A Correlated Random Coefficient Panel Model with Time-Varying Endogeneity

Louise Laage*

This version : March 14, 2020

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.

*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. I acknowledge financial support from the grants ERC POEMH 337665 and ANR-17-EURE-0010, and from the IAAE travel grant. All errors are mine.
1 Introduction

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

\[ y_{it} = x_{it}' \mu_i + \alpha_i + \epsilon_{it}, \quad i \leq n, \ t \leq T, \]  

where the number of periods \( T \) is fixed and the number of units \( n \) is large. The scalar \( \epsilon_{it} \) is a time-varying disturbance. The impact of the vector \( x_{it} \in \mathbb{R}^{d_x} \) of covariates on the scalar dependent variable \( y_{it} \) is linear in \( \mu_i \), vector-valued time-invariant unobserved heterogeneity. In order to depict situations in which the researcher does not know what drives heterogeneity in the impact of \( x_{it} \), a fixed effect approach is adopted, that is, we do not impose assumptions on the joint distribution of \( (\mu_i, (x_{it})_{t \leq T}) \). This positions (1) in the class of correlated random coefficient (CRC) models: attention is given to recovering properties of the distribution of the unobserved heterogeneity \( \mu \), a question complicated by the correlation between this vector and the regressor. For instance, a linear least squares regression computed with a single cross-section will not consistently estimate \( \mathbb{E}(\mu) \).

Model (1) has been studied in the seminal paper by Chamberlain (1992), and more recently in Arellano and Bonhomme (2012) and Graham and Powell (2012). Under various sets of constraints on the variations over time of the regressor and disturbance, we now know how to identify the conditional mean of the unobserved heterogeneity (Chamberlain (1992) for when \( T > d_x + 1 \) and Graham and Powell (2012) when \( T = d_x + 1 \)) and even the conditional distribution of \( \mu \) (Arellano and Bonhomme (2012)). An essential assumption for these identification argument is a strict exogeneity condition on the regressors. In model (1), this strict exogeneity condition can be written \( \mathbb{E}(\epsilon_{it}|x_{i1}, \ldots, x_{iT}) = 0 \). Note that under the strict exogeneity condition, the covariates \( (x_{it})_{t \leq T} \) can be correlated with the vector of unobservables \( (\alpha_i, \mu'_i, (\epsilon_{it})_{t \leq T}) \) through their correlation with \( (\alpha_i, \mu'_i) \). This correlation can be controlled for by a fixed effect transformation because \( (\alpha_i, \mu'_i) \) is time-invariant. Loosely speaking, this condition implies that the endogeneity of the model can be “captured by a fixed effect”. It prohibits the presence of time-varying omitted variables correlated with the regressors \( x_{it} \) and, as pointed out, e.g., in Arellano and Bonhomme (2012) does not allow for sequentially exogeneous regressors. 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

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

This paper seeks to relax the strict exogeneity condition and to allow for what is called here “time-varying endogeneity”. The parameter of interest will be the average partial effect (APE), as defined in Wooldridge (2005b) as $E_{q_t}[\hat{c}\hat{E}(y_t|x_t, q_t)/\hat{c}x_t|x_t=x]$ 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 $q_t$ and in (1), this average partial effect is equal to $E(\mu)$. A discussion on the difficulty of identifying the APE in panel correlated random coefficients models can be found in Graham and Powell (2012).

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. Then instinctively, if one wants to disentangle the regressors from the unobserved heterogeneity, the instrument must impact variations over time of the regressors conditional on the control variables. An invertibility condition formalizes this intuition. Equipped now with such instruments and control variables, identification of the average partial effect is obtained using a two-step approach. First the individual unobserved heterogeneity is “differenced away” using the individual time variation of the regressors, which identifies nonparametrically the conditional expectation of the disturbances given regressors and control variables, that is, the term controlling the endogeneity of the model. Second, the endogeneity is corrected for using this identified nonparametric function and “between-group” variations will allow us to pin down the average effect. This identification argument is constructive and its structure suggests a natural multi-step estimator, following the identification steps after estimating the control variables. We define the estimator in Section 4.1 and show consistency and asymptotic normality in Sections 4.2 through 4.4. 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 directly applicable to our estimator on the asymptotic distribution of sample moments depending on nonparametric two-step sieve estimators.

The identification argument does not rely on an exact specification of the control variables. Interestingly this approach gives flexibility to the two-step method and we show in Section 3 that identification can be obtained in variations of model (1) with different types of violation of the strict exogeneity condition. A panel random coefficient model with sample selection is such an example, and we also apply the idea in a model such as (1) when sequential exogeneity holds, that is, when the regressors are predetermined.
We review here the literature this paper is connected to. Linear panel models with random coefficients, such as (1), are sometimes referred to as models with individual-specific slope or variable-coefficient models. They are surveyed for example in Wooldridge (2010) and Hsiao (2014). Although we focus on the fixed \( T \) framework, such models have been studied when \( T \) is allowed to grow, see, e.g, Pesaran and Smith (1995). For fixed \( T \), Wooldridge (2005a) shows that consistency of the standard fixed-effects estimator in these models requires 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), papers most closely related to ours. Both papers build upon Chamberlain (1992), which studies efficiency bounds in semiparametric models. As an application, 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, the method developed in Chamberlain (1992) cannot be applied. They develop an alternative identification argument exploiting the subsample of “stayers” with little regressor variation as a first step, and construct an estimator. Other recent papers have analyzed nonseparable panel models under a fixed-effect approach. One such paper is Evdokimov (2010) which studies identification and estimation of a model where the outcome equation is additively separable in a nonparametric function of regressors and scalar time-invariant unobserved heterogeneity, and a residual term. Other papers are Chernozhukov, Fernández-Val, Hahn, and Newey (2013) studying partial identification of average structural function and quantile structural function in nonseparable panel models with time-invariant unobserved heterogeneity, and Hoderlein and White (2012). 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 first stage thus allowing for heterogeneity in the impact of the instrument. They retrieve the conditional APE under the assumption that the random coefficient is independent of the instrument and the regressor. Hoderlein, Holzmann, and Meister (2017) analyze a triangular model with random coefficients in both stages, independent of instruments and exogenous regressors: for such a model they show nonidentification of the distribution of the random coefficients in general and that an independence condition between the random coefficients is required for identification. 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.

Section 2 reviews the model and constructs the main two-step identification argument. Some extensions are provided to ease the burden of the curse of dimensionality. The model describes a very generic form of endogeneity and Section 3 leverages this aspect to obtain identification in related models. In Section 4 an estimator is provided for the APE in the main model. 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 5. Finally, Section 6 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.

2 Model and Identification

The following section sets up the model and the control function approach assumption. 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

For a sample of units indexed by \( i \), for \( i \leq n \), the outcome variable in period \( t \), for \( t \leq T \), 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, \( \epsilon_{it} \) is a time-varying disturbance, and \( \mu_i \) is a time-invariant vector which represents individual unobserved heterogeneity. We consider the case where \( T \) is fixed and \( n \) large, and assume \( T \geq d_x + 2 \). More details on this last condition are in Section 2.2.1. Denoting by \( y_i = (y_{i1}, \ldots, y_{iT})' \) the vector of outcomes of unit \( i \), \( X_i = (x_{i1}, \ldots, x_{iT})' \) the matrix of regressors, and \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})' \) the vector of error terms, we can rewrite (1) as

\[
y_i = X_i \mu_i + \alpha_i 1_T + \epsilon_i \text{ where } 1_T \text{ is the vector of size } T \text{ composed of ones.}
\]

The parameters of interest we focus on are the average effects \( \mathbb{E}(\mu_i) \) and \( \mathbb{E}(\alpha_i) \). A standard assumption in the panel correlated random coefficient literature is strict exogeneity of the regressors, which in Model (1) takes the form \( \mathbb{E}(\epsilon_{it}|X_i, \alpha_i, \mu_i) = 0 \). However as pointed out in Section 1 this assumption does not allow for the presence of time-varying omitted variables. We seek to relax the strict exogeneity condition. We will assume the availability of instrumental variables \( z_{it} \in \mathbb{R}^{d_x} \) 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 \( x_{it} = (x_{it,d})_{d \leq d_x} \).

**Assumption 2.1.**

1. \((X_i, Z_i, \epsilon_i, \mu_i, \alpha_i)\) is i.i.d. accross \( i \), and \( \mathbb{P}(\epsilon_{it}) = 0 \),

2. For each \( t \leq T \), there exists an identified function \( C_t \) such that, defining \( v_{it} = C_t(x_{it}, z_{it}) \in \mathbb{R}^{d_x} \),

\[
\mathbb{E}(\epsilon_{it}|x_{i1}, \ldots x_{iT}, v_{i1}, \ldots v_{iT}) = f_t(v_{i1}, \ldots v_{iT}).
\]

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, Powell, and Vella (1999)\(^2\). Define \( V_i = (v_{i1}', \ldots v_{iT}')' \). If for all \( t \leq T \), 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 \( t \leq T \). 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{equation}
\begin{aligned}
y_{it} &= x_{it}^1 \mu_1^1 + x_{it}^2 \mu_2^1 + \epsilon_{it}, \\
x_{it}^2 &= E(x_{it}^2|x_{it}, z_{it}) + v_{it}, \quad E(\epsilon_{it}|V_i, X_i) = f(V_i),
\end{aligned}
\end{equation}
where \(x_{it}^1 \in \mathbb{R}^{d_1}\) is a vector of exogenous regressors and \(x_{it}^2 \in \mathbb{R}^{d_2}\) a vector of potentially endogenous regressors. We can rewrite (2) as (1) taking time invariance of \(\alpha\) separately identifiable. Our identification procedure will indeed prove this fact. Exploiting the random variable \(f_{\alpha}\) \(i\) is random variable correlated with the regressors. The random variable \(f_{\alpha} \) is well-known choice of control variable studied in Blundell and Powell (2003) and Imbens and Newey (2009) is the scalar random variable \(v_{it} = F_{x_{it}^2|x_{it}^1, z_{it}}(x_{it}^2|x_{it}^1, x_{it}^1, z_{it})\).

**Note 2.1.** Unlike this particular case, Model (1) and our general definition of the control variables as \(v_{it} = C_{t}(x_{it}, z_{it})\) in Assumption 2.1 do not make explicit which of the regressors are endogenous. Nor does the condition \(E(\epsilon_{it} | X_i, V_i) = f_{i}(V_i)\). 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 general identification argument applies.

### 2.1.1 Time differencing

Define \(u_{it} = \epsilon_{it} - f_{t}(V_i)\) and \(u_i = (u_{i1}, \ldots, u_{iT})'\). By Assumption 2.1 \(E(u_i|X_i, V_i) = 0\). We also write \(f_{t}(V_i) = (f_{1}(V_i), \ldots, f_{T}(V_i))'\), 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
\[ y_{it} = x_{it}^1 \mu_i + \alpha_i + f_{t}(V_i) + u_{it}. \]

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, \(\alpha_i\), which is time invariant but varies across individuals. Second, \(f_{t}(V_i)\), which varies with both \(i\) and \(t \leq T\). Thus, \(f_{t}(V_i)\) and \(E(\alpha_i|V_i)\) are not separately identifiable. However, we normalized \(E(f(V_i)) = 0\), which should intuitively guarantee that \(f(V_i)\) and \(E(\alpha_i)\) are separately identifiable. Our identification procedure will indeed prove this fact. Exploiting the time invariance of \(\alpha_i\), we take time differences to eliminate this term. We will then later obtain identification of \(E(\alpha_i)\) using \(E(f(V_i)) = 0\).

From now on, we therefore look at a first-differencing transformation of the model, that is, for \(t \leq T - 1\),
\[ y_{it+1} - y_{it} = [x_{it+1} - x_{it}]' \mu_i + f_{t+1}(V_i) - f_{t}(V_i) + u_{it+1} - u_{it}. \]
\[ \dot{y}_{it} = \dot{x}_{it}' \mu_i + g_t(V_i) + \dot{u}_{it}, \]  

(3)

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

We write the model in vector form, defining \( \dot{X}_i = (\dot{x}_{i1}, \ldots, \dot{x}_{iT-1})' \) a \((T-1) \times d_x\) matrices, and the \((T-1) \times 1\) vectors \( \dot{y}_i = (\dot{y}_{i1}, \ldots, \dot{y}_{iT-1})' \), \( g(V_i) = (g_1(V_i), \ldots, g_{T-1}(V_i))' \) and \( \dot{u}_i = (\dot{u}_{i1}, \ldots, \dot{u}_{iT-1})' \). Equation (3) can then be rewritten as

\[ \dot{y}_i = \dot{X}_i \mu_i + g(V_i) + \dot{u}_i, \]

(4)

with by assumption, \( \mathbb{E}(\dot{u}_i | X_i, V_i) = 0 \).

**Note 2.2.** Having \( \epsilon_{it} = f_t(V_i) + u_{it} \) with \( \mathbb{E}(\dot{u}_i | X_i, V_i) = 0 \) does not imply that the control variables are strictly exogenous. Indeed, \( f_t(V_i) \) is allowed to depend on each of the \( v_{is}, s \leq T \). In Section 3, we take advantage of this property of Assumption 2.1[2] and obtain identification in a class of models similar to (1) without contemporaneous endogeneity but where only sequential exogeneity is imposed.

### 2.2 Identification

#### 2.2.1 Two-step identification

We now introduce two matrices commonly used in the panel CRC literature. The first one is the \((T-1) \times (T-1)\) matrix \( M_i = I_{T-1} - \dot{X}_i (\dot{X}_i' \dot{X}_i)^{-1} \dot{X}_i' \), 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 \( \dot{X}_i \dot{X}_i^+ \dot{X}_i = \dot{X}_i \)). \( M_i \) is a projection matrix projecting onto the space orthogonal to the columns of \( \dot{X}_i \). And the second one, defined only if \( \dot{X}_i \) has full column rank, is the \( d_x \times (T-1) \) matrix \( Q_i = (\dot{X}_i' \dot{X}_i)^{-1} \dot{X}_i' \). By definition \( M_i \dot{X}_i \mu_i = 0 \) and \( Q_i \dot{X}_i \mu_i = \mu_i \).

Before going further into the identification argument, we mention here some of the limitations of using the matrices \( M_i \) and \( Q_i \). First, when \( \dot{X}_i \) has full column rank, that is, when \( Q_i \) is defined, if \( T = d_x + 1 \) then \( M_i = 0 \) there is no residual variation to identify common parameters such as the function \( g \). Therefore as in Arellano and Bonhomme (2012), we study panels where \( T \geq d_x + 2 \). Second, the norm of the matrix \( Q_i \) can be very large. Indeed the norm of \( q(X) = (\dot{X}' \dot{X})^{-1} \dot{X} \) goes to infinity as \( \det(\dot{X}' \dot{X}) \) approaches 0. In particular, the norm of \( q(X) \) is not necessarily bounded when \( X \) lies in a compact subset of the space of matrices of size \( T \times k_x \). The identification argument will involve expectations of the product of the vector of outcome variables by \( Q_i \) but this will be properly defined only if \( \mathbb{E}(|Q_i \dot{u}_i|) \) converges. Given the properties of \( Q_i \) that we highlighted, this is a strong condition. This issue is discussed in more detail in Graham and Powell (2012), which studies Model (1) when \( T = d_x + 1 \). The identification method in this paper will impose positive density of “stayers”, that is of individuals with non full column rank \( X_i \). The stayers will be leveraged for
identification of their common parameters, yet under this condition \( \mathbb{E}(||Q_i \tilde{u}_i||) < \infty \) is unlikely to hold. They thus provide an alternative closed-form expression for the average effect as the limit of a conditional expectation conditional on \( \det(\hat{X}_i) > h \), limit taken as \( h \to 0 \). As they point out, even when \( T = d_x + 1 \) and \( \mathbb{E}(||Q_i \tilde{u}_i||) < \infty \) as is the case here, for similar reasons the semiparametric variance bound for \( \mathbb{E}(\mu_i) \) computed by Chamberlain (1992) might not be bounded. That is, \( \mathbb{E}(\mu_i) \) is not regularly identified. Their limit closed-form equation can be instead used to identify the average effect.

We acknowledge these issues and for the sake of simplicity will assume that the needed moments are finite, noting that the identification strategy used in Graham and Powell (2012) could be extended to \( T = d_x + 1 \) and infinite moments. We will show that \( \mathbb{E}(\mu_i | \det(\hat{X}_i) > \delta) \) is also identified under standard assumptions and will actually suggest an estimator for this parameter, the asymptotic properties of which are studied in more details using these standard assumptions. This is also the object estimated in Arellano and Bonhomme (2012). Note that whether or not \( \hat{X}_i' \hat{X}_i \) is invertible, \( M_i \) is an orthogonal projection matrix. This implies that the function \( M(X) = I_{T-1} - \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}' \) is bounded.

**Assumption 2.2.**
1. \( \hat{X}_i \) is of full column rank with probability 1,
2. \( \mathbb{E}(||\tilde{u}_i||) < \infty, \mathbb{E}(||Q_i \tilde{u}_i||) < \infty, \mathbb{E}(||Q_i g(V_i)||) < \infty, \) and \( \mathbb{E}(||\mu_i, \alpha_i||) < \infty \).

With expectations now properly defined, since \( M_i \) and \( Q_i \) are function of \( \hat{X}_i \), we use Assumption 2.1 (2) to obtain

\[
M_i \tilde{y}_i = M_i g(V_i) + M_i \tilde{u}_i, \quad \mathbb{E}(M_i \tilde{u}_i | X_i, V_i) = 0, \tag{5}
\]
\[
Q_i \tilde{y}_i = \mu_i + Q_i g(V_i) + Q_i \tilde{u}_i, \quad \mathbb{E}(Q_i \tilde{u}_i | X_i, V_i) = 0. \tag{6}
\]

Equation (5) is a within-group transformation that allows us to separate \( g \) from \( \mu_i \) and to identify it, while equation (6) isolates \( \mu_i \) from \( \hat{X}_i \) and uses the knowledge of \( g \) to identify \( \mathbb{E}(\mu) \) by taking expectation.

**2.2.2 Identification of \( g_t(.) \)**

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

\[
\mathbb{E}(M_i \tilde{y}_i | X_i, V_i) = M_i g(V_i), \tag{7}
\]

since \( M_i \) is function of \( X_i \). 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 because \( M_i \) is a projection
matrix, it is singular. It is therefore not possible to identify \( g(V_i) \) directly using (7), despite \( M_i \) being observed.

Instead of using \( \mathbb{E}(M_i \hat{y}_i | X_i, V_i) \), we focus on \( k(V) = \mathbb{E}(M \hat{y}_i | V_i = V) \) which satisfies

\[
\mathbb{E}(M_i \hat{y}_i | V_i = V) = \mathbb{E}(M_i | V_i = V) g(V) = \mathcal{M}(V) g(V),
\]

where we write \( \mathcal{M}(V) = \mathbb{E}(M_i | V_i = V) = \mathbb{E}(M(X_i) | V_i = V) = \mathbb{E}(I - \hat{X}_i (\hat{X}_i \hat{X}_i)^{-1} \hat{X}_i | V_i = V). \)

If \( \mathcal{M}(V) \) is invertible for a given value \( V \) on the support 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

\[
g(V_i) = \mathcal{M}(V_i)^{-1} \mathbb{E}(M_i \hat{y}_i | V_i), \quad \mathbb{P}_V \text{ a.s.}
\]

Intuitively, Assumption 2.3 precludes \( g(V_i) \) from being of the form \( \hat{X}_i \beta_i \) and thus from distorting a proper identification of \( \mathbb{E}(\mu_i) \). Indeed, \( g(V_i) = \hat{X}_i \beta_i \Rightarrow M_i g(V_i) = 0 \Rightarrow \mathcal{M}(V_i) g(V_i) = 0 \Rightarrow g(V_i) = 0 \) by invertibility of \( \mathcal{M}(V_i) \), which can hold only with probability 0 since \( \hat{X}_i \) is of full column rank with probability 1. This means that the term \( \hat{X}_i \mu_i \) is separately identifiable from \( g(V_i) \) by Assumption 2.3.

**Note 2.3.** Instead of taking time differences, one could define \( a(V_i) = \mathbb{E}(\alpha_i | V_i), \quad \hat{M}_i = I_T - X_i (X_i' X_i)^{-1} X_i' \) and \( \hat{\mathcal{M}}(V) = \mathbb{E}(...). \)

Equation (8) becomes \( \mathbb{E}(M_i \hat{y}_i | V_i) = \hat{\mathcal{M}}(V_i) [f(V_i) + a(V_i)] + \mathbb{E}(M_i [\alpha_i - a(V_i)]) \), where the second term on the RHS is a priori nonzero. Hence one cannot use the above explained method to identify \( f + a \) and must exploit the time invariance of \( \alpha_i \).

### 2.2.3 Identification of average effects

**Average partial effect** \( \mathbb{E}(\mu) \): Under Assumption 2.1, 2.2, and 2.3, \( g_i(.) \) is identified for \( t \leq T - 1 \) and the matrix \( Q_i \) is well-defined with probability 1. Equation (6) implies

\[
\mu_i = Q_i \hat{y}_i - Q_i g(V_i) - Q_i \hat{u}_i,
\]

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

\[
\mathbb{E}(\mu_i) = \mathbb{E}(Q_i \hat{y}_i - Q_i g(V_i)),
\]

which identifies \( \mathbb{E}(\mu) \) since all elements on the right hand side are observed.
Result 2.1. Under Assumptions 2.1, 2.2 and 2.3, the average effect $E(\mu)$ is identified.

As mentioned in Section 2.2.1, one might be worried that the conditions $E(||Q_i\hat{u}_i||) < \infty$ and $E(||Q_i g(V_i)||) < \infty$ 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 $\delta_i = 1(\det(X_i'X_i) > \delta_0)$ and $Q_i^\delta = \delta_i Q_i$. Then

$$E(\mu|\delta) = \frac{E(\delta_i Q_i\hat{y}_i - \delta_i Q_i g(V_i))}{P(\det(X_i'X_i) > \delta_0)} = \frac{E(Q_i^\delta\hat{y}_i - Q_i^\delta g(V_i))}{P(\det(X_i'X_i) > \delta_0)},$$

(12)

which identifies $E(\mu|\delta)$ since all the terms on the right hand side of (12) are identified. The required conditions are $E(||Q_i^\delta\hat{u}_i||) < \infty$ and $E(||Q_i^\delta g(V_i)||) < \infty$, which one can show are satisfied if for instance $X_i$ has bounded support, $E(||\hat{u}_i||) < \infty$ and $E(||g(V_i)||) < \infty$.

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

$$y_{i1} - x'_{i1} \mu_i = y_{i1} - x'_{i1} [Q_i\hat{y}_i - Q_i g(V_i) - Q_i \hat{u}_i].$$

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

$$E(\alpha_i) = E(y_{i1} - x'_{i1} [Q_i\hat{y}_i - Q_i g(V_i)]),$$

where the right hand side is identified or observed. This identifies $E(\alpha_i)$.

**Note 2.4.** The focus of this paper being on allowing for time-varying endogeneity in correlated random coefficient panel model, the parameter of interest was chosen to be the average effect $E(\mu)$. However more properties of the unobserved heterogeneity can be obtained as is shown in Arellano and Bonhomme (2012). This will be explained in Section 2.4.3 but note that by $E(\hat{u}_i|X_i) = 0$, $E(\mu_i|X_i)$ is also identified with $E(\mu_i|X_i) = E(Q_i\hat{y}_i - Q_i g(V_i)|X_i)$.

**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 study in this paper are $(\mu_i, \alpha_i, (\epsilon_{it})_{t \leq T})$, and an exogenous shift can be a variation $\Delta_i$ independent of $(\alpha_i, \mu_i, (\epsilon_{it})_{t \leq T})$ in which case the average impact of the policy is $E(\mu_i)E(\Delta_i)$. However it might be of interest to consider policy interventions where the variation $\Delta$ is correlated with $x_{it}$, hence correlated with $(\mu_i, \alpha_i)$ while exogenous in the sense that it is independent of $(\epsilon_{it})_{t \leq T}$. For example, consider an exogenous intervention that shifts $x_{it}$ to $l(x_{it})$. The average outcome after this intervention is $E(l(x_{it})'\mu_i + \alpha_i + \epsilon_{it})$ and depends on
the joint distribution of \((\mu_i, x_{it})\) where \(\mu_i\) is unobservable. It could potentially be challenging to obtain since we left this joint distribution unrestricted but because Equation (10) expresses \(\mu_i\) as a function of the primitives it can be once more plugged in to recover average effects. The change in expected outcome is

\[
\mathbb{E}(l(x_{it})'\mu_i + \alpha_i + \epsilon_{it} - [x_{it}'\mu_i + \alpha_i + \epsilon_{it}]) = \mathbb{E}([l(x_{it}) - x_{it}]'\mu_i),
\]

\[
= \mathbb{E}([l(x_{it}) - x_{it}]'(Q_iy_i - Q_i g(V_i) - Q_i \hat{u}_i)),
\]

\[
= \mathbb{E}([l(x_{it}) - x_{it}]'(Q_iy_i - Q_i g(V_i))),
\]

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 \(\text{Int}()\) to refer to the interior of a set, \(\mathcal{S}_{W|V}\) refers to the support\(^3\) of the random variable \(W\) conditional on the variable \(V\) taking the value \(\tilde{V}\). \(\text{Rank}(A)\) refers to the column rank of a matrix \(A\), \(\text{GL}_{T-1}(\mathbb{R})\) is the space of matrices of size \((T - 1) \times (T - 1)\) that are invertible, and \(p_{W|V}(\cdot|\cdot)\) 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 \(A\) conditional on \(B = b\).

#### 2.3.1 High-level Condition

Before giving a formal statement, a brief intuition is given here on what properties of the random variables are used to show invertibility of \(\mathcal{M}(V)\). Recall that \(\mathcal{M}(V) = \mathbb{E}(M|V) = \mathbb{E}(M(X)|V)\) where \(M(X) = I_{T-1} - \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}'\) is an orthogonal projection matrix. \(M(X)\) projects onto the space orthogonal to the \(k_x\) columns of \(\hat{X}\), where each column \(k\) corresponds to the \(T - 1\) values of the scalar r.v \(\hat{x}_{i,k}\). 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 \(V \in \mathcal{S}_V\),

\[\mathcal{M}(V) \notin \text{GL}_{T-1}(\mathbb{R}) \iff \exists c \in \mathbb{R}^{T-1}\setminus\{0\} \mathcal{M}(V)c = 0\]

\[\Rightarrow \exists c \in \mathbb{R}^{T-1}\setminus\{0\}, c' \mathcal{M}(V)c = 0\]

\(^3\)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.
\[ \Rightarrow \exists c \in \mathbb{R}^{T-1} \setminus \{0\}, \quad \mathbb{E}(c' M(X)' M(X) c | V) = 0 \]
\[ \Rightarrow \exists c \in \mathbb{R}^{T-1} \setminus \{0\}, \quad \mathbb{E}(|M(X) c|^2 | V) = 0, \]
and since \(|M(X) c|^2\) is a positive function, this implies that \(|M(X) c| = 0\) with probability 1 on the support of \(\hat{X}\) conditional on \(V\), i.e. \(\mathbb{P}_{\hat{X}|V}\)-a.s. That is,
\[ M(X) c = 0, \quad \mathbb{P}_{\hat{X}|V}\)-a.s. \]

This result is very useful here, it implies that if a sum of orthogonal projections of a given vector \(c\) is zero, then each of the orthogonal projections of \(c\) is zero. Thus the goal of any assumption implying invertibility of \(M(V)\) is to make sure that as the value of \(\hat{X}\) varies on \(S_{\hat{X}|V}\) the null spaces of the matrices \(M(X)\) have trivial intersection. The nullspace of \(M(X)\) being the space spanned by the columns of \(\hat{X}\), intuitively this requirement will be satisfied \(S_{\hat{X}|V}\) contains sufficiently different draws of \(\hat{X}\). The following result provides one way to formalize this explanation.

