cemmap
centre for microdata methods and practice

Network and panel quantile effects via distribution regression

Victor Chernozhukov
Ivan Fernandez-Val
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP21/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

NETWORK AND PANEL QUANTILE EFFECTS VIA DISTRIBUTION REGRESSION

VICTOR CHERNOZHUKOV, IVAN FERNANDEZ-VAL, AND MARTIN WEIDNER

Abstract. This paper provides a method to construct simultaneous confidence bands for quantile functions and quantile effects in nonlinear network and panel models with unobserved two-way effects, strictly exogenous covariates, and possibly discrete outcome variables. The method is based upon projection of simultaneous confidence bands for distribution functions constructed from fixed effects distribution regression estimators. These fixed effects estimators are bias corrected to deal with the incidental parameter problem. Under asymptotic sequences where both dimensions of the data set grow at the same rate, the confidence bands for the quantile functions and effects have correct joint coverage in large samples. An empirical application to gravity models of trade illustrates the applicability of the methods to network data.

1. Introduction

Standard regression analyzes average effects of covariates on outcome variables. In many applications it is equally important to consider distributional effects. For example, a policy maker might be interested in the effect of an education reform not only on the mean but also the entire distribution of test scores or wages. Availability of panel data is very useful to identify *ceteris paribus* average and distributional effects because it allows the researcher to control for multiple sources of unobserved heterogeneity that might cause endogeneity or omitted variable problems. The idea is to use variation of the covariates over time for each individual or over individuals for each time period to account for unobserved individual and time effects. In this paper we develop inference methods for distributional effects in nonlinear models with two-way unobserved effects. They apply not only to traditional panel data models where the unobserved effects correspond to individual and time fixed

Date: March 21, 2018.

Initial Discussion: 7/8/2015. We thank Manuel Arellano and seminar participants at Bonn, Boston College, Bristol, Cemfi, Chicago, 2017 EEA-ESEM Conference, Humboldt, Northwestern, Princeton, UC-Davis, UC-Irvine, USC, York, UvA-Econometrics Panel Data Workshop, and Berkeley/CeMMAP Conference on Networks for comments. Gianluca Russo and Siyi Luo provided capable research assistance. Financial support from the National Science Foundation, and the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and the European Research Council grant ERC-2014-CoG-646917-ROMIA is gratefully acknowledged.

effects, but also to models for other types of data where the unobserved effects reflect some grouping structure such as unobserved sender and receiver effects in network data models.

We develop inference methods for quantile functions and effects. The quantile function corresponds to the marginal distribution of the outcome in a counterfactual scenario where the treatment covariate of interest is set exogenously at a desired level and the rest of the covariates and unobserved effects are hold fixed. The quantile effect is the difference of quantile functions at two different treatment levels. Our methods apply to continuous and discrete treatments by appropriate choice of the treatment levels, and have causal interpretation under standard uncounfoundness assumptions; see for example Chernozhukov Fernandez-Val and Melly (2013). The inference is based upon the generic method of Chernozhukov, Fernandez-Val, Melly, and Wuthrich (2016) that projects joint confidence bands for distributions into joint confidence bands for quantile functions and effects. This method has the appealing feature that applies without modification to any type of outcome, let it be continuous, discrete or mixed.

The key input for the inference method is a joint confidence band for the counterfactual distributions at the treatment levels of interest. We construct this band from fixed effects distribution regression (FE-DR) estimators of the conditional distribution of the outcome given the observed covariates and unobserved effects. In doing so, we extend the distribution regression approach to model conditional distributions with unobserved effects. This version of the DR model is semiparametric because not only the DR coefficients can vary with the level of the outcome as in the cross section case, but also the distribution of the unobserved effects is left unrestricted. We show that the FE-DR estimator can be obtained as a sequence of binary response fixed effects estimators where the binary response is an indicator of the outcome passing some threshold. To deal with the incidental parameter problem associated with the estimation of the unobserved effects (Neyman and Scott (1948)), we extend the analytical bias corrections of Fernandez-Val and Weidner (2016) for single binary response estimators to multiple (possibly a continuum) of binary response estimators. In particular, we establish functional central limit theorems for the fixed effects estimators of the DR coefficients and associated counterfactual distributions, and show the validity of the bias corrections under asymptotic sequences where the two dimensions of the data set pass to infinity at the same rate. As in the single binary response model, the bias corrections remove the asymptotic bias of the fixed effects estimators without increasing their asymptotic variances.

We implement the inference method using multiplier bootstrap (Giné and Zinn, 1984). This version of bootstrap constructs draws of an estimator as weighted averages of its influence function, where the weights are independent from the data. Compared to empirical bootstrap, multiplier bootstrap has the computational advantage that it does not involve any parameter reestimation. This advantage is particularly convenient in our setting because the parameter estimation require multiple nonlinear optimizations that can be highly dimensional due to the fixed effects. Multiplier bootstrap is also convenient to account for

data dependencies. In network data, for example, it might be important to account for reciprocity or pairwise clustering. Reciprocity arises because observational units corresponding to the same pair of agents but reversing their roles as sender and receiver might be dependent even after conditioning on the unobserved effects. By setting the weights of these observational units equal, we account for this dependence in the multiplier bootstrap. In addition to the previous practical reasons, there are some theoretical reasons for choosing multiplier bootstrap. Thus, Chernozhukov, Chetverikov and Kato (2016) established bootstrap functional central limit theorems for multiplier bootstrap in high dimensional settings that cover the network and panel models that we consider.

The methods developed in this paper apply to models that include unobserved effects to capture grouping or clustering structures in the data such as models for panel and network data. These effects allow us to control for unobserved group heterogeneity that might be related to the covariates causing endogeneity or omitted variable bias. They also serve to parsimoniously account for clustering dependences in the data. We illustrate the wide applicability with an empirical example to gravity models of trade. In this case the outcome is the volume of trade between two countries and each observational unit corresponds to a country pair indexed by exporter country (sender) and importer country (receiver). We estimate the distributional effects of gravity variables such as the geographical distance controlling for exporter and importer country effects that pick up unobserved heterogeneity that might be correlated with the gravity variables. We uncover significant heterogeneity in the effects of distance and other gravity variables across the distribution, which is missed by traditional mean methods. We also find that the Poisson model, which is commonly used in the trade literature to deal with zero trade in many country pairs, does not provide a good approximation to the distribution of the volume of trade due to severe overdispersion and outliers.

Literature review. Unlike mean effects, there are different ways to define distributional and quantile effects. For example, we can distinguish conditional effects versus unconditional or marginalized effects, or quantile effects versus quantiles of the effects. Here we give a brief review of the recent literature on distributional and quantile effects in panel data models emphasizing the following aspects: (1) type of effect considered; (2) type of unobserved effects in the model; and (3) asymptotic approximation. For the unobserved effects, we distinguish models with one-way effects versus two-way effects. For the asymptotic approximation we distinguish short panels with large N and fixed T versus long panels with large N and large T , where N and T denote the dimensions of the panel. We focus mainly on fixed effects approaches where the unobserved effects are treated as parameters to be estimated, but also mention some correlated random effects approaches that impose restrictions on the distribution of the unobserved effects. This paper deals with inference on marginalized quantile effects in large panels with two-way effects, which has not being previously considered in the literature.

Koenker (2004) introduced fixed effects quantile regression estimators of conditional quantile effects in large panel models with one-way individual effects using shrinkage to control the variability in the estimation of the unobserved effects. Lamarche (2010) discussed the optimal choice of a tuning parameter in Koenker's method. In the same framework, Kato, Galvao, and Montes-Rojas (2012), Galvao, Lamarche and Lima (2013), Galvao and Kato (2016) and Arellano and Weidner (2016) considered fixed effects quantile regression estimators without shrinkage and developed bias corrections. All these papers require that T pass to infinity faster than N, making difficult to extend the theory to models with two-way individual and time effects. Graham, Hahn and Powell (2009) found a special case where the fixed effects quantile regression estimator does not suffer of incidental parameter problem.

In short panels, Rosen (2012) showed that a linear quantile restriction is not sufficient to point identify conditional effects in a panel linear quantile regression model with unobserved individual effects. Chernozhukov, Fernandez-Val, Hahn and Newey (2013) and Chernozhukov, Fernandez-Val, Hoderlein, Holzmann and Newey (2015) discussed identification and estimation of marginalized quantile effects in nonseparable panel models with unobserved individual effects and location and scale time effects under a time homogeneity assumption. They showed that the effects are point identified only for some subpopulations and characterized these subpopulations. Graham, Hahn, Poirier and Powell (2015) considered quantiles of effects in linear quantile regression models with two-way effects. Finally, Abrevaya and Dahl (2008) and Arellano and Bonhomme (2016) developed estimators for conditional quantile effects in linear quantile regression model with unobserved individual effects using correlated random effects approaches.

Finally, we review previous applications of panel data methods to network data. These include Cruz-Gonzalez, Fernandez-Val and Weidner (2016), Dzemski (2017), Fernandez-Val and Weidner (2016), Graham (2016, 2017), and Yan, Jiang, Fienberg and Leng (2016), which developed methods for models of network formation with unobserved sender and receiver effects for directed and undirected networks. None of these papers consider estimation of quantile effects as the outcome variable is binary, whether or not a link is formed between two agents.

Plan of the paper. Section 2 introduces the distribution regression model with unobserved effects for network and panel data, and describes the quantities of interest including model parameters, distributions, quantiles and quantile effects. Section 3 discusses fixed effects estimation, bias corrections to deal with the incidental parameter problem, and uniform inference methods. Section 4 provides asymptotic theory for the fixed effects estimators, bias corrections, and multiplier bootstrap. Section 5 and 6 report results of the empirical application to the gravity models of trade and a Monte Carlo simulation calibrated to the application, respectively. The proofs of the main results are given in the Appendix, and additional technical results are provided in the Supplementary Appendix.

Notation. For any two real numbers a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a real number a, $\lfloor a \rfloor$ denotes the integer part of a. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality or number of elements of $\mathcal{A}$.

2. Model and Parameters of Interest

2.1. **Distribution Regression Model with Unobserved Effects.** We observe the data set $\{(y_{ij}, x_{ij}) : (i, j) \in \mathcal{D}\}$, where y_{ij} is a scalar outcome variable with region of interest $\mathcal{Y}$, and x_{ij} is a vector of covariates with support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$.[1] The variable y_{ij} can be discrete, continuous or mixed. The subscripts i and j index individuals and time periods in traditional panels, but they might index other dimensions in more general data structures. In our empirical application, for example, we use a panel where y_{ij} is the volume of trade between country i and country j, and x_{ij} includes gravity variables such as the distance between country i and country j. Both i and j index countries as exporters and importers respectively. The set $\mathcal{D}$ contains the indexes of the units that are observed. It is a subset of the set of all possible pairs $\mathcal{D}_0 := \{(i, j) : i = 1, \ldots, I; j = 1, \ldots, J\}$, where I and J are the dimensions of the panel. We introduce $\mathcal{D}$ to allow for missing data that are common in panel and network applications. For example, in the trade application $I = J$ and $\mathcal{D} = \mathcal{D}_0 \setminus \{(i, i) : i = 1, \ldots, I\}$ because we do not observe trade of a country with itself. We denote the total number of observed units by n, i.e. $n = |\mathcal{D}|$.

Let v_i and w_j denote vectors of unspecified dimension that contain unobserved random variables or effects that might be related to the covariates x_{ij}. In traditional panels, v_i are individual effects that capture unobserved individual heterogeneity and w_j are time effects that account for aggregate shocks. More generally, these variables serve to capture some forms of endogeneity and group dependencies in a parsimonious fashion. We specify the conditional distribution of y_{ij} given (x_{ij}, v_i, w_j) using the *distribution regression (DR) model with unobserved effects*

$$F_{y_{ij}}(y \mid x_{ij}, v_i, w_j) = \Lambda_y(P(x_{ij})'\beta(y) + \alpha(v_i, y) + \gamma(w_j, y)), \quad y \in \mathcal{Y}, \quad (i, j) \in \mathcal{D}, \tag{1}$$

where Λ_y is a known link function such as the normal or logistic distribution, which may vary with y, $x \mapsto P(x)$ is a dictionary of transformations of x such us polynomials, b-splines and tensor products that has a one as the first component, $\beta(y)$ is an unknown parameter vector, which can vary with y, and $(v, y) \mapsto \alpha(v, y)$ and $(w, y) \mapsto \gamma(w, y)$ are unspecified measurable functions. This DR model is a semiparametric model for the conditional distribution because $y \mapsto \theta(y) := (\beta(y), \alpha_1(y), \ldots, \alpha_I(y), \gamma_1(y), \ldots, \gamma_J(y))$ is a function-valued parameter and the dimension of $\theta(y)$ varies with I and J, although we do not make this dependence explicit. We shall treat the dimension of $P(x)$ as fixed in the asymptotic analysis

[1] If y_{ij} has unbounded support, then the region $\mathcal{Y}$ is usually a subset of the support to avoid tail estimation.

When y_{ij} is continuous, the model (1) has the following representation as an implicit nonseparable model by the probability integral transform

$$\Lambda_{y_{ij}}(P(x_{ij})'\beta(y_{ij}) + \alpha(v_i, y_{ij}) + \gamma(w_j, y_{ij})) = u_{ij}, \quad u_{ij} \mid x_{ij}, v_i, w_j \sim U(0,1),$$

where the error u_{ij} represents the unobserved ranking of the observation y_{ij} in the conditional distribution. The parameters of the model are related to derivatives of the conditional quantiles. Let $Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j)$ be the u-quantile of y_{ij} conditional on (x_{ij}, v_i, w_j) defined as the left-inverse of $y \mapsto F_{y_{ij}}(y \mid x_{ij}, v_i, w_j)$ at u, namely

$$Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j) = \inf\{y \in \mathcal{Y} : F_{y_{ij}}(y \mid x_{ij}, v_i, w_j) \geq u\} \wedge \sup\{y \in \mathcal{Y}\},$$

and $x_{ij} = (x_{ij}^1, \ldots, x_{ij}^{d_x})$.[2] Then, it can be shown that if $x_{ij} \mapsto Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j)$ is differentiable,

$$\beta_k(y)|_{y=Q_{y_{ij}}(u|x_{ij},v_i,w_j)} \propto -\partial_{x_{ij}^k} Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j), \quad k = 1, \ldots, d_x, \quad \partial_{x_{ij}^k} := \partial/\partial x_{ij}^k,$$

and

$$\left.\frac{\beta_\ell(y)}{\beta_k(y)}\right|_{y=Q_{y_{ij}}(u|x_{ij},v_i,w_j)} = \frac{\partial_{x_{ij}^\ell} Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j)}{\partial_{x_{ij}^k} Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j)}, \quad \ell, k = 1, \ldots, d_x,$$

provided that $\partial_{x_{ij}^k} Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j) \neq 0$. The DR coefficients therefore are proportional to (minus) derivatives of the conditional quantile function, and ratios of DR coefficients correspond to ratios of derivatives.

Remark 1 (Parametric models)**.** There are many parametric models that are special cases of the DR model. For example, the standard linear model with two-way effects

$$y_{ij} = x_{ij}'\beta + \alpha(v_i) + \gamma(w_j) + \sigma\varepsilon_{ij}, \quad \varepsilon_{ij} \mid x_{ij}, v_i, w_j \sim G,$$

is a special case of the model (1) with link function $\Lambda_y = G$ for all y,

$$\beta(y) = (e_1 y - \beta)/\sigma, \quad \alpha(v_i, y) = -\alpha(v_i)/\sigma, \quad \gamma(w_j, y) = -\gamma(w_j)/\sigma,$$

where e_1 is a unit vector of dimension d_x with a one in the first component. This location-shift model imposes that all the parameters, other than the intercept, are fixed with respect to the level y. This greatly restricts the heterogeneity in the conditional quantile effects,

$$\partial_{x^\ell} Q_{y_{ij}}(u \mid x_{ij}, v_i, w_j) = \beta_\ell, \quad \forall (u, x_{ij}, v_i, w_j),$$

which are constant not only with respect to the quantile index but also with respect to the observed and unobserved covariates.

The Cox proportional hazard panel model for duration data

$$t(y_{ij}) = P(x_{ij})'\beta + \alpha(v_i) + \gamma(w_j) + \varepsilon_{ij},$$

[2]We use the convention $\inf\{\emptyset\} = +\infty$.

where ε_{ij} has an extreme value distribution independent of (x_{ij}, v_i, w_j) and $t(y)$ is an unknown monotone transformation, is also a special case of the model (1) with link function $\Lambda_y(\iota) = 1 - \exp(-\exp -\iota)$, the complementary log-log distribution, and parameters

$$\beta(y) = \beta - e_1 t(y), \quad \alpha(v_i, y) = \alpha(v_i), \quad \gamma(w_j, y) = \gamma(w_j),$$

where again only the intercept varies with the level y.

Another example is the Poisson model for count data with two-way effects where, for $\lambda_{ij} = \exp(P(x_{ij})'\beta + \alpha(v_i) + \gamma(w_j))$,

$$F_{y_{ij}}(y \mid x_{ij}, v_i, w_j) = \exp(-\lambda_{ij}) \sum_{k=0}^{\lfloor y \rfloor} \lambda_{ij}^k / k!, \quad y \geq 0,$$

where $\lfloor y \rfloor$ denotes the integer part of y. This conditional distribution is encompassed by the model (1) with link function $\Lambda_y(\iota) = \exp(-\exp(\iota)) \sum_{k=0}^{\lfloor y \rfloor} \exp(\iota)^k / k!$, the incomplete gamma function, and parameters $\beta(y) = \beta$, $\alpha(v_i, y) = \alpha(v_i)$,and $\gamma(w_j, y) = \gamma(w_j)$ that do not vary with y. In this case the link function changes with y. ■

2.2. **Estimands.** We are interested in measuring the effect on the outcome of changing one of the covariates holding the rest of the covariates and the unobserved effects fixed. Let $x = (t, z')'$, where t is the covariate of interest or treatment and z are the rest of the covariates that usually play the role of controls. One effect of interest is the quantile (left-inverse) function (QF)

$$Q_k(\tau) = F_k^{\leftarrow}(\tau) := \inf\{y \in \mathcal{Y} : F_k(y) \geq \tau\} \wedge \sup\{y \in \mathcal{Y}\}, \quad \tau \in (0, 1),$$

where

$$F_k(y) = n^{-1} \sum_{(i,j) \in \mathcal{D}} \Lambda_y(P(t_{ij}^k, z_{ij}')'\beta(y) + \alpha(v_i, y) + \gamma(w_j, y)),$$

t_{ij}^k is a level of the treatment that may depend on t_{ij}, and $k \in \{0, 1\}$. We provide examples below. Note that in the construction of the counterfactual distribution F_k, we marginalize (x_{ij}, v_i, w_j) using the empirical distribution. The resulting effects are *finite population* effects. We shall focus on these effects because conditioning on the covariates and unobserved effects is natural in the trade application.[3] We construct the quantile effect function (QEF) by taking differences of the QF at two treatment levels

$$\Delta(\tau) = Q_1(\tau) - Q_0(\tau), \quad \tau \in (0, 1).$$

[3]The distinction between finite and infinite population effects does not affect estimation, but affects inference (Abadie, Athey, Imbens and Wooldridge, 2014). The estimators of infinite population effects need to account for the additional sampling variation coming from the estimation of the distribution of (x_{ij}, v_i, w_j).

