Minimum Distance Estimation of Quantile Panel Data Models[*]

Blaise Melly[†]and Martina Pons[‡]

First version: November 2020
This version: October 23, 2024

Abstract

We propose a minimum distance estimation approach for quantile panel data models where unit effects may be correlated with covariates. This computationally efficient method involves two stages: first, computing quantile regression within each unit, then applying GMM to the first-stage fitted values. Our estimators are applicable to (i) classical panel data, tracking units over time, and (ii) grouped data, where individual-level data are available, but treatment varies at the group level. Depending on the exogeneity assumptions, this approach provides quantile analogs of classic panel data estimators, including fixed effects, random effects, between, and Hausman-Taylor estimators. In addition, our method offers improved precision for grouped (instrumental) quantile regression compared to existing estimators. We establish asymptotic properties as both the number of units and observations per unit jointly diverge to infinity. Additionally, we introduce an inference procedure that automatically adapts to potentially unknown convergence rates of the estimators. Monte Carlo simulations demonstrate that our estimator and inference procedure perform well in finite samples, even when the number of observations per unit is moderate. In an empirical application, we examine the impact of the food stamp program on birth weights. Our findings reveal that the program's introduction increased birth weights predominantly at the lower end of the distribution, demonstrating our method's ability to capture heterogeneous effects across the outcome distribution.

1   Introduction

Quantile regression, introduced by Koenker and Bassett (1978), is a powerful tool for analyzing the effect of policies on the distribution of an outcome variable. The quantile treatment effect function provides more information than the average treatment effect, allowing, for instance, evaluation of the treatment's impact on inequality. When panel data are available, new identification and estimation strategies become feasible. Researchers can alleviate endogeneity

concerns, for instance, by allowing for correlated group effects. They can obtain more precise estimates using a random-effects estimator, or exploit individual-level variables to identify the impact of group-level variables, e.g., with the Hausman and Taylor (1981) estimator. In this paper, we propose a minimum distance estimation approach to quantile panel data models, which provides quantile analogs of classical least-squares panel data estimators such as fixed effects, random effects, between, and Hausman-Taylor estimators.

We define panel data as a dataset structure where observations are organized along two dimensions. While classical panel data typically consist of individuals repeatedly measured over time periods, our results extend to group data, where individual-level observations are available and treatments often vary at the group level. For example, Autor et al. (2013) use commuting zones in the United States as groups, while Angrist and Lang (2004) employ schools. In both cases, treatments vary only between groups, but individual data are essential for estimating the conditional distribution of outcomes within each group. In this paper, we adopt a general notation (i and j subscripts) applicable to both classical panel and group data. We primarily use terminology (individuals and groups) more common to group data, as our application falls into this category. We present general results before discussing their specific applications to classical panel data in Section 5 and group data in Section 4.

Our estimation approach involves two stages. The first stage consists of group-level quantile regressions using individual-level covariates at each quantile of interest. In the second stage, we regress the fitted values from the first stage on individual-level and group-level variables. If these variables are potentially endogenous, an instrumental variable regression or, more generally, the generalized method of moments (GMM) estimator, can be applied. This approach allows for the straightforward inclusion of external or internal instruments in the second stage. The proposed estimator is simple to implement, flexible, computationally fast, and applicable across various fields. While this two-step procedure may seem unconventional, we demonstrate in Section 2.3 that it is numerically equivalent to standard estimators when least squares is used in the first stage along with appropriate instruments.

As a nonlinear estimator, first-stage quantile regression is subject to a finite-sample bias that diminishes as the number of observations per group increases. Our inference procedures are justified within an asymptotic framework where both the number of observations per group n, and the number of groups m diverge to infinity.[1] The asymptotic variance of the sample moments has two components: one arising from the first-stage quantile regression and another from the second-stage GMM regression. In the first stage, since the regressors vary within groups, the relevant number of observations is mn, making the variance proportional to $1/(mn)$. The second-stage estimation error, on the other hand, stems from the randomness of the group effects, with the corresponding variance proportional to $1/m$. However, this second component

---

[1]Large n (often called large T) asymptotics have been widely applied in the nonlinear and dynamic panel data literature. For seminal contributions, see Phillips and Moon (1999), Hahn and Kuersteiner (2002), and Alvarez and Arellano (2003).

is zero if there are no group effects or if the instrument exploits only within variation. Thus, the asymptotic distribution of the estimator is dominated by the component with the slower rate of convergence, which is the second stage error unless there is no group heterogeneity or the instruments are uncorrelated with group membership. As a result, the asymptotic distribution of the estimator is non-standard, as the convergence rate of a coefficient depends on the presence of group heterogeneity and the variation used to identify that coefficient.

In Section 3.2, we consider three specific scenarios before suggesting adaptive estimation and inference procedures in Section 3.3. In the first case, we assume the presence of group heterogeneity and demonstrate that only the coefficients identified through within-group variation can be estimated at the $\sqrt{mn}$ rate. For example, this applies to the fixed effects estimator. In contrast, coefficients associated with variables that are constant within groups rely on between-group variation for identification and are only estimable at the slower $\sqrt{m}$ rate, and for these coefficients, only the second stage error shows up in the asymptotic variance. Consequently, in this scenario, some coefficients converge faster than others. Case 2 assumes no group heterogeneity, eliminating second-stage variance and yielding a uniform $\sqrt{mn}$ convergence rate for all coefficients. Finally, in the third case, we consider the intermediate case when group-level heterogeneity is present but vanishes exactly at the right rate such that both components of the variance matter asymptotically.

These asymptotic results provide valuable insights into the mechanics of our estimator, but they also reveal several limitations. First, the asymptotic distribution and rate of convergence of the slow coefficients vary depending on the presence of group-level heterogeneity, leading to non-adaptive inference. Second, the first-order asymptotic distribution of the slow coefficients is unaffected by the first-stage estimation error. Third, when a coefficient is identified by both within- and between-group variations, the between-group variation becomes first-order asymptotically irrelevant, rendering the random effects estimator equivalent to the fixed effects estimator. To address these issues, we propose an efficient estimator and inference procedure that adapts to the degree of group-level heterogeneity. This approach enables us to introduce a new quantile random effects estimator that is more precise than the fixed effects estimator in finite samples. Additionally, we enhance the accuracy of the estimated standard errors by accounting for the first-stage estimation error. Notably, we find that clustering the standard errors in the second stage inherently adjusts for the first-stage error, providing an adaptive inference procedure that is uniformly valid across the rate of convergence of the moment conditions and the estimator, including intermediate cases.

This paper contributes to the literature on quantile (traditional) panel data models.[2] Koenker (2004) introduced a penalized quantile fixed effects estimator that treats individual heterogene-

---

ity as a pure location shift. Kato et al. (2012) extend this approach by allowing group effects to depend on the quantile of interest and by contributing to the asymptotic theory of the estimator. Galvao and Wang (2015) propose a two-step minimum distance (MD) estimator as a computationally efficient method for estimating fixed effects quantile panel data models. Our framework nests this estimator. However, their focus is solely on the effects of individual-level covariates, without exploiting variation between individuals.[3] In contrast, we aim to estimate the effects of both individual-level and group-level regressors, while also allowing for the use of both internal and external instruments. Galvao and Poirier (2019) suggest using the usual pooled quantile regression estimator in the presence of random effects. Our random effects estimator is different because we focus on the conditional quantile function that also conditions on the group effect (see Remark 1 for a discussion on conditional effects). In other words, we estimate a different parameter and quantile regression is not consistent for this parameter even if the random effects are not correlated with the covariates.

Chetverikov et al. (2016) introduce an IV quantile regression estimator for grouped data. They focus on the effect of variables that vary only between groups and implicitly assume the presence of group-level heterogeneity.[4] While both their approach and ours share the same first stage of estimation, the second stage differs: we regress the fitted values on all variables, whereas they regress the estimated intercept on the group-level regressors. As a result, their estimator is not invariant to reparametrizations of the individual-level regressors. In Table 1, simulations using the same data-generating process as Chetverikov et al. (2016) show that our minimum distance (MD) estimator exhibits substantially lower variance and mean squared error (MSE) across all sample sizes considered—reducing MSE by up to 20 times. In Section 4, we demonstrate and explain why our estimator is more precise than theirs. Furthermore, we contribute to this literature by providing efficient adaptive estimation and inference procedures that remain valid regardless of the degree of group-level heterogeneity, deriving the limiting distribution of the estimator for the coefficients on the individual-level variables, and relaxing the growth condition of n relative to m.

Our class of estimators also includes the MD estimators of Chamberlain (1994) as a special case. We extend his framework by incorporating individual-level regressors, accommodating endogenous regressors and group effects, but requiring the number of groups to approach infinity.[5] Interestingly, in Chamberlain (1994), all the variance arises from the first-stage estimation, consistent with classical MD estimation, while in Chetverikov et al. (2016), the variance originates entirely from the second stage. In our framework, the variance can stem from either the first or

---

[3]Galvao and Wang (2015) consider a traditional panel data setting; thus, in their terminology, they focus on estimating the effects of time-varying regressors.

[4]Their asymptotic distribution becomes degenerate in the absence of group effects, suggesting that the rate of convergence is faster in such cases.

[5]Chamberlain (1994) uses different terminology because he focuses on cross-sectional regressions. He analyzes a quantile regression model with a finite number of combinations of regressor values, where the number of cells (groups in our terminology) is finite, and the regressors are constant within each cell.

second stage, depending on the presence or absence of group effects, with the estimated standard errors capturing the relevant leading component.

Chernozhukov et al. (2013b) introduced distribution regression as an alternative to quantile regression for estimating the entire conditional distribution of outcomes given covariates. Fernández-Val et al. (2022) extend this approach by developing a dynamic distribution regression panel data model with heterogeneous coefficients across groups. In their framework, the first stage involves regressing the outcome on individual-level covariates using distribution regression, followed by a second stage where the coefficients are projected onto group-level instruments. We build on their proof strategy to demonstrate that our inference procedure remains uniformly valid with respect to the degree of heterogeneity. However, our approach differs in several key aspects: we use quantile regression instead of distribution regression, project the fitted values rather than the coefficients, consider both individual-level and group-level instruments, and optimally combine these instruments using GMM. Although our method is limited to continuous outcomes, we find that quantile regression coefficients are easier and more intuitive to interpret.

As an empirical application, we extend the work of Almond et al. (2011) by estimating the distributional effect of the Food Stamp Program on birth weight. Following the enactment of the Food Stamp Act, the number of counties implementing the program increased substantially in the late 1960s and early 1970s. To apply our minimum distance estimator, we define groups as county-trimester cells. The subscript j indexes a county-trimester cell, while the subscript i defines an individual within this cell. We estimate the model separately for black and white mothers and find that the Food Stamp Program has a positive impact on the lower tail of the birth weight distribution, particularly among black mothers.

The remainder of the paper is structured as follows. Section 2 introduces the model and the estimator and shows that traditional least-squares estimators can be implemented as MD estimators. Section 3 develops the asymptotic theory. Section 4 extends the discussion to grouped data and compares our estimator with the grouped IV quantile regression approach of Chetverikov et al. (2016). Section 5 applies our framework to traditional panel data, proposing quantile analogs of the within, between, and random effects estimators, as well as the Hausman test. Both Sections 4 and 5 include Monte Carlo simulations to evaluate finite sample performance. In Section 6, we present the empirical application, and Section 7 concludes.

2 Model and Minimum Distance Estimator

2.1 Quantile Model

We want to learn the effects of the individual-level variables x_{1ij} and the group-level variables x_{2j} on the distribution of an outcome y_{ij}. We observe these variables for the groups $j = 1, \ldots, m$

and individuals $i = 1, \ldots, n.$[6] For some quantile index $0 < \tau < 1$, we assume that

$$Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j) = x'_{1ij}\beta(\tau) + x'_{2j}\gamma(\tau) + \alpha(\tau, v_j), \tag{1}$$

where $Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j)$ is the τth conditional quantile function of the response variable y_{ij} for individual i belonging to group j given the K_1-vector of individual-level regressors x_{1ij}, the K_2-vector of group-level variables x_{2j}, and an unobserved random vector v_j of unrestricted and unknown dimension. In total, there are $K_1 + K_2 = K$ parameters to estimate. The parameters $\beta(\tau)$, $\gamma(\tau)$ and the unobserved group heterogeneity $\alpha(\tau, v_j)$ can depend on the quantile index τ. Depending on the setting, $\beta(\tau)$ or $\gamma(\tau)$ (or both) might be the parameters of interest. We normalize $\mathbb{E}[\alpha(\tau, v_j)] = 0$, which is not restrictive because x_{2j} includes a constant.

Remark 1 (**Conditional versus unconditional effects**). In contrast to the average effect, the definition of a quantile treatment effect depends on the conditioning variables. In this paper, we model the distribution of y_{ij} conditionally on the covariates and on the group effect $\alpha(\tau, v_j)$. Thus, even if the group effects are independent of the regressors, we identify different parameters than those identified by quantile regression as introduced by Koenker and Bassett (1978) or by instrumental variable quantile regression as introduced by Chernozhukov and Hansen (2005). The following example illustrates the difference between these parameters. Consider an application where each group j corresponds to a region and each unit i to an individual within this region. We do not have any x_{1ij} variable. We are interested in the effect of a binary treatment x_{2j}, which has been randomized and is, therefore, independent from $\alpha(\tau, v_j)$. $\gamma(\tau)$ is the effect of this treatment for individuals that rank at the τ quantile of y_{ij} in *their* region. On the other hand, the quantile regression of y_{ij} on x_{2j} identifies the effect for individuals that rank at the τ quantile in the whole country (given the treatment status). These are different parameters except if $\alpha(\tau, v_j)$ is constant over j or if the treatment effect is homogeneous such that $\gamma(\tau) = \gamma$ for all τ. Whether conditional or unconditional quantile treatment effects are of interest depends on the question at hand. For example, conditional quantile treatment effects are particularly of interest to study within-group inequalities when groups might be regions or industries. For example, Autor et al. (2016), and Engelhardt and Purcell (2021) study the effect of the minimum wage on within-state inequality, while Autor et al. (2021) study the effect of trade shock on wage inequality within local labor markets. If the unconditional effect is of interest, one can naturally obtain the unconditional distribution functions by integrating out the group effects (and possibly the other variables) and then inverting the resulting distribution functions to obtain the unconditional quantile functions, see Chernozhukov et al. (2013b).[7]

When model (1) holds, the τ quantile regression of y_{ij} on x_{1ij} and a constant using only observations for group j identifies the slope $\beta(\tau)$ and the intercept $x'_{2j}\gamma(\tau) + \alpha(\tau, v_j)$. To identify

---

[6]For notational simplicity, we assume a balanced panel. However, the results generalize to unbalanced datasets.

[7]We refer to Frölich and Melly (2013) for a discussion about conditional and unconditional treatment effects.

the coefficient on the group-level variables, we need to consider variation across groups. Note that model (1) implies

$$\mathbb{E}\left[Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j)|x_{1ij}, x_{2j}\right] = x'_{1ij}\beta(\tau) + x'_{2j}\gamma(\tau) + \mathbb{E}\left[\alpha(\tau, v_j)|x_{1ij}, x_{2j}\right].$$

If $\alpha(\tau, v_j)$ is exogenous with respect to x_{1ij} and x_{2j} and the linear model is correctly specified, $\mathbb{E}[\alpha(\tau, v_j)|x_{1ij}, x_{2j}] = 0$ and a linear regression identifies the parameters of interest.[8] The last representation suggests a two-step estimation strategy: (i) group-level quantile regression of y_{ij} on x_{1ij}, (ii) OLS regression of the fitted values from the first stage on x_{1ij} and x_{2j}.

When the group effects $\alpha(\tau, v_j)$ are endogenous (possibly correlated with x_{1ij} and x_{2j}), we assume that there is a L-dimensional vector ($L \geq K$) of valid instruments z_{ij} satisfying

$$\mathbb{E}[z_{ij}\alpha(\tau, v_j)] = \mathbb{E}\left[z_{ij}\left(Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j) - x'_{1ij}\beta(\tau) - x'_{2j}\gamma(\tau)\right)\right] = 0. \tag{2}$$

Note that $\beta(\tau)$ is identified in model (1) as long as there is some variation in x_{1ij} within some groups. For instance, we can include the demeaned regressors, $\dot{x}_{1ij} = x_{1ij} - \bar{x}_{1j}$ with $\bar{x}_{1j} = n^{-1}\sum_{i=1}^{n} x_{1ij}$, in the vector of instruments z_{ij} because this variable will satisfy condition (2) under strict exogeneity.[9] On the other hand, we need additional instruments to identify $\gamma(\tau)$. Equation (2) suggests a similar estimation strategy as in the exogenous case but with the instrumental variable estimator (or more generally the GMM estimator) in the second stage: (i) group-level quantile regression of y_{ij} on x_{1ij}, (ii) GMM regression of the fitted values from the first stage on x_{1ij} and x_{2j} using z_{ij} as instrument.

Remark 2 (Skorohod representation). The following Skorohod representation implies the model defined in equation (1):

$$\begin{aligned} y_{ij} &= x_{1ij}\beta(u_{ij}) + x_{2j}\gamma(u_{ij}) + \alpha(u_{ij}, v_j) \\ &= q(x_{1ij}, x_{2j}, u_{ij}, v_j), \end{aligned}$$

where $q(x_{1ij}, x_{2j}, u_{ij}, v_j)$ is strictly increasing in u (while fixing the other arguments).[10] We normalize $u_{ij}|x_{1ij}, x_{2j}, v_j \sim U(0,1)$ such that $q(x_{1ij}, x_{2j}, u, v_j)$ is the conditional quantile function. v_j ranks the groups while u_{ij} ranks the individuals within a group. In this model, a sufficient condition for equation (2) is $(u_{ij}, v_j) \perp\!\!\!\perp z_{ij}$. If the instrument does not vary within groups, only $v_j \perp\!\!\!\perp z_j$ is sufficient.

Remark 3 (Heterogeneous coefficients). Our model allows only the intercept to differ between groups. Now consider a more general model where we also allow the slopes to differ between groups:

$$y_{ij} = x'_{1ij}\beta(u_{ij}, v_j) + x'_{2j}\gamma(u_{ij}, v_j) + \alpha(u_{ij}, v_j). \tag{3}$$

---

[8]Uncorrelation between $\alpha(\tau, v_j)$ and x_{1ij} and x_{2j} is sufficient to identify the linear projection.

[9]In the special case of traditional panel data, the demeaned regressors correspond to the within transformation.

[10]This is the same model as in Chetverikov et al. (2016), where a similar Skorohod representation is derived in their footnote 7.

If we maintain the conditional strict monotonicity assumption with respect to u_{ij}, this model implies that

$$Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j) = x'_{1ij}\beta(\tau, v_j) + x'_{2j}\gamma(\tau, v_j) + \alpha(\tau, v_j). \qquad (4)$$

In the exogenous case where $(x_{1ij}, x_{2j}) \perp\!\!\!\perp v_j$, this implies

$$\mathbb{E}\left[Q\left(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j\right)|x_{1ij}, x_{2j}\right] = x'_{1ij}\int \beta(\tau, v)dF_V(v) + x'_{2j}\int \gamma(\tau, v)dF_V(v) + \int \alpha(\tau, v)dF_V(v)$$
$$= x'_{1ij}\bar{\beta}(\tau) + x'_{2j}\bar{\gamma}(\tau)$$

because we have normalized $\mathbb{E}[\alpha(\tau, v_j)] = 0$. It follows that the linear projection of $Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j)$ on x_{1ij} and x_{2j} identifies the coefficients $\beta(\tau)$ and $\gamma(\tau)$ when the homogenous model (1) holds and the average effect over all groups at the τ quantile of their conditional distribution when the heterogenous model (4) holds.[11] Naturally, it is also possible to modelize the heterogeneity between groups by estimating more flexible linear projections of $Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j)$. For instance, we can interact x_{1ij} with observable characteristics x_{2j}.

Remark 4 (**Least-squares versus quantile regression in the second stage**). As discussed in Remark 3, the projection identifies the average coefficients across groups when those are heterogeneous. It is possible to analyze the inter-group heterogeneity if we impose model (3), restrict v_j to be a scalar, and impose the strict monotonicity of $x'_{1ij}\beta(u, v) + x'_{2j}\gamma(u, v) + \alpha(u, v)$ with respect to v. When we normalize $v_j|x_{1ij}, x_{2j} \sim U(0, 1)$ and $Q(\theta, \alpha(\tau, \theta)|x_{1ij}, x_{2j}) = 0$, we obtain in the exogeneous case, for any $0 < \theta < 1$,

$$Q(\theta, Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j)|x_{1ij}, x_{2j}) = x'_{1ij}\beta(\tau, \theta) + x'_{2j}\gamma(\tau, \theta).$$

All coefficients have two quantile indices: one for the heterogeneity across groups and one for the heterogeneity within groups.[12] These heterogeneous coefficients are identified by a two-step quantile regression: (i) τ quantile regression of y_{ij} on x_{1ij}, (ii) θ quantile regression of the fitted values from the first-stage on x_{1ij} and x_{2j}. Pons (2024) considers a version of this alternative model, which defines different parameters and is outside the scope of this paper. We focus instead on the model defined by equations (1) and (2), which is the same as in Chetverikov et al. (2016) and nests the fixed effects quantile regression model (studied e.g. in Galvao et al., 2020).

2.2 Quantile Minimum Distance Estimators

Motivated by the representation in equation (2), we suggest the following the two-step procedure. In the first step, for each group j and quantile τ, we regress y_{ij} on individual-level variables x_{1ij} and a constant using quantile regression. The intercept of the first stage regression captures both the group effect $\alpha(\tau, v_j)$ and the term $x'_{2j}\gamma(\tau)$ as these vary only between

---

[11]In the endogenous case, we obtain the instrumental variable projection instead of the standard linear projection. For instance, if x_{2j} is an endogenous binary variable and z_{ij} is a binary instrument, we identify the average treatment effects for the compliers at the τ quantile of their conditional distribution.

[12]This is similar to the instrumental variable model in Chesher (2003) and Ma and Koenker (2006), which also contains two quantile indices: one for the selection equation and one for the outcome equation.

groups. In a second step, we regress the fitted values of the first stage on x_{1ij} and x_{2j}, using GMM with instruments z_{ij}.

Formally, the first stage quantile regression solves the following minimization problem for each group and quantile separately:

$$\hat{\beta}_j(\tau) = \left(\hat{\beta}_{0,j}, \hat{\beta}'_{1,j}\right)' = \underset{(b_0,b_1)\in\mathbb{R}^{K_1+1}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_{ij} - b_0 - x'_{1ij}b_1), \tag{5}$$

where $\rho_\tau(x) = (\tau - 1\{x < 0\})x$ for $x \in \mathbb{R}$ is the check function. The true vector of coefficients for group j is given by $\beta_j(\tau) = (\alpha(\tau, v_j) + x'_{2j}\gamma(\tau), \beta(\tau)')'$. When the model does not contain any x_{1ij} variables, quantile regression computes the sample quantiles in each group.

Notation. Throughout the paper, we will use the following notation. Let $\tilde{x}_{ij} = (1, x'_{1ij})'$ and $x_{ij} = (x'_{1ij}, x'_{2j})'$. For each group j we define the following matrices. The $n \times K_1$ matrix of individual-level regressors $X_{1j} = (x_{11j}, x_{12j}, \ldots, x_{1nj})'$, the $n \times K$ matrix containing all regressors $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$ and the $n \times L$ matrix of instruments $Z_j = (z_{1j}, z_{2j}, \ldots, z_{nj})'$. Further, we define two matrices for all observations. The $mn \times K$ matrix of regressors for all groups $X = (X'_1, \ldots, X'_m)'$ and the $mn \times L$ matrix of instruments for all groups as $Z = (Z'_1, \ldots, Z'_m)'$. We let Y be the $mn \times 1$ vector of the response variable. The fitted value for individual i in group j at quantile τ is $\hat{y}_{ij}(\tau) = \hat{\beta}_{0,j}(\tau) + x'_{1ij}\hat{\beta}_{1,j}(\tau)$. We denote the $n \times 1$ column vector of fitted values for group j by $\hat{Y}_j(\tau) = (\hat{y}_{1j}(\tau), \ldots, y_{nj}(\tau))'$, and the $mn \times 1$ vector of fitted values by $\hat{Y}(\tau) = (\hat{Y}'_1(\tau), \ldots, \hat{Y}'_m(\tau))'$.

Remark 5 (Alternative first-stage estimators). The quantile regression estimator proposed by Koenker and Bassett (1978) is not necessarily efficient. Newey and Powell (1990) suggest a semiparametrically efficient weighted estimator of $\beta_j(\tau)$. However, we opt for the unweighted quantile regression estimator due to the challenges associated with estimating the weights and the complicated interpretation of the estimates in cases of misspecification. In our model (1), the variation within groups is assumed to be exogenous. If this assumption were violated, one could apply an instrumental variable (IV) quantile regression (see, e.g., Chernozhukov and Hansen, 2006) in the first stage, followed by the second-stage GMM regression described below.[13] We do not pursue this (computationally intensive) extension in this paper.

The second stage consists in a GMM regression using $\mathbb{E}[g_j(\delta, \tau)] = 0$ as a moment condition, where $g_j(\delta, \tau) = Z'_j(\tilde{X}_j\hat{\beta}_j(\tau) - X_j\delta(\tau))$ and $\delta(\tau) = (\beta(\tau)', \gamma(\tau)')'$ is the K-dimensional vector of coefficients. This moment restriction depends on the first stage and is the sample counterpart of $\mathbb{E}[Z'_j\alpha_j(\tau, v_j)]$. The closed-form expression of the second-stage estimator is

$$\hat{\delta}(\hat{W}, \tau) = \left(X'Z\hat{W}(\tau)Z'X\right)^{-1}X'Z\hat{W}(\tau)Z'\hat{Y}(\tau), \tag{6}$$

where $\hat{W}(\tau)$ is a $L \times L$ symmetric weighting matrix. If $L = K$, the second step estimator in equation (6) simplifies to the IV estimator using Z as an instrument and we can drop the dependence on $\hat{W}(\tau)$.

---

[13]An IV extension of the MD estimator by Galvao and Wang (2015) is suggested in Dai and Jin (2021).

Our two-step estimator is extremely simple to implement; it requires only routines performing quantile regression and GMM estimation, which are already available in many softwares. In addition, we provide general-purpose packages for both R and Stata. Quantile regression, which is computationally more demanding due to the absence of a closed-form solution, is used only in the first stage, where there are fewer observations and a limited number of parameters to estimate. The first stage is also embarrassingly parallelizable, which further increases the computational speed. For this reason, our estimator remains computationally attractive in large datasets with numerous groups. The second stage is a straightforward GMM estimator, which includes OLS and two-stage least squares as special cases. Traditional panel data methods can also be used in the second stage. For instance, in our application, we observe individuals born in a given trimester in a given county. The subscript j defines a county-trimester cell, while the subscript i defines an individual within this cell. In the second stage, we include trimester, county, and state $\times$ year fixed effects to estimate the effect of food stamps on the birth weight distribution.

Remark 6 (**Interpretation as a minimum distance estimator**). Our estimator can be written as a MD estimator, where the second stage imposes restrictions on the first-stage coefficients. For simplicity, in this remark, we consider the case where all the regressors are exogenous and $Z = X$. Define

$$\underset{(K_1+1)\times K}{R_j} = \begin{pmatrix} 0 & x'_{2j} \\ I_{K_1} & 0 \end{pmatrix} \tag{7}$$

such that $\tilde{X}_j R_j = X_j$. It follows that our MD estimator minimizes

$$
\begin{aligned}
\hat{\delta}(\tau) &= \arg\min_{\delta} \sum_{j=1}^{m} (\tilde{X}_j \hat{\beta}_j(\tau) - X_j \delta)'(\tilde{X}_j \hat{\beta}_j(\tau) - X_j \delta) \\
&= \arg\min_{\delta} \sum_{j=1}^{m} (\tilde{X}_j \hat{\beta}_j(\tau) - \tilde{X}_j R_j \delta)'(\tilde{X}_j \hat{\beta}_j(\tau) - \tilde{X}_j R_j \delta) \\
&= \arg\min_{\delta} \sum_{j=1}^{m} (\hat{\beta}_j(\tau) - R_j \delta)' \tilde{X}'_j \tilde{X}_j (\hat{\beta}_j(\tau) - R_j \delta),
\end{aligned}
\tag{8}
$$

which corresponds to the definition of a weighted minimum distance estimator that imposes the linear restrictions $\beta_j(\tau) = R_j \delta(\tau)$ with weights $\tilde{X}'_j \tilde{X}_j$.

Thus, our estimator is an MD estimator. However, it does not correspond to the textbook definition of a "classical minimum distance" estimator.[14] In the classical MD setup, all the sampling variance arises in the first stage: if we know the first stage coefficients, we know the final coefficients. It follows that the efficient weighting matrix $\tilde{W}(\tau)$ is the inverse of the first-stage variance. In our case, the second stage also contributes to the variance due to the presence of the group effects $\alpha(\tau, v_j)$. Even if we know $\beta_j(\tau)$ (for a finite number of groups), we cannot exactly pinpoint $\gamma(\tau)$. The group effects play a role similar to misspecification in

---

[14]See section 14.6 in Wooldridge (2010).

the classical MD, but with our estimator, the resulting bias disappears asymptotically as the number of groups increases. This is the second important difference: the dimension of our first stage estimates increases with the sample size while it is fixed for classical MD estimators.

2.3   Least Squares Minimum Distance Estimators

In this paper, we propose a two-step estimator in which the first stage involves performing quantile regressions within each group. Although this approach might seem unusual and specific to quantile models, we demonstrate in this subsection that this method yields numerically identical results to traditional least squares panel estimators, provided that OLS is applied in the first stage. A more detailed discussion, including formal statements, is provided in Appendix B.1, with proofs available in Appendix B.2.

Consider first a model with group effects and individual-level regressors

$$y_{ij} = x'_{1ij}\beta + \alpha_j + \varepsilon_{ij}. \tag{9}$$

Typically, when estimating this model with fixed effects, most researchers apply the within transformation, which is widely recognized as being equivalent to a dummy variable regression.[15] Yet, there exists a third equivalent method for computing the least squares fixed effects estimator, which involves exploiting the exogenous within-group variation using instrumental variables. Setting $\dot{x}_{1ij} = x_{1ij} - \bar{x}_{1j}$ as an instrument, with $\bar{x}_{1j} = n^{-1}\sum_{i=1}^n x_{1ij}$ in an instrumental variable regression is numerically identical to the least squares fixed effects.

Corollary 1 in Appendix B presents a fourth way to compute the least squares fixed effects estimator, which aligns with our paper's approach. This minimum distance method involves two steps: first, regressing with OLS the dependent variable on x_{1ij} within each group, then using IV to regress the first-stage fitted values on x_{1ij}, with $\dot{x}_{1ij}$ as the instrument. For quantile estimation, this approach offers the most computationally efficient alternative, as it divides the problem into two convex optimization steps rather than a single nonconvex IV quantile regression, and it avoids the challenges of high-dimensional quantile regression.

The two-step procedure is not specific to fixed effects but applies to a wide range of estimators. Proposition 5 in Appendix B shows that the MD least squares estimator is algebraically identical to the one-step GMM regression of y_{ij} on x_{ij} under the condition that for each group j, the matrix of instruments lies in the column space of the matrix of first stage regressors.[16] For example, $\dot{x}_{1ij}$, $\bar{x}_{1j}$, and x_{2j} satisfy the condition.

---

[15]In the context of traditional panel data, we would refer to the j units "individuals" and the i units "time periods". Thus, this transformation corresponds to the time-demeaning applied in the traditional panel data literature, which eliminates time-invariant individual effects.

[16]The intuition is as follows. The fitted values of the first-stage least squares regression can be written as $P_{X_j}Y_j$ where P_{X_j} is the first-stage least squares projection matrix of group j. If the instrument matrix, Z_j, is in the column space of $\tilde{X}_j$, it follows that $P_{X_j}Z_j = Z_j$. Therefore, $Z'\hat{Y} = Z'Y$ and the two GMM regressions are numerically identical.