**Assumption 2.4.** The following holds almost surely in \(V\).

1. \(\text{Int} \left( S_{\hat{X}|V} \right) \neq \emptyset\),
2. There exists a basis \(e = (e_1, \ldots, e_{T-1})\) of \(\mathbb{R}^{T-1}\) and for each \(t \leq T-1\), there exists \(\hat{X}^{(t)} \in \text{Int} \left( S_{\hat{X}|V} \right)\) such that \(p_{\hat{X}|V}(\hat{X}^{(t)} | V) > 0\), \(\text{Rank}(\hat{X}^{(t)}) = d_x\) and \(\hat{X}^{(t)'} e_t = 0\).

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

**Note 2.5.** Some comments are necessary on the conditions imposed in Assumption 2.4.

1. If \(e\) is the canonical basis of \(\mathbb{R}^{T-1}\), that is, \(e_1 = (1, 0, \ldots, 0)'\), \ldots and \(e_{T-1} = (0, \ldots, 0, 1)'\), then \(\hat{X}' d_t = 0\) is equivalent to \(x_{i t} = x_{i t+1} = \bar{x}\). In this case, because
\[ \mathbb{E}(y_{i t+1} - y_{i t} | x_{i t} = x_{i t+1} = \bar{x}, V = V) = f_{t+1}(V) - f_t(V) = g_t(V). \]
g_t is trivially identified. Note that the condition \(p_{\hat{X}|V}(\hat{X}^{(t)} | V) > 0\) in Assumption 2.4 (2) is similar to a condition in Evdokimov (2010), which studies a panel nonparametric 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 \((T - 1, \text{exactly})\) and we do not need to know these values of \(\bar{x}\) to obtain identification, we simply need their existence.

2. The existence of a draw \(\hat{X}^{(t)}\) such that \(\hat{X}^{(t)}\) is of full column rank \(d_x\) and \(\hat{X}^{(t)} e_t = 0\) implicitly requires the number of columns of \(\hat{X}^{(t)}\) to be lower than \(T - 2\), i.e. \(d_x \leq T - 2\) as we imposed.

\[ ^4 \text{Although it is not explicit in the notation, the vector } c \text{ depends on the draw of } V. \]
Moreover, note that if we do not consider a first-differencing transformation of the model and use a vector form of \((\mathbf{1})\), then the column \((1, \ldots, 1)'\) is always included in the matrix \(M_i\), which would imply that the matrix \(\mathcal{M}(V_i)\) is singular.

3. The statement here is about invertibility of \(\mathcal{M}(V)\), almost everywhere on the support of \(V\).

In Section 4.1 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 \(\lambda_{\text{min}}(\mathcal{M}(V))\) is bounded away from zero uniformly in \(V\).

### 2.3.2 Examples

Assumption 2.3 is implicitly an assumption on the first stage used to construct the control variables \(V\), i.e, it is an assumption on the role of the instruments \((z_t)_{t \leq T}\). To better understand it, some particular cases are discussed here as well as some extensions.

**Example 1:** We focus here on the model \((2)\) when there are no exogenous regressors, i.e, \(d_1 = 0\) and where the regressor is scalar, i.e, \(d_2 = 1\). We write

\[
x_{it} = b_t(z_{it}) + v_{it}, \quad \mathbb{E}(v_{it}|z_{it}) = 0. \tag{13}
\]

For simplicity we also consider the case where the instrument \(z_{it}\) is real-valued. In this case, we show that for any value of \(V\) and as \(Z\) varies on \(\mathcal{S}_Z|V\), it is sufficient that two draws of \(\tilde{X}\) are non collinear. Since \(\tilde{X}\) is the column vector of the values of \(\dot{x}\) over time, the noncollinearity is a condition requiring the instrument to have an impact on the variations over time of \(\dot{x}\). This is also visible in the following assumption.

**Assumption 2.5.**

1. For all \(t \leq T\), \(b_t\) is a continuously differentiable function,
2. Almost surely in \(V\):
   \[
   \text{Int} \left( \mathcal{S}_Z|V \right) \neq \emptyset. \quad \text{Plus, there exists } Z^V \in \text{Int} \left( \mathcal{S}_Z|V \right) \text{ such that } p_{Z|V}(Z^V|V) > 0 \text{ and for some } t \leq T - 1, \frac{db_t(Z_t^V)}{dz_t} \neq 0.
   \]

**Result 2.3.** If \((13)\) holds and Assumption 2.5 is satisfied, \(\mathcal{M}(V)\) is nonsingular \(\mathbb{P}_V\) a.s.

**Example 2:** Consider again Model \((2)\) but where now the first stage is linear. That is, \(x_{it} \in \mathbb{R}^{d_x}\), \(z \in \mathbb{R}^{d_z}\) and

\[
x_{it} = A_t z_{it} + v_{it}, \quad \mathbb{E}(v_{it}|z_{it}) = 0, \tag{14}
\]

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 says that almost surely in \(V\) and for all \(t \leq T - 1\), there exists \(X^{(t)} \in \mathcal{S}_{X|V}\)
such that \( \epsilon^{(t)}_t \dot{X}^{(t)} = 0 \) or equivalently such that \( \dot{x}^{(t)}_t = 0 \). By (14), for a fixed value \( V \) of the vector of control variables, this imposes for all \( t \leq T - 1 \) the existence of \( Z^{(t)} \in S_{Z|V} \) such that \( A_{t+1}z^{(t)}_{t+1} - A_t z^{(t)}_t + \dot{v}_t = 0 \). It is visible that this is a condition on the dynamics of \((A_t z_{t})_{t \leq T}\) conditional on \( V \), which can be translated into conditions on the dynamics of the instrument depending on the matrices \((A_t)_{t \leq T}\).

For instance, this condition will be satisfied if \( A_{t+1} \) has full row rank (which implies \( d_z \geq d_x \)) and if \( A'_{t+1}(A_{t+1}A'_{t+1})^{-1}(A_{t} z_{t} - \dot{v}_{t}) \in S_{z_{t+1}|V,z_{t}} \) for some \( z_{t} \in S_{z_{t}|V} \). The support of \( z_{t+1} \) conditional on \( V \) and \( z_{t} \) needs to be large enough. If for instance the matrix \( A_{t} = A \) does not vary over time and is of full row rank, it is the support of \( z_{t+1} - z_{t}|V \) which must be large enough as this condition becomes \( A'(AA')^{-1}\dot{v}_{t} \in S_{z_{t}-z_{t+1}|V} \).

Another example is if the instrument does not vary over time, \( z_{t} = z_{1} \forall t \leq T \). Condition (2) of Assumption 2.4 would be satisfied in this case if \( \forall t \leq T, A_{t+1} - A_{t} \) is of full row rank and \((A_{t} - A_{t+1})'[(A_{t+1} - A_{t})(A_{t+1} - A_{t})']^{-1}\dot{v}_{t} \in S_{z_{t}|V} \). Observe that having \( z_{t} \) constant over time transfer the time variation requirement onto the matrices \( A_{t} \) as we now need \( A_{t+1} - A_{t} \) to have full row rank. One of the applications in Section 3 studying sequential exogeneity, exploits this possibility as it uses \( x_{1} \) as an instrument for all time periods.

**Extension 1:** The support condition in Assumption 2.4 does not allow for deterministic relation between regressors. If some regressors do depend deterministically on others, it is possible to rewrite the model and obtain sufficient conditions guaranteeing invertibility of the matrix. Writing \( x_{it} = (x_{it}^{1}, \ldots, x_{it}^{s}, x_{it}^{s+1}, \ldots, x_{it}^{d_{x}}) \), where the first \( s \) components of \( x_{it} \) do not have functional dependence, we assume there are \( d_{x} - s \) functions \((l_{k})_{s+1 \leq k \leq d_{x}} \) such that for \( s+1 \leq k \leq d_{x}, x_{it}^{k} = l_{k}(x_{it}^{1}, \ldots, x_{it}^{s}) \). We define \( \dot{X}^{s} = (\dot{x}_{i1}^{s}, \ldots, \dot{x}_{im}^{s}) \) the collection of time differences for the first \( s \) components of \( x_{it} \). With this new setting we can rewrite Assumption 2.4.

**Assumption 2.6.** The following holds almost surely in \( V \).

1. \( \text{Int} \left( S_{\dot{X}^{s}|V} \right) \neq \emptyset \).
2. For all \( s+1 \leq k \leq d_{x}, l_{k} \) is a continuous function,
3. For all \( t \leq T - 1 \), there exists \( \dot{X}^{(t)} \) such that \( (\dot{X}^{(t)})^{s} \in \text{Int} \left( S_{\dot{X}^{s}|V} \right), p_{\dot{X}|V}(\dot{X}^{(t)}|V) > 0, \text{Rank}(\dot{X}^{(t)}) = d_{x} \) and \( \dot{X}^{(t)}/e_{t} = 0 \).

**Result 2.4.** Under Assumption 2.6, \( \mathcal{M}(V) \) is nonsingular almost surely in \( V \).

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.

**Extension 2:** The various sets of assumptions suggested so far do not handle the case where the conditional distribution of \( X \) given \( V \) is discrete. That is, if say 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 we will often assume in the rest of the paper that \( x, z \) and \( v \) are continuously distributed. \footnote{Note that we do not assume that \( V \) is continuously distributed because we directly assume that \( V \) is a control variable satisfying the control function assumption as well as the invertibility assumption. Typically the construction of the control variable will require \( V \) to be continuously distributed (see, e.g, Assumption (ii) of Theorem 1 in Imbens and Newey (2009)).}

Assume that the vector \( Z_i \) conditional on \( V_i = V \) takes \( N(V) \) values with positive probability. For each value \( Z_{(N)} \), \( N \leq N(V) \), we denote by \( X_{(N)} \) 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 \( \mathcal{M}(V) \).

\[
\mathcal{M}(V) \notin GL_{T-1}(\mathbb{R}) \Leftrightarrow \exists c \in \mathbb{R}^{T-1}, \mathcal{M}(V)c = 0 \\
\Rightarrow \exists c \in \mathbb{R}^{T-1}, c' \mathcal{M}(V)c = 0 \\
\Rightarrow \exists c \in \mathbb{R}^{T-1}, \sum_{N \in N(V)} c' M'_{(N)}M_{(N)}c = 0 \\
\Rightarrow \exists c \in \mathbb{R}^{T-1}, \sum_{N \in N(V)} ||M_{(N)}c||^2 = 0 \Rightarrow \forall N \leq N(V), M_{(N)}c = 0.
\]

An assumption yielding invertibility of \( \mathcal{M}(V) \), \( \mathbb{P}_V \) – a.s, is as follows.

**Assumption 2.7.** Almost surely in \( V \), there exists a basis \( e = (e_1, \ldots, e_{T-1}) \) of \( \mathbb{R}^{T-1} \) and for each \( t \leq T-1 \), there exists \( N_t(V) \leq N(V) \) such that \( X'_{(N_t(V))} e_t = 0 \).

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

Note that if Assumption \ref{assumption:invertibility} holds, \( X_{(N)} \) is of full column rank for all \( N \). Thus for a given \( X_{(N)} \) there can be at most \( d_x - (T-1) \) linearly independent vectors in the nullspace of \( X'_{(N)} \). Assumption \ref{assumption:invertibility} therefore implies that the support of \( Z \) conditional on \( V \) must have at least \( \left\lceil \frac{T-1}{d_x} \right\rceil \) points.

### 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

\[
y_{it} = l'_{it} b + x'_{it} \mu_i + \epsilon_{it}, \tag{15}
\]
where \( x_{it} = (x_{1t}, x_{2t}) \in \mathbb{R}^{d_x} \) and where as in Model (2), \( x_{1t} \in \mathbb{R}^{d_1} \) are exogenous regressors and \( x_{2t} \in \mathbb{R}^{d_2} \) are allowed to be endogenous. We also write \( l_{it} = (l_{1it}^1, l_{1it}^2) \in \mathbb{R}^{d_m} \) where \( l_{1it}^1 \in \mathbb{R}^{d_{11}} \) is exogenous, \( l_{1it}^2 \in \mathbb{R}^{d_{12}} \) is endogenous, which are regressors known to have a homogeneous impact.

In the case where the control variables are the residuals of the regression of \( x_{2t} \), these extensions are useful for two reasons. First, if all coefficients are assumed heterogeneous, the procedure described in Section 2.2 requires \( T \) to be at least \( d_x + d_l + 2 \), 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. 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

\[
\mathbb{E}(\epsilon_{it}|Z^L_i,X_i,V_i) = f_i(V_i),
\]

where as before, \( V_i \) is an identified function of the regressors \( X_i \) and the instruments \( Z_i \), and \( Z^L_i \) is composed of \( L^1_i \) and instruments for \( L^2_i \). The matrices \( M_i \) and \( Q_i \) are the same matrices function of \( X_i \).

Using the within-group operation,

\[
M_i \hat{y}_i = M_i \hat{\hat{L}}_i b + M_i g(V_i) + M_i \hat{\hat{u}}_i, \quad \text{with } \mathbb{E}(M_i \hat{\hat{u}}_i|Z^L_i,X_i,V_i) = 0, \tag{16}
\]

\[
\mathbb{E}(M_i \hat{y}_i|V_i) = E(M_i \hat{\hat{L}}_i|V_i) b + M_i \mathcal{M}(V_i) g(V_i)
\Rightarrow M_i \mathcal{M}(V_i)^{-1} \mathbb{E}(M_i \hat{y}_i|V_i) = \mathbb{E}(M_i \hat{\hat{L}}_i|V_i) b + M_i g(V_i).
\]

The left multiplication by \( M_i \mathcal{M}(V_i)^{-1} \) leads to having the term \( M_i g(V_i) \), which also appears in (16). This suggests a modification of the procedure developed in Robinson (1988) for the identification of \( b \). Indeed, defining \( \Delta \hat{y}_i = \hat{y}_i - \mathcal{M}(V_i)^{-1} \mathbb{E}(M_i \hat{\hat{u}}_i|V_i) \) and \( \Delta \hat{L}_i = \hat{\hat{L}}_i - \mathcal{M}(V_i)^{-1} \mathbb{E}(M_i \hat{\hat{L}}_i|V_i) \), we obtain

\[
M_i \Delta \hat{y}_i = M_i \Delta \hat{\hat{L}}_i b + M_i \hat{\hat{u}}_i.
\]

Since \( \mathbb{E}(\hat{\hat{Z}}^L_i M_i \hat{\hat{u}}_i) = \mathbb{E}(\hat{\hat{Z}}^L_i M_i \mathbb{E}(\hat{\hat{u}}_i|Z^L_i,X_i,V_i)) = 0 \) and using \( M_i = M_i' = M_i^2 \), we obtain

\[
b = \mathbb{E}(\hat{\hat{L}}^W_i M_i \Delta \hat{\hat{L}}_i)^{-1} \mathbb{E}(\hat{\hat{Z}}^L_i' M_i \Delta \hat{y}_i), \tag{17}
\]

under the assumption that \( \mathbb{E}(\hat{\hat{Z}}^L_i' M_i \Delta \hat{\hat{L}}_i) \) 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 \( T \geq 3 \). Taking \( v_{it} \) to be scalar and \( T = 3 \), then the dimension of the conditioning set for the nonparametric regressions needed for identification is 3, independently of the number of regressors in \( l_{it} \).

### 2.4.2 Case where \( T > d_x + 2 \)

The dimension of the conditioning set can also be large because the number of periods is large (but fixed) despite a small number of endogenous regressors \( x_{it}^2 \). For identification, it is required that \( T \geq k_x + 2 \), so if \( T > k_x + 2 \), one can select \( k_x + 2 \) time periods among the \( T \) available and obtain identification assuming the control function approach assumption as well as the invertibility condition holds for this subset of periods. However, it is possible\(^6\) to use the \( T \) time periods without increasing the dimension of the conditioning set.

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

\[
y_{i\tau} = X_i^\tau \mu_i + \epsilon_{i\tau}.
\]

Now we modify the control function approach assumption.

**Assumption 2.8.** There exist a set of functions \((h_i^\tau)_{\tau \in T}\) and identified functions \((C_i)_{i \in T}\) such that, defining \( v_{it} = C_t(x_{it}, z_{it}) \in \mathbb{R}^{d_v} \),

\[
\forall \tau \in T, \forall t \in \tau, \mathbb{E}(\epsilon_{it} \mid X_i^\tau, V_i^\tau) = h_i^\tau(V_i^\tau).
\]

Indeed the assumption that we used in the main model is \( \mathbb{E}(\epsilon_{it} \mid X_i, V_i) = f_t(V_i) \) and does not imply Assumption 2.8. If Assumption 2.1 (2) holds, then by the law of iterated expectations \( \mathbb{E}(\epsilon_{it} \mid X_i^\tau, V_i^\tau) = \mathbb{E}(f_t(V_i) \mid X_i^\tau, V_i^\tau) \) which is not necessarily a function of \( V_i^\tau \) only. However, the independence assumptions we make in all our applications directly satisfy Assumption 2.8.

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

\[
\mathbb{E}(\mu_i) = \frac{1}{\binom{T}{k_x+2}} \sum_{\tau \in T} \mathbb{E}(Q_i^t y_i^\tau - Q_i^t g^\tau(V_i^\tau)).
\]

---

\(^6\)I thank Donald Andrews for this suggestion.
2.4.3 Identifying higher-order properties of $\mu_i$

In a model with strict exogeneity, Arellano and Bonhomme (2012) extend the method in Chamberlain (1992) to identify the variance matrix and even the distribution of $(\alpha_i, \mu_i)$ under various restrictions on the time-dependence of $\epsilon_{it}$ and on the joint distribution of $(\epsilon_i, \alpha_i, \mu_i)$. The argument first identifies the common parameters. Then subtracting the common part from the outcome variables, higher order moments of $\mu_i$ are separated from those of $\epsilon_i$ using the above-mentioned restrictions. We note here that their argument can be combined with the assumptions made in the present paper so as to allow for endogeneity of the regressors. Indeed $g(V_i)$ being recovered using the method described in Section 2.2.2, the analysis of Arellano and Bonhomme (2012) can be conducted on $y_i - g(V_i) = X_i\mu_i + u_i$ which takes the same form as in their paper. We refer to the paper for more details on the procedure to recover these moments.

3 Applications and Variations of the Model

In this section, we propose a direct application of the model and also describe some models, different from the main model [1] but where, using the appropriate control variables, the two-step approach also provides identification results under some conditions.

3.1 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 education outcomes, where $x_{it}$ is any type of parental investment, and the random coefficients represent 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 $(\mu_i, \epsilon_{it})$ at the time of the decision. Instead, she has information about it contained in $\eta_{it} \in \mathbb{R}$, scalar random variable. Writing $C_t(x, z)$ a cost function with $z$ the cost shifters, she chooses $x_{it}$ to maximize an expected profit,

$$ x_{it} = \arg\max_x \mathbb{E} (y_{it} - C_t(x, z_{it}) | z_{it}, \eta_{it}). \tag{18} $$

This implies the existence of a function $H_t$ such that $x_{it} = H_t(z_{it}, \eta_{it})$. We assume that for all $t \leq T$, $H_t(z_{it}, \eta)$ is strictly monotonic in $\eta$ with probability 1, $\eta_{it}$ is continuously distributed and its CDF
is strictly increasing. We also assume that $(\eta_{it}, \epsilon_{it})_{t \leq T}$ is independent of $(z_{it})_{t \leq T}$. Defining $v_{it} = F_{z_{it} | z_{it}}(x_{it} | z_{it}) = F_{\eta}(\eta_{it})$, these assumptions as shown in Imbens and Newey (2009) imply that $\epsilon_{it}$ is independent of $Z_i$ conditional on $V_i$. Therefore,

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

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 this result requires the information $\eta_{it}$ to be scalar while the unobserved heterogeneity $(\mu_i, \epsilon_{it})$ in the outcome equation is of higher dimension. On the other hand, it does not require the instrument to be independent of the returns $\mu_i$.

### 3.2 Sample selection

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 and address the endogeneity issue with a selection equation which provides them with a control variable.

The selection equation studied here is similar and some of the arguments closely follow theirs, but the outcome equation differs: as in (1) it is a panel random coefficients specification. The selection model we consider is

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

$$
d_{it} = 1 \left( \eta_{it} \leq C_t(x_{it}, z_{it}) \right),
$$

$$
y_{it} = d_{it} y_{it}^*,
$$

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

$$
\mathbb{E}(\epsilon_{it}|d_i = 1, X_i, P_i) = f_t(P_i).
$$

Note that as pointed out in Das, Newey, and Vella (2003) in the cross-sectional case, this assumption is satisfied in particular if $(\epsilon_{it}, \eta_{it})_{t \leq T}$ is independent of $(X_i, Z_i)$ and if the cdf of $\eta_{it}$, $F_t$, is strictly increasing. Indeed, in this case defining $\nu_{it} = F_t(\eta_{it})$, $\nu_i = (\nu_{it})_{t \leq T}$, $p_{it} = F_t(C_t(x_{it}, z_{it}))$, then $d_{it} = 1 \left( \nu_{it} \leq p_{it} \right)$ and

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

20
as desired (where by an abuse of notation the inequality $\nu_t \leq P_t$ denotes the inequality component by component). Note that 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 $t$. Identification can thus be obtained by a similar two-step argument. 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 $u_{it} = \epsilon_{it} - f_t(P_i)$, $\tilde{x}_{it} = d_{i,t+1} x_{i,t+1} - d_{it} x_{it}$, $\tilde{g}_t(P_i) = d_{i,t+1} f_{t+1}(P_i) - d_{it} f_t(P_i)$, 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 $\tilde{X}_i = \tilde{X}_i$. Hence,

$$
\mathbb{E} \left( \tilde{M}_i \tilde{y}_i \mid d_i = 1, P_i \right) = \mathbb{E} \left( \tilde{M}_i \tilde{X}_i \mu_i + \tilde{M}_i \tilde{g}(P_i) + \tilde{M}_i \tilde{u}_i \mid d_i = 1, P_i \right) \\
= \mathbb{E} \left( \tilde{M}_i \tilde{X}_i \mu_i + \tilde{M}_i g(P_i) + \tilde{M}_i \tilde{u}_i \mid d_i = 1, P_i \right) = \mathcal{M}(P_i) g(P_i),
$$

where we define $\mathcal{M}(P_i) = \mathbb{E} \left( \tilde{M} \tilde{y}_i \mid (d_i = 1), P_i \right)$. This first step equation identifies $g$ on the support of $P_t$ if $\mathcal{M}(P_i)$ is invertible a.s. The second step equation will be given by

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

**Assumption 3.1.** $\mathbb{E} (\epsilon_i \mid d = 1, X, P) = f_t(P)$, and $\mathcal{M}(P)$ is invertible almost surely in $P$.

**Result 3.1.** Under Assumptions 4.7 and 3.1, $\mathbb{E}(\mu \mid d = 1)$ is identified.

The identified object is the average effect conditional on selection, $\mathbb{E}(\mu \mid d = 1)$ which is in general different from $\mathbb{E}(\mu)$ unless $\mu \perp (\eta, X, Z)$. That additional restrictions are needed to identify $\mathbb{E}(\mu)$ is intuitive: when $d_{it} \neq 1$ the econometrician does not have any information on the unobserved 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 $t \leq T$, which can be quite small if $T$ is large (and still fixed).

Other cases studied in Das, Newey, and Vella (2003) can be handled here. For instance, the model allows for regressors $s_{it}$ to be subject to selection as well, that is, to be 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 $\mathbb{E}(\epsilon_i \mid d = 1, X, S, P) = f_t(P)$ 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

$$ y_{it} = d_{it} y_{it}^*, \quad \text{with} \quad y_{it}^* = x_{it}^1 \mu_{i1} + x_{it}^2 \mu_{i2} + \epsilon_{it}, $$

$$ x_{it}^2 = d_{it} x_{it}^{2*}, \quad \text{with} \quad x_{it}^{2*} = \pi_{i1}^2(x_{it}^1, z_{it}^1) + v_{it}, $$

$$ d_{it} = 1 \left( \nu_{it} \leq p_{it}(x_{it}^1, z_{it}^1, z_{it}^2) \right) = 1 \left( \nu_{it} \leq p_{it} \right), $$

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

$$ \mathbb{E}(v_{it} \mid d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2) = \mathbb{E}(\mathbb{E}(v_{it} \mid \nu_{it}) \mid (\nu_{it} \leq p_{it}), x_{it}^1, z_{it}^1, z_{it}^2) := \phi_t(p_{it}), $$

implying

$$ \mathbb{E}(x_{it}^{2} \mid d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2) = \pi_{i1}^2 (x_{it}^1, z_{it}^1) + \phi_t(p_{it}). $$

This gives

$$ x_{it}^{2} - \mathbb{E}(x_{it}^{2} \mid d_{it} = 1, x_{it}^1, z_{it}^1, z_{it}^2) = v_{it} - \phi_t(p_{it}) := \tilde{v}_{it}, $$

where $\tilde{v}_{it}$ 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 $\nu_i = (\nu_{is})_{s \leq T}$, and similarly $d_i, P_i, \tilde{V}_i, X_i = (X_{i1}^1, X_{i2}^2)$ and $\phi(P_i) = (\phi_t(p_{it}))_{t \leq T}$. Note that the function $(V_i, P_i) \mapsto (\tilde{V}_i, P_i)$ is one-to-one. Therefore,

$$ \mathbb{E}(\epsilon_{it} \mid d_i = 1, X_i, \tilde{V}_i, P_i) = \mathbb{E}(\mathbb{E}(\epsilon_{it} \mid \nu_i, V_i, X_{i1}^1, Z_{i1}^1, Z_{i2}^2) \mid (\nu_i \leq P_i), X_i, V_i, P_i) $$

$$ = \mathbb{E}(\mathbb{E}(\epsilon_{it} \mid \nu_i, V_i) \mid (\nu_i \leq P_i), X_i, V_i, P_i) := \mathbb{h}_t(V_i, P_i), $$

$$ = \mathbb{h}_t(\tilde{V}_i + \phi(P_i), P_i) := f_t(\tilde{V}_i, P_i). $$

This conditional expectation is as in Assumption 2.1, where the control variables are $(\tilde{V}_i, P_i)$. This double use of the control function approach is already suggested in Das, Newey, and Vella (2003). We presented here a slight modification such that the identification requires two steps instead of three. This for instance allows one to use the formula for the asymptotic variance given in Section 4.4 (provided $\pi^2$ is not an object of interest) as it is known that increasing the number of steps typically changes the asymptotic variance matrix.
### 3.3 Relaxing a strict exogeneity condition

Instead of focusing on contemporaneous endogeneity, that is the joint dependence of \((\epsilon_{it}, x_{it})\), one can use the framework of this paper to relax restrictions on the joint dependence of \((\epsilon_{it}, x_{it+1}, ..., x_{iT})\). This corresponds to relaxing the strict exogeneity condition imposed in Arellano and Bonhomme (2012) to allow for sequential exogeneity.

The model is as \(\Pi\),

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

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 \(\epsilon_{it}\). 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 and write \(x_{it+1} = m_t(x_{it}) + \eta_{it+1}, 1 \leq t \leq T - 1\), where \(\eta_{it}\) are i.i.d over time. One could alternatively consider \(x_{it+1} = m_{t+1}(x_{it}, \eta_{it+1})\) with \(\eta\) scalar and \(m_t\) strictly monotonic in \(\eta\) and use the control variable suggested in Imbens and Newey (2009). We also assume that \((\epsilon_{it}, \eta_{it+1}, ..., \eta_{iT}) \perp (x_{i1}, x_{i2}, ..., x_{it})\), which implies that \((\epsilon_{it}, \eta_{it+1}, ..., \eta_{iT}) \perp (x_{i1}, ..., x_{it})\). That is, the innovations giving the evolution of \(x\) after time \(t\) and \(\epsilon_t\) are independent of past values of \(x\). However the joint distribution of \((\epsilon_{it}, \eta_{it+1}, ..., \eta_{iT})\) is not restricted and allows for sequential exogeneity. Then, for all \(t\) less than \(T\),

\[
\mathbb{E}(\epsilon_{it} | x_{i1}, ..., x_{iT}, \eta_{i2}, ..., \eta_{iT}) = \mathbb{E}(\epsilon_{it} | x_{i1}, \eta_{i2}, ..., \eta_{iT}) = \mathbb{E}(\epsilon_{it} | \eta_{it+1}, ..., \eta_{iT}) = f_t(\eta_{i2}, .., \eta_{iT}) := f_t(V_i),
\]

where the first equality holds by the Markov structure of \(x\), and the second by the independence assumption on the error terms. We define \(V_i = (\eta_{i2}, .., \eta_{iT})\) to be the control variable. Note that the \(\eta_{it}\) are all identified as the residuals of reduced form regressions. Assuming independence between \(\eta_{it}\) for \(t \geq 2\) and \(x_{i1}\) might restrict the joint distribution of \((x_{it}, \mu_i)\) if \(x_{i1}\) is correlated with \(\mu_i\) but it does not necessarily require independence between \(x_{it}\) and \(\mu_i\).