We can also obtain the average effect using the relationship between averages and distributions. Thus, the average effect is

$$\Delta = \mu_1 - \mu_0,$$

where μ_k is the counterfactual average obtained from F_k as

$$\mu_k = \int_{\mathcal{Y}} [1(y \geq 0) - F_k(y)]\nu(dy), \quad k \in \{0, 1\}, \tag{2}$$

where ν is either the counting measure when $\mathcal{Y}$ is countable or the Lebesgue measure otherwise.[4]

The choice of the levels t_{ij}^0 and t_{ij}^1 is usually based on the scale of the treatment:

- If the treatment is binary, $\Delta(\tau)$ is the τ-quantile treatment effect with $t_{ij}^0 = 0$ and $t_{ij}^1 = 1$.
- If the treatment is continuous, $\Delta(\tau)$ is the τ-quantile effect of a unitary or one standard deviation increase in the treatment with $t_{ij}^0 = t_{ij}$ and $t_{ij}^1 = t_{ij} + d$, where d is 1 or the standard deviation of t_{ij}.
- If the treatment is the logarithm of a continuos treatment, $\Delta(\tau)$ is the τ-quantile effect of doubling the treatment (100% increase) with $t_{ij}^0 = t_{ij}$ and $t_{ij}^1 = t_{ij} + \log 2$.

For example, in the trade application we use the levels $t_{ij}^0 = 0$ and $t_{ij}^1 = 1$ for binary covariates such as the indicators for common legal system and free trade area, and $t_{ij}^0 = t_{ij}$ and $t_{ij}^1 = t_{ij} + \log 2$ for the logarithm of distance.

All the previous estimands have causal interpretation under the standard unconfoundedness or conditional independence assumption for panel data where the conditioning set includes not only the observed controls but also the unobserved effects.

3. Fixed Effects Estimation and Uniform Inference

To simplify the notation in this section we write $P(x_{ij}) = x_{ij}$ without loss of generality, and define $\alpha_i(y) := \alpha(v_i, y)$ and $\gamma_j(y) := \gamma(w_j, y)$.

3.1. Fixed Effects Distribution Regression Estimator.

The parameters of the PDR model can be estimated from multiple binary regressions with two-way effects. To see this, note that the conditional distribution in (1) can be expressed as

$$\Lambda_y(x_{ij}'\beta(y) + \alpha_i(y) + \gamma_j(y)) = \mathbb{E}[1\{y_{ij} \leq y\} \mid x_{ij}, v_i, w_j].$$

[4]The derivation of μ_k assumes that $\int_{\mathcal{Y}} dF_k(y) = 1$, i.e. $\mathcal{Y}$ is the support of dF_k.

Accordingly, we can construct a collection of binary variables,

$$1\{y_{ij} \leq y\}, \quad (i,j) \in \mathcal{D}, \quad y \in \mathcal{Y},$$

and estimate the parameters for each y by conditional maximum likelihood with fixed effects. Thus, $\widehat{\theta}(y) := (\widehat{\beta}(y), \widehat{\alpha}_1(y), \ldots, \widehat{\alpha}_I(y), \widehat{\gamma}_1(y), \ldots, \widehat{\gamma}_J(y))$, the *fixed effects distribution regression estimator* of $\theta(y) := (\beta(y), \alpha_1(y), \ldots, \alpha_I(y), \gamma_1(y), \ldots, \gamma_J(y))$, is obtained as

$$\widehat{\theta}(y) \in \underset{\theta \in \mathbb{R}^{d_x+I+J}}{\operatorname{argmax}} \sum_{(i,j)\in\mathcal{D}} \Big(1\{y_{ij} \leq y\} \log \Lambda_y(x_{ij}'\beta + \alpha_i + \gamma_j) \\ + 1\{y_{ij} > y\} \log[1 - \Lambda_y(x_{ij}'\beta + \alpha_i + \gamma_j)] \Big), \tag{3}$$

for $y \in \mathcal{Y}$. When the link function is the normal or logistic distribution, the previous program is concave and smooth in parameters and therefore has good computational properties. See Fernandez-Val and Weidner (2016) and Cruz-Gonzalez, Fernandez-Val and Weidner (2016) for a discussion of computation of logit and probit regressions with two-way effects.

The quantile functions and effects are estimated via plug-in rule, i.e.,

$$\widehat{Q}_k(\tau) = \widehat{F}_k^{\leftarrow}(\tau) \wedge \sup\{y \in \mathcal{Y}\}, \quad \tau \in (0,1), \quad k \in \{0,1\},$$

where

$$\widehat{F}_k(y) = n^{-1} \sum_{(i,j)\in\mathcal{D}} \Lambda_y((t_{ij}^k, z_{ij}')'\widehat{\beta}(y) + \widehat{\alpha}_i(y) + \widehat{\gamma}_j(y)), \quad y \in \mathcal{Y},$$

and

$$\widehat{\Delta}(\tau) = \widehat{Q}_1(\tau) - \widehat{Q}_0(\tau) \quad \tau \in (0,1).$$

Remark 2 (Computation)**.** When $\mathcal{Y}$ is not finite, we replace $\mathcal{Y}$ by a finite subset $\bar{\mathcal{Y}}$ such that the Hausdorff distance between $\bar{\mathcal{Y}}$ and $\mathcal{Y}$ goes to zero at a rate faster than $1/\sqrt{n}$.

3.2. **Incidental Parameter Problem and Bias Corrections.** Fixed effects estimators can be severely biased in nonlinear models because of the incidental parameter problem (Neyman and Scott, 1948). These models include the binary regressions that we estimate to obtain the DR coefficients and estimands. We deal with the incidental parameter problem using the analytical bias corrections of Fernandez-Val and Weidner (2016) for parameters and average partial effects (APE) in binary regressions with two-way effects. We note here that the distributions $F_0(y)$ and $F_1(y)$ can be seen as APE, i.e. they are averages of functions of the data, unobserved effects and parameters.

The bias corrections are based on expansions of the bias of the fixed effects estimators as $I, J \to \infty$. For example,

$$\mathbb{E}[\widehat{F}_k(y) - F_k(y)] = \frac{I}{n} B_k^{(F)}(y) + \frac{J}{n} D_k^{(F)}(y) + R_k^{(F)}(y), \tag{4}$$

where $nR_k^{(F)}(y) = o(I \vee J)$.[5] In Section 4 we establish that this expansion holds uniformly in $y \in \mathcal{Y}$ and $k \in \{0,1\}$, i.e.

$$\sup_{k\in\{0,1\},y\in\mathcal{Y}} \|nR_k^{(F)}(y)\| = o(I \vee J).$$

This result generalizes the analysis of Fernandez-Val and Weidner (2016) from a single binary regression to multiple (possibly a continuum) of binary regressions. This generalization is required to implement our inference methods for quantile functions and effects.

The expansion (4) is the basis for the bias corrections. Let $\widehat{B}_k^{(F)}(y)$ and $\widehat{D}_k^{(F)}(y)$ be estimators of $B_k^{(F)}(y)$ and $D_k^{(F)}(y)$, which are uniformly consistent in $y \in \mathcal{Y}$ and $k \in \{0,1\}$. Bias corrected fixed effects estimators of F_k and Q_k are formed as

$$\begin{aligned} \widetilde{Q}_k(\tau) &= \widetilde{F}_k^{\leftarrow}(\tau) \wedge \sup\{y \in \mathcal{Y}\}, \\ \widetilde{F}_k(y) &= \widehat{F}_k(y) - \frac{I}{n}\widehat{B}_k(y) - \frac{J}{n}\widehat{D}_k(y), \quad y \in \mathcal{Y}. \end{aligned}$$

We also use the corrected estimators $\widetilde{F}_k$ as the basis for inference and to form a bias corrected estimator of the average effect.

Remark 3 (Monotonization)**.** If the bias corrected estimator $y \mapsto \widetilde{F}_k(y)$ is non-monotone on $\mathcal{Y}$, we can rearrange it into a monotone function by simply sorting the values of function in a nondecreasing order. Chernozhukov, Fernandez-Val and Galichon (2009) showed that the rearrangement improves the finite sample properties of the estimator. ■

3.3. **Uniform Inference.** Our inference goal is to construct confidence bands that cover the QF $\tau \mapsto Q_k(\tau)$ and the QEF $\tau \mapsto \Delta(\tau)$ simultaneously over a set of quantiles $\mathcal{T} \subseteq [\varepsilon, 1-\varepsilon]$, for some $0 < \varepsilon < 1/2$, and treatment levels $k \in \mathcal{K} \subseteq \{0,1\}$. The set $\mathcal{T}$ is chosen such that $Q_k(\tau) \in [\inf\{y \in \mathcal{Y}\}, \sup\{y \in \mathcal{Y}\}]$, for all $\tau \in \mathcal{T}$ and $k \in \mathcal{K}$.

We use the generic method of Chernozhukov, Fernández-Val, Melly and Wüthrich (2016) to construct confidence bands for quantile functions and effects from confidence bands for the corresponding distributions. Let $\mathbb{D}$ denote the space of weakly increasing functions, mapping $\mathcal{Y}$ to $[0,1]$. Assume we have a confidence band $I_k = [L_k, U_k]$ for F_k, with lower and upper endpoint functions $y \mapsto L_k(y)$ and $y \mapsto U_k(y)$ such that $L_k, U_k \in \mathbb{D}$ and $L_k(y) \leq U_k(y)$ for all $y \in \mathcal{Y}$.[6] We say that I_k covers F_k if $F_k \in I_k$ pointwise, namely $L_k(y) \leq F_k(y) \leq U_k(y)$ for all $y \in \mathcal{Y}$. If U_k and L_k are some data-dependent bands, we say that I_k is a confidence band for F_k of level p, if I_k covers F_k with probability at least p. Similarly, we say that the set of bands $\{I_k : k \in \mathcal{K}\}$ is a joint confidence band for the set of

[5]Fernandez-Val and Weidner (2016) considered the case where $n = IJ$, i.e. there is no missing data, so that $I/n = 1/J$ and $J/n = 1/I$.

[6]If $[L'_k, U'_k]$ is a confidence band for F_k that does not obey the constraint $L'_k, U'_k \in \mathbb{D}$, we can transform $[U'_k, L'_k]$ into a new band $[L_k, U_k]$ such that $L_k, U_k \in \mathbb{D}$ using the rearrangement method of Chernozhukov, Fernandez-Val and Galichon (2009).

functions $\{F_k : k \in \mathcal{K}\}$ of level p, if I_k covers F_k with probability at least p simultaneously over $k \in \mathcal{K}$. The index set $\mathcal{K}$ can be a singleton to cover individual confidence bands or $\mathcal{K} = \{0, 1\}$ to cover joint confidence bands. In Section 4 we provide a multiplier bootstrap algorithm for computing joint confidence bands based on the joint asymptotic distribution of the bias corrected estimators $\{\widetilde{F}_k : k \in \mathcal{K}\}$.

The following result provides a method to construct joint confidence bands for $\{Q_k = F_k^{\leftarrow} : k \in \mathcal{K}\}$, from joint confidence bands for $\{F_k : k \in \mathcal{K}\}$.

Lemma 1 (Chernozhukov, Fernandez-Val, Melly, and Wuthrich, 2016)**.** *Consider a set of distribution functions $\{F_k : k \in \mathcal{K}\}$ and endpoint functions $\{L_k : k \in \mathcal{K}\}$ and $\{U_k : k \in \mathcal{K}\}$ with components in the class $\mathbb{D}$. If $\{F_k : k \in \mathcal{K}\}$ is jointly covered by $\{I_k : k \in \mathcal{K}\}$ with probability p, then $\{Q_k = F_k^{\leftarrow} : k \in \mathcal{K}\}$ is jointly covered by $\{I_k^{\leftarrow} : k \in \mathcal{K}\}$ with probability p, where*

$$I_k^{\leftarrow}(\tau) := [U_k^{\leftarrow}(\tau), L_k^{\leftarrow}(\tau)], \quad \tau \in \mathcal{T}, \quad k \in \mathcal{K}.$$

This Lemma establishes that we can construct confidence bands for quantile functions by inverting the endpoint functions of confidence bands for distribution functions. The geometric intuition is that the inversion amounts to rotate and flip the bands, and these operations preserve coverage.

We next construct simultaneous confidence bands for the quantile effect function $\tau \mapsto \Delta(\tau)$ defined by

$$\Delta(\tau) = Q_1(\tau) - Q_0(\tau) = F_1^{\leftarrow}(\tau) - F_0^{\leftarrow}(\tau), \quad \tau \in \mathcal{T}.$$

The basic idea is to take appropriate differences of the bands for the quantile functions Q_1 and Q_0 as the confidence band for the quantile effect. Specifically, suppose we have the set of confidence bands $\{I_k^{\leftarrow} = [U_k^{\leftarrow}, L_k^{\leftarrow}] : k = 0, 1\}$ for the set of functions $\{F_k^{\leftarrow} : k = 0, 1\}$ of level p. Chernozhukov, Fernández-Val, Melly and Wüthrich (2016) showed that a confidence band for the difference $Q_1 - Q_0$ of size p can be constructed as $[U_1^{\leftarrow} - L_0^{\leftarrow}, L_1^{\leftarrow} - U_0^{\leftarrow}]$, i.e. the pointwise Minkowski difference $\ominus$ between $I_1^{\leftarrow}$ and $I_0^{\leftarrow}$.

Lemma 2 (Chernozhukov, Fernandez-Val, Melly, and Wuthrich, 2016)**.** *Consider a set of distribution functions $\{F_k : k = 0, 1\}$ and endpoint functions $\{L_k : k = 0, 1\}$ and $\{U_k : k = 0, 1\}$, with components in the class $\mathbb{D}$. If the set of distribution functions $\{F_k : k = 0, 1\}$ is jointly covered by the set of bands $\{I_k : k = 0, 1\}$ with probability p, then the quantile effect function $\Delta = F_1^{\leftarrow} - F_0^{\leftarrow}$ is covered by $I_\Delta^{\leftarrow}$ with probability at least p, where $I_\Delta^{\leftarrow}$ is defined by:*

$$I_\Delta^{\leftarrow}(\tau) := [U_1^{\leftarrow}(\tau), L_1^{\leftarrow}(\tau)] \ominus [U_0^{\leftarrow}(\tau), L_0^{\leftarrow}(\tau)] = [U_1^{\leftarrow}(\tau) - L_0^{\leftarrow}(\tau), L_1^{\leftarrow}(\tau) - U_0^{\leftarrow}(\tau)], \quad \tau \in \mathcal{T}.$$

4. Asymptotic Theory

This section derives the asymptotic properties of the fixed effect estimators of $y \mapsto \beta(y)$ and $\{F_k : k \in \mathcal{K}\}$, as both dimensions I and J grow to infinity. We focus on the case where the link function is the logistic distribution at all levels, $\Lambda_y = \Lambda$, where $\Lambda(\xi) = (1 + \exp(-\xi))^{-1}$. We choose the logistic distribution for analytical convenience. In this case the Hessian of the log-likelihood function does not depend on y_{it}, what leads to several simplifications in the asymptotic expansions. For the case of single binary regressions, Fernandez-Val and Weidner (2016) showed that the properties of fixed effects estimators are similar for the logistic distribution and other smooth log-concave distributions such as the normal distribution. Accordingly, we expect that our results can be extended to other link functions using conceptually similar arguments but more tedious derivations.

We make the following assumptions:

Assumption 1 (Sampling and Model Conditions)**.**

(i) *Sampling: The outcome variable y_{ij} is independently distributed over i and j conditional on all the observed and unobserved covariates $\mathcal{C}_B := \{(x_{ij}, v_i, w_j) : (i,j) \in \mathcal{D}\}$.*

(ii) *Model: For all $y \in \mathcal{Y}$,*

$$F_{y_{ij}}(y \mid \mathcal{C}_B) = F_{y_{ij}}(y \mid x_{ij}, v_i, w_j) = \Lambda(x_{ij}'\beta(y) + \alpha(v_i, y) + \gamma(w_j, y)),$$

where $y \mapsto \beta(y)$, $y \mapsto \alpha(\cdot, y)$ and $y \mapsto \gamma(\cdot, y)$ are measurable functions.

(iii) *Compactness: $\mathcal{XVW}$, the support of (x_{ij}, v_i, w_j), is a compact set, and $y \mapsto \alpha(v_i, y)$ and $y \mapsto \gamma(w_j, y)$ are a.s. uniformly bounded on $\mathcal{Y}$.*

(iv) *Compactness and smoothness: Either $\mathcal{Y}$ is a discrete finite set, or $\mathcal{Y} \subset \mathbb{R}$ is a bounded interval. In the latter case, we assume that the conditional density function $f_{y_{ij}}(y \mid x_{ij}, v_i, w_j)$ exists, is uniformly bounded above and away from zero, and is uniformly continuous in y, uniformly on $\operatorname{int}(\mathcal{Y}) \times \mathcal{XVW}$, where $\operatorname{int}(\mathcal{Y})$ is the interior of $\mathcal{Y}$.*

(v) *Missing data: There is only a fixed number of missing observations for every i and j, that is, $\max_i(J - |\{(i', j') \in \mathcal{D} : i' = i\}|) \leq c_2$ and $\max_j(I - |\{(i', j') \in \mathcal{D} : j' = j\}|) \leq c_2$ for some constant $c_2 < \infty$ that is independent of the sample size.*

(vi) *Non-collinearity: The regressors x_{ij} are non-collinear after projecting out the two-way fixed effects, that is, there exists a constant $c_3 > 0$, independent of the sample size, such that*

$$\min_{\{\delta \in \mathbb{R}^{d_x} : \|\delta\| = 1\}} \min_{(a,b) \in \mathbb{R}^{I+J}} \left[\frac{1}{n} \sum_{(i,j) \in \mathcal{D}} (x_{ij}'\delta - a_i - b_j)^2 \right] \geq c_3.$$

(vii) *Asymptotics: We consider asymptotic sequences where $I_n, J_n \to \infty$ with $I_n/J_n \to c$ for some positive and finite c, as the total sample size $n \to \infty$. We drop the indexing by n from I_n and J_n, i.e. we shall write I and J.*

Remark 4 (Sufficient conditions for Assumption 1)**.** Part (i) holds if $(y_{ij}, x_{ij}, v_i, w_j)$ is i.i.d. over i and j, but it is more general as it does not restrict the distribution of (x_{ij}, v_i, w_j) nor its dependence across i and j. We show how to relax this assumption allowing for a form of weak conditional dependence in Section 4.4. Part (ii) holds if the observed covariates are strictly exogenous conditional on the unobserved effects and the conditional distribution is correctly specified for all $y \in \mathcal{Y}$. Part (iii) imposes that the support of the covariates and the unobserved effects is a compact set. For fixed values y it is possible to obtain asymptotic results of our estimators without the compact support assumption, see e.g. Yan, Jiang, Fienberg and Leng (2016), but deriving empirical process results that hold uniformly over y is much more involved without this assumption. Part (iv) can be slightly weakened to Lipschitz continuity with uniformly bounded Lipschitz constant, instead of differentiability. For mixed random variables with mass points at the boundary of the support, the data generating process for the mass points can be arbitrarily different from the rest of the support because the density $y \mapsto f_{y_{ij}}(y \mid \cdot)$ only needs to be continuous in the interior of $\mathcal{Y}$. If the panel is balanced, part (vi) can be stated as

$$\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\widetilde{x}_{ij}\widetilde{x}_{ij}' \;\geq\; c_3\,\mathbb{I}_{d_x},$$

where $\widetilde{x}_{ij} = x_{ij} - x_{i\cdot} - x_{\cdot j} + x_{\cdot\cdot}$, $x_{i\cdot} = J^{-1}\sum_{j=1}^{J} x_{ij}$, $x_{\cdot j} = I^{-1}\sum_{i=1}^{I} x_{ij}$, and $x_{\cdot\cdot} = (IJ)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$. This is the typical condition in linear panel models requiring that all the covariates display variation in both dimensions.

