Depth-Weighted Estimation of Heterogeneous Agent Panel Data Models*

YOONSEOK LEE† — Syracuse University

DONGGYU SUL‡ — University of Texas at Dallas

June 2020

Abstract

We develop robust estimation of panel data models, which is robust to various types of outlying behavior of potentially heterogeneous agents. We estimate parameters from individual-specific time-series and average them using data-dependent weights. In particular, we use the notion of data depth to obtain order statistics among the heterogeneous parameter estimates, and develop the depth-weighted mean-group estimator in the form of an L-estimator. We study the asymptotic properties of the new estimator for both homogeneous and heterogeneous panel cases, focusing on the Mahalanobis and the projection depths. We examine relative purchasing power parity using this estimator and cannot find empirical evidence for it.

Keywords: Panel data, Depth, Robust estimator, Heterogeneous agents, Mean group estimator.
JEL Classifications: C23, C33.

*First draft: September 2017. The authors thank to Robert Serfling and participants at numerous seminar/conference presentations for very helpful comments. Lee acknowledges support from Appleby-Mosher Research Fund, Maxwell School, Syracuse University.

†Corresponding author. *Address*: Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244. *E-mail*: ylee41@maxwell.syr.edu

‡*Address*: Department of Economics, University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080. *E-mail*: d.sul@utdallas.edu

1 Introduction

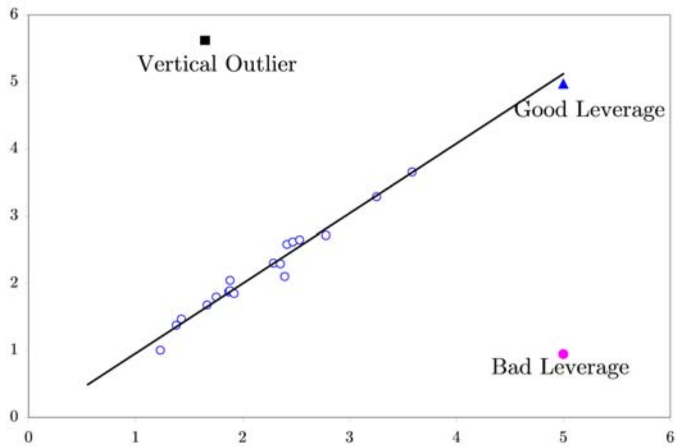
A robust estimator is a statistic that is less influenced by outliers. Many robust estimators are available for regression models, where the robustness is toward outliers in the regression error. Popular examples includes the least absolute deviation regression, which can be also seen as a special case of the quantile regression (Koenker and Bassett (1978)), the least median squares estimator (Rousseeuw (1984)), the least trimmed squares estimator (Agulló et al. (2008)), and the modified M-estimator (Yohai (1987); Kudraszow and Maronna (2011)), to name a few. See Dutta and Genton (2017) for more recent reference.

Unlike the standard regression case, studies on robust estimators for panel regression models are rather limited. Available studies can be classified into two broad categories depending on the method of eliminating unobserved heterogeneity. The first group of studies uses the standard approaches to eliminate fixed effects such as subtracting individual means or taking first difference, which includes Wagenvoort and Waldmann (2002), Cantoni (2004), Lucas et al. (2007), and Aquaroa and Čížek (2013, 2014). The second group of studies uses some robust approaches to eliminate fixed effects (e.g., subtracting individual medians), which includes Bramati and Croux (2007), Baltagi and Bresson (2012), and Dhaene and Zhu (2017). Once the unobserved heterogeneity is eliminated, these studies apply the well-known robust estimation methods to the panel regression, such as Huber (1964)'s robust M-estimation, and least median squares on the weighted sum of squared errors.

Such a limited number of studies in panel robust estimation stems from the ambiguity of the notion of outliers in the panel data structure. In the cross-sectional or time-series regression setup, two types of outliers are considered: vertical outliers (outliers from the regression error) and the bad leverage points (outliers from the regressor) as depicted in Figure 1, which can be readily identified from the scatter plot. In comparison, for the panel regression setup with potential parameter heterogeneity, where either the intercept or the slope parameters are different across individuals, it is not clear which observations are outliers. When a panel regression imposes heterogeneous slope parameters, for instance, some bad leverage points under homogeneous model can be good leverage points or even non-outliers under the heterogeneous case as depicted in Figure 2. Therefore, the panel regression can have different notions of outliers from the conventional regression case, and it is somewhat difficult to stylize the types of outliers or the sources of data contamination.

In this paper, we develop a novel approach of robust estimation that can be used exclusively for panel data regression, which utilizes the distinctive two-way structure of the panel data. The estimator is robust toward any types of outlying individuals in the panel data. The main idea relies on that existence of outlying individuals is very likely to yield heterogeneity in the individual-specific time series estimates. Importantly, the new estimator uses the resulting heterogeneity in the individual-specific estimates directly to construct a robust

Figure 1: Typical Regression Outliers

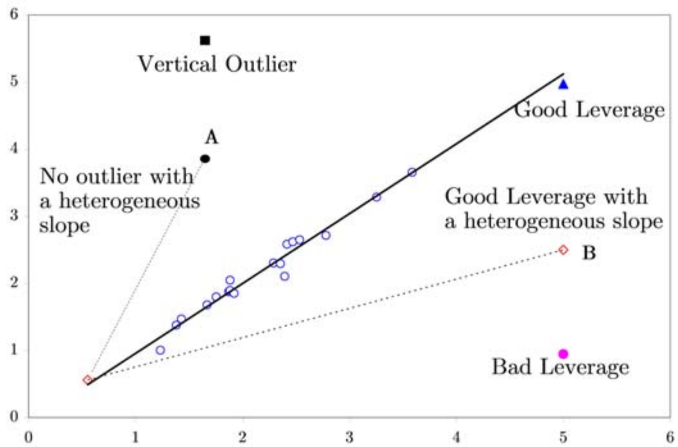


estimator without identifying the source of the heterogeneity. Hence, unlike the conventional robust regression estimation, we do not need to distinguish the types of outliers and consider a specific strategy for each type. For the homogeneous parameter case, it will yield a consistent estimator toward the true parameter as long as the individual-specific estimators are consistent. For the heterogeneous parameter case, it will yield a consistent estimator toward the center of the underlying distribution of the true heterogeneous parameters. We can use this unified approach whether the true model is homogeneous or heterogeneous. This is the strength of our estimator, which is novel in robust regression studies.

In particular, we develop the depth-weighted mean-group (DWMG) estimator, which robustly estimates the central tendency of potentially heterogeneous parameters in panel data models. Similar to the mean-group estimator, we estimate the parameters for each individual time series, even when we have a homogeneous panel data model. We then average these heterogeneous estimates using data-dependent weights, where the weights are based on an outlyingness measure of the multi-dimensional estimates. More precisely, we use the notion of data depth to form order statistics among the heterogeneous parameter estimates, which is based on some nonparametric distance of each estimate toward the center of the true parameter distribution (e.g., Zuo and Serfling (2000)). The resulting estimator is obtained as an L-estimator based on the depth order, and it is robust towards various sources of outliers. In this sense, this estimation method can be understood as a robust way to combine individual-specific parameter estimators.

The idea of depth-weighted mean was originally developed in a different setup. As a robust measure of the central tendency of multivariate random samples, Liu (1990), Liu

Figure 2: Panel Regression Outliers with Heterogeneous Slopes



et al. (1999), and Zuo et al. (2004) consider a weighted mean whose weighting scheme is determined by the statistical depth. However, there is no studies applying this idea to the unobserved parameters or to the observations embedding errors. In this aspect, the DWMG estimator also contributes to this statistics literature by extending the idea of depth-weighted mean to estimation of the central tendency of unobserved parameters. By focusing on the two popular statistical depth functions, the Mahalanobis and the projection depths, we show that the DWMG estimator consistently estimates the population depth-weighted mean, which is a robust measure of the central tendency of heterogeneous parameters, as long as both the cross-section size n and the time-series length T are large regardless of their relative size. We also derive its limiting distribution, based on which the inference of the DWMG estimator is studied.

The remainder of the paper is organized as follows. Section 2 provides an overview of the statistical depth function and introduces the DWMG estimator in the panel data model. Sections 3 studies the asymptotic properties of the DWMG estimator for the heterogeneous and for the homogeneous panel models, respectively. Section 4 considers two examples based on the Mahalanobis and the projection depths and summarizes DWMG estimation procedure. Section 5 examines the finite sample performance of the DWMG estimators by Monte Carlo simulations. Section 6 applies the DWMG estimator to examine relative purchasing power parity with 27 bilateral exchange rates. Section 7 concludes with some remarks. All the proofs are in the Appendix.

2 Depth-Weighted Estimation

2.1 Data depth and depth-weighted mean

Unlike the univariate case, defining a central tendency for a multivariate distribution is not straightforward. Among various approaches, it is popular to define a robust center of a distribution using *data depth* in the nonparametrics literature. The data depth measures the outlyingness of a given multivariate sample point with respect to its underlying joint distribution. It hence leads to a center-outward ordering of each sample point. The deepest point, as a robust measure of the center of a multivariate distribution, is defined as the point with the largest data depth.

In particular, the data depth is formulated as an index between 0 and 1. If the data point is at the center of the distribution, then the depth value of the data point becomes unity. If a data point locates very far from the center, then the depth value of the point becomes near zero. For example, for a radially symmetric multivariate distribution, the data depth at the median is equal to one. In practice, the data depth is used for multivariate ordering, robust estimation, and outlier detection.

Data depths are calculated using statistical depth functions. Zuo and Serfling (2000) define a statistical depth function, which have the following set of axioms. Let Z be a $k \times 1$ generic random vector with a joint distribution $F_Z \in \mathcal{F}$.

Definition 1 (Statistical Depth Function) *A statistical depth function of Z at z, denoting $\mathcal{D}(z; F_Z) : \mathbb{R}^k \times \mathcal{F} \rightarrow [0, 1]$, satisfies*

- *Affine Invariance: $\mathcal{D}(Az + c; F_{AZ+c}) = \mathcal{D}(z; F_Z)$ for any nonsingular $k \times k$ matrix A and any $k \times 1$ vector c.*
- *Maximality at Center: The deepest point z^* satisfies $\mathcal{D}(z^*; F_Z) = \sup_{z \in \mathbb{R}^k} \mathcal{D}(z; F_Z)$.*
- *Monotonicity Relative to Deepest Point: $\mathcal{D}(z; F_Z) \leq \mathcal{D}(z^* + \tau(z - z^*); F_Z)$ for any $\tau \in [0, 1]$.*
- *Vanishing at Infinity: $\mathcal{D}(z; F_Z) \rightarrow 0$ as $||z|| \rightarrow \infty$.*

The first property implies that the statistical depth should not be changed by the scale and the location. The second property of the maximality at the center (i.e., the deepest point) is natural. The third property implies that the depth at z should decrease monotonically as it moves away from the deepest point. Hence, the τth quantile depth curve is never crossing with the $(\tau + \epsilon)$th quantile depth curve for any $\epsilon > 0$. In other words, by connecting the equi-depth data points, we have a well-defined contour plot over the joint distribution. The last property is also important for the robustness against outliers.

As noted above, the statistical depth function is closely related to the notion of the outlyingness, which measures how far a data point is located from the center of the joint distribution. In particular, Zuo and Serfling (2000) show that the following type of the depth function well satisfies the definition above:

$$\mathcal{D}(z; F_Z) = \frac{1}{1 + \mathcal{O}(z; F_Z)},$$

where $\mathcal{O}(z; F_Z) \geq 0$ is the outlyingness function. If the outlyingness of a point is equal to zero, then its depth becomes unity, which is the maximum depth.

Among various outlyingness measures available, we focus on two popular outlyingness functions in this paper: the Mahalanobis and the projection outlyingness. Consider random samples $Z_i \in \mathbb{R}^k$ for $i = 1, \ldots, n$ from a distribution F_Z. The Mahalanobis distance (e.g., Mahalanobis (1936)) or outlyingness of the ith observation Z_i is define as

$$\mathcal{O}^m(Z_i; F_Z) \equiv (Z_i - \mu_Z)' \, \Sigma_Z^{-1} \, (Z_i - \mu_Z) \tag{1}$$

and the Mahalanobis depth is given as $\mathcal{D}^m(Z_i; F_Z) = (1 + \mathcal{O}^m(Z_i; F_Z))^{-1}$, where μ_Z and Σ_Z are some location and scale parameters of F_Z. They are typically chosen as the mean and the variance of Z_i, respectively.[1] The projection outlyingness (e.g., Liu (1992)) is defined as

$$\mathcal{O}^p(Z_i; F_Z) \equiv \sup_{v:||v||=1} \frac{|v'Z_i - \mu_{v'Z}|}{\sigma_{v'Z}} \tag{2}$$

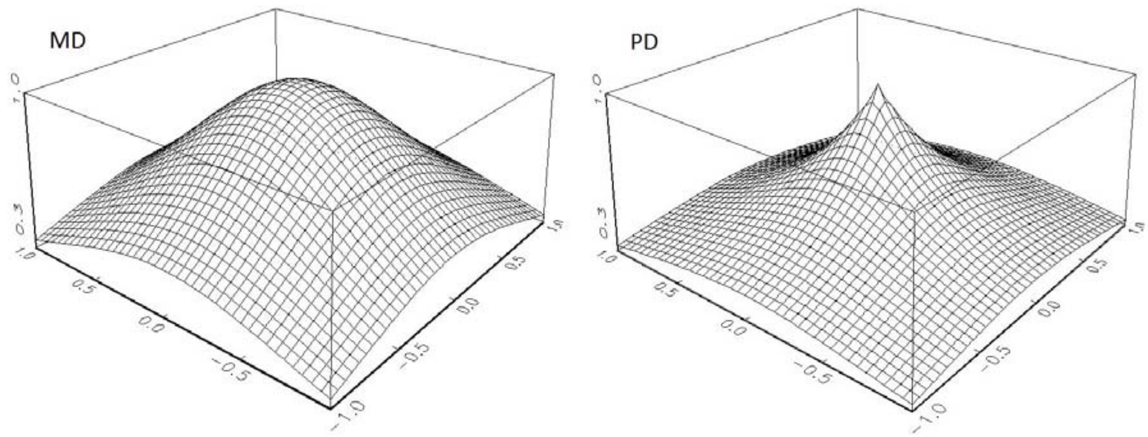
and the projection depth is given as $\mathcal{D}^p(Z_i; F_Z) = (1 + \mathcal{O}^p(Z_i; F_Z))^{-1}$, where v is a $k \times 1$ nonrandom vector with $||v|| = 1$ and $\mu_{v'Z}$ and $\sigma_{v'Z}$ are some location and scale parameters of $F_{v'Z}$, such as the median and the median absolute deviation (MAD) of $v'Z_i$, respectively.[2] For the scalar case, if we use the median and the MAD, the projection outlyingness is simply given as $\mathcal{O}^p(Z_i; F_Z) = |Z_i - \mathsf{med}[Z_i]| / \mathsf{MAD}[Z_i]$, where $\mathsf{med}[Z_i]$ and $\mathsf{MAD}[Z_i]$ are the median and MAD of Z_i, respectively.

As an illustration, Figure 3 depicts these two depth functions for the case of bivariate standard normal, where the Mahalanobis depth (MD; $\mathcal{D}^m(z; F_Z)$) is based on the mean and the variance, whereas the projection depth (PD; $\mathcal{D}^p(z; F_Z)$) is based on the median and the MAD. Note that the PD function satisfies all the defining properties in Definition 1 regardless of the symmetry of the distributions. Meanwhile, the MD function satisfies the properties of

[1] In this case, however, the Mahalanobis depth is not robust, in the sense that it does not have a bounded influence function nor have a high breakdown point. This is because of the lack of robustness of the mean μ_Z itself (see e.g., Liu and Singh (1993)).

[2] When $v'z - \mu_{v'Z} = 0$ and $\sigma_{v'Z} = 0$, we define $\mathcal{O}^p(z; F_Z) = 0$.

Figure 3: MD and PD of bivariate standard normal



the statistical depth function only when the distributions are elliptically symmetric. More precisely, it is a statistical depth function in the sense of Definition 1 above only when (i) the distribution F_Z is symmetric, (ii) the location and scale parameters μ_Z and Σ_Z are affine invariant, and (iii) μ_Z agrees with the point of symmetry of F_Z.[3] But the MD is much easier to calculate compared to the PD; as the dimension of k increases, the computation cost of the PD increases exponentially.[4] Throughout the paper, we assume all the location and scale parameters exist and they are affine invariant, so that the depth functions are well defined.

