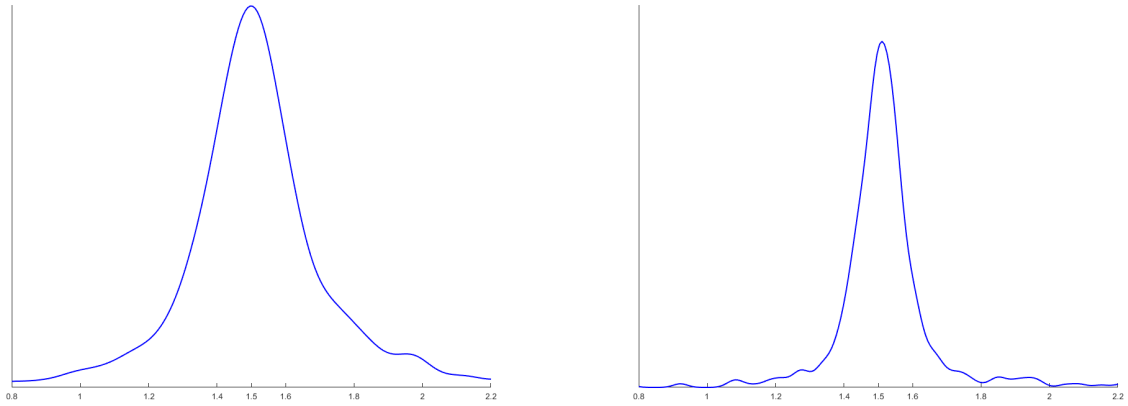

Figure 2: *Estimation of $\mathbb{E}(\mu_i^1)$*
Distribution of $\hat{\mu}^{1r}$, with true value $\mathbb{E}(\mu_i^1) = 1.5$,
when the sample sizes are $n = 1000$ (left) and $n = 2000$ (right).

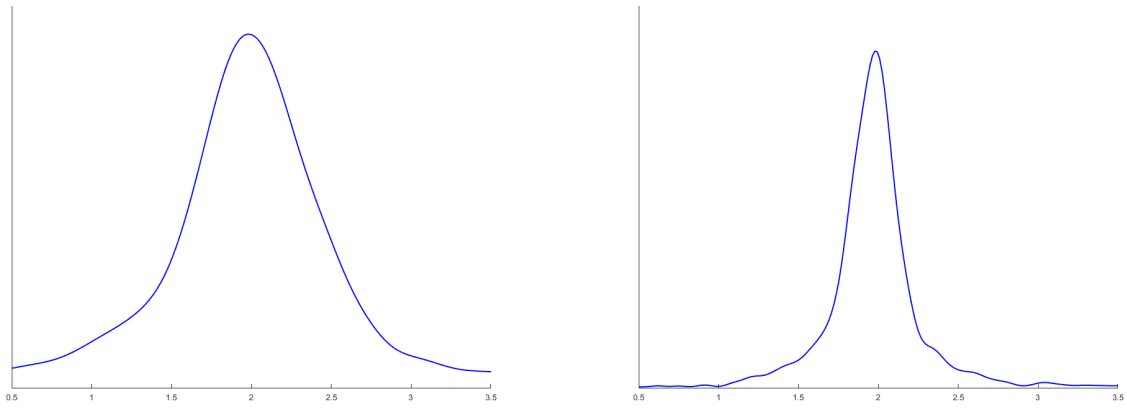

Figure 3: *Estimation of* $\mathbb{E}(\mu_i^2)$
Distribution of $\hat{\mu}^{2r}$*, with true value* $\mathbb{E}(\mu_i^2) = 2$*,*
when the sample sizes are $n = 1000$ *(left) and* $n = 2000$ *(right).*