4.1. **Asymptotic Distribution of the Uncorrected Estimator.** We introduce first some further notation. Denote the qth derivatives of the cdf Λ by $\Lambda^{(q)}$, and define $\Lambda_{ij}^{(q)}(y) = \Lambda^{(q)}(x_{ij}'\beta(y) + \alpha_i(y) + \gamma_j(y))$ and $\Lambda_{ij,k}^{(q)}(y) = \Lambda^{(q)}(\mathbb{x}_{ij,k}'\beta(y) + \alpha_i(y) + \gamma_j(y))$ with $\mathbb{x}_{ij,k} := (t_{ij}^k, z_{ij}')'$ and $q = 1, 2, \ldots$. For $\ell \in \{1, \ldots, d_x\}$ define the following projections of the ℓ'th covariate x_{ij}^ℓ,

$$\left(\alpha_x^\ell(y), \gamma_x^\ell(y)\right) \in \arg\min_{(a,c)\in\mathbb{R}^{I+J}} \left[\sum_{(i,j)\in\mathcal{D}} \Lambda_{ij}^{(1)}(y)\,\left(x_{ij}^\ell - a_i - c_j\right)^2\right], \tag{5}$$

and let $\alpha_{x,i}(y)$ and $\gamma_{x,j}(y)$ be the d_x-vectors with components $\alpha_{x,i}^\ell(y)$ and $\gamma_{x,j}^\ell(y)$, where $\alpha_{x,i}^\ell(y)$ is the ith component of $\alpha_x^\ell(y)$ and $\gamma_{x,j}^\ell(y)$ is the jth component of $\gamma_x^\ell(y)$. Also define $\widetilde{x}_{ij}(y) = x_{ij} - \alpha_{x,i}(y) - \gamma_{x,j}(y)$ and $\widetilde{\mathbb{x}}_{ij,k}(y) = \mathbb{x}_{ij,k} - \alpha_{x,i}(y) - \gamma_{x,j}(y)$. Notice that $\widetilde{\mathbb{x}}_{ij,k}(y)$ is defined using projections of x_{ij} instead of $\mathbb{x}_{ij,k}$. Also, while the locations of $\alpha_{x,i}(y)$ and $\gamma_{x,j}(y)$ are not identified, $\widetilde{x}_{ij}(y)$ and $\widetilde{\mathbb{x}}_{ij,k}(y)$ are uniquely defined. Analogous

to the projection of x_k above, we define $\Psi_{ij,k}(y) = \alpha_i^{\Psi}(y) + \gamma_j^{\Psi}(y)$, where

$$\left(\alpha^{\Psi}(y), \gamma^{\Psi}(y)\right) \in \arg\min_{(a,c)\in\mathbb{R}^{I+J}} \left[\sum_{(i,j)\in\mathcal{D}} \Lambda_{ij}^{(1)}(y) \left(\frac{\Lambda_{ij,k}^{(1)}(y)}{\Lambda_{ij}^{(1)}(y)} - a_i - c_j \right)^2 \right]. \tag{6}$$

For example, if $\mathbb{x}_{ij,k} = x_{ij}$, then $\Psi_{ij,k}(y) = 1$. Furthermore, we define[7]

$$W(y) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij}^{(1)}(y)\, \widetilde{x}_{ij}(y)\, \widetilde{x}_{ij}(y)', \qquad \partial_\beta F_k(y) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij,k}^{(1)}(y)\, \widetilde{\mathbb{x}}_{ij,k}(y)',$$

and

$$\begin{aligned}
B^{(\beta)}(y) &= -\frac{1}{2} W^{-1}(y) \left[\frac{1}{I} \sum_{i=1}^{I} \frac{\sum_{j\in\mathcal{D}_i} \Lambda_{ij}^{(2)}(y)\, \widetilde{x}_{ij}(y)}{\sum_{j\in\mathcal{D}_i} \Lambda_{ij}^{(1)}(y)} \right], \\
D^{(\beta)}(y) &= -\frac{1}{2} W^{-1}(y) \left[\frac{1}{J} \sum_{j=1}^{J} \frac{\sum_{i\in\mathcal{D}_j} \Lambda_{ij}^{(2)}(y)\, \widetilde{x}_{ij}(y)}{\sum_{i\in\mathcal{D}_j} \Lambda_{ij}^{(1)}(y)} \right], \\
B_k^{(\Lambda)}(y) &= \frac{1}{2\,I} \sum_{i=1}^{I} \frac{\sum_{j\in\mathcal{D}_i} \left[\Lambda_{ij,k}^{(2)}(y) - \Lambda_{ij}^{(2)}(y) \Psi_{ij,k}(y) \right]}{\sum_{j\in\mathcal{D}_i} \Lambda_{ij}^{(1)}(y)}, \\
D_k^{(\Lambda)}(y) &= \frac{1}{2\,J} \sum_{j=1}^{J} \frac{\sum_{i\in\mathcal{D}_j} \left[\Lambda_{ij,k}^{(2)}(y) - \Lambda_{ij}^{(2)}(y) \Psi_{ij,k}(y) \right]}{\sum_{i\in\mathcal{D}_j} \Lambda_{ij}^{(1)}(y)},
\end{aligned}$$

where $\mathcal{D}_i := \{(i',j') \in \mathcal{D} : i' = i\}$ and $\mathcal{D}_j := \{(i',j') \in \mathcal{D} : j' = j\}$ are the subsets of observational units that contain the index i and j, respectively. In the previous expressions, $\partial_\beta F_k(y)$ is a $1 \times d_x$ vector for each $k \in \mathcal{K}$ that we stack in the $|\mathcal{K}| \times d_x$ matrix $\partial_\beta F(y) = [\partial_\beta F_k(y) \,:\, k \in \mathcal{K}]$. Similarly, $F_k(y)$, $B_k^{(\Lambda)}(y)$, $D_k^{(\Lambda)}(y)$ and $\Psi_{ij,k}(y)$ are scalars for each $k \in \mathcal{K}$, that we stack in the $|\mathcal{K}| \times 1$ vectors $F(y) = [F_k(y) \,:\, k \in \mathcal{K}]$, $B^{(\Lambda)}(y) = [B_k^{(\Lambda)}(y) \,:\, k \in \mathcal{K}]$, $D^{(\Lambda)}(y) = [D_k^{(\Lambda)}(y) \,:\, k \in \mathcal{K}]$, $\Psi_{ij}(y) = [\Psi_{ij,k}(y) \,:\, k \in \mathcal{K}]$.

Let $\ell^\infty(\mathcal{Y})$ be the space of real-valued bounded functions on $\mathcal{Y}$ equipped with the sup-norm $\|\cdot\|_{\mathcal{Y}}$, and $\rightsquigarrow$ denote weak convergence (in distribution). We establish a functional central limit theorem for the fixed effects estimators of $y \mapsto \beta(y)$ and $y \mapsto F(y)$ in $\mathcal{Y}$. All stochastic statements are conditional on $\{(x_{ij}, v_i, w_j) : (i,j) \in \mathcal{D}\}$.

[7] The FOC of problem (5) imply that $\sum_{(i,j)\in\mathcal{D}} \Lambda_{ij,k}^{(1)}(y)\, \widetilde{x}_{ij}(y)' = 0$, and we can therefore equivalently write $\partial_\beta F_k(y) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij,k}^{(1)}(y) \left[\widetilde{\mathbb{x}}_{ij,k}(y) - \widetilde{x}_{ij}(y)\right]' = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij,k}^{(1)}(y) \left[\mathbb{x}_{ij,k}(y) - x_{ij}(y)\right]'$.

Theorem 1 (FCLT for Fixed Effects DR Estimators)**.** *Let Assumption 1 hold. For all $y_1, y_2 \in \mathcal{Y}$ with $y_1 \geq y_2$ we assume the existence of*

$$\overline{V}(y_1, y_2) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij}(y_1)\,[1 - \Lambda_{ij}(y_2)]\ \widetilde{x}_{ij}(y_1)\,\widetilde{x}_{ij}(y_2)',$$

$$\overline{\Omega}(y_1, y_2) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda_{ij}(y_1)\,[1 - \Lambda_{ij}(y_2)]\ \Xi_{ij}(y_1)\Xi_{ij}(y_2)',$$

where $\Xi_{ij}(y) = \Psi_{ij}(y) + \partial_\beta F(y) W^{-1}(y)\,\widetilde{x}_{ij}(y)$. Let $\overline{V}(y_2, y_1) := \overline{V}(y_1, y_2)'$, $\overline{\Omega}(y_2, y_1) := \overline{\Omega}(y_1, y_2)'$, and $\overline{W}(y_1) := \overline{V}(y_1, y_1)$. Then, in the metric space $\ell^\infty(\mathcal{Y})^{d_x}$,

$$\sqrt{n}\left[\widehat{\beta}(y) - \beta(y) - \frac{I}{n} B^{(\beta)}(y) - \frac{J}{n} D^{(\beta)}(y)\right] \rightsquigarrow Z^{(\beta)}(y),$$

and, in the metric space $\ell^\infty(\mathcal{Y})^{|\mathcal{K}|}$,

$$\sqrt{n}\left\{\widehat{F}(y) - F(y) - \frac{I}{n}\underbrace{\left[B^{(\Lambda)}(y) + (\partial_\beta F(y)) B^{(\beta)}(y)\right]}_{B^{(F)}(y)} - \frac{J}{n}\underbrace{\left[D^{(\Lambda)}(y) + (\partial_\beta F(y)) D^{(\beta)}(y)\right]}_{D^{(F)}(y)}\right\} \rightsquigarrow Z^{(F)}(y),$$

as stochastic processes indexed by $y \in \mathcal{Y}$, where $y \mapsto Z^{(\beta)}(y)$ and $y \mapsto Z^{(F)}(y)$ are tight zero-mean Gaussian processes with covariance functions $(y_1, y_2) \mapsto \overline{W}^{-1}(y_1)\,\overline{V}(y_1, y_2)\,\overline{W}^{-1}(y_2)$ and $(y_1, y_2) \mapsto \overline{\Omega}(y_1, y_2)$, respectively.

Assumption 1(vi) guarantees the invertibility of $W(y)$ and $\overline{W}(y)$. Notice that $\overline{W}(y)$ is equal to the limit of $W(y)$ because $\Lambda^{(1)}_{ij}(y) = \Lambda_{ij}(y)\,[1 - \Lambda_{ij}(y)]$ by the properties of the logistic distribution. This information equality follows by the correct specification condition in Assumption 1(ii). By Assumption 1(v), we could have used $\sqrt{IJ}$ instead of $\sqrt{n}$, $1/J$ instead of I/n, and $1/I$ instead of J/n. We prefer the expressions in the theorem, because they might provide a more accurate finite-sample approximation.

Remark 5 (Comparison with binary response models)**.** Fernandez-Val and Weidner (2016) derived the properties of fixed effects estimators of coefficients and APEs in single binary regressions with two-way effects. We generalize their results to multiple binary regressions where our quantities of interest are function-valued, for example $y \mapsto \beta(y)$, $y \in \mathcal{Y}$, and therefore the corresponding estimators are stochastic processes. ■

Remark 6 (Case $\mathbb{x}_{ij,k} = x_{ij}$)**.** When $\mathbb{x}_{ij,k} = x_{ij}$ the asymptotic bias of $\widehat{F}_k$ vanishes, because $B^{(\Lambda)}_k(y) = D^{(\Lambda)}_k(y) = 0$, and $\partial_\beta F_k(y) = 0$ (see footnote 7). In fact, in that case $\widehat{F}_k$ is equal to the empirical distribution function, namely

$$\widehat{F}_k(y) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda(x'_{ij}\widehat{\beta}(y) + \widehat{\alpha}_i(y) + \widehat{\gamma}_j(y)) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} 1\{y_{ij} \leq y\},$$

by the first order conditions of the fixed effects logit DR estimator with respect to the fixed effect parameters. This property provides another appealing feature to choose the logistic distribution. ■

4.2. **Bias Corrections.** Theorem 1 shows that the fixed effects DR estimator has asymptotic bias of the same order as the asymptotic standard deviation under the approximation that we consider. The finite-sample implications are that this estimator can have substantial bias and that confidence regions constructed around it can have severe undercoverage. We deal with these problems by removing the first order bias of the estimator.

We estimate the bias components using the plug-in rule. Define $\widehat{\Lambda}_{ij}^{(q)}(y) = \Lambda^{(q)}(x_{ij}'\widehat{\beta}(y) + \widehat{\alpha}_i(y) + \widehat{\gamma}_j(y))$ and $\widehat{\Lambda}_{ij,k}^{(q)}(y) = \widehat{\Lambda}^{(q)}(\mathbb{x}_{ij,k}'\widehat{\beta}(y) + \widehat{\alpha}_i(y) + \widehat{\gamma}_j(y))$. Replacing $\Lambda_{ij}^{(1)}(y)$ and $\Lambda_{ij,k}^{(1)}(y)$ by $\widehat{\Lambda}_{ij}^{(1)}(y)$ and $\widehat{\Lambda}_{ij,k}^{(1)}(y)$ in the definitions of $\alpha_x^{\ell}(y)$, $\gamma_x^{\ell}(y)$, $\alpha^{\Psi}(y)$, and $\gamma^{\Psi}(y)$ yields the corresponding estimators. We plug-in these estimators to obtain $\widehat{x}_{ij}(y) = x_{ij} - \widehat{\alpha}_{x,i}(y) - \widehat{\gamma}_{x,j}(y)$, $\widehat{\mathbb{x}}_{ij,k}(y) = \mathbb{x}_{ij,k} - \widehat{\alpha}_{x,i}(y) - \widehat{\gamma}_{x,j}(y)$, and $\widehat{\Psi}_{ij,k}(y) = \widehat{\alpha}_i^{\Psi}(y) + \widehat{\gamma}_j^{\Psi}(y)$. Then we construct

$$\widehat{W}(y) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \widehat{\Lambda}_{ij}^{(1)}(y)\,\widehat{x}_{ij}(y)\,\widehat{x}_{ij}'(y), \qquad \partial_\beta \widehat{F}_k(y) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \widehat{\Lambda}_{ij,k}^{(1)}(y)\,\widehat{\mathbb{x}}_{ij,k}(y)',$$

and

$$\widehat{B}^{(\beta)}(y) = -\frac{1}{2}\widehat{W}^{-1}(y)\left[\frac{1}{I}\sum_{i=1}^{I}\frac{\sum_{j\in\mathcal{D}_i}\widehat{\Lambda}_{ij}^{(2)}(y)\,\widehat{x}_{ij}(y)}{\sum_{j\in\mathcal{D}_i}\widehat{\Lambda}_{ij}^{(1)}(y)}\right],$$

$$\widehat{D}^{(\beta)}(y) = -\frac{1}{2}\widehat{W}^{-1}(y)\left[\frac{1}{J}\sum_{j=1}^{J}\frac{\sum_{i\in\mathcal{D}_j}\widehat{\Lambda}_{ij}^{(2)}(y)\,\widehat{x}_{ij}(y)}{\sum_{i\in\mathcal{D}_j}\widehat{\Lambda}_{ij}^{(1)}(y)}\right],$$

$$\widehat{B}_k^{(\Lambda)}(y) = \frac{1}{2\,I}\sum_{i=1}^{I}\frac{\sum_{j\in\mathcal{D}_i}\left[\widehat{\Lambda}_{ij,k}^{(2)}(y) - \widehat{\Lambda}_{ij}^{(2)}(y)\widehat{\Psi}_{ij,k}(y)\right]}{\sum_{j\in\mathcal{D}_i}\widehat{\Lambda}_{ij}^{(1)}(y)},$$

$$\widehat{D}_k^{(\Lambda)}(y) = \frac{1}{2\,J}\sum_{j=1}^{J}\frac{\sum_{i\in\mathcal{D}_j}\left[\widehat{\Lambda}_{ij,k}^{(2)}(y) - \widehat{\Lambda}_{ij}^{(2)}(y)\widehat{\Psi}_{ij,k}(y)\right]}{\sum_{i\in\mathcal{D}_j}\widehat{\Lambda}_{ij}^{(1)}(y)}.$$

We also define the $|\mathcal{K}| \times d_x$ matrix $\partial_\beta \widehat{F}(y) = [(\partial_\beta \widehat{F}_k(y)) : k \in \mathcal{K}]$, and the $|\mathcal{K}| \times 1$ vectors $\widehat{B}^{(F)}(y) = [\widehat{B}_k^{(F)}(y) : k \in \mathcal{K}]$, $\widehat{D}^{(F)}(y) = [\widehat{D}_k^{(F)}(y) : k \in \mathcal{K}]$, $\widehat{\Psi}_{ij}(y) = [\widehat{\Psi}_{ij,k}(y) : k \in \mathcal{K}]$. Finally, we also construct the estimator of the asymptotic variance of $\widehat{F}(y)$

$$\widehat{\Omega}(y) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \widehat{\Lambda}_{ij}^{(1)}(y)\,\widehat{\Xi}(y)\,\widehat{\Xi}(y)'.$$

where $\widehat{\Xi}(y) = \widehat{\Psi}_{ij}(y) + (\partial_\beta \widehat{F}(y))\widehat{W}^{-1}(y)\,\widehat{x}_{ij}(y)$.

The following theorem shows that the estimators of the asymptotic bias and variance are consistent, uniformly in $y \in \mathcal{Y}$. For a matrix A, we denote by $\|A\|$ the Frobenius norm of A, i.e. $\|A\| = \text{trace}(AA')$.

Theorem 2 (Uniform Consistency of Estimators of Bias and Variance Components)**.** *Let Assumption 1 hold. Then,*

$$\sup_{y\in\mathcal{Y}}\left\|\widehat{W}(y)-\overline{W}(y)\right\| = o_P(1), \qquad \sup_{y\in\mathcal{Y}}\left\|\partial_\beta\widehat{F}(y)-\partial_\beta F(y)\right\| = o_P(1),$$

$$\sup_{y\in\mathcal{Y}}\left\|\widehat{B}^{(\beta)}(y)-B^{(\beta)}(y)\right\| = o_P(1), \qquad \sup_{y\in\mathcal{Y}}\left\|\widehat{D}^{(\beta)}(y)-D^{(\beta)}(y)\right\| = o_P(1),$$

$$\sup_{y\in\mathcal{Y}}\left\|\widehat{B}^{(\Lambda)}(y)-B^{(\Lambda)}(y)\right\| = o_P(1), \qquad \sup_{y\in\mathcal{Y}}\left\|\widehat{D}^{(\Lambda)}(y)-D^{(\Lambda)}(y)\right\| = o_P(1),$$

$$\sup_{y\in\mathcal{Y}}\left\|\widehat{\Omega}(y)-\overline{\Omega}(y)\right\| = o_P(1).$$

Bias corrected estimators of $\beta(y)$ and $F(y)$ are formed as

$$\widetilde{\beta}(y) = \widehat{\beta}(y) - \frac{I}{n}\widehat{B}^{(\beta)}(y) - \frac{J}{n}\widehat{D}^{(\beta)}(y), \tag{7}$$

and

$$\widetilde{F}(y) = \widehat{F}(y) - \frac{I}{n}\underbrace{\left[\widehat{B}^{(\Lambda)}(y) + (\partial_\beta\widehat{F}(y))\widehat{B}^{(\beta)}(y)\right]}_{\widehat{B}^{(F)}(y)} - \frac{J}{n}\underbrace{\left[\widehat{D}^{(\Lambda)}(y) + (\partial_\beta\widehat{F}(y))\widehat{D}^{(\beta)}(y)\right]}_{\widehat{D}^{(F)}(y)}.$$

Alternatively, we could define the bias corrected version of $\widehat{F}(y)$ as

$$\widetilde{F}_k^*(y) = \left[\frac{1}{n}\sum_{(i,j)\in\mathcal{D}}\Lambda\left(\mathbb{x}_{ij,k}'\,\widetilde{\beta}(y) + \widetilde{\alpha}_i(y) + \widetilde{\gamma}_j(y)\right)\right] - \frac{I}{n}\widehat{B}_k^{(\Lambda)}(y) - \frac{J}{n}\widehat{D}_k^{(\Lambda)}(y),$$

where $\widetilde{\xi}(y) := (\widetilde{\alpha}_1(y), \ldots, \widetilde{\alpha}_I(y), \widetilde{\gamma}_1(y), \ldots, \widetilde{\gamma}_J(y))$ is a solution to

$$\max_{\xi\in\mathbb{R}^{I+J}}\sum_{(i,j)\in\mathcal{D}}(1\{y_{ij}\le y\}\log\Lambda(x_{ij}'\widetilde{\beta}(y)+\alpha_i+\gamma_j)+1\{y_{ij}>y\}\log[1-\Lambda(x_{ij}'\widetilde{\beta}(y)+\alpha_i+\gamma_j)]).$$

It can be shown that $\sup_{y\in\mathcal{Y}}\sqrt{n}\left|\widetilde{F}_k^*(y)-\widetilde{F}_k(y)\right| = o_P(1)$, that is, the difference between those alternative bias corrected estimators is asymptotically negligible. There is no obvious reason to prefer one over the other, and we present result for $\widetilde{F}_k$ in the following, which equivalently hold for $\widetilde{F}_k^*$.[8]

Combining Theorem 1 and 2 we obtain the following functional central limit theorem for the bias corrected estimators.