As a robust measure of the central tendency, that is more general than the deepest point, Liu (1990), Liu et al. (1999), and Zuo et al. (2004) study a weighted mean whose weighting scheme is determined by the statistical depth function. More precisely, they define the *depth-weighted mean* as

$$\frac{\int zW(\mathcal{D}(z,F_Z))F_Z(dz)}{\int W(\mathcal{D}(z,F_Z))F_Z(dz)}, \tag{3}$$

where $W(\cdot)$ is some positive weight function. The depth-weighted mean is well defined provided $\int W(\mathcal{D}(z,F_Z))F_Z(dz) > 0$ and $\int ||z||W(\mathcal{D}(z,F_Z))F_Z(dz) < \infty$, and it is general enough to include the popular centrality measures. When the joint distribution is symmetric, for instance, the depth-weighted mean corresponds to the mean (if it exists) or the median of Z_i. Using an empirical data depth $\mathcal{D}(z, F_{Z,n})$, where $F_{Z,n}$ is the empirical distribution of Z_i, we

[3]See Theorems 2.9 and 2.10 in Zuo and Serfling (2000) for the details. Recall that, for any affine transformation $AZ+c$ of Z, a location measure μ_Z is affine equivariant if $\mu_{AZ+c} = A\mu_Z + c$; a covariance measure Σ_Z is affine equivariant if $\Sigma_{AZ+c} = A\Sigma_Z A'$.

[4]See Zuo and Lai (2011) and the R-package 'ExPD2D' for various ways to construct the projection depth, including the fixed random sampling method.

can obtain the depth-weighted mean estimator in the form of an L-estimator given as

$$\frac{\sum_{i=1}^{n} W\left(\mathcal{D}(Z_i, F_{Z,n})\right) Z_i}{\sum_{i=1}^{n} W\left(\mathcal{D}(Z_i, F_{Z,n})\right)}. \tag{4}$$

2.2 Depth-weighted mean-group estimator

As Lee and Sul (2020) point out, panel pooled estimators, which presume the homogeneity restriction on slope coefficients, can be rewritten as weighted mean-group estimators, where weights are proportional to time series variances of regressors. Under the presence of some outlying individuals, however, panel pooled estimators generally perform poorly since the weights do not reflect outlying behaviors of the potentially heterogeneous individual estimators.

We develop a generalized mean-group estimator in panel data models by extending the idea of the depth-weighted mean estimator in (4). We consider a panel regression given by

$$y_{it} = x_{it}'\beta_i + \lambda_i' f_t + u_{it} \tag{5}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$, where the slope parameters $\beta_i \in \mathcal{B} \subset \mathbb{R}^k$ are "potentially" heterogeneous. x_{it} is a $k \times 1$ vector of exogenous regressors of interest, λ_i is a vector of factor loadings, and f_t is a vector of common factors. Without loss of generality, we let $T_i = T$ for all T. It is convenient to rewrite the model as

$$y_{it}^* = x_{it}^{*\prime}\beta_i + u_{it}^*, \tag{6}$$

where y_{it}^*, x_{it}^*, and u_{it}^* are the projected residuals on the individual-specific dummies and estimated common factors (e.g., Pesaran (2006); Bai (2009)). When $\lambda_i = (\alpha_i, 1)'$ and $f_t = (1, \tau_t)'$, the regression in (5) becomes the standard two-way fixed effect regression model, and (6) corresponds to the regression model after the within transformation or the concentrated quasi-likelihood: $x_{it}^* = \widetilde{x}_{it} - n^{-1}\sum_{j=1}^{n} \widetilde{x}_{jt}$ with $\widetilde{x}_{it} = x_{it} - T^{-1}\sum_{s=1}^{T} x_{is}$.

The slope parameter β_i can be either homogeneous (i.e., $\beta_i = \beta$ for all i) or heterogeneous, which is unknown. Whether β_i is homogeneous or not, we obtain its estimator $\widehat{\beta}_i$ from the individual-specific time series regression in (6) for each i. We define the *depth-weighted mean-group (DWMG) estimator* as

$$\widehat{\beta}_{DW} = \frac{\sum_{i=1}^{n} \widehat{\beta}_i W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))}{\sum_{i=1}^{n} W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))} = \frac{\int b W(\mathcal{D}(b, \widehat{F}_n))\widehat{F}_n(db)}{\int W(\mathcal{D}(b, \widehat{F}_n))\widehat{F}_n(db)} \equiv L(\widehat{F}_n), \tag{7}$$

for some positive weight function $W : \mathbb{R} \to [0,1]$, where

$$\widehat{F}_n(b) = \frac{1}{n}\sum_{i=1}^{n} 1\{\widehat{\beta}_i \le b\} \text{ for any } b \in \mathcal{B}$$

and $\mathcal{D}(\cdot, \widehat{F}_n) \in [0,1]$ is the data depth of $\{\widehat{\beta}_i\}_{i=1}^n$. $1\{\cdot\}$ is the binary indicator function. $\widehat{\beta}_{DW}$ can be understood as an L-estimator of generated variables $\{\widehat{\beta}_i\}_{i=1}^n$, in which the order statistic is based on the data depth of each $\widehat{\beta}_i$.[5] Though we consider the linear regression model (5) with exogeneous x_{it} mainly for the sake of presentation simplicity, we easily obtain the DWMG estimator in (7) for more general cases such as nonlinear panel data models and models with sequentially exogeneous or endogeneous regressors, as long as we have the individual estimator $\widehat{\beta}_i$ satisfying Assumption 2 below.

When β_i is indeed heterogeneous, it is important to note that the depth-weighted estimator (7) is different from the following functional:

$$L(F_n) = \frac{\int bW(\mathcal{D}(b, F_n))F_n(db)}{\int W(\mathcal{D}(b, F_n))F_n(db)}, \tag{8}$$

where F_n is the empirical distribution function of β_i. When β_i's are observable, the definition (8) is the standard form of the depth-weighted estimator in the literature like equation (2.1) of Zuo et al. (2004) or as in (4) above. However, we cannot observe β_i's here and we define the depth-weighted estimator using the estimators $\widehat{\beta}_i$ instead, which impose estimation errors. In addition, we do not need to assume heterogeneous β_i when defining the depth-weighted estimator $\widehat{\beta}_{DW}$ in (7).

3 Asymptotics of DWMG Estimator

3.1 Heterogeneous case

Individual responses are often very heterogeneous, which is described as the heterogeneous slope parameter β_i in the regression (5). When one is interested in finding the central tendency of the heterogeneous responses, however, it is not straightforward to identify it especially when the response is multi-dimensional. As a robust measure of the central tendency

[5] The data depth $\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n)$ itself is of interest when we want to identify the outlying individual i's based on their multivariate $\widehat{\beta}_i$ values. It can be used to identify groups or to generate multivariate-version quantiles in $\widehat{\beta}_i$ based on data depth contours.

of β_i, we consider the depth-weighted mean of β_i defined as

$$\overline{\beta} \equiv L(F) = \frac{\int bW(\mathcal{D}(b,F))F(db)}{\int W(\mathcal{D}(b,F))F(db)}, \tag{9}$$

similarly as (3), where F is the joint distribution function of $\beta_i \in \mathbb{R}^k$ and $W : \mathbb{R} \to [0,1]$ is some positive weight function that is used in (7). The statistical depth function $\mathcal{D}(b,F) \in [0,1]$ is defined for F and hence it measures how much each heterogeneous slope parameter vector β_i is distant from the center of its distribution. This definition of the centrality is general enough to include the popular centrality measures. For example, if $\mathbb{E}[\beta_i]$ exists, then $\overline{\beta} = \mathbb{E}[\beta_i]$ when $W(\cdot) = 1$, which is the parameter of interest in the typical random coefficient models. If we define $W(\cdot)$ as a trimming form, then $\overline{\beta}$ can describe the trimmed (depth-weighted) mean. When β_i has a density function that is symmetric about its mode, $\overline{\beta}$ is the median if we define $W(\cdot)$ as an indicator function at the maximal point. Apparently, under the homogeneous slope parameter setup (i.e., $\beta_i = \beta$ for all i), β_i has point mass at β and hence $\overline{\beta} = \beta$.

In the heterogeneous case, we assume the following conditions.

Assumption 1 *(i) β_i is a random draw from a distribution F, whose density function is bounded above zero and elliptically symmetric around zero. (ii) β_i is independent of (x_{it}^*, u_{it}^*) for all i and t.*

Assumption 1-(i) does not require that the moment of β_i exists, and hence it is more general than the conditions of the typical random coefficient models (cf. Swamy (1970)). Assumption 1-(ii) imposes that β_i is not correlated with the regression error u_{it}, which excludes any endogeneity issue of the random coefficient. In addition, unlike the fixed effect, the random coefficient β_i is not allowed to be correlated with the regressor x_{it}, which follows the standard assumption in the random coefficient panel regression models in the literature (e.g., Hsiao and Pesaran (2008)).

In order to study the asymptotic properties of the depth-weighted estimator $\widehat{\beta}_{DW}$ in (7) when β_i is heterogeneous, we suppose that an estimator $\widehat{\beta}_i$ of β_i from the individual-specific time series regression in (6) satisfies the following conditions.

Assumption 2 *(i) $\{x_{it}^*, u_{it}^*, \beta_i\}$ are cross-sectionally independent for sufficiently large n and T. (ii) There exists an increasing sequence of J_T with T such that*

$$\widehat{\beta}_i - \beta_i = O_p(J_T^{-1/2}) \quad \textit{for each } i. \tag{10}$$

Assumption 2-(i) still allows for cross-sectional dependence in x_{it} or u_{it} such as the common correlated effect models of Pesaran (2006). However, after common factors are controlled out, the residuals x_{it}^* and u_{it}^* should be cross-sectionally independent in the limit. Assumption 2-(ii) imposes high level conditions as in Pesaran (2006) or Bai (2009). However, whether β_i is heterogeneous or not, the $\sqrt{J_T}$-consistency of $\widehat{\beta}_i$ generally holds for the standard panel regression with the least squares estimator $\widehat{\beta}_i = (\sum_{t=1}^T x_{it}^* x_{it}^{*\prime})^{-1} \sum_{t=1}^T x_{it}^* y_{it}^*$ in (6) and $J_T = T$ when the regressors are exogeneous. It also allows for weak serial dependence in x_{it}^* and u_{it}^*. Note that we can consider more general models including nonlinear models as long as we have an M-estimator or a GMM estimator $\widehat{\beta}_i$ satisfying (10).[6]

Under Assumption 2, we can show that using an estimator $\widehat{\beta}_i$ instead of the true β_i in estimating the distribution function F (i.e., $\widehat{F}_n$ instead of F_n) still yields a consistent estimator of F as $n, T \to \infty$. Furthermore, we can obtain limiting distribution of $\sqrt{n}(\widehat{F}_n(b) - F(b))$ with further smoothness condition of F. Based on this result, we will show that $\widehat{\beta}_{DW} = L(\widehat{F}_n)$ in (7) is consistent to $\overline{\beta} = L(F)$ in (9) and satisfies asymptotic normality.

Theorem 1 *Suppose Assumptions 1 and 2 hold. Then,* $\sup_{b \in \mathcal{B}} |\widehat{F}_n(b) - F(b)| = o_p(1)$ *as* $n, T \to \infty$*. Furthermore, if* $\sqrt{n}/J_T \to \phi$ *for some constant* $0 \leq \phi < \infty$ *as* $n, T \to \infty$*, and* F *is twice continuously differentiable at* $b \in \mathcal{B}$ *with bounded derivatives, we have*

$$\sqrt{n} \left(\widehat{F}_n(b) - F(b) - J_T^{-1} tr[\ddot{F}(b) \overline{\Omega}/2] \right) \to_d \mathcal{N}\left(0, F(b)[1 - F(b)]\right) \tag{11}$$

as $n, T \to \infty$*, where* $\overline{\Omega} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n var(J_T^{1/2}(\widehat{\beta}_i - \beta_i))$ *and* $\ddot{F}(b)$ *is the Hessian matrix of* $F(b)$*.*

We allow for (conditionally) heteroskedastic regression error u_{it} in the definition of $\overline{\Omega}$, which is bounded from (10). Theorem 1 implies that $\widehat{F}_n(b) - F(b) = O_p(n^{-1/2} + J_T^{-1})$ and hence we need both large n and T to achieve consistency. This is because we need $\widehat{\beta}_i \to_p \beta_i$ as $T \to \infty$ so that we can use $\widehat{\beta}_i$ instead of β_i to estimate the true distribution of β_i. However, the rate of convergence of $\widehat{\beta}_i$ toward β_i is not important as long as it is consistent. Note

[6] For instance, we can consider M-estimation given by

$$\min_{\{\beta_i\}, \{\lambda_i\}, \{f_t\}} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \psi\left(w_{it}; \beta_i, \lambda_i, f_t\right)$$

for panel observations $\{w_{it}\}$ and some objective function ψ. Examples include panel linear regression models (e.g., Pesaran (2006)), panel binary choice models (e.g., Boneva and Linton (2016)), and panel quantile regression models (e.g., Harding et al. (2020)) with fixed effects and potentially heterogeneous slope parameters. However, we may need some bias corrections for nonlinear models when time fixed effects or interactive fixed effects present.

that (11) can be explained from $\sqrt{n}(\widehat{F}_n(\cdot) - F(\cdot)) = \sqrt{n}(F_n(\cdot) - F(\cdot)) + \sqrt{n}(\widehat{F}_n(\cdot) - F_n(\cdot))$, where the first term satisfies the standard Functional Central Limit Theorem (FCLT) and the second term is $O_p(\sqrt{n}/J_T)$. Therefore, we need $\sqrt{n}/J_T \to 0$ to obtain the standard FCLT of the empirical process. On the other hand, when $\sqrt{n}/J_T \to \phi$ for some non-zero constant $0 < \phi < \infty$, we need to correct the bias in the limiting distribution as in (11).[7] As a corollary, general results regarding $Q(F)$ and its estimator $Q(\widehat{F}_n)$ is given in Corollary A in the Appendix A.1, where $Q(\cdot)$ is a regular function that is Hadamard differentiable at F with a bounded derivative.

We now let

$$\widehat{\nu}_n(\cdot) = \sqrt{n}(\widehat{F}_n(\cdot) - F(\cdot)) \quad \text{and} \quad \widehat{H}_n(\cdot) = \sqrt{n}(\mathcal{D}(\cdot, \widehat{F}_n) - \mathcal{D}(\cdot, F))$$

and assume the following conditions. These conditions are similar to the conditions (A1)-(A4) of Zuo et al. (2004), which are satisfied for popular depth functions including the Mahalanobis and the projection depths. We define $\mathcal{B}_0 = \{b : \mathcal{D}(b, F) \geq d_0\}$ for some $d_0 \geq 0$.

Assumption 3 *We assume:*

(i) $W(\cdot)$ *is continuously differentiable with a bounded derivative* $\dot{W}(\cdot)$*; and* $W(d) = 0$ *for* $d \in [0, d_0\kappa]$ *with some* $\kappa > 1$*.*

(ii) $\int W(\mathcal{D}(b, F))F(db) > 0$, $\int ||b||W(\mathcal{D}(b, F))F(db) < \infty$, *and* $\int ||b||^2(W(\mathcal{D}(b, F)))^2F(db) < \infty$.

(iii) There exists $h(b, c)$ *such that* $\widehat{H}_n(b) = \int h(b, c)\widehat{\nu}_n(dc) + o_p(1)$ *uniformly on* $\mathcal{B}_0$.

(iv) $\int[\int ||b|||\dot{W}(\mathcal{D}(b, F))h(b, c)|F(db)]^2F(dc) < \infty$ *and* $\int ||b||^2(\dot{W}(\mathcal{D}(b, F)))^2F(db) < \infty$.

(v) $\sup_{b\in\mathcal{B}} |\widehat{H}_n(b)| = O_p(1)$ *and* $\sup_{b\in\mathcal{B}_0} ||b|||\widehat{H}_n(b)| = O_p(1)$.

(vi) $|\mathcal{D}(\widehat{\beta}_i, F) - \mathcal{D}(\beta_i, F)| = O_p(||\widehat{\beta}_i - \beta_i||)$ *and* $|h(c, \widehat{\beta}_i) - h(c, \beta_i)| = O_p(||\widehat{\beta}_i - \beta_i||)$ *for any* $c \in \mathcal{B}$.