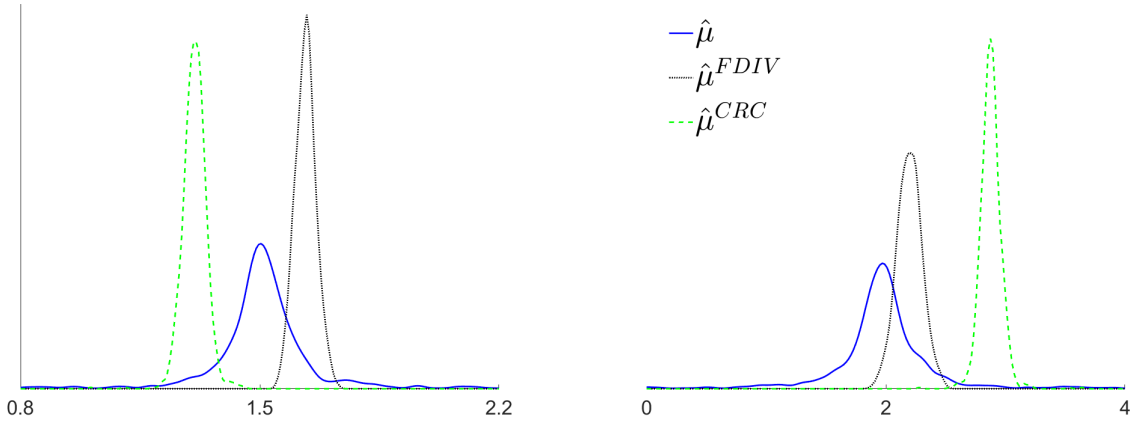

Figure 4: *Comparison of estimators*
For μ^1 *(left) with true value* $\mathbb{E}(\mu_i^1) = 1.5$*, and* μ^2 *(right) with true value* $\mathbb{E}(\mu_i^2) = 2$*,*
Sample size $n = 2000$*.*

9 Empirical Example

As an empirical exercise, we apply our method to a model of labor supply with heterogenous elasticity of intertemporal substitution (EIS). The EIS is an essential object of interest in the study of labor supply as it quantifies how labor supply responds to variations of the wage rate over time. More specifically, the model we consider is

$$\ln h_{it} = \alpha_i + \ln \omega_{it}\, \mu_i + \chi'_{it} b + \epsilon_{it}, \qquad i = 1..n,\ t = 1..T, \tag{46}$$

where h_{it} is the number of annual hours worked, ω_{it} is the hourly wage, χ_{it} is composed of additional demographics. The individual elasticity of intertemporal subsitution μ_i enters the individual utility function, and heterogeneity in preferences may covary with ω_{it} and χ_{it}. This justifies not restricting the joint distribution of these random variables and taking a fixed effect approach to identification and estimation. We allow for the log wage rate variable to be endogenous.

A version of (46) without random coefficient, i.e, where $\mu_i = \mu$ almost surely, is studied in Ziliak (1997) which also focuses on estimation of the EIS. In this paper, the demographics are assumed to satisfy the sequential exogeneity condition $\mathbb{E}(\epsilon_{it}|\chi_{is}) = 0$, for all $s \leqslant t$. On the other hand, the wage variable is considered contemporaneously endogenous due to either nonlinear income taxes, omitted variables or measurement error. The wage is therefore assumed to only satisfy $\mathbb{E}(\epsilon_{it}|\ln \omega_{is}) = 0$ for all $s < t$.

We will estimate $\mathbb{E}(\mu_i)$ under a different set of conditions using the data set used in Ziliak (1997) and the identification results of Section 2. Consider a panel of periods 1 to T, preceded by periods 0, -1, ..., $-\tau$. Define $v_{it} = \ln \omega_{it} - \mathbb{E}(\ln \omega_{it}|\chi_{i1}, \ln \omega_{i0})$ and $V_i = (v_{it})_{t \leqslant T}$. We assume that there exists $(f_t(.))_{t \leqslant T}$ such that

$$\text{for all } t \leqslant T,\ \mathbb{E}(\epsilon_{it}|V_i, \ln \omega_{i0}, \chi_{i1}, \chi_{i0}, .., \chi_{i\,-\tau}) = f_t(V_i), \tag{47}$$

where the normalization condition $\mathbb{E}(f_t(V_i)) = 0$ holds. Here, $(\chi_{i1}, \chi_{i0}, .., \chi_{i-\tau})$ corresponds to the set of additional instruments mentioned in Section 2.4.2. By sequential exogeneity of χ_{it}, its values in s for $2 \leqslant s \leqslant T$ cannot be in this set of instruments.For the same reason, we do not use χ_{it} as insrument to construct the control variables v_{it}. Instead, we use the initial values χ_{i1} and $\ln \omega_{i0}$. This is similar to the approach described in Section 3.4. The conditional expectation equation (47) holds if for instance for all t, $(\ln \omega_{is-1}, \chi_{is}, \chi_{is-1}, .., \chi_{i\,-\tau})_{s \leqslant t} \perp\!\!\!\perp (v_{is'}, \epsilon_{is'})_{s' \geqslant t}$, a condition which would also imply the moment conditions used in Ziliak (1997).

Defining M_i as in Section 2, we need $T \geqslant 3$ for M_i not to be the null matrix with probability 1. Moreover, we use the log wage one period before the beginning of the panel as instrument to construct the control variables. We also use values of χ_{it} drawn before period 1 as instrumental variables to estimate b. These requirements imply that T must be greater than 4.

The dataset constructed in Ziliak (1997) is described in Section 2.1 of the paper. It is a selected sample from the Survey Research Center subsample of the Panel Study of Income Dynamics. It is composed of 532 men aged 22 to 55, married and working at all periods of the panel. We define the demographics χ_{it} as number of children, age and an indicator of bad health. We use a panel of years 1979 to 1982 where period 1 is year 1980, period T is year 1982 and $\tau = 1$. Note that the sample size is not as large as is desirable in semiparametric estimation.