[8]We use the estimator $\widetilde{F}_k^*$ in the numerical examples for computational convenience as the bias correction involves estimating less terms.

Corollary 1 (FCLT for Bias Corrected Fixed Effects DR Estimators)**.** *Let Assumption 1 hold. Then, in the metric space $\ell^\infty(\mathcal{Y})^{d_x}$,*

$$\sqrt{n}\left[\widetilde{\beta}(y) - \beta(y)\right] \rightsquigarrow Z^{(\beta)}(y),$$

and, in the metric space $\ell^\infty(\mathcal{Y})^{|\mathcal{K}|}$,

$$\sqrt{n}\left[\widetilde{F}(y) - F(y)\right] \rightsquigarrow Z^{(F)}(y),$$

as stochastic processes indexed by $y \in \mathcal{Y}$, where $Z^{(\beta)}(y)$ and $Z^{(F)}(y)$ are the same Gaussian processes that appear in Theorem 1.

4.3. **Uniform Confidence Bands and Bootstrap.** We show how to construct pointwise and uniform confidence bands for $y \mapsto \beta(y)$ and $y \mapsto F(y)$ on $\mathcal{Y}$ using Corollary 1. The uniform bands for F can be used as inputs in Lemmas 1 and 2 to construct uniform bands for the QFs $\tau \mapsto Q_k(\tau) = F_k^{\leftarrow}(\tau)$, $k \in \mathcal{K}$, and the QEF $\tau \mapsto \Delta(\tau)$ on $\mathcal{T}$.

Let $\mathcal{B} \subseteq \{1, \dots, d_x\}$ be the set of indexes for the coefficients of interest. For given $y \in \mathcal{Y}$, $\ell \in \mathcal{B}$, $k \in \mathcal{K}$, and $p \in (0,1)$, a pointwise p-confidence interval for $\beta_\ell(y)$, the ℓ'th component of $\beta(y)$, is

$$[\widetilde{\beta}_\ell(y) \pm \Phi^{-1}(1-p/2)\widehat{\sigma}_{\beta_\ell}(y)], \tag{8}$$

and a pointwise p-confidence intervals for $F_k(y)$ is

$$[\widetilde{F}_k(y) \pm \Phi^{-1}(1-p/2)\widehat{\sigma}_{F_k}(y)],$$

where Φ denotes the cdf of the standard normal distribution, $\widehat{\sigma}_{\beta_\ell}(y)$ is the standard error of $\widetilde{\beta}_\ell(y)$ given in (12), and $\widehat{\sigma}_{F_k}(y)$ is the standard error of $\widetilde{F}_k(y)$ given in (13). These intervals have coverage p in large samples by Corollary 1 and Theorem 2.

We construct joint uniform bands for the coefficients and distributions using Kolmogorov-Smirnov type critical values, instead of quantiles from the normal distribution. A uniform p-confidence band joint for the vector of functions $\{\beta_\ell(y) : \ell \in \mathcal{B}, y \in \mathcal{Y}\}$ is

$$I_\beta = \{[\widetilde{\beta}_\ell(y) \pm t^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p)\widehat{\sigma}_{\beta_\ell}(y)] : \ell \in \mathcal{B}, y \in \mathcal{Y}\},$$

where $t^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p)$ is the p-quantile of the maximal t-statistic

$$t^{(\beta)}_{\mathcal{B},\mathcal{Y}} = \sup_{y \in \mathcal{Y},\, \ell \in \mathcal{B}} \frac{\left|Z^{(\beta)}_\ell(y)\right|}{\sigma^{(\beta)}_\ell(y)}, \tag{9}$$

where $\sigma^{(\beta)}_\ell(y) = [\overline{W}(y)^{-1}]^{1/2}_{\ell,\ell}$, the square root of the (ℓ,ℓ) element of the matrix $\overline{W}(y)^{-1}$. Similarly, a uniform p-confidence band joint for the set of distribution functions $\{F_k(y) : k \in \mathcal{K}, y \in \mathcal{Y}\}$ is

$$I_F = \{[\widetilde{F}_k(y) \pm t^{(F)}_{\mathcal{K},\mathcal{Y}}(p)\widehat{\sigma}_{F_k}(y)] : k \in \mathcal{K}, y \in \mathcal{Y}\}, \tag{10}$$

where $t^{(F)}_{\mathcal{K},\mathcal{Y}}(p)$ is the p-quantile of the maximal t-statistic

$$t^{(F)}_{\mathcal{K},\mathcal{Y}} = \sup_{y\in\mathcal{Y},\, k\in\mathcal{K}} \frac{\left|Z^{(F)}_k(y)\right|}{\sigma^{(F)}_k(y)}, \tag{11}$$

where $\sigma^{(F)}_k(y) = [\overline{\Omega}(y)]^{1/2}_{k,k}$, the square root of the (k,k) element of the matrix $\overline{\Omega}(y,y)$. The previous confidence bands also have coverage p in large samples by Corollary 1 and Theorem 2.

The maximal t-statistics used to construct the bands I_β and I_F are not pivotal, but their distributions can be approximated by simulation after replacing the variance functions of the limit processes by uniformly consistent estimators. In practice, however, we find more convenient to use resampling methods. We consider a multiplier bootstrap scheme that resamples the efficient scores or influence functions of the fixed effects estimators $\widehat{\beta}(y)$ and $\widehat{F}(y)$. This scheme is computationally convenient because it does not need to solve the high dimensional nonlinear fixed effects conditional maximum likelihood program (3) or making any bias correction in each bootstrap replication. In these constructions we rely on the uncorrected fixed effects estimators instead of the bias corrected estimators, because they have the same influence functions and the uncorrected estimators are consistent under the asymptotic approximation that we consider.

To describe the standard errors and multiplier bootstrap we need to introduce some notation for the influence functions of $\widehat{\theta}(y)$ and $\widehat{F}(y)$. Let $\theta = (\beta, \alpha_1, \ldots, \alpha_I, \gamma_1, \ldots, \gamma_J)$ be a generic value for the parameter $\theta(y)$, the influence function of $\widehat{\theta}(y)$ is the (d_x+I+J)-vector $\psi^y_{ij}(\theta(y))$, where

$$\psi^y_{ij}(\theta) = H(\theta)^\dagger[\mathbf{1}\{y_{ij}\le y\} - \Lambda(x'_{ij}\beta + \alpha_i + \gamma_j)]w_{ij}, \quad w_{ij} = (x_{ij}, e_{i,I}, e_{j,J}), \quad y\in\mathcal{Y},$$

$e_{i,I}$ is a unit vector of dimension I with a one in the position i, $e_{j,J}$ is defined analogously, $H(\theta)^\dagger$ is the Moore-Penrose pseudo-inverse of $H(\theta)$, and

$$H(\theta) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}(x'_{ij}\beta + \alpha_i + \gamma_j)w_{ij}w'_{ij},$$

is minus the Hessian of the log-likelihood with respect to θ, which does not depend on y in the case of the logistic distribution. The influence function of $\widehat{F}_k(y)$ is $\varphi^y_{ij,k}(\theta(y))$, where

$$\varphi^y_{ij,k}(\theta) = J_k(\theta)'\psi^y_{ij}(\theta),$$

and

$$J_k(\theta) = \frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}(\mathbb{x}'_{ij,k}\beta + \alpha_i + \gamma_j)\mathbb{w}_{ij,k}, \quad \mathbb{w}_{ij,k} = (\mathbb{x}_{ij,k}, e_{i,I}, e_{j,J}).$$

The standard error of $\widetilde{\beta}_\ell(y)$ is constructed as

$$\widehat{\sigma}_{\beta_\ell}(y) = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \psi_{ij}^y(\widehat{\theta}(y))\psi_{ij}^y(\widehat{\theta}(y))'\right]_{\ell,\ell}^{1/2}, \tag{12}$$

the square root of the (ℓ,ℓ) element of the sandwich matrix $n^{-2}\sum_{(i,j)\in\mathcal{D}} \psi_{ij}^y(\widehat{\theta}(y))\psi_{ij}^y(\widehat{\theta}(y))'$. Similarly, the standard error of $\widetilde{F}_k(y)$ is constructed as

$$\widehat{\sigma}_{F_k}(y) = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \varphi_{ij,k}^y(\widehat{\theta}(y))^2\right]^{1/2}. \tag{13}$$

The following algorithm describes a multiplier bootstrap scheme to obtain the critical values for a set of parameters indexed by $\ell \in \mathcal{B} \subseteq \{1,\ldots,d_x\}$ and a set of distributions indexed by $k \in \mathcal{K} \subseteq \{0,1\}$. This scheme is based on perturbing the first order conditions of the fixed effects estimators with random multipliers independent from the data.

Algorithm 1 (Multiplier Bootstrap)**.** *(1) Let $\bar{\mathcal{Y}}$ to be a finite cover of $\mathcal{Y}$ such that $\bar{\mathcal{Y}} = \mathcal{Y}$ if $\mathcal{Y}$ is finite or $\sqrt{n}/|\bar{\mathcal{Y}}| \to 0$ otherwise, e.g., $\bar{\mathcal{Y}}$ is a fine mesh with $\sqrt{n}\log\log n$ points. (2) Draw the bootstrap multipliers $\{\omega_{ij}^m : (i,j)\in\mathcal{D}\}$ i.i.d. from the standard normal distribution, independently from the data, i.e., $\omega_{ij}^m = \tilde{\omega}_{ij}^m - \sum_{(i,j)\in\mathcal{D}} \tilde{\omega}_{ij}^m/n$, $\tilde{\omega}_{ij}^m \sim$ i.i.d. $\mathcal{N}(0,1)$. Here we have normalized the multipliers to have zero mean as a finite-sample adjustment. (3) For each $y \in \bar{\mathcal{Y}}$, obtain the bootstrap draws of the estimator of parameter $\widehat{\theta}^m(y) = \widehat{\theta}(y) + n^{-1}\sum_{(i,j)\in\mathcal{D}} \omega_{ij}^m \psi_{ij}^y(\widehat{\theta}(y))$, and of the estimator of the distribution functions $\widehat{F}_k^m(y) = \widehat{F}_k(y) + n^{-1}\sum_{(i,j)\in\mathcal{D}} \omega_{ij}^m \varphi_{ij,k}^y(\widehat{\theta}(y))$, $k \in \mathcal{K}$. (4) Construct the bootstrap draw of the maximal t-statistic for the parameters, $t_{\mathcal{B},\bar{\mathcal{Y}}}^{(\beta),m} = \max_{y\in\bar{\mathcal{Y}},\ell\in\mathcal{B}} |\widehat{\beta}_\ell^m(y) - \widehat{\beta}_\ell(y)|/\widehat{\sigma}_{\beta_\ell}(y)$, where $\widehat{\sigma}_{\beta_\ell}(y)$ is defined in (12), and $\psi_{ij,\ell}^y(\theta)$ is the component of $\psi_{ij}^y(\theta)$ corresponding to β_ℓ. Similarly, construct the bootstrap draw of the maximal t-statistic for the distributions, $t_{\mathcal{K},\bar{\mathcal{Y}}}^{(F),m} = \max_{y\in\bar{\mathcal{Y}},k\in\mathcal{K}} |\widehat{F}_k^m(y) - \widehat{F}_k(y)|/\widehat{\sigma}_{F_k}(y)$, where $\widehat{\sigma}_{F_k}(y)$ is defined in (13). (5) Repeat steps (1)–(3) M times and index the bootstrap draws by $m \in \{1,\ldots,M\}$. In the numerical examples we set $M = 500$. (6) Obtain the bootstrap estimators of the critical values as*

$$\begin{aligned}
\widehat{t}_{\mathcal{B},\mathcal{Y}}^{(\beta)}(p) &= p-\textit{quantile of } \{t_{\mathcal{B},\bar{\mathcal{Y}}}^{(\beta),m} : 1 \le m \le M\},\\
\widehat{t}_{\mathcal{K},\mathcal{Y}}^{(F)}(p) &= p-\textit{quantile of } \{t_{\mathcal{K},\bar{\mathcal{Y}}}^{(F),m} : 1 \le m \le M\}.
\end{aligned}$$

The next result shows that the multiplier bootstrap provides consistent estimators of the critical values of the inferential statistics. The proof follows from Chernozhukov, Chetverikov and Kato (2016).

Theorem 3 (Consistency of Multiplier Bootstrap Inference)**.** *Let Assumption 1 hold. Then, conditional on the data* $\{(y_{ij}, x_{ij}) : (i,j) \in \mathcal{D}\}$*, as* $n \to \infty$ *and* $M \to \infty$

$$\widehat{t}^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p) \to_{\mathrm{P}} t^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p) \quad \text{and} \quad \widehat{t}^{(F)}_{\mathcal{K},\mathcal{Y}}(p) \to_{\mathrm{P}} t^{(F)}_{\mathcal{K},\mathcal{Y}}(p),$$

where $t^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p)$ *and* $t^{(F)}_{\mathcal{K},\mathcal{Y}}(p)$ *are defined in* (9) *and* (11)*, respectively.*

4.4. **Pairwise Clustering Dependence or Reciprocity.** The conditional independence of Assumption 1(i) can be relaxed to allow for some forms of conditional weak dependence. A form of dependence that is particularly relevant for network data is *pairwise clustering* or *reciprocity* where the observational units with symmetric indexes (i,j) and (j,i) might be dependent due to unobservable factors not accounted by unobserved effects.[9] In the trade application, for example, these factors may include distributional channels or multinational firms operating in both countries.

The presence of reciprocity does not change the bias of the fixed effects estimators, but affects the standard errors and the implementation of the multiplier bootstrap. The standard error of $\widetilde{\beta}_\ell(y)$ becomes

$$\widehat{\sigma}_{\beta_\ell}(y) = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \left\{\psi^y_{ij}(\widehat{\theta}(y)) + \psi^y_{ji}(\widehat{\theta}(y))\right\} \psi^y_{ij}(\widehat{\theta}(y))'\right]^{1/2}_{\ell,\ell}. \tag{14}$$

Similarly, the standard error of $\widetilde{F}_k(y)$ needs to be adjusted to

$$\widehat{\sigma}_{F_k}(y) = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \left\{\varphi^y_{ij,k}(\widehat{\theta}(y)) + \varphi^y_{ji,k}(\widehat{\theta}(y))\right\} \varphi^y_{ij,k}(\widehat{\theta}(y))\right]^{1/2}. \tag{15}$$

In the previous expressions we assume that if $(i,j) \in \mathcal{D}$ then $(j,i) \in \mathcal{D}$ to simplify the notation. The modified multiplier bootstrap algorithm becomes:

Algorithm 2 (Multiplier Bootstrap with Pairwise Clustering)**.** *(1) Let* $\bar{\mathcal{Y}}$ *to be a finite cover of* $\mathcal{Y}$ *such that* $\bar{\mathcal{Y}} = \mathcal{Y}$ *if* $\mathcal{Y}$ *is finite or* $\sqrt{n}/|\bar{\mathcal{Y}}| \to 0$ *otherwise, e.g.,* $\bar{\mathcal{Y}}$ *is a fine mesh with* $\sqrt{n}\log\log n$ *points. (2) Draw the bootstrap multipliers* $\{\omega^m_{ij} : (i,j) \in \mathcal{D}\}$ *i.i.d. from some distribution with zero mean and unit variance, independently from the data such that* $\omega^m_{ji} = \omega^m_{ij}$*. For example,* $\omega^m_{ij} = \tilde{\omega}^m_{ij} - \sum_{(i,j)\in\mathcal{D}} \tilde{\omega}^m_{ij}/n$*,* $\tilde{\omega}^m_{ij} \sim$ *i.i.d.* $\mathcal{N}(0,1)$*,* $i \le j$*. (3) For each* $y \in \bar{\mathcal{Y}}$*, obtain the bootstrap draws of the estimator of parameter* $\widehat{\theta}^m(y) = \widehat{\theta}(y) + n^{-1}\sum_{(i,j)\in\mathcal{D}} \omega^m_{ij}\psi^y_{ij}(\widehat{\theta}(y))$*, and of the estimator of the distribution functions* $\widehat{F}^m_k(y) = \widehat{F}_k(y) + n^{-1}\sum_{(i,j)\in\mathcal{D}} \omega^m_{ij}\varphi^y_{ij,k}(\widehat{\theta}(y))$*,* $k \in \mathcal{K}$*. (4) Construct the bootstrap draw of the maximal t-statistic for the parameters,* $t^{(\beta),m}_{\mathcal{B},\bar{\mathcal{Y}}} = \max_{y\in\bar{\mathcal{Y}},\ell\in\mathcal{B}} |\widehat{\beta}^m_\ell(y) - \widehat{\beta}_\ell(y)|/\widehat{\sigma}_{\beta_\ell}(y)$*, where* $\widehat{\sigma}_{\beta_\ell}(y)$ *is defined in* (14)*, and and* $\psi^y_{ij,\ell}(\theta)$ *is the component of* $\psi^y_{ij}(\theta)$ *corresponding to*

[9] Cameron and Miller (2014) consider other patterns of dependence in linear models for dyadic data.

β_ℓ. Similarly, construct the bootstrap draw of the maximal t-statistic for the distributions, $t^{(F),m}_{\mathcal{K},\bar{\mathcal{Y}}} = \max_{y\in\bar{\mathcal{Y}},k\in\mathcal{K}} |\widehat{F}^m_k(y) - \widehat{F}_k(y)|/\widehat{\sigma}_{F_k}(y)$, where $\widehat{\sigma}_{F_k}(y)$ is defined in (15). (5) Repeat steps (1)–(3) M times and index the bootstrap draws by $m \in \{1,\dots,M\}$. In the numerical examples we set $M = 500$. (6) Obtain the bootstrap estimators of the critical values as

$$\begin{aligned} \widehat{t}^{(\beta)}_{\mathcal{B},\mathcal{Y}}(p) &= p - \textit{quantile of } \{t^{(\beta),m}_{\mathcal{B},\bar{\mathcal{Y}}} : 1 \le m \le M\}, \\ \widehat{t}^{(F)}_{\mathcal{K},\mathcal{Y}}(p) &= p - \textit{quantile of } \{t^{(F),m}_{\mathcal{K},\bar{\mathcal{Y}}} : 1 \le m \le M\}. \end{aligned}$$

The clustered multiplier bootstrap preserves the dependence in the symmetric pairs (i,j) and (j,i) by assigning the same multiplier to each of these pairs.