We extend the model by incorporating group-level regressors, x_{2j}:

$$y_{ij} = x'_{1ij}\beta + x'_{2j}\gamma + \alpha_j + \varepsilon_{ij}, \tag{10}$$

By selecting different instrumental variables for the second-step GMM regression, we can numerically obtain the most common least squares panel data estimators. For example, using the group-averaged variables, $\bar{x}_{1j}$ and x_{2j}, as instruments yields the between estimator. Instrumental variable approaches are also available for random effects estimation. Although FGLS is the most common estimator for the random effects model, Im et al. (1999) show that the overidentified 3SLS estimator, with instruments $\dot{x}_{1ij}$, $\bar{x}_{1j}$, and x_{2j}, is numerically identical to the random effects estimator. Since 3SLS is a special case of a GMM estimator, using the first-stage fitted values as dependent variables does not change the estimates. Alternatively, the random effects estimator can be implemented using the theory of optimal instrument with a just identified 2SLS regression (see Im et al., 1999; Hansen, 2022b). Additionally, the Hausman and Taylor (1981) estimator can be implemented by selecting the following instruments: $\dot{x}_{1ij}$ and the group average of the exogenous regressors. External instruments might also be included.

Notably, clustering standard errors at the group level (or higher) effectively captures the first-stage estimation error, as shown in Proposition 7 in Appendix B. These clustered standard errors are numerically identical to those obtained from a one-step GMM estimator with yij as the dependent variable.

3 Asymptotic Theory

3.1 Preliminaries: Assumptions, Consistency, and Sample Moments

In this section, we state the assumptions and present the asymptotic results. All the proofs are included in Appendix A. For simplicity of notation, in the following, we write $\alpha_j(\tau)$ instead of $\alpha(\tau, v_j)$. We prove weak uniform consistency and weak convergence of the whole quantile regression process for $\tau \in \mathcal{T}$, where $\mathcal{T} \in (0, 1)$ is a compact set of quantile indices of interest. The symbol $\ell^\infty(\mathcal{T})$ denotes the set of component-wise bounded vector valued function of $\mathcal{T}$, $\rightsquigarrow$ denotes weak convergence, and for a random variable h_{ij}, $\mathbb{E}_{i|j}[h_{ij}]$ indicates the expectation over i in group j.

We start by writing the sampling error of $\hat{\delta}(\hat{W}, \tau)$ as a sum of a component arising from the first stage estimation error of $\beta_j(\tau)$ and a component arising from the second stage noise $\alpha_j(\tau)$:

Lemma 1 (Sampling error). *Assume that the model in equation (1) holds, then*

$$\hat{\delta}(\hat{W}, \tau) - \delta(\tau) = \hat{G}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} z_{ij}\left(\tilde{x}'_{ij}(\hat{\beta}_j(\tau) - \beta_j(\tau)) + \alpha_j(\tau)\right),$$

where $\hat{G}(\tau) = \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)$ and $S_{ZX} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} z_{ij}x'_{ij}$.

To keep the notation light, we suppress the dependency of $\hat{G}(\tau)$ on $\hat{W}(\tau)$. We now state assumptions that ensure that both components are well-behaved. For the analysis of the first stage estimator, we rely on results derived in Galvao et al. (2020) and make the assumptions required in their Theorem 2:

Assumption 1 (Sampling). *(i) The processes $\{(y_{ij}, x_{ij}, z_{ij}) : i = 1, \ldots, n\}$ are independent across j. (ii) For each j, the observations $(y_{ij}, x_{ij}, z_{ij})_{i=1,\ldots,n}$ are i.i.d. across i.*

Assumption 2 (Covariates). *(i) For all $j = 1, \ldots, m$ and all $i = 1, \ldots, n$, $\|x_{ij}\| \leq C$ almost surely. (ii) The eigenvalues of $\mathbb{E}_{i|j}[\tilde{x}_{ij}\tilde{x}'_{ij}]$ are bounded away from zero and infinity uniformly across j.*

Assumption 3 (Conditional distribution). *The conditional distribution $F_{y_{ij}|x_{1ij},v_j}(y|x,v)$ is twice differentiable w.r.t. y, with the corresponding derivatives $f_{y_{ij}|x_{1ij},v_j}(y|x,v)$ and $f'_{y_{ij}|x_{1ij},v_j}(y|x,v)$. Further, assume that*

$$f_{max} = \sup_j \sup_{y \in \mathbb{R}, x \in \mathcal{X}} |f_{y_{ij}|x_{1ij},v_j}(y|x,v)| < \infty$$

and

$$\bar{f}' = \sup_j \sup_{y \in \mathbb{R}, x \in \mathcal{X}} |f'_{y_{ij}|x_{1ij},v_j}(y|x,v)| < \infty.$$

where $\mathcal{X}$ is the support of x_{1ij}

Assumption 4 (Bounded density). *There exists a constant $f_{min} < f_{max}$ such that*

$$0 < f_{min} \leq \inf_j \inf_{\tau \in \mathcal{T}} \inf_{x \in \mathcal{X}} f_{y_{ij}|x_{1ij},v_j}(Q(\tau, y_{ij}|x,v)|x,v).$$

These are quite standard assumptions in the quantile regression literature. In Assumption 1, we assume that the processes are independent across j; this assumption can also be relaxed by allowing for clustering between groups. We also assume that the observations are i.i.d. within groups, but this can be relaxed at the cost of a more complex notation by applying Theorem 4 in Galvao et al. (2020), which requires only stationarity and β-mixing. The estimator of the asymptotic variance that we suggest below is consistent in both cases. Assumption 2 requires that the regressors are bounded and that $\mathbb{E}_{i|j}[\tilde{x}_{ij}\tilde{x}'_{ij}]$ is invertible. Assumptions 3 and 4 impose smoothness and boundedness of the conditional distribution, the density, and its derivatives.

For the second stage GMM regression we impose the following assumptions:

Assumption 5 (Instruments). *(i) For all $j = 1, \ldots, m$ and all $i = 1, \ldots, n$, $\|z_{ij}\| \leq C$ a.s. (ii) For all $j = 1, \ldots, m$ and all $i = 1, \ldots, n$, $\mathbb{E}[z_{ij}\alpha_j(\tau)] = 0$. (iii) For all $j = 1, \ldots, m$ and all $i = 1, \ldots, n$, y_{ij} is independent of z_{ij} conditional on (x_{ij}, v_j). (iv) As $m \to \infty$, $m^{-1}\sum_{j=1}^m \mathbb{E}_{i|j}[z_{ij}x'_{ij}] \to \Sigma_{ZX}$ where the singular values of Σ_{ZX} are bounded from below and from above.*

Assumption 6 (Group effects).
(i) For all $j = 1, \ldots, m$, $\mathbb{E}\left[\sup_{\tau \in \mathcal{T}} |\alpha_j(\tau)|^{4+\varepsilon_C}\right] \leq C$ for $\varepsilon_C > 0$. (ii) For some (matrix-valued) function $\Omega_2 : \mathcal{T} \times \mathcal{T} \to \mathbb{R}^{L \times L}$, $m^{-1} \sum_{j=1}^m \mathbb{E}_{i|j}[\alpha_j(\tau_1)\alpha_j(\tau_2)z_{ij}z'_{ij}] \underset{p}{\to} \Omega_2(\tau_1, \tau_2)$ uniformly over $\tau_1, \tau_2 \in \mathcal{T}$. (iii) For all $\tau_1, \tau_2 \in \mathcal{T}$, $|\alpha_j(\tau_2) - \alpha_j(\tau_1)| \leq C|\tau_2 - \tau_1|$.

These assumptions are the same as in Chetverikov et al. (2016). For the instrumental variables, we assume that (i) they are bounded, (ii) they are not correlated with the group effect (exclusion restriction), (iii) they do not affect the first stage estimation (this is often satisfied by construction, e.g. when the instruments do not vary within individuals or are a linear transformation of the first stage regressors), and (iv) they satisfy the relevance conditions. For the group effects we assume that they have a finite fourth moment, and the average variance of $z_{ij}\alpha_j(\tau)$ converges to a well-defined matrix.

Since the unobserved heterogeneity $\alpha_j(\tau)$ is group-specific, we require that the number of groups m diverges to infinity. The first stage quantile regression estimator is a nonlinear estimator that is potentially biased in finite samples. Hence, for consistency, the number of observations per group, n, must also diverge to infinity. Galvao et al. (2020) show that the bias is approximately of order $1/n$. For unbiased asymptotic normality, we need the bias to shrink faster than the standard deviation of the estimator. We will see that some elements of $\hat{\delta}(\hat{W}, \tau)$ converge at the $\sqrt{m}$ rate such that we need that n goes to infinity more quickly than $\sqrt{m}$. On the other hand, other elements converge at the $\sqrt{mn}$ rate so that n must go to infinity more quickly than m. We state these three different relative growth rates in the following assumption:

Assumption 7 (Growth rates). *As $m \to \infty$, we have*

(a) $\frac{\log m}{n} \to 0$,

(b) $\frac{\sqrt{m}\log n}{n} \to 0$,

(c) $\frac{m(\log n)^2}{n} \to 0$.

Finally, we assume that the estimated weighting matrix uniformly converges to a strictly positive definite matrix that is continuous in the quantile index.

Assumption 8 (Full-rank weighting matrix). *Uniformly in $\tau \in \mathcal{T}$, $\hat{W}(\tau) \underset{p}{\to} W(\tau)$ where $W(\tau)$ is strictly positive definite and, for all $\tau_1, \tau_2 \in \mathcal{T}$, $\|W(\tau_2) - W(\tau_1)\| \leq C|\tau_2 - \tau_1|$.*

In our first result, we establish uniform consistency of our estimator.

Theorem 1 (Uniform consistency). *Let the model in equation (1), Assumptions 1-6, 7(a), and 8 hold. Then,*

$$\sup_{\tau \in \mathcal{T}} \|\hat{\delta}(\tau) - \delta(\tau)\| = o_p(1).$$

We now study the asymptotic distribution of our estimator. In Lemma 1, we see that the sample moment condition is the sum of two terms. It is useful to consider them separately:

$$\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} z_{ij} \tilde{x}_{ij}' \left(\hat{\beta}_j(\tau) - \beta_j(\tau) \right) \tag{11}$$

$$\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} z_{ij} \alpha_j(\tau) \tag{12}$$

such that total moment condition is the sum of both components: $\bar{g}_{mn}(\hat{\delta}, \tau) = \bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) + \bar{g}_{mn}^{(2)}(\hat{\delta}, \tau) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} g_{ij}(\hat{\delta}, \tau)$. Lemma 2 establishes joint asymptotic normality for the entire moment condition processes.

Lemma 2 (**Asymptotic distribution of the sample moments**). *Let the model in equation (1), and Assumptions 1-6 hold.*

(i) Under Assumption 7(c), as $m \to \infty$,

$$\sqrt{mn}\bar{g}_{mn}^{(1)}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}_1(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{13}$$

where $\mathbb{Z}_1(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega_1(\tau, \tau') = \mathbb{E}_j \left[\Sigma_{ZXj} V_j(\tau, \tau') \Sigma_{ZXj}' \right]$ with $\Sigma_{ZXj} = \mathbb{E}_{i|j}[z_{ij}\tilde{x}_{ij}']$ and $V_j(\tau, \tau')$ is the asymptotic variance-covariance matrix of $\hat{\beta}_j(\tau)$ and $\hat{\beta}_j(\tau')$:

$$V_j(\tau, \tau') = \mathbb{E}_{i|j}[f_{y|x}(Q_{y|x,\nu_j}(\tau|\tilde{x}_{ij})|\tilde{x}_{ij})\tilde{x}_{ij}\tilde{x}_{ij}']^{-1}(\min(\tau, \tau') - \tau\tau')\mathbb{E}_{i|j}[\tilde{x}_{ij}\tilde{x}_{ij}']$$
$$\times \mathbb{E}_{i|j}[f_{y|x}(Q_{y|x,\nu_j}(\tau'|\tilde{x}_{ij})|\tilde{x}_{ij})\tilde{x}_{ij}\tilde{x}_{ij}']^{-1}$$

(ii) Under Assumption 7(b), As $m \to \infty$,

$$\sqrt{m}\bar{g}_{mn}^{(2)}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}_2(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{14}$$

where $\mathbb{Z}_2(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega_2(\tau, \tau')$, which is defined in Assumption 6(ii).

(iii) Under Assumption 7(c), as $m \to \infty$, $\sup_{\tau, \tau' \in \mathcal{T}} \left\| \text{Cov} \left(\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau), \bar{g}_{mn}^{(2)}(\hat{\delta}, \tau') \right) \right\| = o_p \left(\frac{1}{\sqrt{mn}} \right).$

$\bar{g}_{mn}^{(1)}(\hat{\delta}, \cdot)$ reflects the estimation error that arises in the first stage quantile regression estimation. Since the first-stage regressors vary within groups, the relevant number of observations is mn and, correspondingly, the variance is proportional to $1/(mn)$. On the other hand, since the expected bias of the first-stage quantile regression is of order $1/n$, for asymptotic unbiasedness, we must require that n goes to infinity slightly faster than m. In the proof we build on results derived in Volgushev et al. (2019) and in Galvao et al. (2020). $\bar{g}_{mn}^{(2)}(\hat{\delta}, \cdot)$ reflects the estimation error due to the randomness in $\alpha_j(\tau)$. This moment can also be interpreted as the moment that would be relevant if we knew $\beta_j(\tau)$. Since $\alpha_j(\tau)$ varies only between groups, the relevant number of observations here is m and, accordingly, the variance of this moment converges at the slower rate of $1/m$. For asymptotic unbiasedness, we need only the weaker condition 7(b), which requires that n goes to infinity slightly faster than $\sqrt{m}$.

15

3.2 Asymptotic Distribution when the Degree of Heterogeneity is Known

The sample moment condition $\bar{g}_{mn}(\hat{\delta}, \tau)$ is thus the sum of two components that converge to zero at different rates. The asymptotic distribution of the estimator is dominated by the component with the slower rate of convergence, $\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)$, except its variance is zero, which is the case if $\text{Var}(\alpha_j(\tau)) = 0$ or if $\bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} = 0$. Since the degree of group-level heterogeneity affects this variance, it is useful to consider three cases: strong, no, and weak heterogeneity. In this subsection, we derive the asymptotic distribution of our estimator when the degree of heterogeneity is known. We suggest adaptive estimation and inference procedures in the next subsection.

Case 1: Strong group-level heterogeneity. We start with the case of strong heterogeneity that we define to be $\text{Var}(\alpha_j(\tau)) > \varepsilon > 0$ uniformly in τ. The variance of an element of the vector $\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau) = \frac{1}{mn} \sum_{j=1}^{m} \bar{z}_j \alpha_j(\tau)$ equals to zero when the corresponding instrument satisfies $\bar{z}_j = 0$ for all j. For this reason, we distinguish between two sorts of instruments: L_1 instruments in z_{1ij} satisfy $\bar{z}_{1j} = n^{-1} \sum_{i=1}^{n} z_{1ij} = 0$ for all j, while L_2 instruments in z_{2ij} satisfy $\bar{z}_{2j} \neq 0$ at least for some groups j.[17] We order the instruments such that $z_{ij} = (z'_{1ij}, z'_{2ij})'$. It follows that

$$\bar{g}_{mn}(\hat{\delta}, \tau) = \begin{pmatrix} \bar{g}_{mn,1}(\hat{\delta}, \tau) \\ \bar{g}_{mn,2}(\hat{\delta}, \tau) \end{pmatrix} = \begin{pmatrix} \bar{g}_{mn,1}^{(1)}(\hat{\delta}, \tau) \\ \bar{g}_{mn,2}^{(1)}(\hat{\delta}, \tau) + \bar{g}_{mn,2}^{(2)}(\hat{\delta}, \tau) \end{pmatrix}$$

where $g_{mn,1}(\hat{\delta}, \tau)$ is a $L_1 \times 1$ vector and $g_{mn,2}(\hat{\delta}, \tau)$ is a $L_2 \times 1$ vector. Thus, in the case of strong heterogeneity, some moments converge at the fast rate $\sqrt{mn}$ while others converge at the slow rate $\sqrt{m}$.

We order the coefficients in $\delta(\tau)$ (and the corresponding regressors in x_{ij}) such that the first M_1 elements are identified using only the L_1 fast moments while the remaining M_2 elements requires the L_2 slow moments for identification. We denote by $\delta_1(\tau)$ the former and by $\delta_2(\tau)$ the latter coefficients. Formally, we partition Σ_{ZX} such that it is block lower triangular:

$$\Sigma_{ZX} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{15}$$

where Σ_{11} is a full column rank $L_1 \times M_1$ matrix, Σ_{12} is $L_1 \times M_2$, Σ_{21} is $L_2 \times M_1$ and Σ_{22} is $L_2 \times M_2$. Assumption 5(iv) implies that Σ_{22} has also full cokumn rank. Note that L_1, L_2, M_1, and M_2 can be equal to zero such that equation (15) is a definition and not an assumption. Accordingly, the adaptive estimation and testing procedures suggested in Section 3.3 do not require the user to classify the instruments or the regressors.[18]

---

[17]Note that all the instruments that vary only within groups can be normalized to have mean zero. For instance, we can identify the effect of the individual-level variable x_{1ij} by using the instrument $\dot{x}_{1ij}$, which has a zero mean in all groups.

[18]In this subsection, we discuss the asymptotic distribution of $\hat{\delta}(\tau)$ when we know the degree of heterogeneity and the types of the instruments and regressors. However, our estimation and testing procedures are shown to be adaptive in the next subsection.

In the exactly identified case, our estimator simplifies to the instrumental variable estimator such that

$$\hat{G}(\tau) = S_{ZX}^{-1} \underset{p}{\to} \Sigma_{ZX}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22}^{-1} \end{pmatrix}$$

and the first-order asymptotic distributions can be written as

$$\sqrt{mn}(\hat{\delta}_1(\tau) - \delta_1(\tau)) \underset{d}{\to} \Sigma_{11}^{-1}\sqrt{mn}\bar{g}_{mn,1}^{(1)}(\hat{\delta}, \tau)$$

$$\sqrt{m}(\hat{\delta}_2(\tau) - \delta_2(\tau)) \underset{d}{\to} \Sigma_{22}^{-1}\sqrt{m}\bar{g}_{mn,2}^{(2)}(\hat{\delta}, \tau)$$

$\hat{\delta}_1(\tau)$ is only a function of $\bar{g}_{mn}^{(1)}$ and converges thus at the $\sqrt{mn}$ rate. $\hat{\delta}_1(\tau)$ depends on both $\bar{g}_{mn}^{(1)}$ and $\bar{g}_{mn}^{(2)}$ but its first-order asymptotic distribution is dominated by the slower $\bar{g}_{mn}^{(2)}$ term.

Example 1. We can illustrate this notation with a simple example. Consider the case where there is one individual-level variable x_{1ij}, one group-level variable $\tilde{x}_{2j}$, and a constant. As instrumental variables, we use $(\dot{x}_{1ij}, \tilde{x}_{2j}, 1)$. This corresponds to using the fixed effects estimator for the coefficient on x_{1ij} and the between estimator for the other two coefficients. In this case, only the first instrument has mean zero in all groups such that $L_1 = 1$ and $L_2 = 2$. By construction, $\dot{x}_{1ij}$ is uncorrelated with $\tilde{x}_{2j}$ and with the constant such that Σ_{ZX} is block lower diagonal as defined in equation (15) with $M_1 = 1$ and $M_2 = 2$. The coefficient on x_{1ij} converges at the fast rate because it is not affected by the group-level effects $\alpha_j(\tau)$. On the other hand, the intercept and the coefficient on the group-level variable $\tilde{x}_{2j}$ converge only at the slow rate of convergence $\sqrt{m}$ because they are affected by the group effects. In a many application, we see that $M_1 = K_1$, but this is not always true. For instance, in the previous example, all coefficients would converge at the slow rate if we used $(x_{1ij}, \tilde{x}_{2j}, 1)$ as instrumental variables such that L_1 and M_1 would be equal to zero.

In the case of an overidentified model, employing a full-rank weighting matrix $W(\tau)$ can lead to the contamination of the entire parameter vector $\hat{\delta}(\tau)$ by the slower-converging moments. This is because, in such a case, $\hat{G}(\tau)$ will not have a block-lower-triangular structure, even if Σ_{ZX} does. In Example 2 below, this phenomenon occurs in the 2SLS estimator of $\delta_1(\tau)$, which converges at the $\sqrt{m}$ rate, while the exactly identified IV estimator and the efficient GMM estimator converge at the $\sqrt{mn}$ rate. To avoid contamination, we must give asymptotically infinitely higher weights to fast moments than slow ones. The following assumption imposes this critical condition on the weighting matrix.[19]

Assumption 8′ (Heterogeneous weighting matrix). *Uniformly in $\tau \in \mathcal{T}$,*

$$\hat{W}(\tau) = \underbrace{\begin{pmatrix} W_{11}(\tau) & a_n(\tau)W_{12}(\tau) \\ a_n(\tau)W_{21}(\tau) & a_n(\tau)W_{22}(\tau) \end{pmatrix}}_{W_{mn}(\tau)} + \begin{pmatrix} o_p(1) & o_p\left(\sqrt{a_n(\tau)}\right) \\ o_p\left(\sqrt{a_n(\tau)}\right) & o_p(a_n(\tau)) \end{pmatrix}$$

---

[19]The intuition behind the sequence $a_n(\tau)$ will be elucidated in the following subsection, where we delve into a comprehensive discussion of our efficient GMM estimator.

where $a_n(\tau) = \frac{1}{1+\text{Var}(\alpha_j(\tau))n}$. $W_{11}(\tau)$ and $W_{22}(\tau)$ are respectively a $L_1 \times L_1$ and a $L_2 \times L_2$ full rank matrix. For all $l_1, l_2 \in \{1, 2\}$ and $\tau_1, \tau_2 \in \mathcal{T}$, $\|W_{l_1 l_2}(\tau_2) - W_{l_1 l_2}(\tau_1)\| \leq C|\tau_2 - \tau_1|$.

Example 2. Consider an extension of Example 1 where the vector of instrumental variable is now $(\dot{x}_{1ij}, \bar{x}_{1j}, x_{2j}, 1)$. This model is overidentified with $L_1 = 1$, $L_2 = 3$, and $K = 3$. When we impose Assumption 8, $W(\tau)$ is a full rank weighting matrix, and the estimator converges only at the $\sqrt{m}$ rate. We can obtain a faster-converging estimator by giving asymptotically infinitely more weight to $\dot{x}_{ij}$ than to the other instruments. When we impose Assumption 8′, the coefficient on x_{1ij} is estimated at the $\sqrt{nm}$ rate. We formalize these results in Theorem 2 below.

Case 2: no group-level heterogeneity. We now consider the case where there is no heterogeneity across groups, i.e., when $\alpha_j(\tau) = 0$ uniformly in j and τ. When this occurs, $\bar{g}_{mn}^{(2)}(\hat{\delta}, \cdot) = 0$ such that all the variance arises from the first stage moment, which can be estimated at the fast rate $\sqrt{mn}$. This case corresponds to the textbook definition of a classical minimum distance estimator. It follows that all the coefficients converge at the fast $\sqrt{mn}$ rate, and the asymptotic distribution of $\hat{\delta}(\tau)$ can be derived in a straightforward way. By comparing Case 1 and 2, we notice that the limiting distribution and even the rate of convergence of the M_2 coefficients $\hat{\delta}_2(\tau)$ differ. Their asymptotic distribution is discontinuous at 0 in the variance of $\alpha_j(\tau)$, and in many applications, we are not sure whether there is group-level heterogeneity. For this reason, we discuss adaptive estimation and inference in the next subsection.

Remark 7 (Related literature). Chamberlain (1994) considers a setup with a finite number of design points (groups in our terminology), exogenous group-level variables, and no individual-level variables. His correctly specified case corresponds to our Case 2 (no heterogeneity). Accordingly, we obtain the same asymptotic distribution in this special case.[20] Interestingly, Chamberlain (1994) also considers the misspecified case. If we interpret his misspecification errors as group effects, then we obtain the same asymptotic variance in our Case 1 as he obtains with misspecification. However, in his setup, the estimator converges to pseudo-true values that absorb the bias, while we allow the number of groups to go to infinity to avoid bias.

Chetverikov et al. (2016) do not consider Case 2 (no heterogeneity) in their theoretical results even if, interestingly, it corresponds to one of their data-generating processes in their simulations. Notice that, in Case 2, their matrix-valued function $J(u_1, u_2)$ defined in their Assumption 6(ii) is uniformly equal to zero. This implies that their asymptotic covariance function $\mathcal{C}(u_1, u_2)$ defined in their Theorem 1 is also uniformly equal to 0. This degenerate asymptotic distribution indicates that the convergence rate is faster than $\sqrt{m}$ in this case.

Case 3: weak group-level heterogeneity. In Case 1, the second-stage variance dominates, while in Case 2, the first-stage variance dominates the asymptotic distribution of $\hat{\delta}_2(\tau)$.

---

[20]In the absence of group-level heterogeneity, the asymptotic distribution of $\hat{\delta}(\tau)$ stated in part (ii) of Theorem 2 is also valid when the number of groups is fixed.

It is interesting to consider the intermediate case when group-level heterogeneity is present but vanishes exactly at the right rate such that both components of the variance matter asymptotically. This should provide a good approximation for the applications where the first-stage and the second-stage variances are similar. Formally, we assume that $\Omega_2(\tau_1, \tau_2)$, the covariance function of $\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)$ defined in Assumption 6(ii), converges to zero at the n rate:

$$\Omega_2(\tau_1, \tau_2) = n^{-1}\bar{\Omega}_2(\tau_1, \tau_2).$$

Under this assumption, $\bar{g}_{mn}^{(1)}(\hat{\delta}, \cdot)$ and $\bar{g}_{mn}^{(2)}(\hat{\delta}, \cdot)$ converge at the same rate. Lemma 2 implies then

$$\sqrt{mn}\bar{g}_{mn}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{16}$$

where $\mathbb{Z}(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega_1(\tau, \tau') + \bar{\Omega}_2(\tau, \tau')$.

Theorem 2 formally states the results for these three cases. Parts (i) provides the asymptotic distribution for the fast and slow coefficients when there is strong heterogeneity, part (ii) in the absence of heterogeneity, and part (iii) when there is weak heterogeneity.

Theorem 2 (Asymptotic distribution when the degree of heterogeneity is known). *Let Assumptions 1-6 hold.*

(i) Case 1 (strong heterogeneity): $\text{Var}(\alpha_j(\tau)) > \varepsilon > 0$ *uniformly in* τ *and Assumption 8' holds.*

(a) In addition, let Assumption 7(c) hold. Then,

$$\sqrt{mn}(\hat{\delta}_1(\hat{W}(\cdot), \cdot) - \delta_1(\cdot)) \rightsquigarrow G_{11}(\cdot)\mathbb{Z}_{11}(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{17}$$

where $G_{11}(\tau) = (\Sigma_{11}'W_{11}(\tau)\Sigma_{11})^{-1}\Sigma_{11}'W_{11}(\tau)$ *and* $\mathbb{Z}_{11}(\cdot)$ *is the Gaussian process consisting of the first* L_1 *elements of* $\mathbb{Z}_1(\cdot)$ *defined in Lemma 2(i).*

(b) In addition, let Assumption 7(b) hold. Then,

$$\sqrt{m}(\hat{\delta}_2(\hat{W}(\cdot), \cdot) - \delta_2(\cdot)) \rightsquigarrow G_{22}(\cdot)\mathbb{Z}_{22}(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{18}$$

where $G_{22}(\tau) = (\Sigma_{22}'W_{22}(\tau)\Sigma_{22})^{-1}\Sigma_{22}'W_{22}(\tau)$ *and* $\mathbb{Z}_{22}(\cdot)$ *is the Gaussian process consisting of the last* L_2 *elements of* $\mathbb{Z}_2(\cdot)$ *defined in Lemma 2(ii).*

(ii) Case 2 (no heterogeneity): $\alpha_j(\tau) = 0$ *uniformly in* j *and* τ. *In addition, Assumption 7(c) and 8 hold. Then,*

$$\sqrt{mn}(\hat{\delta}(\hat{W}(\cdot), \cdot) - \delta(\cdot)) \rightsquigarrow G(\cdot)\mathbb{Z}_1(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{19}$$

where $G(\tau) = (\Sigma_{ZX}'W(\tau)\Sigma_{ZX})^{-1}\Sigma_{ZX}'W(\tau)$ *and* $\mathbb{Z}_1(\cdot)$ *is defined in Lemma 2(i).*

(iii) Case 3 (weak heterogeneity): $\Omega_2(\tau, \tau') = n^{-1}\bar{\Omega}_2(\tau, \tau')$. *Assumption 8 and 7(c) hold. Then,*

$$\sqrt{mn}(\hat{\delta}(\hat{W}(\cdot), \cdot) - \delta(\cdot)) \rightsquigarrow G(\cdot)\mathbb{Z}(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{20}$$

where $G(\tau) = (\Sigma_{ZX}'W(\tau)\Sigma_{ZX})^{-1}\Sigma_{ZX}'W(\tau)$ *and* $\mathbb{Z}(\cdot)$ *is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function* $\Omega_1(\tau, \tau') + \bar{\Omega}_2(\tau, \tau')$.

These asymptotic results are useful for understanding the mechanics behind our estimator, but they have several weaknesses. First, the asymptotic distribution of $\hat{\delta}_1(W, \tau)$ in part (i)-(a) is only a function of the fast instruments. The same asymptotic distribution can be obtained by ignoring the slow instruments. Consider Example 2, including $\bar{x}_{1j}$ as an instrument does not reduce the asymptotic variance of $\hat{\delta}_1(W, \tau)$ even if this instrument is valid and the between-group variation in x_{1ij} is non-negligible. In other words, the random effects estimator is asymptotically equivalent to the fixed effects estimator.[21] In some applications, the random effects estimator is appreciably more precise such that we would like to exploit the between-group variation efficiently. Second, the asymptotic distribution and the convergence rate of the slow coefficients $\hat{\delta}_2(W, \tau)$ are different, depending on whether there is group-level heterogeneity or not. Thus, performing inference based on Theorem 2 requires knowing which case is relevant for the specific application. In other words, inference based directly on this result is not adaptive. Third, in Case 1, the variance coming from the first stage estimation does not appear in the asymptotic distribution of $\hat{\delta}_2(W, \tau)$ because it converges to zero at a quicker rate. Consequently, inference may have poor properties. We solve these issues in the following subsection by suggesting an efficient estimator and adaptive inference that are both valid in all three cases and more generally uniformly valid in the variance of $\alpha_j(\tau)$.

3.3   Adaptive Estimation and Inference

In this subsection, we derive asymptotic results that are uniformly valid in the variance of $\alpha_j(\tau)$. In other words, we allow $\text{Var}(\alpha_j(\tau))$ (and thus $\Omega_2(\tau, \tau)$) to be exactly zero, bounded away from zero, or a sequence that converges to zero at an arbitrary rate. This approach nests all three pointwise cases described in the previous section. The sequence $a_n(\tau) = \frac{1}{1 + \text{Var}(\alpha_j(\tau))n}$ defined in Assumption 8′ plays a central role because it is proportional to the rate of convergence of the slow moments relative to the fast moments $\| \text{Var}(\bar{g}_{mn,2}(\delta, \tau)) \|^{-1} \| \text{Var}(\bar{g}_{mn,1}(\delta, \tau) \| = O_p(a_n(\tau))$. It is equal to 1 without heterogeneity, converges to 0 at the n^{-1} rate with strong heterogeneity, and is always bounded between 0 and 1. All matrices that are functions of $a_n(\tau)$ may also depend on the sample size. When necessary, we make the notation explicit with the subscript mn. For example, let

$$G_{mn}(\tau) = \left(\Sigma'_{ZX} W_{mn}(\tau) \Sigma_{ZX} \right)^{-1} \Sigma'_{ZX} W_{mn}(\tau), \tag{21}$$

where $W_{mn}(\tau)$ is defined in Assumption 8′.