We start by estimating the generated covariates v_{it}, writing

$$\ln \omega_{it} = \gamma_{1t} \ln \omega_{i0} + \gamma'_{2t} \chi_i^{\mathrm{GC}} + v_{it},$$

where χ_i^{GC} includes χ_{i1} and age_{i1}^2. We choose this linear specification with a quadratic in age instead of a fully nonparametric one to avoid the curse of dimensionality which potentially has a strong impact given our small sample size. We then estimate successively the vector b, the functions $g_t(.) = f_{t+1}(.) - f_t(.)$ for $t \leqslant T-1$, and the average partial effect $\mathbb{E}(\mu_i)$. These steps require estimation of conditional expectation functions conditional on V. We choose the same basis of approximating functions of V (power series) and the same number of approximating terms for each of these functions. The exact choice of approximating functions is decided using a leave-one-out cross-validation (CV) criterion. By design, the estimator of $\mathbb{E}(\mu_i)$ depends on the inverse of the matrix function $\mathcal{M}(.) = \mathbb{E}(M_i | V_i = .)$ while it depends linearly on the other conditional expectations. For that reason, we chose as a criterion function the mean square forecast error of the random variable M_i. The set of conditioning variables is (v_{i1}, v_{i2}, v_{i3}), hence the terms that can be included in the sieve basis are v_{it} raised to various powers and interactions of those (in addition to a constant term). We report the CV values for some specifications in Table 1. Our choice will be the power series basis of degree 2.

Terms included	CV values
$(v_{it}, v_{it}^2)_{t \leqslant T}$, $v_{i1}v_{i2}$, $v_{i2}v_{i3}$	252
$(v_{it}, v_{it}^2)_{t \leqslant T}$	262
$(v_{it}, v_{it}^2, v_{it}^3)_{t \leqslant T}$	332
$(v_{it}, v_{it}^2, v_{it}^3)_{t \leqslant T}$, $v_{i2}v_{i3}$	278
$(v_{it})_{t \leqslant T}$, $v_{i1}v_{i2}$, $v_{i2}v_{i3}$	269
$(v_{it})_{t \leqslant T}$	264

Table 1: Cross-validation values

We follow the method developed in Section 2.4.1 to estimate the vector of coefficients b. The set of instruments is $Z_i^{\chi} = (\chi_{i1}, \chi_{i0}, \mathrm{age}_{i1}^2, \mathrm{age}_{i0}^2)$. Defining the differences $\Delta \dot{\ln} h_i$ and $\Delta \dot{\chi}_i$ as in

Section 2.4.1, and their estimators as $\widehat{\Delta \ln \dot{h}_i}$ and $\widehat{\Delta \dot{\chi}_i}$, our estimator of b is

$$\hat{b} = \left(\sum_{i=1}^{n} \widehat{\Delta \dot{\chi}_i}' Z_i \sum_{i=1}^{n} Z_i' Z_i \sum_{i=1}^{n} Z_i' \widehat{\Delta \dot{\chi}_i} \right)^{-1} \left(\sum_{i=1}^{n} \widehat{\Delta \dot{\chi}_i}' Z_i \sum_{i=1}^{n} Z_i' Z_i \sum_{i=1}^{n} Z_i' \widehat{\Delta \ln \dot{h}_i} \right),$$

and we obtain $\hat{b} = (-0.035, 0.219, -0.025)$. Finally we estimate $\mathbb{E}(\mu_i)$ by the two-step approach as explained in the main body of the paper. We first estimate $\mathcal{M}(.)$ and $k(.) = \mathbb{E}(M_i[\dot{\ln} h_i - \dot{\chi}_i' b] \,|V_i = .)$ using a series approximation and plugging in the estimate $\hat{b}$. Using these estimators $\hat{\mathcal{M}}(.)$ and $\hat{k}(.)$, our estimate of g is $\hat{g} = \hat{\mathcal{M}}^{-1}\hat{k}$. The final step to obtain the estimate of the average partial effect, that is, of the average elasticity of intertemporal substitution, entails computing the sample analog of the moment equality $\mathbb{E}(Q_i[\dot{\ln} h_i - \dot{\chi}_i' b - g(V_i)]) = \mathbb{E}(\mu_i)$. This gives $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Q_i[\dot{\ln} h_i - \dot{\chi}_i'\hat{b} - \hat{g}(\hat{V}_i)] = 0.251$. We note that this value is in the range of those reported in Ziliak (1997).