4.5. **Average Effect.** A bias corrected estimator of the average effect can be formed as

$$\widetilde{\Delta} = \widetilde{\mu}_1 - \widetilde{\mu}_0, \tag{16}$$

where

$$\widetilde{\mu}_k = \int_{\mathcal{Y}} [1(y \ge 0) - \widetilde{F}_k(y)]\nu(dy), \quad k \in \{0,1\}.$$

The following central limit theorem for the bias corrected estimator of the average effect is a corollary of Theorem 1 and 2 together with the functional delta method.

Corollary 2 (CLT for Bias Corrected Fixed Effects Estimators of Average Effect)**.** *Let Assumption 1 hold and $\int_{\mathcal{Y}} dF_k(y) = 1$, $k \in \{0,1\}$. Then,*

$$\sqrt{n}\left(\widetilde{\Delta} - \Delta\right) \to_d -\int_{\mathcal{Y}} \left[Z^{(F)}_1(y) - Z^{(F)}_0(y)\right] \nu(dy) =: Z^{(\Delta)}, \tag{17}$$

where $Z^{(F)}(y) = [Z^{(F)}_0(y), Z^{(F)}_1(y)]'$ is the same Gaussian process that appears in Theorem 1 with $\mathcal{K} = \{0,1\}$.

Remark 7 (Support of Y)**.** The condition that $\int_{\mathcal{Y}} dF_k(y) = 1$ guarantees that $\mathcal{Y}$ is the support of the potential outcome corresponding to the distribution F_k, so that (2) yields the average potential outcome under F_k. Together with Assumption 1, this condition is satisfied when Y is discrete with finite support $\mathcal{Y}$, or continuous or mixed with bounded support $\mathcal{Y}$ and conditional density bounded away from zero in the interior of $\mathcal{Y}$. This support condition is not required for the estimation of the quantile effects.

We can construct confidence intervals for the average effect using Corollary 2. Let

$$\widehat{\sigma}_\Delta = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \widehat{\varphi}^2_{ij}\right]^{1/2}, \quad \widehat{\varphi}_{ij} = -\int_{\mathcal{Y}} \left[\varphi^y_{ij,1}(\widehat{\theta}(y)) - \varphi^y_{ij,0}(\widehat{\theta}(y))\right] \nu(dy).$$

Then, $\widehat{\sigma}_\Delta$ is an estimator of σ_Δ, the standard deviation of the limit process $Z^{(\Delta)}$ in (17), and

$$I_\Delta = [\widetilde{\Delta} \pm \Phi^{-1}(1-p/2)\widehat{\sigma}_\Delta],$$

is an asymptotic p-confidence interval for Δ. The normal critical value $\Phi^{-1}(1-p/2)$ can be replaced by a multiplier bootstrap critical value $\widehat{t}^{(\Delta)}(p)$ obtained from Algorithm 1 as

$$\widehat{t}^{(\Delta)}(p) = p - \text{quantile of } \{t^{(\Delta),m} : 1 \leq m \leq M\}$$

where $t^{(\Delta),m} = |\widehat{\Delta}^m - \widehat{\Delta}|/\widehat{\sigma}_\Delta$ and $\widehat{\Delta}^m = \widehat{\Delta} + n^{-1}\sum_{(i,j)\in\mathcal{D}} \omega_{ij}^m \widehat{\varphi}_{ij}$.

The standard errors and critical values of the average effects can be adjusted to account for pairwise clustering following the procedure described in Section 4.4. Thus, the pairwise clustering robust standard error is

$$\widehat{\sigma}_\Delta = n^{-1}\left[\sum_{(i,j)\in\mathcal{D}} \{\widehat{\varphi}_{ij} + \widehat{\varphi}_{ji}\}\,\widehat{\varphi}_{ij}\right]^{1/2}.$$

5. Quantile Effects in Gravity Equations for International Trade

We consider an empirical application to gravity equations for bilateral trade between countries. We use data from Helpman, Melitz and Rubinstein (2008), extracted from the Feenstra's World Trade Flows, CIA's World Factbook and Andrew Rose's web site. These data contain information on bilateral trade flows and other trade-related variables for 157 countries in 1986.[10] The data set contains network data where both i and j index countries as senders (exporters) and receivers (importers), and therefore $I = J = 157$. The outcome y_{ij} is the volume of trade in thousands of constant 2000 US dollars from country i to country j, and the covariates x_{ij} include determinants of bilateral trade flows such as the logarithm of the distance in kilometers between country i's capital and country j's capital and indicators for common colonial ties, currency union, regional free trade area (FTA), border, legal system, language, and religion. Following Anderson and van Wincoop (2003), we include unobserved importer and exporter country effects.[11] These effects control for other country specific characteristics that may affect trade such as GDP, tariffs, population, institutions, infrastructures or natural resources. We allow for these characteristics to affect differently the imports and exports of each country, and be arbitrarily related with the observed covariates.

Table 1 reports descriptive statistics of the variables used in the analysis. There are $157 \times 156 = 24,492$ observations corresponding to different pairs of countries. The observations

[10] The original data set includes 158 countries. We exclude Congo because it did not export to any other country in 1986.

[11] See Harrigan (1994) for an earlier empirical international trade application that includes unobserved country effects.

with $i = j$ are missing because we do not observe trade flows from a country to itself. The trade variable in the first row is an indicator for positive volume of trade. There are no trade flows for 55% of the country pairs. The volume of trade variable exhibits severe overdispersion relative to the mean. Since this variable is bounded below at zero, the overdispersion indicates the presence of a very heavy upper tail in the distribution. This feature also makes quantile methods specially well-suited for this application on robustness grounds.[12]

Table 1. Descriptive Statistics

	Mean	Std. Dev.
Trade	0.45	0.50
Trade Volume	84,542	1,082,219
Log Distance	4.18	0.78
Legal	0.37	0.48
Language	0.29	0.45
Religion	0.17	0.25
Border	0.02	0.13
Currency	0.01	0.09
FTA	0.01	0.08
Colony	0.01	0.10
Country Pairs	24,492	

Source: Helpman, Melitz and Rubinstein (08)

The previous literature estimated nonlinear parametric models such as Poisson, Negative Binomial, Tobit and Heckman-selection models to deal with the large number of zeros in the volume of trade (e.g., Eaton and Kortum, 2001, Santos Silva and Tenreyro, 2006, and Helpman, Melitz and Rubinstein, 2008).[13] These models impose strong conditions on the process that generates the zeros and/or on the conditional heteroskedasticity of the volume of trade. The DR model deals with zeros and any other fixed censoring points in a very flexible and natural fashion as it specifies the conditional distribution separately at the mass point. In particular, the model coefficients at zero can be arbitrarily different from the model coefficients at other values of the volume of trade. Moreover, the DR model can also accommodate flexible patterns of conditional heteroskedasticity.

Figure 1 shows estimates and 95% pointwise confidence intervals for the DR coefficients of log distance and common legal system plotted against the quantile indexes of the volume of trade. We report uncorrected and bias corrected fixed effects estimates obtained from (3) and (7), respectively. The confidence intervals are constructed using (8). The x-axis starts at .54, the maximum quantile index corresponding to zero volume of trade. The

[12]In results not reported, we find that estimates of average effects are very sensitive to the trimming of outliers at the top of the distribution.

[13]See Head and Mayer (2014) for a recent survey on gravity equations in international trade.

region of interest $\mathcal{Y}$ corresponds to the interval between zero and the 0.95-quantile of the volume of trade. As predicted by the theory in Section 4, the difference between the uncorrected and bias corrected estimates is the same order of magnitude as the width of the confidence intervals, specially for the coefficient of log distance. We find the largest estimated biases for both coefficients at highest quantiles of the volume of trade, where the indicators $1\{y_{ij} \leq y\}$ take on many ones. The signs of the DR coefficients indicate that increasing distance has a negative effect and having a common legal system has a positive effect on the volume of trade throughout the distribution. Recall that the sign of the effect in terms of volume of trade, y_{ij}, is the opposite to the sign of the DR coefficient.

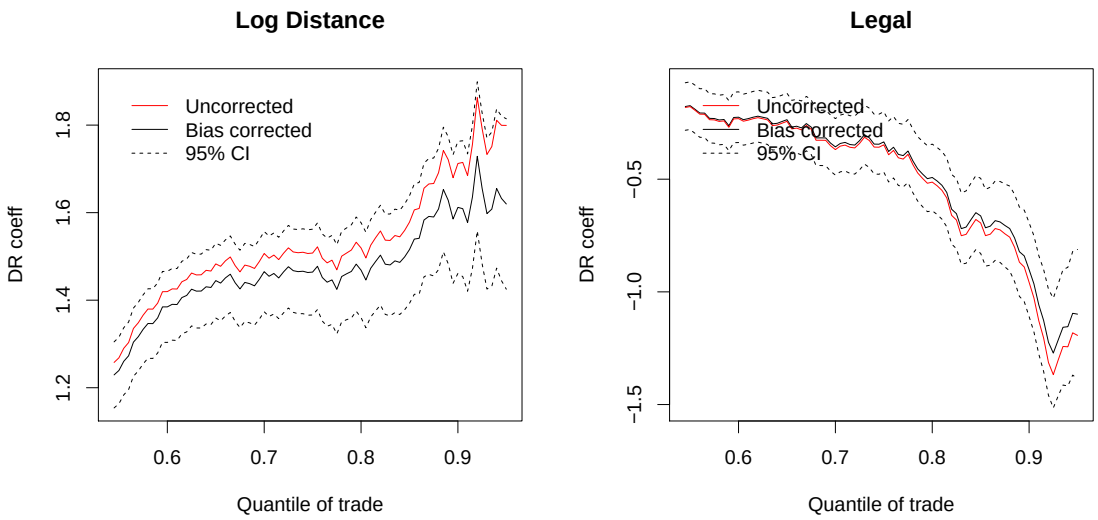

FIGURE 1. Estimates and 95% pointwise confidence intervals for the DR-coefficients of log distance and common legal system.

Figures 2 and 3 show estimates and 95% uniform confidence bands for distribution and quantile functions of the volume of trade at different values of the log of distance and the common legal system. The left panels plot the functions when distance takes the observed levels (dist) and two times the observed values (2*dist), i.e. when we counterfactually double all the distances between the countries. The right panels plot the functions when all the countries have the same legal system (legal=1) and different systems (legal=0). The confidence bands for the distribution are obtained by Algorithm 1 with 500 bootstrap replications and standard normal multipliers, and a grid of values $\bar{\mathcal{Y}}$ that includes the sample quantiles of the volume of trade with indexes $\{.54, .55, \ldots, .95\}$. The bands are joint for the two functions displayed in each panel. The confidence bands for the quantile functions are obtained by inverting and rotating the bands for the corresponding distribution functions using Lemma 1.

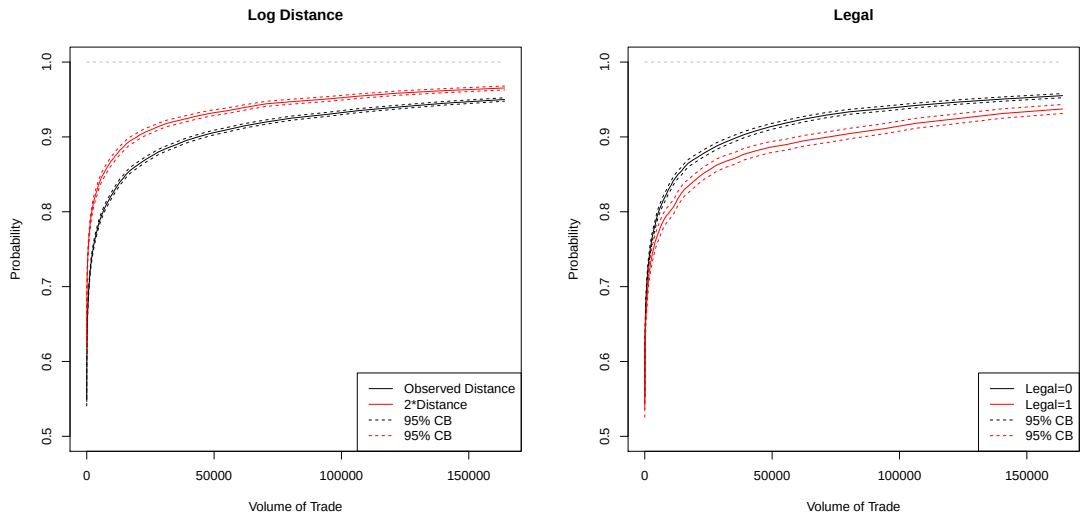

FIGURE 2. Estimates and 95% uniform confidence bands for distribution functions of the volume of trade.

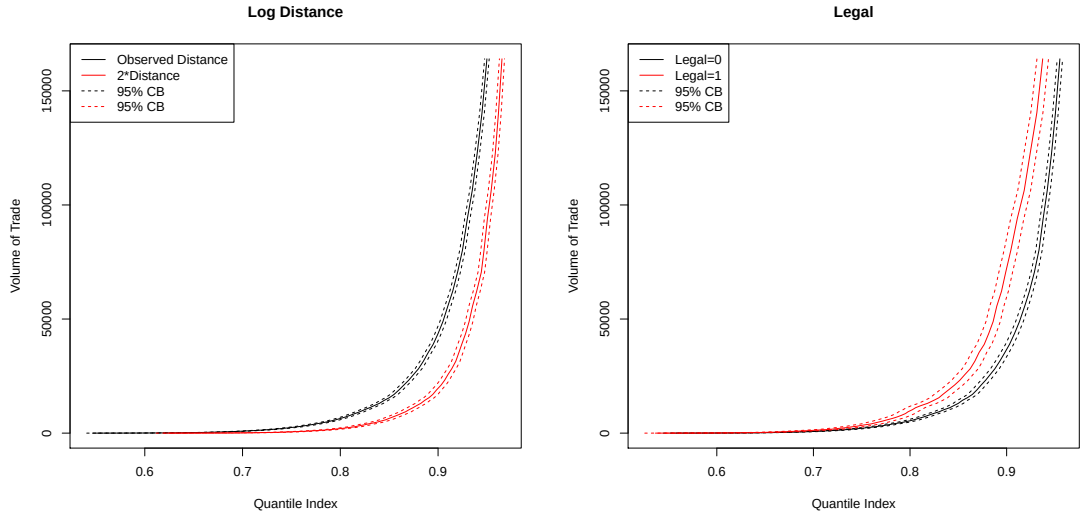

FIGURE 3. Estimates and 95% uniform confidence bands for quantile functions of the volume of trade.

Figure 4 displays estimates and 95% uniform confidence bands for the quantile effects of the log of distance and the common legal system on the volume of trade, constructed using Lemma 2. For comparison, we also include estimates from a Poisson model. Here,

we replace the DR estimators of the distributions by

$$\widehat{F}_k(y) = \frac{1}{n} \sum_{(ij)\in\mathcal{D}} \exp \lambda_{ij,k} \sum_{\tilde{y}=0}^{\lfloor y \rfloor} \frac{\lambda_{ij,k}^{\tilde{y}}}{\tilde{y}!}, \quad k \in \mathcal{K}, \tag{18}$$

where $\lfloor y \rfloor$ is the integer part of y, $\lambda_{ij,k} = \exp(\mathbb{x}'_{ij,k}\widehat{\beta} + \widehat{\alpha}_i + \widehat{\gamma}_j)$, and $\widehat{\theta} = (\widehat{\beta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_I, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_J)$ is the Poisson fixed effects conditional maximum likelihood estimator

$$\widehat{\theta} \in \arg \max_{\theta \in \mathbb{R}^{d_x+I+J}} \sum_{(ij)\in\mathcal{D}} [y_{ij}(x'_{ij}\beta + \alpha_i + \gamma_j) - \exp(x'_{ij}\beta + \alpha_i + \gamma_j)].$$

We find that distance and common legal system have heterogenously increasing effects along the distribution. For example, the negative effects of doubling the distance grows more than proportionally as we move up to the upper tail of the distribution of volume of trade. Putting all the countries under the same legal system has little effects in the extensive margin of trade, but has a strong positive effect at the upper tail of the distribution. The Poisson estimates lie outside the DR confidence bands reflecting overdispersion in the conditional distribution of the volume of trade that is missed by the Poisson model.[14] Figure 5 shows confidence bands of the quantile effects that account for pairwise clustering. The bands are constructed from confidence bands from the distributions using Algorithm 2 with 500 bootstrap draws and standard normal multipliers. Accounting for unobservables that affect symmetrically to the country pairs has very little effect on the width of the bands in this case.

6. Montecarlo Simulation

We conduct a Montecarlo simulation calibrated to the empirical application of Section 5. The outcome is generated by the censored logistic process

$$y^s_{ij} = \max\{x'_{ij}\widehat{\beta} + \widehat{\alpha}_i + \widehat{\gamma}_j + \widehat{\sigma}\Lambda^{-1}(u^s_{ij})/\sigma_L, 0\}, \quad (i,j) \in \mathcal{D},$$

where $\mathcal{D} = \{(i,j) : 1 \leq i,j \leq 157, i \neq j\}$, x_{ij} is the value of the covariates for the observational unit (i,j) in the trade data set, $\sigma_L = \pi/\sqrt{3}$, the standard deviation of the logistic distribution, and $(\widehat{\beta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_I, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_J, \widehat{\sigma})$ are Tobit fixed effect estimates of the parameters in the trade data set with lower censoring point at zero.[15] We consider two designs: independent errors with $u^s_{ij} \sim$ i.i.d $\mathcal{U}(0,1)$, and pairwise dependent errors with $u^s_{ij} = \Phi(0.75e^s_{ij} + \sqrt{1-0.75^2}e^s_{ji})$, where $e^s_{ij} \sim$ i.i.d $\mathcal{N}(0,1)$ and Φ is the standard normal

[14]This overdispersion problem with the Poisson model is well-known in the international trade literature. The Poisson estimator is treated as a quasilikelihood estimator and standard errors robust to misspecification are reported (Santos Silva and Tenreyro, 2006).

[15]We upper winsorize the volume of trade y_{ij} at the 95.5% quantile to reduce the effect of outliers in the Tobit estimation of the parameters.

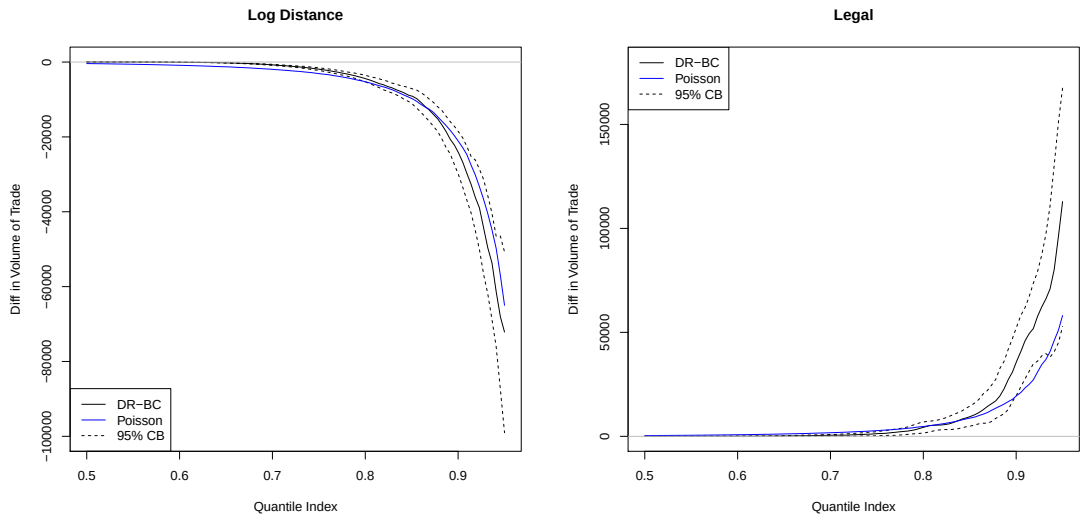

FIGURE 4. Estimates and 95% uniform confidence bands for the quantile effects of log distance and common legal system on the volume of trade.