Similarly to Fernández-Val et al. (2022), we define the covariance kernel of the limiting process of $\hat{\delta}(\cdot) - \delta(\cdot)$. For a given integer $T > 0$, let $\mathcal{T}_T = (\tau_1, \ldots, \tau_T)$ be an arbitrary T-dimensional vector on $\otimes_{t=1}^T \mathcal{T}$. Let

$$\Sigma_{mn}(\tau, \tau') = G_{mn}(\tau) \left(\frac{\Omega_1(\tau, \tau')}{mn} + \frac{\Omega_2(\tau, \tau')}{m} \right) G_{mn}(\tau')'$$

---

[21]This is not specific to quantile models and also affects least squares models with large n (see Ahn and Moon, 2014).

and $\Sigma_{mn}(\tau) = \Sigma_{mn}(\tau, \tau)$. The covariance kernel is now given by the limit of the elements of the following $(KT) \times (KT)$ matrix

$$H_{mn} = \left(H_{mn}(\tau, \tau') \right)_{T \times T},$$

where

$$H_{mn}(\tau, \tau') = \text{diag}(\Sigma_{mn}(\tau))^{-1/2} \Sigma_{mn}(\tau, \tau') \, \text{diag}(\Sigma_{mn}(\tau'))^{-1/2}.$$

Assumption 9 (Covariance kernel). *For any integer $T > 0$ and any T-dimensional vector $\mathcal{T}_T = (\tau_1, \ldots, \tau_T)$ on $\otimes_{t=1}^T \mathcal{T}$, there is a $(KT) \times (KT)$ matrix H, such that, almost surely,*

$$\lim_{m,n \to \infty} H_{mn} = H.$$

In addition, there is $c_{\mathcal{T}_T} > 0$ such that for the smallest eigenvalue we have

$$\lambda_{\min}(H) > c_{\mathcal{T}_T}.$$

We can now state the asymptotic distribution of our estimator that is uniformly valid in $\text{Var}(\alpha_j(\tau))$.

Theorem 3 (Adaptive asymptotic distribution). *Assumptions 1-7(c) and 9 hold. In addition, either Assumption 8 or Assumption 8' holds. Then,*

$$\text{diag}(\Sigma_{mn}(\cdot))^{-1/2}(\hat{\delta}(\cdot) - \delta(\cdot)) \rightsquigarrow \mathbb{G}(\cdot)$$

where $\mathbb{G}$ is a centered Gaussian process with a covariance function $H(\tau_k, \tau_l)$ as the (k, l) submatrix of H.

Theorem 3 allows for both types of weighting matrices: the asymptotically full-ranked weighting matrix of Assumption 8 or the heterogeneous weighting matrix of Assumption 8'. The first case covers the exactly identified case and 2SLS. As we show in Proposition 3 below, our estimated efficient weighting matrix satisfies the second condition.

In order to use these results for inference, we must provide a consistent estimator of the asymptotic variance $\Sigma_{mn}(\tau, \tau')$. The crucial ingredient is the asymptotic variance of the sample moments that we denote by

$$\Omega_{mn}(\tau, \tau') = \Omega_1(\tau, \tau')/n + \Omega_2(\tau, \tau')$$

Remember that $\Omega_1(\tau, \tau')$ arises from the first stage estimation and $\Omega_2(\tau, \tau')$ from the presence of the group-level heterogeneity $\alpha_j(\cdot)$. The difficulty resides in that the leading term may arise from the first-stage or second-stage estimation depending on the variance of $\alpha_j(\cdot)$ and the type of instrument (L_1 or L_2). This expression may suggest that we must estimate these two components separately. However, and perhaps surprisingly, we find that we can estimate the leading term of $\Omega_{mn}(\tau, \tau')$ adaptively with a traditional cluster-robust estimator of the variance:

$$\hat{\Omega}(\tau, \tau') = \frac{1}{m} \sum_{j=1}^m \left\{ \left(\frac{1}{n} \sum_{i=1}^n z_{ij} \hat{u}_{ij}(\tau) \right) \left(\frac{1}{n} \sum_{i=1}^n z_{ij} \hat{u}_{ij}(\tau') \right)' \right\} \tag{22}$$

where $\hat{u}_{ij}(\tau)$ are the second-stage residuals.

Proposition 1 (Properties of $\hat{\Omega}(\tau, \tau')$). *Let assumptions 1-6 and 7(c) hold. Further, assume that as $m \to \infty$, for each $l, l' \in \{1, \ldots, L\}$ and uniformly in $\tau, \tau' \in \mathcal{T}^2$,*

$$m^{-1} \sum_{j=1}^{m} \mathbb{E}\left[\left(\bar{z}_{jl} \bar{z}_{jl'} \alpha_j(\tau) \alpha_j(\tau') - \Omega_{mn,2ll'}(\tau, \tau') \right)^2 \right] \to C_{l,l'}(\tau, \tau') < \infty \tag{23}$$

where $C_{l,l'}(\tau, \tau')$ is continuous in τ and τ'. The estimator used to compute $\hat{u}_{ij}(\tau)$ satisfies (i) $\hat{\beta}(\tau) - \beta(\tau) = O_p(1/\sqrt{mn})$, (ii) $\hat{\gamma}(\tau) - \gamma(\tau) = O_p(1/\sqrt{mn}) + O_p(\sqrt{\mathrm{Var}(\alpha_j(\tau))}/\sqrt{m})$ uniformly in τ. Then, for any ll' entry of the $\hat{\Omega}(\tau, \tau')$ matrix with $l, l' \in \{1, \ldots, L\}$ we have uniformly in $\tau, \tau' \in \mathcal{T}^2$

$$\hat{\Omega}_{ll'}(\tau, \tau') = \Omega_{mn,ll'}(\tau, \tau') + o_p\left(\sqrt{\Omega_{mn,ll}(\tau) \Omega_{mn,l'l'}(\tau')} \right).$$

Beyond the numbered assumptions, two additional conditions are necessary. First, equation (23) requires the average variance of $\bar{z}_j \alpha_j(\tau)$ to converge to a constant. This holds automatically when we sample identically distributed groups but is required because we allow for heterogeneity across groups. Second, we require that the coefficients on the individual-level variables are estimated at the $1/\sqrt{mn}$ rate and the coefficients on the group-level variable at the $1/\sqrt{mn} + \sqrt{\mathrm{Var}(\alpha_j(\tau))}/\sqrt{m}$ rate. This corresponds to assuming that $M_1 = K_1$ so that $\beta = \delta_1$. This assumption ensures that estimation errors in slow coefficients do not contaminate the variance estimate of fast ones. This condition can always be achieved, for instance, by using $\dot{x}_{1ij}$ as an instrument for x_{1ij} and a weighting matrix that satisfies Assumption 8'. We show below that our efficient weighting matrix satisfies this restriction.[22]

The suggested estimator of the variance is straightforward to implement because it does not require estimating directly the variance of the first-stage quantile regression coefficients. This is a noteworthy advantage because the variance of these coefficients depends on the conditional density of the outcome given the covariates, which can be difficult to estimate. To understand this surprising result, consider the case where $\alpha_j(\tau) = 0$ for all groups. This implies that the entire variance comes from the first-stage estimation. The clustered estimator of the variance can be interpreted as a subsampling estimator where the groups represent the subsamples. The variance of the coefficients across groups provides information about the precision of the first-stage estimates. Similar results for the cross-sectional bootstrap are provided in Liao and Yang (2018), Lu and Su (2022), and Fernández-Val et al. (2022).

Proposition 1 shows we can consistently estimate the diagonal elements of $\Omega_{mn}(\tau)$ in the sense that the error that we make vanishes more quickly than the true variance. On the other hand, if two sample moments vanish at different rates, we might not be able to estimate their covariance consistently. However, this does not preclude testing hypotheses about combinations of coefficients because the error that we make when we estimate the covariance vanishes faster

---

[22]The only exception in this paper is the between estimator from Section 5. In that case, all the coefficients are estimated at the slow rate $1/\sqrt{mn} + \sqrt{\mathrm{Var}(\alpha_j(\tau))}/\sqrt{m}$. Proposition 1' in the Appendix demonstrates the consistency of $\hat{\Omega}(\tau, \tau')$ in this case as well.

than the variance of the slowest coefficient. We formalize this result in Proposition 2 below for a single linear null hypothesis.[23]

Define the $TK \times 1$ vector $\delta(\tau_1, \ldots, \tau_T) = (\delta(\tau_1)', \ldots, \delta(\tau_T)')'$ whose covariance matrix is given by the elements of the following $(KT) \times (KT)$ matrix:

$$\Sigma_{mn} = \left(\Sigma_{mn}(\tau, \tau') \right)_{T \times T}.$$

For each $\tau, \tau' \in \mathcal{T} \times \mathcal{T}$, $\Sigma_{mn}(\tau, \tau')$ is estimated by

$$\hat{\Sigma}(\tau, \tau') = \hat{G}(\tau) \frac{\hat{\Omega}(\tau, \tau')}{m} \hat{G}(\tau')$$

where $\hat{\Omega}(\tau, \tau')$ is defined in equation (22). The following proposition shows that a straightforward z-test provides valid inference that is adaptive in the degree of heterogeneity.

Proposition 2 (**Asymptotic normality of the test statistic**). *The conditions for Theorem 3 and Proposition 1 hold. Let $\eta \in \mathbb{R}^{T \times K}$ with $||\eta|| > \varepsilon > 0$. Then,*

$$\frac{\eta' \left(\hat{\delta}(\tau_1, \ldots, \tau_T) - \delta(\tau_1, \ldots, \tau_T) \right)}{\sqrt{\eta' \hat{\Sigma} \eta}} \xrightarrow{d} N(0, 1).$$

The asymptotic variance derived in Theorem 3 depends on the weighting matrix. Following standard GMM arguments, the efficient weighting matrix is given by

$$W(\tau)^* = \Omega_{mn}(\tau)^{-1} = \left(\Omega_1(\tau)/n + \Omega_2(\tau) \right)^{-1}. \tag{24}$$

This weighting matrix automatically takes into account the different rates of convergence of the different moments. If some moments converge faster than others, then this matrix, asymptotically, gives infinitely more weight to the fast moments than the slow moments, so the parameters identified by the fast moments will converge at the faster rate.

Usually, we would simply plug in a consistent estimator of $W^*(\tau)$ and obtain a feasible efficient GMM estimator, but Proposition 1 shows that the off-diagonal elements of $\Omega_{mn}(\tau)$ might not be consistently estimated when the rates of convergence of the corresponding coefficients are different. Proposition 3 shows that the estimated efficient weighting matrix nevertheless satisfies Assumption $8'$ such that Theorem 3 and Proposition 2 apply to the feasible efficient GM estimator. In addition, Theorem 3 reveals that the (first-order) asymptotic distribution of the estimator that uses $W^*(\tau)$ as the weighting matrix is the same as that of the estimator that uses $\hat{W}^*(\tau)$ as the weighting matrix.

Proposition 3 (**Adaptive Efficiency of the GMM Estimator**). *Assume that the conditions for Proposition 1 hold. Then, the estimated efficient weighting matrix $\hat{W}^*(\tau) = \hat{\Omega}(\tau)^{-1}$ satisfies Assumption $8'$.*

---

[23]We can similarly extend this proposition to multiple non-linear hypotheses.

Figure 1: First Stage Regressions

(a) Extrapolation

(b) Misspecification
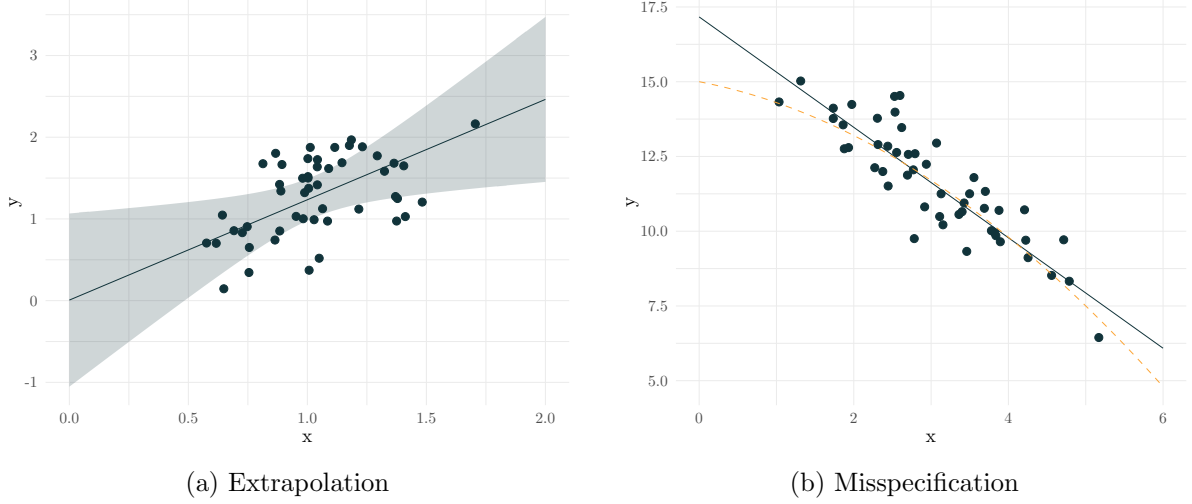
The figure illustrates the venerability of the intercept to extrapolation and misspecification. Both panels show generated data for one group and a first-stage fit. Panel (a) uses the same DGP as in the simulation of CLP. The solid line shows the regression line estimated by median regression, and the shaded area shows the 95% confidence interval. Panel (b) uses a different DGP where $y = 15 - 0.5x - 0.2x^2 + u$, where $u \sim N(0,1), x \sim N(3,1)$. The solid line is estimated by median regression without the quadratic term (misspecified). The dashed orange line is the true regression line.

If there are more moment conditions than parameters to estimate $(L > K)$, it is possible to implement overidentification tests in the spirit of Sargan (1958) and Hansen (1982). More precisely, we can test the validity of the instrumental variables while maintaining the other assumptions. The test statistic is the GMM objective function evaluated at the efficient GMM estimator:

$$J(\hat{\delta}, \tau) = m\bar{g}_{mn}\left(\hat{\delta}\left(\hat{W}^*(\tau), \tau\right), \tau\right)' \hat{\Omega}(\tau)^{-1}\bar{g}_{mn}\left(\hat{\delta}\left(\hat{W}^*(\tau), \tau\right), \tau\right) \tag{25}$$

Proposition 4 demonstrates the adaptive validity of the J-test implemented with the clustered covariance matrix.

Proposition 4 (Overidentification Test). *Under the $\mathbb{H}_0 : \mathbb{E}[z_{ij}\alpha_i(\tau)] = 0$ and the Assumptions required for Proposition 1, as n and $m \to \infty$, $J(\hat{\delta}(\tau_1, \ldots, \tau_T)) \xrightarrow{d} \chi^2_{TL-TK}$.*

We prove the result for a single quantile but the proof easily extends to multiple quantiles. In Section 5.3, we show that this overidentification test can be used to generalize the Hausman test comparing the random effects and the fixed effects estimators.

4   Grouped (IV) Quantile Regression Model

4.1   Chetverikov et al. (2016)

Chetverikov et al. (2016, hereafter CLP) propose an IV quantile regression estimator for group-level treatments. Their first model aligns with our model (1). Since their simulations

and all applications of their estimator adhere to this initial model, we begin by comparing their estimator with ours within this framework. Following that, we examine whether their second, more general model addresses some of the limitations of their estimator.

For model (1), CLP suggest an alternative two-step estimator. In the first stage, they perform quantile regression separately for each group j and quantile τ, regressing y_{ij} on x_{1ij} and a constant. In the second stage, they regress the *intercept* from the first stage on x_{2j}. The key difference between their approach and ours is that CLP use the estimated intercept from the first stage as the dependent variable in the second stage, whereas we use the fitted values.

Let's first compare the theoretical results. CLP focus exclusively on $\gamma(\tau)$, the coefficients on the group-level variables, and do not directly estimate $\beta(\tau)$. They assume strong group-level heterogeneity,[24] assuming $\text{Var}(\alpha_j(\tau)) > 0$ uniformly in τ. As a result, their theoretical findings are not adaptive with respect to $\text{Var}(\alpha_j(\tau))$. In other words, their asymptotic distribution corresponds to case 1-a) in our Theorem 2, where the variance from the first stage diminishes more rapidly than that from the second stage. Consequently, their asymptotic distribution is the same as if the true first-stage coefficients $\beta_j(\tau)$ were known. On a more technical note, when we focus on the coefficients of the group-level variables and assume group-level heterogeneity, we are able to relax the growth rate condition imposed by CLP, building on recent results from Volgushev et al. (2019).

Next, we compare the variance of our estimator to that of the CLP estimator. To simplify notation, we consider the exogenous case where x_{2j} can serve as its own instrumental variable, though the results extend to the endogenous case as well. As discussed in Section 3, the asymptotic variance consists of two components: one accounting for first-stage error and the other for second-stage noise. We will examine each component separately. To isolate the variance arising in the second stage, we consider the scenario where the true first-stage coefficients are known. In this scenario, the CLP point estimates can be obtained numerically within our MD framework by regressing the true first-stage fitted values on x_{1ij} and x_{2j}, using $\dot{x}_{1ij}$ and x_{2j} as instruments. Thus, the second-stage variance resulting from the randomness of $\alpha_j(\tau)$ is identical for both estimators. Our framework allows exploiting, in addition, the between-group variation in $\bar{x}_{1j}$ using the efficient GMM estimator, which will (weakly) reduce the variance of the estimator. Other than this advantage, our approach is neither better nor worse than the CLP estimator concerning second-stage variance.

We can isolate the first-stage error by setting $\text{Var}(\alpha_j(\tau)) = 0$. Under this condition, our estimator becomes a classical MD estimator. To facilitate a comparison between both estimators, we can express the CLP estimator as a minimum distance estimator based on the same first-stage estimates:

$$\hat{\delta}_{CLP}(\tau) = \underset{\delta \in \mathbb{R}^{K_1 \cdot m + K_2}}{\arg\min} \frac{1}{m} \sum_{j=1}^{m} \left(\hat{\beta}_j(\tau) - \tilde{R}_j \delta \right)' \left(\hat{\beta}_j(\tau) - \tilde{R}_j \delta \right), \tag{26}$$

---

[24]Although this assumption is not explicitly stated, their asymptotic distribution would become degenerate without it, as they discuss in footnote 9.

where

$$\underset{(K_1+1)\times(K_1\cdot m+K_2)}{\tilde{R}_j} = \begin{pmatrix} 0 & x'_{2j} \\ l'_j \otimes I_{K_1} & 0 \end{pmatrix},$$

and l_j is a m-dimensional vector of zeros with a 1 in the j position. The restriction matrix $\tilde{R}_j$ is different from the restriction matrix of our estimator defined in equation (7), as it does not impose equality of the first stage coefficients implied by the model. Thus, our estimator imposes $K_1 \cdot (m-1)$ additional restrictions that are correct. A second difference is that CLP use an identity weighting matrix while we use $\tilde{X}'_j \tilde{X}_j$. When there is no group heterogeneity, we are in the classical MD framework such that the efficient weighting matrix is the inverse of the first-stage variance. It follows that weighting by $\tilde{X}'_j \tilde{X}_j$ is efficient when the first-stage error is separable and the density of y given x at the τ quantile is the same across groups. When, in addition, $\tilde{X}'_j \tilde{X}_j$ is constant across groups (balanced panel, identical distribution of x_{1ij}), then equal weighting is efficient.

Adding valid constraints within an efficient MD framework reduces variance. Therefore, if $\tilde{X}'_j \tilde{X}_j$ is the efficient weighting matrix, our estimator will necessarily have lower first-stage variance than the CLP estimator. However, with an inefficient weighting matrix, adding valid constraints might, in some relatively pathological cases, increase the variance.[25] While it is possible to develop a more efficient estimator by following the ideas discussed in Section 5.1, we prefer to avoid estimating the first-stage variance and instead opt for a more interpretable estimator in cases of misspecification.

To summarize the comparison, the MD estimator using $\dot{x}_{1ij}$ and x_{2j} as instrumental variables exhibits the same variance due to the randomness of $\alpha_j(\tau)$ but a lower variance due to estimation of $\hat{\beta}_j(\tau)$ compared to the CLP estimator—this holds formally when the efficient weighting matrix is used. Note that CLP assume strong group heterogeneity so that only the variance arising in the second stage matters asymptotically. Consequently, it's not surprising that they did not optimize their estimator to minimize the variance arising in the first stage. Intuitively, the first-stage intercept used by CLP represents the fitted value at an x_{1ij} value that may lie outside of the variable's support. For example, in a wage regression where one of the individual-level variables is *age*, the variance of the fitted value at *age* $= 0$ is higher than that at a value near the center of the *age* distribution. Additionally, even a slight misspecification of the conditional quantile function is more likely to have a significant impact when we extrapolate beyond the support of the variable. Figure 1 illustrates these issues with artificial data.

CLP also consider a generalization of model (1) in which they assume

$$Q(\tau, y_{ij}|x_{1ij}, x_{2j}, v_j) = \tilde{x}'_{1ij}\beta_j(\tau) \tag{27}$$

$$\beta_{j,1}(\tau) = x'_{2j}\gamma(\tau) + \alpha(\tau, v_j). \tag{28}$$

In contrast to model (1), this approach allows the coefficient on x_{1ij} to vary across groups. By default, $\beta_{j,1}(\tau)$ is the first element of the vector $\beta_j(\tau)$, but it could represent any element of

---

[25]See the discussion in Section 8 of Hansen (2022b).

this vector. This allows the researcher to also estimate the interaction effects of the group-level treatment and a micro-level covariate. However, if the effect of a group-level variable varies with individual-level variables, the effect of x_{2j} on the intercept will not represent an average effect, but the effect evaluated at $x_{1ij} = 0$, which may not be of interest. In addition, the estimator of the effect on the intercept suffers from all the issues discussed above concerning the variance of the CLP estimator. To solve the problem, the researcher should estimate all interaction effects and combine them. This approach has not been discussed in CLP and has not been implemented in their simulations or in any application of their estimator.

4.2 Simulations

This subsection presents Monte Carlo simulations comparing the MD and CLP estimators. The simulations are based on the same data generating process and sample sizes as in Chetverikov et al. (2016). That is, $(m, n) = \{(25, 25), (200, 25), (25, 200), (200, 200)\}$. For both estimator we use a OLS (or 2SLS) second stage. The generated data include one individual-level regressor, one group-level regressor, and one instrument. Heterogeneity is introduced via a rank variable u_{ij}, and the data is generated as follows:

$$y_{ij} = \beta_0(u_{ij}) + x_{1ij}\beta(u_{ij}) + x_{2j}\gamma(u_{ij}) + \alpha_j(u_{ij}), \tag{29}$$

$$z_j = x_{2j} + \eta_j + \nu_j, \tag{30}$$

$$\alpha_j(u_{ij}) = u_{ij}\eta_j - \frac{u_{ij}}{2}, \tag{31}$$

where x_{1ij}, x_{2j} and ν_j are distributed $\exp(0.25 \cdot N[0, 1])$ and η_j as well as the rank variable u_{ij} are $U[0, 1]$ distributed. The data generating process implies that $\mathbb{E}[\alpha(u_{ij})|x_{2j}] = \mathbb{E}[u_{ij}\eta_j - \frac{u_{ij}}{2}|x_{2j}] = \mathbb{E}[\frac{u_{ij}}{2} - \frac{u_{ij}}{2}|x_{2j}] = 0$. At quantiles $\tau \in (0, 1)$, the true parameters $\gamma(\tau)$ and $\beta(\tau)$ equal $\sqrt{\tau}$ and, $\alpha_1(\tau) = \frac{\tau}{2}$. Consequently, $\gamma(u_{ij}) = \beta(u_{ij}) = \sqrt{u_{ij}}$ and $\beta_0(u_{ij}) = \frac{u_{ij}}{2}$.

The simulations consider three cases. In the first one, $\alpha_j(\tau) = 0$ for all j and all τ. In this case, as there are no group effects, conditioning on the group does not affect the quantile function, and quantile regression is consistent for the same parameter. Further, both estimators are $\sqrt{mn}$-consistent. In the second case, there are group-specific effects ($\alpha_j(\tau) \neq 0$), which are uncorrelated with the regressors. As individual heterogeneity is multiplied with the rank variable u_{ij}, there is only weak group heterogeneity in the lower tail of the distribution, and we see faster convergence there. In the third case, $\alpha_j(\tau)$ is correlated with the regressor of interest, such that x_{2j} is endogenous. For the implementation of the MD estimator, we use $\dot{x}_{1ij}$ as an instrument for x_{1ij} so that, as with the CLP estimator, we do not exploit the between variation of the regressor.[26] In the third case, as x_{2j} is endogenous, we instrument x_{2j} with z_j. Since the data generating process of Chetverikov et al. (2016) has a weak instrument when m is small,

---

[26]Using the entire variation of x_{1ij} does not affect the results since there is no between variation in x_{1ij}.

Table 1: Bias, Standard Deviation and MSE of $\hat{\gamma}(\tau)$

Quantile	No group-level heterogeneity			Exogenous			Endogenous		
	MD	CLP	Rel. MSE	MD	CLP	Rel. MSE	MD	CLP	Rel. MSE
				$(m, n) = (25, 25)$					
0.1	0.022	-0.011	0.051	0.022	-0.010	0.052	0.036	0.001	0.160
	(0.192)	(0.858)		(0.195)	(0.860)		(2.025)	(5.062)	
0.5	-0.010	-0.001	0.061	-0.011	0.000	0.088	-0.029	0.039	0.132
	(0.166)	(0.673)		(0.204)	(0.691)		(1.992)	(5.491)	
0.9	-0.019	-0.003	0.049	-0.020	-0.004	0.216	-0.063	-0.011	0.176
	(0.094)	(0.435)		(0.227)	(0.490)		(2.123)	(5.065)	
				$(m, n) = (200, 25)$					
0.1	0.024	0.003	0.060	0.024	0.004	0.063	0.023	0.006	0.057
	(0.066)	(0.284)		(0.067)	(0.285)		(0.106)	(0.456)	
0.5	-0.006	-0.001	0.059	-0.006	0.000	0.086	-0.009	-0.003	0.071
	(0.056)	(0.232)		(0.069)	(0.238)		(0.097)	(0.366)	
0.9	-0.017	-0.004	0.060	-0.017	-0.003	0.223	-0.022	-0.009	0.142
	(0.031)	(0.145)		(0.075)	(0.164)		(0.086)	(0.234)	
				$(m, n) = (25, 200)$					
0.1	0.003	-0.002	0.059	0.003	-0.001	0.066	-0.028	-0.076	0.141
	(0.070)	(0.289)		(0.074)	(0.291)		(2.107)	(5.618)	
0.5	-0.001	-0.002	0.060	-0.001	-0.001	0.233	-0.083	-0.094	0.628
	(0.060)	(0.247)		(0.134)	(0.278)		(3.627)	(4.575)	
0.9	-0.002	0.000	0.061	-0.001	0.001	0.769	-0.120	-0.114	1.218
	(0.030)	(0.121)		(0.217)	(0.247)		(3.930)	(3.561)	
				$(m, n) = (200, 200)$					
0.1	0.003	-0.003	0.057	0.003	-0.003	0.062	0.002	-0.004	0.058
	(0.024)	(0.100)		(0.025)	(0.101)		(0.039)	(0.162)	
0.5	-0.001	0.000	0.059	-0.001	-0.001	0.222	-0.004	-0.004	0.141
	(0.020)	(0.084)		(0.044)	(0.093)		(0.051)	(0.136)	
0.9	-0.002	0.000	0.067	-0.003	-0.001	0.762	-0.009	-0.007	0.617
	(0.010)	(0.040)		(0.071)	(0.082)		(0.074)	(0.095)	

Note:
The table reports mean bias, standard deviation and relative MSE from the simulations for $\gamma(\tau)$ from 10,000 Monte Carlo simulations using the MD estimator and the CLP estimator. The relative MSE gives the MSE of the MD estimator relative to that of the CLP estimator.

one should pay attention when looking at the simulation results for the endogenous case.[27] In empirical research, it is straightforward to construct confidence intervals that are valid even if identification is weak. We perform 10,000 Monte Carlo replications for the set of quantiles $\tau \in \{0.1, 0.5, 0.9\}$. Since the CLP estimator does not directly provide an estimate for $\beta(\tau)$, we present only results for $\gamma(\tau)$.

Table 1 reports the bias, standard deviation, and relative MSE of the estimators. The relative MSE reports the MSE of the MD estimator relative to that of the CLP estimator. Thus, a number smaller than 1 indicates that the MD estimator has a lower MSE. The CLP estimator seems to have a smaller bias than the MD estimator when $n = 25$. When n increases to 200, the difference disappears. There are more remarkable differences in the variance of the estimators. The standard deviation of the MD estimator is four times smaller compared to

---

[27]With $m = 25$ in over 40% of the draws, the F-statistics of the first stage of the IV estimations is below 10. The issue disappears when $m = 200$.

Table 2: Coverage Probability of the 95% Confidence Intervals

	No group-level heterogeneity			Exogenous			Endogenous		
	Rel. length	Coverage Rate		Rel. length	Coverage Rate		Rel. length	Coverage Rate	
Quantile	MD/CLP	MD	CLP	MD/CLP	MD	CLP	MD/CLP	MD	CLP
				(m, n) = (25, 25)					
0.1	0.232	0.941	0.938	0.235	0.939	0.938	0.227	0.966	0.972
0.5	0.244	0.940	0.942	0.301	0.943	0.945	0.262	0.963	0.972
0.9	0.223	0.941	0.949	0.501	0.940	0.946	0.373	0.956	0.972
				(m, n) = (200, 25)					
0.1	0.230	0.932	0.947	0.233	0.932	0.948	0.231	0.942	0.953
0.5	0.245	0.946	0.944	0.296	0.945	0.946	0.267	0.952	0.949
0.9	0.220	0.926	0.947	0.475	0.941	0.945	0.368	0.953	0.952
				(m, n) = (25, 200)					
0.1	0.241	0.943	0.940	0.256	0.943	0.943	0.240	0.968	0.974
0.5	0.242	0.937	0.944	0.495	0.938	0.944	0.370	0.949	0.971
0.9	0.248	0.948	0.941	0.884	0.933	0.945	0.771	0.939	0.955
				(m, n) = (200, 200)					
0.1	0.241	0.944	0.944	0.254	0.947	0.945	0.246	0.951	0.950
0.5	0.244	0.946	0.945	0.483	0.952	0.948	0.377	0.957	0.951
0.9	0.246	0.942	0.953	0.872	0.950	0.950	0.772	0.954	0.955

Note:
Results based on 10,000 Monte Carlo simulations. The table provides the coverage rate and median length of the confidence intervals of $\gamma(\tau)$. The relative length provides the length of the confidence interval of the MD estimator relative to that of the CLP estimator. Robust standard errors are used for the CLP estimator, and clustered standard errors at the group level are used for the MD estimator.

that of the CLP estimator in the homogeneous case. The difference is somewhat smaller in the exogenous and endogenous cases but remains substantial.[28] The differences in the standard deviations of the estimator in the different cases also reflect the precision improvement of the MD estimator compared to the CLP estimator discussed above. The disparity is most remarkable in the case without group-level heterogeneity, where both estimators converge at the $\sqrt{mn}$ rate as we provide a first-order improvement compared to CLP. Differently, in the exogenous case, where $\text{Var}(\alpha(\tau))$ increases over τ and therefore the convergence rate approaches $\sqrt{m}$, we see that the improvement becomes smaller at higher quantiles and with larger n. This difference in precision explains the large discrepancies in MSE. The MSE of the CLP estimator is over ten times larger than that of the MD estimator when $\alpha_j(\tau) = 0$ and remains substantially larger in all scenarios considered. If $\alpha_j(\tau) = 0$, quantile regression is a consistent estimator for $\beta(\tau)$. Although not shown here, simulation results comparing our estimator with traditional quantile regression show that the two estimators are indistinguishable in terms of bias and variance in large samples.