Assumption 3-(i) supposes a sufficiently smooth weight function W in a neighborhood of zero. The first two conditions in Assumption 3-(ii) ensures that the depth-weighted mean $\overline{\beta}$ is well-defined; the last condition is required for the existence of the covariance matrix of the depth-weighted estimator. For each $b \in \mathcal{B}$, we may suppose that $\mathcal{D}(b, \cdot)$ is Fréchet differentiable at F with respect to the supremum metric $|| \cdot ||_\infty$, where the derivative is bounded and $||\widehat{F}_n - F||_\infty = o_p(1)$. Then, from Proposition 2.19 of Huber and Ronchetti

[7] It should be noted that, however, the bias term in (11) can be ignored when $\ddot{F}(b) = 0$. For instance, if the density function of β_i is symmetric around the specific location b, then $\ddot{F}(b) = 0$ and hence we can ignore this bias at b.

(2009), the Fréchet differentiability ensures the existence of bounded function $h(b,c)$ such that the Fréchet derivative $\dot{\mathcal{D}}(b,F;\widehat{F}_n - F)$ satisfies $\dot{\mathcal{D}}(b,F;\widehat{F}_n - F) = \lim_{\epsilon\to 0}\epsilon^{-1}[\mathcal{D}(b,F+\epsilon(\widehat{F}_n - F)) - \mathcal{D}(b,F)] = \int h(b,c)(\widehat{F}_n - F)(dc)$ for each $b \in \mathcal{B}$, which yields the expression of $\widehat{H}_n(b)$ in Assumption 3-(iii). The first condition in Assumption 3-(iv) ensures that the Fréchet derivative $\dot{\mathcal{D}}(b,F;\widehat{F}_n - F)$ is bounded in the sense that $\int b\dot{W}(\mathcal{D}(b,F))\dot{\mathcal{D}}(b,F;\widehat{F}_n - F)F(db) < \infty$; the second condition is new and it is needed to control for the estimation error from using $\widehat{\beta}_i$ instead of β_i in the limiting distribution of $\widehat{\beta}_{DW}$. Assumption 3-(v) can be verified under some moment restrictions for the Mahalanobis and the projection depths (cf. Zuo, 2003, Theorem 2.2; Zuo et al, 2004, Lemma 3.1). Assumption 3-(vi) readily holds when $\mathcal{D}(b,F)$ and $h(c,b)$ are continuous in b for any c. Note that $|\mathcal{D}(\widehat{\beta}_i,F) - \mathcal{D}(\beta_i,F)| = O_p(J_T^{-1/2})$ and $|h(c,\widehat{\beta}_i) - h(c,\beta_i)| = O_p(J_T^{-1/2})$ since $\widehat{\beta}_i - \beta_i = O_p(J_T^{-1/2})$ in (10).

The following theorem shows the asymptotic properties of the DWMG estimator $\widehat{\beta}_{DW}$. We see that $\widehat{\beta}_{DW}$ is $\sqrt{n}$-consistent to $\overline{\beta}$ and its limiting distribution is mean-zero normal as $n,T\to\infty$. The form of $h(\cdot,\cdot)$ in (12) is to be given in Section 4 for the Mahalanobis and the projection depth functions.

Theorem 2 (Heterogeneous Case) *Under Assumptions 1-3,* $\widehat{\beta}_{DW} - \overline{\beta} = O_p(n^{-1/2})$ *and* $\sqrt{n}(\widehat{\beta}_{DW} - \overline{\beta}) \to_d \mathcal{N}(0,V_F)$ *as* $n,T\to\infty$*, where* $V_F = \int K_F^0(b)K_F^0(b)'F(db)$ *with*

$$K_F^0(b) = \frac{\int c\dot{W}(\mathcal{D}(c,F))h^0(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))}{\int W(\mathcal{D}(c,F))F(dc)} \tag{12}$$

and $h^0(c,b) = h(c,b) - \int h(c,b)F(db)$*.*

Theorem 2 shows that, even for the static panel case, we need both large n and T to obtain the limiting distribution of the depth-weighted estimator $\widehat{\beta}_{DW}$. To understand this point, we decompose

$$\sqrt{n}(\widehat{\beta}_{DW} - \overline{\beta}) = \sqrt{n}(L(\widehat{F}_n) - L(F)) = \sqrt{n}(L(F_n) - L(F)) + \sqrt{n}(L(\widehat{F}_n) - L(F_n)) \tag{13}$$

using expressions in (9), (7), and (8). The first term in (13) satisfies the standard Central Limit Theorem (CLT), whereas the second term is $O_p(J_T^{-1/2})$.[8] It is important to note that,

[8] For the case of Mahalanobis depth and β_i with the bounded second moment, we can verify that the second term in (13) is asymptotically normal with zero mean. In this case, we can instead obtain the limiting distribution of $\widehat{\beta}_{DW}$ only using n-asymptotics:

$$\sqrt{n}(\widehat{\beta}_{DW} - \overline{\beta}) \to_d \mathcal{N}(0, V_F + J_T^{-1}V_T) \quad \text{as } n\to\infty \tag{14}$$

for some $0 < V_T < \infty$.

unlike Theorem 1 or Corollary A, we do not need the condition $\sqrt{n}/J_T \to 0$ to have the mean-zero limiting distribution of $\widehat{\beta}_{DW}$.

This result, on the other hand, does not mean that we also need $n, T \to \infty$ for consistency of $\widehat{\beta}_{DW}$; we only need $n \to \infty$ for consistency (cf. Maronna and Yohai (1995); Zuo et al. (2004)). This is because $\widehat{\beta}_{DW} - \overline{\beta} = O_p(n^{-1/2}) + O_p(n^{-1/2} J_T^{-1/2}) = O_p(n^{-1/2})$, and hence the depth-weighted estimator $\widehat{\beta}_{DW}$ is consistent to the depth-weighted mean $\overline{\beta}$ as long as $n \to \infty$, regardless of the size of J_T. Intuitively, this is because the DWMG estimator is basically a weighted average of $\widehat{\beta}_i$'s over $i = 1, \ldots, n$, in which the weighted average of the estimation error $\eta_i = \widehat{\beta}_i - \beta_i$ converges to zero in probability as long as $n \to \infty$.[9]

For the asymptotic variance V_F in Theorem 2, since $\widehat{\beta}_{DW} - \overline{\beta} = o_p(1)$, $\widehat{\beta}_i - \beta_i = o_p(1)$ for all i, and $\sup_{b \in \mathcal{B}} |\mathcal{D}(b, \widehat{F}_n) - \mathcal{D}(b, F)| = n^{-1/2} \sup_{b \in \mathcal{B}} |\widehat{H}_n(b)| = o_p(1)$ as $n, T \to \infty$, we can estimate it as[10]

$$\widehat{V}_F = \frac{1}{n} \sum_{i=1}^{n} \widehat{K}_F^0(\widehat{\beta}_i) \widehat{K}_F^0(\widehat{\beta}_i)', \tag{15}$$

where

$$\widehat{K}_F^0(\widehat{\beta}_i) = \frac{n^{-1} \sum_{j=1}^{n} \widehat{\beta}_j \dot{W}(\mathcal{D}(\widehat{\beta}_j, \widehat{F}_n)) \widehat{h}^0(\widehat{\beta}_j, \widehat{\beta}_i) + (\widehat{\beta}_i - \widehat{\beta}_{DW}) W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))}{n^{-1} \sum_{j=1}^{n} W(\mathcal{D}(\widehat{\beta}_j, \widehat{F}_n))} \tag{16}$$

and

$$\widehat{h}^0(\widehat{\beta}_j, \widehat{\beta}_i) = \widehat{h}(\widehat{\beta}_j, \widehat{\beta}_i) - \frac{1}{n} \sum_{\ell=1}^{n} \widehat{h}(\widehat{\beta}_j, \widehat{\beta}_\ell)$$

for some consistent estimator $\widehat{h}(\cdot, \cdot)$ of $h(\cdot, \cdot)$. See Section 4 for examples.

3.2 Homogeneous case

In Theorem 2, we obtain $\widehat{\beta}_{DW} - \overline{\beta} = O_p(n^{-1/2})$ for the heterogeneous case, which is the same rate of convergence as the mean-group estimator. When all β_i are the same and hence $\beta_i = \beta = \overline{\beta}$ for all i, however, we can improve this rate of convergence. In this case, we know that the standard within-group or the maximum likelihood estimator satisfies $\widehat{\beta}_{WG} - \beta = O_p((nT)^{-1/2})$ as $n, T \to \infty$ and we achieve faster rate of convergence, which is quite natural since we take average over i and t at the same time. This section studies how the depth-weighted estimator $\widehat{\beta}_{DW}$ behaves under this homogeneous parameter case. We show that the depth-weighted estimator achieves the $\sqrt{nJ_T}$-rate of convergence and it is

[9] For instance, if we consider the equally-weighted average, $n^{-1} \sum_{i=1}^{n} \widehat{\beta}_i = n^{-1} \sum_{i=1}^{n} \beta_i + n^{-1} \sum_{i=1}^{n} \eta_i$, we have $n^{-1} \sum_{i=1}^{n} \beta_i \to_p \overline{\beta}$ and $n^{-1} \sum_{i=1}^{n} \eta_i \to_p 0$ as $n \to \infty$ since $\mathbb{E}[\eta_i] = 0$.

[10] Even for the variance with n-asymptotics in (14), $\widehat{V}_F$ in (15) is a consistent estimator of $V_F + J_T^{-1} V_T$ as $n \to \infty$.

more robust than the conventional within-group or mean-group estimators.

For the homogeneous case, we let $\beta_i = \beta$ (a.s.) for all i instead of Assumption 1. We also define

$$\widehat{\xi}_i = J_T^{1/2}(\widehat{\beta}_i - \beta) \tag{17}$$

for each i. It is important to note that, since $\beta_i = \beta$ for all i in this case, we cannot define the depth of β_i. Therefore, the data depth of $\widehat{\beta}_i$ no longer estimates the depth of $\beta_i = \beta$. Instead, the heterogeneity of $\widehat{\beta}_i$ is now solely from the estimation error and the depth based on $\widehat{\beta}_i$ describes that of the scaled estimation error $\widehat{\xi}_i$ in (17). For this reason, the meaning of statistical depth of the homogeneous case is different from that of the heterogeneous case. For the heterogeneous β_i in the previous section, the data depth of $\widehat{\beta}_i$ consistently estimates the depth of β_i.[11]

To study the statistical properties of the DWMG estimator $\widehat{\beta}_{DW}$ for the homogeneous case, we suppose the following condition that replaces Assumption 2.

Assumption 4 *(i) $\{x_{it}^*, u_{it}^*\}$ are cross-sectionally independent. (ii) For each i, there exists ξ_i satisfying $\widehat{\xi}_i - \xi_i = o_p(1)$ as $T \to \infty$, where $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i \xi_i'] < \infty$.*

Under the regularity conditions and with $J_T = T$, Assumption 4-(ii) generally holds for any extremum estimators, where ξ_i is the k-dimensional multivariate normal random vector. We let $\mathcal{C} \subset \mathbb{R}^k$ be the support of ξ_i and $G(r) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathbb{P}\{\xi_i \leq r\}$ for any $r \in \mathcal{C}$, which is a well-defined distribution function. When ξ_i is identically distributed, G is simply the distribution function of ξ_i. We also define $\widehat{G}_n(r) = n^{-1} \sum_{i=1}^n 1\{\widehat{\xi}_i \leq r\}$ and

$$\widehat{\gamma}_n(\cdot) = \sqrt{n}(\widehat{G}_n(\cdot) - G(\cdot)) \quad \text{and} \quad \widehat{M}_n(\cdot) = \sqrt{n}(\mathcal{D}(\cdot; \widehat{G}_n) - \mathcal{D}(\cdot; G)).$$

We suppose the following conditions, where most of them are the same as Assumption 3 with F replaced by G. We let $\mathcal{C}_0 = \{r : \mathcal{D}(r, G) \geq d_0\}$ for some $d_0 \geq 0$.

Assumption 5 *We assume:*

(i) $W(\cdot)$ is continuously differentiable with bounded derivative $\dot{W}(\cdot)$; and $W(d) = 0$ for $d \in [0, d_0\kappa]$ with some $\kappa > 1$.

(ii) $\int W(\mathcal{D}(r, G))G(dr) > 0$, $\int ||r|| W(\mathcal{D}(r, G))G(dr) < \infty$, $\int ||r||^2 (W(\mathcal{D}(r, G)))^2 G(dr) < \infty$.

(iii) There exists $m(r, s)$ such that $\widehat{M}_n(r) = \int m(r, s)\widehat{\gamma}_n(ds) + o_p(1)$ uniformly on $\mathcal{C}_0$, satisfying

[11] More precisely, for $\widehat{\beta}_i = \beta_i + J_T^{-1/2}\widehat{\xi}_i$, the distribution of the $O_p(J_T^{-1/2})$ estimation error is dominated by that of β_i for the heterogeneous case.

$\int [\int ||r|| |\dot{W}(\mathcal{D}(r, G)) m(r, s)| G(dr)]^2 G(ds) < \infty$.

(iv) $\sup_{r \in \mathcal{C}} |\widehat{M}_n(r)| = O_p(1)$ *and* $\sup_{r \in \mathcal{C}_0} ||r|| |\widehat{M}_n(r)| = O_p(1)$.

The following theorem gives the asymptotic properties of $\widehat{\beta}_{DW}$ under the homogeneous parameter case.

Theorem 3 (Homogeneous Case) *We suppose Assumptions 4 and 5 hold, where* $\beta_i = \beta$ *for all* i*. Then,* $\widehat{\beta}_{DW} - \beta = O_p(n^{-1/2} J_T^{-1/2})$ *and* $\sqrt{nJ_T}(\widehat{\beta}_{DW} - \beta) \to_d \mathcal{N}(0, V_G)$ *as* $n, T \to \infty$*, where* $V_G = \int K_G^0(r) K_G^0(r)' G(dr)$ *with*

$$K_G^0(r) = \frac{\int s \dot{W}(\mathcal{D}(s, G)) m^0(s, r) G(ds) + r W(\mathcal{D}(r, G))}{\int W(\mathcal{D}(s, G)) G(ds)}$$

and $m^0(s, r) = m(s, r) - \int m(s, r) G(dr)$.

Unlike the heterogeneous parameter case in Theorem 2, Theorem 3 requires both $n, T \to \infty$ for consistency. Note that Theorem 3 is general enough to obtain limiting distributions of other types of average estimators in the form of $\sum_{i=1}^n \omega_i \widehat{\beta}_i$, including the mean-group (MG) estimator with $\omega_i = n^{-1}$ and the within-group (WG) estimator with $\omega_i = (\sum_{i=1}^n \sum_{t=1}^T x_{it}^* x_{it}^{*\prime})^{-1} \sum_{t=1}^T x_{it}^* x_{it}^{*\prime}$.