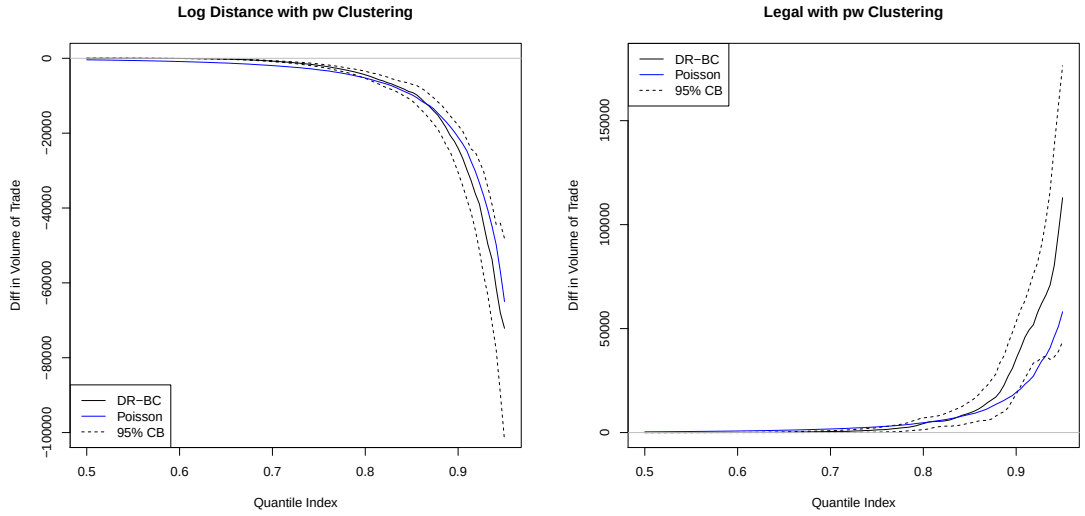

FIGURE 5. Estimates and 95% uniform confidence bands for the quantile effects of log distance and common legal system on the volume of trade.

CDF.[16] In both cases the conditional distribution function of y_{ij}^s is a special case of the

[16]The Spearman rank correlation between u_{ij}^s and u_{ji}^s in the design with pairwise-dependent errors is 0.73.

DR model (1) with link function $\Lambda_y = \Lambda$, the logistic distribution, for all y,

$$\beta(y) = \sigma_L(e_1 y - \widehat{\beta})/\widehat{\sigma}, \quad \alpha_i(y) = -\sigma_L \widehat{\alpha}_i/\widehat{\sigma}, \text{ and } \gamma_j(y) = -\sigma_L \widehat{\gamma}_j/\widehat{\sigma},$$

where e_1 is a unit vector of dimension d_x with a one in the first component. As in the empirical application, the region of interest $\mathcal{Y}$ is the interval between zero and the 0.95-quantile of the volume of trade in the data set. All the results are based on 500 simulated panels $\{(y_{ij}^s, x_{ij}) : (i,j) \in \mathcal{D}\}$.

Figures 6 and 7 report the biases, standard deviations and root mean square errors (rmses) of the fixed effects estimators of the DR coefficients of log-distance and legal system as a function of the quantiles of y_{ij} in the design with independent errors.[17] All the results are in percentage of the true value of the parameter. As predicted by the large sample theory, the fixed effects estimator displays a bias of the same order of magnitude as the standard deviation. As in fig. 1, the bias is more severe for the coefficient of log distance. The bias correction removes most of the bias and does not increase the standard deviation, yielding a reduction in rmse of about 5% for the coefficient of log distance at the highest quantile indexes.

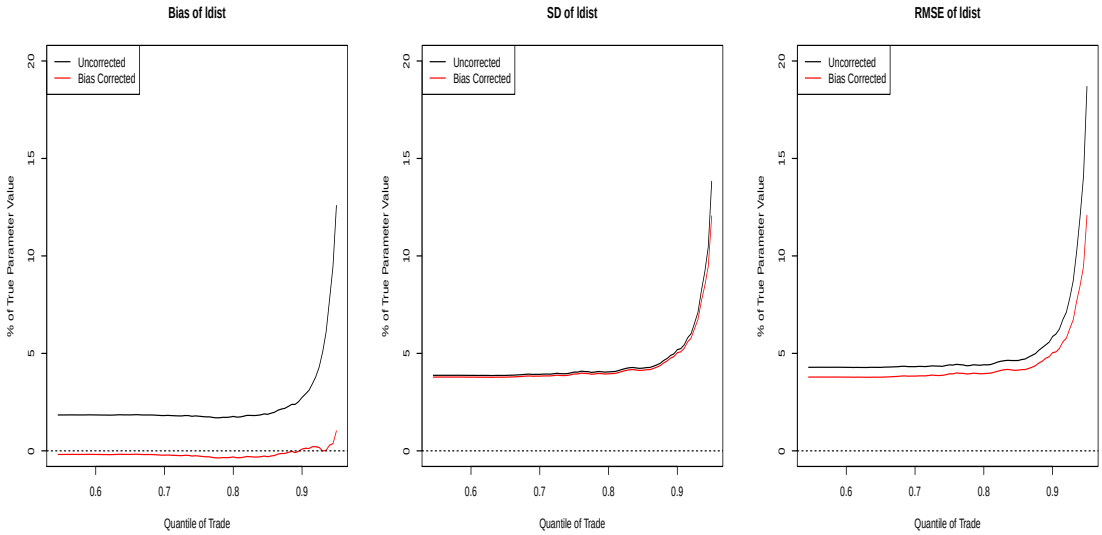

FIGURE 6. Bias, standard deviation and root mean squared error for the estimators of the DR-coefficients of log-distance.

Figure 8 reports the biases, standard deviations and rmses of the estimators of the counterfactual distributions at two levels of log-distance as a function of the quantiles of

[17]The design with pairwise dependent errors produces similar results, which are not reported for the sake of brevity.

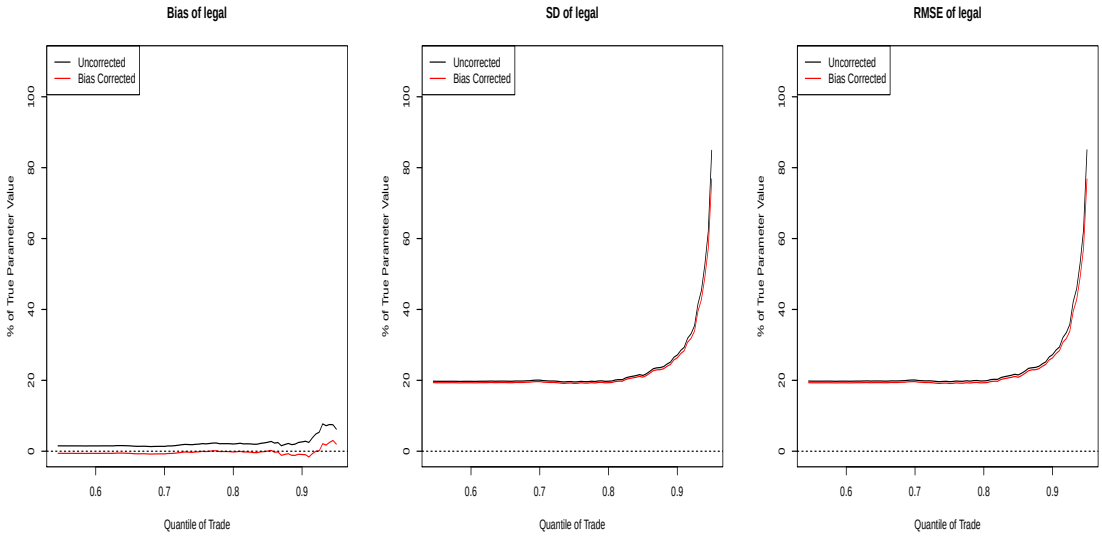

FIGURE 7. Bias, standard deviation and root mean squared error for the estimators of the DR-coefficients of same legal system.

y_{ij} in the design with independent errors. The levels of distance in these distributions are the same as in the empirical application, i.e. $k = 0$ and $k = 1$ correspond to the observed values and two times the observed values, respectively. All the results are in percentage of the true value of the functions. In this case we find that the uncorrected and bias corrected estimators display small biases relative to their standard deviations, and have similar standard deviations and rmses at both treatment levels. Indeed the standard deviations and rmses are difficult to distinguish in the figure as they almost superposed. In results not reported, we find very similar patterns in the design with pairwise dependent errors and for the estimators of the counterfactual distributions at the same two levels of legal as in the empirical application.

Table 2 shows results on the finite sample properties of 95% confidence bands for the DR coefficients and counterfactual distributions in the design with independent errors. The confidence bands are constructed by multiplier bootstrap with 500 draws, standard normal weights, and a grid of values $\bar{\mathcal{Y}}$ that includes the sample quantiles of the volume of trade with indexes $\{.54, .55, \ldots, .95\}$ in the trade data set. For the coefficients, it reports the average length of the confidence bands integrated over threshold values, the average value of the estimated critical values, and the empirical coverages of the confidence bands. For the distributions, it report the same measures averaged also over the two treatment levels and where the coverage of the bands is joint for the two counterfactual distributions.[18] For

[18]The joint coverage of the bands for the quantile functions and quantile effect is determined by the joint coverage of the bands of the distribution functions in our construction.

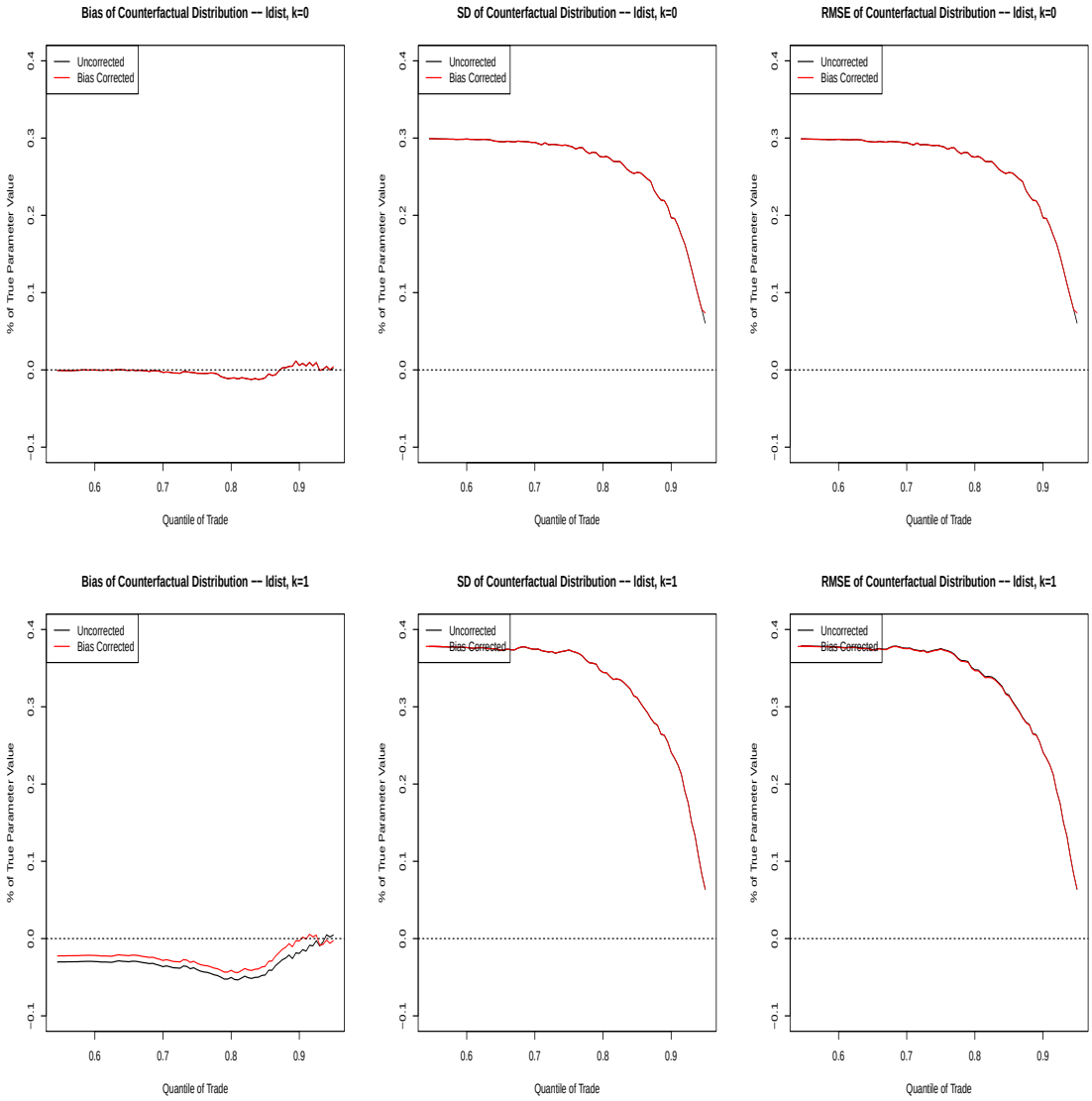

FIGURE 8. Bias, standard deviation and root mean squared error for the estimators of the counterfactual distributions of log-distance.

comparison, it also reports the coverage of pointwise confidence bands using the normal distribution, i.e. with critical value equal to 1.96. The last row computes the ratio of the standard error averaged across simulations to the simulation standard deviation, integrated over threshold values for the coefficients and over thresholds and treatment levels for the distributions. We consider standard errors and confidence bands with and without

TABLE 2. 95% Confidence Bands – Design with Independent Errors

	Uncorrected				Bias Corrected			
	β_{ldist}	β_{legal}	F_{ldist}	F_{legal}	β_{ldist}	β_{legal}	F_{ldist}	F_{legal}
Unclustered Inference								
Average Length	0.24	0.35	0.01	0.02	0.24	0.35	0.01	0.02
Average Critical Value	2.90	2.89	3.10	3.13	2.90	2.89	3.10	3.13
Coverage uniform band (%)	83	91	94	93	95	94	94	94
Coverage pointwise band (%)	35	58	35	29	60	64	35	29
Average SE/SD	0.97	1.01	0.99	1.01	1.00	1.04	0.99	1.01
Pairwise Clustered Inference								
Average Length	0.23	0.35	0.01	0.02	0.23	0.35	0.01	0.02
Average Critical Value	2.89	2.89	3.09	3.12	2.89	2.89	3.09	3.12
Coverage uniform band (%)	82	92	93	93	94	93	93	93
Coverage pointwise band (%)	35	57	35	30	59	63	36	29
Average SE/SD	0.97	1.01	0.99	1.01	1.00	1.04	0.99	1.01

Notes: Nominal level of critical values is 95%. 500 simulations with 500 multiplier bootstrap draws.

accounting for pairwise clustering. All the results are computed for confidence bands centered at the uncorrected fixed effects estimates and at the bias corrected estimates. For the coefficients, we find that the bands centered at the uncorrected estimates undercover the true coefficients, whereas the bands centered at the bias corrected estimates have coverages close to the nominal level. The joint coverage of the bands for the distributions is close to the nominal level regardless of whether they are centered at the uncorrected or bias corrected estimates. We attribute this similarity in coverage to the small biases in the uncorrected estimates of the distributions found in fig. 8. As expected, pointwise bands severely undercover the entire functions. The standard errors based on the asymptotic distribution provide a good approximation to the sampling variability of both the uncorrected and bias corrected estimators. Accounting for pairwise clustering in this design where it is not necessary has very little effect on the quality of the inference.

Table 3 reports the same results as table 2 for the design with pairwise dependent errors. The bands that do not account for pairwise clustering undercover the functions because the standard errors underestimate the standard deviations of the estimators. Compared to the design with independent errors, the critical values are similar but the bands that account for clustering are wider due to the increase in the standard errors. To sum up, inference methods robust to pairwise clustering perform well in both designs, whereas inference methods that do not account for clustering undercover in the presence of pairwise dependence. The bias corrections are effective in reducing bias and bringing the coverage probabilities of the bands close to their nominal level for the coefficients, whereas they have little effect for the distributions.

Table 3. 95% Confidence Bands – Design with Pairwise Dependent Errors

	Uncorrected				Bias Corrected			
	β_{ldist}	β_{legal}	F_{ldist}	F_{legal}	β_{ldist}	β_{legal}	F_{ldist}	F_{legal}
Unclustered Inference								
Average Length	0.24	0.35	0.01	0.02	0.24	0.35	0.01	0.02
Average Critical Value	2.90	2.89	3.10	3.13	2.90	2.89	3.10	3.13
Coverage uniform band (%)	64	73	73	68	80	78	74	68
Coverage pointwise band (%)	21	27	11	8	32	36	12	8
Average SE/SD	0.77	0.76	0.77	0.77	0.79	0.78	0.77	0.77
Pairwise Clustered Inference								
Average Length	0.30	0.44	0.02	0.02	0.30	0.44	0.02	0.02
Average Critical Value	2.82	2.82	3.02	3.05	2.82	2.82	3.02	3.05
Coverage uniform band (%)	86	92	93	92	96	93	93	92
Coverage pointwise band (%)	47	59	44	37	67	66	43	37
Average SE/SD	1.00	0.99	1.00	0.99	1.03	1.01	1.00	0.99

Notes: Nominal level of critical values is 95%. 500 simulations with 500 multiplier bootstrap draws.

Appendix A. Proofs of Main Text Results

We present the proofs of Theorems 1 and 2, and relegate various technical details to the supplementary appendix. Once Theorems 1 and 2 are shown, the proof of Theorem 3 for the multiplier bootstrap follows from results in Chernozhukov, Chetverikov and Kato (2016). The uniform confidence bands I_F for the cdfs in (10) obtained by the multiplier bootstrap can then be inverted and differenced to obtain uniform confidence bands for the quantile function and quantile effects, see Chernozhukov, Fernández-Val, Melly and Wüthrich (2016) and also Lemma 1 and 2 above. This appendix thus contains the proofs of all the main results that are new to the current paper. The proofs for all of the lemmas below are given in the supplementary appendix. All stochastic statements in the following are conditional on $\{(x_{ij}, v_i, w_j) : (i,j) \in \mathcal{D}\}$.