Table 2 show the performance of the 95% confidence intervals suggested with our inference procedure. The table reports the coverage rate and the median length of the intervals of our

---

[28]The standard deviations in the endogenous case with $m = 25$ should be interpreted with caution due to the weak instrument.

estimator relative to that of the CLP estimator.[29] Our suggested inference procedure has coverage close to 95% in all cases. Compared to the CLP estimator, our confidence bands are substantially shorter. In most cases, our estimator yields confidence bands less than half the length of those for the CLP estimator. This difference becomes even more pronounced in our empirical application in Section 6, where the CLP estimator produces confidence bands that are, on average, 14 times wider than those of the MD estimator.

5 Traditional Quantile Panel Data Estimators

5.1 Fixed Effects, Random Effects and Between Estimators

In this subsection, we apply our results to derive quantile analogs of the fixed effects, between, random effects, and Hausman-Taylor estimators. As discussed in Section 2.3, these estimators can be obtained by selecting appropriate instrumental variables in the second stage. For fixed effects estimation, model (1) implies that $\dot{x}_{1ij}$ is a valid instrument since it varies only within groups and is uncorrelated with the group effects. This instrument automatically satisfies Assumption 5. In this special case, the approach corresponds to the traditional MD estimator, where all variance originates in the first stage.

In the first stage, $\beta(\tau)$ is estimated separately for each group. The second stage then averages these group-level coefficients using weights proportional to $\tilde{X}_j'\tilde{X}_j$ (see equation 8). Galvao and Wang (2015) propose an alternative approach, suggesting efficient weights that are proportional to the inverse of the variance of the first-stage estimators. Their weights are equivalent to ours when

$$f_{y_{ij}|x_{1ij},v_j}(Q(\tau,y_{ij}|x_{1ij},v)|x,v) = f_{y_{ij}|x_{1ij},v_j}(Q(\tau,y_{ij}|x_{1ij},v')|x,v') \qquad (32)$$

for any v and v', i.e. when the conditional distribution of the group effects is the same across groups. Outside this specific case, our estimator may be less efficient but avoids the need to estimate the first-stage variance, which depends on the conditional densities in equation (32). A third approach is to take the unweighted average of $\hat{\beta}_j(\tau)$, which, while not efficient when $\beta_j(\tau)$ are homogeneous, remains straightforward to interpret even if the model is misspecified.

To exploit only the variation across groups, we can implement a quantile between estimator using $\bar{x}_{1j}$ as an instrument. On the other hand, combining within and between variation is more complex for quantile models than for least squares models. Applying quantile regression to the whole population without controlling for groups identifies parameters that differ from our intended parameters (see Remark 1). Using our MD estimator with x_{1ij} as an instrument, which corresponds to using OLS in the second stage, consistently estimates $\beta(\tau)$ but only at the slow $\sqrt{m}$ rate because x_{1ij} also varies between groups. The same occurs if both $\dot{x}_{1ij}$ and $\bar{x}_{1j}$ are combined with 2SLS, as the weights attributed to $\bar{x}_{1j}$ do not vanish asymptotically.

---

[29]For the CLP estimator, we use heteroskedasticity robust standard errors as suggested in Chetverikov et al. (2016). Recall that the CLP estimator uses only one observation per group in the second stage. Hence, we would attain the same standard errors if we kept all observations and clustered the standard errors at the group level.

Instead, we propose two efficient random effects estimators: an efficient GMM estimator and one using optimal instruments. It is worth highlighting that these estimators not only reduce the asymptotic variance of the estimator but also increase the rate of convergence from $\sqrt{m}$ to $\sqrt{mn}$ compared to 2SLS.

Given the first-stage estimation, we have the following moment condition:

$$\mathbb{E}[Z_j'(\tilde{X}_j \hat{\beta}_j(\tau) - X_j \delta(\tau)] = 0. \tag{33}$$

When the instrument includes both the within-group variation $\dot{x}_{1ij}$ and the between-group average $\bar{x}_{1j}$, the efficient GMM estimator will optimally combine these two sources of variation. The weighting matrix is computed as shown in equation (22). According to Propositions 3 and Theorem 2, this estimator is guaranteed to be both $\sqrt{mn}$ consistent and efficient. This random effects estimator has the same first-order asymptotic distribution as the fixed effects estimator that uses only the within-group variation $\dot{x}_{1ij}$. However, the random effects estimator is expected to have a lower variance in finite samples because it also incorporates the between-group variation. As the number of observations n increases, the influence of the between-group variation diminishes, causing the random effects estimator to converge to the fixed effects estimator. This behavior is similar to what is observed in least squares models, as discussed by Baltagi (2021) and Ahn and Moon (2014).

If we impose the stronger assumption that the moment restriction in equation (33) holds conditionally on Z_j, we can use the theory of optimal instruments to derive a more efficient random effects estimator. Optimal instruments are relevant when a researcher has a conditional moment restriction of the form $\mathbb{E}[g_j(\delta, \tau)|Z_j] = 0$. When a moment condition holds conditional on Z_j, an infinite set of valid moments exist, and one could use additional moments to increase efficiency. The goal is to select the instrument that minimizes the asymptotic variance, which takes the form $Z_j^* = \mathbb{E}[g_j(\delta, \tau)g_j(\delta, \tau)'|Z_j]^{-1}R_j(\delta, \tau)$, with $R_j(\delta, \tau) = \mathbb{E}[\frac{\partial}{\partial \delta}g_j(\delta, \tau)|Z_j]$ (see, e.g., Chamberlain, 1987 and Newey, 1993). To implement the random effect estimator with optimal instruments, we set $Z_j = X_j$. Under the additional assumption that $\mathbb{E}[\alpha_j^2(\tau)|X_j] = \sigma_\alpha^2(\tau)$,[30] the optimal instrument simplifies to

$$Z_j^*(\tau) = \left(\tilde{X}_j \frac{V_j(\tau)}{n} \tilde{X}_j' + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2(\tau) \right)^+ X_j, \tag{34}$$

where $V_j(\tau)$ is the asymptotic variance from the first stage for a group j, $\mathbf{l}_n$ is a n-dimenstional vector of ones, and $^+$ denotes the Moore-Penrose inverse.[31]

A few remarks about the optimal instruments follow. First, under standard random effects assumptions, the optimal instrument applied to least squares models is numerically identical to

---

[30]We assume homoskedasticity of $\alpha_j(\tau)$ to obtain a simple estimator, similar to the classical least squares random effects estimator. If we were to drop this assumption, we would need to estimate $\mathbb{E}[\alpha_j(\tau)|X_j]$. Note that we do not assume homoscedasticity in the group-level model such that this assumption does not constrain the heterogeneity of $\beta(\tau)$ across different values of τ.

[31]Since the matrix $(\tilde{X}_j \frac{V_j(\tau)}{n} \tilde{X}_j' + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2(\tau))$ is singular, we use the Moore-Penrose inverse.

the FGLS estimator. Second, the optimal instrument depends on n analogously to the efficient weighting matrix of the GMM estimator. As n increases, the first stage variance converges to zero, and the generalized inverse will give infinitely more weights to the within variation and asymptotically converge to the fixed effects estimator. Third, if $\sigma_\alpha(\tau) = 0$, then all the variance arises in the first stage and this estimator is identical to the efficient MD estimator (see Proposition 8 in Appendix C).

5.2 Hausman and Taylor Model

The Hausman-Taylor model provides a method for finding instrumental variables within the model itself. It represents a middle ground between the random effects model, which assumes orthogonality between $\alpha_j(\tau)$ and x_{ij}, and the fixed effects model, which only identifies the effect of individual-level variables. To estimate the effect of group-level variables, Hausman and Taylor (1981) assume that some elements of x_{1ij} are uncorrelated with $\alpha_j(\tau)$. We consider model (1) but partition the regressors into four types of variables, $x_{ij} = [x_{1ij}^{ex} \ x_{1ij}^{en} \ x_{2j}^{ex} \ x_{2j}^{en}]$, where the superscript ex indicates exogenous variables and the superscript en indicates potentially endogenous variables. Thus,

$$\mathbb{E}[x_{1ij}^{ex}\alpha_j(\tau)] = 0,$$
$$\mathbb{E}[x_{2j}^{ex}\alpha_j(\tau)] = 0.$$

The assumptions imply that we can estimate $\delta(\tau)$ using the instrument $z_{ij} = (\dot{x}_{1ij}^{ex}, \dot{x}_{1ij}^{en}, \bar{x}_{1ij}^{ex}, x_{2j}^{ex})$. While x_{2j}^{en} is potentially endogenous, the within variation is uncorrelated with $\alpha_j(\tau)$ as it varies only within j. Identification requires at least as many instruments as parameters to estimate, hence $dim(x_{1ij}^{ex}) \geq dim(x_{2j}^{en})$. In overidentified models, efficient GMM can be implemented, and if conditional moment restrictions are available, optimal instruments can be used. However, implementing optimal instruments is not straightforward as it requires estimating $\mathbb{E}[x_{ij}|z_{ij}]$, typically done nonparametrically (see Newey, 1993). We do not contribute to this aspect in this paper.

5.3 Hausman Test

The random effects estimator's consistency relies on stronger orthogonality conditions compared to the fixed effects estimator. Under these stronger assumptions, both estimators are consistent, but the fixed effects estimator is inefficient. Hausman (1978) proposed a test for the null hypothesis of random effects against the alternative of fixed effects. Various generalizations of the Hausman test have been suggested in the literature (e.g., Chamberlain (1982); Mundlak (1978); Wooldridge (2019)). Arellano (1993) considers a heteroskedasticity and autocorrelation robust generalization based on a Wald test. Ahn and Low (1996) propose a GMM test based on a 3SLS regression as an equivalent method for the Hausman test.

This subsection explains how we can use the overidentification test presented in Section 3 as a quantile version of the Hausman test for our two-step estimator. The assumption of correct specification of the first stage is maintained under both the null and alternative hypotheses. Compared to the fixed effects estimator, consistency of the random effects estimator additionally requires that x_{ij} is uncorrelated with $\alpha_j(\tau)$, so that $\mathbb{E}[\dot{x}'_{1ij}\alpha_j(\tau)] = 0$ and $\mathbb{E}[\bar{x}'_{ij}\alpha_j(\tau)] = 0$ are valid moment conditions. By contrast, the fixed effects estimator relies only on the moment condition $\mathbb{E}[\dot{x}'_{1ij}\alpha_j(\tau)] = 0$. Consequently, the overidentification test suggested in Proposition 4 can be used as a test of the random effects orthogonality conditions. Compared to the traditional Hausman test, our test does not rely on the assumption of conditional homoskedasticity of the errors and is robust to clustering.

Galvao and Poirier (2019) also note that a quantile regression with all observations and fixed effects quantile regression identify different parameters. They propose a Hausman test based on an auxiliary quantile regression model that incorporates both x_{1ij} and $\bar{x}_{1j}$ as regressors, and tests for the significance of $\bar{x}_{1j}$. However, this test differs from our approach in several ways: it starts from a different model (not conditional on the groups), relies on the correct specification of the auxiliary regression, and does not directly compare random effects and fixed effects estimators.

5.4 Simulations

This section presents simulation results for the different panel data estimators and the Hausman-type test presented in the previous subsections. These simulations focus on the estimation of $\beta(\tau)$, while the next section includes results for $\gamma(\tau)$. We consider the following data generating process where all variables are scalars:

$$y_{ij} = \beta x_{1ij} + \alpha_j + (1 + 0.1x_{1ij})\nu_{ij}. \tag{35}$$

We let $\beta = 1$ and $\nu_{ij} \sim \mathcal{N}(0,1)$. The regressor is defined by $x_{1ij} = h_j + 0.5u_{ij}$, with $u_{ij} \sim \mathcal{N}(0,1)$ and

$$\begin{pmatrix} h_j \\ \alpha_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}\right).$$

If $\lambda \neq 0$, x_{1ij} is correlated with α_j. For the simulation of the panel data estimators, we let $\lambda = 0$ so that all estimators are consistent. In contrast, in the Monte Carlo study of the Hausman test, we set $\lambda = \{0, 0.1, 0.2, 0.3, 0.4\}$. The true coefficient takes the values $\beta(\tau) = \beta + 0.1F^{-1}(\tau)$ where F is the standard normal CDF. We consider the samples with $n = \{10, 25, 200\}$ and $m = \{25, 200\}$ and focus on the set of quantiles $\mathcal{T} = \{0.1, 0.5, 0.9\}$. All simulation results are based on 10,000 replications.

We compare the performance of four MD estimators, all based on the same first-stage regression but differing in their choice of instruments in the second stage. The 'Pooled' estimator uses x_{1ij} as the instrument,[32] the BE estimator uses $\bar{x}_{1j}$ as the instrument, the RE-GMM esti-

---

[32]We refer to this estimator as 'Pooled' because the second stage involves a pooled OLS regression. It should not be confused with the one-step pooled quantile regression, which does not account for group structure.

Table 3: Bias and Standard Deviation of $\hat{\beta}(\tau)$

Quantile	Pooled	BE	FE	RE-GMM	RE-OI
		(m, n) = (25, 10)			
0.1	0.009	0.002	0.037	0.019	0.044
	(0.193)	(0.235)	(0.261)	(0.183)	(0.177)
0.5	0.000	0.000	-0.001	-0.001	0.000
	(0.182)	(0.224)	(0.172)	(0.142)	(0.168)
0.9	-0.010	-0.003	-0.039	-0.021	-0.045
	(0.195)	(0.235)	(0.259)	(0.184)	(0.181)
		(m, n) = (200, 10)			
0.1	0.011	0.005	0.040	0.021	0.046
	(0.068)	(0.080)	(0.092)	(0.061)	(0.067)
0.5	0.001	0.001	0.001	0.001	0.001
	(0.063)	(0.076)	(0.059)	(0.047)	(0.063)
0.9	-0.010	-0.003	-0.040	-0.019	-0.045
	(0.067)	(0.080)	(0.091)	(0.060)	(0.068)
		(m, n) = (25, 25)			
0.1	0.003	0.000	0.015	0.011	0.016
	(0.175)	(0.222)	(0.141)	(0.122)	(0.120)
0.5	-0.003	-0.004	0.000	-0.001	-0.002
	(0.171)	(0.218)	(0.102)	(0.094)	(0.106)
0.9	-0.009	-0.007	-0.017	-0.014	-0.018
	(0.177)	(0.223)	(0.138)	(0.121)	(0.120)
		(m, n) = (200, 25)			
0.1	0.006	0.004	0.015	0.012	0.017
	(0.061)	(0.075)	(0.049)	(0.041)	(0.042)
0.5	0.000	0.000	0.000	0.000	0.000
	(0.059)	(0.073)	(0.036)	(0.032)	(0.036)
0.9	-0.006	-0.004	-0.015	-0.012	-0.017
	(0.061)	(0.075)	(0.049)	(0.041)	(0.042)
		(m, n) = (25, 200)			
0.1	0.001	0.002	0.002	0.002	0.002
	(0.163)	(0.211)	(0.049)	(0.049)	(0.047)
0.5	0.001	0.001	0.000	0.000	0.000
	(0.163)	(0.210)	(0.035)	(0.035)	(0.035)
0.9	0.000	0.001	-0.002	-0.002	-0.002
	(0.163)	(0.211)	(0.049)	(0.048)	(0.046)
		(m, n) = (200, 200)			
0.1	0.000	0.000	0.002	0.002	0.002
	(0.058)	(0.073)	(0.017)	(0.017)	(0.016)
0.5	0.000	0.000	0.000	0.000	0.000
	(0.058)	(0.072)	(0.013)	(0.012)	(0.012)
0.9	-0.001	-0.001	-0.002	-0.002	-0.002
	(0.058)	(0.073)	(0.017)	(0.017)	(0.017)

Note:
The table reports bias and standard deviation (in parentheses) of the simulations for $\beta(\tau)$ from 10,000 Monte Carlo simulations.

Table 4: Coverage Probability of the 95% Confidence Invervals

Quantile	Pooled	BE	FE	RE-GMM	RE-OI
(m, n) = (25, 10)					
0.1	0.948	0.922	0.946	0.920	0.914
0.5	0.946	0.920	0.948	0.923	0.914
0.9	0.950	0.924	0.950	0.918	0.912
(m, n) = (200, 10)					
0.1	0.942	0.941	0.927	0.930	0.874
0.5	0.947	0.943	0.952	0.948	0.943
0.9	0.946	0.942	0.932	0.934	0.877
(m, n) = (25, 25)					
0.1	0.949	0.921	0.949	0.928	0.933
0.5	0.946	0.918	0.948	0.933	0.931
0.9	0.946	0.917	0.951	0.931	0.931
(m, n) = (200, 25)					
0.1	0.947	0.945	0.942	0.939	0.928
0.5	0.950	0.945	0.954	0.948	0.948
0.9	0.948	0.946	0.938	0.938	0.927
(m, n) = (25, 200)					
0.1	0.950	0.923	0.950	0.938	0.947
0.5	0.949	0.926	0.952	0.941	0.948
0.9	0.947	0.925	0.950	0.938	0.948
(m, n) = (200, 200)					
0.1	0.948	0.943	0.947	0.946	0.950
0.5	0.947	0.943	0.951	0.951	0.950
0.9	0.948	0.944	0.949	0.947	0.949

Note:
Results based on 10,000 Monte Carlo simulations. The table reports the coverage probabilities of the confidence intervals of $\beta(\tau)$.

Table 5: Hausman Test

Quantile	λ				
	0.0	0.1	0.2	0.3	0.4
(m,n) = (25, 10)					
0.1	0.052	0.058	0.076	0.116	0.178
0.5	0.048	0.060	0.095	0.162	0.268
0.9	0.050	0.066	0.094	0.144	0.218
(m,n) = (200, 10)					
0.1	0.062	0.085	0.277	0.579	0.844
0.5	0.050	0.177	0.532	0.871	0.987
0.9	0.058	0.194	0.483	0.782	0.949
(m,n) = (25, 25)					
0.1	0.046	0.058	0.098	0.173	0.281
0.5	0.043	0.060	0.114	0.204	0.340
0.9	0.045	0.062	0.110	0.186	0.306
(m,n) = (200, 25)					
0.1	0.051	0.166	0.554	0.897	0.994
0.5	0.050	0.232	0.689	0.963	0.999
0.9	0.049	0.230	0.645	0.938	0.997
(m,n) = (25, 200)					
0.1	0.043	0.062	0.123	0.230	0.399
0.5	0.043	0.063	0.125	0.242	0.412
0.9	0.044	0.063	0.124	0.235	0.403
(m,n) = (200, 200)					
0.1	0.054	0.259	0.771	0.985	1.000
0.5	0.054	0.272	0.788	0.988	1.000
0.9	0.052	0.270	0.785	0.987	1.000

Note:
The table reports rejection rates of the Hausman test. The results are based on 10,000 Monte Carlo simulations. The first column shows the empirical size, while the other columns show the power of the test.

mator combines $\dot{x}_{1ij}$ and $\bar{x}_{1j}$ using efficient GMM, and the RE-OI estimator combines the same exogenous variation using the single optimal instrument defined in equation (34). We use the estimator of Powell (1991) for $V_j(\tau)$ and the estimator of Nerlove (1971) for $\sigma_\alpha^2(\tau)$.

Table 3 shows that the estimators perform well also when both m and n are small. The RE-GMM estimator performs similarly to the RE-OI estimator except with small n when the RE-GMM estimator outperforms the RE-OI both in terms of bias and variance. As expected, the RE-GMM, the RE-OI, and the fixed effects (FE) estimators become indistinguishable as n increases. Whereas with small n, there is an apparent gain in using a random effects estimator. From the standard deviations, it is possible to see the different rates of convergence of the estimators. The precision of the fixed effects and random effects estimators increases in similar magnitude when m or n increases. In contrast, the standard deviation of the pooled and between estimators decreases only when m increases. The pooled and the between estimators have the smallest bias but, in most cases, also the largest variance.

The coverage probabilities of the 95% confidence intervals are provided in Table 4. The confidence intervals of the pooled and the fixed effects estimator perform well in all sample sizes considered. On the other hand, the confidence bands of the random effects estimators slightly undercover the true parameter mostly when n is small. In larger samples, all the coverage probabilities are close to the theoretical level.

Table 5 shows the rejection probabilities of the overidentification test for different values of λ. When $\lambda = 0$, the $\mathbb{H}_0$ is satisfied, so we should reject the null at a rate close to 5%. If $\lambda \neq 0$, x_{ij} is correlated with α_j, and some moment conditions used by the RE-GMM estimator are not valid. In this case, higher rejection probabilities suggest a more powerful test. The first column shows that the empirical sizes of the test are close to the theoretical levels in most sample sizes. Columns 2-5 show that, as expected, the power of the test is higher in large samples and increases with the correlation between $\bar{x}_{1j}$ and the unobserved heterogeneity α_j. An increase in m substantially improves the power of the test, while a larger number of time periods n improves the results to a lesser extent. In general, the test performs better both in terms of size and power when m is large, which is most often the case in empirical applications. Even if the random effects estimator converges to the fixed effects estimator as n increases and the random effects estimator of β will be consistent even if $\lambda \neq 0$, the size and power of the test do not deteriorate. This result is consistent with the findings in Ahn and Moon (2014).

6   Empirical Application: The Effect of the Food Stamps Program on Birth Weight

In this section, we apply our minimum distance approach to estimate the impact of the food stamp program on the birth weight distribution using grouped data. We complement the analysis of Almond et al. (2011) by providing distributional effects. Food stamps constitute an important means-tested program that gives entitled households coupons they can redeem at approved retail food stores. The Food Stamp Act (FSA) was introduced in 1964 and enabled counties to start their own federally funded food stamp program (FSP). In the subsequent years, counties increasingly adopted such programs, and in 1973, an amendment to the FSA required all counties to establish a FSP by 1975. Thus, the share of counties with an FSP increased steadily from 1964 to 1974, and identification exploits the variation in the timing of the adoption across counties. Almond et al. (2011) use data from 1968 (when about 40% of the counties had introduced the program) to 1977 (two years after the FSP was implemented everywhere) to analyze the effect of the program.

Given the negative consequences of low birth weight, besides estimating the effect of the policy on average weight, Almond et al. (2011) estimate the effect on the probability that birth weight falls below a certain threshold. As discussed in Melly and Santangelo (2015), this procedure leads to biased results unless there is no time effect or group effect or the outcome is

Figure 2: Impact of Food Stamp Introduction on the Distribution of Birth Weight
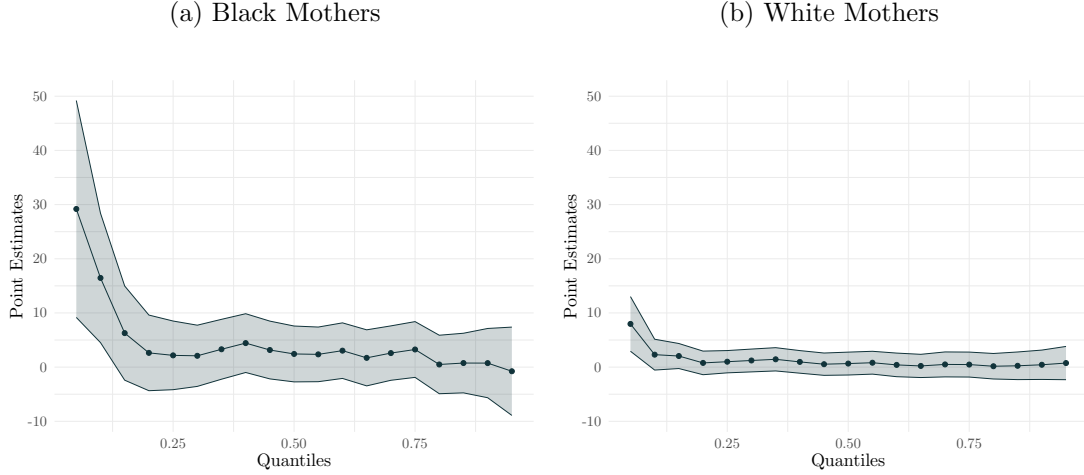
(a) Black Mothers

(b) White Mothers

The figure shows the impact of the food stamp introduction on the conditional distribution of birth weight. The panels show point estimates and 95% confidence bands (shaded area) computed using standard errors clustered at the county level. The panel on the left (right) shows the effects for blacks (whites). The regressions include county, time, and state-year fixed effects.

uniformly distributed.

In this section, we use the subscripts i, c, and t to denote the birth, the county, and the trimester of birth, respectively.[33] The variable of interest is a binary variable that is coded 1 if there was a food stamp program in place three months before birth. Hence, the treatment is assigned to county-month cells, and in around 1% of cases, it also varies within groups.

We consider the following model separately for black and white mothers:

$$Q(\tau, bw_{ict}|fsp_{ct}, x_{1ict}, x_{2ct}, v_{ct}) = fsp_{ct}\gamma_1(\tau) + x_{1ict}\beta(\tau) + x_{2ct}\gamma_2(\tau) + \alpha(\tau, v_{ct}), \quad (36)$$

where $Q(\tau, bw_{ict}|fsp_{ct}, x_{1ict}, x_{2ct}, v_{ct})$ is the τth conditional quantile function of the outcome given all the variables. The variable fsp_{ct} indicates whether there is a food stamp program in place, x_{1ict} are variables related to the individual births, such as gender, mother age, and its square as well as the legitimacy status of the birth. Group-level control variables x_{2ct} include annual county-level controls (real per capita income, government transfers to individuals, medical spending, and retirement and disability payments) and 1960 county-level characteristics (county population and the shares of urban population, black population, and of farmland) interacted with a linear time trend. Further, x_{2j} also includes county, state-year, and time fixed effects.

For the estimation, we drop groups that have less than 25 degrees of freedom.[34] The estimations are performed using a sample of 2,822,091 individual observations divided into 19,482

---

[33]Using the same notation as in the paper, the j units are county-trimester combinations, and the i units index individual births within a county in a given trimester. In this section, we use three subscripts for clarity.

[34]If there are K_1 individual level variables we drop all groups with less than $K_1 + 1 + 25$ observations. Since some variables might vary at the individual level in some groups only, this threshold is group-specific.

groups for blacks and 16,038,235 individual births divided into 80,289 groups for whites.[35] Figure 2 illustrates the results. The results for black mothers are in the left panel, while the results for white mothers are in the right panel. The effect is substantially larger among black mothers. The results suggest a positive effect of the food stamp program on the lower tail of the conditional distribution. For instance, the estimates suggest that the food stamp program is associated with an increase in birth weight by almost 30 grams for blacks at the 5th percentile of the conditional distribution. For whites, there seems to be an effect only in the left tail of the distribution, and the effects are small. For blacks, the coefficients are large in the left tail and remain positive, albeit of small magnitude, until the 75% percentile. However, for higher quantiles, the effects are not statistically different from zero.

To compare our MD estimator with the CLP estimator on observational data, we also perform the estimation using the CLP procedure. For this comparison, we focus on the sample of black mothers. Before implementing the CLP estimator on this data, we need to solve a few issues that affect the CLP estimator but not the MD estimator. First, the treatment variable exhibits within-group variation in around 1% of the groups. If there are variables that vary only within some groups, the intercept is not comparable over groups, thus creating a problem for the CLP estimator. Hence, for the comparison, we drop groups for which the variable fsp is varying within the group. A second but similar issue is due to the covariate controlling for the legitimacy status of the birth, as there are groups that do not contain all levels of the categorical variable. To solve this issue, we drop the variable from the conditioning set. Third, for one observation (out of over 2 million), one of the group-level control variables is minimally different from the value assigned to other members of the group. This is most likely due to a coding error. In this case, we replace the values with the group average. While these features of the data do not pose any problem with our estimation approach, they can lead to substantially different results using the CLP estimation method. Figure 3 shows the estimation results performed with our approach compared to the CLP approach using the sample of black mothers. Given that we estimate a slightly different model on a marginally different sample, we re-estimate the effects using the MD estimator to ensure a meaningful comparison. The results obtained using the MD estimator remain almost identical to those in Figure 2. On the other hand, using the CLP estimator, we obtain substantially different results from those attained using the MD estimator, and the confidence intervals are, on average, 14 times larger, precluding any possibility of drawing informative conclusions. Further, the figure shows that the MD point estimates and confidence intervals are always inside the confidence intervals of the CLP estimator.

---

[35]We have a different number of groups compared to Almond et al. (2011) due to multiple reasons. First, they give higher weights to births in groups where only 50% of the births are coded in the natality data; thus, when they drop groups with less than 25 births, the number of births in these groups is inflated. Second, since they take the group average, they keep births with missing values for birth weight. We drop those births as we work with individual-level data.

Figure 3: Impact of Food Stamp Introduction estimated with the MD and CLP estimators
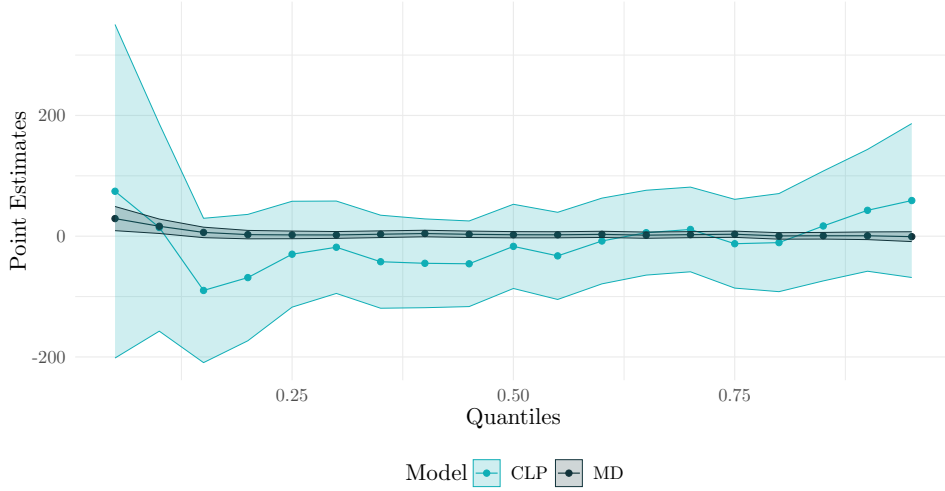
The figure shows the impact of the food stamp introduction on the conditional distribution of birth weight in the sample of black mothers computed with the MD (in black) and the CLP (in blue) estimators. The sample consists of 2,787,509 divided into 17,141 groups. The panels show point estimates and 95% confidence bands (shaded area) computed using standard errors clustered at the county level. For this estimation, we drop groups for which the treatment variables vary within groups and do not control for the legitimacy status of the birth.

7 Conclusion

This paper suggests a MD estimator for quantile panel data models. The estimator is of practical relevance with classical panel data settings where the units are observed over time and with grouped data, where individuals are divided into groups, and the treatment varies at the group level. The coefficient on the individual-level and group-level variables can be estimated. The estimator is computationally fast and straightforward to compute and consists of first-stage individual-level quantile regressions, followed by a GMM regression with the first-stage fitted values as the dependent variable. We show that our two-step procedure applied to linear estimators is algebraically identical to traditional one-step estimators. We suggest a quantile counterpart to traditional panel data estimators, including the pooled, the fixed effects, and the random effects estimators. In the second stage, both internal and external instruments can be used in a Hausman and Taylor or traditional instrumental variables framework. Further, an overidentification test can be implemented in overidentified models.