We let $\widetilde{\xi}_i = \sqrt{J_T}(\widehat{\beta}_i - \widehat{\beta}_{DW})$. Then it holds that $\widetilde{\xi}_i = \sqrt{J_T}(\widehat{\beta}_i - \beta) - \sqrt{J_T}(\widehat{\beta}_{DW} - \beta) = \widehat{\xi}_i + o_p(1)$ since $\widehat{\beta}_{DW} - \beta = O_p((nJ_T)^{-1/2})$. Hence, the asymptotic variance V_G can be estimated as

$$\widehat{V}_G = \frac{1}{n} \sum_{i=1}^n \widehat{K}_G^0(\widetilde{\xi}_i) \widehat{K}_G^0(\widetilde{\xi}_i)', \tag{18}$$

where

$$\widehat{K}_G^0(\widetilde{\xi}_i) = \frac{n^{-1} \sum_{j=1}^n \widetilde{\xi}_j \dot{W}(\mathcal{D}(\widetilde{\xi}_j, \widetilde{G}_n)) \widehat{m}^0(\widetilde{\xi}_j, \widetilde{\xi}_i) + \widetilde{\xi}_i W(\mathcal{D}(\widetilde{\xi}_i, \widetilde{G}_n))}{n^{-1} \sum_{j=1}^n W(\mathcal{D}(\widetilde{\xi}_j, \widetilde{G}_n))}, \tag{19}$$

and $\widehat{m}^0(\widetilde{\xi}_j, \widetilde{\xi}_i) = \widehat{m}(\widetilde{\xi}_j, \widetilde{\xi}_i) - n^{-1} \sum_{\ell=1}^n \widehat{m}(\widetilde{\xi}_j, \widetilde{\xi}_\ell)$. Here, $\widetilde{G}_n(r) = n^{-1} \sum_{i=1}^n 1\{\widetilde{\xi}_i \leq r\}$ and $\widehat{m}(\cdot, \cdot)$ is the sample analogue of $m(\cdot, \cdot)$ whose form is determined by the specific statistical depth function $\mathcal{D}(\cdot, \cdot)$. In practice, however, because of the affine-invariance property of the depth function, we have $K_G^0(\xi_i) = \sqrt{T} K_F^0(\beta_i)$ and we can simply use $\sqrt{T} \widehat{K}_F^0(\widehat{\beta}_i)$ from (16) instead of $\widehat{K}_G^0(\widetilde{\xi}_i)$.[12]

Since the depth-weighted estimator $\widehat{\beta}_{DW}$ for the homogeneous panel case depends on the depth of $\widehat{\xi}_i$ by construction, it is naturally robust against the outliers in $\widehat{\xi}_i$. From (6), for

[12] In ganeral, $\xi_i \sim \mathcal{N}(0, \Omega_i)$ for some $\Omega_i > 0$, and hence we can instead obtain $\widetilde{\xi}_i = \widehat{\Omega}_i^{1/2} z_i$ with z_i being a random draw from the k-dimensional multivariate standard normal distribution and for some consistent estimator $\widehat{\Omega}_i$.

instance, we have

$$\widehat{\xi}_i = \left(\frac{1}{T} \sum_{t=1}^{T} x_{it}^* x_{it}^{*\prime} \right)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{it}^* u_{it}^*. \tag{20}$$

In this case, we can understand that outlying behaviors of the ith agent mainly depend on the conditional distribution of the regression error u_{it} or u_{it}^*. Examples include the cases (i) when x_{it} has little time-variation for some i and hence the denominator in (20) is near singular; (ii) when x_{it} has minor measurement error or lagged dependent variables resulting in non-zero (or local-to-zero) $\mathbb{E}[x_{it}^* u_{it}^*]$ for some i and hence the limiting distribution of the numerator in (20) has non-zero mean;[13] (iii) when $var(x_{it}^* u_{it}^*)$ is very large for some i under heteroskedasticity and hence the limiting distribution of the numerator in (20) has large variance. In such cases, the standard MG and WG estimators may not be consistent. It is also possible that a poolability test rejects the null of homogeneous panel, not because the true parameters β_i's are indeed heterogeneous, but because $\widehat{\beta}_i$'s impose estimation errors that renders the test concludes an incorrect result.

The robustness of $\widehat{\beta}_{DW}$ can be formalized by observing the influence function. In the homogeneous case, we can obtain the influence function of the scaled estimation error:

$$\begin{aligned}
\sqrt{J_T}(\widehat{\beta}_{DW} - \beta) &= \frac{\sum_{i=1}^n \sqrt{J_T}(\widehat{\beta}_i - \beta) W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))}{\sum_{i=1}^n W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))} \\
&= \frac{\sum_{i=1}^n \widehat{\xi}_i W(\mathcal{D}(\widehat{\xi}_i, \widehat{G}_n))}{\sum_{i=1}^n W(\mathcal{D}(\widehat{\xi}_i, \widehat{G}_n))} = \frac{\int r W(\mathcal{D}(r, \widehat{G}_n)) \widehat{G}_n(dr)}{\int W(\mathcal{D}(r, \widehat{G}_n)) \widehat{G}_n(dr)} \equiv L(\widehat{G}_n),
\end{aligned}$$

where the second equality is because of the affine invariance property of the statistical depth function. We let δ_r be the point-mass distribution at r and $G(\varepsilon, \delta_{r_0}) = (1-\varepsilon)G + \varepsilon \delta_{r_0}$ be a version of G that is contaminated by an ε amount of an arbitrary point-mass distribution at r_0, where $0 \leq \varepsilon \leq 1$. Then, the influence function of L is obtained as

$$\begin{aligned}
\mathbb{IF}(r_0; L(G), G) &= \lim_{\varepsilon \to 0+} \varepsilon^{-1} \{ L(G(\varepsilon, \delta_{r_0})) - L(G) \} \\
&= \frac{\int r \dot{W}(\mathcal{D}(r, G)) \, m^0(r, r_0) G(dr) + r_0 W(\mathcal{D}(r_0, G))}{\int W(\mathcal{D}(r, G)) G(dr)}
\end{aligned} \tag{21}$$

as Zuo et al. (2004, Theorem 3.3), provided that the influence function of the depth $\mathcal{D}(r, G)$, $m^0(r, r_0)$ defined in Theorem 3, exists. Note that the influence function (21) corresponds to $K_G^0(r)$ in Theorem 3.

[13] The minor measurement error here means that x_{it} is contaminated only for a small number of individuals or over a small number of time periods. As we demonstrate in Section 5, even such minor measurement errors can distort the conventional MG and WG estimators, though the proposed depth-weighted estimator is robust toward such mild contamination.

The influence function measures the local robustness of $L(\cdot)$; a robust statistic has a uniformly bounded influence function (i.e., $\sup_r ||\mathbb{IF}(r; L(G), G)|| < \infty$). In this case, the uniform boundedness of $\mathbb{IF}(r_0; L(G), G)$ depends on that of $m^0(r, r_0)$, which limits the choice of the measures of the location and scale. The mean and the variance are not robust statistics because the breakdown point of mean is zero and its influence function is unbounded. For this reason, if the robustness is of the main concern, we use robust statistics for the location and scale, such as the median and the MAD, when we define the depth $\mathcal{D}$ (e.g., Zhou et al., 2004, Theorem 3.2).[14]

Remark (Mixture of heterogeneous and homogeneous parameters) We suppose that $\beta_i = (\beta'_{Ii}, \beta'_{IIi})' \in \mathbb{R}^{k_I + k_{II}}$, where $\beta_{Ii} \neq \beta_{Ij}$ for some $i \neq j$ but $\beta_{IIi} = \beta_{IIj}$ for all i, j. We also let $\overline{\beta} = (\overline{\beta}'_I, \overline{\beta}'_{II})'$ is the depth-weighted mean of β_i. In this case, Theorems 2 and 3 imply that

$$\sqrt{n}\left(\widehat{\beta}_{DW} - \overline{\beta}\right) = \sqrt{n}\begin{pmatrix} \widehat{\beta}_{I,DW} - \overline{\beta}_I \\ \widehat{\beta}_{II,DW} - \overline{\beta}_{II} \end{pmatrix} \to_d \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_I & 0 \\ 0 & 0 \end{pmatrix}\right) \tag{22}$$

as $n, T \to \infty$, where $V_I > 0$ is defined as in Theorem 2 based on the marginal distribution of β_{Ii}. This finding suggests that we may only need to consider the depth of the heterogeneous component β_{Ii} and their depth-weighted estimator $\widehat{\beta}_{I,DW}$ for inferences of $\overline{\beta}_I$. In fact, if we denote the joint distribution of $\beta_i = (\beta'_{Ii}, \beta'_{IIi})'$ as F and the marginals as (F_I, F_{II}), we can readily verify that the following two estimators are both consistent to $\overline{\beta}_I$:

$$\widehat{\beta}_{I,DW} = \frac{\sum_{i=1}^n \widehat{\beta}_{Ii} W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))}{\sum_{i=1}^n W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n))} \quad \text{and} \quad \widetilde{\beta}_{I,DW} = \frac{\sum_{i=1}^n \widetilde{\beta}_{Ii} W(\mathcal{D}(\widetilde{\beta}_{Ii}, \widetilde{F}_{In}))}{\sum_{i=1}^n W(\mathcal{D}(\widetilde{\beta}_{Ii}, \widetilde{F}_{In}))}, \tag{23}$$

where $\widehat{\beta}_i = (\widehat{\beta}'_{Ii}, \widehat{\beta}'_{IIi})'$ is an estimator using all the regressors and $\widetilde{\beta}_{Ii}$ is an estimator using only the relevant regressors. The first estimator $\widehat{\beta}_{I,DW}$ uses the depth from the joint empirical distribution $\widehat{F}_n$ of F (i.e., based on $\widehat{\beta}_i$), whereas the second estimator $\widetilde{\beta}_{I,DW}$ uses the depth from the marginal empirical distribution $\widetilde{F}_{In}$ of F_I (i.e., based on $\widetilde{\beta}_{Ii}$). An interesting example is when $\beta_{IIi} = 0$ for all i, which is the case that the corresponding regressors are irrelevant

[14] For further robustness, we could use a robust estimator for β_i (for each i) as well, instead of the least squares estimator. Examples of the robust estimators that have high breakdown points include the least median of squares regression estimator (Rousseeuw (1984)), which minimizes the median of squared residuals, and the least trimmed squares regression estimator (Rousseeuw (1985)), which minimizes the sum of trimmed smallest squared residuals. However, unlike Bramati and Croux (2007), we still conduct within transformation using the sample mean instead of the sample median for the two-way effect models. This is because the median is not a linear operator and $y_{it} - \mathsf{med}_{1 \leq s \leq T}\{y_{is}\}$ is not necessarily $(\tau_t - \mathsf{med}_{1 \leq s \leq T}\{\tau_s\}) + (x_{it} - \mathsf{med}_{1 \leq s \leq T}\{x_{is}\})'\beta + (u_{it} - \mathsf{med}_{1 \leq s \leq T}\{u_{is}\})$. Note that we use the within transformation instead of the first-difference so that the DWMG estimator is robust toward potential measurement errors (e.g., Griliches and Hausman (1986)).

and hence redundant. In this case, whether we use the joint depth or marginal depth, we have the same limit of the depth-weighted estimator.

4 Examples

The functions $h(\cdot,\cdot)$ in Assumption 3 and $m(\cdot,\cdot)$ in Assumption 5 are determined based on the choice of a specific statistical depth function $\mathcal{D}(\cdot,\cdot)$. The following examples give the forms of $h(\cdot,\cdot)$ and $m(\cdot,\cdot)$ for the Mahalanobis and the projection depths.

4.1 Mahalanobis depth

Heterogeneous Case: For the heterogeneous case in Section 3.1, we suppose the location and scatter functionals of F, $\mu(F)$ and $\Sigma(F)$, satisfy

$$\sqrt{n}(\mu(\widehat{F}_n) - \mu(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_1(b) + o_p(1), \tag{24}$$

$$\sqrt{n}(\Sigma(\widehat{F}_n) - \Sigma(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_2(b) + o_p(1) \tag{25}$$

for some $h_1(b)$ and $h_2(b)$ with $\int h_1(b)F(db) = \int h_2(b)F(db) = 0$. Then, it is straightforward to check that Assumption 3 holds for the Mahalanobis depth with $d_0 = 0$ and

$$h(c,b) = \frac{2(c-\mu(F))'\Sigma(F)^{-1}h_1(b) + (c-\mu(F))'\Sigma(F)^{-1}h_2(b)\Sigma(F)^{-1}(c-\mu(F))}{(1+(c-\mu(F))'\Sigma(F)^{-1}(c-\mu(F)))^2}, \tag{26}$$

provided $W(\cdot)$ is continuously differentiable on $[0,1]$ and $\int ||b||^2 F(db) < \infty$.

For instance, we consider the location and scale parameters $(\mu,\Sigma) = (mean, variance)$ that are affine invariant. We also suppose that the population distribution F is elliptically symmetric. We denote $\mu(F) = \mathbb{E}[\beta_i]$ and $\Sigma(F) = var(\beta_i)$ as the population mean and variance of β_i, which are assumed to exist.[15] Similarly as Example 2.3 of Zuo et al. (2004), we can verify that

$$h_1(b) = b - \mu(F), \tag{27}$$

$$h_2(b) = (b-\mu(F))(b-\mu(F))' - \Sigma(F). \tag{28}$$

[15] We can consider other robust measures of the location and scale as long as they are affine invariant; or simply use robust estimators of the mean and the variance based on trimmed or Winsorized samples. When the distribution is not symmetric, a trimmed sample could be useful to define the Mahalanobis depth. In such cases, $h_1(\cdot)$ and $h_2(\cdot)$ functions are to be modified accordingly.

To construct a consistent estimator of $h(\cdot,\cdot)$ for (16), we let $\mu(\widehat{F}_n) = \widehat{\mu} = n^{-1}\sum_{i=1}^n \widehat{\beta}_i$ and $\Sigma(\widehat{F}_n) = \widehat{\Sigma} = (n-k)^{-1}\sum_{i=1}^n (\widehat{\beta}_i - \widehat{\mu})(\widehat{\beta}_i - \widehat{\mu})'$ respectively as the sample mean and variance using observed $\widehat{\beta}_i$. Then the sample Mahalanobis outlyingness and the sample Mahalanobis depth of $\widehat{\beta}_i$ are obtained as

$$\mathcal{O}^m(\widehat{\beta}_i, \widehat{F}_n) = (\widehat{\beta}_i - \widehat{\mu})'\widehat{\Sigma}^{-1}(\widehat{\beta}_i - \widehat{\mu}) \quad \text{and} \quad \mathcal{D}^m(\widehat{\beta}_i, \widehat{F}_n) = \frac{1}{1 + \mathcal{O}^m(\widehat{\beta}_i, \widehat{F}_n)}.$$

Furthermore, for each $i, j = 1, \ldots, n$, we can obtain

$$\begin{aligned}\widehat{h}(\widehat{\beta}_j, \widehat{\beta}_i) &= \frac{2\widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_i^0 + \widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\left(\widehat{\beta}_i^0\widehat{\beta}_i^{0\prime} - \widehat{\Sigma}\right)\widehat{\Sigma}^{-1}\widehat{\beta}_j^0}{\left[1 + \widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_j^0\right]^2} \\ &= \frac{2\widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_i^0 + \left[\widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_i^0\right]^2 - \widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_j^0}{\left[1 + \widehat{\beta}_j^{0\prime}\widehat{\Sigma}^{-1}\widehat{\beta}_j^0\right]^2}\end{aligned}$$

from (26), where $\widehat{\beta}_i^0 = \widehat{\beta}_i - \widehat{\mu}$ for each i.

Homogeneous Case: For the homogeneous case in Section 3.2, we note that $\mu(G) = 0$ and we denote $\Sigma(G) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbb{E}[\xi_i\xi_i'] \equiv \Omega$. Hence, we can readily verify that Assumption 5 holds and we have

$$m(s, r) = \frac{2s'\Omega^{-1}r + s'\Omega^{-1}(rr' - \Omega)\Omega^{-1}s}{(1 + s'\Omega^{-1}s)^2}$$

as (26) since h_1 and h_2 terms are given as $h_1(r) = r$ and $h_2(r) = rr' - \Omega$. The estimator $\widehat{m}(s, r)$ can be obtained using some consistent estimator of Ω. For instance, for the individual time-series least squares estimator $\widehat{\beta}_i$ in (6), $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{\Omega}_i$ with $\widehat{\Omega}_i = (\sum_{t=1}^T x_{it}^* x_{it}^{*\prime}/T)^{-1}(\sum_{t=1}^T\sum_{s=1}^T x_{it}^*\widehat{u}_{it}^*\widehat{u}_{is}^* x_{is}^{*\prime}/T)(\sum_{t=1}^T x_{it}^* x_{it}^{*\prime}/T)^{-1}$ and $\widehat{u}_{it}^* = y_{it}^* - x_{it}^{*\prime}\widehat{\beta}_i$.

4.2 Projection depth

Heterogeneous Case: For the heterogeneous case in Section 3.1, we let v be a $k \times 1$ nonrandom vector with $||v|| = 1$, and $\mu(F; v)$ and $\sigma(F; v)$ be the location and scatter functionals of the distribution of $v'\beta_i$, respectively. We suppose

$$\sqrt{n}(\mu(\widehat{F}_n; v) - \mu(F; v)) \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^n h_1(b; v) + o_p(1), \tag{29}$$

$$\sqrt{n}(\sigma(\widehat{F}_n; v) - \sigma(F; v)) \quad = \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_2(b; v) + o_p(1) \tag{30}$$

hold uniformly in v for some $h_1(b; v)$ and $h_2(b; v)$, where they satisfy $\mathbb{E}[h_j(\beta_i; v)] = 0$, $\mathbb{E}[\sup_{||v||=1} h_j^2(\beta_i; v)] < \infty$, and $\mathbb{E}[\sup_{||v_1 - v_2|| \leq \delta, ||v_1|| = ||v_2|| = 1} |h_j(\beta_i; v_1) - h_j(\beta_i; v_2)|^2] \to 0$ as $\delta \to 0$ for $j = 1, 2$. Then, from Theorem 3.1 of Zuo et al. (2004), we can verify that Assumption 3 holds for the projection depth with $d_0 = 0$ and

$$h\,(c, b) = \frac{h_1(b; v^*(c)) + \mathcal{O}^p(c, F) h_2(b; v^*(c))}{\sigma(F; v^*(c))\,(1 + \mathcal{O}^p(c, F))^2}, \tag{31}$$

where $v^*(c)$ is such that $\mathcal{O}^p(c, F) = |v^*(c)'c - \mu(F; v^*(c))| / \sigma(F; v^*(c))$, provided $W(\cdot)$ is continuously differentiable on $[0, 1]$.