As explained in Section 4, we consider the logistic cdf $\Lambda_y(\pi) = \Lambda(\pi) = (1 + \exp(-\pi))^{-1}$ for all our theorems. In the following we indicate the dependence on $y \in \mathcal{Y}$ as a subscript, for example, we write θ_y instead of $\theta(y)$ from now on. We use the column vector $w_{ij} = (x'_{ij}, e'_{i,I}, e'_{j,J})'$, as in Section 4.3, and can then write the single index $\pi_{y,ij} := x'_{ij}\beta_y + \alpha_{y,i} + \gamma_{y,j}$ simply as $\pi_{y,ij} = w'_{ij}\theta_y$. The corresponding estimator is $\widehat{\pi}_{y,ij} = w'_{ij}\widehat{\theta}_y$. We also define minus the log-likelihood function as $\ell_{y,ij}(\pi) := -1\{y_{ij} \leq y\} \log \Lambda(\pi) - 1\{y_{ij} > y\} \log[1 - \Lambda(\pi)]$. Let π_y be a n-vector containing $\pi_{y,ij}$, $(i,j) \in \mathcal{D}$. For a given $y \in \mathcal{Y}$ we

can then rewrite the estimation problem in (3) as

$$\widehat{\pi}_y = \arg\min_{\pi_y \in \mathbb{R}^n} \sum_{(i,j)\in\mathcal{D}} \ell_{y,ij}(\pi_{y,ij}), \qquad \text{s.t.} \qquad \exists\, \theta \in \mathbb{R}^{d_x+I+J} : \pi_{y,ij} = w_{ij}'\theta_y. \tag{A.1}$$

In the following we denote the true parameter values by θ^0, and correspondingly we write $\pi^0_{y,ij} = w_{ij}'\theta^0_y$, in order to distinguish the true value from generic values like the argument $\pi_{y,ij}$ in the last display. For the k'th derivative of $\ell_{y,ij}(\pi_{y,ij})$ with respect to $\pi_{y,ij}$ we write $\partial_{\pi^k}\ell_{y,ij}(\pi_{y,ij})$, and we drop the argument when the derivative is evaluated at $\pi^0_{y,ij}$, that is, $\partial_{\pi^k}\ell_{y,ij} = \partial_{\pi^k}\ell_{y,ij}(\pi^0_{y,ij})$. The normalized score for observation i,j then reads

$$s_{y,ij} := [\partial_{\pi^2}\ell_{y,ij}]^{-1/2}\,\partial_\pi \ell_{y,ij} = \left(\Lambda^{(1)}_{y,ij}\right)^{-1/2}\partial_\pi \ell_{y,ij},$$

where $\Lambda^{(1)}_{y,ij} = \Lambda^{(1)}(\pi^0_{y,ij}) = \partial_\pi \Lambda(\pi^0_{y,ij})$, as defined in Section 4.1. Note that $\mathbb{E}s_{y,ij} = 0$ and $\mathbb{E}s^2_{y,ij} = 1$.

Let s_y be the n-vector obtained by stacking the elements $s_{y,ij}$ across all observations $(i,j)\in\mathcal{D}$. Similarly, let $\Lambda^{(1)}_y$ be the $n\times n$ diagonal matrix with diagonal elements given by $\Lambda^{(1)}_{y,ij}$, $(i,j)\in\mathcal{D}$. Finally, let w be the $n\times(d_x+I+J)$ matrix with rows given by w_{ij}', $(i,j)\in\mathcal{D}$. We define the $n\times n$ symmetric idempotent matrix

$$Q_y := \left(\Lambda^{(1)}_y\right)^{1/2} w \left(w'\Lambda^{(1)}_y w\right)^{\dagger} w' \left(\Lambda^{(1)}_y\right)^{1/2},$$

where $\dagger$ is the Moore-Penrose pseudoinverse. For the elements of this matrix we write $Q_{y,ij,i'j'}$. We have $(Q_y s_y)_{ij} = \sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'} s_{y,i'j'}$. The constraint $\exists\,\theta : \pi_{y,ij} = w_{ij}'\theta_y$ in (A.1) can then equivalently be written as[19]

$$Q_y \left(\Lambda^{(1)}_y\right)^{1/2} \pi_y = \left(\Lambda^{(1)}_y\right)^{1/2} \pi_y. \tag{A.2}$$

The matrix Q_y projects onto the column span of $\left(\Lambda^{(1)}_y\right)^{1/2} w$. This projector acts in the space of weighted index vectors $\left[\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\pi_{y,ij} \,:\, (i,j)\in\mathcal{D}\right]$, and the weighting of each $\pi_{y,ij}$ by $\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}$ is natural, because $\Lambda^{(1)}_{y,ij}$ is simply the expected Hessian for observation (i,j).

[19] In matrix notation the constraint can be written as $\pi_y = w\,\theta_y$, and we thus have $Q_y\left(\Lambda^{(1)}_y\right)^{1/2}\pi_y = Q_y\left(\Lambda^{(1)}_y\right)^{1/2} w\,\theta_y = \left(\Lambda^{(1)}_y\right)^{1/2} w\,\theta_y = \left(\Lambda^{(1)}_y\right)^{1/2}\pi_y$, where we also used that $Q_y\left(\Lambda^{(1)}_y\right)^{1/2} w = \left(\Lambda^{(1)}_y\right)^{1/2} w$, which follows from the definition of Q_y.

A.1. **Technical Lemmas.** We require some results for the proofs of the main theorems below. The following lemma provides an asymptotic expansion of $\widehat{\pi}_{y,ij} - \pi^0_{y,ij}$.

Lemma 3 (Score expansion of fixed effect estimates)**.** *Under Assumption 1, for $y \in \mathcal{Y}$ and $(i,j) \in \mathcal{D}$, we have*

$$\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\left(\widehat{\pi}_{y,ij} - \pi^0_{y,ij}\right) = -\left(Q_y s_y\right)_{ij} - \frac{1}{2}\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\,\frac{\Lambda^{(2)}_{y,i'j'}}{\left(\Lambda^{(1)}_{y,i'j'}\right)^{3/2}}\left[\left(Q_y s_y\right)_{i'j'}\right]^2 + r_{y,ij},$$

and the remainder $r_{y,ij}$ satisfies $\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}|r_{y,ij}| = o_P(n^{-1/2})$.

The expansion in the preceding lemma is a second-order stochastic expansion, because it does not only describe the terms linear in the score s_y, but also the terms quadratic in s_y. We need to keep track of those quadratic terms, because they yield the leading order incidental parameter biases that appear in Theorem 1. The remainder $r_{y,ij}$ contains higher-order terms in s_y (cubic, quartic, etc), which turn out not to matter for the result in Theorem 1. Note also that $\Lambda^{(2)}_{y,ij} = \partial_{\pi^3}\ell_{y,ij}$. Thus, the term quadric in the score is proportional to the third derivative of the objective function.

We now want to decompose the projector Q_y into the parts stemming from x_{ij}, $e_{i,I}$ and $e_{j,J}$, respectively. We have already introduced the d_x-vector $\widetilde{x}_{y,ij} = \widetilde{x}_{ij}(y)$ in Section 4.1. Let $\widetilde{x}_y$ be the $n \times d_x$ matrix with rows given by $\widetilde{x}'_{y,ij}$, $(i,j) \in \mathcal{D}$. The $d_x \times d_x$ matrix $W_y = W(y) = n^{-1}\widetilde{x}'_y\Lambda^{(1)}_y\widetilde{x}_y$ was also introduced in Section 4.1. Invertibility of W_y is guaranteed by Assumption 1(vi), and uniform boundedness of $\Lambda^{(1)}_{y,ij}$ and $\left(\Lambda^{(1)}_{y,ij}\right)^{-1}$, as formalized by the following lemma.

Lemma 4 (Invertibility of W_y)**.** *Let Assumption 1 hold. Then $\sup_{y\in\mathcal{Y}}\|W_y^{-1}\| = O_P(1)$.*

Next, define $w^{(2)}_{ij} = e_{i,I}$ and $w^{(3)}_{ij} = e_{j,J}$, and let $w^{(2)}$ and $w^{(3)}$ be the corresponding $n \times I$ and $n \times J$ matrices with rows given by $w^{(2)'}_{ij}$ and $w^{(3)'}_{ij}$, respectively. Let

$$Q^{(1)}_y := n^{-1}\left(\Lambda^{(1)}_y\right)^{1/2}\widetilde{x}_y\,W_y^{-1}\,\widetilde{x}'_y\left(\Lambda^{(1)}_y\right)^{1/2},$$

$$Q^{(\mathrm{FE})}_y := \left(\Lambda^{(1)}_y\right)^{1/2}\left[w^{(2)}, w^{(3)}\right]\left(\left[w^{(2)}, w^{(3)}\right]'\Lambda^{(1)}_y\left[w^{(2)}, w^{(3)}\right]\right)^{\dagger}\left[w^{(2)}, w^{(3)}\right]'\left(\Lambda^{(1)}_y\right)^{1/2}.$$

$\widetilde{x}_{y,ij}$ is defined as the part of $x_{y,ij}$ that is orthogonal to the fixed effects under a metric given by $\Lambda^{(1)}_{y,ij}$. We have $Q^{(\mathrm{FE})}_y\left(\Lambda^{(1)}_y\right)^{1/2}\widetilde{x}_y = 0$, which implies that

$$Q_y = Q^{(1)}_y + Q^{(\mathrm{FE})}_y \tag{A.3}$$

and also $Q_y^{(1)} Q_y^{(\mathrm{FE})} = Q_y^{(\mathrm{FE})} Q_y^{(1)} = 0$. Also, because $Q_y^{(1)} \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y = \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y$ and $Q_y^{(\mathrm{FE})} \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y = 0$, we obtain

$$Q_y \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y = \left(Q_y^{(1)} + Q_y^{(\mathrm{FE})}\right) \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y = \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y. \tag{A.4}$$

We have thus decomposed Q_y into the component stemming from the regressors and a component stemming from the fixed effects. For the elements of $Q_y^{(1)}$,

$$Q_{y,ij,i'j'}^{(1)} = n^{-1} \left(\Lambda_{y,ij}^{(1)} \Lambda_{y,i'j'}^{(1)}\right)^{1/2} \widetilde{x}_{y,ij}' \, W_y^{-1} \, \widetilde{x}_{y,i'j'}. \tag{A.5}$$

Next, define the projection matrices

$$Q_y^{(2)} := \left(\Lambda_y^{(1)}\right)^{1/2} w^{(2)} \left(w^{(2)\prime} \Lambda_y^{(1)} w^{(2)}\right)^{-1} w^{(2)\prime} \left(\Lambda_y^{(1)}\right)^{1/2},$$
$$Q_y^{(3)} := \left(\Lambda_y^{(1)}\right)^{1/2} w^{(3)} \left(w^{(3)\prime} \Lambda_y^{(1)} w^{(3)}\right)^{-1} w^{(3)\prime} \left(\Lambda_y^{(1)}\right)^{1/2}.$$

Notice that $w^{(2)\prime} \Lambda_y^{(1)} w^{(2)}$ and $w^{(3)\prime} \Lambda_y^{(1)} w^{(3)}$ are simply diagonal $I \times I$ and $J \times J$ matrices with diagonal entries $\sum_{j \in \mathcal{D}_i} \Lambda_{y,ij}^{(1)}$ and $\sum_{i \in \mathcal{D}_j} \Lambda_{y,ij}^{(1)}$, respectively, and therefore

$$Q_{y,ij,i'j'}^{(2)} = 1(i = i') \frac{\left(\Lambda_{y,ij}^{(1)} \Lambda_{y,ij'}^{(1)}\right)^{1/2}}{\sum_{j'' \in \mathcal{D}_i} \Lambda_{y,ij''}^{(1)}}, \qquad Q_{y,ij,i'j'}^{(3)} = 1(j = j') \frac{\left(\Lambda_{y,ij}^{(1)} \Lambda_{y,i'j}^{(1)}\right)^{1/2}}{\sum_{i'' \in \mathcal{D}_j} \Lambda_{y,i''j}^{(1)}}. \tag{A.6}$$

It is not exactly true that $Q_y^{(\mathrm{FE})}$ equals $Q_y^{(2)} + Q_y^{(3)}$, but Lemma 5 shows that this is approximately true in a well-defined sense.

Lemma 5 (Properties of Q_y)**.** *Under Assumption 1,*

(i) $Q_y = Q_y^{(1)} + Q_y^{(\mathrm{FE})}$ *and* $Q_y^{(\mathrm{FE})} = Q_y^{(2)} + Q_y^{(3)} + Q_y^{(\mathrm{rem})}$, *where*
$$\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \max_{(i',j') \in \mathcal{D}} \left| Q_{y,ij,i'j'}^{(\mathrm{rem})} \right| = O_P(n^{-1}).$$

(ii) $\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \sum_{(i',j') \in \mathcal{D}} \left| Q_{y,ij,i'j'} \right| = O_P(1)$, *and*
$\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \sum_{(i',j') \in \mathcal{D}} \left| Q_{y,ij,i'j'}^{(\mathrm{FE})} \right| = O_P(1)$.

(iii) $\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \max_{(i',j') \in \mathcal{D}} \left| Q_{y,ij,i'j'} \right| = O_P(n^{-1/2})$.

Remark 8 (Bias of $\widehat{\pi}_{y,ij}$)**.** According to part (i) of this lemma the remainder term $Q_y^{(\mathrm{rem})} = Q_y^{(\mathrm{FE})} - Q_y^{(2)} - Q_y^{(3)}$ has elements uniformly bounded of order n^{-1}, and it can easily be seen from (A.5) that the same is true for $Q_y^{(1)}$, because the elements of $\widetilde{x}_y$ are also uniformly bounded under our assumptions. By contrast, $Q_y^{(2)}$ and $Q_y^{(3)}$ have elements of order J^{-1}

and I^{-1}, respectively, that is, of order $n^{-1/2}$. Using this and the fact that $s_{y,ij}$ has variance one and is independent across observations (i,j) we find

$$\begin{aligned}\mathbb{E}\left[(Q_y s_y)_{ij}\right]^2 &= \sum_{(i',j')\in\mathcal{D}} [Q_{y,ij,i'j'}]^2 = Q_{y,ij,ij} = Q^{(2)}_{y,ij,ij} + Q^{(3)}_{y,ij,ij} + O_P(n^{-1}) \\ &= \frac{\Lambda^{(1)}_{y,ij}}{\sum_{j'\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij'}} + \frac{\Lambda^{(1)}_{y,ij}}{\sum_{i'\in\mathcal{D}_j}\Lambda^{(1)}_{y,i'j}} + O_P(n^{-1}), \end{aligned} \tag{A.7}$$

where we use that Q_y is idempotent in the second step, and (A.6) in the third step. Combining this with Lemma 3 one finds that the leading order bias term in $\widehat{\pi}_{y,ij} - \pi^0_{y,ij}$ is given by

$$-\frac{1}{2}\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\,\frac{\Lambda^{(2)}_{y,i'j'}}{\Lambda^{(1)}_{y,i'j'}}\left[\frac{1}{\sum_{j'\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij'}} + \frac{1}{\sum_{i'\in\mathcal{D}_j}\Lambda^{(1)}_{y,i'j}}\right],$$

which then translates into corresponding bias terms for all other estimators as well.

For the following lemma, let $Z^{(\beta)}_y = Z^{(\beta)}(y)$, $Z^{(F)}_y = Z^{(F)}(y)$, $B^{(\beta)}_y = B^{(\beta)}(y)$, $D^{(\beta)}_y = D^{(\beta)}(y)$, $B^{(\Lambda)}_{y,k} = B^{(\Lambda)}_k(y)$ and $D^{(\Lambda)}_{y,k} = D^{(\Lambda)}_k(y)$ be as defined in and before Theorem 1 in the main text.

Lemma 6 (Properties of score averages)**.** *Under Assumption 1,*

(i) $\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|(Q_y s_y)_{ij}\right| = o_P(n^{-1/6})$.

(ii) $-W_y^{-1}\, n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\widetilde{x}_{y,ij}\,\partial_\pi \ell_{y,ij} \rightsquigarrow Z^{(\beta)}_y$, *in* $\ell^\infty(\mathcal{Y})^{d_x}$.

(iii) $-\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\left[\Psi_{y,ij} + (\partial_\beta F_y)\, W_y^{-1}\,\widetilde{x}_{y,ij}\right]\partial_\pi \ell_{y,ij} \rightsquigarrow Z^{(F)}_y$, *in* $\ell^\infty(\mathcal{Y})^{|\mathcal{K}|}$.

(iv) $-\frac{1}{2}W_y^{-1}\,\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\widetilde{x}_{y,ij}\left(\Lambda^{(1)}_{y,ij}\right)^{-1}\Lambda^{(2)}_{y,ij}\left[(Q_y s_y)_{ij}\right]^2 - \left(\frac{I}{\sqrt{n}}B^{(\beta)}_y + \frac{J}{\sqrt{n}}D^{(\beta)}_y\right) \to_P 0$, *uniformly in* $y\in\mathcal{Y}$.

(v) $\frac{1}{2\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\left(\Lambda^{(1)}_{y,ij}\right)^{-1}\left(\Lambda^{(2)}_{y,ij,k} - \Lambda^{(2)}_{y,ij}\Psi_{y,ij,k}\right)\left[(Q_y s_y)_{ij}\right]^2 - \left(\frac{I}{\sqrt{n}}B^{(\Lambda)}_{y,k} + \frac{J}{\sqrt{n}}D^{(\Lambda)}_{y,k}\right) \to_P 0$, *uniformly in* $y\in\mathcal{Y}$.

Regarding part (i) of this lemma, notice that pointwise we have $(Q_y s_y)_{ij} = O_P(n^{-1/4})$, because (A.7) implies that $\mathbb{E}\left[(Q_y s_y)_{ij}\right]^2 = O_P(n^{-1/2})$. However, after taking the supremum over y, i, j the term is growing faster than $n^{-1/4}$. The rate $o_P(n^{-1/6})$ in part (i) of the lemma is crude, but sufficient for our purposes.

A.2. **Proof of Main Text Theorems.**

Proof of Theorem 1. # Part 1: FCLT for $\widehat{\beta}_y = \widehat{\beta}(y)$.

The definition of $\widetilde{x}_y$ implies that $\sum_{i\in\mathcal{D}_j} \Lambda^{(1)}_{y,ij}\widetilde{x}_{y,ij} = 0$ and $\sum_{j\in\mathcal{D}_i} \Lambda^{(1)}_{y,ij}\widetilde{x}_{y,ij} = 0$, and $n^{-1}\sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{y,ij}\widetilde{x}_{y,ij}x'_{ij} = n^{-1}\sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{y,ij}\widetilde{x}_{y,ij}\widetilde{x}'_{ij} = W_y$. Using this and

$$\widehat{\pi}_{y,ij} - \pi^0_{y,ij} := x'_{ij}\left(\widehat{\beta}_y - \beta^0_y\right) + \left(\widehat{\alpha}_{y,i} - \alpha^0_{y,i}\right) + \left(\widehat{\gamma}_{y,j} - \gamma^0_{y,j}\right)$$

we obtain

$$n^{-1}\sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij}\,\Lambda^{(1)}_{y,ij}\,\left(\widehat{\pi}_{y,ij} - \pi^0_{y,ij}\right) = n^{-1}\sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij}\,\Lambda^{(1)}_{y,ij}\,x'_{ij}\left(\widehat{\beta}_y - \beta^0_y\right) = W_y\left(\widehat{\beta}_y - \beta^0_y\right),$$

and therefore

$$\widehat{\beta}_y - \beta^0_y = W_y^{-1}\,n^{-1}\sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij}\,\Lambda^{(1)}_{y,ij}\,\left(\widehat{\pi}_{y,ij} - \pi^0_{y,ij}\right).$$

By combining this with Lemma 3 we obtain

$$\sqrt{n}\left(\widehat{\beta}_y - \beta^0_y\right) = T^{(1,\beta)}_y + T^{(2,\beta)}_y + r^{(\beta)}_y, \tag{A.8}$$

where

$$T^{(1,\beta)}_y := -n^{-1/2}\,W_y^{-1}\sum_{(i,j)\in\mathcal{D}}\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\,\widetilde{x}_{y,ij}\,\left(Q_y s_y\right)_{ij},$$

$$T^{(2,\beta)}_y := -\frac{1}{2}n^{-1/2}\,W_y^{-1}\sum_{(i,j)\in\mathcal{D}}\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\,\widetilde{x}_{y,ij}\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\,\frac{\Lambda^{(2)}_{y,i'j'}}{\left(\Lambda^{(1)}_{y,i'j'}\right)^{3/2}}\left[\left(Q_y s_y\right)_{i'j'}\right]^2,$$

and $r^{(\beta)}_y := W_y^{-1}\,n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\widetilde{x}_{y,ij}\,\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\,r_{y,ij}$ satisfies

$$\sup_{y\in\mathcal{Y}}\left|r^{(\beta)}_y\right| \le \underbrace{\left(\sup_{y\in\mathcal{Y}} W_y^{-1}\,n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\left|\widetilde{x}_{y,ij}\right|\left|\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}\right|\right)}_{=O_P(n^{1/2})}\underbrace{\left(\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|r_{y,ij}\right|\right)}_{=o_P(n^{-1/2})} = o_P(1),$$

where we also use that $\Lambda^{(1)}_{y,ij}$ and $\widetilde{x}_{y,ij}$ are uniformly bounded under our assumptions. For the term linear in the score we find

$$\begin{aligned} T^{(1,\beta)}_y &= -n^{-1/2}W_y^{-1}\,\widetilde{x}'_y\,\left(\Lambda^{(1)}_y\right)^{1/2}Q_y s_y = -n^{-1/2}W_y^{-1}\,\widetilde{x}'_y\,\left(\Lambda^{(1)}_y\right)^{1/2}\,s_y \\ &= -W_y^{-1}\,n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\widetilde{x}_{y,ij}\,\partial_\pi\ell_{y,ij} \rightsquigarrow Z^{(\beta)}_y, \end{aligned}$$

where in the second step we used (A.4), and the final step follows from part (ii) of Lemma 6.