The asymptotic distribution of the estimator is non-standard because the speed of convergence is not the same for all coefficients. The speed of convergence depends on the moment conditions that are used to identify a parameter. For the coefficients converging at the faster $(\sqrt{mn})$ rate, only the variance coming from the first stage enters the first-order asymptotic distribution. On the other hand, since n diverges to infinity, the first stage variance does not appear in the first-order asymptotic variance of the coefficients converging at the slower $(\sqrt{m})$ rate. In

other words, the first-order asymptotic distribution is the same as if we knew the true first stage. We suggest an inference procedure that is uniformly valid regardless of the convergence rate of the estimator and, importantly, takes the first stage variance into account, thus providing better inference. Monte Carlo simulations show that our estimator and the suggested standard errors perform well in finite samples. Compared to the grouped estimator of Chetverikov et al. (2016), the MD estimator has a much smaller MSE. Finally, in an empirical application, we study the effect of the food stamp program on birth weight, and we find positive effects for blacks in the lower tail of the conditional distribution.

References

AHN, S. C. AND S. LOW (1996): "A reformulation of the Hausman test for regression models with pooled cross-section-time-series data," *Journal of Econometrics*, 71, 309–319.

AHN, S. C. AND H. R. MOON (2014): "Large-N and Large-T Properties of Panel Data Estimators and the Hausman Test," in *Festschrift in Honor of Peter Schmidt*, Springer.

ALMOND, D., H. W. HOYNES, AND D. W. SCHANZENBACH (2011): "Inside the war on poverty: The impact of food stamps on birth outcomes," *Review of Economics and Statistics*, 93, 387–403.

ALVAREZ, J. AND M. ARELLANO (2003): "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, 71, 1121–1159.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

ANGRIST, J. D. AND K. LANG (2004): "Does school integration generate peer effects? Evidence from Boston's metco program," *American Economic Review*, 94, 1613–1634.

ARELLANO, M. (1993): "On the testing of correlated effects with panel data," *Journal of Econometrics*, 59, 87–97.

ARELLANO, M. AND S. BONHOMME (2016): "Nonlinear panel data estimation via quantile regressions," *Econometrics Journal*, 19, C61–C94.

——— (2017): "Quantile Selection Models With an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85, 1–28.

AUTOR, D. H., D. DORN, AND G. H. HANSON (2013): "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 103, 2121–2168.

——— (2021): "When Work Disappears: Manufacturing Decline and the Falling Marriage Market Value of Young Men," *American Economic Review: Insights*, 1, 161–178.

AUTOR, D. H., A. MANNING, AND C. L. SMITH (2016): "The contribution of the minimum wage to US wage inequality over three decades: A reassessment," *American Economic Journal: Applied Economics*, 8, 58–99.

BALTAGI, B. H. (2021): *Econometric Analysis of Panel Data*, Springer, 6 ed.

BROWN, B. M. (1971): "Martingale central limit theorems," *The Annals of Mathematical Statistics*, 59–66.

CHAMBERLAIN, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

——— (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

——— (1994): "Quantile Regression, Censoring, and the Structure of Wages," *Advances in econometrics*, 1, 171–209.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. K. NEWEY (2013a): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013b): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

CHERNOZHUKOV, V. AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

CHESHER, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441.

CHETVERIKOV, D., B. LARSEN, AND C. PALMER (2016): "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade," *Econometrica*, 84, 809–833.

DAI, X. AND L. JIN (2021): "Minimum distance quantile regression for spatial autoregressive panel data models with fixed effects," *PLoS ONE*, 16, 1–13.

ENGELHARDT, G. V. AND P. J. PURCELL (2021): "The minimum wage and annual earnings inequality," *Economics Letters*, 207, 110001.

FERNÁNDEZ-VAL, I., W. Y. GAO, Y. LIAO, AND F. VELLA (2022): "Dynamic Heterogeneous Distribution Regression Panel Models, with an Application to Labor Income Processes," 1–45.

FRÖLICH, M. AND B. MELLY (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business and Economic Statistics*, 31, 346–357.

GALVAO, A. AND A. POIRIER (2019): "Quantile Regression Random Effects," *Annals of Economics and Statistics*, 109–148.

GALVAO, A. F., J. GU, AND S. VOLGUSHEV (2020): "On the unbiased asymptotic normality of quantile regression with fixed effects," *Journal of Econometrics*, 218, 178–215.

GALVAO, A. F. AND L. WANG (2015): "Efficient Minimum Distance Estimator for Quantile Regression Fixed Effects Panel Data," *Journal of Multivariate Analysis*, 133, 1–26.

HAHN, J. AND G. KUERSTEINER (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large," *Econometrica*, 70, 1639–1657.

HANSEN, B. (2022a): *Probability and Statistics for Economists*, Princeton University Press.

HANSEN, B. E. (2022b): *Econometrics*, Princeton University Press.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

HAUSMAN, J. AND W. E. TAYLOR (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377.

HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

IM, K. S., S. C. AHN, P. SCHMIDT, AND J. M. WOOLDRIDGE (1999): "Efficient estimation of panel data models with strictly exogenous explanatory variables," *Journal of Econometrics*, 93, 177–201.

KATO, K., A. GALVAO JR., AND G. V. MONTES-ROJAS (2012): "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics*, 170, 76–91.

KNIGHT, K. (1998): "Limiting Distributions for L1 Regression Estimators under General Conditions," *Annals of Statistics*, 26, 755–770.

KOENKER, R. (2004): "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91, 74–89.

KOENKER, R. AND G. BASSETT (1978): "Regression Quantiles," *Econometrica*, 46, 33.

LIAO, Y. AND X. YANG (2018): "Uniform Inference for Characteristic Effects of Large Continuous-Time Linear Models," 1–52.

LU, X. AND L. SU (2022): "Uniform inference in linear panel data models with two-dimensional heterogeneity," *Journal of Econometrics*.

MA, L. AND R. KOENKER (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134, 471–506.

MELLY, B. AND G. SANTANGELO (2015): "The changes-in-changes model with covariates," 1–32.

MUNDLAK, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

NERLOVE, M. (1971): "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, 39, 359.

NEWEY, W. K. (1993): "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics*, Elsevier, vol. 11, chap. 16, 419–454.

NEWEY, W. K. AND J. L. POWELL (1990): "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions," *Econometric Theory*, 6, 295–317.

PHILLIPS, P. C. AND H. R. MOON (1999): "Linear regression limit theory for nonstationary panel data," *Econometrica*, 67, 1057–1111.

PONS, M. (2024): "Quantile on Quantiles," *Working Paper*.

POWELL, J. L. (1991): "Estimation of Monotonic Regression Models under Quantile Restric-

tion," in *Nonparametric and Semiparametric Model in Economics*, Cambridge: Cambridge University Press.

SARGAN, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393–415.

VOLGUSHEV, S., S.-K. CHAO, AND G. CHENG (2019): "Distributed inference for quantile regression processes," *The Annals of Statistics*, 47, 1634–1662.

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

——— (2019): "Correlated random effects models with unbalanced panels," *Journal of Econometrics*, 211, 137–150.

A   Proofs

A.1   Proof of Lemma 1: Sampling error

Proof of lemma 1. Starting from the definition of the estimator we obtain

$$
\begin{aligned}
\hat{\delta}(\tau) &= \left(X'Z\hat{W}(\tau)Z'X\right)^{-1}X'Z\hat{W}(\tau)Z'\hat{Y}(\tau)\\
&= \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}\hat{\beta}_j(\tau)\\
&= \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\left(\tilde{x}'_{ij}\left(\hat{\beta}_j(\tau)-\beta_j(\tau)\right)+\tilde{x}'_{ij}\beta_j(\tau)\right)\\
&= \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\left(\tilde{x}'_{ij}\left(\hat{\beta}_j(\tau)-\beta_j(\tau)\right)+x'_{ij}\delta(\tau)+\alpha_j(\tau)\right)\\
&= \delta(\tau)+\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\left(\tilde{x}'_{ij}\left(\hat{\beta}_j(\tau)-\beta_j(\tau)\right)+\alpha_j(\tau)\right).
\end{aligned}
$$

∎

A.2   Proof of Theorem 1: Uniform consistency

A.2.1   Lemma 3

As a preliminary step to prove the uniform consistency of our estimator, we show uniform (in τ and j) consistency of the group-level quantile regressions.

Lemma 3 (Uniform consistency of $\hat{\beta}_j(\tau)$). *Under Assumptions 1-4 and 7(a), we have*

$$
\sup_{\tau\in\mathcal{T}}\max_{1\le j\le m}\|\hat{\beta}_j(\tau)-\beta_j(\tau)\|=o_p(1).
$$

Proof of Lemma 3. Angrist et al. (2006) show uniform consistency of the quantile regression estimator in τ but not in j (see their Theorem 3) while Galvao and Wang (2015) show uniform consistency in j but not in τ (see their Lemma 1). We show uniformity in both dimensions by following the steps of the proof in Galvao and Wang (2015) and extending it.

We define $\mathbb{Q}_{nj}(\beta,\tau)=\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_{ij}-\tilde{x}'_{ij}\beta)-\rho_\tau(y_{ij}-\tilde{x}'_{ij}\beta_j(\tau))$ and $Q_j(\beta,\tau)=\mathbb{E}_{i|j}[\rho_\tau(y_{ij}-\tilde{x}'_{ij}\beta)-\rho_\tau(y_{ij}-\tilde{x}'_{ij}\beta_j(\tau))]$. As shown in Angrist et al. (2006), the empirical processes $(\beta,\tau)\mapsto\mathbb{Q}_{nj}(\beta,\tau)$ for all groups j are stochastically equicontinuous because $|\mathbb{Q}_{nj}(\beta',\tau')-\mathbb{Q}_{nj}(\beta'',\tau'')|\le C_1\cdot|\tau'-\tau''|+C_2\cdot\|\beta'-\beta''\|$ where $C_1=2\cdot C\cdot\sup_{\beta\in\mathcal{B}}\|\beta\|$ for any compact set $\mathcal{B}$ and $C_2=2\cdot C$. The constant C is defined in Assumption 2. Note that C_1 and C_2 are neither functions of j nor τ.

Fix any $\zeta>0$. Let $B_j(\zeta,\tau)=\{\beta:\|\beta-\beta_j(\tau)\|\le\zeta\}$, the ball with center $\beta_j(\tau)$ and radius ζ. For each $\beta\notin B_j(\zeta,\tau)$, define $\tilde{\beta}=r_j\beta+(1-r_j)\beta_j(\tau)$ where $r_j=\frac{\zeta}{\|\beta-\beta_j(\tau)\|}$. So $\tilde{\beta}\in\partial B_j(\zeta,\tau)=\{\beta:\|\beta-\beta_j(\tau)\|=\zeta\}$, the boundary of $B_j(\zeta,\tau)$. Since $\mathbb{Q}_{nj}(\beta,\tau)$ is convex in

β for all τ, and $\mathbb{Q}_{nj}(\beta_j(\tau), \tau) = 0$, we have

$$r_j \mathbb{Q}_{nj}(\beta, \tau) \geq \mathbb{Q}_{nj}(\tilde{\beta}, \tau) = Q_j(\tilde{\beta}, \tau) + \mathbb{Q}_{nj}(\tilde{\beta}, \tau) - Q_j(\tilde{\beta}, \tau) > \epsilon_\zeta + \mathbb{Q}_{nj}(\tilde{\beta}, \tau) - Q_j(\tilde{\beta}, \tau) \quad (37)$$

uniformly in j and τ, where

$$\epsilon_\zeta = \inf_{\tau \in \mathcal{T}} \inf_{1 \leq j \leq m} \inf_{\|\beta - \beta_j(\tau)\| = \zeta} \mathbb{E}_{i|j} \left[\int_0^{\tilde{x}'_{ij}(\beta - \beta_j(\tau))} \left(1(y_{ij} - \tilde{x}'_{ij}\beta_j(\tau) \leq s) - 1(y_{ij} - \tilde{x}'_{ij}\beta_j(\tau) \leq 0) \right) ds \right]$$

by the identity of Knight (1998) and $\epsilon_\zeta > 0$ by Assumptions 3 and 4.

Thus, we have the following

$$\left\{ \sup_{\tau \in \mathcal{T}} \max_{1 \leq j \leq m} \|\hat{\beta}_j(\tau) - \beta_j(\tau)\| > \zeta \right\} \overset{(a)}{\subseteq} \{\exists \tau_j \in \mathcal{T}, \exists \beta_j \notin B_j(\zeta, \tau_j) : \mathbb{Q}_{nj}(\beta_j, \tau_j) \leq 0\}$$

$$\overset{(b)}{\subseteq} \cup_{j=1}^m \left\{ \sup_{\tau \in \mathcal{T}} \sup_{\beta_j \in B_j(\zeta, \tau_j)} |\mathbb{Q}_{nj}(\beta_j, \tau_j) - Q_j(\beta_j, \tau_j)| \geq \epsilon_\zeta \right\}$$

Relation (a) holds because, by definition, $\hat{\beta}_j(\tau)$ minimizes $\mathbb{Q}_{nj}(\beta, \tau)$, and $\mathbb{Q}_{nj}(\beta_j(\tau), \tau) = 0$. Relation (b) holds by the rightmost inequality of line (37). Then, it follows that

$$P\left\{ \sup_{\tau \in \mathcal{T}} \max_{1 \leq j \leq m} \|\hat{\beta}_j(\tau) - \beta_j(\tau)\| > \zeta \right\} \leq P\left\{ \cup_{j=1}^m \left\{ \sup_{\tau \in \mathcal{T}} \sup_{\beta_j \in B_j(\zeta, \tau)} |\mathbb{Q}_{nj}(\beta_j, \tau) - Q_j(\beta_j, \tau)| \geq \epsilon_\zeta \right\} \right\}$$

$$\leq \sum_{j=1}^m P\left\{ \sup_{\tau \in \mathcal{T}} \sup_{\beta_j \in B_j(\zeta, \tau)} |\mathbb{Q}_{nj}(\beta_j, \tau) - Q_j(\beta_j, \tau)| \geq \epsilon_\zeta \right\}$$

$$\leq m \max_{1 \leq j \leq m} P\left\{ \sup_{\tau \in \mathcal{T}} \sup_{\beta_j \in B_j(\zeta, \tau)} |\mathbb{Q}_{nj}(\beta_j, \tau) - Q_j(\beta_j, \tau)| \geq \epsilon_\zeta \right\}$$

Therefore, if we can show that

$$\max_{1 \leq j \leq m} P\left\{ \sup_{\tau \in \mathcal{T}} \sup_{\beta_j \in B_j(\zeta, \tau)} |\mathbb{Q}_{nj}(\beta_j, \tau) - Q_j(\beta_j, \tau)| \geq \epsilon_\zeta \right\} = o\left(\frac{1}{m} \right)$$

the proof of the lemma will be completed.

Without loss of generality, we assume $\beta_j(\tau) = 0$ for all j and $\tau \in \mathcal{T}$. Then the balls $B_j(\zeta, \tau)$ for all j and $\tau \in \mathcal{T}$ are identical and we denote them by $B(\delta)$. Because the closed ball $B(\zeta)$ is compact, there exist K balls with center β^k, $k = 1, ..., K$, and radius $\frac{\epsilon_\zeta}{3C_2}$ such that the collection of them covers $B(\zeta)$. Since $\epsilon_\zeta > 0$, we can find a finite K that satisfies this condition and is independent of j and τ. Therefore, for any $\beta \in B(\delta)$, there is some $k \in \{1, ..., K\}$ such that

$$|\mathbb{Q}_{nj}(\beta, \tau) - Q_j(\beta, \tau)| - |\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| \leq |\mathbb{Q}_{nj}(\beta, \tau) - Q_j(\beta, \tau) - \mathbb{Q}_{nj}(\beta^k, \tau) + Q_j(\beta^k, \tau)|$$

$$\leq |\mathbb{Q}_{nj}(\beta, \tau) - \mathbb{Q}_{nj}(\beta^k, \tau)| + |Q_j(\beta, \tau) - Q_j(\beta^k, \tau)|$$

$$\leq C_2 \frac{\epsilon_\zeta}{3C_2} + C_2 \frac{\epsilon_\zeta}{3C_2} = \frac{2\epsilon_\zeta}{3}$$

uniformly in j and $\tau \in \mathcal{T}$. The third line is justified by the stochastic equicontinuity of $\mathbb{Q}_{nj}(\beta, \tau)$.

It then follows that,

$$\sup_{\tau \in \mathcal{T}} \sup_{\beta \in B(\zeta)} |\mathbb{Q}_{nj}(\beta, \tau) - Q_j(\beta, \tau)| \le \sup_{\tau \in \mathcal{T}} \max_{1 \le k \le K} |\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| + \frac{2\epsilon_\zeta}{3}$$

and

$$\begin{aligned}
\mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \sup_{\beta \in B(\delta)} |\mathbb{Q}_{nj}(\beta, \tau) - Q_j(\beta, \tau)| > \epsilon_\zeta\right\} &\le \mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \max_{1 \le k \le K} |\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| + \frac{2\epsilon_\zeta}{3} > \epsilon_\zeta\right\} \\
&= \mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \max_{1 \le k \le K} |\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| > \frac{\epsilon_\zeta}{3}\right\} \\
&\le \sup_{\tau \in \mathcal{T}} \sum_{k=1}^K \mathrm{P}\left\{|\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| > \frac{\epsilon_\zeta}{3}\right\}
\end{aligned}$$

For each $\tau \in \mathcal{T}$, $\mathbb{Q}_{nj}(\beta^k, \tau)$ is the sample mean of n i.i.d. terms bounded in absolute values by $2 \cdot C \cdot \zeta$. By Hoeffding's inequality, it follows that

$$\begin{aligned}
\sum_{k=1}^K \mathrm{P}\left\{|\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| > \frac{\epsilon_\zeta}{3}\right\} &\le 2K \exp\left\{-\frac{2n\epsilon_\zeta^2}{3^2 2^2 C^2 \zeta^2}\right\} \\
&= 2K \exp\left\{-\frac{n\epsilon_\zeta^2}{18 C^2 \zeta^2}\right\} \\
&= O(\exp(-n))
\end{aligned}$$

Since $\frac{\log m}{n} \to 0$ by Assumption 7(a), it follows that $O_p(\exp(-n)) = o_p(1/m)$. Note that the bound $\epsilon_\zeta/3 > 0$ is uniform in τ, and K is finite. By stochastic equicontinuity of $\mathbb{Q}_{nj}(\beta, \tau)$, as $n \to \infty$, uniformly in $\tau \in \mathcal{T}$,

$$\mathbb{Q}_{nj}(\beta^k, \tau) = Q_j(\beta^k, \tau) + o_p(1)$$

It follows that

$$\sup_{\tau \in \mathcal{T}} \sum_{k=1}^K \mathrm{P}\left\{|\mathbb{Q}_{nj}(\beta^k, \tau) - Q_j(\beta^k, \tau)| > \frac{\epsilon_\zeta}{3}\right\} = o_p\left(\frac{1}{m}\right)$$

∎

A.2.2 Lemma 4

Lemma 4 (Uniform consistency of $\hat{G}(\tau)$ with a full-rank weighting matrix). *Assumptions 1, 2, 5, and 8 hold. Then,*

$$\sup_{\tau \in \mathcal{T}} \left\| \left(S'_{ZX} \hat{W}(\tau) S_{ZX}\right)^{-1} S'_{ZX} \hat{W}(\tau) - \left(\Sigma'_{ZX} W(\tau) \Sigma'_{ZX}\right)^{-1} \Sigma'_{ZX} W(\tau) \right\| = o_p(1) \qquad (38)$$

and $\left(\Sigma'_{ZX} W(\tau) \Sigma'_{ZX}\right)^{-1} \Sigma'_{ZX} W(\tau)$ is uniformly bounded and continuous.

Proof of Lemma 4. First, it follows from Assumptions 1(ii), 2(i) and 5(i) that $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n z_{ij} x'_{ij}\right) = o_p\left(\frac{1}{n}\right)$ and $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n z_{ij} x'_{ij}\right] = \mathbb{E}[z_{ij} x'_{ij}]$. Hence, by Assumption 1(i), $\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^m \frac{1}{n}\sum_{i=1}^n x_{ij} z'_{ij}\right) =$

$o_p\left(\frac{1}{mn}\right)$. By Chebyshev's inequality,

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}-\mathbb{E}[z_{ij}x'_{ij}]\right)\underset{p}{\to}0.$$

In addition, by Assumption 5(iii), $m^{-1}\sum_{j=1}^{m}\mathbb{E}[z_{ij}x'_{ij}]\to\Sigma_{ZX}$. It follows that

$$S_{ZX}\underset{p}{\to}\Sigma_{ZX}$$

By assumption 8, uniformly in $\tau\in\mathcal{T}$, $\hat{W}(\tau)\underset{p}{\to}W(\tau)$ where $W(\tau)$ is uniformly continuous and strictly positive definite. Since Σ_{ZX} is bounded and has full column rank, it implies that $\Sigma'_{ZX}W(\tau)\Sigma_{ZX}$ is also uniformly continuous and invertible. The result of the lemma follows. $\blacksquare$

A.2.3 Theorem 1

Proof of Theorem 1. By Lemma 1

$$\hat{\delta}(\tau)-\delta(\tau)=\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\left(\tilde{x}'_{ij}\left(\hat{\beta}_j(\tau)-\beta_j(\tau)\right)+\alpha_j(\tau)\right).$$

By Lemma 4, the first factor converges uniformly to $G(\tau)$:

$$\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\underset{p}{\to}\left(\Sigma'_{ZX}W(\tau)\Sigma'_{ZX}\right)^{-1}\Sigma'_{ZX}W(\tau)=G(\tau)\tag{39}$$

By lemma 3, $\hat{\beta}_j(\tau)$ is consistent for $\beta_j(\tau)$ uniformly in j and τ. Together with the boundedness of x_{ij} in Assumption 2(i) and of z_{ij} in Assumption 5(i), it follows that

$$\sup_{\tau\in\mathcal{T}}\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau))\underset{p}{\to}0.\tag{40}$$

By Assumption 5(ii), $\mathbb{E}[z_{ij}\alpha_j(\tau)]=0$ uniformly in τ. By Assumption 5, $\text{Var}(z_{ij}\alpha_j(\tau))$ is uniformly bounded. In addition, z_{ij} is bounded and $\alpha_j(\tau)$ is uniformly continuous in τ. Hence,

$$\sup_{\tau\in\mathcal{T}}\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau)\underset{p}{\to}0.\tag{41}$$

The result of the Theorem follows from equations 39, 40, and 41. $\blacksquare$

A.3 Proof of Lemma 2 - Asymptotic distribution of sample moments

Proof of Lemma 2. **Part (i)**

Lemma 3 in Galvao et al. (2020) provides the uniform Bahadur representation for the group-level quantile regression coefficient under our assumptions:

$$\hat{\beta}_j(\tau)-\beta_j(\tau)=\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij},y_{ij})+R^{(1)}_{nj}(\tau)+R^{(2)}_{nj}(\tau),\tag{42}$$

where

$$\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) = -B_{j,\tau}^{-1}\tilde{x}_{ij}(1(y_{ij} \leq \tilde{x}_{ij}\beta_j(\tau)) - \tau), \tag{43}$$

with $B_{j,\tau} = \mathbb{E}_{i|j}[f_{y|x}(Q_{y|x,\nu_j}(\tau|\tilde{x}_{ij})|\tilde{x}_{ij})\tilde{x}_{ij}\tilde{x}_{ij}']$ and

$$\sup_j \sup_{\tau \in \mathcal{T}} \left\| R_{nj}^{(2)}(\tau) \right\| = O_p\left(\frac{\log n}{n}\right), \tag{44}$$

$$\sup_j \sup_{\tau \in \mathcal{T}} \left\| \mathbb{E}_{i|j}\left[R_{nj}^{(1)}(\tau) \right] \right\| = O\left(\frac{\log n}{n}\right), \tag{45}$$

$$\sup_j \sup_{\tau \in \mathcal{T}} \left\| \mathbb{E}_{i|j}\left[\left(R_{nj}^{(1)}(\tau) - \mathbb{E}_{i|j}[R_{nj}^{(1)}(\tau)] \right) \left(R_{nj}^{(1)}(\tau) - \mathbb{E}_{i|j}[R_{nj}^{(1)}(\tau)] \right)' \right] \right\| = O\left(\left(\frac{\log n}{n}\right)^{3/2}\right). \tag{46}$$

It follows that

$$\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\left(\hat{\beta}_j(\tau) - \beta_j(\tau)\right) = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) \tag{47}$$

$$+ \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(1)}(\tau) \tag{48}$$

$$+ \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(2)}(\tau). \tag{49}$$

Consider first the third term (49). By assumptions 2(i) and 5(i), x_{ij} and z_{ij} are bounded by C such that the sample mean of their product is also bounded. Therefore, (44) implies that

$$\sup_{\tau \in \mathcal{T}} \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(2)}(\tau) = O_p\left(\frac{\log n}{n}\right) \tag{50}$$

Consider now the second term (48). Since $\mathrm{Var}\left(R_{nj}^{(1)}(\tau)\right) = o\left(\frac{1}{n}\right)$ by (46), x_{ij} and z_{ij} are bounded by assumptions 2(i) and 5(i), and observations are independent across groups, it follows that $\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(1)}(\tau)\right) = o_p\left(\frac{1}{mn}\right)$. In addition, by (45), $\sup_j \sup_{\tau \in \mathcal{T}} \mathbb{E}_{i|j}\left[R_{nj}^{(1)}(\tau) \right] = O\left(\frac{\log n}{n}\right)$ such that $\sup_{\tau \in \mathcal{T}} \mathbb{E}_{i|j}\left[\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(1)}(\tau)\right] = O\left(\frac{\log n}{n}\right)$. Putting this together, by the Chebyshev inequality and under Assumption 7(c),

$$\sup_{\tau \in \mathcal{T}} \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right) R_{nj}^{(1)}(\tau) = o_p\left(\frac{1}{\sqrt{mn}}\right) \tag{51}$$

It follows that both remainder terms are $o_p\left(\frac{1}{\sqrt{mn}}\right)$ uniformly over τ.

Consider now the term (47). Let $\Sigma_{ZXj} = \mathbb{E}_{i|j}[z_{ij}\tilde{x}_{ij}']$.

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) =$$

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}\tilde{x}_{ij}' - \Sigma_{ZXj}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) + \frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) \tag{52}$$

49

By the boundedness of z_{ij} and x_{ij} and the independence of the observations over time, it follows that $\left\|\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}' - \Sigma_{ZXj}\right\| = o(1)$ uniformly in j. In addition, $\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) = O\left(\frac{1}{n}\right)$. Hence,

$$\text{Var}\left(\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}' - \Sigma_{ZXj}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{i,\tau}(\tilde{x}_{ij}, y_{ij})\right)\right) = o\left(\frac{1}{mn}\right)$$

The model in equation (1) and Assumption 5(iii) imply that $\mathbb{E}_{i|j}\left[1(y_{ij} \leq \tilde{x}_{ij}\beta_j(\tau))|\tilde{x}_{ij}, z_{ij}\right] = \tau$, which implies that

$$\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}' - \Sigma_{ZXj}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right)\right] = 0$$

uniformly in τ. Therefore, by Chebyshev inequality,

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}' - \Sigma_{ZXj}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) = o_p\left(\frac{1}{\sqrt{mn}}\right) \quad (53)$$

uniformly in τ.

Since all other terms are $o_p\left(\frac{1}{\sqrt{mn}}\right)$ uniformly over τ, the limiting distribution of the process $\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}'\left(\hat{\beta}_j(\tau) - \beta_j(\tau)\right)$ is the same as the limiting distribution of

$$\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right)$$

$$= \frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{-B_{j,\tau}^{-1}}{n}\sum_{i=1}^{n}\tilde{x}_{ij}(1(y_{ij} \leq \tilde{x}_{ij}\beta_j(\tau)) - \tau)\right) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}s_{ij}(\tau) \quad (54)$$

This is a sample mean over mn independent (but not necessarily identically distributed) observations denoted by $s_{ij}(\tau)$. The model in equation (1) and Assumption 5(iii) imply that $\mathbb{E}\left[1(y_{ij} \leq \tilde{x}_{ij}\beta_j(\tau))|\tilde{x}_{ij}, z_{ij}, v_j\right] = \tau$, which implies that $\mathbb{E}[s_{ij}(\tau)] = 0$. In addition,

$$\text{Var}(s_{ij}(\tau)) = \mathbb{E}_j[\Sigma_{ZXj}\text{Var}(\phi_{j,\tau})\Sigma_{ZXj}'] = \mathbb{E}_j[\Sigma_{ZXj}B_{j,\tau}^{-1}\tau(1-\tau)\mathbb{E}_{i|j}[\tilde{x}_{ij}\tilde{x}_{ij}']B_{j,\tau}^{-1}\Sigma_{ZXj}'] \quad (55)$$

Pointwise asymptotic normality follows by an application of the Lindeberg CLT.

Next we note that $\left\{\Sigma_{ZXj}\left(\frac{-B_{j,\tau}^{-1}}{n}\sum_{i=1}^{n}\tilde{x}_{ij}(1(y_{ij} \leq \tilde{x}_{ij}\beta) - \tau)\right), \tau \in \mathcal{T}, \beta \in \mathcal{B}\right\}$ is a Donsker class for any compact set $\mathcal{B}$. This follows by noting that $\{1(y_{ij} \leq \tilde{x}_{ij}\beta_j(\tau)), \tau \in \mathcal{T}, \beta \in \mathcal{B}\}$ is a VC subgraph class and hence a bounded Donsker class. Hence,

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ij}(1(y_{ij} \leq \tilde{x}_{ij}\beta) - \tau), \tau \in \mathcal{T}, \beta \in \mathcal{B}\right\}$$

is also bounded Donsker with a square-integrable envelope $2 \cdot \max_{i \in 1,\ldots,n}\|\tilde{x}_{ij}\| \leq 2 \cdot C$. The whole function is then Donsker by the boundedness of Σ_{ZXj} and $B_{j,\tau}^{-1}$. The weak convergence result

50

follows by application of the functional central limit theorem for independent but not identically distributed random variables, see for instance Theorem 3 in Brown (1971).

Part (ii) follows directly by Lemma 3 in Chetverikov et al. (2016).

Part (iii) The first moment is asymptotically equivalent to (up to a term, which is uniformly $o_p\left(\frac{1}{\sqrt{mn}}\right)$)

$$\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{-B_{j,\tau}^{-1}}{n}\sum_{i=1}^{n}\tilde{x}_{ij}(1(y_{ij}\leq\tilde{x}_{ij}\beta_j(\tau))-\tau)\right).$$

We have already shown that both moments have mean zero. By Assumption 1, the observations are independent across i and j such that we only need to consider the correlation between both moments for the same individual and group.

$$\begin{aligned}\mathrm{Cov}(\tilde{x}_{ij}&(1(y_{ij}\leq\tilde{x}_{ij}\beta_j(\tau))-\tau),z_{ij}\alpha_j(\tau'))\\&=\mathbb{E}[\tilde{x}_{ij}(1(y_{ij}\leq\tilde{x}_{ij}\beta_j(\tau))-\tau)z'_{ij}\alpha_j(\tau')]\\&=\mathbb{E}[\tilde{x}_{ij}\mathbb{E}_{i|j}[(1(y_{ij}\leq\tilde{x}_{ij}\beta_j(\tau))-\tau)|x_{ij},z_{ij}]z'_{ij}\alpha_j(\tau)]=0\end{aligned}$$

It follows that

$$\sup_{\tau,\tau'\in\mathcal{T}}\left\|\mathrm{Cov}\left(\bar{g}_{mn}^{(1)}(\hat{\delta},\tau),\bar{g}_{mn}^{(2)}(\hat{\delta},\tau')\right)\right\|=o_p\left(\frac{1}{\sqrt{mn}}\right)$$

∎

A.4 Lemma 5: Uniform consistency of the G matrix with a heterogeneous weighting matrix

Lemma 5. *Assumptions 1, 2, 5, and 8′ hold. Then, uniformly in τ, we can partition $\hat{G}(\tau)$ into $\hat{G}_{11}(\tau)$ (with dimensions $M_1\times L_1$), $\hat{G}_{12}(\tau)$ ($M_1\times L_2$), $\hat{G}_{21}(\tau)$ ($M_2\times L_1$), and $\hat{G}_{22}(\tau)$ ($M_2\times L_2$) such that*

$$\hat{G}(\tau)=\begin{pmatrix}\hat{G}_{11}(\tau)&\hat{G}_{12}(\tau)\\\hat{G}_{21}(\tau)&\hat{G}_{22}(\tau)\end{pmatrix}=\begin{pmatrix}G_{mn,11}(\tau)&G_{mn,12}(\tau)\\G_{mn,21}(\tau)&G_{mn,22}(\tau)\end{pmatrix}+\begin{pmatrix}o_p(1)&o_p\left(\sqrt{a_n(\tau)}\right)\\o_p\left(1/\sqrt{a_n(\tau)}\right)&o_p(1)\end{pmatrix}$$

where $G_{mn}(\tau)$ is defined in equation (21), $\sup_{\tau\in\mathcal{T}}G_{mn,12}(\tau)=O_p(a_n(\tau))$ and the other elements of $G_{mn}(\tau)$ are $O_p(1)$ uniformly in τ.