For instance, we consider the location and scale parameters $(\mu, \sigma) = (\mathsf{med}, \mathsf{MAD})$. For a given v, we let $\mu(F; v) = \mathsf{med}[v'\beta_i]$ and $\sigma(F; v) = \mathsf{MAD}[v'\beta_i]$ as the population median and median absolute deviation (MAD) of $v'\beta_i$, which are assumed to exist. From Lemma 3.2 of Zuo et al. (2004), we have

$$h_1(b; v) \quad = \quad \frac{\sqrt{v'\Sigma(F)v}}{p(0)} \left(\frac{1}{2} - 1\left\{ v'(b - \overline{\beta}) \leq 0 \right\} \right), \tag{32}$$

$$h_2(b; v) \quad = \quad \frac{\sqrt{v'\Sigma(F)v}}{2p(\varphi_0)} \left(\frac{1}{2} - 1\left\{ |v'(b - \overline{\beta})| \leq \varphi_0 \sqrt{v'\Sigma(F)v} \right\} \right). \tag{33}$$

Here, $\Sigma(F)$ is some positive definite matrix such that $(v'\Sigma(F)v)^{-1/2} v'(\beta_i - \overline{\beta})$ is a univariate symmetric variable (e.g., we let $\Sigma(F)$ be $var(\beta_i)$ if it exists); $p(\cdot)$ is the density function of $(v'\Sigma(F)v)^{-1/2} v'(\beta_i - \overline{\beta})$ and φ_0 is its MAD, satisfying $p(0) > 0$ and $p(\varphi_0) > 0$.

To construct a consistent estimator of $h(\cdot, \cdot)$ for (16), we let $\mu(\widehat{F}_n; v) = \mathsf{med}_{1 \leq i \leq n}\{v'\widehat{\beta}_i\}$ and $\sigma(\widehat{F}_n; v) = \mathsf{MAD}_{1 \leq i \leq n}\{v'\widehat{\beta}_i\} = \mathsf{med}_{1 \leq i \leq n}\{|v'\widehat{\beta}_i - \mathsf{med}_{1 \leq j \leq n}\{v'\widehat{\beta}_j\}|\}$ be the sample median and MAD using observed $v'\widehat{\beta}_i$. For each $i, j = 1, \ldots, n$, we obtain the sample projection depth of $\widehat{\beta}_i$ and $\widehat{h}(\widehat{\beta}_j, \widehat{\beta}_i)$ as follows:

1. We generate v_r from a k-dimensional multivariate standard normal for $r = 1, \ldots, R$. For each r, redefine v_r as $v_r / \sqrt{v_r' v_r}$ so that $\|v_r\| = 1$. Recall that since the standard normal density function is rotationally symmetric, standard-normal-distributed random coordinates yields a uniform distribution of directions and hence it generates random points on the surface of the unit circle.

2. For each $r = 1, \ldots, R$, we let $\mathsf{med}_{1 \leq \ell \leq n}\{v_r'\widehat{\beta}_\ell\}$ be the sample median of $v_r'\widehat{\beta}_\ell$ over $\ell = 1, \ldots, n$; and $\mathsf{MAD}_{1 \leq \ell \leq n}\{v_r'\widehat{\beta}_\ell\} = \mathsf{med}_{1 \leq \ell \leq n}\{|v_r'\widehat{\beta}_\ell - \mathsf{med}_{1 \leq \ell' \leq n}\{v_r'\widehat{\beta}_{\ell'}\}|\}$ be the sample MAD of $v_r'\widehat{\beta}_\ell$.

3. For each $j = 1, \ldots, n$, we find $v^*(j)$ such that

$$v^*(j) = \arg\max_{v_r:1\leq r\leq R} \frac{\left|v_r'\widehat{\beta}_j - \mathsf{med}_{1\leq \ell\leq n}\{v_r'\widehat{\beta}_\ell\}\right|}{\mathsf{MAD}_{1\leq \ell\leq n}\{v_r'\widehat{\beta}_\ell\}}.$$

Then, the sample outlyingness and the depth of $\widehat{\beta}_j$ are defined as

$$\mathcal{O}^p(\widehat{\beta}_j, \widehat{F}_n) = \frac{\left|v^*(j)'\widehat{\beta}_j - \mathsf{med}_{1\leq \ell\leq n}\{v^*(j)'\widehat{\beta}_\ell\}\right|}{\mathsf{MAD}_{1\leq \ell\leq n}\{v^*(j)'\widehat{\beta}_\ell\}} \text{ and } \mathcal{D}^p(\widehat{\beta}_j, \widehat{F}_n) = \frac{1}{1+\mathcal{O}^p(\widehat{\beta}_j, \widehat{F}_n)}.$$

4. For each $j, \ell = 1, \ldots, n$, we let

$$\zeta_\ell(j) = \left[v^*(j)'\widehat{\Sigma}v^*(j)\right]^{-1/2} v^*(j)'\left(\widehat{\beta}_\ell - \widehat{\beta}_{DW}\right),$$

where $\widehat{\Sigma} = (n-k)^{-1}\sum_{i=1}^n(\widehat{\beta}_i - n^{-1}\sum_{j=1}^n \widehat{\beta}_j)(\widehat{\beta}_i - n^{-1}\sum_{j=1}^n \widehat{\beta}_j)'$ is the sample variance using observed $\widehat{\beta}_i$, and $\widehat{\beta}_{DW}$ is defined using the projection depth. Furthermore, we define

$$\begin{aligned} \widehat{\varphi}_0(j) &= \mathsf{MAD}_{1\leq \ell\leq n}\{\zeta_\ell(j)\} \\ \widehat{p}(j;z) &= \frac{1}{n\lambda_j}\sum_{\ell=1}^n \Upsilon\left(\frac{\zeta_\ell(j)-z}{\lambda_j}\right) \end{aligned} \tag{34}$$

for some kernel function $\Upsilon(\cdot)$ and a bandwidth λ_j that satisfies the conventional conditions for the consistent kernel density estimator.

5. Then, for each $i, j = 1, \ldots, n$, we can obtain

$$\widehat{h}(\widehat{\beta}_j, \widehat{\beta}_i) = \frac{\widehat{h}_1(\widehat{\beta}_i; v^*(j)) + \mathcal{O}^p(\widehat{\beta}_j, \widehat{F}_n)\widehat{h}_2(\widehat{\beta}_i; v^*(j))}{\mathsf{MAD}_{1\leq \ell\leq n}\{v^*(j)'\widehat{\beta}_\ell\}\left(1+\mathcal{O}^p(\widehat{\beta}_j, \widehat{F}_n)\right)^2} \tag{35}$$

from (31), where

$$\begin{aligned} h_1(\widehat{\beta}_i; v^*(j)) &= \frac{\sqrt{v^*(j)'\widehat{\Sigma}v^*(j)}}{\widehat{p}(j;0)}\left(\frac{1}{2} - 1\left\{v^*(j)'(\widehat{\beta}_i - \widehat{\beta}_{DW}) \leq 0\right\}\right) \\ h_2(\widehat{\beta}_i; v^*(j)) &= \frac{\sqrt{v^*(j)'\widehat{\Sigma}v^*(j)}}{2\widehat{p}(j;\widehat{\varphi}_0(j))}\left(\frac{1}{2} - 1\left\{|\zeta_i(j)| \leq \widehat{\varphi}_0(j)\right\}\right). \end{aligned}$$

Homogeneous Case: For the homogeneous case in Section 3.2, we can readily verify that Assumption 5 holds. In this case, we have

$$m(s,r) = \frac{h_1(r; v^*(s)) + \mathcal{O}^p(s,G) h_2(r; v^*(s))}{\sigma(G; v^*(s)) \left(1 + \mathcal{O}^p(s,G)\right)^2}$$

as (31), where $v^*(s)$ is such that $\mathcal{O}^p(s,G) = |v^*(s)'s - \mu(G; v^*(s))| / \sigma(G; v^*(s))$, and h_1 and h_2 terms are now given as

$$\begin{aligned} h_1(r;v) &= \frac{\sqrt{v'\Omega v}}{p(0)} \left(\frac{1}{2} - 1\left\{ v'r \leq 0 \right\} \right) \\ h_2(r;v) &= \frac{\sqrt{v'\Omega v}}{2p(\varphi_0)} \left(\frac{1}{2} - 1\left\{ |v'r| \leq \varphi_0 \sqrt{v'\Omega v} \right\} \right) \end{aligned}$$

with the density function $p(\cdot)$ of $(v'\Omega v)^{-1/2} v' z_i$ and φ_0 is its MAD, satisfying $p(0) > 0$ and $p(\varphi_0) > 0$. The estimator $\widehat{m}(s,r)$ can be obtained using a consistent estimator $\widehat{\Omega}$ and the density estimator $\widehat{p}(\cdot)$ as in (34).

5 Simulations

In this section, we examine finite sample performance of depth-weighted estimators. We consider a two-way effect panel regression model

$$y_{it} = x_{it}'\beta_i + \alpha_i + \tau_t + u_{it}$$

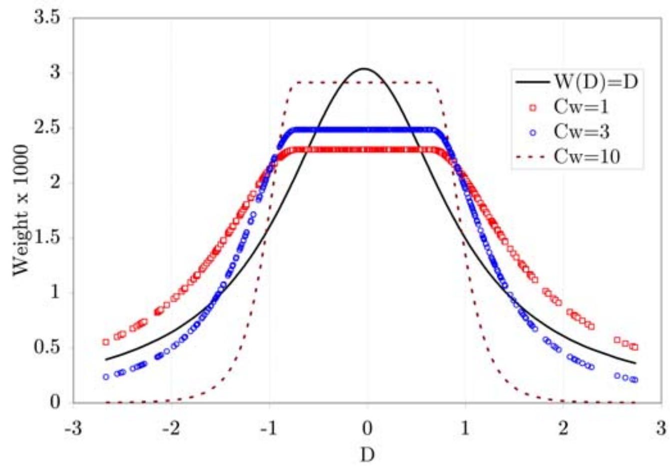
in (5) with $x_{it} \in \mathbb{R}^2$. The fixed effects are generated as $\alpha_i = \alpha_i^0 + T^{-1} \sum_{t=1}^T (x_{1,it} + x_{2,it})$ with $\alpha_i^0 \sim iid\mathcal{U}[0,1]$ and $\tau_t = \tau_t^0 + n^{-1} \sum_{i=1}^n (x_{1,it} + x_{2,it})$ with $\tau_t^0 \sim iid\mathcal{U}[0,1]$. The regressors $x_{it} = (x_{1,it}, x_{2,it})'$ and the error term u_{it} are uncorrelated and respectively generated as

$$x_{it} \sim iid\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix} \right) \quad \text{and} \quad u_{it} \sim iid\mathcal{N}\left(0, \sigma_i^2\right),$$

where we consider both the homoskedastic error with $\sigma_i^2 = 1$ and the heteroskedastic error with $\sigma_i^2 \sim iid\mathcal{X}_1^2$. We examine the following six data generating processes:

- DGP1 (Homogeneous panel without data contamination): $\beta_i = (1,1)'$ for all i.
- DGP2 (Homogeneous panel with vertical outliers for some i): $\beta_i = (1,1)'$ for all i and 5% cross-sectional error terms are outlying as $u_{it} + u^0$ for $i = 1, \ldots, \lfloor 0.05n \rfloor$, where $u^0 \sim iid\mathcal{N}\left(10, 5^2\right)$ is independent of u_{it} and $\lfloor \cdot \rfloor$ stands for a nearest integer.

Figure 4: Weight function



- DGP3 (Homogeneous panel with endogeneity for some i): $\beta_i = (1,1)'$ for all i and 5% cross-sectional observations of $x_{1,it}$ are correlated with u_{it} by letting $x_{1,it} = 10u_{it}$ for $i = 1, \ldots, \lfloor 0.05n \rfloor$.
- DGP4 (Some heterogeneous individuals): $\beta_i = (\beta_{1,i}, 1)'$ for all i, where $\beta_{1,i} \sim iid\mathcal{N}(1,1)$ for $i = 1, \ldots, \lfloor 0.05n \rfloor$ but $\beta_{1,i} = 1$ for the rest of i.
- DGP5 (One heterogeneous parameter): $\beta_i = (\beta_{1,i}, 1)'$ for all i, where $\beta_{1,i} \sim iid\mathcal{N}(1,1)$ for all i.
- DGP6 (Heterogeneous panel):[16] For all i, $\beta_i = (\beta_{1,i}, \beta_{2,i})'$ is bivariate normal with mean $(1,1)'$ and variance $I_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

DGP1 is the ideal case under which the within-group estimator (WG) should perform the best. DGP2 and DGP3 describe the situation that a few individuals (5% of the cross-section in this experiment) have data contamination problem such as measurement errors in y_{it} (DGP2) or x_{it} (DGP3). DGP4 describes the situation that a few individuals (5% of the cross-section in this experiment) have heterogeneous slope parameters, whereas DGP5 and DGP6 consider the case that all the individuals have heterogeneous slope parameters.

For all the cases, we consider the WG, the mean-group estimator (MG), the depth-weighted estimators based on the projection depth (DWp) and the Mahalanobis depth

[16] In this simulation, we only report the case that β_i is normally distributed and hence its moments exist. When β_i is Cauchy, which is still symmetric but does not have any moments, the performance of the depth-weighted estimators are noticeably superior. For instance, the mean absolute error of the depth-weighted estimators are 25% of those of the WG and MG estimators for any T.

Table 1: Relative Efficiency Compared with WG

n	T	Relative Efficiency					RMSE
		MG	DWp(0)	DWp(3)	DWm(0)	DWm(3)	WG
100	5	0.514	0.717	0.720	0.706	0.718	0.0723
100	10	0.819	0.842	0.852	0.847	0.841	0.0478
100	25	0.949	0.913	0.933	0.926	0.917	0.0294
100	50	0.982	0.921	0.944	0.936	0.922	0.0203
200	5	0.493	0.708	0.713	0.698	0.706	0.0508
200	10	0.835	0.848	0.860	0.857	0.846	0.0334
200	25	0.950	0.908	0.930	0.918	0.909	0.0205
200	50	0.979	0.931	0.951	0.940	0.924	0.0144

Note: The relative efficiency is defined as RMSE(WG)/RMSE(estimator). The empirical RMSE of WG is given in the last column of the table.

(DWm). For the depth-weighted estimators, we use the most typical weight function used in this literature (e.g., Zuo et al. (2004)), which is given by

$$W(d) = \begin{cases} \dfrac{\exp(-c_W \left(1 - d/\overline{\mathcal{D}}\right)^2) - \exp(-c_W)}{1 - \exp(-c_W)} & \text{if } d < \overline{\mathcal{D}}, \\ 1 & \text{if } d \geq \overline{\mathcal{D}}. \end{cases} \quad (36)$$

for some positive constant $c_W > 0$, where $\overline{\mathcal{D}}$ is the median of depths. When $c_W = 0$, we let $W(d) = d$. The constant c_W determines how heavily the weight function penalizes as d gets away from $\overline{\mathcal{D}}$. We consider two cases with $c_W = 0$ and 3, and give these values in the parenthesis like DWp(0) or DWp(3). Figure 4 depicts the weight function (36), normalized as $W(\mathcal{D}) / \int W(r)\, dr$, and shows how the weight function changes with c_W. The simulation results are based on 2000 iterations for different combinations of sample sizes $n = (100, 200)$ and $T = (5, 10)$. Simulation results are summarized in Tables 1 to 3.

First, Table 1 reports relative efficiency of all estimators, which is obtained by dividing the root mean square errors (RMSE) of WG by those of other estimators. For efficiency comparison, we consider DGP1 with the homoskedastic error u_{it} with $\sigma_i^2 = 1$ for all i, under which we expect WG should yield the smallest RMSE by the Gauss-Markov theorem. So, value 1 of the relative efficiency means the estimator is as efficient as WG; as the value gets smaller, it becomes less efficient than WG. The result shows that the depth-weighted estimators lose some efficiency compared to WG, but the efficiency improves as T increases. They outperform the equally weighted average estimator, MG, when T is small.