Defining as previously \(M_i = I - \hat{X}_i(\hat{X}_i'\hat{X}_i)^{-1}\hat{X}_i'\), \(\mathcal{M}(V_i) = \mathbb{E}(M_i|V_i), u_{it} = \epsilon_{it} - f_t(V_i)\) 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,

\[
\mathbb{E}(M_i\hat{y}_{it}|V_i) = \mathcal{M}(V_i)g(V_i), \quad \text{and} \quad \mathbb{E}(Q_i\hat{y}_i) = \mathbb{E}(\mu_i) + \mathbb{E}(Q_ig(V_i)).
\]

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.

---

Note that as for the sample selection example, independence is stronger than needed. Conditional mean independence of \(\mathbb{E}(\epsilon_{it}|x_{i1}, \eta_{i2}, ..., \eta_{iT})\) with respect to \(x_{i1}\) is sufficient.
That $\mathcal{M}(V)$ can be nonsingular is nontrivial here. Indeed by definition, $V_i = (\eta_1, \ldots, \eta_T)$ 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. To show that this invertibility condition can actually hold, let us look more closely at the case where $x_i$ is a scalar AR(1) process, that is, $x_{i t + 1} = \rho x_{i t} + \eta_{i t + 1}$ with $\rho \neq 1$ and $\rho \neq 0$, and $(\epsilon_{i t}, \eta_{i t + 1}, \ldots, \eta_{i T})_\perp(x_{i1}, \eta_{i2}, \ldots, \eta_{it})$. By definition, $M_i = I - \hat{X}_i \hat{X}_i^T/(\hat{X}_i^T \hat{X}_i)$. Moreover, $x_{it} = \rho^{t-1} x_{i1} + \sum_{s=2}^{t} \rho^{t-s} \eta_{is}$, therefore defining the two vectors $C_1 = [\rho - 1] (1, \rho, \ldots, \rho^{T-2})' \in \mathbb{R}^{T-1}$ and $C_2(V) = (\eta_1, [\rho - 1] \eta_2 + \eta_3, \ldots, [\rho - 1] \sum_{s=2}^{T-1} \rho^{T-s-1} \eta_{is} + \eta_{iT})' \in \mathbb{R}^{T-1}$, we can write

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

For a given value $\bar{V} \in S_V$, we proved in Section 2.3 that

$$\mathcal{M}(\bar{V}) \notin GL_{T-1}(\mathbb{R}) \iff \exists a \in \mathbb{R}^{T-1} \setminus \{0\}, \ \mathcal{M}(\bar{V}) a = 0,$$

$$\Rightarrow \exists a \in \mathbb{R}^{T-1} \setminus \{0\}, \ M a = 0 \ \mathbb{P}_{X|V=\bar{V}} a.s.,$$

$$\Rightarrow \exists a \in \mathbb{R}^{T-1} \setminus \{0\}, \ \hat{X} \text{ is collinear to } a \ \mathbb{P}_{X|V=\bar{V}} \text{ a.s.}$$

The draws of $\hat{X}$ from $\mathbb{P}_{X|V=\bar{V}}$, 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 $x_1 C_1$ proportional to the constant vector $C_1$. Note that $S_{x_1|V=\bar{V}} = S_{x_1}$ since $x_1 \perp V$. If there are two nonzero points $x_1$ and $\bar{x}_1$ in $S_{x_1}$ such that $a$ is collinear to $\hat{X} = x_1 C_1 + C_2(\bar{V})$ and to $\bar{X} = \bar{x}_1 C_1 + C_2(\bar{V})$, then since $a \neq 0 \ X$ and $\bar{X}$ are collinear. Since $C_1 \neq 0$, 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 $\bar{V} \in \left\{ b \left( \begin{array}{c} 1 \\ 1 \end{array} \right) \mid b \in \mathbb{R} \right\} =: \mathcal{D}$. Hence, the existence of such $x_1$ and $\bar{x}_1$ implies $\bar{V} \in \mathcal{D}$. $\mathcal{D}$ is a subset of $\mathbb{R}^{T-1}$ with $\mathbb{P}_V$ 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. (25) holds and for all $t \leq T$, $x_{i t + 1} = \rho x_{i t} + \eta_{i t + 1}$ with $\rho \neq 1$, $(\epsilon_{i t}, \eta_{i t + 1}, \ldots, \eta_{iT})_\perp(x_{i1}, \eta_{i2}, \ldots, \eta_{it})$, and $(X_i, \mu_i, \alpha_i, \epsilon_i, \eta_i)$ is i.i.d.

2. Either $x_{i1}$ has a discrete distribution with at least two support points, or $x_{i1}$ is continuously distributed and $\operatorname{Int}(S_{x_1}) \neq \emptyset$. Moreover, $(\eta_1, \ldots, \eta_T)$ is continuously distributed on $\mathbb{R}^{T-1}$, and $f_t$ is continuous.
**Result 3.2.** Under Assumptions [4.7 and 3.2, \( \mathbb{E}(\mu_i', \alpha_i) \) is identified.

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 each time period \( x_{it} \) creates sufficient time variation of the regressor as a result of the condition \( \rho \neq 1 \).

**Note 3.1.** While the model exposed above allows for the regressor to be correlated with past disturbances, the assumptions are not compatible with a lagged dependent variable as regressor. Indeed writing \( x_{i(t+1)} = m_t(x_{it}) + \eta_{it+1} \) implicitly imposes a homogeneous dependence on the past value that is not consistent with the specification \( y_{it+1} = \mu_i y_{it} + \epsilon_{it} \).

**Note 3.2.** Using the first value of the regressor as an instrument for the correlation between the residual \( \epsilon \) at a given time and the future values of the regressors can be extended to models where in addition to contemporaneous endogeneity, there is a such a feedback effect. First, note that Assumption 2.1 (2) that \( \mathbb{E}(\epsilon_{it} | X_i, V_i) = f_t(V_i) \) does not exclude feedback for the control variables themselves as all time periods of \( (v_{it})_{s \leq T} \) are arguments of the function \( f_t \). One could however wonder if a feedback effect in the instruments \( z_{it} \) is allowed within the framework of this paper. For instance, the production function example given in Section 3.1 imposed \( (\eta_{it}, \epsilon_{it})_{t \leq T} \) \( \mathbb{P}(z_{it})_{t \leq T} \). As in the example studied above, it is actually possible to use the first observation of the instruments to construct the control variables. We show this briefly in the same production function model but with a Markov structure on the instrument.

Recall \( x_{it} = H_t(z_{it}, \eta_{it}) \) and assume that the instrument satisfies \( z_{it+1} = m_{t+1}(z_{it}, \nu_{it+1}) \), where the \( \nu_{it} \in \mathbb{R} \) are i.i.d over \( t \). Assume moreover that \( m_{t+1}(z_{it}, \cdot) \) is strictly increasing with probability \( 1, \nu_t \) is continuously distributed, its CDF is strictly continuous for all \( t \), and

\[
(\epsilon_{it}, \eta_{i1}, \ldots, \eta_{iT}, \nu_{it+1}, \ldots, \nu_{iT}) \mathbb{P}(z_{i1}, \nu_{i2}, \ldots, \nu_{iT}).
\]

This assumption guarantees \( \epsilon_{it}, \mathbb{P}(z_{it}) \) but does not restrict the joint distribution of \( \epsilon_{it} \) and \( \nu_{is}, s \geq t + 1 \). Define \( r_{it} = F_{z_{i+1}|z_i}(z_{it} | z_{it}), s_{it} = F_{x_{i}|z_i}(x_{it} | z_{it}), \) and \( V_i = (s_{i1}, r_{i2}, s_{i2}, \ldots, r_{iT}, s_{iT}) \).

The vector \( V_i \) is identified as a collection of residuals of cross-section reduced form regressions. Moreover,

\[
\mathbb{E}(\epsilon_{it} | X_i, V_i) = \mathbb{E}(\epsilon_{it} | H(z_{i1}, \eta_{i1}), \ldots, H(z_{iT}, \eta_{iT}), s_{i1}, r_{i2}, s_{i2}, \ldots, r_{iT}, s_{iT})
\]

\[
= \mathbb{E}(\epsilon_{it} | r_{it}, \ldots, r_{iT}, s_{i1}, \ldots, s_{iT}) := f_t(V_i),
\]

by the strict monotonicity of the functions \( (m_t)_{2 \leq t \leq T}, (H_t)_{t \leq T}, (F_{\nu_t})_{2 \leq t \leq T} \) and \( (F_{\eta_t})_{t \leq T} \). Under invertibility of \( \mathcal{M}(V) \mathbb{P}_V \) almost surely, identification of the average effect is obtained.
4 Estimation

As seen in Section 2, the proof of identification of \( \mathbb{E}(\mu) \) in Model (1) are constructive. An estimator can therefore naturally constructed by following the identification steps and 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 \( M(V) = \mathbb{E}(M_i|V = V) \) and \( k(V) = \mathbb{E}(M_i y_i|V = V) \) are estimated nonparametrically using the generated values \( \hat{V} \) as regressors and the function \( g = M^{-1} \) is estimated plugging in the estimators \( \hat{M}(V) \) and \( \hat{k}(V) \) in this formula. 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 \( \delta_i = \Delta(\det(\hat{X}_t'\hat{X}_t) > \delta_0) \). The estimator for \( \mathbb{E}(\mu|\delta) \) will be a sample analog of Equation (12), plugging in the estimator of \( g \) and \( V \).

It is clear that the asymptotic properties of this estimator will depend on the definition of the control variables, that is, on \( C_t \). The focus of the asymptotic analysis will thus be on an important example in the class of models satisfying (1) and Assumption 2.1. More specifically, the model is

\[
y_{it} = x_{it}^1 \mu_1 + x_{it}^2 \mu_2 + \alpha_i + \epsilon_i, \tag{26}
\]

\[
x_{it}^2 = b_t(x_{it}^1, z_{it}) + v_{it}, \quad \mathbb{E}(v_{it}|x_{it}^1, z_{it}) = 0,
\]

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 control variables in this model are the residuals of the nonparametric regression of the endogenous regressors on the exogenous regressors and the instruments.

The estimators used in the 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 proceed in this section with an explicit definition of the estimators and a stepwise proof of asymptotic normality of \( \hat{\mu} \). All proofs are in the Appendix.

4.1 Definition of the estimators

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

\[
\hat{\beta}_t = (R_t R_t')^{-1} \sum_i r^L(\xi_{it}) x_{it}^{2r} = (R_t R_t')^{-1} R_t X_t^{2r}, \tag{27}
\]
where $R_t = (r_{1t}, \ldots, r_{nt})$ is $L \times n$ and $X_t^2 = (x_{1t}^2, \ldots, x_{nt}^2)$ is $d_2 \times n$.

The control variables are defined as the residuals of the regression of $x_t^2$. Later, the support $S_V$ of $V$ will be assumed bounded. However, the values obtained using the estimated residuals might not be in $S_V$: it will be convenient for the asymptotic analysis to introduce a transformation $\tau$ of the generated variables to ensure that their transformed values lie in $S_V$. Specifically, we assume that the support of $v_t$ is of the form $\prod_{d=1}^{d_2} [\bar{v}_{td}; \bar{v}_{td}]$ so that the support of $V$ is $S_V = \prod_{d \in d_2, t \in T} [\bar{v}_{td}; \bar{v}_{td}]$. We define $\tau$ such that for $V = (v_1', \ldots, v_T') \in \mathbb{R}^{Td_2}$, then $\tau(V) \in S_V$ and the $(d_2(t - 1) + d)^{th}$ component of $\tau(V)$ satisfies

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

where $v_{t,d}$ is the $d^{th}$ component of $v_t$. Write $r_{it} = r^k(\xi_{it})$ and $b_{it} = b_t(\xi_{it})$ and define $b_{it} = b_t(\xi_{it})$ and $\hat{b} = \beta_t r_t$. We also define the residuals $\hat{v}_{it} = x_i^2 - \hat{b}_{it}$, and $\hat{V}_i = (\hat{v}_{i1}', \ldots, \hat{v}_{iT}')'$. Our estimator for $V_i$ will then be $\hat{V}_i = \tau(\hat{V}_i)$: $\tau(\hat{V}_i)$ is the projection of $\hat{V}_i$ onto $S_V$ such that $\hat{V}_i$ lies outside of $S_V$, the function $\tau(\hat{V})$ 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 $||\hat{V}_i - V_i|| \leq ||\hat{V}_i - V_i||$.

Let $p^K(V) = (p_{1K}(V), \ldots, p_{KK}(V))'$ denote a $K \times 1$ vector of approximating functions, $p_i = p^K(V_i)$ and $\hat{p}_i = p^K(\hat{V}_i)$. An estimator of $h^W(V) = \mathbb{E}(w_i|V_i = V)$ for a generic scalar random variable $w_i$ using the generated $\hat{V}$ is $p^K(V)' \hat{\pi}^W$ where $\hat{\pi}^W$ is a vector of size $K$ given by

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

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

Using this general definition, we construct component by component estimators $\hat{M}$ and $\hat{k}$ for the matrix and vector valued functions $M$ and $k$. We obtain $p^K(V)' \hat{\pi}^M^s$ an estimator of the $(s,t)$ component of the matrix $M$, taking $w_i$ to be $(M_i)_{s,t}$. Similarly, an estimator of the $s^{th}$ component of $k$ will be $p^K(V)' \hat{\pi}^k_s$, choosing $w_i = (M_i y_{i1})_s$. Under Assumptions 2.1 and 2.3 we have $g(V) = M(V)^{-1} k(V)$. A straightforward estimator of $g$ is thus

$$
\hat{g}(V) = \hat{M}(V)^{-1} \hat{k}(V).
$$

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

$$
\hat{\mu} = \frac{1}{\sum_{i=1}^{n} \delta_i} \sum_{i=1}^{n} \delta_i Q_i [\hat{y}_i - \hat{g}(\hat{V}_i)] = \frac{1}{\sum_{i=1}^{n} \delta_i} \sum_{i=1}^{n} Q_i [\hat{y}_i - \hat{g}(\hat{V}_i)].
$$
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 to eventually obtain the asymptotic behavior of \( \hat{\mu} \). This type of asymptotic analysis is the subject of a wide literature on nonparametric and semiparametric estimation with generated covariates. Before laying out the main results of our asymptotic analysis, we give here a brief overview of this literature.

Papers studying asymptotic normality of semiparametric estimators, such as Newey (1994a), Chen, Linton, and Van Keilegom (2003), Ai and Chen (2003) and Ichimura and Lee (2010) among many other references, have a level of generality which encompasses the case where the regressors are themselves estimated. However, the conditions given in these papers are “high-level” conditions and are not easily applied to the composition of nonparametrically estimated infinite-dimensional nuisance parameters. Examples of asymptotic derivations in specific models with generated regressors are papers already citation such as Newey, Powell, and Vella (1999), Imbens and Newey (2009), and Das, Newey, and Vella (2003), as the use of a nonparametric control function approach naturally suggests an estimator with generated covariates. Others are, e.g., Ahn and Powell (1993), Blundell and Powell (2004), Newey (2009) and Escanciano, Jacho-Chávez, and Lewbel (2016). Moreover, recent contributions have focused on obtaining general asymptotic results for such semiparametric estimators. Among important recent contributions, Hahn and Ridder (2013) derives in the spirit of Newey (1994a) a general formula of the asymptotic variance of estimators with generated regressors. However they do not provide results on how to obtain asymptotic normality for particular classes of estimators. For estimators with generated regressors depending on a nonparametrically estimated function, this type of analysis can be found for instance in Escanciano, Jacho-Chávez, and Lewbel (2014), Mammen, Rothe, and Schienle (2016) and Hahn, Liao, and Ridder (2018). Escanciano, Jacho-Chávez, and Lewbel (2014) obtain a uniform expansion of a weighted sample average of residuals obtained from kernel-estimated nonparametric regressions with generated covariates, which can be then be used to prove asymptotic normality of a class of semiparametric estimators. Mammen, Rothe, and Schienle (2016) study the asymptotic normality of a general class of semiparametric GMM estimators depending on a nonparametric nuisance parameter, also constructed 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. Moreover we use series to construct the nonparametric estimates while the infinite dimensional nuisance parameter in Mammen, Rothe, and Schienle (2016) is a conditional expectation estimated with local polynomial estimator and they do not specify an estimator for the generated covariates. Estimators in Hahn, Liao, and Ridder (2018) have a structure closer to that of \( \hat{\mu} \): they study nonparametric two-step sieve M estimators, but focus on known functionals. They show asymptotic normality of
their estimator when standardized by a finite sample variance and give a practical estimator of this variance. They do not however provide an explicit formula of the asymptotic variance. The estimator we analyze in this section is instead an estimated functional of the two-step nonparametric estimators. Using a different type of proof techniques with lower level conditions on the primitives of a more specific class of models, we show asymptotic normality and obtain the asymptotic variance of a generic class of estimators to which ours belongs. See, e.g., Mammen, Rothe, and Schienle (2016) for a literature review on semiparametric estimation with generated covariates and explanation on the specificity of this type of estimation.

4.2 Convergence rates of the nonparametric two-step 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 $A \in \mathcal{M}_p(\mathbb{R})$, $||A||_F = \left( \sum_{i,j \leq p} a_{ij}^2 \right)^{1/2} = \text{tr}(A'A)^{1/2}$. 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 $\lambda_1(A) \leq \ldots \leq \lambda_p(A)$ all the eigenvalues ranked by increasing order (when they exist).

We will use the following results. First, for all $A \in \mathcal{M}_p(\mathbb{R})$, $||A||_2 \leq ||A||_F$. 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 $||.||_2$, $||Aa|| \leq |||A|||_2 ||a||$.

We also write, for $g$ a vector of functions of $x \in \mathcal{S}_x \subset \mathbb{R}^k$, $||g||_\infty = \sup_{x \in \mathcal{S}_x} ||g(.)||$. For $l = (l_1, \ldots, l_k) \in \mathbb{N}^k$, we define $|l| = \sum_{j=1}^k l_j$, and the partial derivative $\partial^l g(x) = \partial^{|l|} g(x)/\partial^{l_1} x_1 \cdots \partial^{l_k}$. We will use the norm $|g|_d = \max_{|l| \leq d} \sup_{x \in \mathcal{S}_x} ||\partial^l g(x)||$ when $g$ is $d$ times differentiable. We denote by $\partial g(x)$ the Jacobian matrix $(\partial g(x)/\partial x_1, \ldots, \partial g(x)/\partial x_k)$. In what follows, LLN denotes the weak law of large numbers, $C$ a generic constant (whose value can change from one line to another) and for a sequence $(c_n)_{n \in \mathbb{N}} \in \mathbb{R}^n$, the notation $c_n \rightarrow 0$ should be understood as $c_n \rightarrow_{n \rightarrow \infty} 0$. We now proceed step by step to derive uniform and MSE convergence rates of all nonparametric estimators.

**Sample mean square error for estimator of the control variables**: The generated covariates $\hat{V}_t$ are constructed as the estimated residuals of $T$ regressions. The method to generate regressors is as in Newey, Powell, and Vella (1999) and we use some of their results.

**Assumption 4.1.** There exists $\gamma_1 > 0$ and $a_1(L)$ such that $\sqrt{L/n} a_1(L) \rightarrow_{n \rightarrow \infty} 0$ and for all $t \leq T$, 1. $(x_{it}^2, \xi_t)$ is i.i.d over $i$, continuously distributed and $\text{Var}(x_{it}^2|\xi_t)$ is bounded, 2. There exists a $L \times L$ nonsingular matrix $\Gamma_{1t}$ such that for $R^L(\xi_t) = \Gamma_{1t} r^L(\xi_t)$, $E(R^L(\xi_t) R^L(\xi_t)^\prime)$ has smallest eigenvalue bounded away from zero uniformly in $L$, 29
3. There exists $\beta^L_t$ such that $\sup_{S_t} \|b_t(\xi_t) - \beta^L_t r^L(\xi_t)\| \leq CL^{-\gamma_1}$.

4. $\sup_{S_t} \|R^L(\xi_t)\| \leq a_1(L)$.

**Result 4.1.** Under Assumption 4.1,

\[
\frac{1}{n} \sum_{i=1}^{n} \| V_i - \hat{V}_i \|^2 = O_p \left( \frac{L}{n} + L^{-2\gamma_1} \right) = O_p \left( \Delta_n^2 \right),
\]

\[
\max_i \| V_i - \hat{V}_i \| = O_p (a_1(L) \Delta_n).
\]

If for instance $b_t$ is continuously differentiable up to order $p$, writing $d_\xi = d_1 + d_z$, then Assumption 4.1 holds with $\gamma_1 = p/d_\xi$ for different choices of sieve basis.

Conditions satisfying Assumption 4.1 typically require the support of $\xi_t$ to be bounded and the density of $\xi_t$ 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 $S_\xi$ where the density is bounded away from 0. However, we are interested here in the average effect $E(\mu)$ and counterfactuals involving population means. Trimming arbitrarily on regressors to estimate a conditional effect is contrary to this goal. 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) in assuming a polynomial lower bound on the rate of decrease of the density. Formally, we assume that $S_\xi$ is of the form $\times_{d=1}^{d_\xi} [\xi_{td}; \tilde{\xi}_{td}]$. Recall that $\xi_{td}$ is the $d^{th}$ component of $\xi_t$. In this case, the set of conditions in Assumption 4.1 can be modified as follows.

**Assumption 4.2.** There exists $\alpha_1 > 0$ and $\gamma_1 > 0$ such that $\sqrt{L/n} L^{\alpha_1+1} \xrightarrow{n \to \infty} 0$ and $\forall t \leq T$,

1. $(x_{it}^2, \xi_{it})$ is i.i.d over $i$, continuously distributed and $\text{Var}(x_{it}^2|\xi_t)$ is bounded,

2. $r^L(.)$ is the power series basis, and $\forall \xi_t \in S_\xi$, $f_{\xi}(\xi_t) \geq \prod_{d=1}^{d_\xi} (\xi_{td} - \xi_{td})^{\alpha_1} (\xi_{td} - \xi_{td})^{\alpha_1},$

3. There exists $\beta^L_t$ such that $\sup_{S_t} \|b_t(\xi_t) - \beta^L_t r^L(\xi_t)\| \leq CL^{-\gamma_1}$.

**Result 4.2.** Under Assumption 4.2, (29) and (30) hold.

Allowing for unbounded support is a desirable extension as well and is made possible using a method similar to Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2008), but is outside the scope of this paper.

**Convergence rates of two-step series estimators:** The estimator of $E(w_i|V_i = V) = h^W(V)$ defined in (28) using the generated control variables is as in Newey, Powell, and Vella (1999), aside from the panel aspect. However, they impose an orthogonality condition which by definition does not hold for our specific choices of $w_i$ and this rules out a direct application of their asymptotic results. More specifically, writing

\[
w = h^W(V) + e^W, \; \mathbb{E}(e^W|V) = 0,
\]
an additional assumption required to apply directly Newey, Powell, and Vella (1999) would be \(\mathbb{E}(e^W|X^1, V, Z) = 0\), that is, \(e^W\) is conditionally mean-independent of all variables involved in the first step, that is in the construction of the control variables. However this condition does not hold when \(w_i\) is either a component of the matrix \(M = I - \dot{X}'(\dot{X}'\dot{X})^{-1}\dot{X}'\) or of the vector \(M\dot{y}\), because

\[
\mathbb{E}(M|X^1, X^2, Z) = M \neq \mathbb{E}(M|V) = \mathcal{M}(V),
\]

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

This difference has been documented for instance in Hahn and Ridder (2013) and Mammen, Rothe, and Schienle (2016). It has implications for the convergence rate of the two-step estimator which will have an extra term and, as will be clear in a later part of the paper, on the asymptotic variance of a linear functional of this estimator. To account for the extra term \(\mathbb{E}(e^W|X^1, V, Z) = \mathbb{E}(e^W|X, V)\) where \(X = (X^1, X^2)\), we write

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

\[
w = h^W(V) + \rho^W(X, Z) + e^{W*}, \quad \mathbb{E}(e^{W*}|X, Z) = 0,
\]

where \(\rho^W(X, Z) = \mathbb{E}(e^W|X, Z) = \mathbb{E}(w|X, Z) - \mathbb{E}(w|V)\). As was done for estimation of \(b_t\), we first state results on \(\hat{h}\) under generic assumptions then show that these assumptions are satisfied when the regressors have bounded support and their joint density goes to 0 at the boundary of the support.

**Assumption 4.3.**

1. \(e^{W*}_i\) is i.i.d and \(\mathbb{E}((e^{W*}_i)^2 | X, Z)\) is bounded on \(S_{X,Z}\),
2. \(h^W\) is Lipschitz on \(S_V\) and \(\rho^W\) is bounded on \(S_{X,Z}\),
3. There exists a \(K \times K\) nonsingular matrix \(\Gamma_2\), 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 \(K\),
4. There exists \(\gamma_2\) and \(\pi^K\) such that \(\sup_{S_V} |h^W(V) - p^K(V)'\pi^K_W| \leq CK^{-\gamma_2}\),
5. For \(\sup_{S_V} \|P^K(V)\| \leq b_1(K)\) and \(\sup_{S_V} \|\partial P^K(V)/\partial V\| \leq b_2(K)\), \(\sqrt{K} b_2(K) \Delta_n \xrightarrow{n \to \infty} 0\) and \(\sqrt{K/n} b_1(K) \xrightarrow{n \to \infty} 0\).

**Result 4.3.** Under Assumption 4.1 and 4.3,

\[
\int \left| h^W(V) - h^W(V) \right|^2 dF(V) = O_p(K/n + K^{-2\gamma_2} + \Delta^2_n b_2(K)^2),
\]

\[
\sup_{V \in S_V} |h^W(V) - h^W(V)| = O_p \left( b_1(K)(K/n + K^{-2\gamma_2} + \Delta^2_n b_2(K)^2)^{1/2} \right).
\]

The additional term in the mean-squared error convergence rate compared to, e.g, Newey, Powell, and Vella (1999) or Hahn, Liao, and Ridder (2018) (Section 3 of the Online Appendix) is \(\Delta^2_n b_2(K)^2\). It comes from the correlation between \(e^W\) and \(\hat{P} - P\).
We insert here a corollary stating the rate of convergence of the first order partial derivative of the two-step nonparametric estimator. This will be used in the proof of asymptotic normality. Its proof follows from the proof of Result 4.3.

**Corollary 4.1.** Under Assumptions 4.1 and 4.3, if \( \sup_{S_V} |\hat{c}h^W(V) - \hat{c}p^K(V)| \leq CK^{-\gamma_2} \),

\[
\sup_{V \in S_V} |\hat{c}h^W(V) - \hat{c}h^W(V)| = O_p \left( b_2(K)(K/n + K^{-2\gamma_2} + \Delta_n^2b_2(K)^2)^{1/2} \right).
\]

As mentioned earlier, Assumption 4.3 (3) 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 \( S_V = \bigtimes_{d \in d_2, t \in T} [\underline{y}_{td}, \bar{y}_{td}] \).

**Assumption 4.4.**

1. \( e^W \) is i.i.d and \( E((e^W)^2 | X, Z) \) is bounded on \( S_{X,Z} \),
2. \( h^W \) is Lipschitz on \( S_V \) and \( \rho^W \) is bounded on \( S_{X,Z} \),
3. \( p^K(\cdot) \) is the power series basis, and \( \forall V \in S_V, f_V(V) = \Pi_{d \in d_2, t \in T} (v_{t,d} - \underline{y}_{td})^{\alpha_2}(\bar{y}_{td} - v_{t,d})^{\alpha_2} \),
4. There exists \( \gamma_2 \) and \( \pi^K \) such that \( \sup_{S_V} |h^W(V) - \pi^K p^K(V)| \leq CK^{-\gamma_2} \),
5. \( b_1(K) = K^{\alpha_2 + 7/2}\Delta_n \rightarrow n \rightarrow 0 \) and \( b_2(K) = K^{3/2}/\sqrt{n} \rightarrow n \rightarrow 0 \).