Proof. To simplify the notation, we omit the dependency on τ. We partition S'_{ZX} into four submatrices that have the same dimensions as the submatrices of $\hat{G}$. As in proposition 3, we also partition $\hat{W}$ into four submatrices such that the diagonal submatrices are square matrices of dimensions L_1 and L_2. Note that

$$S'_{ZX}\hat{W}=\begin{pmatrix}S'_{11}&S'_{21}\\0&S'_{22}\end{pmatrix}\begin{pmatrix}\hat{W}_{11}&\hat{W}_{12}\\\hat{W}_{21}&\hat{W}_{22}\end{pmatrix}=\begin{pmatrix}S'_{11}\hat{W}_{11}+S'_{21}\hat{W}_{21}&S'_{11}\hat{W}_{12}+S'_{21}\hat{W}_{22}\\S'_{22}\hat{W}_{21}&S'_{22}\hat{W}_{22}\end{pmatrix}$$

Furthermore,

$$S'_{ZX}\hat{W}S_{ZX} = \begin{pmatrix} S'_{11} & S'_{21} \\ 0 & S'_{22} \end{pmatrix} \begin{pmatrix} \hat{W}_{11} & \hat{W}_{12} \\ \hat{W}_{21} & \hat{W}_{22} \end{pmatrix} \begin{pmatrix} S_{11} & 0 \\ S_{21} & S_{22} \end{pmatrix}$$

$$= \begin{pmatrix} \hat{A}_{11} + \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{pmatrix}$$

where $\hat{A}_{11} = S'_{11}\hat{W}_{11}S_{11}$, $\hat{B}_{11} = S'_{11}\hat{W}_{12}S_{21} + S'_{21}\hat{W}_{21}S_{11} + S'_{21}\hat{W}_{22}S_{21}$, $\hat{B}_{12} = S'_{11}\hat{W}_{12}S_{22} + S'_{21}\hat{W}_{22}S_{22}$, $\hat{B}_{21} = S'_{22}\hat{W}_{21}S_{11} + S'_{22}\hat{W}_{22}S_{21}$, and $\hat{B}_{22} = S'_{22}\hat{W}_{22}S_{22}$.

For the non-zero elements of the matrix S_{ZX}, we have

$$S_{11} - \Sigma_{11} = o_p(1),$$
$$S_{22} - \Sigma_{22} = o_p(1),$$
$$S_{21} - \Sigma_{21} = o_p(1).$$

Together with proposition 3, this implies that uniformly over τ,

$$\hat{A}_{11} = \Sigma'_{11}W_{11}\Sigma_{11} + o_p(1), \tag{56}$$

$$\hat{B}_{11} = a_n\Sigma'_{11}W_{12}\Sigma_{21} + a_n\Sigma'_{21}W_{21}\Sigma_{11} + a_n\Sigma'_{21}W_{22}\Sigma_{21} + o_p(\sqrt{a_n}) = a_nB_{11} + o_p(\sqrt{a_n}), \tag{57}$$

$$\hat{B}_{12} = a_n\Sigma'_{11}W_{12}\Sigma_{22} + a_n\Sigma'_{21}W_{22}\Sigma_{22} + o_p(\sqrt{a_n}) = a_nB_{12} + o_p(\sqrt{a_n}), \tag{58}$$

$$\hat{B}_{21} = a_n\Sigma'_{22}W_{21}\Sigma_{11} + a_n\Sigma'_{22}W_{22}\Sigma_{21} + o_p(\sqrt{a_n}) = a_nB_{21} + o_p(\sqrt{a_n}), \tag{59}$$

$$\hat{B}_{22} = a_n\Sigma'_{22}W_{22}\Sigma_{22} + o_p(a_n) = a_nB_{22} + o_p(a_n). \tag{60}$$

By the inverse of a partitioned matrix

$$\left(S'_{ZX}\hat{W}S_{ZX}\right)^{-1} = \begin{pmatrix} \hat{C} + \hat{C}\hat{B}_{12}\hat{D}\hat{B}_{21}\hat{C} & -\hat{C}\hat{B}_{12}\hat{D} \\ -\hat{D}\hat{B}_{21}\hat{C} & \hat{D} \end{pmatrix}$$

where $\hat{C} = \left(\hat{A}_{11} + \hat{B}_{11}\right)^{-1}$ and $\hat{D} = \left(\hat{B}_{22} - \hat{B}_{21}\hat{C}\hat{B}_{12}\right)^{-1}$. Hence uniformly over τ,

$$\hat{C} = C + o_p(1), \tag{61}$$
$$\hat{D} = a_n^{-1}D + o_p\left(a_n^{-1}\right), \tag{62}$$
$$\hat{D}\hat{B}_{21}\hat{C} = DB_{21}C + o_p(a_n^{-1/2}), \tag{63}$$

where C and D are strictly positive and bounded. Equation (62) holds because

$$a_n\hat{D} = \left(a_n^{-1}\hat{B}_{22} - a_n^{-1}\hat{B}_{21}\hat{C}\hat{B}_{12}\right)^{-1} = (B_{22} - a_nB_{21}CB_{12})^{-1} + o_p(1) = D + o_p(1).$$

Combining equation (59) with equations (61)-(62), we obtain similarly

$$\hat{C}\hat{B}_{12}\hat{D} = CB_{12}D + o_p(a_n^{-1/2}). \tag{64}$$

Now, we can consider each submatrix of $\hat{G}$ separately and derive their convergence rate to the corresponding element of the G_{mn} matrix defined in equation (21):

$$G_{mn}(\tau) = \left(\Sigma'_{ZX}W_{mn}(\tau)\Sigma_{ZX}\right)^{-1}\Sigma'_{ZX}W_{mn}(\tau).$$

For the upper left term, we have

$$\begin{aligned}
\hat{G}_{11} &= (\hat{C} + \hat{C}\hat{B}_{12}\hat{D}\hat{B}_{21}\hat{C})(S'_{11}\hat{W}_{11} + S'_{21}\hat{W}_{21}) - \hat{C}\hat{B}_{12}\hat{D}S'_{22}\hat{W}_{21} \\
&= (C + a_n CB_{12}DB_{21}C + o_p(1))\left(\Sigma'_{11}W_{11} + a_n\Sigma'_{21}W_{21} + o_p(1)\right) \\
&\quad + \left(CB_{12}D + o_p(a_n^{-1/2})\right)\left(a_n\Sigma'_{22}W_{21} + o_p(\sqrt{a_n})\right) \\
&= (C + a_n CB_{12}DB_{21}C)(\Sigma'_{11}W_{11} + a_n\Sigma'_{21}W_{21}) - a_n CB_{12}D\Sigma'_{22}W_{21} + o_p(1) \\
&= G_{mn,11} + o_p(1),
\end{aligned}$$

uniformly over τ.

For the upper right term, we have uniformly over τ,

$$\begin{aligned}
\hat{G}_{12} &= (\hat{C} + \hat{C}\hat{B}_{12}\hat{D}\hat{B}_{21}\hat{C})\left(S'_{11}\hat{W}_{12} + S'_{21}\hat{W}_{22}\right) - \hat{C}\hat{B}_{12}\hat{D}S'_{22}\hat{W}_{22} \\
&= (C + a_n CB_{12}DB_{21}C + o_p(1))\left(a_n\left(\Sigma'_{11}W_{12} + \Sigma'_{21}W_{22}\right) + o_p(\sqrt{a_n})\right) \\
&\quad - \left(CB_{12}D + o_p(a_n^{-1/2})\right)\left(a_n\Sigma'_{22}W_{22} + o_p(a_n)\right) \\
&= a_n(C + a_n CB_{12}DB_{21}C)\left(\Sigma'_{11}W_{12} + \Sigma'_{21}W_{22}\right) - a_n CB_{12}D\Sigma'_{22}W_{22} + o_p(\sqrt{a_n}) \\
&= G_{mn,12} + o_p(\sqrt{a_n}).
\end{aligned}$$

For the lower right term,

$$\begin{aligned}
\hat{G}_{22} &= -\hat{D}\hat{B}_{21}\hat{C}\left(S'_{11}\hat{W}_{12} + S'_{21}\hat{W}_{22}\right) + \hat{D}S'_{22}\hat{W}_{22} \\
&= -\left(DB_{21}C + o_p(a_n^{-1/2})\right)\left(a_n\Sigma'_{11}W_{12} + a_n\Sigma'_{21}W_{22} + o_p(\sqrt{a_n})\right) \\
&= -a_n DB_{21}C\left(\Sigma'_{11}W_{12} + \Sigma'_{21}W_{22}\right) + D\Sigma'_{22}W_{22} + o_p(1) \\
&= G_{mn,22} + o_p(1),
\end{aligned}$$

uniformly over τ.

Finally, for the lower left term, uniformly of τ,

$$\begin{aligned}
\hat{G}_{21} &= -\hat{D}\hat{B}_{21}\hat{C}(S'_{11}\hat{W}_{11} + S'_{21}\hat{W}_{21}) + \hat{D}S'_{22}\hat{W}_{21} \\
&= -\left(DB_{21}C + o_p(a_n^{-1/2})\right)\left(\Sigma'_{11}W_{11} + a_n\Sigma'_{21}W_{21} + o_p(\sqrt{a_n})\right) \\
&\quad + \left(a_n^{-1}D + o_p(a_n^{-1})\right)\left(\Sigma'_{22}a_n W_{21} + o_p(\sqrt{a_n})\right) \\
&= -DB_{21}C\left(\Sigma'_{11}W_{11} - \Sigma'_{21}a_n W_{21}\right) + D\Sigma'_{22}W_{21} + o_p(a_n^{-1/2}) \\
&= G_{mn,21} + o_p\left(a_n^{-1/2}\right).
\end{aligned}$$

Hence,

$$\sup_\tau \left\|\hat{G}(\tau) - G_{mn}(\tau)\right\| = \begin{pmatrix} o_p(1) & o_p\left(\sqrt{a_n(\tau)}\right) \\ o_p\left(1/\sqrt{a_n(\tau)}\right) & o_p(1) \end{pmatrix}$$

where $G_{mn,12}(\tau) = O_p(a_n(\tau))$ and the other elements of $G_{mn}(\tau)$ are $O_p(1)$ uniformly over τ.

■

A.5 Proof of Theorem 2: Degree of heterogeneity is known

Proof of Theorem 2. From the definition of the estimator,

$$\hat{\delta}(\tau) - \delta(\tau) = \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau)\bar{g}_{mn}(\hat{\delta}, \tau) \tag{65}$$

In part (i), Lemma 5 applies uniformly in τ with $a_n(\tau) = O_p(1/n)$ such that

$$\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau) = \begin{pmatrix} \hat{G}_{11}(\tau) & \hat{G}_{12}(\tau) \\ \hat{G}_{21}(\tau) & \hat{G}_{22}(\tau) \end{pmatrix}$$
$$= \begin{pmatrix} G_{11}(\tau) + o_p(1) & o_p(1/\sqrt{n}) \\ o_p(\sqrt{n}) & G_{22}(\tau) + o_p(1) \end{pmatrix}.$$

By the definition of the fast instruments, the first L_1 elements of $\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)$ are equal to zero. It follows from this fact and Lemma 2(i) that

$$\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}_{11}(\cdot) \tag{66}$$

where $\bar{g}_{mn,1}(\hat{\delta}, \cdot)$ contains the first L_1 elements of $\bar{g}_{mn}(\hat{\delta}, \cdot)$. For the remaining L_2 elements, we have $\sqrt{m}\bar{g}_{mn,2}(\hat{\delta}, \cdot)O_p(1)$. It follows that

$$\begin{aligned}
\sqrt{mn}\left(\hat{\delta}_1(\cdot) - \delta_1(\cdot)\right) &= \hat{G}_{11}\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + \hat{G}_{12}\bar{g}_{mn,2}(\hat{\delta}, \cdot) \\
&= G_{11}\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + o_p(1)\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + o_p(1/\sqrt{n})\sqrt{mn}\bar{g}_{mn,2}(\hat{\delta}, \cdot) \\
&= G_{11}\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + o_p(1)\sqrt{mn}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + o_p(1)\sqrt{m}\bar{g}_{mn,2}(\hat{\delta}, \cdot) \\
&\rightsquigarrow G_{11}\mathbb{Z}_{11}(\cdot) + o_p(1),
\end{aligned}$$

which proves part (i)-(a) of the theorem.

For the remaining M_2 coefficients, applying the same results, we obtain

$$\begin{aligned}
\sqrt{m}\left(\hat{\delta}_2(\cdot) - \delta_2(\cdot)\right) &= \hat{G}_{21}\sqrt{m}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + \hat{G}_{22}\sqrt{m}\bar{g}_{mn,2}(\hat{\delta}, \cdot) \\
&= o_p(\sqrt{n})\sqrt{m}\bar{g}_{mn,1}(\hat{\delta}, \cdot) + G_{22}\sqrt{m}\bar{g}_{mn,2}(\hat{\delta}, \cdot) + o_p(1)\sqrt{m}\bar{g}_{mn,2}(\hat{\delta}, \cdot) \\
&\rightsquigarrow G_{22}(\cdot)\mathbb{Z}_{22}(\cdot) + o_p(1).
\end{aligned}$$

The result of part (i)-(b) of the theorem follows.

In part (ii) and (iii), Lemma 4 applies such that

$$\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1}S'_{ZX}\hat{W}(\tau) \underset{p}{\rightarrow} G(\tau) \tag{67}$$

uniformly in $\tau \in \mathcal{T}$. In part (ii), $\bar{g}_{mn}(\hat{\delta}, \cdot) = \bar{g}_{mn}^{(1)}(\hat{\delta}, \cdot)$. It follows from Lemma 2(i) that

$$\sqrt{mn}\bar{g}_{mn}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}_1(\cdot). \tag{68}$$

The result of part (ii) of the theorem follows from equations 65, 67, and 68.

In part (iii), $\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau)$ and $\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)$ converge at the same rate such that Lemma 2 implies

$$\sqrt{mn}\bar{g}_{mn}(\hat{\delta}, \cdot) \rightsquigarrow \mathbb{Z}(\cdot) \tag{69}$$

where $\mathbb{Z}(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega_1(\tau, \tau') + \bar{\Omega}_1(\tau, \tau')$. The result of part (iii) of the theorem follows from equations 65, 67, and 69.

$\blacksquare$

A.6 Proof of Theorem 3: Adaptive Asymptotic Distribution

Proof of Theorem 3. We prove the theorem when Assumption $8'$ holds. The proof when instead Assumption 8 holds is simpler because all coefficients converge at the same rate; it can actually be considered as a special case of the proof below when we set $a_n(\tau) = 1$.

By Lemma 5, uniformly in $\tau \in \mathcal{T}$,

$$\hat{G}(\tau) = G_{mn}(\tau) + \begin{pmatrix} o_p(1) & o_p\left(\sqrt{a_n(\tau)}\right) \\ o_p\left(1/\sqrt{a_n(\tau)}\right) & o_p(1) \end{pmatrix}$$

where $G_{mn,12}(\tau) = O_p(a_n(\tau))$ and the other elements are $O_p(1)$.

Since the rate of convergence may differ across the quantile index and the regressors, we first consider the scalar $\hat{\delta}_k(\tau)$, which is the k-th element of $\hat{\delta}(\tau)$ for $k \in \{1, \dots, K\}$ and $\tau \in \mathcal{T}$. From Lemma 1,

$$\hat{\delta}_k(\tau) - \delta_k(\tau) = \hat{G}_k(\tau)\bar{g}_{mn}(\hat{\delta}, \tau) = \hat{G}_k(\tau)\left(\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) + \bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)\right)$$

where $\hat{G}_k(\tau)$ is the k-th row of $\hat{G}(\tau)$. From the proof of Lemma 2(i), we have

$$\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = \frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) + o_p\left(\frac{1}{\sqrt{mn}}\right)$$

uniformly in $\tau \in \mathcal{T}$.

We deal separately with the fast and slow coefficients and show that the result holds for both types of coefficients. For $k \in \{1, \dots, M_1\}$ (fast coefficients), $\hat{G}_k(\tau)$ is uniformly bounded in probability, which implies that

$$\hat{G}_k(\tau)\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = \hat{G}_k(\tau)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) + o_p\left(\frac{1}{\sqrt{mn}}\right)$$

uniformly in $\tau \in \mathcal{T}$. In addition, 2(i) implies that $\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) = O_p\left(\frac{1}{\sqrt{mn}}\right)$ uniformly in $\tau \in \mathcal{T}$. For these coefficients, $\sup_{\tau \in \mathcal{T}}\left\|\hat{G}_k(\tau) - G_{mn,k}(\tau)\right\| = o_p(1)$. This, in turn, implies that

$$\left(\hat{G}_k(\tau) - G_{mn,k}(\tau)\right)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) = o_p\left(\frac{1}{\sqrt{mn}}\right)$$

uniformly in $\tau \in \mathcal{T}$. Combining the last two displayed equations, we obtain

$$\hat{G}_k(\tau)\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = G_{mn,k}(\tau)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) + o_p\left(\frac{1}{\sqrt{mn}}\right) \quad (70)$$

uniformly in $\tau \in \mathcal{T}$.

For $k \in \{M_1+1, \ldots, K\}$ (slow coefficients), by Lemma 5, uniformly in τ, $\hat{G}_k(\tau) - G_{mn,k}(\tau) = o_p(1/\sqrt{a_n(\tau)})$ and $G_{mn,k} = O_p(1)$ such that $\hat{G}_k = O_p(1/\sqrt{a_n(\tau)})$. In addition, Lemma 2(i) implies that

$$\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) = O_p\left(\frac{1}{\sqrt{mn}}\right).$$

It follows that

$$\hat{G}_k(\tau)\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = \hat{G}_k(\tau)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) + o_p\left(\frac{1}{\sqrt{mna_n(\tau)}}\right)$$

and

$$\left(\hat{G}_k(\tau) - G_{mn,k}(\tau)\right)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) = o_p\left(\frac{1}{\sqrt{mna_n(\tau)}}\right)$$

uniformly in $\tau \in \mathcal{T}$. Combining the last two displayed equations, we obtain

$$\hat{G}_k(\tau)\bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) = G_{mn,k}(\tau)\frac{1}{m}\sum_{j=1}^{m}\Sigma_{ZXj}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_{j,\tau}(\tilde{x}_{ij}, y_{ij})\right) + o_p\left(\frac{1}{\sqrt{mna_n(\tau)}}\right)$$

uniformly in $\tau \in \mathcal{T}$.

Lemma 2(ii) implies that

$$\bar{g}_{mn,l}^{(2)}(\delta, \tau) = \frac{1}{m}\sum_{j=1}^{m}\bar{z}_{j,l}\alpha_j(\tau) = O_p\left(\frac{1}{\sqrt{m}}\right)\sqrt{\Omega_{2,ll}(\tau)} \quad (71)$$

uniformly in $\tau \in \mathcal{T}$, where $g_{mn,l}^{(2)}(\delta, \tau)$ is the l-th element of $g_{mn}^{(2)}(\delta, \tau)$, $z_{j,l}$ is the l-th element of z_j, and $\Omega_{2,ll}(\tau)$ it the element in the l-th row and l-th column of $\Omega_2(\tau)$. Let $\hat{G}_{kl}(\tau)$ and $G_{mn,kl}(\tau)$ be the element in the k-th row and l-th column of $\hat{G}(\tau)$ and $G_{mn}(\tau)$, respectively. Note that $\hat{G}_{kl}(\tau) - G_{mn,kl}(\tau) = o_p(\sqrt{G_{mn,kl}(\tau)})$ for $l \in \{L_1+1, \ldots, L\}$ and $k \in \{1, \ldots, K\}$. Thus, we have

$$(\hat{G}_k(\tau) - G_{mn}(\tau))\bar{g}_{mn}^{(2)}(\hat{\delta}, \tau) = \sum_{l=L_1+1}^{L}(\hat{G}_{kl}(\tau) - G_{mn,kl}(\tau))\bar{g}_{mn,l}^{(2)}(\hat{\delta}, \tau)$$

$$= \sum_{l=L_1+1}^{L}o_p\left(\sqrt{G_{mn,kl}}\right)O_p\left(\frac{1}{\sqrt{m}}\right)\sqrt{\Omega_{2,ll}(\tau)}$$

$$= o_p\left(\frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}}\right)\sum_{l=L_1+1}^{L}\sqrt{G_{mn,kl}(\tau)})$$

uniformly in $\tau \in \mathcal{T}$.

If $k \in \{1, \ldots, M_1\}$, define

$$\zeta(k, \tau) = \frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}} \sum_{l=L_1}^{L} \sqrt{G_{mn,kl}(\tau)}. \tag{72}$$

If $k \in \{M_1 + 1, \ldots, K\}$, define

$$\zeta(k, \tau) = \frac{1}{\sqrt{mna_n(\tau)}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}} \sum_{l=L_1+1}^{L} \sqrt{G_{mn,kl}(\tau)}. \tag{73}$$

Thus, uniformly in $\tau \in \mathcal{T}$,

$$\hat{\delta}_k(\tau) - \delta_k(\tau) = \hat{G}_k(\tau) \bar{g}_{mn}^{(1)}(\hat{\delta}, \tau) + \hat{G}_k(\tau) \bar{g}_{mn}^{(2)}(\hat{\delta}, \tau)$$

$$= G_{mn,k}(\tau) \left(\frac{1}{m} \sum_{j=1}^{m} \Sigma_{ZXj} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) \right) + \frac{1}{m} \sum_{j=1}^{m} \bar{z}_j \alpha_j(\tau) \right) + o_p(\zeta(k, \tau))$$

Thus, we have

$$\hat{\delta}_k(\tau) - \delta_k(\tau) = \sum_{j=1}^{m} d_j(k, \tau) + o_p(\zeta(k, \tau))$$

where

$$d_j(k, \tau) = G_{mn,k}(\tau) \left(\frac{1}{m} \Sigma_{ZXj} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_{j,\tau}(\tilde{x}_{ij}, y_{ij}) \right) + \frac{1}{m} \bar{z}_j \alpha_j(\tau) \right)$$

Let D_j be the $TK \times 1$ vector $(\text{diag}(\Sigma_{mn}(\tau_1))^{-1/2} d_j(\tau_1), \ldots, \text{diag}(\Sigma_{mn}(\tau_T))^{-1/2} d_j(\tau_T))'$ where $d_j(\tau) = (d_j(1, \tau), d_j(2, \tau), \ldots, d_j(K, \tau))'$. It follows that

$$\text{Var}\left(\sum_{j=1}^{m} D_j \right) = H_{mn}$$

Then, it follows from the proof of Lemma 2 and Assumption 9,

$$H_{mn}^{-1/2} \sum_{j=1}^{m} D_j \xrightarrow{d} N(0, I_{TK})$$

By Slutsky's theorem,

$$H^{-1/2} \sum_{j=1}^{m} D_j = H_{mn}^{-1/2} \sum_{j=1}^{m} D_j + o_p(1) \xrightarrow{d} N(0, I_{TK})$$

In the proof of Lemma 2 we show that $\sum_j \text{diag}(\Sigma_{mn}(\cdot))^{-1/2} d_j(\cdot))$ is asymptotically tight in $l^{\infty}(\mathcal{T})$. It follows that the process $\sum_j \text{diag}(\Sigma_{mn}(\cdot))^{-1/2} d_j(\cdot))$ weakly converges to $\mathbb{Z}$, a centered Gaussian process with covariance kernel $H(\tau, \tau')$.

Finally, we show that $o_p(1) \sup_{\tau \in \mathcal{T}, k \in \{1, \ldots, K\}} \zeta(k, \tau) \Sigma_{mn,k}(\tau)^{-1/2} = o_p(1)$, where $\Sigma_{mn,k}(\tau)$ is the (k, k) element of $\Sigma_{mn}(\tau)$. We consider separately the fast and slow coefficients For $k \in \{1, \ldots, M_1\}$ (fast coefficients), note that

$$\Sigma_{mn,k}(\tau)^{1/2} = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{n\sqrt{m}} \right) = O_p\left(\frac{1}{\sqrt{mn}} \right)$$

and

$$\zeta(k,\tau) = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha(\tau))}}{\sqrt{m}} \sum_{l=L_1}^{L} \sqrt{G_{mn,kl}(\tau)}\right)$$

$$= O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}} \frac{1}{\sqrt{1 + \text{Var}(\alpha_j(\tau))n}}\right)$$

$$= O_p\left(\frac{1}{\sqrt{mn}}\right)$$

both uniformly in τ. It follows that $o_p(1)\sup_{\tau\in\mathcal{T}}\zeta(k,\tau)\Sigma_{mn,k}(\tau)^{-1/2} = o_p(1)$.

For $k \in \{M_1+1,\ldots,K\}$ (slow coefficients), note that

$$\Sigma_{mn,k}(\tau)^{1/2} = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}}\right)$$

and

$$\zeta(k,\tau) = \frac{1}{\sqrt{mna_n(\tau)}} + \frac{\sqrt{\text{Var}(\alpha(\tau))}}{\sqrt{m}} \sum_{l=L_1}^{L} \sqrt{G_{mn,kl}(\tau)}.$$

For the first term of $\zeta(k,\tau)$, we obtain

$$\frac{1}{\sqrt{mna_n(\tau)}}\frac{1}{\Sigma_{mn,k}(\tau)^{1/2}} = O_p\left(\frac{1}{\sqrt{mna_n(\tau)}}\frac{1}{\frac{1}{\sqrt{mn}} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}}}\right)$$

$$= O_p\left(\frac{1}{\sqrt{a_n(\tau)}}\frac{1}{1 + \sqrt{\text{Var}(\alpha_j(\tau))n}}\right)$$

$$= O_p\left(\frac{\sqrt{1 + \text{Var}(\alpha_j(\tau))n}}{1 + \sqrt{\text{Var}(\alpha_j(\tau))n}}\right)$$

$$= O_p(1)$$

For the second term of $\zeta(k,\tau)$, we obtain

$$\frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}} \sum_{l=L_1}^{L} \sqrt{G_{mn,kl}(\tau)}\Sigma_{mn,k(\tau)}^{-1/2} = O_p\left(\frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}}\frac{1}{\sqrt{mn} + \frac{\sqrt{\text{Var}(\alpha_j(\tau))}}{\sqrt{m}}}\right)$$

$$= O_p(1)$$

It follows that, also in this second case, $o_p(1)\sup_{\tau\in\mathcal{T}}\zeta(k,\tau)\Sigma_{mn,k}(\tau)^{-1/2} = o_p(1)$.

Hence, uniformly in τ,

$$\text{diag}(\Sigma_{mn}(\tau))^{-1/2}(\hat{\delta}(\tau) - \delta(\tau)) = \text{diag}(\Sigma_{mn}(\tau))^{-1/2}\left(\sum_{j=1}^{m} d_j(\tau) + o_p\left(\zeta(\tau)\right)\right)$$

$$= \text{diag}(\Sigma_{mn}(\tau))^{-1/2}\sum_{j=1}^{m} d_j(\tau) + o_p(1)$$

$$\rightsquigarrow \mathbb{G}(\tau)$$

58

where $\zeta(\tau) = (\zeta(1,\tau),\ldots,\zeta(K,\tau))'$.

∎

A.7   Proof of Propositions 1 and 1′: Properties of $\hat{\Omega}(\tau,\tau')$

A.7.1   Proposition 1

Proof of Proposition 1. We use $\hat{u}_{ij}(\tau) = \tilde{x}'_{ij}\hat{\beta}_j(\tau) - x'_{ij}\hat{\delta}(\tau) = \tilde{x}'_{ij}(\hat{\beta}_j(\tau) - \beta_j(\tau)) + x'_{ij}(\delta(\tau) - \hat{\delta}(\tau)) + \alpha_j(\tau)$ to obtain

$$
\begin{aligned}
\hat{\Omega}(\tau,\tau') &= \frac{1}{m}\sum_{j=1}^{m}\left\{ \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\hat{u}_{ij}(\tau)\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\hat{u}_{ij}(\tau')\right)' \right\} \\
&= \frac{1}{m}\sum_{j=1}^{m}\left\{ \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\left[\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau)) + x_{ij}(\delta(\tau)-\hat{\delta}(\tau)) + \alpha_j(\tau)\right]\right) \right. \\
&\qquad\qquad \left. \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\left[\tilde{x}'_{ij}(\hat{\beta}_j(\tau')-\beta_j(\tau')) + x'_{ij}(\delta(\tau')-\hat{\delta}(\tau')) + \alpha_j(\tau')\right]\right)' \right\} \\
&= \frac{1}{m}\sum_{j=1}^{m}\left\{ \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau')-\beta_j(\tau'))\right)' \right. \\
&\qquad + \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau)\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau')\right)' \\
&\qquad + \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}(\hat{\delta}(\tau)-\delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z'_{ij}x_{ij}(\hat{\delta}(\tau')-\delta(\tau'))\right)' \\
&\qquad - \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}(\hat{\delta}(\tau')-\delta(\tau'))\right)' \\
&\qquad - \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}(\hat{\delta}(\tau)-\delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau')-\beta_j(\tau'))\right)' \\
&\qquad + \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau')\right)' \\
&\qquad + \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau)\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau')-\beta_j(\tau'))\right)' \\
&\qquad - \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}(\hat{\delta}(\tau)-\delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau')\right)' \\
&\qquad \left. - \left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j(\tau)\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}x'_{ij}(\hat{\delta}(\tau')-\delta(\tau'))\right)' \right\}
\end{aligned}
$$

We will show that the first term converges to $\Omega_1(\tau,\tau')/n$, the second term to $\Omega_2(\tau,\tau')$, and the remaining terms vanish at a rate faster than the leading term.