Table 2: Root Mean Square Error

	n	T	WG	MG	DWp(0)	DWp(3)	DWm(0)	DWm(3)
DGP1	100	5	0.0710	0.1404	0.0525	0.0586	0.0662	0.0657
	100	10	0.0482	0.0587	0.0297	0.0329	0.0345	0.0348
	200	5	0.0504	0.1044	0.0360	0.0410	0.0469	0.0466
	200	10	0.0337	0.0404	0.0195	0.0223	0.0234	0.0238
DGP2	100	5	0.0787	0.1537	0.0562	0.0626	0.0713	0.0705
	100	10	0.0540	0.0661	0.0320	0.0354	0.0376	0.0377
	200	5	0.0563	0.1170	0.0387	0.0439	0.0510	0.0505
	200	10	0.0378	0.0455	0.0209	0.0238	0.0254	0.0255
DGP3	100	5	0.0934	0.1379	0.0499	0.0546	0.0624	0.0617
	100	10	0.0867	0.0568	0.0310	0.0328	0.0343	0.0344
	200	5	0.0884	0.0977	0.0356	0.0390	0.0443	0.0439
	200	10	0.0853	0.0395	0.0217	0.0229	0.0240	0.0239
DGP4	100	5	0.0735	0.1407	0.0554	0.0615	0.0682	0.0681
	100	10	0.0496	0.0587	0.0339	0.0375	0.0377	0.0387
	200	5	0.0522	0.1044	0.0382	0.0433	0.0483	0.0485
	200	10	0.0350	0.0404	0.0228	0.0258	0.0259	0.0269
DGP5	100	5	0.1121	0.1426	0.0886	0.0852	0.0918	0.0922
	100	10	0.0736	0.0590	0.0593	0.0545	0.0558	0.0606
	200	5	0.0806	0.1049	0.0616	0.0589	0.0639	0.0642
	200	10	0.0517	0.0405	0.0422	0.0386	0.0395	0.0433
DGP6	100	5	0.1431	0.1439	0.1146	0.1077	0.1125	0.1118
	100	10	0.0938	0.0595	0.0760	0.0685	0.0719	0.0770
	200	5	0.1025	0.1056	0.0796	0.0750	0.0785	0.0782
	200	10	0.0663	0.0406	0.0529	0.0477	0.0506	0.0548

Table 2 fully reports the RMSE of all the estimators in DGP1 to DGP6[17] when u_{it} is heteroskedastic with $\sigma_i^2 \sim iid\mathcal{X}_1^2$. The homoskedastic case gives similar results except for DGP1 as described in Table 1, hence omitted. Since we consider heteroskedastic case, WG does not perform the best even for DGP1. The depth-weighted estimators outperform in reducing RMSE for all the cases. The performance gets improved as the sample size increases, noticeably with T. The depth-weighted estimator is designed robust to any types of heterogeneity including heteroskedastic errors.

Finally, Table 3 compares rejection probabilities of the Wald test for $\overline{\beta} = (1,1)'$, where we use the $\mathcal{X}_2^2$ critical values based on the asymptotic normality in Theorems 2 and 3. For

[17] For the heterogeneous cases of DGP4 to DGP6, we calculate the empirical MSE as $tr\{R^{-1}\sum_{r=1}^{R}(\widehat{\beta} - \overline{\beta})(\widehat{\beta} - \overline{\beta})'\}$ for each estimator $\widehat{\beta}$.

Table 3: Rejection Probabilty of Wald Test

	n	T	WG	MG	DWp(0)	DWp(3)	DWm(0)	DWm(3)
DGP1	100	5	0.097	0.058	0.107	0.059	0.029	0.037
	100	10	0.076	0.051	0.132	0.081	0.039	0.046
	200	5	0.096	0.042	0.114	0.065	0.034	0.050
	200	10	0.069	0.046	0.115	0.075	0.039	0.043
DGP2	100	5	0.089	0.051	0.043	0.022	0.026	0.032
	100	10	0.075	0.057	0.037	0.023	0.035	0.041
	200	5	0.094	0.046	0.036	0.027	0.033	0.046
	200	10	0.068	0.043	0.035	0.022	0.033	0.041
DGP3	100	5	0.909	0.045	0.179	0.096	0.035	0.050
	100	10	0.965	0.055	0.237	0.139	0.063	0.068
	200	5	0.990	0.043	0.197	0.115	0.042	0.056
	200	10	1.000	0.052	0.236	0.134	0.071	0.070
DGP4	100	5	0.114	0.058	0.086	0.053	0.026	0.037
	100	10	0.107	0.054	0.059	0.038	0.043	0.049
	200	5	0.105	0.043	0.103	0.060	0.035	0.049
	200	10	0.103	0.046	0.055	0.038	0.039	0.046
DGP5	100	5	0.209	0.053	0.175	0.101	0.053	0.051
	100	10	0.253	0.050	0.157	0.093	0.049	0.042
	200	5	0.185	0.042	0.156	0.087	0.049	0.044
	200	10	0.234	0.046	0.171	0.087	0.039	0.039
DGP6	100	5	0.238	0.056	0.198	0.112	0.061	0.057
	100	10	0.364	0.058	0.217	0.117	0.059	0.057
	200	5	0.232	0.048	0.193	0.106	0.050	0.055
	200	10	0.339	0.046	0.205	0.113	0.047	0.046

Note: The null hypothesis is $\overline{\beta} = (1,1)'$; test is based on $\mathcal{X}_2^2$ critical values. The nominal size is 5%.

all the cases, we use the heteroskedasticity-robust standard error estimator for WG and the sample cross-sectional covariance estimator for MG. For the depth-weighted estimators, we use the variance estimator in (18) as the null hypothesis implies a homogeneous panel model. In general, we can see that the depth-weighted estimators have the rejection probabilities close to the nominal size 5%, even when the size of WG is barely controlled like in DGP3. One interesting finding is that MG controls size reasonably well for all the cases, which seem quite similar to that of DWm. Understanding that MG is also an average estimator like the depth-weighted estimators, an average estimator in general seems to well control the rejection probability.

Unlike the Mahalanobis depth case, the projection-depth-weighted estimator DWp shows over-rejection tendency that barely improves even with large n. It seems mainly because of under-estimation of the variance of DWp that relies on kernel estimation. Some adjustment of bandwidth could reduce the over-rejection probabilities of DWp. Furthermore, from the results in Cui and Tian (1994), we can see that the terms $\widehat{h}_1(\widehat{\beta}_i; v^*(j))$ and $\widehat{h}_2(\widehat{\beta}_i; v^*(j))$ in (35) has very slow rate of convergence as $n^{-1/4}\ln n$, because they are functions of the sample median and sample MAD. It thus results in a slow convergence rate of the variance estimator of the DWp estimator. In comparison, the same terms of the DWm estimator are functions of the sample mean and sample variance, which have the standard $n^{-1/2}$ rate of convergence.

6 Empirical Illustration: Purchasing Power Parity

Though the purchasing power parity (PPP) has been one of the most studied topics in international economics, the empirical results for the PPP is still mixed. One potential reason can be severe heterogeneity among the currencies and the conventional pooled estimator may yield misleading results. As an illustration, we re-examine the relative PPP in the idiosyncratic level and show how the heterogeneity could affect the PPP model estimation results. We also demonstrate how the depth-weighted estimators handle the outlying currencies and yield different PPP estimates.

To this end, we consider the following factor augmented regression:

$$\Delta s_{it}^{\mathrm{us}} = a_i + \beta_i(\pi_{it} - \pi_t^{\mathrm{us}}) + \lambda_i' f_t + e_{it}, \tag{37}$$

where $\Delta s_{it}^{\mathrm{us}}$ is the depreciation rate of the ith currency against the USD at time t, π_{it} is the monthly inflation rate in the country i, and π_t^{us} is the monthly inflation rate in the U.S. $f_t = (f_{y,t}', f_{x,t}')'$, where $f_{y,t}$ is the vector of the common factors to $\Delta s_{it}^{\mathrm{us}}$ and $f_{x,t}$ is the vector of the common factors to π_{it}. The number of the common factors is selected based on Bai and Ng (2002)'s IC_2 criterion with a maximum number of 8. The selected numbers are 2 for both $f_{y,t}$ and $f_{x,t}$. We include all these chosen common factors in the regression, so that the cross sectional dependence among the nominal and relative inflation rates are fully controlled. In this equation, $\beta_i = 1$ if the relative PPP holds for the ith currency in the short run; $\beta_i = 0$ if there is no relationship between the depreciation rate and the relative inflation rate. We use a monthly panel data set of 27 bilateral spot exchange rates and CPI from 1999.M1 to 2015.M6. The source of the data is Global Insight (GI) at Information Handling Service (IHS). The list of the currencies are in the note of Table 4.

Table 4 reports the factor augmented least squares estimate for each currency, its standard error, the two t-statistics for the null of $\beta_i = 0$ and $\beta_i = 1$, and the calculated weights when obtaining the pooled estimates in Table 5 for three estimators: the pooled least square estimator (WG), the Mahalanobis depth weighted estimator (DWm), and the projection

Table 4: Relative Purchaing Power Parity Estimation of Individual Countries

	individual PPP				weight in %		
	$\widehat{\beta}_i$	s.e.	t-stat $\beta_i = 0$	t-stat $\beta_i = 1$	WG	DWm	DWp
AUS	-1.40	1.08	-1.30	-2.22*	0.68	1.25	1.36
NZ	-1.10	1.64	-0.67	-1.28	0.58	1.66	1.62
HUN	-0.71	0.58	-1.22	-2.95*	2.87	2.46	2.16
SIN	-0.34	0.27	-1.26	-4.96*	2.97	3.57	3.13
GBR	-0.27	0.72	-0.37	-1.76	0.70	3.83	3.44
JPN	-0.26	0.49	-0.54	-2.57*	1.41	3.85	3.48
TWN	-0.22	0.13	-1.71	-9.38*	8.83	4.00	3.70
SWE	-0.19	0.63	-0.30	-1.89*	1.02	4.10	3.86
KOR	-0.14	0.61	-0.23	-1.87*	1.58	4.27	4.17
THA	-0.07	0.29	-0.25	-3.69*	3.32	4.48	4.68
ROM	-0.05	0.36	-0.14	-2.92*	9.99	4.55	4.87
MEX	-0.03	0.60	-0.04	-1.72	1.28	4.62	5.12
PHI	0.04	0.35	0.11	-2.74*	2.67	4.79	5.87
EURO	0.13	0.94	0.13	-0.93	0.34	4.98	7.39
IND	0.17	0.21	0.80	-3.95*	7.49	5.04	6.65
COL	0.17	1.00	0.17	-0.83	1.35	5.05	6.59
CZE	0.18	0.78	0.23	-1.05	1.98	5.06	6.40
SUI	0.29	0.88	0.33	-0.81	0.64	5.13	5.02
POL	0.39	0.95	0.41	-0.64	1.51	5.08	4.20
NOR	0.40	0.58	0.69	-1.03	2.71	5.07	4.13
BRA	0.62	1.34	0.47	-0.28	3.54	4.58	3.02
RSA	0.92	0.88	1.04	-0.09	2.92	3.59	2.24
CAN	1.13	0.64	1.77	0.20	1.36	2.90	1.88
TUR	1.17	0.57	2.08*	0.30	27.38	2.78	1.83
ICE	1.58	0.62	2.54*	0.94	5.39	1.82	1.41
ISR	2.25	0.47	4.82*	2.66*	2.84	0.99	1.03
CHI	3.19	0.68	4.72*	3.22*	2.65	0.51	0.75

Note: * denotes significant at 5% from each t-test. The countries in the table are ordered by the individual relative PPP estimates. The currencies are (in alphabetical order) of Australia (AUS), Brazil (BRA), Canada (CAN), Chile (CHI), Columbia (COL), the Czech Republic (CZE), the Euro (EUR), Hungary (HUN), Iceland (ICE), India (IND), Israel (ISR), Japan (JPN), Korea (KOR), Mexico (MEX), Norway (NOR), New Zealand (NZL), the Philippines (PHI), Poland (POL), Romania (ROM), Singapore (SIN), South Africa (RSA), Sweden (SWE), Switzerland (SUI), Taiwan (TWN), Thailand (THA), Turkey (TUR), and the U.K. (GBR), relative to the U.S. Dollar (USA).

Table 5: Relative Purchaing Power Parity Estimation

	full sample		without TUR	
	$\widehat{\beta}$	p-value	$\widehat{\beta}$	p-value
WG	0.574	0.005	0.348	0.036
MG	0.291	0.057	0.257	0.082
DWm	0.168	0.119	0.130	0.165
DWp	0.135	0.472	0.097	0.401

Note: The p-values is for the null of $\beta = 0$, which implies the relative PPP does not hold. For the sample excluding TUR, the common factors are obtained from the full sample.

depth weighted estimator (DWp). The Mean group estimator (MG) simply uses the equal weights. We set $c_W = 0$ for the depth-weighted estimators. The standard error is calculated by Newey-West HAC estimator with the lag length selection of $\lfloor T^{1/3} \rfloor$. The overall results do not change much with other lag choices. At the individual level, the null of $\beta_i = 1$ (which implies the relative PPP holds) is very often rejected: 13 out of 27 cases are rejected at the 5% level. Meanwhile, the null of $\beta_i = 0$ (which implies the relative PPP does not hold) is hardly rejected: only 4 out of 27 cases are rejected at the 5% level. Also note that the sign of $\widehat{\beta}_i$ is negative in 12 out of 27 cases and positive in 15 cases.

Table 5 reports the pooled PPP estimates. For the full sample of 27 currencies, WG is 0.574 and significantly different from zero. This finding is rather puzzling since the null of $\beta_i = 0$ was hardly rejected at the individual level in Table 4. We can see that the weight used for WG in Turkey (TUR) is extremely high (27%) and hence its large $\widehat{\beta}_i$ value (1.17) results in such an unexpected WG. In comparison, weights based on depths are very different from those used for WG and yield very different results. The weights for Turkey are even less than the equal weight percentage 3.7%. The highest Mahalanobis depth is found for Switzerland (SUI), whose weight is 5.13% and $\widehat{\beta}_i$ is 0.29; the highest projection depth is found for Euro (EURO), whose weight is 7.39% and $\widehat{\beta}_i$ is 0.13. Contrast to the conventional estimates WG and MG, the depth weighted estimators DWm and DWp are not significantly different from zero even at the 10% level. For the sample excluding TUR, Table 5 shows that WG drops from 0.57 to 0.35, though it is still significantly different from zero. However, the depth-weighted estimators change little. Based on these findings, we conclude that the mixed evidence on the relative PPP by the typical WG estimation would be because of over-weighting on a few outliers. Depth-weighted estimators automatically impose very low weights on such outlying currencies, and conclude that there is weak evidence for the relative PPP.

7 Concluding Remarks

As a way of estimating a robust central tendency of potentially heterogeneous parameters in a panel regression, we propose an averaging idea based on the depth of each parameter estimates, the depth-weighted mean-group estimator. Whether the underlying model is indeed heterogeneous or not, this new idea produces a consistent estimator that is asymptotically normal. It relies on the individual-specific time series estimators, and it can be applied for any nonlinear panel data models as long as the individual-specific estimators are consistent (cf. Boneva and Linton (2016)). The new estimator shows promising finite sample performance in that it is quite robust to any types of outlying behavior of heterogeneous agents in panel data.

The idea of averaging estimators can be found under different contexts, such as Sawa (1973), Lee and Zhou (2015), and Chen et al. (2016), which mainly focus on the efficiency gain by allocating optimal weights. In comparison, we focus more on the robustness issue in this paper and we do not consider optimality. However, as we partly show in the simulations, both the robustness and the efficiency can improve by properly choose the weight function $W(\cdot)$, such as the value c_W in (36).

Finally, the weight function $W(\cdot)$ considered in this paper is smooth and it does not completely eliminate the outliers. Similar to Zuo (2006), we can further robustify $\widehat{\beta}_{DW}$ by considering a weight that completely trims out the outliers, which yields the ϖ-trimmed depth-weighted mean estimator:

$$\widehat{\beta}_{TDW}^{\varpi} = \frac{\sum_{i=1}^{n} \widehat{\beta}_i W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n)) 1\{\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n) \geq \varpi\}}{\sum_{i=1}^{n} W(\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n)) 1\{\mathcal{D}(\widehat{\beta}_i, \widehat{F}_n) \geq \varpi\}}$$

for some $\varpi \in (0, \varpi^*)$, where $\varpi^* = \max_{1 \leq i \leq n} \mathcal{D}(\widehat{\beta}_i, \widehat{F}_n)$. When $\varpi = 0$, it is simply $\widehat{\beta}_{DW}$; when $W(\cdot) = 1$, it is the ϖ-trimmed mean where the trim is based on the data depth. In a similar vein, Lee and Sul (2020) considers trimmed mean group estimator.