**Result 4.4.** Under Assumptions 4.2 and 4.4, (32) and (33) hold.

For these results to apply to our choice of \( w_i = (M_i)_{s,t} \) and \( w_i = (M_i\hat{y}_i)_t \) for \( 1 \leq s, t \leq T - 1 \), we adapt Assumption 4.4 to the model primitives.

**Assumption 4.5.**

1. \( E(\hat{u}|X, Z) \) and \( \text{Var}(||\hat{u}|||X, Z) \) are bounded on \( S_{X,Z} \),
2. \( M \) and \( k \) are Lipschitz and \( g \) is bounded on \( S_V \),
3. \( p^K(\cdot) \) is the power series basis, and \( \forall V \in S_V, f_V(V) = \Pi_{d \in d_2, t \in T} (v_{t,d} - \underline{y}_{td})^{\alpha_2}(\bar{y}_{td} - v_{t,d})^{\alpha_2} \),
4. There exists \( \gamma_2 \) and \( \pi^K_{M,sl} \) and \( \pi^K_{k,sl} \) such that \( \sup_{S_V} |M_{sl}(V) - \pi^K_{M,sl} p^K(V)| \leq CK^{-\gamma_2} \)
and \( \sup_{S_V} |k_{sl}(V) - \pi^K_{k,sl} p^K(V)| \leq CK^{-\gamma_2} \) for all \( s, t \leq T - 1 \),
5. \( K^{\alpha_2 + 7/2} \Delta_n \rightarrow n \rightarrow 0 \) and \( K^{3/2}/\sqrt{n} \rightarrow n \rightarrow 0 \).

Under Assumptions 2.1, 4.2 and 4.5 the convergence rate of \( \hat{M} \) and \( \hat{k} \) in sup norm and mean square norms are therefore given by (32) and (33).

**Convergence rate for \( \hat{g} \):** Recall that \( \hat{g}(V) = \hat{M}(V)^{-1} \hat{k}(V) \). The rate of convergence of \( \hat{g}(\cdot) \) is obtained using continuity arguments. We will assume the following set of conditions.

**Assumption 4.6.** \( M \) and \( g \) are continuous on \( S_V \), \( S_V \) is a compact set and \( M(V) \) is invertible for all values \( V \in S_V \).
This implies that \( k = M g \) is continuous as well and that \( ||m||_\infty, ||M||_\infty \) and \( ||g||_\infty \) exist. Note that the continuity assumption somewhat overlaps with Assumption 4.5 (1) and (2) as the existence of a linear approximation relies on smoothness assumptions. Moreover while Assumption 2.3 requires the matrix to be invertible only \( \mathbb{P}_V \) a.s, we assume here that \( M(V) \) is invertible for all values in the support. Under these conditions, the following MSE and sup norm rates are obtained for \( \hat{g} \).

**Result 4.5.** Under Assumptions 4.2, 4.5 and 4.6, assuming \( b_1(K)^2(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2) \to 0, \)

\[
\int ||\hat{g}(V) - g(V)||^2 dF(V) = O_p(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2),
\]
\[
||\hat{g}(V) - g(V)||_\infty = O_p(b_1(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2}).
\]

### 4.3 Consistency of \( \hat{\mu} \)

Equipped with the convergence rate results on the nonparametric estimators, we can now show consistency of the APE estimator \( \hat{\mu} \). Recall that, writing \( \delta_i = \mathbb{1}(\text{det}(\tilde{X}_i^t \tilde{X}_i) > \delta_0) \) and \( Q_i^\delta = \delta_i Q_i \), we defined the estimator for \( \mathbb{E}(\mu|\delta) \) to be

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

Write \( \gamma_n = b_1(K)(K/n + K^{-2\gamma_2} + \Delta_n^2 b_2(K)^2)^{1/2} \).

**Assumption 4.7.** Assume \( \mathbb{E}(||Q^\delta||) < \infty \) and \( \mathbb{E}(||Q^\delta \hat{y}||) < \infty \).

**Result 4.6.** Suppose Assumptions 4.2, 4.4, 4.6, and 4.7 hold. Assume also \( \gamma_n \to 0, a_1(L) \Delta_n \to 0 \) and that \( g \) is continuously differentiable on \( S_V \). Then \( \hat{\mu} \to_{\mathbb{P}} \mathbb{E}(\mu|\delta) \).

### 4.4 Asymptotic normality

We now derive the asymptotic normality of \( \hat{\mu} \). The analysis is carried out in several steps. First, we modify the trimming function. We then explain how to linearize our estimator as a function of the nonparametric two-step sieve estimators. We obtain an asymptotic expansion of a general linear functional of nonparametric two-step sieve estimator which we then apply to the obtained linearization of our estimator. Finally we prove that the linear approximation is valid and derive asymptotic normality of \( \hat{\mu} \).

#### 4.4.1 Trimming

Recall that we defined \( \tilde{V}_i = \tau(\tilde{V}_i) \), where \( \tilde{V}_i = (\tilde{v}_{it})_{t \in T} \) is the vector of residuals from the sieve regression of \( x_{it}^2 \) on \( \xi_{it} = (x_{it}^1, z_{it}) \) and \( \tau \) projects onto \( S_V = \times_{d \in d_2, t \in T} [\xi_{id}; \tilde{v}_{it}] \). The proof of
asymptotic normality will use smoothness properties of $\tau$ and will require it to be twice differentiable, which is not the case when $\tau$ is the projection defined in the previous section. We thus change the definition of $\tau$ so that it now projects onto a bounded superset of $S_V$. Importantly we will not focus anymore on allowing for the density to be 0 on the boundary of $\tau$. 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 $V \in \mathbb{R}^{T_d}$, the $(d_2(t-1)+d)^{th}$ component of $\tau(V)$ is given by

$$
\tau(V)_{(t-1)k_2+d} = \begin{cases} 
v_{ld}, & \text{if } v_{ld} \in \mathbb{Z}_{ld}; \bar{v}_{ld}, \\
v_{ld} + \tau_\varsigma(v_{ld} - \bar{v}_{ld}), & \text{if } v_{ld} \leq \bar{v}_{ld}, \\
\bar{v}_{ld} - \tau_\varsigma(\bar{v}_{ld} - v_{ld}), & \text{if } v_{ld} \geq \bar{v}_{ld},
\end{cases}
$$

and define as before $\tilde{V}_i = \tau(\tilde{V}_i)$. The support of $\tau$ is $\mathbb{R}^{T_{d_2}}$ and we now have $\tilde{V}_i \in S_\tilde{V} = \times_{d \leq d_2, t \in T}[\bar{v}_{ld} - \varsigma; \bar{v}_{ld} + \varsigma]$. We will refer to $S_\tilde{V}$ as the “extended support”.

Each component of $\tau$ is a twice differentiable function of $V$, implying that $\tau$ itself is twice continuously differentiable. Moreover for all $V \in S_V$, $\partial \tau / \partial V = I_{T_{k_2}}$, which will imply that the derivative of a function $m$ composed with $\tau$ evaluated at $V$, $m(\tau(V))$, is equal to the derivative of $m(V)$ whenever $V \in S_V$. On the extended support, that is for all $V \in S_\tilde{V}$, $|\partial \tau / \partial V| \leq C$ and $|\partial^2 \tau / \partial V^2| \leq C$ 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 $m : S_V \to \mathbb{R}^p$ (for any given $p \in \mathbb{N}$) such that $m$ is twice continuously differentiable on $S_V$, we define $m^\varsigma : S_\tilde{V} \to \mathbb{R}^p$ an extension of $m$, twice continuously differentiable. That is, for all $V$ in $S_V$, $m^\varsigma(V) = m(V)$, and $m^\varsigma$ must be twice continuously differentiable on the extended support $S_\tilde{V}$. Note that if there exists a sequence of functions $(m_n)_{n \in \mathbb{N}}$ converging uniformly to $m^\varsigma$ on the extended support $S_\tilde{V}$, the sequence of restrictions of $(m_n)_{n \in \mathbb{N}}$ on $S_V$ converges uniformly to $m$. We previously used, for $g$ a function of the variable $V$, the norm $|g|_d = \max_{|l| \leq d} \sup_{V \in S_V} |\partial^l g(.)|$. A corresponding norm for the extended functions will change the supremum to a supremum over the extended support, i.e., $|g|^\varsigma_\tilde{d} = \max_{|l| \leq d} \sup_{V \in S_\tilde{V}} |\partial^l g(.)|$.

As was the case with our previous definition of $\tau$, $||\tilde{V}_i - V_i|| \leq ||\tilde{V}_i - V_i||$. This guarantees that our results on the sup-norm convergence rates of the nonparametric two-step estimators $\hat{M}$ and $\hat{k}$ and of their derivatives remain valid, provided some changes are made to the definition of the vector of basis functions $p^K(.)$ and to the approximation condition [4] of Assumption 4.3. 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 $M^\varsigma$ and $k^\varsigma$. Under these modified conditions, because the extended functions remain Lipschitz, the rates of convergence of the nonparametric two-step estimators to

34
the extended functions are the same, the rate of convergence of \( \hat{g} \) is unchanged and consistency of \( \hat{\mu} \) holds.

We point out here that we will not show that our asymptotic normality result applies to cases where the density of the regressors goes to zero on the boundaries of their support, as we did for consistency (see Assumption 4.2 and 4.4). Indeed in contrast to the consistency proof, we will use rates on the sup norm of the nonparametric estimates as well as of their derivatives when the suprema are defined over the extended support. This rules out a direct application of the approach allowing the density of the regressors to go to zero on the boundaries of their support. This approach would require Condition (3) of Assumption 4.4 to hold on \( S_V \), which cannot be true if the density of the regressors is 0 on the boundary of the original support. Computing rates on the extended support allowing for this case is beyond the scope of this paper. We therefore remain silent on the choice of the basis.

### 4.4.2 Linearization

We study asymptotic normality of \( \sqrt{n}(\hat{\mu} - E(\mu)) \) where we rewrite

\[
\hat{\mu} = \frac{1}{\sum_{i=1}^{n} \delta_i/n} \hat{\mu}^\delta,
\]

with \( \hat{\mu}^\delta = \sum_{i=1}^{n} Q_i^\delta [\hat{y}_i - \hat{g}(\hat{V}_i)]/n \). We will first study \( \hat{\mu}^\delta - E(\mu^\delta) \). We write \( G = ((b_t)_{t \in T}, k, \mathcal{M}) \) for a vector of generic functions with \( b_t : \mathcal{S}_t \rightarrow \mathbb{R}^d, k : \mathcal{S}_V \rightarrow \mathbb{R}^{T-1} \) and \( \mathcal{M} : \mathcal{S}_V \rightarrow \mathcal{M}_{T-1}(\mathbb{R}) \).

For clarity we choose to write \( G_0 = ((b_{0t})_{t \in T}, k_0, \mathcal{M}_0) \), 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 exponent \( \zeta \) and will display it to avoid confusion whenever necessary. We decompose

\[
\sqrt{n}(\hat{\mu}^\delta - E(\mu^\delta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i^\delta [\hat{y}_i - \hat{g}(\hat{V}_i)] - E(\mu^\delta),
\]

\[
= \frac{1}{\sqrt{n}} \left[ \sum_{i=1}^{n} [\delta_i \mu_i - E(\mu_i \delta_i)] + \sum_{i=1}^{n} Q_i^\delta \hat{u}_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Q_i^\delta g_0(\hat{V}_i) - E(Q^\delta g_0(V))] + \frac{1}{\sqrt{n}} [Q_i^\delta \hat{g}(\hat{V}_i) - E(Q^\delta g_0(V))] \right],
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\delta_i \mu_i - E(\mu_i \delta_i)] + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i^\delta \hat{u}_i - \sqrt{n} \mathcal{X}_n(G) - \mathcal{X}_n(G_0),
\]

where we define

\[
\chi(W_t, G) = Q_i^\delta \mathcal{M} \left( \tau \left[ (x_{it}^2 - b_t(\xi_{it}))_{t \in T} \right] \right)^{-1} k \left( \tau \left[ (x_{it}^2 - b_t(\xi_{it}))_{t \in T} \right] \right),
\]

\[
\mathcal{X}_n(G) = \frac{1}{n} \sum_{i=1}^{n} [\chi(W_i, G) - E(\chi(W_i, G_0))].
\]
where $\tau$ is as defined in Section 4.4.1 and ensures that the argument of $\mathcal{M}$ and $k$ lies in $\mathcal{S}_V$. We recall that $W_i = (X_i, V_i, Z_i, u_i, \mu_i, \alpha_i)$ stands for the whole vector of primitive variables. We write the variables as column vectors, e.g $V_i = (v_{i1}', \ldots, v_{iT}')'$. We also define

$$\mathcal{X}(\mathcal{G}) = \mathbb{E}(\chi(W_i, \mathcal{G})) - \mathbb{E}(\chi(W_i, \mathcal{G}_0)).$$

Note that $\mathcal{X}(\mathcal{G}_0) = 0$.

The decomposition as well as our choice of arguments in $\mathcal{X}$ make explicit the dependence of our estimator on the functions $(b_t)_{t \leq T}$. 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 $(b_t)_{t \leq T}$ highlights the latter aspect.

The two first terms in Equation (34) 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, $\sqrt{n}[\mathcal{X}_n(\mathcal{G}) - \mathcal{X}_n(\mathcal{G}_0)]$, which has a standard form except for its dependence on a composition of the infinite dimensional nuisance parameters. Specifically, we define a class of continuous functions $\mathcal{H}$ endowed with a pseudometric $||.||_\mathcal{H}$ such that $\mathcal{G} \in \mathcal{H}$. Arguments yielding asymptotic normality typically require the following set of conditions for our asymptotic analysis. Define $\mathcal{H}_\delta = \{\mathcal{G} \in \mathcal{H} : ||\mathcal{G} - \mathcal{G}_0||_0 \leq \delta\}$.

**Assumption 4.8.**

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

**Result 4.7.** Under Assumption 4.8,

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

Under Assumption 4.8, the asymptotic distribution of $\hat{\mu}$ depends on the asymptotic behavior of $\sqrt{n}\mathcal{X}_0'(\hat{\mathcal{G}})[\hat{\mathcal{G}} - \mathcal{G}_0]$, where we write $\mathcal{X}_0'(\mathcal{G})$ for the pathwise derivative of $\mathcal{X}$ at $\mathcal{G}_0$. It is a linear functional of a vector of nonparametric estimators.

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 [I] of Assumption 4.8, following Chen, Linton, and Van Keilegom (2003) and the choice of $||.||_\mathcal{H}$ will be driven by Condition [2]. The structure of our asymptotic analysis is as follows: we first derive the asymptotic
distribution of \( \sqrt{n} \lambda_0^{(G)}[\hat{G} - G_0] \) 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 4.8 holds for this choice. Using Result 4.7, we obtain the asymptotic distribution of the standardized \( \hat{\mu} \). We therefore focus here 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 \( \chi_0^{(k)}(W_i)[\hat{b}_i], (\chi_0^{(bt)}(W_i)[\hat{b}_i])_{t \leq T} \) and \( \chi_0^{(M)}(W_i)[\hat{M}] \) to be the partial pathwise derivatives of \( \chi \) with respect to \( k, b_t \) and \( \mathcal{M} \) (respectively) at the true value \( G_0 \), evaluated (respectively) at \( \hat{k}, \hat{b}_t \) and \( \hat{M} \). We have

\[
\chi_0^{(M)}(W_i)[\hat{M}] := \chi_0^{(M)}(W_i, G_0)[\hat{M}] = -Q_i^\delta M_0(V_i)^{-1}\hat{M}(V_i)M_0(V_i)^{-1}k_0(V_i)
\]

\[
= -Q_i^\delta M_0(V_i)^{-1}\hat{M}(V_i)g_0(V_i) = -[g_0(V_i)' \otimes (Q_i^\delta M_0(V_i)^{-1})] \text{Vec}(\hat{M}(V_i)),
\]

\[
\chi_0^{(k)}(W_i)[\hat{k}] := \chi_0^{(k)}(W_i, G_0)[\hat{k}] = Q_i^\delta M_0(V_i)^{-1}\hat{k}(V_i),
\]

\[
\chi_0^{(bt)}(W_i)[\hat{b}_i] := \chi_0^{(bt)}(W_i, G_0)[\hat{b}_i] = -Q_i^\delta \frac{\partial g_0}{\partial v_t}(V_i)\hat{b}_i(\xi_t)
\]

where \( v_0 \) denotes the \( t^{th} \) component of \( V \), and where \( \frac{\partial g_0}{\partial v_t}(V_i) \) is a Jacobian matrix of size \((T - 1) \times d_2\). Note that the function \( \tau \) 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 \( \tau \) simplifies to the identity function on \( S_V \) and its Jacobian is the identity matrix.

We define \( \lambda_0^{(k)}[\hat{k}] \) the partial pathwise derivative of \( \lambda \) with respect to \( k \) at \( G_0 \) and evaluated at \( \hat{k} \), and similarly \( \lambda_0^{(M)}[\hat{M}] \) and \( (\lambda_0^{(bt)}[\hat{b}_i])_{t \leq T} \). Assuming one can interchange expectation and differentiation, we follow Mammen et al (2016) and write

\[
\lambda_0^{(G)}[G - G_0] = \lambda_0^{(k)}[k - k_0] + \lambda_0^{(M)}[M - M_0] + \sum_{t=1}^{T} \lambda_0^{(bt)}[b_t - b_{0,t}],
\]

\[
= \int_{V} [\lambda_M(v) \ \text{Vec}((M - M_0)(v))] dF_V(v) + \int_{V} [\lambda_k(v) (k - k_0)(v)] dF_V(v)
\]

\[
+ \sum_{t=1}^{T} \int \lambda_{bt}(\xi_t) (b_t - b_{0,t})(\xi_t) dF_{\xi_t}(\xi_t),
\]

where the functions \( \lambda \) are defined using the partial pathwise derivatives as

\[
\lambda_M(v) = -E \left( g(V_i)' \otimes (Q_i^\delta M_0(V_i)^{-1}) | V_i = v \right) = -g(v)' \otimes \left[ E(Q_i^\delta | V_i = v) M_0(v)^{-1} \right],
\]

\[
\lambda_k(v) = E(Q_i^\delta M_0(V_i)^{-1} | V_i = v) = E(Q_i^\delta | V_i = v) M_0(v)^{-1},
\]

\[
\lambda_{bt}(\xi_t) = -E \left( Q_i^\delta \frac{\partial g(V_i)}{\partial v_t} | \xi_t = \xi_t \right).
\]
4.4.3 Linear application of a nonparametric two-step sieve estimator

To obtain the asymptotic properties of a linear functional of nonparametric two-step series estimators, we now return to Model (31) and treat the general case. The object of interest in this section is the value of a linear function \( a \) evaluated at \( h^W \) where \( h^W (v) = \mathbb{E}(W|V = v) \). We use the nonparametric two-step sieve estimator \( \hat{h}^W \).

Functionals of nonparametric estimators have been widely studied for different types of nonparametric estimators (see, e.g., Newey (1994b) for kernel estimators and Newey (1997) for series estimators). However the linear functional here is also evaluated at the more complicated two-step nonparametric estimators constructed in the previous sections. Its asymptotic distribution cannot be derived directly from the aforementioned results. Hahn, Liao, and Ridder (2018) derive asymptotic normality results for nonlinear functionals of two-step nonparametric sieve estimators when the sieve estimators are from a general class of nonlinear sieve regression estimators. Characterizing the finite sample variance, they provide a practical estimator arguing that the asymptotic variance of our estimator can be obtained for a class of models where the orthogonality condition between the first and second stage does not hold.

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 sieve estimators.

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 \( W, V, e^W \) and the functions \( h^W \) and \( \rho^W \), the term \( \sqrt{n} (a(h^W) - a(\hat{h}^W)) \) as \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s^W_{i,n} + o_P(1) \). We will then apply the derived results to \( X_0^{(k)} \{ \hat{k} - k_0 \} \) and \( X_0^{(M)} \{ \hat{M} - M_0 \} \).

We consider the case where \( w \in \mathbb{R}, V \in \mathbb{R}^{T_d^2}, a(h) \in \mathbb{R}^{d_a} \). Define \( \rho_i^W = \rho^W(X_i, Z_i) \) and the following matrices,

\[
\Pi_t^W = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i(\partial h_i^W(V_i)/\partial v_t \otimes r'_{it}), \\
A = (a(p_{1K}), \ldots, a(p_{KK})), \\
d\Pi_t^W = \frac{1}{n} \sum_{i=1}^{n} \rho_i^W \left( \frac{\partial p_i^K(V_i)}{\partial v_t} \otimes r'_{it} \right), \\
dP_t^W = \mathbb{E}(\rho_i^W \left( \frac{\partial p_i^K(V_i)}{\partial v_t} \otimes r'_{it} \right)),
\]

where \( h_i^W \) is the functional extension of \( h^W \) and where the equalities on the first two matrices hold because \( h^s \) and \( h \) are equal on \( S_V \). Recall that by \( ||\tau(v_1) - \tau(v_2)|| \leq ||v_1 - v_2|| \), under Assumption
we have $\frac{1}{n} \sum_{i=1}^{n} ||v_i - \hat{v}_i||^2 = O_P \left( L/n + L^{-2\gamma_1} \right) = O_P(\Delta_n^2)$.

Assumption 4.9.

1. The data $W_i$ is i.i.d.
2. $||a(g)|| \leq C|g|_0$.
3. $h^W$ is twice continuously differentiable with bounded first and second derivatives, and $p^W$ is bounded,
4. There exists $\gamma_1$ and $\beta_L^2$ such that for all $t \leq T$, $\sup \mathbb{S}_{\xi_t} ||b_1(\xi_t) - \beta_L^t r_L(x_t, z_t)|| \leq C L^{-\gamma_1}$.
   There exists $\gamma_2$, $\pi^K_W$ such that $\sup \mathbb{S}_{\xi_t} ||h^W_t(v) - p^K(v) \pi^K_W|| \leq C K^{-\gamma_2}$.
5. For all $t \leq T$, there exists $\Gamma_1$, a $L \times L$ nonsingular matrix such that for $R^L_t(\xi_t) = \Gamma_1 r^L_t(\xi_t)$, $\mathbb{E}(R^L_t(\xi_t) R^L_t(\xi_t)' )$ has smallest eigenvalue bounded away from 0 uniformly in $L$. There exists $\Gamma_2$, 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 0 uniformly in $K$.
6. $||A||$ is bounded.
7. For $|R^L_t(\xi_t)|_0 \leq a_1(L)$, $|P^K(V)|_0 \leq b_1(K)$, $|P^K(V)|_2 \leq b_2(K)$, $|P^K(V)|_2 \leq b_3(K)$, we have $\sqrt{n} K^{-\gamma_2} = o(1)$, $\max(\sqrt{K}, \sqrt{L} b_2(K)) a_1(L) \sqrt{L/n} = o(1)$, $b_2(K) \sqrt{n L^{-\gamma_1}} = o(1)$, $b_2(K) (\sqrt{L/n} + L^{-\gamma_1}) K = o(1)$, $b_3(K) [L/\sqrt{n} + \sqrt{L^{-2\gamma_1}}] = o(1)$, $b_2(K)^2 \sqrt{K} [\sqrt{L/n} + L^{-\gamma_1}] = o(1)$.
8. $\mathbb{E}(||v_t||^2 | \xi_t)$ and $\text{Var}(e^W \gamma | X, Z)$ are bounded on $\mathbb{S}_{\xi_t}$ and $\mathbb{S}_{X,Z}$ respectively.

As stated above, we do not specify the sieve basis.

Lemma 4.1. Under Assumptions 4.1 and 4.9

\[
\sqrt{n}[a(\hat{W}) - a(W)] = \\
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A \mathbb{E}(p_t p_t')^{-1} \left[ p_t e^W_t + \sum_{t=1}^{T} (H^W_t - d P^W_t)(I_{k_2} \otimes \mathbb{E}(r_t r_t')^{-1}) (v_t \otimes r_t) \right] + o_P(1).
\]

We use the proof techniques of Lemma 2 of Newey, Powell, and Vella (1999) to obtain this approximation. However as mentioned previously, an essential orthogonal condition they assumed, namely the conditional mean independence of $w - \mathbb{E}(w | V)$ of $(X, Z)$, does not hold in our model. For this reason we obtain an extra term depending on $\rho^W$, the term $d P^W_t$ which would be zero if $\rho^W(X, Z) = 0$. Another difference is the summation over $t$ of the $H_t^W$ and $d P^W_t$ terms due to vector of control variables being composed of $T$ estimated residuals coming from $T$ different cross-section regressions.

Assumption 4.9 (6) is a condition on the generic functional $a$ applied to the elements of the approximating basis. The functionals appearing in $\hat{\mu}$ are derived from the linearization of $X$. They all take the form of an expectation $a(W) = \int \lambda_a(v) h^W(v) dF_V(v)$. This is exactly the mean square continuity condition of Newey (1997), which he shows is sufficient to obtain $\sqrt{n}$ asymptotic
normality of linear functionals of linear sieve estimators. We similarly exploit properties implied by this specification of $a$ but instead obtain an intermediary result on the mean square convergence of the term in bracket on the RHS of Lemma 4.1. We will use this in the following section to show that Condition (6) of Assumption 4.9 holds and to then obtain the total asymptotic variance matrix of $\sqrt{n}\lambda_0^{(G)}[\hat{G} - G_0]$.

**Assumption 4.10.**

1. There exists a function $\lambda_a: \mathbb{R}^{T_2d} \rightarrow \mathbb{R}^{d_a}$ such that $a(h^W) = \int \lambda_a(v) h^W(v) dF_Y(v)$,
2. There exists $t^K_a$ such that $|\lambda_a(V) - t^K_a p^K(V)|_1 = O(K^{-\gamma})$, and $(t^K_{aht}, t^K_{aopt})$ such that as $L \rightarrow \infty$,
$$E\left(||E\left[\lambda_a(V) \frac{\partial h^W(V)}{\epsilon_{v_t}} | \xi_t\right] - t^{L}_{aht} r^L(\xi_t)||^2\right) \rightarrow 0 \text{ and } E\left(||E\left[p_i^W \frac{\partial \lambda_a(V)}{\epsilon_{v_t}} | \xi_t\right] - t^{L}_{aopt} r^L(\xi_t)||^2\right) \rightarrow 0,$$
3. $b_2(K)K^{-\gamma} = o(1)$,
4. $\text{Var}(e_i^W | V)$ is bounded on $S_V$.

Define $\tilde{\lambda}_a(v) = A E(p_i p_i')^{-1} p^K(v)$, $\tilde{\lambda}^{\partial}_{a,td}(\xi_t) = A E(p_i p_i')^{-1} H^{W}_{td} E(r_{it'} r_{it})^{-1} r^L(\xi_t)$ and $\tilde{\lambda}_{a,td}(\xi_t) = A E(p_i p_i')^{-1} dP^W_{td} E(r_{it'} r_{it})^{-1} r^L(\xi_t)$. Write Assumption 4.9 for Assumption 4.9 without Condition (5).

**Lemma 4.2.** Under Assumptions 4.9 and 4.10, as $K, L \rightarrow \infty$, $E(||e^W(\tilde{\lambda}_a(V) - \lambda_a(V))||^2) \rightarrow 0$, $E\left(||v_{td}\left[\tilde{\lambda}_{a,td}(\xi_t) - E\left(\lambda_a(V) \frac{\partial h^W(V)}{\epsilon_{v_t}} | \xi_t\right)\right]||^2\right) \rightarrow 0$ and $E\left(||v_{td}\left[\tilde{\lambda}_{a,td} - E\left[p_i^W \frac{\partial \lambda_a(V)}{\epsilon_{v_t}} | \xi_t\right]\right]||^2\right) \rightarrow 0$.

Note that Condition (2) is a sup norm rate condition on both $\lambda_a$ and $\partial \lambda_a/\partial V$, stronger than only assuming $E(||\lambda_a(V) - t^K_a p^K(V)||^2) \rightarrow 0$ as is assumed in Newey (1997) (Assumption 7) to obtain $\sqrt{n}$ asymptotic normality. Loosely speaking, because the $dP^W_i$ includes derivatives of the vector of basis functions, left multiplication of $dP^W_i$ by the matrix $A$, which is the matrix of expectations of $\lambda_a$ multiplied by the functions, will yield under Condition (2) an approximation of the derivative of $\lambda_a$. This will then appear in the asymptotic variance of our estimator when applied to the specific functionals.