By the proof of Lemma 2(i), it follows for the first term that

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}'_{ij}(\hat{\beta}_j(\tau')-\beta_j(\tau'))\right)'$$

$$=\mathbb{E}\left[\left(\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{i,\tau}(\tilde{x}_{ij},z_{ij})\right)\left(\Sigma_{ZXj}\frac{1}{n}\sum_{i=1}^{n}\phi_{i,\tau'}(\tilde{x}_{ij},z_{ij})\right)'\right]+o_p\left(\frac{1}{mn}\right)$$

$$=\frac{\Omega_1(\tau,\tau')}{n}+O_p\left(\frac{1}{mn}\right)$$

uniformly in τ. For the second term, we consider each element of the $L \times L$ matrix separately. For $l, l' \in \{1, \ldots, L\}$, we have

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}\alpha_j(\tau)\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl'}\alpha_j(\tau')\right)'=\frac{1}{m}\sum_{j=1}^{m}\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau')$$

If either l or l' (or both) are $\in \{1, \ldots, L_1\}$, then both $\Omega_{2ll'}(\tau,\tau')=0$ and $\sum_{j=1}^{m}\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau')=0$. Thus, we need only to consider the case that both $\bar{z}_{jl}$ and $\bar{z}_{jl'}$ are not equal to 0 uniformly across all groups. We apply Theorem 9.2 in Hansen (2022a). His condition (9.3) is satisfied by the boundedness of z_{ij} in Assumption 5(i) and the uniform boundedness of the $4+\varepsilon_C$ moment of $\alpha_j(\tau)$ in Assumption 6(i). Condition (9.5) in Hansen (2022a) is satisfied by the assumption in equation (23). It follows that

$$\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m}\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau')-\Omega_{2ll'}(\tau,\tau')\right)\underset{d}{\rightarrow}N(0,C_{l,l'}(\tau,\tau'))$$

Then, since the condition for Theorem 18.3 in Hansen (2022a) are satisfied, we have that $\sup_{\tau,\tau'}\left\|\frac{1}{m}\sum_{j=1}^{m}\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau')-\Omega_{2ll'}(\tau,\tau')\right\|=O_p\left(\frac{\sqrt{C_{l,l'}(\tau,\tau')}}{\sqrt{m}}\right)$. Let C_l denotes a uniform bound on $|z_{ijl}|$. From the definition of $C_{l,l'}(\tau,\tau')$ we have

$$C_{l,l'}(\tau,\tau')\leq C_l^2 C_{l'}^2\,\mathrm{Var}(\alpha_j(\tau)\alpha_j(\tau'))\leq C_l^2 C_{l'}^2\mathbb{E}[\alpha_j(\tau)^2\alpha_j(\tau')^2]\leq C_l^2 C_{l'}^2\sqrt{\mathbb{E}[\alpha_j(\tau)^4]\mathbb{E}[\alpha_j(\tau')^4]}$$

Finally, note that $\frac{\mathbb{E}[\alpha_j(\tau)^4]}{Var(\alpha_j(\tau))^2}$ is bounded. It follows that $\frac{1}{m}\sum_{j=1}^{m}\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau')=\Omega_{2ll'}(\tau,\tau')+O_p\left(\frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))\,\mathrm{Var}(\alpha_j(\tau'))}}{\sqrt{m}}\right)=\Omega_{2ll'}(\tau,\tau')+O_p\left(\frac{\sqrt{\Omega_{2,ll}(\tau,\tau)\Omega_{2,l'l'}(\tau'\tau')}}{\sqrt{m}}\right)$.

Also for the third term, we consider separately each element l, l' of the matrix:

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}x'_{ij}(\hat{\delta}(\tau)-\delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl'}x'_{ij}(\hat{\delta}(\tau')-\delta(\tau'))\right)=$$

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ijl}x_{ijk}(\hat{\delta}_k(\tau)-\delta_k(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ijl'}x_{ijk}(\hat{\delta}_k(\tau')-\delta_k(\tau'))\right)$$

We split $\sum_{k=1}^{K}z_{ijl'}x_{ijk}(\hat{\delta}_k(\tau')-\delta_k(\tau'))$ into the the individual-level and group-level variables. Since the estimation error of the coefficients on the individual-level variables is $O_p(1/\sqrt{mn})$ and

z_{ijl} and x_{ijk} are bounded, we obtain

$$\sum_{k=1}^{K_1} \frac{1}{n} \sum_{i=1}^{n} z_{ijl} x_{1ijk} (\hat{\delta}_k(\tau) - \delta_k(\tau)) = O_p\left(\frac{1}{\sqrt{mn}}\right)$$

uniformly in τ. For the group-level variables, we obtain

$$\sum_{k=K_1+1}^{K} \frac{1}{n} \sum_{i=1}^{n} z_{ijl} x_{2jk} (\hat{\delta}_k(\tau) - \delta_k(\tau)) = \sum_{k=K_1+1}^{K} \bar{z}_{jl} x_{2jk} (\hat{\delta}_k(\tau) - \delta_k(\tau))$$

It follows that $\sum_{k=K_1+1}^{K} \frac{1}{n} \sum_{i=1}^{n} z_{ijl} x_{jk} (\hat{\delta}_k(\tau) - \delta_k(\tau)) = 0$ if $\bar{z}_{jl} = 0$ for all $j = 1, \ldots, m$, i.e. if $l \in \{1, \ldots, L_1\}$. If $l \in \{L_1 + 1, \ldots, L\}$, uniformly in τ, $\sum_{k=K_1+1}^{K} \frac{1}{n} \sum_{i=1}^{n} z_{ijl} x_{jk} (\hat{\delta}_k(\tau) - \delta_k(\tau)) = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{Var(\alpha_j(\tau))}}{\sqrt{m}}\right)$. Since $\Omega_{2,ll}(\tau) = Var(\bar{z}_{jl}\alpha_j(\tau))$, in both cases we can write

$$\sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} z_{ijl} x_{jk} (\hat{\delta}_k(\tau) - \delta_k(\tau)) = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\Omega_{2,ll}(\tau)}}{\sqrt{m}}\right)$$

uniformly in τ. Combining these results, we get uniformly in τ, τ'

$$\frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl} x'_{ij} (\hat{\delta}(\tau) - \delta(\tau))\right) \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl'} x'_{ij} (\hat{\delta}(\tau') - \delta(\tau'))\right) \tag{74}$$

$$= O_p\left(\frac{\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau')}}{m}\right). \tag{75}$$

For the fourth term, similar arguments and the fact that $\sup_\tau \|\hat{\beta}_j(\tau) - \beta_j(\tau)\| = O_p\left(\frac{1}{\sqrt{n}}\right)$ imply that uniformly in τ, τ'

$$\frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl} \tilde{x}'_{ij} (\hat{\beta}_j(\tau) - \beta_j(\tau))\right) \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl'} x'_{ij} (\hat{\delta}(\tau') - \delta(\tau'))\right) \tag{76}$$

$$= O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\Omega_{2,l'l'}(\tau')}}{\sqrt{mn}}\right). \tag{77}$$

Similarly, for the fifth term

$$\frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl'} x'_{ij} (\hat{\delta}(\tau') - \delta(\tau'))\right) \left(\frac{1}{n} \sum_{i=1}^{n} z_{ijl} \tilde{x}'_{ij} (\hat{\beta}_j(\tau) - \beta_j(\tau))\right) \tag{78}$$

$$= O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\Omega_{2,ll}(\tau)}}{\sqrt{mn}}\right) \tag{79}$$

uniformly in τ, τ'. Thus, the sum of the fourth and fifth terms is

$$O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\Omega_{2,ll}(\tau)}}{\sqrt{m}\sqrt{n}} + \frac{\sqrt{\Omega_{2,l'l'}(\tau')}}{\sqrt{m}\sqrt{n}}\right) = O_p\left(\frac{\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau')}}{\sqrt{m}}\right)$$

uniformly in τ.

For the sixth term, we obtain, uniformly in τ, τ'

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}\tilde{x}_{ij}'(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl'}\alpha_j(\tau')\right)$$

$$=\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}\tilde{x}_{ij}'(\hat{\beta}_j(\tau)-\beta_j(\tau))\right)\bar{z}_{jl'}\alpha_j(\tau')$$

$$=O_p\left(\frac{\sqrt{\Omega_{2,l'l'}(\tau')}}{\sqrt{mn}}\right)$$

and similarly, we can show that the seventh term is $O_p\left(\frac{\sqrt{\Omega_{2,ll}(\tau)}}{\sqrt{mn}}\right)$.

For the eighth term we obtain

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}x_{ij}'(\hat{\delta}(\tau)-\delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl'}'\alpha_j(\tau')\right)$$

$$=\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}z_{ijl}x_{ij}'(\hat{\delta}(\tau)-\delta(\tau))\right)\bar{z}_{jl'}\alpha_j(\tau')$$

$$=O_p\left(\frac{\sqrt{\Omega_{2,ll}(\tau)\Omega_{2,l'l'}(\tau')}}{m}+\frac{\sqrt{\Omega_{2,l'l'}(\tau')}}{m\sqrt{n}}\right)$$

and the ninth term is $O_p\left(\frac{\sqrt{\Omega_{2,ll}(\tau)\Omega_{2,l'l'}(\tau')}}{m}+\frac{\sqrt{\Omega_{mn,l'l'}(\tau')}}{m\sqrt{n}}\right)$ such that the sum of the eight and ninth term is $O_p\left(\frac{\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau')}}{m}\right)$, where both results are uniformly in τ, τ'.

Combining all the terms, we find that for each l, l' entry

$$\hat{\Omega}=\frac{\Omega_{1,ll'}(\tau,\tau')}{n}+\Omega_{2,ll'}(\tau,\tau')+O_p\left(\frac{\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau)}}{\sqrt{m}}\right)$$

$$=\Omega_{mn,ll'}(\tau,\tau')+o_p\left(\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau')}\right)$$

uniformly in τ, τ'.

■

A.7.2   Proposition 1′

When the coefficients of some individual-level variables converge at a slow rate, while those of others converge at a fast rate, the third term in the proof of Proposition 1 may not be $O_p\left(\frac{\sqrt{\Omega_{mn,ll}(\tau)\Omega_{mn,l'l'}(\tau')}}{m}\right)$. The slowly converging coefficients can introduce an estimation error in the variance of the faster moments that diminishes more slowly than the true value. Proposition 1', stated below, provides a more general result that does not assume the coefficients of individual-level variables converge at the $O_p(1/\sqrt{mn})$ rate. Consequently, $\hat{\Omega}(\tau,\tau')$ is consistent as long as all individual-level variable coefficients converge at the same rate. One example of this is the between estimator for individual-level variables. Another example is the 2SLS estimator

applied to the random effects model. In this case, the weighting matrix is full rank, and all coefficients converge at the rate of $1/\sqrt{mn} + \sqrt{\mathrm{Var}(\alpha_j(\tau))}/\sqrt{m}$.

Proposition 1$'$. *Let assumptions 1-6 and 7(c) hold. As $m \to \infty$, for each $l, l' \in \{1, \ldots, L\}$ and uniformly in $\tau, \tau' \in \mathcal{T}^2$,*

$$m^{-1} \sum_{j=1}^{m} \mathbb{E}\left[\left(\bar{z}_{jl}\bar{z}_{jl'}\alpha_j(\tau)\alpha_j(\tau') - \Omega_{2ll'}(\tau, \tau')\right)^2\right] \to C_{l,l'}(\tau, \tau') < \infty$$

The estimator used to compute $\hat{u}_{ij}(\tau)$ satisfies

$$\hat{\delta}(\tau) - \delta(\tau) = O_p\left(\frac{1}{\sqrt{mn}} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))}}{\sqrt{m}}\right)$$

uniformly in τ. Then, for any ll' entry of the $\hat{\Omega}(\tau, \tau')$ matrix with $l, l' \in \{1, \ldots, L\}$ we have

$$\hat{\Omega} = \Omega_{mn,ll'}(\tau, \tau') + o_p\left(\frac{1}{n} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))}}{\sqrt{n}} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau'))}}{\sqrt{n}} + \sqrt{\mathrm{Var}(\alpha_j(\tau))\,\mathrm{Var}(\alpha_j(\tau'))}\right)$$

Proof. The proof follows the same steps as the proof of Lemma 1 but requires some modifications each time $\hat{\delta}(\tau) - \delta(\tau)$ is involved. This term is now $O_p\left(\frac{1}{\sqrt{mn}} + \frac{\mathrm{Var}(\alpha_j(\tau))}{\sqrt{m}}\right)$ instead of $O_p\left(\frac{1}{\sqrt{mn}} + \frac{\Omega_2(\tau)}{\sqrt{m}}\right)$. For the third term, we obtain

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} z_{ijl}x_{ij}'(\hat{\delta}(\tau) - \delta(\tau))\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_{ijl'}'x_{ij}'(\hat{\delta}(\tau') - \delta(\tau'))\right)$$

$$= O_p\left(\frac{1}{mn} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))\,\mathrm{Var}(\alpha_j(\tau'))}}{m} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))}}{\sqrt{n}m} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau'))}}{\sqrt{n}m}\right)$$

The sum of the fourth and fifth terms is now

$$O_p\left(\frac{1}{n\sqrt{m}} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau))}}{\sqrt{nm}} + \frac{\sqrt{\mathrm{Var}(\alpha_j(\tau'))}}{\sqrt{nm}}\right)$$

The sum of the eight and ninth terms is now

$$O_p\left(\frac{\sqrt{\Omega_{2,ll}(\tau)}}{m\sqrt{n}} + \frac{\sqrt{\Omega_{2,l'l'}(\tau')}}{m\sqrt{n}} + \frac{\sqrt{\mathrm{Var}(\alpha(\tau'))\Omega_{2,ll}(\tau)}}{m} + \frac{\sqrt{\mathrm{Var}(\alpha(\tau))\Omega_{2,l'l'}(\tau')}}{m}\right).$$

The result of the lemma follows. ∎

A.8 Proof of Proposition 2: Adaptive Inference

Proof of Proposition 2. We prove the results for $T = 2$ as the proof trivially extends to $T > 2$. We consider the $2K \times 1$ coefficient vector $\hat{\delta}(\tau, \tau') = (\hat{\delta}(\tau)', \hat{\delta}(\tau')')'$ whose asymptotic covariance matrix is

$$\Sigma_{mn} = \begin{pmatrix} \Sigma_{mn}(\tau) & \Sigma_{mn}(\tau, \tau') \\ \Sigma_{mn}(\tau, \tau') & \Sigma_{mn}(\tau') \end{pmatrix}$$

This covariance matrix is estimated by

$$\hat{\Sigma} = \frac{1}{m} \begin{pmatrix} \hat{G}(\tau) & 0 \\ 0 & \hat{G}(\tau') \end{pmatrix} \begin{pmatrix} \hat{\Omega}(\tau,\tau) & \hat{\Omega}(\tau,\tau') \\ \hat{\Omega}(\tau',\tau) & \hat{\Omega}(\tau') \end{pmatrix} \begin{pmatrix} \hat{G}(\tau) & 0 \\ 0 & \hat{G}(\tau') \end{pmatrix}'$$

$$= \frac{1}{m} \begin{pmatrix} \hat{G}(\tau)\hat{\Omega}(\tau)\hat{G}(\tau) & \hat{G}(\tau)\hat{\Omega}(\tau,\tau')\hat{G}(\tau') \\ \hat{G}(\tau')\hat{\Omega}(\tau',\tau)\hat{G}(\tau) & \hat{G}(\tau')\hat{\Omega}(\tau')\hat{G}(\tau'). \end{pmatrix}$$

For each $k, k' \in \{1, \ldots, 2K\}$, we will show that

$$\hat{\Sigma}_{kk'} = \Sigma_{mn,kk'} + o_p\left(\sqrt{\Sigma_{mn,kk}\Sigma_{mn,k'k'}}\right).$$

To simplify the notation, we show this result for the $K \times K$ top-left submatrix of Σ_{mn} such that we can drop the dependence on τ. The proof for the other parts is similar. Our analysis relies on two key ingredients:

First, by Proposition 1, for any ll' entry of the $\hat{\Omega}$ matrix with $l, l' \in \{1, \ldots, L\}$, we have

$$\hat{\Omega}_{ll'} = \Omega_{mn,ll'} + o_p\left(\sqrt{\Omega_{mn,ll}\Omega_{mn,l'l'}}\right)$$

where Ω can be split into four submatrices where the dimensions of the top-left part are $L_1 \times L_1$ and the bottom-right are $L_2 \times L_2$:

$$\Omega = \begin{pmatrix} \Omega_{mn,11} & \Omega_{mn,12} \\ \Omega_{mn,21} & \Omega_{mn,22} \end{pmatrix}$$

such that $\Omega_{mn,11}$, $\Omega_{mn,12}$, and $\Omega_{mn,21}$ are $O_p\left(\frac{1}{n}\right)$ while $\Omega_{mn,22} = O_p\left(\frac{1}{na_n}\right)$.

Second, by Lemma 5,

$$\hat{G} = \begin{pmatrix} G_{mn,11} & G_{mn,12} \\ G_{mn21} & G_{mn,22} \end{pmatrix} + \begin{pmatrix} o_p(1) & o_p(\sqrt{a_n}) \\ o_p(1/\sqrt{a_n}) & o_p(1) \end{pmatrix}$$

where $G_{mn,12} = O_p(a_n)$ and the other elements of G_{mn} are $O_p(1)$.

With this notation,

$$\Sigma_{mn} = \frac{1}{m} \begin{pmatrix} G_{mn,11} & G_{mn,12} \\ G_{mn,21} & G_{mn,22} \end{pmatrix} \begin{pmatrix} \Omega_{mn,11} & \Omega_{mn,12} \\ \Omega_{mn,21} & \Omega_{mn,22} \end{pmatrix} \begin{pmatrix} G'_{mn,11} & G'_{mn,21} \\ G'_{mn,12} & G'_{mn,22} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{mn,11} & \Sigma_{mn,12} \\ \Sigma_{mn,21} & \Sigma_{mn,22} \end{pmatrix}$$

where

$$\Sigma_{mn,11} = \frac{1}{m}\left(G_{mn,11}\Omega_{mn,11}G'_{mn,11} + G_{mn,11}\Omega_{mn,12}G'_{mn,12} + G_{mn,12}\Omega_{mn,21}G'_{mn,11} + G_{mn,12}\Omega_{mn,22}G'_{mn,12}\right)$$

$$\Sigma_{mn,12} = \frac{1}{m}\left(G_{mn,11}\Omega_{mn,11}G'_{mn,21} + G_{mn,11}\Omega_{mn,12}G'_{mn,22} + G_{mn,12}\Omega_{mn,21}G'_{mn,21} + G_{mn,12}\Omega_{mn,22}G'_{mn,22}\right)$$

$$\Sigma_{mn,21} = \frac{1}{m}\left(G_{mn,21}\Omega_{mn,11}G'_{mn,11} + G_{mn,21}\Omega_{mn,12}G'_{mn,12} + G_{mn,22}\Omega_{mn,21}G'_{mn,11} + G_{mn,22}\Omega_{mn,22}G'_{mn,12}\right)$$

$$\Sigma_{mn,22} = \frac{1}{m}\left(G_{mn,21}\Omega_{mn,11}G'_{mn,21} + G_{mn,21}\Omega_{mn,12}G'_{mn,22} + G_{mn,22}\Omega_{mn,21}G'_{mn,21} + G_{mn,22}\Omega_{mn,22}G'_{mn,22}\right)$$

It follows that $\Sigma_{mn,11} = O_p\left(\frac{1}{mn}\right)$, $\Sigma_{mn,12} = O_p\left(\frac{1}{mn}\right)$, $\Sigma_{mn,21} = O_p\left(\frac{1}{mn}\right)$, and $\Sigma_{mn,22} = O_p\left(\frac{1}{mna_n}\right)$.

We consider the estimation error of these four parts separately, each of them being composed of four parts.

$$\hat{G}_{11}\hat{\Omega}_{11}\hat{G}'_{11} = (G_{mn,11} + o_p(1))(\Omega_{mn,11} + o_p(\Omega_{mn,11}))(G'_{mn,11} + o_p(1))$$

$$= G_{mn,11}\Omega_{mn,11}G'_{mn,11} + o_p(\Omega_{mn,11})$$

$$= G_{mn,11}\Omega_{mn,11}G'_{mn,11} + o_p\left(\frac{1}{n}\right)$$

$$\hat{G}_{11}\hat{\Omega}_{12}\hat{G}'_{12} = (G_{mn,11} + o_p(1))\left(\Omega_{mn,12} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)\right)(G'_{mn,12} + o_p(\sqrt{a_n}))$$

$$= G_{mn,11}\Omega_{mn,12}G'_{mn12} + o_p(\Omega_{mn,12}) + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}a_n}\right)$$

$$= G_{mn,11}\Omega_{mn,12}G'_{mn,12} + o_p\left(\frac{1}{n}\right)$$

$$\hat{G}_{12}\hat{\Omega}_{21}\hat{G}'_{11} = (\hat{G}_{11}\hat{\Omega}_{12}\hat{G}'_{12})'$$

$$= G_{mn,12}\Omega_{mn,21}G'_{mn,11} + o_p\left(\frac{1}{n}\right)$$

$$\hat{G}_{12}\hat{\Omega}_{22}\hat{G}'_{12} = (G_{mn,12} + o_p(\sqrt{a_n}))(\Omega_{mn,22} + o_p(\Omega_{mn,22}))(G'_{mn,12} + o_p(\sqrt{a_n}))$$

$$= G_{mn,12}\Omega_{mn,22}G'_{mn,12} + o_p(a_n\Omega_{mn,22})$$

$$= G_{mn,12}\Omega_{mn,22}G'_{mn,12} + o_p\left(\frac{1}{n}\right)$$

It follows that $\hat{\Sigma}_{11} = \Sigma_{mn,11} + o_p\left(\frac{1}{mn}\right) = \Sigma_{mn,11} + o_p(\Sigma_{mn,11})$.

For the second part,

$$\hat{G}_{11}\hat{\Omega}_{11}\hat{G}'_{21} = (G_{mn,11} + o_p(1))(\Omega_{mn,11} + o_p(\Omega_{mn,11}))(G'_{mn,21} + o_p(1/\sqrt{a_n}))$$

$$= G_{mn,11}\Omega_{mn,11}G'_{mn,21} + o_p(\Omega_{mn,11}/\sqrt{a_n})$$

$$= G_{mn,11}\Omega_{mn,11}G'_{mn,21} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)$$

$$\hat{G}_{11}\hat{\Omega}_{12}\hat{G}'_{22} = (G_{mn,11} + o_p(1))\left(\Omega_{mn,12} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)\right)(G'_{mn,22} + o_p(1))$$

$$= G_{mn,11}\Omega_{mn,12}G'_{mn,22} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)$$

$$\hat{G}_{12}\hat{\Omega}_{21}\hat{G}'_{21} = (G_{mn,12} + o_p(\sqrt{a_n}))\left(\Omega_{mn,21} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)\right)(G'_{mn,21} + o_p(1/\sqrt{a_n}))'$$

$$= G_{mn,12}\Omega_{mn,21}G'_{mn,21} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)$$

$$\hat{G}_{12}\hat{\Omega}_{22}\hat{G}'_{22} = (G_{mn,12} + o_p(\sqrt{a_n}))(\Omega_{mn,22} + o_p(\Omega_{mn,22}))(G'_{mn,22} + o_p(1))$$

$$= G_{mn,12}\Omega_{mn,22}G'_{mn,22} + o_p(\sqrt{a_n}\Omega_{mn,22})$$

$$= G_{mn,12}\Omega_{mn,22}G'_{mn,22} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)$$

It follows that $\hat{\Sigma}_{12} = \Sigma_{mn,12} + o_p(m^{-1}\sqrt{\Omega_{mn,11}\Omega_{mn,22}}) = \Sigma_{mn,12} + o_p(\sqrt{\Sigma_{mn,11}\Sigma_{mn,22}})$. The third part, $\hat{\Sigma}_{21}$, is the transpose of $\hat{\Sigma}_{12}$.

For the fourth part,

$$\hat{G}_{21}\hat{\Omega}_{11}\hat{G}'_{21} = (G_{mn,21} + o_p(1/\sqrt{a_n}))(\Omega_{mn,11} + o_p(\Omega_{mn,11}))\left(G'_{mn,21} + o_p(1/\sqrt{a_n})\right)$$

$$= G_{mn,21}\Omega_{mn,11}G'_{mn,21} + o_p(\Omega_{mn,11}/a_n)$$

$$= G_{mn,21}\Omega_{mn,11}G'_{mn,21} + o_p(\Omega_{mn,22})$$

$$\hat{G}_{21}\hat{\Omega}_{12}\hat{G}'_{22} = (G_{mn,21} + o_p(1/\sqrt{a_n}))\left(\Omega_{mn,12} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)\right)\left(G'_{mn,22} + o_p(1)\right)$$

$$= G_{mn,21}\Omega_{mn,12}G'_{mn,22} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}/a_n}\right)$$

$$= G_{mn,21}\Omega_{mn,12}G'_{mn,22} + o_p(\Omega_{mn,22})$$

$$\hat{G}_{22}\hat{\Omega}_{21}\hat{G}'_{21} = (G_{mn,22} + o_p(1))\Omega_{mn,21} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}}\right)\left(G'_{mn,21} + o_p(1/\sqrt{a_n})\right)'$$

$$= G_{mn,22}\Omega_{mn,21}G'_{mn,21} + o_p\left(\sqrt{\Omega_{mn,11}\Omega_{mn,22}/a_n}\right)$$

$$= G_{mn,22}\Omega_{mn,21}G'_{mn,21} + o_p(\Omega_{mn,22})$$

$$\hat{G}_{22}\hat{\Omega}_{22}\hat{G}'_{22} = (G_{mn,22} + o_p(1))(\Omega_{mn,22} + o_p(\Omega_{mn,22}))(G'_{mn,22} + o_p(1))$$

$$= G_{mn,22}\Omega_{mn,22}G'_{mn,22} + o_p(\Omega_{mn,22})$$

It follows that $\hat{\Sigma}_{22} = \Sigma_{mn,22} + o_p\left(m^{-1}\Omega_{mn,22}\right) = \Sigma_{mn,22} + o_p(\Sigma_{mn,22})$. The results for these four submatrices imply that

$$\hat{\Sigma}_{kk'} = \Sigma_{mn,kk'} + o_p(\sqrt{\Sigma_{mn,kk}\Sigma_{mn,k'k'}}),$$

which implies that

$$\eta'\hat{\Sigma}\eta = \eta'\Sigma_{mn}\eta + o_p\left(\eta'\Sigma_{mn}\eta\right).$$

The proposition follows from Theorem 3.

■

A.9   Proof of Proposition 3: Properties of the weighting matrix

Proof. Since this proof focuses on a case where $\tau = \tau'$, we suppress the dependency on τ for simplicity. We partition the Ω matrix as follows:

$$\Omega_{mn} = \begin{pmatrix} \Omega_{mn,11} & \Omega_{mn,12} \\ \Omega_{mn,21} & \Omega_{mn,22} \end{pmatrix}$$

where $\Omega_{mn,11}$ is $L_1 \times L_1$, $\Omega_{mn,12}$ is $L_1 \times L_2$, $\Omega_{mn,21}$ is $L_2 \times L_1$ and $\Omega_{mn,22}$ is $L_2 \times L_2$.

From Proposition 1 we have, uniformly over τ

$$\hat{\Omega} = \begin{pmatrix} \frac{\Omega_{1,11}}{n} & \frac{\Omega_{1,12}}{n} \\ \frac{\Omega_{1,21}}{n} & \Omega_{mn,22} \end{pmatrix} + \begin{pmatrix} o_p(\|\Omega_{mn,11}\|) & +o_p\left(\frac{\sqrt{\|\Omega_{mn,22}\|}}{\sqrt{n}}\right) \\ o_p\left(\frac{\sqrt{\|\Omega_{mn,22}\|}}{\sqrt{n}}\right) & o_p(\|\Omega_{mn,22}\|) \end{pmatrix}.$$

By the inverse of a partitioned matrix

$$\hat{W} = \frac{1}{n} \cdot \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \hat{\Psi} & -\hat{\Psi}\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} \\ -\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\hat{\Psi} & n^{-1} \cdot \hat{\Omega}_{22}^{-1} + \hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\Psi\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} \end{pmatrix} \tag{80}$$

where $\hat{\Psi} = (n \cdot \hat{\Omega}_{11} - n \cdot \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21})^{-1}$. To address the possibility that Ω_{22} could converge to zero, we rescale both elements in $\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}$, which ensures that both elements are well behaved, and show that the expression is self-normalizing. Hence, uniformly in τ

$$\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} = \left(\parallel \Omega_{mn,22} \parallel^{-1} \hat{\Omega}_{12} \right) \left(\parallel \Omega_{mn,22} \parallel^{-1} \hat{\Omega}_{mn,22} \right)^{-1}$$

$$= \Omega_{mn,12}\Omega_{mn,22}^{-1} + o_p \left(\sqrt{n^{-1} \parallel \Omega_{mn,22} \parallel^{-1}} \right)$$

$$= n^{-1}\Omega_{1,12}\Omega_{mn,22}^{-1} + o_p(\sqrt{a_n}),$$

since

$$\parallel \Omega_{mn,22} \parallel^{-1} \hat{\Omega}_{12} = n^{-1} \parallel \Omega_{mn,22} \parallel^{-1} \Omega_{1,12} + o_p \left(\sqrt{n^{-1} \parallel \Omega_{mn,22} \parallel^{-1}} \right),$$

$$\parallel \Omega_{mn,22} \parallel^{-1} \hat{\Omega}_{22} = \parallel \Omega_{mn,22} \parallel^{-1} \Omega_{mn,22} + o_p(1),$$

and $\parallel \Omega_{mn,22} \parallel^{-1} \hat{\Omega}_{22}$ is invertible. It then follows that uniformly in τ

$$\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21} = \frac{\Omega_{1,11}}{n} + o_p \left(\frac{1}{n} \right) - \left(n^{-1}\Omega_{mn,12}\Omega_{mn,22}^{-1} + o_p \left(\sqrt{n^{-1} \parallel \Omega_{mn,22} \parallel^{-1}} \right) \right)$$

$$\left(\frac{\Omega_{1,21}}{n} + o_p \left(\frac{\sqrt{\parallel \Omega_{mn,22} \parallel}}{\sqrt{n}} \right) \right)$$

$$= \frac{\Omega_{1,11}}{n} - \frac{\Omega_{1,12}}{n}\Omega_{mn,22}^{-1}\frac{\Omega_{1,21}}{n} + o_p \left(\frac{1}{n} \right).$$

Hence, uniformly in τ,

$$\hat{\Psi}^{-1} = n \cdot \hat{\Omega}_{11} - n \cdot \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21} = \Omega_{1,11} - \frac{1}{n}\Omega_{1,12}\Omega_{mn,22}^{-1}\Omega_{1,21} + o_p(1)$$

so that for the submatrix in the upper left of equation (80), we have

$$\sup_\tau \left\| \hat{\Psi} - \Psi \right\| = o_p(1),$$

where Ψ is strictly positive definite.

For the upper right term of $\hat{W}$, we have uniformly in τ,

$$-\hat{\Psi}\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} = -\Psi\frac{\Omega_{1,12}}{n}\Omega_{mn,22}^{-1} + o_p(\sqrt{a_n}),$$

where $\Psi\frac{\Omega_{1,12}}{n}\Omega_{mn,22}^{-1} = O_p(a_n)$.