A Appendix

A.1 Bias in heterogeneous panel regression estimators

We suppose a regular function $Q(\cdot)$ that is Hadamard differentiable at F with a bounded derivative. We consider a parameter $Q(F)$ and its estimator $Q(\widehat{F}_n)$. Then, by the functional delta method, the second result in Theorem 1 implies that $\sqrt{n}(Q(\widehat{F}_n) - Q(F))$ is asymptotically normal but it has nonzero mean unless $\sqrt{n}/J_T \to 0$.

For instance, we consider $\theta = \int q(b)F(b) = \mathbb{E}[q(\beta_i)]$ for some smooth $q(\cdot)$ and its estimator $\widehat{\theta} = \int q(b)\widehat{F}_n(b) = n^{-1}\sum_{i=1}^n q(\widehat{\beta}_i)$. The following corollary summarizes the limiting distribution of this averaging estimator $\widehat{\theta}$. The proof is in Appendix A.2.

Corollary A *We suppose $q(\cdot)$ is second-order differentiable with finite derivatives and satisfies $\int q(b)q(b)'F(b) < \infty$, $\int \ddot{q}(b)F(b) < \infty$, and $\int \ddot{q}(b)F(b) \neq 0$, where $\ddot{q}(\cdot)$ is the Hessian matrix of $q(\cdot)$. Then, under the same conditions in Theorem 1, we have $\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathcal{N}(\phi\Psi, \mathbb{E}[(q(\beta_i) - \mathbb{E}[q(\beta_i)])(q(\beta_i) - \mathbb{E}[q(\beta_i)])'])$ as $n, T \to \infty$, where $\Psi = tr[\mathbb{E}[\ddot{q}(\beta_i)]\overline{\Omega}/2]$.*

It is important to note that we have $\Psi = 0$ in Corollary A when $q(b)$ is linear and hence $\ddot{q}(b) = 0$ for any b. It follows that $\sqrt{n}(\widehat{\theta} - \theta)$ is asymptotically normal with mean zero, regardless of the value of $\phi = \lim_{n,T\to\infty} \sqrt{n}/J_T$. In this case, we can instead obtain the limiting distribution only using n-asymptotics as

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d \mathcal{N}\left(0, \mathbb{E}[(q(\beta_i) - \mathbb{E}[q(\beta_i)])(q(\beta_i) - \mathbb{E}[q(\beta_i)])'] + J_T^{-1}\Lambda_T\right)$$

as $n \to \infty$ for some $0 < \Lambda_T < \infty$. The conventional mean-group (MG) estimator, $\widehat{\beta}_{MG} = n^{-1}\sum_{i=1}^n \widehat{\beta}_i$, is an example of this case with $J_T = T$, where we consider the least squares estimator $\widehat{\beta}_i$ in a static panel regression model.

In comparison, for the dynamic panel regression case, Hsiao et al. (1999) show that the conventional MG estimator (i.e., even when $q(b)$ is linear) has an $O_p(\sqrt{n}/T)$ bias term and we need $\lim_{n,T\to\infty} \sqrt{n}/T = 0$ to have a mean-zero limiting distribution. Unlike the static case, the source of this bias is the $O_p(1/T)$ bias term in $\widehat{\beta}_i$ from demeaning on the autoregressive form.

Recall that such a phenomenon is quite common in the panel econometrics literature. For instance, the maximum likelihood (ML) estimator for the "linear dynamic" fixed-effect panel regression is known to have non-zero mean of order $\sqrt{n/T}$ in the $\sqrt{nT}$-normalized limiting distribution (e.g., Hahn and Kuersteiner (2002)). However, the ML estimator for the "nonlinear" (either static or dynamic) fixed-effect panel regression is also known to have the $O(\sqrt{n/T})$-mean term in the $\sqrt{nT}$-normalized limiting distribution, unless the fixed-effect estimator and the slop-parameter estimator are asymptotically orthogonal (e.g., Hahn and Newey (2004)). Corollary A above establishes that a similar result holds for the nonlinear MG-type estimator in heterogeneous panel models.

In practice, we can correct the leading bias term in Corollary A using the panel jackknife estimator by Hahn and Newey (2004) or the split-panel jackknife bias correction method by Dhaene and Jochmans (2015).[18] For instance, we let $\widehat{\theta}_{-t} = n^{-1}\sum_{i=1}^{n} q(\widehat{\beta}_{i(-t)})$ be the leave-one-out estimator of θ using the subsample excluding all the observations on the tth period. Then, the jackknife estimator is defined as $\widehat{\theta}_{Jack} = T\widehat{\theta} - ((T-1)/T)\sum_{t=1}^{T}\widehat{\theta}_{-t}$, in which the $O_p(J_T^{-1})$ bias term Ψ/J_T is deleted and hence the leading bias term becomes of $O_p(1/J_T^2)$. Therefore, the jackknife estimator $\widehat{\theta}_{Jack}$ will have a limiting distribution centered at zero even when $\sqrt{n}/J_T \to \phi \in (0,\infty)$.

A.2 Proofs

Proof of Theorem 1 We define $\eta_i = \widehat{\beta}_i - \beta_i$ for each i. Since η_i is independent of β_i from Assumption 1-(ii), the distribution of $\widehat{\beta}_i = \beta_i + \eta_i$ hence can be obtained as the convolution of the distributions of β_i and η_i. More precisely, we denote F_{β,η_i}, $F(\equiv F_\beta)$, F_{η_i}, and $F_{\beta|\eta_i}$ be the joint distribution of (β_i,η_i), the marginal distribution of β_i, the marginal distribution of η_i, and the conditional distribution of β_i given η_i. Note that we allow η_i to be non-identically distributed over i and hence keep the index i in F_{β,η_i}, F_{η_i}, and $F_{\beta|\eta_i}$. Then, for each i, the distribution function of $\widehat{\beta}_i$ can be expressed as

$$\begin{aligned}
\int 1\{a \le b\} F_{\beta,\eta_i}(da) &= \int\int 1\{a_1 + a_2 \le b\} F_{\beta|\eta_i}(da_1|a_2)F_{\eta_i}(da_2) \\
&= \int \left[\int 1\{a_1 \le b - a_2\} F(da_1)\right] F_{\eta_i}(da_2) \\
&= \int F(b-a_2) F_{\eta_i}(da_2),
\end{aligned}$$

where the second equality is because η_i is independent of β_i. Now, for a given $b \in \mathcal{B}$, we write

$$\begin{aligned}
\widehat{F}_n(b) - F(b) &= \frac{1}{n}\sum_{i=1}^{n}\left(1\{\widehat{\beta}_i \le b\} - F(b)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(1\{\widehat{\beta}_i \le b\} - 1\{\beta_i \le b\}\right) - \left(\int F(b-c) F_{\eta_i}(dc) - F(b)\right)\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1\{\beta_i \le b\} - F(b)) + \frac{1}{n}\sum_{i=1}^{n}\left(\int F(b-c) F_{\eta_i}(dc) - F(b)\right)
\end{aligned}$$

[18] In the context of dynamic heterogeneous panel, the split-panel jackknife bias correction approach is used in a recent work by Okui and Yanagi (2019), which is developed independently of this paper.

$$\equiv \quad A_{n,1}(b) + A_{n,2}(b) + A_{n,3}(b). \tag{A.1}$$

First, since the binary indicator function $1\{\cdot \leq b\}$ is Donsker and $||\widehat{\beta}_i - \beta_i|| \to_p 0$ as $T \to \infty$ for each i from (10), $\sup_{b\in\mathcal{B}} |A_{n,1}(b)| = o_p(n^{-1/2})$ as $n, T \to \infty$ by Theorem 2.1 of van der Vaart and Wellner (2007). Second, it is well known that $\sup_{b\in\mathcal{B}} |A_{n,1}(b)| = o_p(1)$ as $n \to \infty$ by the Glivenko-Cantelli. For the last term, we note that $\mathbb{E}[||\widehat{\beta}_i - \beta_i||^2] = \mathbb{E}[||\eta_i||^2] \to_p 0$ as $T \to \infty$ for each i from (10). This L^2-convergence implies that $\widehat{\beta}_i$ converges to β_i in distribution for each i, which yields $\sup_{b\in\mathcal{B}} |A_{n,3}(b)| = o_p(1)$ as $T \to \infty$. The first results hence follows.

Now, for $b = (b_1, \cdots, b_k)'$, we denote $\dot{F}(b^*) = (\partial F(b)/\partial b_1|_{b=b^*}, \ldots, \partial F(b)/\partial b_k|_{b=b^*})'$ and $\ddot{F}(b^*) = (\partial \dot{F}(b)/\partial b_1'\big|_{b=b^*}, \ldots, \partial \dot{F}(b)/\partial b_k'\big|_{b=b^*})$ when F is twice differentiable at b^*. We then note that $F(b - \eta_i) - F(b) = -\eta_i'\dot{F}(b) + (1/2)\eta_i'\ddot{F}(b_i^*)\eta_i$ for some b_i^* between $b - \eta_i$ and b. However, as $\eta_i = O_p(J_T^{-1/2})$ and $\ddot{F}(\cdot)$ is assumed to be continuous, $\sup_{b\in\mathcal{B}} ||\ddot{F}(b_i^*) - \ddot{F}(b)|| = o_p(1)$ and hence $F(b - \eta_i) - F(b) = -\eta_i'\dot{F}(b) + (1/2)\eta_i'\ddot{F}(b)\eta_i + o_p(J_T^{-1})$. From (A.1), it follows that

$$\begin{aligned}
\sqrt{n}A_{n,3}(b) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int \{F(b-c) - F(b)\} F_{\eta_i}(dc) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int \left\{-c'\dot{F}(b) + \frac{1}{2}c'\ddot{F}(b)c + o_p\left(J_T^{-1}\right)\right\} F_{\eta_i}(dc) \\
&= \frac{1}{2\sqrt{n}}\sum_{i=1}^{n} tr\left[\ddot{F}(b)var(\eta_i)\right] + o_p\left(\frac{\sqrt{n}}{J_T}\right) \\
&= \frac{\sqrt{n}}{2J_T} tr\left[\ddot{F}(b) \times \frac{1}{n}\sum_{i=1}^{n} var\left(J_T^{1/2}\eta_i\right)\right] + o_p\left(\frac{\sqrt{n}}{J_T}\right),
\end{aligned} \tag{A.2}$$

where the first equality is because $\int F_{\eta_i}(dc) = 1$ and $F(b) = \int F(b) F_{\eta_i}(dc)$ for β_i is independent of η_i; the third equality is because $\int cF_{\eta_i}(dc) = 0$. Therefore, since $\sup_{b\in\mathcal{B}} |\sqrt{n}A_{n,1}(b)| = o_p(1)$ as $n, T \to \infty$ by Theorem 2.1 of van der Vaart and Wellner (2007) and $\sqrt{n}A_{n,2}(b) \to_d \mathcal{N}(0, F(b)(1 - F(b))$ as $n \to \infty$ by the functional central limit theorem, the second result follows. ■

Proof of Corollary A We note that

$$\begin{aligned}
\sqrt{n}(\widehat{\theta} - \theta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{q(\widehat{\beta}_i) - \mathbb{E}[q(\beta_i)]\right\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{q(\beta_i) - \mathbb{E}[q(\beta_i)]\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{q(\widehat{\beta}_i) - q(\beta_i)\right\} \\
&\equiv B_{n,1} + B_{n,2},
\end{aligned}$$

where $B_{n,1} \to_d \mathcal{N}(0, \mathbb{E}[(q(\beta_i) - \mathbb{E}[q(\beta_i)])(q(\beta_i) - \mathbb{E}[q(\beta_i)])'])$ as $n \to \infty$ by CLT. For $B_{n,2}$, for $\eta_i = \widehat{\beta}_i - \beta_i$, we have

$$B_{n,2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \eta_i' \dot{q}(\beta_i) + \frac{1}{2} \eta_i' \ddot{q}(\beta_i) \eta_i + o_p(||\eta_i^2||) \right\},$$

where $o_p(||\eta_i^2||) = o_p(J_T^{-1})$ from (10). Furthermore, $\mathbb{E}[\eta_i' \dot{q}(\beta_i)] = \mathbb{E}[\eta_i]' \, \mathbb{E}[\dot{q}(\beta_i)] = 0$ and $\mathbb{E}[\eta_i' \ddot{q}(\beta_i) \eta_i] = tr(\mathbb{E}[\ddot{q}(\beta_i)] var(\eta_i))$ since β_i and η_i are independent. Hence, similarly as (A.2),

$$\begin{aligned}
B_{n,2} &= \frac{1}{\sqrt{J_T}} \times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(J_T^{1/2} \eta_i \right)' \dot{q}(\beta_i) \\
&\quad + \frac{1}{2J_T} \times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \left(J_T^{1/2} \eta_i \right)' \ddot{q}(\beta_i) \left(J_T^{1/2} \eta_i \right) - \mathbb{E}\left[\left(J_T^{1/2} \eta_i \right)' \ddot{q}(\beta_i) \left(J_T^{1/2} \eta_i \right) \right] \right\} \\
&\quad + \frac{\sqrt{n}}{2J_T} \times \frac{1}{n} \sum_{i=1}^{n} tr \left(\mathbb{E}\left[\ddot{q}(\beta_i) \right] var \left(J_T^{1/2} \eta_i \right) \right) + o_p\left(\frac{\sqrt{n}}{J_T} \right) \\
&= O_p\left(\frac{1}{\sqrt{J_T}} \right) + O_p\left(\frac{1}{J_T} \right) + O\left(\frac{\sqrt{n}}{J_T} \right) + o_p\left(\frac{\sqrt{n}}{J_T} \right)
\end{aligned}$$

as $n, T \to \infty$, where the first two terms are by the CLT as $\mathbb{E}[\eta_i' \dot{q}(\beta_i)] = 0$; the third $O_p(\sqrt{n}/J_T)$ term is by the LLN. Therefore, unless $\mathbb{E}[\ddot{q}(\beta_i)] = 0$, $B_{n,2} = O_p(1)$ when $\sqrt{n}/J_T \to \phi < \infty$ as $n, T \to \infty$ since $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} tr(\mathbb{E}[\ddot{q}(\beta_i)] var(J_T^{1/2} \eta_i)) = tr[\mathbb{E}[\ddot{q}(\beta_i)] \overline{\Omega}] < \infty$.[19] The desired result follows, where the limiting distribution has (potentially) non-zero mean given as $\phi tr[\mathbb{E}[\ddot{q}(\beta_i)] \overline{\Omega}/2]$. ■

Proof of Theorem 2 We prove the second result here; the proof for the consistency is similar and omitted. For further details, see Massé (2004, Proposition 3.1). We first observe that

$$\begin{aligned}
& \sqrt{n} \int (b - \overline{\beta}) W(\mathcal{D}(b, \widehat{F}_n)) \widehat{F}_n(db) \\
&= \int (b - \overline{\beta}) W(\mathcal{D}(b, F)) \sqrt{n} (\widehat{F}_n - F)(db) \\
&\quad + \int (b - \overline{\beta}) \sqrt{n} \left\{ W(\mathcal{D}(b, \widehat{F}_n)) - W(\mathcal{D}(b, F)) \right\} \widehat{F}_n(db) \\
&= \int (b - \overline{\beta}) W(\mathcal{D}(b, F)) \widehat{\nu}_n(db) + \int (b - \overline{\beta}) \dot{W}(\delta_n(b)) \widehat{H}_n(b) \widehat{F}_n(db) \qquad \text{(A.3)}
\end{aligned}$$

for some $\widehat{\delta}_n(b)$ between $\mathcal{D}(b, \widehat{F}_n)$ and $\mathcal{D}(b, F)$, where $\sup_{b \in \mathcal{B}} |\widehat{\delta}_n(b) - \mathcal{D}(b, F)| = O_p(n^{-1/2})$. Similarly as the proofs of Massé (2004, Theorem 3.2) or Zuo et al. (2004, Theorem 2.1), we

[19] If $q(b)$ is linear in b, then $\ddot{q}(b) = 0$ for any b and hence $B_{n,2} = O_p(J_T^{-1/2}) = o_p(1)$ even when $\sqrt{n}/J_T \to \phi < \infty$ as $n, T \to \infty$.