4.4.4 Application to the model

4.4.4.1 Asymptotics of the linear part

By Result 4.7, under Assumption 4.8, $\sqrt{n}[\lambda_n(\hat{G}) - \lambda_n(G_0)]$ is asymptotically equivalent to $\sqrt{n}\lambda_0^{(G)}[\hat{G} - G_0]$. $\lambda_n^{(G)}$ is a sum of linear functionals applied to the components of $\hat{G} - G_0$, where $\hat{G} = ((b_t)_{t \in T}, k, \mathcal{M})$, see (35). We thus apply Lemma 4.1 to these functionals, choosing $w_i$ to be either a component of $M_i \hat{y}_i$ or of $M_i$. By analogy with the general model, to apply Lemma 4.1 to $\lambda_0^M$ we define the following objects

$$e_i^M = M_i - E(M_i | V_i) = M_i - M_0(V_i),$$
\[e_i^M = M_i - \mathbb{E}(M_i | X_i, Z_i) = 0,\]
\[\rho^M(X_i, Z_i) = \mathbb{E}(M_i | X_i, Z_i) - \mathbb{E}(M_i | V_i) = M_i - \mathcal{M}_0(V_i),\]
and similarly for \(\mathcal{X}_0^k[\hat{k}].\)

\[e_i^k = M_i \hat{y}_i - \mathbb{E}(M_i \hat{y}_i | V_i) = [M_i - \mathcal{M}_0(V_i)]g(V_i) + M_i \hat{u}_i,\]
\[e_i^{k*} = M_i \hat{y}_i - \mathbb{E}(M_i \hat{y}_i | X_i, Z_i) = M_i [\hat{u}_i - \mathbb{E}(\hat{u}_i | X_i, Z_i)],\]
\[\rho^k(X_i, Z_i) = \mathbb{E}(M_i \hat{y}_i | X_i, Z_i) - \mathbb{E}(M_i \hat{y}_i | V_i) = [M_i - \mathcal{M}_0(V_i)]g(V_i) + M_i \mathbb{E}(\hat{u}_i | X_i, Z_i).\]

It will be convenient to assume \(\mathbb{E}(\hat{u}_i | X_i^1, X_i^2, Z_i) = 0.\) We define now the analogs of the matrices \(A, dP_t^W\) and \(H_t^W.\) For a given \(v\) \(\lambda_M^j(v)\) is the \(j^{th}\) column of the matrix \(\lambda_M(v)\) and

\[\Lambda^M = \int_v \left[ \lambda_M^1(v)p^K_1(v), \ldots, \lambda_M^s(v)p^K_s(v), \ldots, \lambda_M^{(T-1)^2}(v)p^K_2(v), \ldots, \lambda_M^{(T-1)^2}(v)p^K_s(v) \right] dF_V(v),\]
of dimension \(d_x \times K(T - 1)^2.\) Define similarly the matrix \(\Lambda^k.\) We will interchangeably index the columns of \(\lambda_M\) as \(\lambda_M^d\) with \(d \leq (T - 1)^2\) and as \(\lambda_M^s\) with \(1 \leq s, t \leq T - 1.\) We will also need

\[H_t^M = \mathbb{E}\left[ \frac{\partial \text{Vec}(\mathcal{M}_0(V_i))}{\partial v_t} \otimes p_i \otimes r_i' \right],\]
\[H_t^k = \mathbb{E}\left[ \frac{\partial^2 V_0(V_i)}{\partial v_t} \otimes p_i \otimes r_i' \right],\]
\[dP_t^M = \mathbb{E}\left[ \text{Vec}(\rho_t^M) \otimes \frac{\partial p^K(V_i)}{\partial v_t} \otimes r_i' \right],\]
\[dP_t^k = \mathbb{E}\left[ \rho_t^k \otimes \frac{\partial p^K(V_i)}{\partial v_t} \otimes r_i' \right].\]

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

\[\Lambda^t = \int_{\xi_t} \left[ \lambda_{bt}^1(\xi_t)r_1^t(\xi_t), \ldots, \lambda_{bt}^1(\xi_t)r_L^t(\xi_t), \ldots, \lambda_{bt}^{T-1}(\xi_t)r_1^t(\xi_t), \ldots, \lambda_{bt}^{T-1}(\xi_t)r_L^t(\xi_t) \right] dF_{\xi_t}(\xi_t).\]

We now state the assumptions required to apply Lemma 4.1 on the functionals \(\mathcal{X}_0^{(M)}\) and \(\mathcal{X}_0^{(k)}\) applied respectively to \(\mathcal{M}\) and \(\mathcal{h}\) where as mentioned in the discussion before Assumption 4.9 we do not specify the basis of approximating functions.

**Assumption 4.11.**

1. \(\mathcal{M}_0\) and \(g_0\) are twice continuously differentiable with bounded first and second order derivatives,
2. \(\mathbb{E}(||Q_i^1||) < \infty, \mathbb{E}(||\hat{u}_i|| | X, Z)\) is bounded on \(S_{X,Z}\), and \(S_{\xi_t}\) for all \(t \leq T\) and \(S_V\) are bounded,
3. There exists \(\gamma_1\) and \(\beta_t^k\) such that for all \(t \leq T, \sup_{S_{\xi_t}} ||b_{0t}(\xi_t) - \beta_t^k r_L^t(\xi_t)|| \leq C L^{-\gamma_1}.\) There exists \(\gamma_2, \pi_M^K\) and \(\pi_k^K\) such that \(\sup_{S_V} ||\mathcal{M}_0^{(v)}(v) - p^K(v)\pi_M^K|| \leq C K^{-\gamma_2}\) and \(\sup_{S_V} ||b_{0t}^{(v)}(v) - p^K(v)\pi_k^K|| \leq C K^{-\gamma_2},\)
4. For all $t \leq T$, there exists $\Gamma_{1t}$, a $L \times L$ nonsingular matrix such that for $R_{iL}^t(\xi_t) = \Gamma_{1t}r_{iL}^t(\xi_t)$, $\mathbb{E}(R_{iL}^t(\xi_t)R_{iL}^t(\xi_t)'$ has smallest eigenvalue bounded away from 0 uniformly in $L$. There exists $\Gamma_2$, a $K \times K$ nonsingular matrix such that for $P^K(V) = \Gamma_2p^K(V)$, $\mathbb{E}(P^K(V)P^K(V)')$ has smallest eigenvalue bounded away from 0 uniformly in $K$.

5. $||\Lambda^M||$, $||\Lambda^k||$, and $||\Lambda^M||$ are bounded.

6. For $|R_{iL}^t(\xi_t)| \leq a_1(L)$, $|P^K(V)|_0 \leq b_1(K)$, $|P^K(V)|_1 \leq b_2(K)$, $|P^K(V)|_2 \leq b_3(K)$, we have $\sqrt{n}K^{-\gamma_2} = o(1)$, $\max(\sqrt{K}, \sqrt{L}b_2(K))a_1(L)\sqrt{L/n} = o(1)$, $b_2(K)\sqrt{L}K^{-\gamma_1} = o(1)$, $b_2(K)\sqrt{L}K^{-\gamma_1} = o(1)$, $b_3(K)[L/\sqrt{n} + \sqrt{n}L^{-2\gamma_1}] = o(1)$, $b_2(K)^2\sqrt{K}[\sqrt{L/n} + L^{-\gamma_1}] = o(1)$.

7. $\mathbb{E}(\hat{u}|X, Z) = 0$, and $\mathbb{E}(||v_t||^2|\xi_t)$ and $\mathbb{E}(||\hat{u}||^2|X, Z)$ are bounded on $S_{\xi_t}$ and $S_{X, Z}$ respectively.

Assumption 4.11 (7) is imposed to simplify computations. It amounts to strengthening the control function assumption, that is, Assumption 2.1 (2). Applying Lemma 4.1, we obtain the following linearization.

**Result 4.8.** Under Assumptions 2.1, 4.6 and 4.11,

\[
\sqrt{n}\lambda_0^{(G)}(\hat{G} - G_0) = o_P(1)
\]

\[
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda^M(I_{(T-1)^2} \otimes \Theta) \text{Vec}(e_i^M) \otimes p_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda^k(I_{T-1} \otimes \Theta) e_i^k \otimes p_i
\]

\[
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} \left[ \Lambda^M(I_{(T-1)^2} \otimes \Theta)(H_{iL}^T - dP_{iL}^T) + \Lambda^k(I_{T-1} \otimes \Theta)(H_{iL}^k - dP_{iL}^k) + \Lambda^M \right] (I_{k2} \otimes \Theta_1) v_{it} \otimes r_{it},
\]

where we define $\Theta = \mathbb{E}(p_i\hat{r}_i)$ and $\Theta_1 = \mathbb{E}(r_i\hat{r}_i)$.

Note that we can now write $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \delta_i\mu_i - \mathbb{E}(\mu\delta) \right] + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i \hat{u}_i = \sqrt{n}\lambda_0^{(G)}(\hat{G} - G_0)$ as $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{i,n} + o_P(1)$ where

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

\[
+ \sum_{t=1}^{T} \left[ \Lambda^M(I_{(T-1)^2} \otimes \Theta)(dP_{iL}^T - H_{iL}^T) + \Lambda^k(I_{T-1} \otimes \Theta)(dP_{iL}^k - H_{iL}^k) + \Lambda^M \right] (I_{k2} \otimes \Theta_1) v_{it} \otimes r_{it}.
\]

We define $\Omega = \text{Var}(s_{i,n})$ and the following objects, $\hat{Q}_i = Q_i - \mathbb{E}(Q_i|V_i)M_0(V_i)^{-1}M_i$, and $\Omega_0 = \text{Var} \left( \left[ \delta_i\mu_i - \mathbb{E}(\mu\delta) \right] + \hat{Q}_i \hat{u}_i + \sum_{t=1}^{T} \mathbb{E} \left( \hat{Q}_i \hat{u}_i | v_{it} \right) v_{it} \right)$.

**Assumption 4.12.**

1. For each function $\lambda_a(\cdot)$, column of $\lambda_M(\cdot)$ or $\lambda_k(\cdot)$, $\lambda_a(\cdot)$ is continuously differentiable and there exists $i_{aL}^K$ such that $|\lambda_a(V) - i_{aL}^K p^K(V)|_1 = O(K^{-\gamma_3})$, and $(i_{aht}, i_{apL})$ such that as $L \rightarrow \infty$,

\[
\mathbb{E} \left( ||||\lambda_a(V) - i_{aht} p^K(V)\xi_t||_1^2 \right) \rightarrow 0 \quad \text{and} \quad \mathbb{E} \left( ||||\lambda_a(V) - i_{apL} p^K(V)\xi_t||_1^2 \right) \rightarrow 0.
\]

For all $t \leq T$, $b_t$ is continuous and there exists $i_{aL}^L$ such that $\mathbb{E} \left( ||\lambda_a(V) - i_{aL}^L p^K(V)\xi_t||_1^2 \right) \rightarrow 0$ as $L \rightarrow \infty$. 

42
2. \(b_2(K)K^{-\gamma_3} = o(1)\),  
3. \(\mathbb{E}(Q^2_t) < \infty, \mathbb{E}(\|\mu_t\|^2) < \infty\), and there exists \(C > 0\) such that \(\Omega_0 \geq CI_{d_e}\). 

The condition \(\Omega_0 \geq CI_{d_e}\) holds if for instance \(\text{Var}(\mu_i|X_i, Z_i, u_i, V_i) \geq CI_{d_e}\) for some \(C > 0\), or if a similar condition holds on the conditional variance of \(\hat{u}_i\), as is typically assumed. We now state the result giving the asymptotic variance of the estimator and guaranteeing that Assumption 4.11 (5) holds. The boundedness of the two last matrices is added for later results on asymptotic normality. We will write Assumption 4.11 for Assumption 4.11 without its condition (5).

**Result 4.9.** Under Assumptions 2.1, 4.6, 4.11, 4.12, \(\Omega^{-1/2} \to n \to \infty \Omega_0^{-1/2}\). Moreover, \(\|\Lambda^M\|, \|\Lambda^k\|, \|\Lambda^M(I(T-1)^2 \otimes \Theta)(H^M_t - dP^M_t)\|, \text{ and } \|\Lambda^k(I_{T-1} \otimes \Theta)(H^k_t - dP^k)\|\) are bounded.

We now know that under Assumption 4.12, Assumption 4.11 (5) holds. Thus, under Assumption 4.11 and Assumption 4.12, Equation (36) on \(\sqrt{n}\chi^2_0[G - G_0]\) holds.

### 4.4.4.2 Asymptotic distribution of \(\hat{\mu}\)

We now assemble the arguments of Section 4.4.2 and 4.4.4.1. Recall that if Assumption 4.8 holds, Result 4.7 will guarantee that \(\sqrt{n}(\hat{\mu} - \mathbb{E}(\mu|\delta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{i,n} + o_p(1)\). We thus focus now on showing that Assumption 4.8 does hold.

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

Our choice of the norm on \(H, \|\cdot\|_{\mathcal{H}}\), is justified by Condition (2). The functional \(\chi\) is a function of \(\mathcal{M}\), \(k\) and \((b_t)_{t \in T}\), where \(\mathcal{M}\) and \(k\) are composed with \((b_t)_{t \in T}\). 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 \(\chi(G)\) and \(\chi(G) - \chi(G_0)[G - G_0]\) can be easily controlled by, among other terms, the distance between first order partial derivatives of these functions. A natural norm on \(H_{e,c,c'}\) is therefore \(\|\cdot\|_{\mathcal{H}} = \sum_{j=1}^{(T-1)^2} |\mathcal{M}_j - \mathcal{M}_{0,j}|^i + \sum_{j=1}^{T-1} |k_j - k_{0,j}|^i \geq \sum_{t \in T} \sum_{j=1}^{d_{c'}} |b_{t,j} - b_{0,t,j}|^\infty\) where by an abuse of notation \(\mathcal{M}_j\) is the \(j^{th}\) component of \(\text{Vec}(\mathcal{M})\). This norm is our choice of norm in the remainder of this section.
Assumption 4.13. \( \varrho > \max(Td_2, d_z + d_1)/2. \)

Result 4.10. Defining \( \mathcal{H} = \mathcal{H}_{c,c}^0 \) and \( ||.||_{\mathcal{H}} \) as described, if Assumption 4.11', 4.11, 4.12 and 4.13 hold, then Assumption 4.11 and 4.12 hold.

We chose \( \mathcal{H} \) and \( ||.||_{\mathcal{H}} \) and provided a set of conditions guaranteeing that Conditions 4.11' and 4.12 of Assumption 4.1 hold. Condition 4.11' is a condition on the convergence rate of the estimators \( \hat{b}_t, \hat{k} \) and \( \hat{\mathcal{M}}_j \). The rate of convergence of \( ||\hat{b}_{t,j} - b_{0,t,j}||_{\infty} \) for all \( (t, j) \) is given by Equation 4.1, see the Proof of Result 4.1. The rates of convergence of \( |\hat{k}_j - k_{0,j}|_1 \) and \( |\hat{\mathcal{M}}_j - \mathcal{M}_{0,j}|_1 \) are given by Assumption 4.11. The conditions required to apply this corollary must be adapted to the extended support, as we did for other results. Assumption 4.11 already includes most of these conditions, specifying an approximation rate of \( \mathcal{M}_0 \) and \( k_0 \) over the extended support and defining the rates \( b_1, b_2 \) and \( b_3 \) as bounds on sup-norms of derivatives of \( p^K \) defined over the extended support. Only a slight modification of Condition 4.11' needs to be added.

“There exists \( \gamma_1 \) and \( \beta_1 \) such that for all \( t \leq T \), \( \sup_{\xi_t} ||g^{2i}(\xi_t) - g^{2i,t'}(\xi_t)|| \leq C \sqrt{\gamma_1}. \) There exists \( \gamma_2, \pi_{M,1}^K \) and \( \pi_{k,1}^K \) such that \( |\mathcal{M}_0(.)|_{st} - p^K(.)\pi_{M,1}^K|^{1/2} \leq C \sqrt{\gamma_2} \) and \( |k_0(.)|_{st} - p^K(.)\pi_{k,1}^K|^{1/2} \leq C \sqrt{\gamma_2}, \) for all \( 1 \leq s, t \leq T - 1. \)

This modification is a stronger assumption, changing the approximation rate to be over the \( |.||_1 \) norm instead of the sup norm. Assumption 4.11 is the modified version of Assumption 4.11. We can now state the following result.

Result 4.11. Under Assumptions 2.1, 4.6, 4.11, 4.12 and 4.13, assuming moreover that \( a_1(L)\Delta_n = o(n^{-1/4}) \) and \( b_2(K)[K/n + K^{-2\gamma_2} + \Delta_n^2b_2(K)^2]^{1/2} = o(n^{-1/4}) \), then

\[ \left[ \hat{\mu} - \mathbb{E}(\mu|\delta) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{i,n} + o_P(1). \]

Assumption 4.14. \( \mathbb{E}( ||\mu_i - \mathbb{E}(\mu)|^4 || ) < \infty, \mathbb{E}( ||\xi_t^\delta||^4 ) < \infty \) and \( \text{Var}(\delta_i) > 0. \) Also, \( \mathbb{E}( ||v_t||^4 |\xi_t \) and \( \mathbb{E}( ||v_{it}||^4 |X, Z) \) are bounded on \( \mathcal{S}_{\xi_t} \) for all \( t \leq T \) and on \( \mathcal{S}_{X,Z} \) respectively.

We can now state the main result of this section.

Result 4.12. Under Assumptions 2.1, 4.6, 4.11, 4.12, 4.13 and 4.14, assuming moreover that \( a_1(L)\Delta_n = o(n^{-1/4}) \) and \( b_2(K)[K/n + K^{-2\gamma_2} + \Delta_n^2b_2(K)^2]^{1/2} = o(n^{-1/4}), \)

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

where \( \Xi = \Omega_0 + \mathbb{E}( (\delta_i - \Phi)s_i )\mathbb{E}(\mu|\delta)' + \mathbb{E}(\mu|\delta)\mathbb{E}( (\delta_i - \Phi)s_i)' + (\Phi - \Phi^2)\mathbb{E}(\mu|\delta)\mathbb{E}(\mu|\delta)'. \)
5 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.

\[
y_{it} = x_{it}^1 \mu_i^1 + x_{it}^2 \mu_i^2 + \sin(3v_{it}) + u_{it}, \quad i = 1..n, \ t \leq T.
\]

where the random coefficients are drawn according to

\[
\mu_i = A \nu_i, \quad \text{with } A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \quad \text{and } \nu_i^1 \sim U[0, 1], \ \nu_i^2 \sim U[0, 1], \ \nu_i^1 \parallel \nu_i^2,
\]

and the specification for the covariates, instruments and time-varying disturbances are, for all \(t \leq T\),

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

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 \(\mu^1\) has support \([0,3]\), \(\mathbb{E}(\mu_i^1) = 1.5\), \(\text{Var}(\mu_i^1) = 5/12\), and that \(\mu^2\) has support \([0,4]\), \(\mathbb{E}(\mu_i^2) = 2\) and \(\text{Var}(\mu_i^2) = 5/6\). The design implies that the support of \(x_i^1\) is \([0, 6.58]\) and the support of \(x_i^2\) is \([-0.5, 4.2]\). 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 \(\mathbb{E}(|Q_i \nu_i|) < \infty\) holds, we imposed \(z_{it}\) and \(x_{it}^1\) to depend multiplicatively on \(\mu_i^1\) and \(\mu_i^2\) raised to the power \(1/4\), following an observation made in Graham and Powell (2012).

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 \(x^1\) and \(z\), and estimation of the functions \(\mathcal{M}(.) = \mathbb{E}(M_i|V_i = .)\) and \(k(.) = \mathbb{E}(M_i \hat{y}_i|V_i = .)\). The conditional expectation of \(x^2\) is used to construct the generated covariates \(V_i\). Recall that \(g(V) = \mathcal{M}^{-1}(V)k(V)\). For each of the simulation draws \(r\), an estimate \(\hat{g}^r\) of the function \(g\) is computed. We report in Figure 1 the pointwise average of these estimates \(\bar{g}(V) = \sum_{r \in R} \hat{g}^r(V)/R\) as well as the 5th and 95th quantiles \(g^{0.05}(V)\) and \(g^{0.95}(V)\) 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 scale for each coefficient. These plots are compatible with the asymptotic normality result of Section 4.4. It is noticeable that the variance of the estimator of $E(\mu_2)$, which is the average partial effect of the endogenous variable, is larger than the variance of the estimator of $E(\mu_1)$. 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, $\hat{\mu}^{CRC} = \sum_{i=1}^{n} Q_{it}y_{it}/n$: 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.

6 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

$$\ln h_{it} = \alpha_i + \ln \omega_{it} \mu_i + \chi_{it}x + \epsilon_{it}, \quad i = 1..n, \ t = 1..T,$$

(37)

where $h_{it}$ is the number of annual hours worked, $\omega_{it}$ is the hourly wage, $\chi_{it}$ is composed of additional demographics. The individual elasticity of intertemporal substitution $\mu_i$ enters the individual utility function, and heterogeneity in preferences may covary with $\omega_{it}$ and $\chi_{it}$. 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 (37) 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 $E(\epsilon_{it} | \chi_{is}) = 0$, for all $s < t$. 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 $E(\epsilon_{it} | \ln \omega_{is}) = 0$ for all $s < t$. 

46
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 $0$, $-1$, $\ldots$, $-\tau$. Define $v_{it} = \ln \omega_{it} - \mathbb{E}(\ln \omega_{it}|\chi_{i1}, \ln \omega_{i0})$ and $V_t = (v_{it})_{t \leq T}$. We assume that there exists $(f_t(.))_{t \leq T}$ such that

$$
\text{for all } t \leq T, \quad \mathbb{E}(\varepsilon_{it}|V_t, \ln \omega_{i0}, \chi_{i1}, \chi_{i0}, \ldots, \chi_{i-\tau}) = f_t(V_t),
$$

where the normalization condition $\mathbb{E}(f_t(V_t)) = 0$ holds. Here, $(\chi_{i1}, \chi_{i0}, \ldots, \chi_{i-\tau})$ corresponds to the set of additional instruments mentioned in Section 2.4.2. By sequential exogeneity of $\chi_{it}$, its values in $s$ for $2 \leq s \leq T$ cannot be in this set of instruments. For the same reason, we do not use $\chi_{it}$ as an instrument to construct the control variables $v_{it}$. Instead, we use the initial values $\chi_{i1}$ and $\ln \omega_{i0}$. This is similar to the approach described in Section 3.3. The conditional expectation equation (38) holds if for instance for all $t$, $(\ln \omega_{is-1}, \chi_{is}, \chi_{is-1}, \ldots, \chi_{i-\tau})_{s \leq t} \| (v_{is'}, \varepsilon_{is'})_{s' \geq t}$, a condition which would also imply the moment conditions used in Ziliak (1997).

Defining $M_i$ as in Section 2, we need $T \geq 3$ 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 $\chi_{it}$ 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 $\chi_{it}$ 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

$$
\ln \omega_{it} = \gamma_{1t} \ln \omega_{i0} + \gamma_{2t} \chi_{i}^{GC} + v_{it},
$$

where $\chi_{i}^{GC}$ includes $\chi_{i1}$ and $\text{age}_{i1}$. 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 $t \leq T - 1$, 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 $M(.) = \mathbb{E}(M_i|V_i = .)$ 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 |
|----------------|-----------|
| $(v_{it}, v_{it}^2)_{t \leq T}, v_{i1}v_{i2}, v_{i2}v_{i3}$ | 252 |
| $(v_{it}, v_{it}^2)_{t \leq T}$ | 262 |
| $(v_{it}, v_{it}^2, v_{it}^3)_{t \leq T}$ | 332 |
| $(v_{it}, v_{it}^2, v_{it}^3)_{t \leq T}, v_{i1}v_{i2}, v_{i2}v_{i3}$ | 278 |
| $(v_{it})_{t \leq T}, v_{i1}v_{i2}, v_{i2}v_{i3}$ | 269 |
| $(v_{it})_{t \leq T}$ | 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 $Z_i^X = (\chi_{i1}, \chi_{i0}, \text{age}_{i1}^2, \text{age}_{i0}^2)$. Defining the differences $\Delta \hat{h}_i$ and $\Delta \hat{\chi}_i$ as in Section 2.4.1 and their estimators as $\Delta \hat{\ln} \hat{h}_i$ and $\Delta \hat{\chi}_i$, our estimator of $b$ is

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

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 $k(.) = \mathbb{E}(M_i [\hat{\ln} h_i - \hat{\chi}' b] | V_i = .)$ 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 $\mathbb{E}(Q_i [\hat{\ln} h_i - \hat{\chi}' b - g(V_i)]) = \mathbb{E}(\mu_i)$. This gives $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{m} Q_i [\hat{\ln} h_i - \hat{\chi}' b - \hat{g}(V_i)] = 0.251$. We note that this value is in the range of those reported in Ziliak (1997).

7 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 $\mathbb{E}(\mu_i \mid \det(\hat{X}_i' \hat{X}_i) > \delta_0)$, 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 $\delta_0$. However $\delta_0$ 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 $\mathbb{E}(\mu_i | \text{det} \hat{X}_i^T \hat{X}_i > \delta_n)$ with $\delta_n \to 0$. This could give a sense of an optimal choice for $\delta_n$ 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

Abadir, K. M., and J. R. Magnus (2005): Matrix algebra, vol. 1. Cambridge University Press.

Ahn, H., and J. L. Powell (1993): “Semiparametric estimation of censored selection models with a nonparametric selection mechanism,” Journal of Econometrics, 58(1-2), 3–29.

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

Altonji, J. G., and R. L. Matzkin (2005): “Cross section and panel data estimators for non-separable 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, pp. 307–345.

Arellano, M., and S. Bonhomme (2012): “Identifying Distributional Characteristics in Random Coefficients Panel Data Models,” The Review of Economic Studies, 79(3), 987–1020.

Arellano, M., and S. Bonhomme (2016): “Nonlinear panel data estimation via quantile regressions,” The Econometrics Journal, 19(3), C61–C94.

Bester, C. A., and C. Hansen (2009): “Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model,” Journal of Business & Economic Statistics, 27(2), 235–250.

Blundell, R. W., and J. L. Powell (2003): Endogeneity in Nonparametric and Semiparametric Regression Models vol. 2 of Econometric Society Monographs, p. 312357. Cambridge University Press.

——— (2004): “Endogeneity in semiparametric binary response models,” The Review of Economic Studies, 71(3), 655–679.

Chamberlain, G. (1992): “Efficiency Bounds for Semiparametric Regression,” Econometrica, 60(3), 567–596.

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

Chen, X., H. Hong, and A. Tarozzi (2008): “Semiparametric efficiency in GMM models with auxiliary data,” The Annals of Statistics, 36(2), 808–843.
Chen, X., O. Linton, and I. Van Keilegom (2003): “Estimation of semiparametric models when the criterion function is not smooth,” *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): “Average and quantile effects in nonseparable panel models,” *Econometrica*, 81(2), 535–580.

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

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016): “Identification and estimation of semiparametric two-step models,” *Quantitative Economics*, 7(2), 561–589.

Escanciano, J. C., D. T. Jacho-Chávez, and A. Lewbel (2014): “Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing,” *Journal of Econometrics*, 178, 426–443.

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

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

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

Hahn, J., Z. Liao, and G. Ridder (2018): “Nonparametric two-step sieve M estimation and inference,” *Econometric Theory*, pp. 1–44.

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

Heckman, J., and E. Vytlacil (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*, pp. 974–987.

Hoderlein, S., H. Holzmann, and A. Meister (2017): “The triangular model with random coefficients,” *Journal of econometrics*, 201(1), 144–169.

Hoderlein, S., and H. White (2012): “Nonparametric identification in nonseparable panel data models with generalized fixed effects,” *Journal of Econometrics*, 168(2), 300–314.