Similarly,

$$\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\hat{\Psi} = \Omega_{mn,22}^{-1}\frac{\Omega_{1,21}}{n}\Psi + o_p(\sqrt{a_n}).$$

67

Finally, for the lower right term, using an appropriate normalization, we find that uniformly in τ

$$n^{-1}\hat{\Omega}_{22}^{-1} = n^{-1}\Omega_{mn,22}^{-1} + o_p\left(\frac{\|\Omega_{mn,22}\|^{-1}}{n}\right) = n^{-1}\Omega_{mn,22}^{-1} + o_p\left(a_n\right),$$

and combining with the results for the other submatrices, for the lower right submatrix, we obtain

$$\sup_{\tau}\left\|n^{-1}\cdot\hat{\Omega}_{22}^{-1} + \hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\Psi\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} - n^{-1}\cdot\Omega_{mn,22}^{-1} + \Omega_{mn,22}^{-1}\frac{\Omega_{1,21}}{n}\Psi\frac{\Omega_{1,12}}{n}\Omega_{mn,22}^{-1}\right\| = o_p(a_n).$$

Hence, uniformly in $\tau \in \mathcal{T}$,

$$\hat{W}(\tau) = \begin{pmatrix} W_{11}(\tau) & a_n(\tau)W_{12}(\tau) \\ a_n(\tau)W_{21}(\tau) & a_n(\tau)W_{22}(\tau) \end{pmatrix} + \begin{pmatrix} o_p(1) & o_p\left(\sqrt{a_n(\tau)}\right) \\ o_p\left(\sqrt{a_n(\tau)}\right) & o_p(a_n(\tau)) \end{pmatrix}.$$

■

A.10   Proof of Proposition 4: Overidentification Test

Proof of Proposition 4. We prove this result for $T = 1$ as the proof trivially extends to $T > 1$. We suppress the dependency on τ for simplicity, as this proof focuses on a case where $\tau = \tau'$. First, we want to rewrite the J-statistics, in a way that accounts for the different convergence rates of the moments conditions. Let $d_\Omega = I_L \cdot \text{diag}(\Omega)$. Then

$$\begin{aligned} J(\hat{\delta}) &= m\bar{g}_{mn}(\hat{\delta})'\hat{\Omega}^{-1}\bar{g}_{mn}(\hat{\delta}) \\ &= \left(d_{\Omega,mn}^{-1/2}\sqrt{m}\bar{g}_{mn}(\hat{\delta})\right)'\left[d_{\Omega,mn}^{-1/2}\hat{\Omega}d_{\Omega,mn}^{-1/2}\right]^{-1}\left(d_{\Omega,mn}^{-1/2}\sqrt{m}\bar{g}_{mn}(\hat{\delta})\right) \end{aligned}$$

Second, we want to show that $\bar{g}_{mn}(\hat{\delta}) = \hat{B}\bar{g}_{mn}(\delta)$. Recall that $\hat{Y}_j = X_j\delta + \alpha_j + \tilde{X}_j(\hat{\beta}_j - \beta_j)$. Hence, we can write

$$\begin{aligned} Z_j'\hat{Y}_j &= Z_j'X_j\delta + Z_j'\alpha_j + Z_j'\tilde{X}_j(\hat{\beta}_j - \beta) \\ S_{Z\hat{Y}} &= S_{ZX}\delta + \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}z_{ij}\alpha_j + \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}z_{ij}\tilde{x}_{ij}'(\hat{\beta}_j - \beta_j) \\ S_{Z\hat{Y}} &= S_{ZX}\delta + \bar{g}_{mn}(\delta). \end{aligned}$$

Then, note that

$$\begin{aligned} \bar{g}_{mn}(\hat{\delta}) &= \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}z_{ij}(\hat{y}_{ij} - x_{ij}'\hat{\delta}) \\ &= S_{Z\hat{Y}} - S_{ZX}\hat{\delta} \\ &= S_{Z\hat{Y}} - S_{ZX}\left(S_{ZX}\hat{\Omega}^{-1}S_ZX\right)^{-1}S_{ZX}\hat{\Omega}^{-1}S_{Z\hat{Y}} = \hat{B}S_{Z\hat{Y}}, \end{aligned}$$

where $\hat{B} = \left(I_L - S_{ZX} \left(S'_{ZX} \hat{\Omega}^{-1} S_{ZX} \right)^{-1} S'_{ZX} \hat{\Omega}^{-1} \right)$.

Thus,

$$\begin{aligned}
\bar{g}_{mn}(\hat{\delta}) &= \hat{B} S_{Z\hat{Y}} \\
&= \left(I_L - S_{ZX} \left(S'_{ZX} \hat{\Omega}^{-1} S_{ZX} \right)^{-1} S'_{ZX} \hat{\Omega}^{-1} \right) \left(S_{ZX} \delta + \bar{g}_{mn}(\delta) \right) \\
&= \hat{B} \bar{g}_{mn}(\delta).
\end{aligned}$$

Since $\left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]$ is positive definite there exist a matrix $\hat{Q}$ such that

$$\left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]^{-1} = \hat{Q}' \hat{Q}.$$

We define $\hat{A} = \hat{Q} d_{\Omega,mn}^{-1/2} S'_{ZX}$ and $\hat{M} = I_L - \hat{A} (\hat{A}' \hat{A})^{-1} \hat{A}'$.

In this third part, we show that

$$\hat{B}' d_{\Omega,mn}^{-1/2} \left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]^{-1} d_{\Omega,mn}^{-1/2} \hat{B} = d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{M} \hat{Q} d_{\Omega,mn}^{-1/2},$$

where $\hat{B}' d_{\Omega,mn}^{-1/2} \left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]^{-1} d_{\Omega,mn}^{-1/2} \hat{B} = \hat{B}' d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{Q} d_{\Omega,mn}^{-1/2} \hat{B}$. Note that

$$\begin{aligned}
\hat{Q} d_{\Omega,mn}^{-1/2} \hat{B} &= \hat{Q} d_{\Omega,mn}^{-1/2} \left(I_L - S_{ZX} \left(S'_{ZX} \hat{\Omega}^{-1} S_{ZX} \right)^{-1} S'_{ZX} \hat{\Omega}^{-1} \right) \\
&= \left(\hat{Q} d_{\Omega,mn}^{-1/2} - \hat{Q} d_{\Omega,mn}^{-1/2} S_{ZX} \left(S'_{ZX} \hat{\Omega}^{-1} S_{ZX} \right)^{-1} S'_{ZX} \hat{\Omega}^{-1} \right) \\
&= \left(\hat{Q} d_{\Omega,mn}^{-1/2} - \hat{Q} d_{\Omega,mn}^{-1/2} S_{ZX} \left(S'_{ZX} d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{Q} d_{\Omega,mn}^{-1/2} S_{ZX} \right)^{-1} S'_{ZX} d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{Q} d_{\Omega,mn}^{-1/2} \right) \\
&= \left(I_L - \hat{A} \left(\hat{A}' \hat{A} \right)^{-1} \hat{A}' \right) \hat{Q} d_{\Omega,mn}^{-1/2} \\
&= \hat{M} \hat{Q} d_{\Omega,mn}^{-1/2}.
\end{aligned}$$

Where the third line uses $d_{\Omega,mn}^{1/2} \hat{\Omega}^{-1} d_{\Omega,mn}^{1/2} = \hat{Q}' \hat{Q}$. In the last two lines, we use the definitions of $\hat{A}$ and $\hat{M}$.

$\hat{M}$ is symmetric and idempotent. Thus

$$\begin{aligned}
\hat{B}' d_{\Omega,mn}^{-1/2} \hat{\Omega}^{-1} d_{\Omega,mn}^{-1/2} \hat{B} &= \hat{B}' d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{Q} d_{\Omega,mn}^{-1/2} \hat{B} \\
&= \left(\hat{Q} d_{\Omega,mn}^{-1/2} \hat{B} \right)' \hat{Q} d_{\Omega,mn}^{-1/2} \hat{B} \\
&= \left(\hat{M} \hat{Q} d_{\Omega,mn}^{-1/2} \right)' \hat{M} \hat{Q} d_{\Omega,mn}^{-1/2} \\
&= d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{M} \hat{Q} d_{\Omega,mn}^{-1/2}.
\end{aligned}$$

The rank of $\hat{M}$ is the trace of $\hat{M}$, which is $L - K$.

Since $d_{\Omega,mn}^{-1/2} \Omega_{mn} d_{\Omega,mn}^{-1/2}$ is positive definite, there exist a matrix Q_{mn} such that

$$Q'_{mn} Q_{mn} = \left[d_{\Omega,mn}^{-1/2} \Omega_{mn} d_{\Omega,mn}^{-1/2} \right]^{-1}.$$

It is easy to show that

$$d_\Omega^{1/2} \hat{\Omega}^{-1} d_\Omega^{1/2} = d_\Omega^{1/2} \Omega_{mn}^{-1} d_\Omega^{1/2} + o_p(1)$$

where

$$\left(d_\Omega^{-1/2} \Omega_{mn} d_\Omega^{-1/2} \right)^{-1} = \begin{pmatrix} 1 & \frac{\Omega_{1,12}}{\sqrt{n}\sqrt{\Omega_{1,11}\Omega_{22}}} \\ \frac{\Omega_{1,21}}{\sqrt{n}\sqrt{\Omega_{22}\Omega_{1,11}}} & 1 \end{pmatrix}.$$

It follows directly that $\hat{Q} = Q_{mn} + o_p(1)$.

Further, we have that

$$d_{\Omega,mn}^{-1/2} \sqrt{m} \bar{g}_{mn}(\delta) \xrightarrow{d} N\left(0, d_{\Omega,mn}^{-1/2} \Omega_{mn} d_{\Omega,mn}^{-1/2} \right).$$

Define $\hat{v}_{mn} = \sqrt{m} \hat{Q} d_{\Omega,mn}^{-1/2} \bar{g}_{mn}(\delta)$ and note that

$$\hat{v}_{mn} \xrightarrow{d} N(0, Q_{mn} d_{\Omega,mn}^{-1/2} \Omega_{mn} d_{\Omega,mn}^{-1/2} Q'_{mn}) = N(0, Q_{mn}(Q'_{mn}Q_{mn})^{-1}Q'_{mn}) = N(0, I_L).$$

Now we can come back to our test statistic:

$$
\begin{aligned}
J(\hat{\delta}) &= m \bar{g}_{mn}(\hat{\delta})' \hat{\Omega}^{-1} \bar{g}_{mn}(\hat{\delta}) \\
&= \left(\sqrt{m} d_{\Omega,mn}^{-1/2} \bar{g}_{mn}(\hat{\delta}) \right)' \left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]^{-1} \left(\sqrt{m} d_{\Omega,mn}^{-1/2} \bar{g}_{mn}(\hat{\delta}) \right) \\
&= \left(\sqrt{m} d_{\Omega,mn}^{-1/2} \hat{B} \bar{g}_{mn}(\delta) \right)' \left[d_{\Omega,mn}^{-1/2} \hat{\Omega} d_{\Omega,mn}^{-1/2} \right]^{-1} \left(\sqrt{m} d_{\Omega,mn}^{-1/2} \hat{B} \bar{g}_{mn}(\delta) \right) \\
&= \left(\sqrt{m} \bar{g}_{mn}(\delta) \right)' \hat{B}' \hat{\Omega}^{-1} \hat{B} \left(\sqrt{m} \bar{g}_{mn}(\delta) \right) \\
&= \left(\sqrt{m} \bar{g}_{mn}(\delta) \right)' d_{\Omega,mn}^{-1/2} \hat{Q}' \hat{M} \hat{Q} d_{\Omega,mn}^{-1/2} \left(\sqrt{m} \bar{g}_{mn}(\delta) \right) \\
&= \left(\sqrt{m} \hat{Q} d_{\Omega,mn}^{-1/2} \bar{g}_{mn}(\delta) \right)' \hat{M} \left(\sqrt{m} \hat{Q} d_{\Omega,mn}^{-1/2} \bar{g}_{mn}(\delta) \right) \\
&= \hat{v}'_{mn} \hat{M} \hat{v}_{mn}.
\end{aligned}
$$

Since $\hat{M}$ is idempotent with rank $L - K$, it follows that

$$J(\hat{\delta}) \xrightarrow{d} \chi^2_{L-K}.$$

$\blacksquare$

B Least Squares Panel Data Models

B.1 Formal results

This section complements subsection 2.3 by discussing more in detail the relationship between least squares estimator and the minimum distance approach. Throughout the section, we define

the $n \times (K_1+1)$ matrix of first-stage regressors $\tilde{X}_{1j} = (\tilde{x}_{1j}, \tilde{x}_{2j}, \ldots, \tilde{x}_{nj})'$, and the $mn \times K_1$ matrix of individual-level regressors $X_1 = (X_{1j}', \ldots, X_{1m}')'$. Further, we use the matrices $P_j = l(l'l)^{-1}l'$ and $Q_j = I_j - P_j$, where l is a $n \times 1$ vector of ones. Thus, $P_j X_j = \bar{X}_j$ and $Q_j X_{1j} = \dot{X}_{1j}$. We consider a linear version of our estimator, where OLS instead of quantile regression is used in the first stage and we focus on model (10). In this section, we show that mean models can be estimated using a two-step procedure. Notation is the same as in the paper, except that the fitted values are computed using an OLS regression. More precisely, the vector of fitted values of group j is

$$\hat{Y}_j = \tilde{X}_j \hat{\beta}_j = \tilde{X}_j \left(\tilde{X}_j' \tilde{X}_j \right)^{-1} \tilde{X}_j' Y_j.$$

The following Proposition states the equivalence of the two-step procedure using the fitted values and the conventional one-step estimator in mean models.

Proposition 5. *Denote $\hat{\delta}_{GMM}^{MD}$ the coefficient vector of a linear GMM regression of $\hat{Y}$ on X with instrument Z. Let $\hat{\delta}_{GMM}$ be the coefficient vector of the same GMM regression but with regressand Y. Assume that Z_j lies in the column space of $\tilde{X}_j$, then $\hat{\delta}_{GMM}^{MD} = \hat{\delta}_{GMM}$.*

The proof of this Proposition and all subsequent proofs are in Appendix B.2. Proposition 5 implies that any linear model can be computed by a two-step estimator as long as the matrix of instruments of group j, Z_j lies in the column space of the matrix of first-stage regressors of group j, $\tilde{X}_j$.[36] This result applies to a wide range of estimators. Since OLS is a special case of GMM, the result for pooled OLS follows directly, while the result for the within estimator is summarized in the following Corollary.

Corollary 1. *Denote $\hat{\delta}_{FE}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on X_1 with instruments $\dot{X}_1$. Let $\hat{\delta}_{FE}$ be the coefficient vector of the within estimator, that is, of a regression of $\hat{Y}$ on $\dot{X}_1$. Then $\hat{\delta}_{FE}^{MD} = \hat{\delta}_{FE}$.*

The between estimator is usually computed by regressing $\bar{Y}$ on $\bar{X}$. Alternatively, it can be estimated by an IV regression of Y (or $\hat{Y}$) on X using $\bar{X}$ as an instrument, where it exploits only the variation between individuals.

Corollary 2. *Denote $\hat{\delta}_{BE}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on X with instruments $\bar{X}$. Let $\hat{\delta}_{BE}$ be the coefficient vector of the between estimator, that is, of a regression of $\bar{Y}$ on $\bar{X}$. Then $\hat{\delta}_{BE}^{MD} = \hat{\delta}_{BE}$.*

It is worth noting that the IV approach to these panel data estimators also works in one stage with Y as the dependent variable. Further, it is clearly possible to estimate between (within) models using average (demeaned) fitted values and regressors.

The pooled OLS and the between estimators can estimate both β and γ but are not efficient. The random effects estimator optimally combines between and the within variation to find a more

---

[36] Since $\tilde{X}_j$ includes a constant, the presence of group-level variables in Z_j will not affect its column space.

efficient estimator. While FGLS is the most common estimator for the random effects model, Im et al. (1999) show that the overidentified 3SLS estimator, with instruments $Z_j = (\dot{X}_{1j}, \bar{X}_j)$, is identical to the random effects estimator. The 3SLS estimator is a special case of GMM with weighting matrix $W = \mathbb{E}[Z_j' \tilde{\Omega} Z_j]$ where $\tilde{\Omega}$ follows the usual random effects covariance structure. Thus, by Proposition 5, the random effects estimator can also be computed in two steps using the fitted values in the second stage.

Corollary 3. *Denote $\hat{\delta}_{RE}^{MD}$ the coefficient vector of a 3SLS regression of $\hat{Y}$ on X with instruments $(\dot{X}_{1j}, \bar{X}_j)$. Let $\hat{\delta}_{RE}$ be the coefficient vector of a random effects regression of Y on X. Then $\hat{\delta}_{RE}^{MD} = \hat{\delta}_{RE}$.*

Alternatively, the random effects estimator can be implemented using the theory of optimal instruments and a just identified 2SLS regression. Starting from a conditional moment restriction, the idea of optimal instruments is to select an instrument and weights that minimize the asymptotic variance (see, e.g. Newey, 1993). Relevant to our two-step procedure, under homoskedasticity of the errors, the conditional moments $\mathbb{E}[Y_j - X_j \delta | X_j] = 0$ and $\mathbb{E}[\hat{Y}_j - X_j \delta | X_j] = 0$ imply the same optimal instrument. To put it differently, under random effects assumptions, the matrix $\tilde{X}_j \frac{V_j}{n} \tilde{X}_j' + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2$ simplifies to the usual random effects structure.

Proposition 6. *Assume $\mathbb{E}[\varepsilon_{ij}^2 | X_j] = \sigma_\varepsilon^2$ and $\mathbb{E}[\alpha_j^2 | X_j] = \sigma_\alpha^2$. The conditional moments $\mathbb{E}[\hat{Y}_j - X_j \delta | X_j] = 0$ and $\mathbb{E}[Y_j - X_j \delta | X_j] = 0$ imply the same optimal instrument.*

The Hausman-Taylor model (Hausman and Taylor, 1981) is a middle ground between the fixed effects and the random effects models where some regressors are assumed to be uncorrelated with α_j. In contrast, no restriction is placed on the relationship between the other regressors and the unobserved heterogeneity. The matrix of regressors X is partitioned as $X = [X_1^{ex} \ X_1^{en} \ X_2^{ex} \ X_2^{en}]$ where X_1^{ex} and X_2^{ex} are orthogonal to α_j. No assumption is placed on the relationship between α_j and X_1^{ex} and X_2^{en}. The model can be estimated by IV using instruments $Z = (\dot{X}_1^{ex}, \dot{X}_1^{en}, \bar{X}_1^{ex}, X_2^{ex})$ (see, e.g., Hansen, 2022b). Thus, it follows by Proposition 5 that the Hausman-Taylor model can be estimated in two stages.

Corollary 4. *Denote $\hat{\delta}_{HT}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on X with instruments $(\dot{X}_1^{ex}, \dot{X}_1^{en}, \bar{X}_1^{ex}, X_2^{ex})$. Let $\hat{\delta}_{HT}$ be the coefficient vector of the Hausman Taylor Estimator based on a regression Y on X. Then $\hat{\delta}_{HT}^{MD} = \hat{\delta}_{HT}$.*

Finally, we show that not only the point estimates but also the standard errors can be obtained using the two-stage minimum distance approach. This requires clustering the standard errors in the second stage at a level weakly higher than the group j. Let $g = 1, \ldots, G$ index the clusters and assume that each of the clusters has N_g observations. This nests the case where one wishes to cluster at the individual level or at a higher level. For example, if j are county-year combinations, one might cluster at the county level. For an estimator $\hat{\delta}$ the clustered covariance

matrix is estimated by

$$\hat{V}_\delta = \left(\frac{1}{G} \sum_{g=1}^{G} X_g' Z_g \hat{W} \frac{1}{G} \sum_{g=1}^{G} Z_g' X_g \right)^{-1} \frac{1}{G} \sum_{g=1}^{G} X_g' Z_g \hat{W} \left(\frac{1}{G} \sum_{g=1}^{G} Z_g' \tilde{u}_g \tilde{u}_g' Z_g \right)$$

$$\cdot \hat{W} \frac{1}{G} \sum_{g=1}^{G} Z_g' X_g \left(\frac{1}{G} \sum_{g=1}^{G} X_g' Z_g \hat{W} \frac{1}{G} \sum_{g=1}^{G} Z_g' X_g \right)^{-1},$$

where $\tilde{u}_g$ is a N_g-dimensional vector of estimated errors for the observations in cluster g, X_g is the $K \times N_g$ matrix of regressors of cluster g, and Z_g is the $L \times N_g$ matrix of instruments.

Proposition 7. *Denote $\hat{V}_\delta$ the clustered covariance matrix of $\hat{\delta}$ estimated by a GMM regression of Y on X with instrument Z. Let $\hat{V}_{\delta^{MD}}$ be the clustered covariance matrix of $\hat{\delta}^{MD}$ estimated by GMM regression of $\hat{Y}$ on X with instrument Z, where $\hat{Y}$ are estimated by an OLS first-stage. Let the clusters be at weakly higher level than j. Then, $\hat{V}_{\delta^{MD}} = \hat{V}_\delta$.*

B.2 Proofs of the least squares results

Proof of Proposition 5. Define the projection matrix $\tilde{P} = \tilde{X}_j (\tilde{X}_j' \tilde{X}_j)^{-1} \tilde{X}_j'$. Since Z_j is in the column space of $\tilde{X}_j$,

$$\tilde{P} Z_j = Z_j \tag{81}$$

The MD estimator with a GMM second stage is:

$$\hat{\delta}_{GMM}^{MD} = \left(X' Z \hat{W} Z' X \right)^{-1} X' Z \hat{W} Z' \hat{Y}.$$

For $\hat{\delta}_{GMM}^{MD}$ to be equal to $\hat{\delta}_{GMM}$, it suffices that $Z' \hat{Y} = Z' Y$. Note that

$$
\begin{aligned}
Z' \hat{Y} &= \sum_{j=1}^{m} Z_j' \hat{Y}_j \\
&= \sum_{j=1}^{m} Z_j' \tilde{X}_j \hat{\beta}_j \\
&= \sum_{j=1}^{m} Z_j' \tilde{X}_j (\tilde{X}_j' \tilde{X}_j)^{-1} \tilde{X}_j' Y_j \\
&= \sum_{j=1}^{m} (\tilde{P} Z_j)' Y_j \\
&= \sum_{j=1}^{m} Z_j' Y_j = Z' Y,
\end{aligned}
$$

where the third line uses $\hat{Y}_j = \tilde{X}_j \hat{\beta}_j$, the fourth line uses the definition of the OLS estimator in the first stage and the last line uses equation (81). Thus, it follows directly that $\hat{\delta}_{MD}$ equals $\hat{\delta}_{GMM}$. ∎

73

Proof of Corollary 1. First, note that since $Q_j X_{1j} = \dot{X}_{1j}$, $\dot{X}_{1j}$ lies in the column space of X_{1j}. Then, we apply Proposition 5 and since $K = L$, the 2SLS estimator reduces to the IV estimator. It follows that a 2SLS (or IV) regression of $\hat{Y}$ on X_{1j} with instrument Z_j is algebraically identical to a 2SLS regression with Y_j as dependent variable. Then,

$$
\begin{aligned}
\hat{\delta}_{FE}^{MD} &= \left(\sum_{j=1}^{m} Z_j' X_{1j} \right)^{-1} \sum_{j=1}^{m} Z_j' Y_j \\
&= \left(\sum_{j=1}^{m} \dot{X}_{1j}' X_{1j} \right)^{-1} \sum_{j=1}^{m} \dot{X}_{1j}' Y_j \\
&= \left(\sum_{j=1}^{m} X_{1j}' Q_j X_{1j} \right)^{-1} \sum_{j=1}^{m} X_{1j}' Q_j Y_j \\
&= \left(\sum_{j=1}^{m} \dot{X}_{1j}' \dot{X}_{1j} \right)^{-1} \sum_{j=1}^{m} \dot{X}_{1j}' \dot{Y}_j = \hat{\delta}_{FE},
\end{aligned}
$$

where the second line follows since $Z_j = \dot{X}_{1j}$, the third and last line by $Q_j X_{1j} = \dot{X}_{1j}$, $Q_j Y_j = \dot{Y}_j$ and since Q_j is idempotent. $\blacksquare$

Proof of Corollary 2. First, note that since $P_j \tilde{X}_j = \bar{X}_j$, $\bar{X}_j$ lies in the column space of $\tilde{X}_j$. Then, we apply Proposition 5 and since $K = L$, the 2SLS estimator reduces to an IV estimator. It follows that a 2SLS regression of $\hat{Y}_j$ on X_j with instrument Z_j is algebraically identical to a 2SLS regression with Y_j as dependent variable. Then,

$$
\begin{aligned}
\hat{\delta}_{BE}^{MD} &= \left(\sum_{j=1}^{m} Z_j' X_j \right)^{-1} \sum_{j=1}^{m} Z_j' Y_j \\
&= \left(\sum_{j=1}^{m} \bar{X}_j' X_j \right)^{-1} \sum_{j=1}^{m} \bar{X}_j' Y_j \\
&= \left(\sum_{j=1}^{m} X_j' P_j X_j \right)^{-1} \sum_{j=1}^{m} X_j' P_j Y_j \\
&= \left(\sum_{j=1}^{m} \bar{X}_j' \bar{X}_j \right)^{-1} \sum_{j=1}^{m} \bar{X}_j' \bar{Y}_j = \hat{\delta}_{BE}
\end{aligned}
$$

where the second line follows since $Z_j = \bar{X}_j$, the third and last line by $P_j X_j = \bar{X}_j$, $P_j Y_j = \bar{Y}_j$ and, since P_j is idempotent. $\blacksquare$

Proof of Proposition 6. The optimal instrument takes the form $Z_j^* = \mathbb{E}[g_j(\delta) g_j(\delta)' | Z_j]^{-1} R_j(\delta, \tau)$, where $R_j(\delta, \tau) = \mathbb{E}[\frac{\partial}{\partial \delta} g_j(\delta, \tau) | Z_j]$. For both moment conditions, $R_j(\delta, \tau)$ is identical. Then for

the first moment restriction, we have:

$$\mathbb{E}[(\hat{Y}_j - X_j\delta)(\hat{Y}_j - X_j\delta)'|X_j] = \mathbb{E}[(\tilde{X}_j(\hat{\beta}_j - \beta) + \tilde{X}_j\beta - X_j\delta)(\tilde{X}_j(\hat{\beta}_j - \beta) + \tilde{X}_j\beta - X_j\delta)'|X_j]$$

$$\tag{82}$$

$$= \mathbb{E}[(\tilde{X}_j(\hat{\beta}_j - \beta) + \alpha_j)(\tilde{X}_j(\hat{\beta}_j - \beta) + \alpha_j)'|X_j]$$

$$= \tilde{X}_j \frac{V_j}{n} \tilde{X}_j' + \mathbf{l}_n \mathbf{l}_n' \sigma_\alpha^2.$$

The matrix $\tilde{X}_j \frac{V_j}{n} \tilde{X}_j' + \mathbf{l}_n \mathbf{l}_n' \sigma_\alpha^2$ is singular, so that we suggest using the Moore-Penrose inverse to construct the optimal instrument.

For the second moment restriction, we have:

$$\mathbb{E}[(Y_j - X_j\delta)(Y_j - X_j\delta)'|X_j] = \mathbb{E}[(\alpha_j + \varepsilon_{ij})(\alpha_j + \varepsilon_{ij})'|X_j]$$

$$= \left(\mathbf{I}_n \sigma_\varepsilon^2 + \mathbf{l}_n \mathbf{l}_n' \sigma_\alpha^2 \right).$$

Then note that $\left(\mathbf{I}_n \sigma_\varepsilon^2 + \mathbf{l}_n \mathbf{l}_n' \sigma_\alpha^2 \right)^{-1} = (\tilde{X}_j \tilde{X}_j^+ \sigma_\varepsilon^2 + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2)^+ = (\tilde{X}_j (\tilde{X}_j' \tilde{X}_j)^{-1} \tilde{X}_j' \sigma_\varepsilon^2 + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2)^+ = (\tilde{X}_j \frac{V_j}{n} \tilde{X}_j' + \mathbf{l}_n' \mathbf{l}_n \sigma_\alpha^2)^+ X_j$, where $V_j = (\frac{1}{n} \tilde{X}_j' \tilde{X}_j)^{-1} \sigma_\varepsilon^2$ and since for a full column rank matrix $\tilde{X}_j$, $\tilde{X}_j \tilde{X}_j^+ = I_n$ and $\tilde{X}_j^+ = (\tilde{X}_j' \tilde{X}_j)^{-1} \tilde{X}_j'$. ∎

Proof of Proposition 7. Define $Z_g = (z_{1g}, \ldots, z_{n_g g})'$, $X_g = (x_{1g}, \ldots, x_{n_g g})'$, $Y_g = (y_{1g}, \ldots, y_{n_g g})'$ and $\hat{Y}_g = (\hat{y}_{1g}, \ldots, \hat{y}_{n_g g})'$. The first and third terms of the expression are identical for both estimators. Thus, we focus on the middle term. Let $\hat{u}_g = Y_g - X_g \hat{\delta}$ be the vector of residuals from the regression using Y as dependent variable, and let $\hat{u}_g^{MD} = \hat{Y}_g - X_g \hat{\delta}^{MD}$ be the vector of residuals of the estimator using the fitted values as regressand. We show that $Z_g' \hat{u}_g = Z_g' \hat{u}_g^{MD}$ for all g. By Proposition 5, $\hat{\delta}^{MD} = \hat{\delta}$. Thus, the fitted values of both estimators are identical. Next, define $\breve{X}_g = \text{diag}\{\tilde{x}_{1g}, \ldots, \tilde{x}_{n_g g}\}$ and recall that regressing Y_g on $\breve{X}_g$ is the same as performing G separate regressions. Let $\breve{\beta}_g$ be the coefficient vector of a OLS regression of Y_g on $\breve{X}_g$. Note that Z_g is in the column space of $\tilde{X}_g$. Define the projection matrix $\breve{P} = \breve{X}_g (\breve{X}_g' \breve{X}_g)^{-1} \breve{X}_g'$. Since Z_j is in the column space of $\breve{X}_g$,

$$\breve{P} Z_g = Z_g. \tag{83}$$

Then,

$$Z_g' \hat{u}_g^{MD} = Z_g' \left(\hat{Y}_g - X_g \hat{\delta}^{MD} \right)$$

$$= Z_g' \tilde{X}_g \breve{\beta}_g - Z_g X_g \hat{\delta}$$

$$= Z_g' \tilde{X}_g (\breve{X}_g' \breve{X}_g)^{-1} \breve{X}_g' Y_g - Z_g X_g \hat{\delta}$$

$$= Z_g' (Y_g - X_g \hat{\delta}) = Z_g' \hat{u}_g,$$

where the fourth line follows by (83). Since this holds for all g, the desired result follows directly.
∎

C Optimal Instruments and Minimum Distance

In this section, we show that if $\alpha_j(\tau) = 0$ for all j and τ, efficient minimum distance can be implemented by optimal instruments. From equation (82) we have that if $\alpha_j(\tau) = 0$ for all j and all τ, $\mathbb{E}[(\tilde{X}_j\hat{\beta}_j(\tau) - X_j\delta(\tau))(\tilde{X}_j\hat{\beta}_j(\tau) - X_j\delta(\tau))'|X_j] = \tilde{X}_j\frac{V_j(\tau)}{n}\tilde{X}_j'$. This implies the optimal instrument $Z_j^* = (\tilde{X}_j\frac{V_j(\tau)}{n}\tilde{X}_j')^+X_j$. Since n is a scalar, using $Z_j^*(\tau) = (\tilde{X}_jV_j(\tau)\tilde{X}_j')^+X_j$ leads to the same results.

Proposition 8. *The IV regression with instrument $Z_j^*(\tau) = (\tilde{X}_jV_j(\tau)\tilde{X}_j')^+X_j$ is the efficient MD estimator.*

Proof.

$$
\begin{aligned}
\hat{\delta}_{EMD}(\tau) &= \left(\sum_{j=1}^m R_j'\hat{V}_j^{-1}(\tau)R_j\right)^{-1}\left(\sum_{j=1}^m R_j'\hat{V}_j^{-1}(\tau)\hat{\beta}_j(\tau)\right) \\
&= \left(\sum_{j=1}^m X_j'\tilde{X}_j\left(\tilde{X}_j'\tilde{X}_j\hat{V}_j(\tau)\tilde{X}_j'\tilde{X}_j\right)^{-1}\tilde{X}_j'X_j\right)^{-1}\left(X_j'\tilde{X}_j\left(\tilde{X}_j'\tilde{X}_j\hat{V}_j(\tau)\tilde{X}_j'\tilde{X}_j\right)^{-1}\tilde{X}_j'\hat{Y}_j(\tau)\right) \\
&= \left(\sum_{j=1}^m X_j'\left(\tilde{X}_j\hat{V}_j(\tau)\tilde{X}_j'\right)^+X_j\right)^{-1}\left(X_j'\left(\tilde{X}_j\hat{V}_j(\tau)\tilde{X}_j'\right)^+\hat{Y}_j(\tau)\right) = \hat{\delta}_{OI}(\tau).
\end{aligned}
$$

The second line follows as $\tilde{X}_jR_j = X_j$ and the third line follows since for a full column rank matrix $\tilde{X}_j$, $\tilde{X}_j^+ = (\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'$. ∎