note that

$$\left|\int (b-\overline{\beta})\dot{W}(\widehat{\delta}_n(b))\widehat{H}_n(b)\widehat{F}_n(db) - \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{F}_n(db)\right|$$
$$\leq \sup_{b\in\mathcal{B}}\left|\widehat{H}_n(b)\right| \int_{\{\mathcal{D}(b,F)+O_p(n^{-1/2})\geq d_0\}} \left\|b-\overline{\beta}\right\| \left|\dot{W}(\widehat{\delta}_n(b)) - \dot{W}(\mathcal{D}(b,F))\right| \widehat{F}_n(db) = o_p(1)$$

by Assumptions 3-(i), (ii), and (v), which yields

$$\int (b-\overline{\beta})\dot{W}(\delta_n(b))\widehat{H}_n(b)\widehat{F}_n(db) = \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{F}_n(db) + o_p(1).$$

Furthermore, we similarly have

$$\begin{aligned}
&\left|\int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{F}_n(db) - \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)F(db)\right| \\
\leq\ & \frac{1}{\sqrt{n}}\left|\int_{\mathcal{B}_0} (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{\nu}_n(db)\right| \\
&+\frac{1}{\sqrt{n}}\left|\int_{\mathcal{B}_0^c} (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{\nu}_n(db)\right| \\
=\ & o_p(1)
\end{aligned}$$

and hence

$$\int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)\widehat{F}_n(db) = \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)F(db) + o_p(1).$$

By Fubini's theorem and Assumptions 3-(iii) and 3-(iv), it follows that the second term in (A.3) can be rewritten as

$$\begin{aligned}
\int (b-\overline{\beta})\dot{W}(\widehat{\delta}_n(b))\widehat{H}_n(b)\widehat{F}_n(db) &= \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\widehat{H}_n(b)F(db) + o_p(1) \\
&= \int (b-\overline{\beta})\dot{W}(\mathcal{D}(b,F))\left\{\int h(b,c)\widehat{\nu}_n(dc)\right\}F(db) + o_p(1) \\
&= \iint (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h(c,b)F(dc)\widehat{\nu}_n(db) + o_p(1).
\end{aligned}$$

Therefore, by putting this expression into (A.3), we have

$$\begin{aligned}
&\sqrt{n}\int (b-\overline{\beta})W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db) \qquad \text{(A.4)} \\
=\ & \int\left\{\int (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))\right\}\widehat{\nu}_n(db) + o_p(1).
\end{aligned}$$

Likewise, we can also show that

$$\int W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db) = \int W(\mathcal{D}(b,F))F(db) + o_p(1). \tag{A.5}$$

From (A.4) and (A.5), therefore,

$$\begin{aligned}
&\sqrt{n}(\widehat{\beta}_{DW} - \overline{\beta}) \\
=\ & \frac{\sqrt{n}\int (b-\overline{\beta})W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db)}{\int W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db)} \\
=\ & \frac{\int \left\{\int (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))\right\}\sqrt{n}(F_n - F)(db)}{\int W(\mathcal{D}(b,F))F(db)} \\
& + \frac{\int \left\{\int (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))\right\}\sqrt{n}(\widehat{F}_n - F_n)(db)}{\int W(\mathcal{D}(b,F))F(db)} + o_p(1) \\
\equiv\ & \Phi_{n,1} + \Phi_{n,2} + o_p(1).
\end{aligned} \tag{A.6}$$

The first term $\Phi_{n,1}$ satisfies

$$\Phi_{n,1} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(K_F(\beta_i) - \mathbb{E}[K_F(\beta_i)]\right) \to_d \mathcal{N}\left(0, \mathbb{E}[K_F^0(\beta_i)K_F^0(\beta_i)']\right) \tag{A.7}$$

from Theorem A of Serfling (1980; p.226), where

$$K_F(b) = \frac{\int (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))}{\int W(\mathcal{D}(c,F))F(dc)}$$

and

$$K_F^0(b) = K_F(b) - \mathbb{E}[K_F(\beta_i)] = \frac{\int (c-\overline{\beta})\dot{W}(\mathcal{D}(c,F))h^0(c,b)F(dc) + (b-\overline{\beta})W(\mathcal{D}(b,F))}{\int W(\mathcal{D}(c,F))F(dc)}$$

with $h^0(c,b) = h(c,b) - \int h(c,b)F(db)$, because $\int (b-\overline{\beta})W(\mathcal{D}(b,F))F(db) = 0$ by construction.

For the second term $\Phi_{n,2}$, we note that

$$\begin{aligned}
\mathbb{E}\left\|\Phi_{n,2}\right\|^2 &= \frac{1}{n}\mathbb{E}\left\|\sum_{i=1}^{n}\left(K_F(\widehat{\beta}_i) - K_F(\beta_i)\right)\right\|^2 \\
&= \mathbb{E}\left\|K_F(\widehat{\beta}_i) - K_F(\beta_i)\right\|^2 \\
&\le \frac{\int \left\|c-\overline{\beta}\right\|^2 \dot{W}(\mathcal{D}(c,F))^2\mathbb{E}\left[h(c,\widehat{\beta}_i) - h(c,\beta_i)\right]^2 F(dc)}{\left(\int W(\mathcal{D}(c,F))F(dc)\right)^2}
\end{aligned}$$

$$+\frac{\mathbb{E}\left\|(\widehat{\beta}_i-\overline{\beta})W(\mathcal{D}(\widehat{\beta}_i,F))-(\beta_i-\overline{\beta})W(\mathcal{D}(\beta_i,F))\right\|^2}{\left(\int W(\mathcal{D}(c,F))F(dc)\right)^2}.$$

From Assumptions 2 and 3-(vi), we have

$$\mathbb{E}\left[h(c,\widehat{\beta}_i)-h(c,\beta_i)\right]^2=O\left(J_T^{-1}\right).$$

Hence the first term is $O(J_T^{-1})$ from Assumption 3-(iv) and $\int W(\mathcal{D}(c,F))F(dc)<\infty$ by construction. Furthermore, since $|\mathcal{D}(\widehat{\beta}_i,F)-\mathcal{D}(\beta_i,F)|=O_p(J_T^{-1/2})$ from Assumption 3-(vi), we have $|W(\mathcal{D}(\widehat{\beta}_i,F))-W(\mathcal{D}(\beta_i,F))|=O_p(J_T^{-1/2})$ by Assumption 3-(i) and hence

$$\begin{aligned}
&\mathbb{E}\left\|(\widehat{\beta}_i-\overline{\beta})W(\mathcal{D}(\widehat{\beta}_i,F))-(\beta_i-\overline{\beta})W(\mathcal{D}(\beta_i,F))\right\|^2\\
=\;&\mathbb{E}\left\|(\widehat{\beta}_i-\beta_i)W(\mathcal{D}(\beta_i,F))+O_p(J_T^{-1/2})\right\|^2\\
\leq\;&2\mathbb{E}\left\|\eta_i\right\|^2\mathbb{E}[W(\mathcal{D}(\beta_i,F))^2]+O\left(J_T^{-1}\right)=O\left(J_T^{-1}\right).
\end{aligned}$$

It follows that $\Phi_{n,2}\rightarrow_p 0$ as $n,T\rightarrow\infty$, which yields the desired result. ■

Proof of Theorem 3 First note that, using the affine invariance property of the statistical depth function, we have

$$\mathcal{D}(\widehat{\beta}_i,\widehat{F}_n)=\mathcal{D}(\widehat{\xi}_i,\widehat{G}_n)\tag{A.8}$$

from (17). Therefore, we can write

$$\widehat{\beta}_{DW}-\beta=\frac{\int(b-\beta)W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db)}{\int W(\mathcal{D}(b,\widehat{F}_n))\widehat{F}_n(db)}=\frac{\int(J_T^{-1/2}r)W(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr)}{\int W(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr)}\tag{A.9}$$

from (A.8) and by the change of variables with $r=J_T^{1/2}(b-\beta)$. However, similarly to the proof of Theorem 2 above, we have

$$\begin{aligned}
&\sqrt{n}\int rW(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr) &&\text{(A.10)}\\
=\;&\int\left\{\int s\dot{W}(\mathcal{D}(s,G))m(s,r)G(ds)+rW(\mathcal{D}(r,G))\right\}\widehat{\gamma}_n(dr)+o_p(1)
\end{aligned}$$

and

$$\int W(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr)=\int W(\mathcal{D}(r,G))G(dr)+o_p(1).\tag{A.11}$$

From (A.9), (A.10), and (A.11), we thus have

$$\sqrt{nJ_T}\left(\widehat{\beta}_{DW}-\beta\right)$$

$$
\begin{aligned}
&= \frac{\sqrt{n}\int rW(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr)}{\int W(\mathcal{D}(r,\widehat{G}_n))\widehat{G}_n(dr)} \\
&= \frac{\int\left\{\int s\dot{W}(\mathcal{D}(s,G))m(s,r)G(ds)+rW(\mathcal{D}(r,G))\right\}\sqrt{n}(\widehat{G}_n-G)(dr)}{\int W(\mathcal{D}(r,G))G(dr)}+o_p\left(1\right).
\end{aligned}
$$

The desired result follows using the same argument in the proof of Theorem 2 as

$$
\sqrt{nJ_T}\left(\widehat{\beta}_{DW}-\beta\right)\rightarrow_d \mathcal{N}\left(0,\mathbb{E}[K_G^0(\xi_i)K_G^0(\xi_i)']\right),
$$

where

$$
K_G^0(r)=\frac{\int s\dot{W}(\mathcal{D}(s,G))m^0(s,r)G(dr)+rW(\mathcal{D}(r,G))}{\int W(\mathcal{D}(s,G))G(ds)} \tag{A.12}
$$

since $\int rW(\mathcal{D}(r,G))G(dr)=0$ by construction. ■

References

Agulló J., C. Croux, and S. Van Aelst (2008). The multivariate least-trimmed squares estimator, *Journal of Multivariate Analysis*, 99, 311-338.

Aquaroa, M. and P. Čížek (2013). One-step robust estimation of fixed-effects panel data models, *Computational Statistics and Data Analysis*, 57, 536-548.

Aquaroa, M. and P. Čížek (2014). Robust estimation of dynamic fixed-effects panel data models. *Statististical Papers*, 55, 169-186.

Bai, J. (2009). Panel data models with interactive fixed effects, *Econometrica*, 77, 1229-1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models, *Econometrica*, 70, 191-221.

Baltagi, B.H. and G. Bresson (2012). A robust Hausman–Taylor estimator, in B.H. Baltagi, R.C. Hill, W.K. Newey, H.L. White (ed.) Essays in Honor of Jerry Hausman, *Advances in Econometrics*, 29, Emerald Group Publishing Limited, 175-214.

Boneva, L. and O. Linton (2016). A discrete-choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance, *Journal of Applied Econometrics*, 32, 1226-1243.

Bramati, M.C. and C. Croux (2007). Robust estimators for the fixed effects panel data model, *Econometrics Journal*, 10, 521-540.

Cantoni, E. (2004). A robust approach to longitudinal data analysis, *The Canadian Journal of Statistics*, 32, 169-180.

Chen, X., D.T. Jacho-Chavez, and O. Linton (2016). Averaging of an increasing number of moment condition estimators, *Econometric Theory*, 32, 30-70.

Cui, H. and Y. Tian (1994). Estimation of the projection absolute median deviation and its application (in Chinese), *Journal of Systems Science and Mathematical Sciences*, 14, 63-72.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed effects models, *Review of Economic Studies*, 82, 991-1030.

Dhaene, G. and Y. Zhu (2017). Median-based estimation of dynamic panel models with fixed effects, *Computational Statistics and Data Analysis*, 113, 398-423.

Dutta, S. and M.G. Genton (2017). Depth-weighted robust multivariate regression with application to sparse data, *Canadian Journal of Statistics,* 45(2), 164-184.

Griliches, Z. and J.A. Hausman (1986). Errors in variables in panel data, *Journal of Econometrics*, 31, 93-118.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects, *Econometrica*, 70, 1639-1657.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models, *Econometrica*, 72, 1295-1319.

Harding, M., C. Lamarche, and M.H. Pesaran (2020). Common correlated effects estimation of heterogeneous dynamic panel quantile regression models, *Journal of Applied Econometrics*, Forthcoming.

Hsiao, C. and M.H. Pesaran (2008). Random Coefficients Models, in (eds.) L. Matayas and P. Sevestre, *The Econometrics of Panel Data*, 3rd ed., Kluwer, 187-216.

Hsiao, C., M.H. Pesaran, and A.K. Tahmiscioglu (1999). Bayes estimation of short-run coefficients in dynamic panel data models, in (eds.) C. Hsiao, K. Lahiri, L.-F. Lee and M.H. Pesaran, *Analysis of Panels and Limited Dependent Variables: A Volume in Honour of G.S. Maddala*, Cambridge University Press, 268-296.

Huber, P.J. (1964). Robust estimation of a location parameter, *Annual of Mathematical Statistics*, 35, 73-101.

Huber, P.J. and E.M. Ronchetti (2009). *Robust Statistics*, 2nd ed., Wiley.

Koenker, R. and G. Bassett, Jr. (1978). Regression quantiles, *Econometrica*, 46, 33-50.

Kudraszow, N.L. and R.A. Maronna (2011). Estimates of MM type for the multivariate linear model, *Journal of Multivariate Analysis*, 102(9), 1280-1292.

Lee, Y. and D. Sul (2020). Trimmed mean group estimation, *mimeo*.

Lee, Y. and Y. Zhou (2015). Averaged instrumental variables estimators, *mimeo*.

Liu, R.Y. (1990). On a notation of data depth based on random simplices, *Annals of Statistics*, 18, 405-414.

Liu, R.Y. (1992). Data depth and multivariate rank tests, in (ed.) Y. Dodge, *L1-Statistical Analysis and Related Methods*, North-Holland, 279-294.

Liu, R.Y., J.M. Parelius, and K. Singh (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion), *Annals of Statistics*, 27, 783-858.

Liu, R.Y. and K. Singh (1993). A quality index based on data depth and multivariate rank tests, *Journal of the American Statistical Association*, 88 252-260.

Lucas, A., R. van Dijik, and T. Kloek (2007). Outlier robust GMM estimation of leverage determinants in linear dynamic panel data models, *mimeo*.

Mahalanobis, P.C. (1936). On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India*, 2, 49-55.

Maronna, R.A. and V.J. Yohai (1995). The behavior of the Stahel–Donoho robust multivariate estimator, *Journal of the American Statistical Association*, 90, 330-341.

Massé, J.-C. (2004). Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean, *Bernoulli*, 10, 397-419.

Okui, R. and T. Yanagi (2019). Panel data analysis with heterogeneous dynamics, *Journal of Econometrics*, 212, 451-475.

Pesaran, M.H. (2006). Estimation and inference in large heterogenous panels with a multi-factor error structure, *Econometrica*, 74, 967-1012.

Rousseeuw, P.J. (1984). Least Median of Squares Regression, *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point, in (eds.) W. Grossmann, G. Pflug, I. Vincze and W. Wertz, *Mathematical Statistics and Applications*, B, Reidel Publishing Company, 283-297.

Sawa, T. (1973). Almost unbiased estimator in simultaneous equation systems, *International Economic Review*, 14, 97-106.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley.

Swamy, P.A.V.B. (1970). Efficient inference in a random coefficient regression model, *Econometrica*, 38, 311-323.

van der Vaart, A.W. and J.A. Wellner (2007). Empirical processes indexed by estimated functions. *Asymptotics: Particles, Processes and Inverse Problems.* Institute of Mathematical Statistics Lecture Notes - Monograph Series 55, 234-252.

Wagenvoort, R. and R. Waldmann (2002). On B-robust instrumental variable estimation of the linear model with panel data, *Journal of Econometrics*, 106, 297-324.

Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics*, 642-656.

Zuo, Y. (2003). Projection-based depth functions and associated medians, *Annals of Statistics*, 31, 1460-1490.

Zuo, Y. (2006). Multidimensional trimming based on projection depth, *Annals of Statistics*, 34, 2211-2251.

Zuo, Y., H. Cui, and X. He (2004). On the Stahel-Donoho estimator and depth-weighted mean of multivariate data, *Annals of Statistics*, 32, 167-188.

Zuo, Y. and S. Lai (2011). Exact computation of bivariate projection depth and the Stahel–Donoho estimator, *Computational Statistics & Data Analysis*, 55, 1173-1179.

Zuo, Y. and R. Serfling (2000). General notions of statistical depth function, *Annals of Statistics*, 28, 461-482.