10 Conclusion

In this paper, we studied a correlated random coefficient panel model and relaxed the strict exogeneity condition imposed in the literature to allow for time-varying endogeneity. We proved identification of the average partial effect $\mathbb{E}(\mu_i)$. Moreover, we provided an estimator of $\mathbb{E}(\mu_i | \det(\dot{X}_i'\dot{X}_i) > \delta_0)$, showed its asymptotic normality and computed its asymptotic variance.

We highlight two directions for future research. First, our estimation focuses on $\mathbb{E}(\mu|\delta)$, which depends on a constant δ_0. However δ_0 is arbitrarily fixed in the paper and we do not give directions on how to choose a value when implementing the estimator suggested in this paper. It would be of interest to follow Graham and Powell (2012) and study the asymptotic properties of $\mathbb{E}(\mu_i | \det(\dot{X}_i'\dot{X}_i) > \delta_n)$ with $\delta_n \to 0$. This could give a sense of an optimal choice for δ_n as a function of the sample size n. Note that extending the asymptotic analysis of Graham and Powell (2012) is nontrivial, as our estimation procedure includes an additional step with computation of nonparametric two-step series estimators.

The identification argument required $T > d_x + 1$. This can be quite restrictive, and long enough panels might not be available to identify average partial effects in models with multiple covariates with random coefficients. A second direction for future work would be to relax this condition and obtain identification in the case $T = d_x + 1$, as is done in Graham and Powell (2012).

References

Ai, C., & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, *71*(6), 1795–1843.

Altonji, J. G., & Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, *73*(4), 1053–1102.

Andrews, D. W. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica: Journal of the Econometric Society*, 307–345.

Arellano, M., & Bonhomme, S. [Stéphane]. (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, *79*(3), 987–1020.

Arellano, M., & Bonhomme, S. [Stéphane]. (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, *19*(3), C61–C94.

Bester, C. A., & Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics*, *27*(2), 235–250.

Blundell, R., & Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. In M. Dewatripont, L. P. Hansen, & S. J. Turnovsky (Eds.), *Advances in economics and econometrics: Theory and applications, eighth world congress* (Vol. 2, pp. 312–357). Econometric Society Monographs. Cambridge University Press.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, *60*(3), 567–596.

Chen, X., Hong, H., & Tamer, E. (2005). Measurement error models with auxiliary data. *The Review of Economic Studies*, *72*(2), 343–366.

Chen, X., Hong, H., Tarozzi, A., et al. (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, *36*(2), 808–843.

Chen, X., Linton, O., & Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, *71*(5), 1591–1608.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33–58.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*.

Graham, B. S., Hahn, J., Poirier, A., & Powell, J. L. (2018). A quantile correlated random coefficients panel data model. *Journal of Econometrics*, *206*(2), 305–335.

Graham, B. S., & Powell, J. L. (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, *80*(5), 2105–2152.

Hahn, J., Liao, Z., & Ridder, G. (2018). Nonparametric two-step sieve m estimation and inference. *Econometric Theory*, 1–44.

Hahn, J., & Ridder, G. (2013). Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, *81*(1), 315–340.

Heckman, J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, 974–987.

Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Econometric Society Monographs. Cambridge University Press.

Ichimura, H., & Lee, S. (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics*, *159*(2), 252–266.

Imbens, G. W., & Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, *77*(5), 1481–1512.

Mammen, E., Rothe, C., & Schienle, M. (2016). Semiparametric estimation with generated covariates. *Econometric Theory*, *32*(5), 1140–1177.

Mammen, E., Rothe, C., Schienle, M., et al. (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, *40*(2), 1132–1170.

Masten, M. A., & Torgovitsky, A. (2016). Identification of instrumental variable correlated random coefficients models. *Review of Economics and Statistics*, *98*(5), 1001–1005.

Mundlak, Y. (1978). Models with variable coefficients: Integration and extension. In *Annales de l'insee* (pp. 483–509). JSTOR.

Murtazashvili, I., & Wooldridge, J. M. (2008). Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *Journal of Econometrics*, *142*(1), 539–552.

Murtazashvili, I., & Wooldridge, J. M. (2016). A control function approach to estimating switching regression models with endogenous explanatory variables and endogenous switching. *Journal of Econometrics*, *190*(2), 252–266.

Newey, W. K. (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, *10*(2), 1–21.

Newey, W. K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, *79*(1), 147–168.

Newey, W. K., Powell, J. L., & Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, *67*(3), 565–603.

Wooldridge, J. M. (1997). On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics letters*, *56*(2), 129–133.

Wooldridge, J. M. (2003). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics letters*, *79*(2), 185–191.

Wooldridge, J. M. (2005a). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics*, *87*(2), 385–390.

Wooldridge, J. M. (2005b). Unobserved heterogeneity and estimation of average partial effects. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of thomas rothenberg* (pp. 27–55). Cambridge University Press.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Ziliak, J. P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business & Economic Statistics*, *15*(4), 419–431.