Horn, R. A., and C. R. Johnson (2012): *Matrix analysis*. Cambridge university press.
HSIAO, C. (2014): *Analysis of Panel Data*, Econometric Society Monographs. Cambridge University Press, 3 edn.

ICHIMURA, H., AND S. LEE (2010): “Characterization of the asymptotic distribution of semiparametric M-estimators,” *Journal of Econometrics*, 159(2), 252–266.

IMBENS, G. W., AND W. K. NEWEY (2009): “Identification and estimation of triangular simultaneous equations models without additivity,” *Econometrica*, 77(5), 1481–1512.

MAMMEN, E., C. ROTH, AND M. SCHIENLE (2016): “Semiparametric estimation with generated covariates,” *Econometric Theory*, 32(5), 1140–1177.

MASTEN, M. A., AND A. TORGOVITSKY (2016): “Identification of instrumental variable correlated random coefficients models,” *Review of Economics and Statistics*, 98(5), 1001–1005.

MURTAZASHVILI, I., AND J. M. WOOLDRIDGE (2008): “Fixed effects instrumental variables estimation in correlated random coefficient panel data models,” *Journal of Econometrics*, 142(1), 539–552.

——— (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): “The asymptotic variance of semiparametric estimators,” *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

——— (1994b): “Kernel estimation of partial means and a general variance estimator,” *Econometric Theory*, 10(2), 1–21.

——— (1997): “Convergence rates and asymptotic normality for series estimators,” *Journal of econometrics*, 79(1), 147–168.

——— (2009): “Two-step series estimation of sample selection models,” *The Econometrics Journal*, 12, S217–S229.

NEWEY, W. K., J. L. POWELL, AND F. VELLA (1999): “Nonparametric estimation of triangular simultaneous equations models,” *Econometrica*, 67(3), 565–603.

PESARAN, M. H., AND R. SMITH (1995): “Estimating long-run relationships from dynamic heterogeneous panels,” *Journal of econometrics*, 68(1), 79–113.

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

——— (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.

——— (2005b): *Unobserved Heterogeneity and Estimation of Average Partial Effects*. p. 2755. Cambridge University Press.

——— (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.
A Simulation Results

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 $g(V) = -\sin(3v_1)$).

Figure 2: Estimation of $E(\mu_1)$
Distribution of $\hat{\mu}^{1r}$, with true value $E(\mu_1) = 1.5$,
when the sample sizes are $n = 1000$ (left) and $n = 2000$ (right).
Figure 3: Estimation of $\mathbb{E}(\mu_1^2)$

Distribution of $\hat{\mu}^2$, with true value $\mathbb{E}(\mu_1^2) = 2$, when the sample sizes are $n = 1000$ (left) and $n = 2000$ (right).

Figure 4: Comparison of estimators

For $\mu^1$ (left) with true value $\mathbb{E}(\mu_1^1) = 1.5$, and $\mu^2$ (right) with true value $\mathbb{E}(\mu_1^2) = 2$,

Sample size $n = 2000$.  

55
B Proofs of Results in Section 2 and 3

Proof of Result 2.2: Consider a draw of \( V \in \mathcal{S}_V \) satisfying Assumption 2.4(1) and (2). \( \text{Int} \left( \mathcal{S}_{X|V} \right) \neq \emptyset \) and there exists a basis \( e = (e_1, ..., e_{T-1}) \) and for each \( t \leq T - 1 \), \( X^{(t)} \in \text{Int} \left( \mathcal{S}_{X|V} \right) \) such that \( p_{X|V}(\hat{X}^{(t)}|V) > 0 \), \( \text{Rank}(\hat{X}^{(t)}) = d_x \) and \( \hat{X}^{(t)'}e_t = 0 \).

\( \mathcal{S}_{X|V} \) 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 \( t \leq T - 1 \). \( \hat{X}^{(t)} \) is of full column rank \( k_x \), which implies that \( \det(\hat{X}^{(t)' \hat{X}^{(t)})} \neq 0 \). The determinant function being continuous, as well as the density \( p_{X|V}(.|V) \), this implies that there exists an open ball \( B_t \subset \text{Int} \left( \mathcal{S}_{X|V} \right) \) such that (1) \( \hat{X}^{(t)} \in B_t \), (2) \( \forall \hat{X} \in B_t \), \( p_{X|V}(\hat{X}|V) > 0 \), and (3) \( \forall \hat{X} \in B_t \), \( \text{Rank}(\hat{X}) = k_x \).

Take \( c \in \mathbb{R}^{T-1} \) such that \( \mathcal{M}(V)c = 0 \). Then we know by the argument given in Section 2.3.1 that \( M(X)c = 0 \), \( \mathbb{P}_{X|V} \)-a.s. Since the density \( p_{X|V}(.|V) \) is strictly positive on \( B_t \), \( M(X)c = 0 \) for all \( \hat{X} \) in \( B_t \) except on a set of measure 0. Additionally, since for all \( \hat{X} \in B_t \), \( \hat{X} \) is of full rank, then \( M(X)c \) is a continuous function of \( \hat{X} \). Those two facts imply that \( M(X)c \) is uniformly 0 on \( B_t \), and in particular, \( M(X^{(t)})c = 0 \). Moreover \( \hat{X}^{(t)'}e_t = 0 \) implies \( M(X^{(t)})e_t = M(X^{(t)})'e_t = e_t \). Thus,

\[
\forall t \leq T - 1, \quad M(X^{(t)})c = 0 \Rightarrow \forall t \leq T - 1, \quad e_t'M(X^{(t)})c = 0,
\]
\[
\Rightarrow \forall t \leq T - 1, \quad e_t'c = c_t = 0,
\]
\[
\Rightarrow c = 0.
\]

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

Proof of Result 2.3: We write \( \dot{z}_{it} = \dot{b}(z_{i,t+1}, z_{it}) + \dot{a}_{it} \) and in vector form \( \dot{X}_t = \dot{B}(Z_t) + \dot{V}_t \), where \( \dot{X}_t \) is a column vector of size \( T - 1 \). We define for a draw of \( V \) and for the corresponding \( Z^V \) defined in Condition (2) of Assumption 2.5, the variable \( \dot{X}^V = \dot{B}(Z^V) + \dot{V} \). Wlog we can assume that \( \dot{X}^V \neq 0 \) since \( db_t(z^V_{it})/dz_t \neq 0 \). We also define an open ball \( B \) around \( \dot{X}^V \) such that for all \( \dot{X} \in B, \dot{X} \neq 0 \) 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 columns of \( \dot{X} \), is continuous on \( B \) by the same argument as in the proof of Result 2.2. Take \( c \) such that \( \mathcal{M}(V)c = 0 \). As in this proof, we then have \( \| M(X)c \| = 0 \) for all \( X \) such that \( \dot{X} \in B \).

For this same draw of \( V \) and the corresponding \( Z^V \), there exists \( t \leq T - 1 \) such that \( db_t(z^V_{it})/dz_t \neq 0 \). For a given \( \delta > 0 \), define \( Z^V_{\delta} = (z^V_{i1}, ..., z^V_{i,t-1}, z^V_{it} + \delta, z^V_{it+1}, ..., z^V_{iT})' \) and \( \dot{X}^V_{\delta} = \dot{B}(Z^V_{\delta}) + V \). For \( \delta \) small enough, \( \dot{X}^V_{\delta} \in B \). Note that \( \dot{X}^V_{\delta} \) are column vectors. All components of \( \dot{X}^V_{\delta} \) are the same as those of \( \dot{X}^V \) but two. If \( T \geq 4 \), one can deduce directly that there exists \( \delta \) such that \( \dot{X}^V_{\delta} \) and \( \dot{X}^V \)
are not collinear and \( \hat{X}^V_Y \neq 0 \). If \( T = 3 \) and \( \delta \) small enough, one can show that \( \hat{X}^V_Y \) and \( \hat{X}^V \) can be collinear only if \( \hat{x}^V_{i1} = \hat{x}^V_{i2} \). Since \( B \) is an open ball, one can change \( \hat{X}^V \in B \) to ensure \( \hat{x}^V_{i1} \neq \hat{x}^V_{i2} \).

We now have \( \hat{X}^V_Y \) and \( \hat{X}^V \) two noncollinear vectors of \( B \). However \( M(\hat{X}^V)c = 0 \) implies that \( c \) and \( \hat{X}^V \) are collinear, and \( M(\hat{X}^V_Y)c = 0 \) also implies that \( c \) and \( \hat{X}^V_Y \) are collinear, which would imply, if \( c \neq 0 \), that \( \hat{X} \) and \( \hat{X} \) are collinear. Therefore \( c \) must be 0. This implies that \( M(V) \) is nonsingular.

\[\square\]

**Proof of Result 2.4** The proof of invertibility of \( M(V) \) under Assumption 2.6 follows the same steps as the proof of Result 2.2, using additionally continuity of the functions \((l_k)_{s+1 \leq k \leq d_x} \).

\[\square\]

**Proof of Result 2.5** The proof of this result follows as in the proof of Result 2.2 without the continuity arguments.

\[\square\]

**Proof of Result 3.2** Under Assumption 3.2 for a given value \( \hat{V} \), if \( M(\hat{V}) \) is not invertible, there exist two nonzero draws of \( x_1, x_2 \) and \( p_1 \), with positive density in the continuously distributed case, or probability in the discretely distributed case, such that \( x_1C_1 + x_2C_2(\hat{V}) \) and \( p_1C_1 + p_2C_2(\hat{V}) \) are proportional. Thus \( M(\hat{V}) \) not invertible implies \( \hat{V} \in \mathcal{D} \). The function \( g \) is then identified over \( \mathcal{S}_V \setminus \mathcal{D} \). 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 \( E(\mu_i) \) and \( E(\alpha_i) \).

\[\square\]

**C Proofs of Results in Section 4**

**C.1 Proof of Consistency of \( \hat{\mu} \)**

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 \( \mathbb{E} \left( |a_n| \right| b_n \right) = O_p(r_n) \) then \( |a_n| = O_p(r_n) \). For a sequence \((c_n)_{n \in \mathbb{N}} \in \mathbb{R}^N \), the notation \( c_n \to 0 \) should be understood as \( c_n \to_{n \to \infty} 0 \).

**Proof of Result 4.1** Under Assumption 4.1, using Theorem 1 of Newey (1997) and Lemma A1 of Newey, Powell, and Vella (1999) (see e.g Equations A.3 and A.5), we have \( \forall t \leq T, \frac{1}{n} \sum_{i=1}^n \| v_{it} - \tilde{v}_{it} \|^2 = O_p(\Delta_n^2) \), as well as \( \max_i \| v_{it} - \tilde{v}_{it} \| = O_p(a_1(L)\Delta_n) \), where \( \Delta_n = \sqrt{L/n + L^{-\gamma}} \), and

\[
\sup_{\mathcal{S}_V} \| b_t - \tilde{b}_t \| = O_p(a_1(L)\Delta_n) \tag{39}
\]

Define \( \tilde{V}_i = (\tilde{v}_{i1}, ..., \tilde{v}_{iT}) \). Since \( \| \tilde{V}_i - V_i \| \leq \| \tilde{V}_i - V_i \| \), the result applies.

\[\square\]

57
Proof of Result 4.2 Define $a_1(L) = L^{\alpha_1+1}$. Under these conditions, Lemma S.3 of Imbens and Newey (2009) can be modified using Andrews (1991) (Equations 3.14 or A.40) to account for the fact that $\xi_t$ is not scalar. One obtains that since $r^L$ is the power series basis of functions, there exists a nonsingular $L \times L$ matrix $\hat{\Gamma}_{\xi_t}$ such that for $\hat{r}^L(\xi_t) = \hat{\Gamma}_{\xi_t} r^L(\xi_t)$ then $E(\hat{r}^L(\xi_t) \hat{r}^L(\xi_t)') = I_L$, implying that Assumption 4.1 (2) holds. One also obtain $\sup_{S_t} ||\hat{r}^L(\xi_t)|| \leq a_1(L)$ with $a_1(L) = CL^{\alpha_1+1}$ as is required in Assumption 4.1 (4). Thus, Assumption 4.1 is satisfied and Result 4.1 applies.

Proof of Result 4.3 Instead of applying the general results of Section 5.2 of the Online Appendix of Hahn, Liao, and Ridder (2018), we directly extend the proof of Theorem 12 of Imbens and Newey (2009) (IN09 thereafter) because it uses lower level conditions similar to the ones we seek to impose. We adapt some of their claims to our model where $E(e_i^W | X_i^1, X_i^2, Z_i) \neq 0$.

Define $P = (p_1, \ldots, p_n)$, $Q = PP'/n$, $\hat{Q} = \hat{P}\hat{P}'/n$, $p_k^W = \rho^W(X_i^1, X_i^2, Z_i)$, as well as the vectors $e^W = (e_1^W, \ldots, e_n^W)'$, $\hat{\rho}^W = (p_1^W, \ldots, p_n^W)'$ and $e^W* = (e_1^W*, \ldots, e_n^W*)'$. Note that $e^W = \hat{\rho}^W + e^W*$. 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 4.3 $E(p_i^Kp_i'^K) = I_K$. By construction, $\hat{V}_i \in S_V$, and under Assumption 4.3 (29) holds. Therefore, as in Lemma S.5 of Imbens and Newey (2009), we have

$$||Q - I_K|| = O_P(b_1(K)\sqrt{K/n}),$$
(40) $$||P'e^W/n|| = O_P(\sqrt{K/n}),$$
(41) $$||\hat{P} - P||^2/n = O_P(b_2(K)^2\Delta^2_n),$$
(42) $$||\hat{Q} - Q|| = O_P(b_2(K)^2\Delta^2_n + \sqrt{Kb_2(K)\Delta_n}).$$
(43)

Hence by Assumption 4.3 (5), $||\hat{Q} - I_K|| = o_P(1)$ and as in Lemma S.6 of Imbens and Newey (2009), with probability going to 1, $\lambda_{\min}(\hat{Q}) \geq C$ and $\lambda_{\min}(Q) \geq C$.

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 $||\hat{\pi}^W - \pi^K_W||$, where we recall $\hat{\pi}^W = \hat{Q}^{-1}\hat{P}W$. We define $H^W = (h^W(1), \ldots, h^W(V_n))'$, $\hat{H}^W = (h^W(\hat{V}_1), \ldots, h^W(\hat{V}_n))'$, $\hat{\pi}^W = \hat{Q}^{-1}\hat{P}\hat{H}^W/n$, $\pi^W = \hat{Q}^{-1}\hat{P}H^W/n$. We decompose

$$||\hat{\pi}^W - \pi^K_W|| \leq \underbrace{||\hat{\pi}^W - \pi^W||}_{(A)} + \underbrace{||\pi^W - \hat{\pi}^W||}_{(B)} + \underbrace{||\hat{\pi}^W - \pi^K_W||}_{(C)}.$$

The first term can in turn be decomposed as

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

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

$$E(||\hat{Q}^{-1/2}\hat{P}e^W*/n||^2|X_1, Z_1, \ldots, X_n, Z_n) = \text{tr}(\hat{Q}^{-1/2}\hat{P}E(e^W*e^W*/|X_1, Z_1, \ldots, X_n, Z_n)\hat{P}'\hat{Q}^{-1/2}/n^2),$$

58
\[ \leq C \text{tr}(\hat{P}'(\hat{P}P')^{-1}\hat{P})/n \leq CK/n. \]

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 \( \lambda_{\text{min}}(\hat{Q}) \geq C \) 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 since \( e^{W^*} \) is by definition conditionally mean-independent of the regressors generating \( V \). As for the second term appearing in (A), we write

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

Since \( \mathbb{E}(\rho^W_i|V_i) = 0 \) and \( (\rho^W_i, V_i) \) is i.i.d, we know that as in [41], \( \|P \tilde{\rho}^W/n\|^2 = O_{\mathbb{P}}(K/n) \). Therefore, by \( \lambda_{\text{min}}(Q) \geq C \) w. p. a 1, \( \|\hat{Q}^{-1}P \tilde{\rho}^W/n\| = O_{\mathbb{P}}(\sqrt{K/n}) \).

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

\[ \frac{1}{n} \left\| \sum_{i=1}^{n} (p^K_i - \hat{p}^{K}_i) \rho^W_i \right\| \leq \frac{1}{n} \sum_{i=1}^{n} \left\| (p^K_i - \hat{p}^{K}_i) \rho^W_i \right\| \leq C \left( \frac{1}{n} \sum_{i=1}^{n} \left\| (p^K_i - \hat{p}^{K}_i) \right\|^2 \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} |\rho^W_i|^2 \right)^{1/2}, \]

\[ \leq Cb_2(K) \left( \frac{1}{n} \sum_{i=1}^{n} \left\| (\hat{V}_i - V_i) \right\|^2 \right)^{1/2} \|\tilde{\rho}\|_\infty = O_{\mathbb{P}}(b_2(K)\Delta_n). \]

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

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 [4,3] [4]. This implies that

\[ \|\tilde{\pi}^W - \pi^W\|^2 = O_{\mathbb{P}}(K/n + b_2(K)^2\Delta_n^2 + \Delta_n^2 + K^{-2\gamma_2}) = O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + b_2(K)^2\Delta_n^2), \]

which differs from the rate \( (K/n + K^{-2\gamma_2} + \Delta_n^2) \) obtained in IN09. 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 required linearizing the term \( \hat{P} - P \).

We can now conclude that

\[ \int \left| \hat{h}^W(V) - h^W(V) \right|^2 dF_V(V) \leq \int \left| \hat{h}^W(V) - p^K(V)\pi^W \right|^2 dF_V(V) + \int \left| p^K(V)\pi^W - h^W(V) \right|^2 dF_V(V) \]

\[ \leq (\hat{\pi}^W - \pi^W)' \left[ \int p^K(V)p^K(V)'dF_V(V) \right] (\hat{\pi}^W - \pi^W) + CK^{-2\gamma_2} \]

\[ = O_{\mathbb{P}}(K/n + K^{-2\gamma_2} + \Delta_n^2b_2(K)^2), \]

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

\[ \sup_{V \in \mathcal{S}_V} \left| \hat{h}^W(V) - h^W(V) \right| \leq \sup_{V \in \mathcal{S}_V} \left| \hat{h}^W(V) - p^K(V)\pi^W \right| + O_{\mathbb{P}}(K^{-\gamma_2}) \]

59
Assumption 4.3 (3). Then Assumption 4.4 (5) implies Assumption 4.3 (5), and together with Assumption 4.4 (3), implies

\[ D \]

is continuous by Lemma C.1, it has a minimum and reaches it. This minimum value cannot be 0,

\[ p \]

where we index the eigenvalues

\[ \lambda \]

Proof of Result 4.5. Under Assumptions 4.2 and 4.4, writing \( \Gamma \) for all

\[ V \]

Proof of Lemma C.2. Under Assumption 4.6,

\[ \text{Lemma C.2.} \]

Under Assumption 4.6, there exists \( c > 0 \), such that for all \( V \in S_V \), \( \lambda_{\min}(M(V)) \geq c \).

Proof of Lemma C.1. Under Assumption 4.6, the function \( V \in S_V \mapsto \lambda_{\min}(M(V)) \) is continuous.

Proof of Result 4.4. As in the proof of Result 4.2, define \( b_1(K) = K^{\alpha_2 + 1} \) and \( b_2(K) = K^{\alpha_2 + 3} \). Then Assumption 4.4 (5) implies Assumption 4.3 (5), and together with Assumption 4.4 (3), implies Assumption 4.3 (3).

To prove Result 4.5, we will need the two following Lemmas.

\[ \text{Lemma C.1.} \]

Under Assumption 4.6, \( \text{Lemma C.1.} \) for two values \( V \) and \( V' \),

\[ \sum_{i=1}^{T-1} \left| \lambda_i(M(V)) - \lambda_i(M(V')) \right|^2 \leq \|M(V) - M(V')\|_F^2, \]

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

\[ |\lambda_{\min}(M(V)) - \lambda_{\min}(M(V'))| \leq \|M(V) - M(V')\|_F^{1/2}. \]  

(44)

Since \( M(.) \) is a continuous function, this concludes the argument. Note that the Lipschitz inequality \( (44) \) will be used in the proof for the convergence rates.

\[ \text{Lemma C.2.} \]

Under Assumption 4.6, there exists \( c > 0 \), such that for all \( V \in S_V \), \( \lambda_{\min}(M(V)) \geq c \).

Proof of Lemma C.2. Under Assumption 4.6, \( M(V) \) is nonsingular for all \( V \in S_V \). This implies that for all \( V \in S_V \), \( \lambda_{\min}(M(V)) > 0 \). Since \( S_V \) is a compact set and the function \( V \mapsto \lambda_{\min}(M(V)) \) is continuous by Lemma C.1, it has a minimum and reaches it. This minimum value cannot be 0, hence \( \exists c > 0, \forall V \in S_V, \lambda_{\min}(M(V)) \geq c \).

Proof of Result 4.5. Under Assumptions 4.2 and 4.4 writing \( \Gamma^2_n \) and \( \gamma_n \) respectively the mean square and sup norm rates of convergence, we have

\[ \int \left\| \bar{k}(V) - k(V) \right\|^2 dF(V) = O_p(\Gamma^2_n), \sup_{V \in S_V} \left\| \bar{k}(V) - k(V) \right\| = O_p(\gamma_n), \]

\[ \int \left\| \hat{M}(V) - M(V) \right\|_F^2 dF(V) = O_p(\Gamma^2_n), \sup_{V \in S_V} \left\| \hat{M}(V) - M(V) \right\|_F = O_p(\gamma_n), \]

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 \( M(V) \) and \( k(V) \).

We write \( 1_n = 1 \left( \min_{V \in S_V} \lambda_{\min}(\hat{M}(V)) > \frac{\epsilon}{2} \right) \). Using \( (44) \), we have

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

60
Using $T$, the norm inequality and definition of induced norm, this gives
Assumptions 4.2 and 4.4,
which implies
To prove consistency, we need to show that
Proof of Result 4.6.
We have
to obtain the sup norm rate, we write

$$
\gamma_n \to 0 \text{ which implies } 1_n = 1 \text{ w. p. a. 1.}
$$

To obtain the sup norm rate, we write

$$(1)$$

$$
\|\hat{g}(V) - g(V)\|_\infty = O_P(\gamma_n).
$$

To obtain the mean square error rate, we write

$$
\frac{1}{n} \int \|\hat{g}(V) - g(V)\|^2 dF(V)
= \frac{1}{n} \int \left\| \mathcal{M}^{-1} \left[ \hat{\mathcal{M}}(V) - \mathcal{M}(V) \right] \hat{\mathcal{M}}(V)^{-1} k(V) + \hat{\mathcal{M}}(V)^{-1} \left[ k(V) - \hat{k}(V) \right] \right\|^2 dF(V)
\leq \frac{1}{n} \frac{2}{c} \|\| k\|_\infty \int \left\| \hat{\mathcal{M}}(V) - \mathcal{M}(V) \right\|^2 dF(V) + \frac{1}{n} \frac{2}{c} \int \left\| \hat{k}(V) - k(V) \right\|^2 dF(V),
$$

which implies $\|\hat{g}(V) - g(V)\|^2 dF(V) = O_P(\gamma_n^2)$.

Proof of Result 4.6. To prove consistency, we need to show that $\frac{1}{n} \sum_{i=1}^n Q_i^\delta \hat{g}(\hat{V}_i) \to_P E(Qg(V)\delta)$. Indeed then by the LLN, we have $\frac{1}{n} \sum_{i=1}^n \delta_i \to_P P(\det(\hat{X}_i'\hat{X}_i) > \delta)$, and by Assumption 4.7 and the LLN, $\frac{1}{n} \sum_{i=1}^n Q_i^\delta \hat{g}_i \to_P E(Q\hat{g}\delta)$ also holds. Then consistency would follow from Equation (12).

To obtain $\frac{1}{n} \sum_{i=1}^n Q_i^\delta \hat{g}(\hat{V}_i) \to_P E(Qg(V)\delta)$, we decompose

$$
\frac{1}{n} \sum_{i=1}^n Q_i^\delta \hat{g}(\hat{V}_i) - E(Qg(V)\delta) = \frac{1}{n} \sum_{i=1}^n Q_i^\delta [\hat{g}(\hat{V}_i) - g(V_i)] + \frac{1}{n} \sum_{i=1}^n Q_i^\delta [g(V_i) - E(Qg(V)\delta)]
:= A_n + B_n.
$$

We have

$$
\|A_n\| = \|\frac{1}{n} \sum_{i=1}^n Q_i^\delta [\hat{g}(\hat{V}_i) - g(V_i)] + \frac{1}{n} \sum_{i=1}^n Q_i^\delta [g(V_i) - g(V_i)]\|
\leq \|\hat{g} - g\|_\infty \frac{1}{n} \sum_{i=1}^n \|Q_i^\delta\| + C \max_i \|\hat{V}_i - V_i\| \frac{1}{n} \sum_{i=1}^n \|Q_i^\delta\|_2
= O_P(\gamma_n + a_1(L)\Delta_n) \frac{1}{n} \sum_{i=1}^n \|Q_i^\delta\|_2,
$$

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 (4.5). We assumed $\gamma_n \to 0$ and $a_1(L)\Delta_n \to 0$ as $n$ goes to infinity, thus we obtain $\|A_n\| = O_P(1)$.

\qed
C.2 Proof of Asymptotic Normality of $\hat{\mu}$

We first introduce some more notations. We define $\bar{b}_t = (b_{1t}, ..., b_{nt}) = (b_t(\xi_{1t}), ..., b_t(\xi_{nt}))$, $\bar{\nu}_t = (\nu_{1t}, ..., \nu_{nt})$, $\bar{V} = (V_1, ..., V_n)$, $\bar{x} = (X_{1t}, ..., X_{nt})$ and similarly $\bar{z}$. For the results in Section 4.4.3, we also define the vector $\bar{h}_W^\gamma = (h_W^\gamma(V_1), ..., h_W^\gamma(V_n)) = (h_W(V_1), ..., h_W(V_n))$ since $V_i \in S_V \forall i \leq n$, and the vector $\bar{h}_W^\gamma = (h_W^\gamma(\bar{V}_1), ..., h_W^\gamma(\bar{V}_n))$.

Proof of Result 4.7. This proof is a special case of Theorem 2 in Chen, Linton, and Van Keilegom (2003). By Assumption 4.8 (3), there exists $\delta_n = o(1)$ such that $P(||\hat{G} - G_0|| > \delta_n) \rightarrow 0$. Take $1_n = 1(||\hat{G} - G_0|| \leq \delta_n)$. $1_n||X_n(\hat{G}) - X_n(G_0) - X'(G_0)[\hat{G} - G_0]|| \leq 1_n||X_n(\hat{G}) - X'(\hat{G}) - X_n(G_0)|| + 1_n||X'(G_0)[\hat{G} - G_0]|| = o_P(n^{-1/2})$ by Assumption 4.8. Hence $||X_n(\hat{G}) - X_n(G_0) - X'(G_0)[\hat{G} - G_0]|| = o_P(n^{-1/2})$.

Proof of Lemma 4.7. As argued in the proof of Result 4.3, under Assumption 4.9 (5), we can impose without loss of generality the normalization $E(p_ir'_i) = K$ and $E(r_ir'_i) = I_L$. We define $\Delta_p = b_2(K)\sqrt{L/n}$, $\Delta_Q = b_2(K)^2\Delta_n^2 + \sqrt{K}b_2(K).\Delta_n + b_2(K)\sqrt{L/n}$, $\Delta_Q = a_1(L)\sqrt{L/n}$ and $\Delta_H = b_2(K)\Delta_n^2 + a_1(L)\sqrt{K\bar{L}/n}$. Recall that $\Delta_n^2 = L/n + L^{-2\gamma_1}$ and note that $b_2(K) \leq b_2(K) \leq b_2(K)$ under Assumption 4.9 (7). $\sqrt{nK^{-\gamma_2}} = o(1)$ and

$$\sqrt{K}\Delta_Q = O(\sqrt{K}[\sqrt{K}b_2(K).\Delta_n + b_2(K)\sqrt{L/n}]) = O(Kb_2(K)[\sqrt{L/n} + L^{-\gamma_1}]) = o(1),$$

$$\Delta_p = b_2(K)\sqrt{L/n} = o(1), \quad b_2(K)\Delta_Q = b_2(K)a_1(L)L/K\sqrt{n} = o(1),$$

$$b_2(K)\Delta_n = o(1), \quad \sqrt{L}\Delta_H = O(b_2(K)L(\sqrt{L/n} + L^{-\gamma_1}) + a_1(L)\sqrt{K\bar{L}/n}) = o(1),$$

$$\Delta_Q b_2(K) = O\left(b_2(K)^2\sqrt{K}[L^{-\gamma_1} + \sqrt{L/n}] + b_2(K)^2\sqrt{K\bar{L}/n}\right) = o(1),$$

$$\sqrt{nb_3(K)\Delta_n^2} = b_3(K)[L/n + \sqrt{n}L^{-2\gamma_1}] = o(1).$$

These results imply in particular that $\Delta_p = o(1), \Delta_Q = o(1), \Delta_Q = o(1)$ and $\Delta_H = o(1)$. All these rates will be used in the steps of the proof.

We define for all $t \leq T$, $Q_t = R_t R'_t/n$. As in the proof of Result 4.3, we obtain $||Q_t - I_L|| = O_P(\Delta_Q) = o_P(1)$, $t \leq T$, and $||Q - I_K || = O_P(\Delta_Q) = o_P(1)$. This implies, as argued in NPV99, that the eigenvalues of $\hat{Q}$ are bounded away from 0 w.p.a 1, therefore $||B\hat{Q}^{-1}|| \leq ||B||O_P(1)$ and $||B\hat{Q}^{-1/2}|| \leq ||B||O_P(1)$ for any matrix $B$. Using

$$||\Pi_t^W - H_t^W|| \leq \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - p_i)(\hat{h}_W^\gamma(V_i)/\hat{v}_t \otimes r'_it)$$

$$+ \frac{1}{n} \sum_{i=1}^n p_i (\hat{h}_W^\gamma(V_i)/\hat{v}_t \otimes r'_it) - E[p_i (\hat{h}_W^\gamma(V_i)/\hat{v}_t \otimes r'_it)]$$

with

$$\frac{1}{n} \sum_{i=1}^n (\hat{p}_i - p_i)(\hat{h}_W^\gamma(V_i)/\hat{v}_t \otimes r'_it) \leq b_2(K)\Delta_n \sup_{S_V} ||h_W^\gamma|| \text{tr}(R_t R'_t)^{1/2}/\sqrt{n} = O_P(b_2(K)\Delta_n \sqrt{L}).$$
and
\[
\mathbb{E}(\| \frac{1}{n} \sum_{i=1}^{n} p_i (\partial h^w(V_i)/\partial v_t \otimes r_{it}') - \mathbb{E}[p_i (\partial h^w(V_i)/\partial v_t \otimes r_{it}')] \|^2 ) \\
\leq \mathbb{E} \left[ \text{tr} \left( \frac{1}{n^2} \sum_{i=1}^{n} p_i (\partial h^w(V_i)/\partial v_t \otimes r_{it}') (\partial h^w(V_i)/\partial v_t \otimes r_{it}')' p_i \right) \right] \\
\leq C \frac{1}{n} \mathbb{E}(\text{tr}(p_i r_{it}' r_{it}')) \leq Ca_1(L)^2 K/n,
\]
we obtain \( \| \mathbb{P}^W_t - H^W_t \| = O_P(\Delta_H) \). Thus by Assumption 4.9(7), \( \| \mathbb{P}^W_t - H^W_t \| = o_P(1) \). Moreover, since \( \rho^W(.) \) is a bounded function and \( \mathbb{E}(r_{it}' r_{it}) = \mathbb{E}(\text{tr}(I_L)) = L \), we have for all \( t \leq T \),
\[
\mathbb{E}(\| d\mathbb{P}^W_t - dP^W_t \|^2 ) \leq C \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \left[ \text{tr} \left( \frac{1}{n} \sum_{i=1}^{n} p_i (\partial h^w(V_i)/\partial v_t \otimes r_{it}')' r_{it}' \right) \right] \leq Cb_2(K)^2 L/n,
\]
which implies by M that \( \| d\mathbb{P}^W_t - dP^W_t \| = O_P(\Delta_{\mathbb{P}}) = o_P(1) \).

Since \( a(.) \) is linear, \( a(\hat{h}^W) = A\tilde{\pi}_W \). Using Assumption 4.9(2),
\[
||a(p^K)(.) \tilde{\pi}_W^K - a(h^W(.))|| \leq C \sup_{\tilde{S}_v} ||p^K(.) \tilde{\pi}_W^K - h^W(.)|| \leq C \sup_{\tilde{S}_v} ||p^K(.) \tilde{\pi}_W^K - h^W(.)|| \leq CK^{-\gamma},
\]
which implies using Assumptions 4.9(7) that
\[
\sqrt{n}[a(\hat{h}^W) - a(h^W)] = \sqrt{n}A[\tilde{\pi}_W - \pi^K] + o_P(1),
\]
\[
= A\hat{Q}^{-1}P(W - \tilde{P}'\tilde{\pi}_W) / \sqrt{n} + o_P(1).
\]

Since
\[
||A\hat{Q}^{-1}P(h^W - \tilde{P}'\tilde{\pi}_W)|| \leq ||A\hat{Q}^{-1}P|| ||h^W - \tilde{P}'\tilde{\pi}_W|| \leq \sqrt{n}||A\hat{Q}^{-1/2}|| \sqrt{n} \sup_{\tilde{S}_v} ||p^K(.) \tilde{\pi}_W^K - h^W(.)||,
\]
then \( A\hat{Q}^{-1}P(h^W - \tilde{P}'\tilde{\pi}_W) / \sqrt{n} = o_P(1) \) by \( ||A|| \) bounded. Therefore we obtain as in NPV99,
\[
\sqrt{n}(a(\hat{h}^W) - a(h^W)) = \underbrace{A\hat{Q}^{-1}P \tilde{e}^W / \sqrt{n}}_{(B)} + \underbrace{A\hat{Q}^{-1}(\tilde{P} - P)e^W/\sqrt{n}}_{(B1)} + \underbrace{A\hat{Q}^{-1}(\hat{P} - P)\tilde{e}^W / \sqrt{n}}_{(B2)} + \underbrace{A\hat{Q}^{-1}(\hat{P} - P)\hat{e}^W / \sqrt{n}}_{(B3)} + o_P(1). \tag{45}
\]

We first focus on the term (B), which we decompose
\[
(B) = \underbrace{AP' e^W / \sqrt{n}}_{(B1)} + \underbrace{A(\hat{Q}^{-1} - I)P e^W / \sqrt{n}}_{(B2)} + \underbrace{A\hat{Q}^{-1}(\tilde{P} - P)e^W* / \sqrt{n}}_{(B2)} + \underbrace{A\hat{Q}^{-1}(\hat{P} - P)\tilde{e}^W / \sqrt{n}}_{(B3)}.
\]

Since
\[
\mathbb{E}(||P e^W / \sqrt{n}||^2) = \text{tr}[\mathbb{E}(P e^W(e^W)' P')]/n = \text{tr}[\mathbb{E}(e^W(e^W)'P' P'])/n \leq C \text{tr}[I_k] = O(K),
\]
by Assumption 4.9(8), by M we have \( ||P e^W / \sqrt{n}|| = O_P(K^{1/2}) \). Therefore,
\[
||(B1)|| \leq ||A\hat{Q}^{-1}|| ||I_K - \hat{Q}|| ||P e^W / \sqrt{n}|| = ||A||O_P(1) ||I_K - \hat{Q}|| ||P e^W / \sqrt{n}|| = O_P(\Delta_{\mathbb{Q}}K^{1/2})
\]

63
\[ (B1) = o_{\mathbb{P}}(1), \]

under Assumption 4.9 (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 \( \mathbb{E}(e^W|\bar{x}, \bar{z}) = 0 \),

\[
\mathbb{E}((\hat{P} - P)e^W/\sqrt{n}|\bar{x}, \bar{z}) = 1/n \text{tr} \mathbb{E}((\hat{P} - P)e^W(e^W)'(\hat{P} - P)'|\bar{x}, \bar{z})
\]

\[
= 1/n \text{tr} ((\hat{P} - P)\mathbb{E}(e^W(e^W)'|\bar{x}, \bar{z})(\hat{P} - P)' \leq C/n\|\hat{P} - P\|^2.
\]

By the proof of Result 4.3, \( \|\hat{P} - P\|^2/n = O_{\mathbb{P}}(b_2(K)^2\Delta^2_n) \) (the difference is that now \( b_2(K) \) is defined as the sup rate over the extended support \( S_\nu \)) hence by CM, \( \|(\hat{P} - P)e^W/\sqrt{n}\| = O_{\mathbb{P}}(b_2(K)\Delta_n) \).

\[
\|A\hat{Q}^{-1}(\hat{P} - P)e^W/\sqrt{n}\| \leq \|A\hat{Q}^{-1}\| \|\hat{P} - P\|e^W/\sqrt{n} = \|A\|O_{\mathbb{P}}(1)(\hat{P} - P)e^W/\sqrt{n}
\]

\[ \Rightarrow (B2) = O_{\mathbb{P}}(b_2(K)\Delta_n) = o_{\mathbb{P}}(1). \]

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

\[
||\hat{p}^K_i - p^K_i|| - \frac{\partial p^K_i(V_i)}{\partial V}(V_i - V_i)|| \leq Cb_3(K)||V_i - V_i||^2,
\]

which can be rewritten as \( ||\hat{p}^K_i - p^K_i|| - \frac{\partial p^K_i(V_i)}{\partial V}(V_i - V_i)|| \leq Cb_3(K)||V_i - V_i||^2 \), since \( V_i \in S_\nu \) and we chose \( \tau \) so that its Jacobian matrix is the identity matrix on \( S_\nu \). Hence

\[
\|A\hat{Q}^{-1}\frac{1}{\sqrt{n}}[(\hat{P} - P)\rho^W - \sum_{i=1}^n \frac{\partial p^K_i(V_i)}{\partial V}(V_i - V_i)\rho^W_i] \| \leq CO_{\mathbb{P}}(1)b_3(K)\sum_{i=1}^n ||V_i - V_i||^2/\sqrt{n}
\]

\[ = O_{\mathbb{P}}(\sqrt{mb_3(K)\Delta^2_n}) = o_{\mathbb{P}}(1), \]

by Assumption 4.9 (7). Therefore \( (B3) = A\hat{Q}^{-1}\sum_{i=1}^n \frac{\partial p^K_i(V_i)}{\partial V}(V_i - V_i)\rho^W_i/\sqrt{n} + o_{\mathbb{P}}(1) \). We can decompose

\[ (-1)(\bar{v}_{it} - v_{it}) = \beta_t^L r_{it} - b_{it} = (Q_{it}^{-1}R_{t}[X^2_t - R_t\beta_t^L]/n)'r_{it} + \beta_t^L r_{it} - b_{it}, \]

\[ = [(Q_{it}^{-1}R_{t}b_{it}'n)'r_{it}] + [(Q_{it}^{-1}R_{t}[t_{i}' - R_t\beta_t^L]/n)'r_{it}] + [\beta_t^L r_{it} - b_{it}], \]

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

\[
(B3.1) = \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \hat{p}_{it}^W \frac{\partial p^K_i(V_i)}{\partial v_t} \left[ Q_{it}^{-1}R_{t}[\hat{b}_{it}' - R_t\beta_t^L]/n \right]'r_{it}/\sqrt{n}
\]

\[
(B3.2) = \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \hat{p}_{it}^W \frac{\partial p^K_i(V_i)}{\partial v_t} \left[ \beta_t^L r_{it} - b_{it} \right]/\sqrt{n}
\]

\[
(B3.3) = \sum_{t=1}^T A\hat{Q}^{-1}\sum_{i=1}^n \hat{p}_{it}^W \frac{\partial p^K_i(V_i)}{\partial v_t} \left[ Q_{it}^{-1}R_{t}b_{it}'n \right]'r_{it}/\sqrt{n}
\]

64
The first term in this expression of (B.3) can be rewritten

\[
(B3.1) = \sum_{i=1}^{n} A\hat{Q}^{-1} \sum_{i=1}^{n} \rho_{i}^{W} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \left[ Q_{1i}^{-1} R_{t} \left[ \hat{b}_{t} - R_{t}\beta_{t}^{L} \right] / n \right] / \sqrt{n} \\
= T \sum_{i=1}^{n} A\hat{Q}^{-1} \sum_{i=1}^{n} \rho_{i}^{W} \left( \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \otimes t_{it}' \right) \text{Vec}(Q_{1i}^{-1} R_{t} \left[ \hat{b}_{t} - R_{t}\beta_{t}^{L} \right]) / \sqrt{n} \\
= T \sum_{i=1}^{n} A\hat{Q}^{-1} \text{d}\hat{P}_{t}^{W} \left( I_{d_{2}} \otimes Q_{1i}^{-1} R_{t} \right) \text{Vec}(\hat{b}_{t} - R_{t}\beta_{t}^{L}) / \sqrt{n} ,
\]

(see e.g. p282 Abadir and Magnus (2005)) where \(||\text{Vec}(\hat{b}_{t} - R_{t}\beta_{t}^{L})|| \leq \sqrt{n} \sup_{t} ||b_{t}(.) - \beta_{t}^{L}r_{t}(.)|| \leq \sqrt{nL^{-\gamma}}\). Defining, for \(d \leq d_{2}\), the matrix \(\hat{d}_{P_{t}^{W}} = \frac{1}{n} \sum_{i=1}^{n} \rho_{i}^{W} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} r_{it}'\) where \(v_{it}\) is the \(d\)th component of \(v_{i}\), then \(\hat{d}_{P_{t}^{W}'} = (\hat{d}_{P_{t1}^{W}}, \ldots, \hat{d}_{P_{td_{2}}^{W}})\), and we can write

\[
||A\hat{Q}^{-1} \hat{d}_{P_{t}^{W}} (I_{d_{2}} \otimes Q_{1i}^{-1} R_{t})||^{2} = \sum_{d=1}^{d_{2}} ||A\hat{Q}^{-1} \hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}}||^{2} \\
= \sum_{d=1}^{d_{2}} \text{tr}(A\hat{Q}^{-1} \hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}} R_{t}' \hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}} A') = n \sum_{d=1}^{d_{2}} \text{tr}(A\hat{Q}^{-1} \hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}} \hat{d}_{P_{td}'} Q_{1i}^{-1} A'),
\]

and \(\hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}} \hat{d}_{P_{td}'} = \left( \frac{1}{n} \sum_{i=1}^{n} \rho_{i}^{W} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} r_{it}' \right) \left( \frac{1}{n} \sum_{i=1}^{n} r_{it} r_{it}' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \rho_{i}^{W} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} r_{it}' \right)'\).

By \(R_{t}'(R_{t}')^{-1}R_{t}\) being an orthogonal projection matrix,

\[
\hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}} \hat{d}_{P_{td}'} \leq \frac{1}{n} \sum_{i=1}^{n} (\rho_{i}^{W})^{2} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \left( \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \right)' ,
\]

implying \(||\hat{d}_{P_{td}^{W} Q_{1i}^{-1} R_{t}}||^{1/2} \leq b_{2}(K)\) and

\[
\text{tr}(A\hat{Q}^{-1} \hat{d}_{P_{td}^{W} Q_{1i}^{-1} \hat{d}_{P_{td}'} Q_{1i}^{-1} A'}) \leq \frac{1}{n} \sum_{i=1}^{n} (\rho_{i}^{W})^{2} \text{tr}(A\hat{Q}^{-1} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \left( \frac{\partial p_{K}(V_{i})}{\partial v_{i}} \right)') \hat{Q}^{-1} A' \\
\leq \frac{1}{n} ||A\hat{Q}^{-1}||^{2} \sum_{i=1}^{n} (\rho_{i}^{W})^{2} ||\frac{\partial p_{K}(V_{i})}{\partial v_{i}}||^{2} = O_{P}(1) b_{2}(K)^{2},
\]

since \(\rho^{W}(.)\) is bounded. Hence \(||A\hat{Q}^{-1} \hat{d}_{P_{t}^{W}} (I_{d_{2}} \otimes Q_{1i}^{-1} R_{t})||^{2} = O_{P}(nb_{2}(K)^{2})\) and we obtain by Assumption 4.3 (7),

\[
||(B3.1)|| = O_{P}(\sqrt{n}b_{2}(K)\sqrt{n}L^{-\gamma}/\sqrt{n}) = O_{P}(b_{2}(K)\sqrt{n}L^{-\gamma} / \sqrt{n}) = o_{P}(1).
\]

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

\[
||(B3.2)|| = || \sum_{i=1}^{n} A\hat{Q}^{-1} \sum_{i=1}^{n} \rho_{i}^{W} \frac{\partial p_{K}(V_{i})}{\partial v_{i}} [\beta_{t}^{L} r_{it}' - b_{it}]/\sqrt{n} ||/ \sqrt{n} \leq ||A\hat{Q}^{-1}|| nb_{2}(K) C L^{-\gamma} / \sqrt{n} \\
= O_{P}(1) b_{2}(K) \sqrt{n} L^{-\gamma} = o_{P}(1),
\]

65
again by Assumption 4.9 (7). This implies that \((B3) = -(B3.3) + o_{\bar{\tau}}(1)\), with

\[
(B3.3) = \sum_{t=1}^{T} A\hat{Q}^{-1}d\tilde{P}_{t}^{W} (I_{d_2} \otimes Q_{1t}^{-1}) \text{Vec}(R_{t}\tilde{v}_{t}^i)/\sqrt{n}.
\]

First, by \(\mathbb{E}(||v_{it}||^2|\xi_{it} = \xi_t)\) bounded, \(||R_{t}\tilde{v}_{t}^i/\sqrt{n}|| = O_{\bar{\tau}}(\sqrt{L})\), which gives

\[
||\sum_{t=1}^{T} A\hat{Q}^{-1}d\tilde{P}_{t}^{W} [I_{d_2} \otimes Q_{1t}^{-1} - I_{d_2}] \text{Vec}(R_{t}\tilde{v}_{t}^i)/\sqrt{n}||,
\]

\[
\leq ||A\hat{Q}^{-1}|| \sum_{t=1}^{T} ||d\tilde{P}_{t}^{W} (I_{d_2} \otimes Q_{1t}^{-1/2})|| \ |I_{d_2} \otimes Q_{1t}^{-1/2}|| \ |I_{d_2} \otimes (I_{L} - Q_{1t})|| \ |R_{t}\tilde{v}_{t}^i/\sqrt{n}||,
\]

\[
\leq O_{\bar{\tau}}(1)Cb_2(K)O_{\bar{\tau}}(1)\Delta Q_1\sqrt{L} = O_{\bar{\tau}}(b_2(K)\Delta Q_1\sqrt{L}) = o_{\bar{\tau}}(1),
\]

by Assumption 4.9 (7). Similarly

\[
||\sum_{t=1}^{T} A\hat{Q}^{-1}(d\tilde{P}_{t}^{W} - dP_{t}^{W}) \text{Vec}(R_{t}\tilde{v}_{t}^i)/\sqrt{n}|| \leq C||A\hat{Q}^{-1}|| \ |d\tilde{P}_{t}^{W} - dP_{t}^{W}|| \ |R_{t}\tilde{v}_{t}^i/\sqrt{n}||
\]

\[
= O_{\bar{\tau}}(\Delta_{\tilde{P}}\sqrt{L}) = o_{\bar{\tau}}(1).
\]

Finally, we write \(dP_{td}^{W} = \mathbb{E}(\rho_{t}^{i} \hat{\varphi} K(V_{i})/\hat{\varphi} v_{td} r_{it})\), as well as \(v_{td}\) the \(d^{th}\) component of \(v_{it}\) and \(\tilde{v}_{td} = (v_{1td}, ..., v_{ndtd})'\). Then

\[
\mathbb{E}(||dP_{td}^{W} R_{t}\tilde{v}_{td}/\sqrt{n}||^2) = \text{tr}(dP_{td}^{W} \mathbb{E}(R_{t}\tilde{v}_{td}^{i} \tilde{v}_{td}^{i} r_{it}^{i} r_{it}^{i}) dP_{td}^{W} r_{it}^{i} r_{it}^{i}) /n \leq C \text{tr}(dP_{td}^{W} \mathbb{E}(R_{t} R_{t}^{i}) dP_{td}^{W} r_{it}^{i} r_{it}^{i}) /n
\]

\[
\leq C ||dP_{td}^{W} dP_{td}^{W} r_{it}^{i}|| = C\mathbb{E}(\rho_{t}^{i} \hat{\varphi} K(V_{i})/\hat{\varphi} v_{td} r_{it}^{i}) \mathbb{E}(r_{it}^{i} r_{it}^{i})^{-1} \mathbb{E}(\rho_{t}^{i} \hat{\varphi} K(V_{i})/\hat{\varphi} v_{td})
\]

\[
\leq C\mathbb{E}(\rho_{t}^{i} \hat{\varphi} K(V_{i})/\hat{\varphi} v_{td}) \mathbb{E}(\rho_{t}^{i} \hat{\varphi} K(V_{i})/\hat{\varphi} v_{td}) \leq Cb_2(K)^2,
\]

where the second to last inequality follows from taking the orthogonal projection matrix argument to the limit. This implies that \(||dP_{td}^{W} R_{t}\tilde{v}_{td}/\sqrt{n}|| = O_{\bar{\tau}}(b_2(K))\), and

\[
||\sum_{t=1}^{T} A(\hat{Q}^{-1} - I_{d_2})dP_{td}^{W} \text{Vec}(R_{t}\tilde{v}_{td})/\sqrt{n}|| \leq \sum_{t=1}^{T} \sum_{d=1}^{d_2} ||A\hat{Q}^{-1}(I_{d_2} - \hat{Q})dP_{td}^{W} R_{t}\tilde{v}_{td})/\sqrt{n}||
\]

\[
\leq ||A\hat{Q}^{-1}|| O_{\bar{\tau}}(\Delta_{\tilde{Q}})O_{\bar{\tau}}(b_2(K)) = O_{\bar{\tau}}(\Delta_{\tilde{Q}}b_2(K)) = o_{\bar{\tau}}(1),
\]

by Assumption 4.9 (7). We can now write

\[
(B3.3) = \frac{1}{\sqrt{n}} A \sum_{t=1}^{T} dP_{t}^{W} \text{Vec}(R_{t}\tilde{v}_{t}^i)/\sqrt{n} + o_{\bar{\tau}}(1) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} A \sum_{t=1}^{T} dP_{t}^{W} v_{it} \otimes r_{it} + o_{\bar{\tau}}(1),
\]

where the term appearing in the sum over \(n\), were the weight matrices not normalized, would become \(\sum_{t\leq T} A\mathbb{E}(p_{it} p_{it}^{-1})^{-1} dP_{t}^{W} (I_{d_2} \otimes (r_{it} r_{it}^{-1}) v_{it} \otimes r_{it})\). Adding all terms appearing in \((B)\), one obtains,

\[
(B) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} A \left[ p_{it} e_{i}^{W} - \sum_{t=1}^{T} dP_{t}^{W} (v_{it} \otimes r_{it}) \right] + o_{\bar{\tau}}(1).
\]
The remaining term in the expression of \( \sqrt{n}(a(h^W) - a(h^W)) \) is \((C) = A\hat{Q}^{-1} \hat{P}^{W}(\hat{h}^W - \bar{h}^W)/\sqrt{n} \). This term is similar to the second term in equation (A.16) of NPV99, p598, where the regression function is becomes \( h^W \). Since \( v \mapsto h^W(\tau(v)) \) is by composition twice continuously differentiable and has bounded second order derivative on the extended support, one obtains using NPV99

\[
(C) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A \sum_{t=1}^{T} H^W_t(v_{it} \otimes r_{it}) + o_P(1),
\]

adapting to the fact that \( h^W \) is here function of \( T \) generated covariates instead of one and using \( \sqrt{n}L^{-\gamma_1}, \sqrt{n}b_1(K)\Delta_n^2, \sqrt{T}\Delta_Q, \sqrt{K}\Delta_Q \) and \( \sqrt{T}\Delta_H \) converge to zero as \( n \) goes to infinity. Note that, absent the normalization of the weight matrices, the term summed over \( ? \) adapting to the fact that \( h^W \) is the mean square projection of \( \lambda_a \) on the functional space spanned by \( p || \lambda_a \) using Assumption 4.10 (3) and (4).

Proof of Lemma 4.2 By Assumption 4.9[5], we can assume wlog that \( E(p_i p'_i) = I_K \) and \( E(r_{it} r'_{it}) = I_L \). Note that \( \hat{\lambda}_a(v) = A p^K(v) = E(\hat{\lambda}_a(V)p^K(V)' \hat{p}^K(v) \) is the mean square projection of \( \lambda_a \) on the functional space spanned by \( p^K \). As in the proof of Theorem 3 in Newey (1997), this implies that \( E(||\hat{\lambda}_a(V) - \lambda_a(V)||^2) \leq E(||a p^K(V) - \lambda_a(V)||^2) \), which gives \( E(||e^W(\hat{\lambda}_a(V) - \lambda_a(V)||^2) \leq C K^{-\gamma_3} \to 0 \), using Assumption 4.10[3] and (4).

Following NPV99, writing \( \hat{\lambda}_{a,tld}(\xi_t) = A H^W_t r^L(\xi_t) = E \left( \hat{\lambda}_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \), and since \( E \left( \hat{\lambda}_a(V) - \lambda_a(V) \right) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \) is the mean square projection of the function \( E \left( \hat{\lambda}_a(V) - \lambda_a(V) \right) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \) on the functional space spanned by \( r^L \), by properties of projection we have

\[
E \left( ||\hat{\lambda}_{a,tld}(\xi_t) - E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) ||^2 \right) 
\]

\[
\leq E \left( ||\hat{\lambda}_a(V) - \lambda_a(V)||^2 \right) \leq CE \left( ||\hat{\lambda}_a(V) - \lambda_a(V)||^2 \right) \to 0,
\]

where the last inequality holds by Assumption 4.9[3].

Since \( E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \) is the mean square projection of \( E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \), then

\[
E \left( ||E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) - E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) ||^2 \right) 
\]

\[
\leq E \left( ||\lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) - E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) ||^2 \right) \to 0.
\]

This implies \( E \left( ||\hat{\lambda}_{a,tld}(\xi_t) - E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) ||^2 \right) \to 0 \), and by Assumption 4.9[8]

\[
E \left( ||v_{td} \left[ \hat{\lambda}_{a,tld}(\xi_t) - E \left( \lambda_a(V) \frac{\partial h^W(V)}{\partial v_{td}} \otimes r^L(\xi_t)' \right) r^L(\xi_t) \right] ||^2 \right) \to 0.
\]
For the third result, we need
\[
\mathbb{E} \left( \left\| \frac{\partial \hat{\lambda}_a(V)}{\partial v_{td}} - \frac{\partial \lambda_a(V)}{\partial v_{td}} \right\|^2 \right) \leq 2 \mathbb{E} \left( \left\| \mathbb{E} \left[ (\lambda_a(V)p^K(V)^\prime) \frac{\partial p^K(V)}{\partial v_{td}} - \frac{\partial \lambda_a(V)p^K(V)}{\partial v_{td}} \right] \right\|^2 \right) \\
+ 2 \mathbb{E} \left( \left\| \frac{\partial \lambda_a(V)p^K(V)}{\partial v_{td}} \right\|^2 \right),
\]
where the second term in the sum converges to 0 by Assumption 4.10 (3) and the result obtained above. Moreover,\[ E \]where the last equality is obtained by the same argument as in the previous proof. This implies that
\[ E \]which implies \[ E \]
where the last equality is obtained by the same argument as in the previous proof. This implies that
\[ E \]where the last equality is obtained by the same argument as in the previous proof. This implies that
\[ E \]
where the last equality is obtained by the same argument as in the previous proof. This implies that
\[ E \]
by Assumption 4.9 (3) and the result obtained above. Moreover,
\[ E \]
which implies \[ E \]and \[ E \]by Assumption 4.11 (1) and (2), this guarantees that \[ ||\lambda_a^{(b)}|| = ||\mathbb{E}(Q_i M_0(V_i) V_i^{-1} \hat{k}(V_i))|| \leq C||b_t||. \] The other required conditions hold by Assumption 4.11.

Proof of Result 4.8. We first focus on the functional \[ \lambda_0^{(k)}[b_t] = \int_{\xi} \lambda_k(\xi) b_t(\xi) dF(\xi). \] Newey (1997) shows in the proof of Theorem 2 (equation (A.7) p164 and the subsequent text) that if \[ ||\lambda_0^{(b)}|| \leq C||b_t||, \] then \[ \sqrt{nL^{-\gamma_1}} \to 0, \] and \[ \Delta Q_1 = a_1(\sqrt{L/n}) \to 0, \] and \[ ||\lambda_a|| \] is bounded, then \[ \sqrt{n} \sum_{i=1}^n r_{it} \to \mathbb{E} \] and \[ ||\lambda_0^{(b)}[b_t]|| = ||\mathbb{E}(Q_i \hat{M}_0(V_i) b_t(\xi_i))|| \leq C||b_t||. \] The other required conditions hold by Assumption 4.11.

We now check that the conditions of Assumption 4.9 hold for the functionals \[ \lambda_a^{(k)} \] and \[ \lambda_a^{(M)} \] applied to the two-step estimators \( \hat{k} \) and \( \hat{M} \). Under Assumption 4.4, we have by Lemma C.2 that \( \lambda_{\min}(\mathcal{M}(V)) \geq C \). Together with Assumption 4.11 (1) and (2), this guarantees that \[ ||\lambda_a^{(k)}[\hat{k}]|| = ||\mathbb{E}(Q_i M_0(V_i) V_i^{-1} \hat{k}(V_i))|| \leq C||\hat{k}|| \] and \[ ||\lambda_0^{(M)}[\hat{M}]|| = ||\mathbb{E}(Q_i M_0(V_i) M_0(V_i) g_0(V_i))|| \leq C||\hat{M}||. \] Hence Assumption 4.9 (2) holds for each functional.

Moreover, \[ \rho^M(X_i, Z_i) = M_i - M_0(V_i) \] where \( M_i = I - \hat{X}_i \hat{X}_i^+ \) if \( \hat{X}_i \) is of full rank, or \( M_i = I - \hat{X}_i \hat{X}_i^+ \) if not, with \( \hat{X}_i^+ \) is the Moore Penrose inverse. In either case, \[ ||M_i||_2 \leq 1 \] implying
\[ ||M_i||_F \leq C \text{ and } ||M_0(V_i)||_F = ||E(M_i|V_i)||_F \leq C, \] ensuring that \( \rho^M \) is a bounded function.

By the same argument, Assumption 4.11(1), (2) and (7), \( \rho^k(X_i, Z_i) = [M_i - M_0(V_i)]g_0(V_i) \) is uniformly bounded. By Assumption 2.1 and 2.3, \( k \) is twice continuously differentiable, implying that Assumption 4.9(3) holds for each component of the vector \( \text{Vec}_{12} \).

Thus, \( s_i \) is a bounded function.

**Proof of Result 4.9** Under Assumptions 4.11 and 4.12, Conditions (1), (2) and (3) of Assumption 4.10 holds for \( \rho^M \) and \( \rho^k \) are bounded, as well as \( \mathbb{E}(||e^M||_2^2|X, Z) \) and \( \mathbb{E}(||e^k||_2^2|X, Z) \): this implies that for all \( V, \text{Var}(e^M|V) \leq C \) and \( \text{Var}(e^k|V) \leq C \). Condition (4) of Assumption 4.10 is also satisfied for our choices of \( \rho^M \), hence we can apply Lemma 4.2.

We now use Equation (36) to construct the asymptotic variance. Define

\[
s_i = [\delta_i \mu_i - E(\mu \delta)] + Q^k \tilde{u}_i - \lambda_M(V_i) \text{Vec}(e_i^M) - \lambda_k(V_i) e_i^k
\]

\[
+ \sum_{t=1}^{T} \left[ \mathbb{E} \left[ \frac{\partial \lambda_M(V_i)}{\partial v_t} \text{Vec}(\rho^M_i) \xi_{it} \right] - \mathbb{E} \left[ \lambda_M(V_i) \frac{\partial M_0(V_i)}{\partial v_t} | \xi_{it} \right] \right] v_{it} + \mathbb{E} \left[ \frac{\partial \lambda_k(V_i)}{\partial v_t} \rho_i^k | \xi_{it} \right] - \mathbb{E} \left[ \lambda_k(V_i) \frac{\partial k_0(V_i)}{\partial v_t} | \xi_{it} \right] - \lambda_{bl}(\xi_{it}) v_{it},
\]

where, by a convenient abuse of notation, we denote with \( \frac{\partial \lambda_M(V_i)}{\partial v_t} \text{Vec}(\rho^M_i) \) the sum \( \sum_{j \in (T-1)^2} \rho^M_{ij} \frac{\partial \lambda_i^M(V)}{\partial v_t} \) with \( \rho^M_{ij} \) the \( j^{th} \) component of the vector \( \text{Vec}(\rho^M(X_i, Z_i)) \), and similarly for \( \lambda_k \). We will, in a later step of this proof, simplify the formula for \( s_i \).

We conveniently decompose the difference \( s_{i,n} - s_i \) as

\[
\begin{align*}
s_{i,n} - s_i &= \lambda_M(V_i) \text{Vec}(e_i^M) - \Lambda^M(I_{(T-1)^2} \otimes \Theta) \text{Vec}(e_i^M) \otimes p_i \\
&+ \lambda_k(V_i) e_i^k - \Lambda^k(I_{(T-1)^2} \otimes \Theta) e_i^k \otimes p_i \\
+ \sum_{t=1}^{T} \mathbb{E} \left[ \lambda_M(V_i) \frac{\partial M_0(V_i)}{\partial v_t} | \xi_{it} \right] v_{it} - \Lambda^M(I_{(T-1)^2} \otimes \Theta) H^{M}^t(I_{k_2} \otimes \Theta_1) v_{it} \otimes r_{it} \\
+ \sum_{t=1}^{T} \mathbb{E} \left[ \lambda_k(V_i) \frac{\partial k_0(V_i)}{\partial v_t} | \xi_{it} \right] v_{it} - \Lambda^k(I_{(T-1)^2} \otimes \Theta) H^{k}^t(I_{k_2} \otimes \Theta_1) v_{it} \otimes r_{it} \\
+ \sum_{t=1}^{T} \Lambda^M(I_{(T-1)^2} \otimes \Theta) dP^{M}(I_{k_2} \otimes \Theta_1) v_{it} \otimes r_{it} - \mathbb{E} \left[ \frac{\partial \lambda_M(V_i)}{\partial v_t} \text{Vec}(\rho^M_i) | \xi_{it} \right] v_{it} \\
+ \sum_{t=1}^{T} \Lambda^k(I_{(T-1)^2} \otimes \Theta) dP^{k}(I_{k_2} \otimes \Theta_1) v_{it} \otimes r_{it} - \mathbb{E} \left[ \frac{\partial \lambda_k(V_i)}{\partial v_t} | \xi_{it} \right] v_{it} \\
+ \sum_{t=1}^{T} \lambda_{bl}(\xi_{it}) v_{it} - \Lambda^{bl}(I_{k_2} \otimes \Theta_1) v_{it} \otimes r_{it},
\end{align*}
\]

69
where each line in this sum is of one of the three types of elements analyzed in Lemma 4.2 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 up to a multiplicative constant. To show that it converges to 0 as \( n \) goes to infinity, we use the fact that Assumption 4.10 holds for each \( \lambda_a \), where \( \lambda_a \) is a column of either \( \lambda_M \) or \( \lambda_k \). By Assumption 4.12 and Assumption 4.11, 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 4.8 and Assumption 4.12, the functions multiplying the residuals appearing in the definition of \( s_i \) are all bounded. Together with Assumption 4.12, this guarantees \( \mathbb{E}(|s'_c|^2) < \infty \). For a constant vector \( c \in \mathbb{R}^{d_x} \), \( |c' [\mathbb{E}(s_{in}'s_{in}) - \mathbb{E}(s_is_i')]| \leq \mathbb{E}([s'_c c - s'_c c]^2) + 2 \mathbb{E}([s'_c c - s'_c c]^2)^{1/2} \mathbb{E}([s'_c c - s'_c c]^2)^{1/2} \). Hence, \( |c' [\mathbb{E}(s_{in}'s_{in}) - \mathbb{E}(s_is_i')]| \to 0 \) for all \( c \), implying \( \mathbb{E}(s_{in}'s_{in}) - \mathbb{E}(s_is_i') \to 0 \). That is, \( \Omega \to \text{Var}(s_i) \) as \( n \to \infty \).

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

\[
\begin{align*}
\lambda_M(V_i) \text{Vec}(e_i^M) + \lambda_k(V_i)e_i^k &= \mathbb{E}(Q^M_i | V_i) M_0(V_i)^{-1} M_i \hat{u}_i, \\
\lambda_M(V_i) \frac{\partial M_0(V_i)}{\partial v_t} + \lambda_k(V_i) \frac{\partial k_0(V_i)}{\partial v_t} &= \mathbb{E}(Q^M_i | V_i) \frac{\partial g_0(V_i)}{\partial v_t}, \\
\frac{\partial \lambda_M(V_i)}{\partial v_t} \text{Vec}(\rho_i^M) + \frac{\partial \lambda_k(V_i)}{\partial v_t} \rho_i^k &= -\mathbb{E}(Q^M_i | V_i) M_0(V_i)^{-1} (M_i - M_0(V_i)) \frac{\partial g_0(V_i)}{\partial v_t},
\end{align*}
\]

and since \( \lambda_{bt}(\xi_t) = -\mathbb{E} \left( Q^M_i \frac{\partial g_0(V_i)}{\partial v_t} | \xi_t \right) \), we obtain

\[
\begin{align*}
s_i &= \delta_i \mu_i - \mathbb{E}(\mu \delta) + Q^M_i \hat{u}_i - \mathbb{E}(Q^M_i | V_i) M_0(V_i)^{-1} M_i \hat{u}_i \\
&\quad + \sum_{t=1}^T \mathbb{E} \left( Q^M_i - \mathbb{E}(Q^M_i | V_i) M_0(V_i)^{-1} M_i \right) \frac{\partial g_0(V_i)}{\partial v_t} | \xi_t \right) v_{it}, \\
&= \delta_i \mu_i - \mathbb{E}(\mu \delta) + Q^M_i \hat{u}_i + \sum_{t=1}^T \mathbb{E} \left( Q^M_i \frac{\partial g_0(V_i)}{\partial v_t} | \xi_t \right) v_{it}.
\end{align*}
\]

Thus \( \text{Var}(s_i) = \Omega_0 \) and \( \Omega \to n \to \infty \). By \( \Omega_0 \geq CI_{d_x} \), we obtain \( \Omega^{-1/2} \to n \to \infty \). \( \Omega_0^{-1/2} \leq C^{-1/2} I_{k_x} \).

We again use Lemma 4.2 to prove that \( ||\Lambda^M||, ||\Lambda^k||, \) and \( ||\Lambda^{bt}|| \) are bounded. Indeed, since wlog we can assume \( \Theta_1 = I_L \), we have \( ||\Lambda^{bt}||^2 = \text{tr}(\Lambda^{bt} \Lambda^{bt}) = \text{tr}(\Lambda^{bt} (I_{k_2} \otimes \Theta_1) \Lambda^{bt}) \). Using the notation of Lemma 4.2 with \( \lambda^L_{bt}(\xi) = \mathbb{E}(\lambda^L_{bt}(\xi_t) r^L(\xi_t)^t) v^t(\xi) \), this gives \( ||\Lambda^{bt}||^2 = \text{tr} \left( \sum_{j=1}^{T-1} \mathbb{E} \left[ \lambda^L_{bt}(\xi_t) \lambda^L_{bt}(\xi_t)^t \right] \right) \). However, by Lemma 4.2 we know that \( \mathbb{E}(||\lambda^L_{bt}(\xi_t) - \lambda^L_{bt}(\xi_t)||^2) \to 0 \), under Assumption 4.12. The same reasoning we used for \( \mathbb{E}(s_{in}'s_{in}) - \mathbb{E}(s_is_i') \) applies, and since \( \lambda^L_{bt}(\cdot) \) is a bounded function, we obtain \( \mathbb{E} \left( \lambda^L_{bt}(\xi_t) \lambda^L_{bt}(\xi_t)^t \right) \to n \to \infty \). Therefore

\[
||\Lambda^{bt}||^2 \to n \to \infty \text{ tr} \left( \sum_{j=1}^{T-1} \mathbb{E} \left( \lambda^L_{bt}(\xi_t) \lambda^L_{bt}(\xi_t)^t \right) \right) \leq C.
\]

Hence \( ||\Lambda^{bt}||^2 \) is bounded.
The same arguments applied to the functions $\lambda_M$, $\lambda_k$, as well as to
\[
\mathbb{E} \left[ \frac{\partial \lambda_M(V_i)}{\partial v_t} \text{Vec}(\rho_t^M) | \xi_{it} \right], \quad \mathbb{E} \left[ \frac{\partial \lambda_M(V_i)}{\partial v_t} | \xi_{it} \right], \quad \mathbb{E} \left[ \frac{\partial \lambda_k(V_i)}{\partial v_t} - \rho_t^k | \xi_{it} \right] \quad \text{and} \quad \mathbb{E} \left[ \frac{\partial k_0(V_i)}{\partial v_t} | \xi_{it} \right],
\]
would imply that $||\Lambda^M||, ||\Lambda^k||, ||\Lambda^M(I_{(T-1)} \otimes \Theta)(R_t^M - dP_t^M)||$, and $||\Lambda^k(I_{(T-1)} \otimes \Theta)(R_t^k - dP_t^k)||$ are bounded.

\begin{proof}[Proof of Result 4.10]
We start by showing that the stochastic equicontinuity condition, Condition \[4.10], holds. Lemma 1 of CLVK shows that if $(W_i, \eta_i)_{i=1}^n$ is i.i.d, Assumption 4.8 holds if: (A) the class $\mathcal{F} = \{\chi(W, \mathcal{G}) : \mathcal{G} \in \mathcal{H}^\theta_{c,c'}\}$ is $\mathbb{P}$-Donsker, i.e it satisfies $\int_0^\infty \log N(\epsilon, \mathcal{F}, ||\cdot||_{L^2(\mathbb{P})}) d\epsilon < \infty$, where $N(\epsilon, \mathcal{F}, ||\cdot||_{L^2(\mathbb{P})})$ is the covering number with bracketing, and if (B) $\chi(., \mathcal{G})$ is $L_2(\mathbb{P})$
continuous at $\mathcal{G}_0$, that is, $\mathbb{E}(||\chi(\mathcal{W}, \mathcal{G}) - \chi(\mathcal{W}, \mathcal{G}_0)||^2) \to 0$ as $||\mathcal{G} - \mathcal{G}_0||_{H} \to 0$. We now check that each of these conditions is satisfied under our assumptions.

Condition (A): We use $j,l$ to index components of vectors. As in CLVK, it is enough to prove that $\mathcal{F}_l = \{\chi_l(W, \mathcal{G}) : \mathcal{G} \in \mathcal{H}^\theta_{c,c'}\}$ is $\mathbb{P}$-Donsker for each component $l$ of $\chi(\cdot)$. Recall that $\chi(W_i, \mathcal{G}) = Q_1^b \mathcal{M} (\tau ((x_{it}^2 - b_t(\xi_{it})))_{t \leq T})^{-1} k (\tau ((x_{it}^2 - b_t(\xi_{it})))_{t \leq T})$. 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, $V_i = (x_{it}^2 - b_t(\xi_{it}))_{t \leq T}$ and $V_{0,i} = (x_{it}^2 - b_{0,i}(\xi_{it}))_{t \leq T}$. Note that $V_0 = \tau(V_0)$. We decompose
\[
\chi(W_i, \mathcal{G}) - \chi(W_i, \mathcal{G}_0) = Q_1^b \mathcal{M} (\tau[V])^{-1} k (\tau[V]) - k_0 (\tau[V])
\]
\[
+ Q_1^b \mathcal{M} (\tau[V])^{-1} [\mathcal{M}_0 (\tau[V]) - \mathcal{M} (\tau[V])] \mathcal{M}_0 (\tau[V])^{-1} k_0 (V_0)
\]
\[
+ Q_1^b \mathcal{M} (\tau[V])^{-1} [\mathcal{M}_0 (V_0) - \mathcal{M} (V_0)] \mathcal{M}_0 (V_0)^{-1} k_0 (V_0)
\]
\[
+ Q_1^b \mathcal{M} (\tau[V])^{-1} [k_0 (\tau[V]) - k_0 (V_0)].
\]

Since $(\mathcal{G}, \mathcal{G}_0) \in \mathcal{H}^\theta_{c,c'} \times \mathcal{H}^\theta_{c,c'}$, the norms of each functional and its first order derivatives are bounded. Moreover the derivatives of $\tau$ are bounded. This implies that $||\mathcal{M}_0 (V_0) - \mathcal{M}_0 (\tau[V])|| \leq c ||V_0 - V||$, and the same result holds for $k$. Hence, using (46), $||\chi_l(W_i, \mathcal{G}) - \chi_l(W_i, \mathcal{G}_0)|| \leq C ||Q_1^b|| (\sum_{j=1}^{T-1} |\mathcal{M}_j - \mathcal{M}_0| ||k_j - k_0||_0 + \sum_{l \leq T} \sum_{j=1}^{d_2} |b_{l,j} - b_{0,l,j}||_x)$, where the constant $C$ depends on $c$ and $c'$. By Assumption 4.13 $\mathbb{E}(||Q_1^b||^2) < \infty$ which implies by the proof of Theorem 3 of CLVK that
\[
N(\epsilon, \mathcal{F}, ||\cdot||_{L^2(\mathbb{P})}) \leq N (\epsilon/c^Q, C^\theta_c(S^\gamma_{\xi}), ||\cdot||_x)^{(T-1)^2 + T - 1} \Pi_{l \leq T} N (\epsilon/c^Q, C^\theta_c(S^\gamma_{\xi}), ||\cdot||_x)^{d_2},
\]
where $N(\epsilon, C^\theta_c(S^\gamma_{\xi}), ||\cdot||_x)$ denotes the covering number of the class $C^\theta_c(S^\gamma_{\xi})$, and $c^Q = 2[(T - 1)^2 + T - 1 + T d_2] \mathbb{E}(||Q_1^b||^2)$, is the size of the brackets constructed in CLVK.

It is known that for $S^\gamma_{\xi}$ a bounded subset of $\mathbb{R}^k$, $\log N(\epsilon, C^\theta_c(S^\gamma_{\xi}), ||\cdot||_x) \leq -\epsilon^{-k/\theta}$. By Assumption 4.13 $q > \max(T d_2, d_z + d_1)/2$, which implies that $\mathcal{F}_j$ is $\mathbb{P}$-Donsker. Therefore, Condition (A) is satisfied.

71
Condition (B): By $\mathbb{E}(|Q^g_0|^2) \leq C$ and using once more the decomposition given by (46),
$\mathbb{E}(|\chi(W_t, G) - \chi(W_t, G_0)|^2) \leq C||G - G_0||^2_H$ which gives the wanted result.

We now show that Assumption 4.8 (2) holds. This condition is on the remainder of the linearization, $||\mathcal{X}(G) - \mathcal{X}^G(G_0)[G - G_0]||$. Note that

\[
\mathcal{X}(G) - \mathcal{X}^G(G_0)[G - G_0] = \mathbb{E}(Q^g_0 \mathcal{M}^g_0(V_0)^{-1} - \mathcal{M}^g_0(V_0)^{-1}[k(V_0) - k_0(V_0)])
\]

We use this decomposition and bound each line separately. We show how to find upper bounds for the first and second lines, both of which will be less than $||G - G_0||^2_H$ up to a multiplicative constant. The upper bounds for the third and fourth lines of this decomposition can be obtained in a similar fashion. By the triangular inequality, this will give $||\mathcal{X}(G) - \mathcal{X}^G(G_0)[G - G_0]|| \leq C||G - G_0||^2_H$, as desired. First,

\[
||\mathbb{E}(Q^g_0 \mathcal{M}^g_0(V_0)^{-1} - \mathcal{M}^g_0(V_0)^{-1}[k(V_0) - k_0(V_0)])|| \leq C\mathbb{E}(|Q^g_0||^2)
\]

As for the second line of the decomposition of $\mathcal{X}(G) - \mathcal{X}^G(G_0)[G - G_0]$, we write

\[
||\mathbb{E}(Q^g_0 \mathcal{M}^g_0(V_0)^{-1} - \mathcal{M}^g_0(V_0)^{-1}[k(V_0) - k_0(V_0)])|| \leq C||G - G_0||^2_H.
\]
\[
\begin{aligned}
&+ \| \mathbb{E}(Q^\delta I M_0(V_0)^{-1} [k_0(\tau[V]) - k_0(V_0) - \frac{\partial k_0}{\partial V_0}(V_0)[V-V_0]]) \|
\leq C \mathbb{E}(||Q^\delta||) \left( \left( \sum_{t \leq T} \sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_x \right)^2 + (\sum_{t \leq T} \sum_{j=1}^{d_2} ||b_{t,j} - b_{0,t,j}||_x)(\sum_{j=1}^{(T-1)^2} |M_j - M_{0,j}|^2) \right)
\leq C ||G - G_0||_t^2,
\end{aligned}
\]

where the inequality for the third term in this equation holds by Assumption 4.11 (1), by the Jacobian of \( \tau \) being the identity matrix when evaluated at \( V_0 \) (since \( V_0 \in \mathcal{S}_V \)) and by the second order derivative of \( \tau \) being bounded.

\( \square \)

**Proof of Result 4.12.** Define \( \Phi = \mathbb{P}(\det(\hat{X}_i^t \hat{X}_i) > \delta_0)^{-1} \) and \( \phi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i \). The estimator of the average effect \( \mathbb{E}(\mu|\delta) \) is \( \hat{\mu} = \frac{\hat{\mu}^\delta}{\phi_n} \). We define \( \Sigma_0 = \text{Var}(\hat{s}', \delta_i') \). Since \( \Omega_0 > 0 \) by Assumption 4.12, then Assumption 4.14 guarantees that \( \Sigma_0 > 0 \). We decompose

\[
\begin{aligned}
\sqrt{n} \phi_n [\hat{\mu} - \mathbb{E}(\mu|\delta)] &= \sqrt{n} [\hat{\mu}^\delta - \mathbb{E}(\mu|\delta)] + \sqrt{n} \mathbb{E}(\mu|\delta) [\Phi - \phi_n],
\end{aligned}
\]

\[
\begin{aligned}
= \sqrt{n} (I_{d_x}; \mathbb{E}(\mu|\delta)) \left[ (\hat{\mu}^\delta - \frac{\mathbb{E}(\mu|\delta)}{\phi_n}) \right], \tag{47}
\end{aligned}
\]

where \( (I_{d_x}; \mathbb{E}(\mu|\delta)) \) is of size \( d_x \times (d_x + 1) \).

We first show

\[
\sqrt{n} \Sigma_0^{-1/2} \left[ (\hat{\mu}^\delta - \frac{\mathbb{E}(\mu|\delta)}{\phi_n}) \right] \to^d \mathcal{N}(0, I_{d_x+1}). \tag{48}
\]

By Result 4.11, \( \sqrt{n} \Sigma_0^{-1/2} \left[ (\hat{\mu}^\delta - \frac{\mathbb{E}(\mu|\delta)}{\phi_n}) \right] \to^d \mathcal{N}(0, I_{d_x+1}) \). We obtain the asymptotic distribution in two steps. We prove first that \( \sqrt{n} \Sigma_0^{-1/2} \sum_{i=1}^{n} (\hat{s}_i - \bar{s}_i - \Phi) \to^d \mathcal{N}(0, I_{d_x+1}) \). We show in a second step that \( \Sigma_n \to \Sigma_0 \), which will yield the desired result. We follow Newey, Powell, and Vella (1999) in proving a Lindeberg condition for \( c^\prime \Sigma_n (s_i', \delta_i - \Delta)' \) for any constant vector \( c \in \mathbb{R}^{d_x+1} \) such that \( \|c\| = 1 \). More precisely, if for all such \( c \), \( \frac{1}{\sqrt{n}} c^\prime \Sigma_n^{-1/2} \sum_{i=1}^{n} (s_i', \delta_i - \Phi)' \to^d \mathcal{N}(0, 1) \), this first result will be a consequence of the Cramér-Wold theorem. Write \( S_{i,n} = c^\prime \Sigma_n (s_i', \delta_i - \Phi)' \), then \( \mathbb{E}(S_{i,n}) = 0 \) and \( \text{Var}(S_{i,n}) = 1 \). 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 \( \rho^M \) and \( \rho^K \) bounded under Assumption 4.11, \( \mathbb{E}([|\hat{u}_i|X_i, Z_i]) = 0 \) and \( \mathbb{E}([|\hat{u}_i|^4|X_i = X, Z_i = Z]) \leq C \) for all \( (X, Z) \), then \( \mathbb{E}([|e_i^M|^4|V_i = V]) \leq C \) and \( \mathbb{E}([\|\text{Vec}(e_i^M)|^4|V_i = V]) \leq C \). Fix \( \epsilon > 0 \). We normalize \( \Theta = I_K \) and \( \Theta_1 = I_L \), and obtain

\[
ne^2 \mathbb{E}(S_{i,n}^4 \mathbb{1}[|S_{i,n}| > \epsilon \sqrt{n}]) \leq \mathbb{E}(S_{i,n}^4 \mathbb{1}[|S_{i,n}| > \epsilon \sqrt{n}]) \leq \mathbb{E}(S_{i,n}^4),
\]

73
\begin{align*}
&\leq C \left( \mathbb{E}[|\mu_{i} - \mathbb{E}(\mu)|^4] + \mathbb{E}[|Q_{i}^3 u_{i}|^4] + ||\Lambda^M||^4 \mathbb{E}[||\text{Vec}(e_{i}^M) \otimes p_{i}||^4] + ||\Lambda^k||^4 \mathbb{E}[||e_{i}^k \otimes p_{i}||^4] \\
&\quad + \sum_{t \leq T} ||[\Lambda^M (H_t^M - dP_t^M) + \Lambda^k (H_t^k - dP_t^k) + \Lambda^t]||^4 \mathbb{E}[||v_{it} \otimes r_{it}||^4] + \mathbb{E}[||\delta_{i} - \Phi||^4] \right).
\end{align*}

We can bound \( \mathbb{E}(||v_{it} \otimes r_{it}||^4) = \mathbb{E}(||r_{it}||^4 ||v_{it}||^4) \leq C \mathbb{E}(||r_{it}||^4) \) by Assumption 4.14, and \( \mathbb{E}(||r_{it}||^4) \leq a_1(L)^2 \text{tr}(\mathbb{E}(r_{it}'r_{it})) = a_1(L)^2 L \). Similarly, by Assumption 4.14, \( \mathbb{E}(||e_{i}^k \otimes p_{i}||^4) = O(b_1(K)^2 K) \) and \( \mathbb{E}(||\text{Vec}(e_{i}^M) \otimes p_{i}||^4) = O(b_1(K)^2 K) \). Therefore, by Result 4.9, \( n \epsilon^2 \mathbb{E}(S_{i,n}^2 (|S_{i,n}| > \epsilon \sqrt{n})) = O(b_1(K)^2 K + a_1(L)^2 L) \).

Assumption 4.11 (6) 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 \( \mathbb{E}(S_{i,n}^2 (|S_{i,n}| > \epsilon \sqrt{n})) \to 0 \) holds.

The second step to obtain (48) requires \( \Sigma_n \to \Sigma_0 \). This is a consequence of the proof of Result 4.9. Now we can use (47) with (48) to obtain by a delta method argument
\begin{align*}
\sqrt{n} \phi_n [\hat{\mu} - \mathbb{E}(\mu|\delta)] \to^d \mathcal{N} (0, (I_{d_x}; \mathbb{E}(\mu|\delta)) \Sigma_0 (I_{d_x}; \mathbb{E}(\mu|\delta))'),
\end{align*}

hence \( \sqrt{n} [\hat{\mu} - \mathbb{E}(\mu|\delta)] \to^d \mathcal{N} (0, \Phi^{-2} (I_{d_x}; \mathbb{E}(\mu|\delta)) \Sigma_0 (I_{d_x}; \mathbb{E}(\mu|\delta))'). \) \( \Box \)