Employing again (A.4) we find

$$T_y^{(2,\beta)} := -\frac{1}{2}\, W_y^{-1}\, n^{-1/2} \sum_{(i',j')\in\mathcal{D}} \underbrace{\sum_{(i,j)\in\mathcal{D}} \left(\Lambda_{y,ij}^{(1)}\right)^{1/2} \widetilde{x}_{y,ij}\, Q_{y,ij,i'j'}}_{=\left(\Lambda_{y,i'j'}^{(1)}\right)^{1/2} \widetilde{x}_{y,i'j'}} \frac{\Lambda_{y,i'j'}^{(2)}}{\left(\Lambda_{y,i'j'}^{(1)}\right)^{3/2}} \left[(Q_y s_y)_{i'j'}\right]^2$$

$$= -\frac{1}{2} W_y^{-1}\, n^{-1/2} \sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij} \frac{\Lambda_{y,ij}^{(2)}}{\Lambda_{y,ij}^{(1)}} \left[(Q_y s_y)_{ij}\right]^2,$$

and according to part (iv) of Lemma 6 we thus have

$$T_y^{(2,\beta)} - \left(\frac{I}{n^{1/2}}\, B_y^{(\beta)} + \frac{J}{n^{1/2}}\, D_y^{(\beta)}\right) \to_P 0,$$

uniformly in $y \in \mathcal{Y}$. Combining the above gives the result for $\sqrt{n}\left(\widehat{\beta}_y - \beta_y^0\right)$ in the theorem.

<u># Part 2: FCLT for $\widehat{F}_{y,k} = \widehat{F}_k(y)$.</u>

Let $\pi_{y,ij,k}^0 := \pi_{y,ij}^0 + (\mathbb{x}_{ij,k} - x_{ij})'\beta_y^0$ and $\widehat{\pi}_{y,ij,k} := \widehat{\pi}_{y,ij} + (\mathbb{x}_{ij,k} - x_{ij})'\widehat{\beta}_y$. Because $\mathbb{x}_{ij,k} - x_{ij} = \widetilde{\mathbb{x}}_{y,ij,k} - \widetilde{x}_{y,ij}$ we have

$$\widehat{\pi}_{y,ij,k} - \pi_{y,ij,k}^0 = \widehat{\pi}_{y,ij} - \widehat{\pi}_{y,ij} + (\widetilde{\mathbb{x}}_{y,ij,k} - \widetilde{x}_{y,ij})'(\widehat{\beta}_y - \beta_y^0). \tag{A.9}$$

Using (A.2) and $Q_y^{(1)}\left(\Lambda_y^{(1)}\right)^{1/2} \pi_y = \left(\Lambda_y^{(1)}\right)^{1/2} \widetilde{x}_y \beta_y$ for any $\pi_y = w\theta_y$,

$$\widehat{\pi}_y - \pi_y^0 = \left(\Lambda_y^{(1)}\right)^{-1/2} \underbrace{\left(Q_y^{(1)} + Q_y^{(\mathrm{FE})}\right)}_{=Q_y} \left(\Lambda_y^{(1)}\right)^{1/2} (\widehat{\pi}_y - \pi_y^0)$$

$$= \left(\Lambda_y^{(1)}\right)^{-1/2} Q_y^{(\mathrm{FE})} \left(\Lambda_y^{(1)}\right)^{1/2} (\widehat{\pi}_y - \pi_y^0) + \widetilde{x}_y(\widehat{\beta}_y - \beta_y^0).$$

Combining the above gives

$$\widehat{\pi}_{y,ij,k} - \pi_{y,ij,k}^0 = \left[\left(\Lambda_y^{(1)}\right)^{-1/2} Q_y^{(\mathrm{FE})} \left(\Lambda_y^{(1)}\right)^{1/2} (\widehat{\pi}_y - \pi_y^0)\right]_{ij} + \widetilde{\mathbb{x}}_{y,ij,k}'(\widehat{\beta}_y - \beta_y^0).$$

Using Lemma 3 and the properties of Q_y, $Q_y^{(1)}$ and $Q_y^{(\mathrm{FE})}$, we thus find

$$\left(\Lambda_{y,ij}^{(1)}\right)^{1/2} \left(\widehat{\pi}_{y,ij,k} - \pi_{y,ij,k}^0\right) = -\left(Q_y^{(\mathrm{FE})} s_y\right)_{ij} - \frac{1}{2} \sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}^{(\mathrm{FE})} \frac{\Lambda_{y,i'j'}^{(2)}}{\left(\Lambda_{y,i'j'}^{(1)}\right)^{3/2}} \left[(Q_y s_y)_{i'j'}\right]^2$$

$$+ \left(Q^{(\mathrm{FE})} r_y\right)_{ij} + \left(\Lambda_{y,ij}^{(1)}\right)^{1/2} \widetilde{\mathbb{x}}_{y,ij,k}'(\widehat{\beta}_y - \beta_y^0). \tag{A.10}$$

Next, by expanding $\Lambda(\widehat{\pi}_{y,ij,k})$ in $\widehat{\pi}_{y,ij,k}$ around $\pi^0_{y,ij,k}$ we find

$$\begin{aligned}\widehat{F}_{y,k} - F_{y,k} &= n^{-1} \sum_{(i,j)\in\mathcal{D}} \left[\Lambda(\widehat{\pi}_{y,ij,k}) - \Lambda(\pi^0_{y,ij,k})\right] \\ &= n^{-1} \sum_{(i,j)\in\mathcal{D}} \Bigg[\Lambda^{(1)}_{y,ij,k}\left(\widehat{\pi}_{y,ij,k} - \pi^0_{y,ij,k}\right) + \frac{1}{2}\Lambda^{(2)}_{y,ij,k}\left(\widehat{\pi}_{y,ij,k} - \pi^0_{y,ij,k}\right)^2 \\ &\qquad + \frac{1}{6}\Lambda^{(3)}(\widetilde{\pi}_{y,ij,k})\left(\widehat{\pi}_{y,ij,k} - \pi^0_{y,ij,k}\right)^3\Bigg],\end{aligned}$$

where $\widetilde{\pi}_{y,ij,k}$ is some value between $\widehat{\pi}_{y,ij,k}$ and $\pi^0_{y,ij,k}$, and we use the notation $\Lambda^{(\ell)}_{y,ij,k} = \Lambda^{(\ell)}(\pi^0_{y,ij,k})$, which corresponds to $\Lambda^{(\ell)}_{ij,k}(y)$ in the main text. By appropriately inserting (A.9) and (A.10) into this expansion, also using (A.8), and sorting by terms linear in s_y, quadratic in s_y, and remainder, we find

$$\sqrt{n}\left(\widehat{F}_{y,k} - F_{y,k}\right) = T^{(1,F)}_{y,k} + T^{(2,F)}_{y,k} + r^{(F)}_{y,k}, \tag{A.11}$$

where the terms linear in s_y read

$$T^{(1,F)}_{y,k} = -\frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{y,ij,k} \left[\frac{\left(Q^{(\mathrm{FE})}_y s_y\right)_{ij}}{\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}} + \widetilde{\mathbb{x}}'_{y,ij,k} W_y^{-1} \frac{1}{n} \sum_{(i',j')\in\mathcal{D}} \widetilde{x}_{y,i'j'}\, \partial_\pi \ell_{y,i'j'}\right],$$

with $\partial_\pi \ell_{y,i'j'} = \left(\Lambda^{(1)}_{y,i'j'}\right)^{1/2} s_{y,i'j'}$.

The projection $\Psi_{y,ij,k} = \Psi_{ij,k}(y)$, defined just before (6) in the main text, can be written in terms of the matrix $Q^{(\mathrm{FE})}_y$ as

$$\Psi_{y,ij,k} = \left(\Lambda^{(1)}_{y,ij}\right)^{-1/2} \sum_{(i',j')\in\mathcal{D}} Q^{(\mathrm{FE})}_{y,ij,i'j'} \frac{\Lambda^{(1)}_{y,i'j',k}}{\left(\Lambda^{(1)}_{y,i'j'}\right)^{1/2}}, \tag{A.12}$$

which implies that $\sum_{(i,j)\in\mathcal{D}} \Psi_{y,ij,k}\partial_\pi \ell_{y,ij} = \sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{y,ij,k}\left(\Lambda^{(1)}_{y,ij}\right)^{-1/2}\left(Q^{(\mathrm{FE})}_y s_y\right)_{ij}$. Using $\partial_\beta F_{y,k} = \partial_\beta F_k(y) = n^{-1}\sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{ij,k}(y)\,\widetilde{\mathbb{x}}'_{y,ij,k}$ we obtain

$$T^{(1,F)}_{y,k} = -\frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} \left(\Psi_{y,ij,k} + \partial_\beta F_{y,k} W_y^{-1} \widetilde{x}_{y,ij}\right) \partial_\pi \ell_{y,ij}. \tag{A.13}$$

According to part (iii) of Lemma 6 the vector $T^{(1,F)}_y = \left[T^{(1,F)}_{y,k} \,:\, k \in \mathcal{K}\right]$ therefore satisfies $T^{(1,F)}_y \rightsquigarrow Z^{(F)}_y$ asymptotically.

The terms quadratic in s_y read

$$T_{y,k}^{(2,F)} = -\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\Lambda_{y,ij,k}^{(1)}\left(\Lambda_{y,ij}^{(1)}\right)^{-1/2}\sum_{(i',j')\in\mathcal{D}}Q_{y,ij,i'j'}^{(\mathrm{FE})}\frac{\Lambda_{y,i'j'}^{(2)}}{\left(\Lambda_{y,i'j'}^{(1)}\right)^{3/2}}\left[(Q_y s_y)_{i'j'}\right]^2$$
$$+\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\frac{\Lambda_{y,ij,k}^{(2)}}{\Lambda_{y,ij}^{(1)}}\left[(Q_y s_y)_{ij}\right]^2 + (\partial_\beta F_{y,k})T_y^{(2,\beta)},$$

where for the term quadratic in $\widehat{\pi}_{y,ij,k} - \pi_{y,ij,k}^0$ in the expansion of $\widehat{F}_{y,k} - F_{y,k}$ we do not insert (A.10) but rather insert (A.9), and we ignore the terms involving $\widehat{\beta}_y - \beta_y^0$ here — they give contributions quadratic in the score s_y, but only of smaller order, and we therefore rather include those in the remainder term $r_{y,k}^{(F)}$ below. Using again (A.12) we find

$$T_{y,k}^{(2,F)} = \frac{1}{2}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\frac{\Lambda_{y,ij,k}^{(2)} - \Lambda_{y,ij}^{(2)}\Psi_{y,ij,k}}{\Lambda_{y,ij}^{(1)}}\left[(Q_y s_y)_{ij}\right]^2 + (\partial_\beta F_{y,k})T_y^{(2,\beta)}.$$

Using part (v) of Lemma 6, and our previous result for $T_y^{(2,\beta)}$, we thus obtain

$$T_{y,k}^{(2,F)} - \frac{I}{n^{1/2}}\left[B_{y,k}^{(\Lambda)} + (\partial_\beta F_{y,k})B_y^{(\beta)}\right] - \frac{J}{n^{1/2}}\left[D_{y,k}^{(\Lambda)} + (\partial_\beta F_{y,k})D_y^{(\beta)}\right] \to_P 0, \tag{A.14}$$

uniformly in $y \in \mathcal{Y}$ and $k \in \mathcal{K}$.

The remainder term of the expansion reads

$$r_{y,k}^{(F)} = n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\Bigg\{\Lambda_{y,ij,k}^{(1)}\left(\Lambda_{y,ij}^{(1)}\right)^{-1/2}\left(Q^{(\mathrm{FE})}r_y\right)_{ij} + n^{-1/2}\Lambda_{y,ij,k}^{(1)}\widetilde{\mathbb{x}}_{y,ij,k}'r_y^{(\beta)}$$
$$+\frac{1}{8}\Lambda_{y,ij,k}^{(2)}\left(\Lambda_{y,ij}^{(1)}\right)^{-1}\left[\textstyle\sum_{(i',j')\in\mathcal{D}}Q_{y,ij,i'j'}\left(\Lambda_{y,i'j'}^{(1)}\right)^{-3/2}\Lambda_{y,i'j'}^{(2)}(Q_y s_y)_{i'j'}^2\right]^2$$
$$+\frac{1}{2}\Lambda_{y,ij,k}^{(2)}\left(\Lambda_{y,ij}^{(1)}\right)^{-1}(r_{y,ij})^2 + \frac{1}{2}\Lambda_{y,ij,k}^{(2)}\left[(\mathbb{x}_{ij,k} - x_{ij})'(\widehat{\beta}_y - \beta_y^0)\right]^2$$
$$+\frac{1}{6}\Lambda^{(3)}(\widetilde{\pi}_{y,ij,k})\left[\widehat{\pi}_{y,ij} - \widehat{\pi}_{y,ij} + (\mathbb{x}_{ij,k} - x_{ij})'(\widehat{\beta}_y - \beta_y^0)\right]^3\Bigg\}.$$

Our assumptions guarantee that $\Lambda_{y,ij}^{(\ell)}$ and $\Lambda_{y,ij,k}^{(\ell)}$, $\ell \in \{1,2,3\}$, and $\left(\Lambda_{y,ij}^{(1)}\right)^{-1}$ are all uniformly bounded. Lemma 3 guarantees that $r_{y,ij} = o_P(n^{-1/2})$, uniformly over y, i, j, and using Lemma 5(ii) this also implies that $\left(Q^{(\mathrm{FE})}r_y\right)_{ij} = o_P(n^{-1/2})$, uniformly over y, i, j. Above we have shown $r_y^{(\beta)} = o_P(1)$, uniformly over y. Lemma 5(ii) and Lemma 6(i) imply that

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left[\textstyle\sum_{(i',j')\in\mathcal{D}}Q_{y,ij,i'j'}\left(\Lambda_{y,i'j'}^{(1)}\right)^{-3/2}\Lambda_{y,i'j'}^{(2)}(Q_y s_y)_{i'j'}^2\right]^2 = o_P(n^{-1+1/3}) = o_P(n^{-1/2}).$$

Our asymptotic result for $\widehat{\beta}_y$ from part 1 of this proof guarantees that $\sup_{y\in\mathcal{Y}} \|\widehat{\beta}_y - \beta_y^0\|^2 = o_P(n^{-1/2})$. Lemma 3 together with Lemma 5(ii) and Lemma 6(i) guarantee that $\widehat{\pi}_{y,ij} - \pi_{y,ij}^0 = o_P(n^{-1/6})$, uniformly over y, i, j. We thus find, uniformly over $y \in \mathcal{Y}$ and $k \in \mathcal{K}$,

$$
\begin{aligned}
&\left| r_{y,k}^{(F)} \right| \\
&\leq \frac{1}{\sqrt{n}} \underbrace{\left[\sum_{(i,j)\in\mathcal{D}} \left| \frac{\Lambda_{y,ij,k}^{(1)}}{\left(\Lambda_{y,ij}^{(1)}\right)^{1/2}} \right| \right]}_{=O_P(n)} \underbrace{\left[\max_{(i,j)\in\mathcal{D}} \left| \left(Q^{(\mathrm{FE})} r_y \right)_{ij} \right| \right]}_{=o_P(n^{-1/2})} + \underbrace{\left(\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \left\| \Lambda_{y,ij,k}^{(1)} \widetilde{\mathbb{x}}_{y,ij,k} \right\| \right)}_{=O_P(1)} \underbrace{\left\| r_y^{(\beta)} \right\|}_{=o_P(1)} \\
&+ \frac{1}{8\sqrt{n}} \underbrace{\left(\sum_{(i,j)\in\mathcal{D}} \left| \frac{\Lambda_{y,ij,k}^{(2)}}{\Lambda_{y,ij}^{(1)}} \right| \right)}_{=O_P(n)} \underbrace{\left\{ \max_{(i,j)\in\mathcal{D}} \left[\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'} \left(\Lambda_{y,i'j'}^{(1)}\right)^{-3/2} \Lambda_{y,i'j'}^{(2)} \left(Q_y s_y\right)_{i'j'}^2 \right]^2 \right\}}_{=o_P(n^{-1/2})} \\
&+ \frac{1}{2\sqrt{n}} \underbrace{\left(\sum_{(i,j)\in\mathcal{D}} \left| \frac{\Lambda_{y,ij,k}^{(2)}}{\Lambda_{y,ij}^{(1)}} \right| \right)}_{=O_P(n)} \underbrace{\left[\max_{(i,j)\in\mathcal{D}} \left(r_{y,ij}\right)^2 \right]}_{=o_P(n^{-1})} + \frac{1}{2\sqrt{n}} \underbrace{\sum_{(i,j)\in\mathcal{D}} \left| \Lambda_{y,ij,k}^{(2)} \right| \|\mathbb{x}_{ij,k} - x_{ij})\|^2}_{=O_P(n)} \underbrace{\left\| \widehat{\beta}_y - \beta_y^0 \right\|^2}_{=o_P(n^{-1/2})} \\
&+ \frac{4}{3\sqrt{n}} \underbrace{\left(\sum_{(i,j)\in\mathcal{D}} \left| \Lambda^{(3)}(\widetilde{\pi}_{y,ij,k}) \right| \right)}_{=O_P(n)} \left\{ \underbrace{\max_{(i,j)\in\mathcal{D}} \left| \widehat{\pi}_{y,ij} - \widehat{\pi}_{y,ij} \right|^3}_{=o_P(n^{-1/2})} + \underbrace{\max_{(i,j)\in\mathcal{D}} \|\mathbb{x}_{ij,k} - x_{ij}\|^3}_{=O_P(1)} \underbrace{\left\| \widehat{\beta}_y - \beta_y^0) \right\|^3}_{=o_P(n^{-1/2})} \right\},
\end{aligned}
$$

and therefore

$$
\sup_{y\in\mathcal{Y}, k\in\mathcal{K}} \left| r_{y,k}^{(F)} \right| = o_P(1). \tag{A.15}
$$

Combing (A.11), (A.13), (A.14) and (A.15) gives the statement for $\widehat{F}(y) - F(y)$ in the theorem. ■

Proof of Theorem 2. Let

$$
\widetilde{x}_y(\pi_y) = x_y - \left[w^{(2)}, w^{(3)} \right] \left(\left[w^{(2)}, w^{(3)} \right]' \Lambda^{(1)}(\pi_y) \left[w^{(2)}, w^{(3)} \right] \right)^{\dagger} \left[w^{(2)}, w^{(3)} \right]' \left(\Lambda^{(1)}(\pi_y) \right) x_y,
$$

and

$$
W_y(\pi_y) = \frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}(\pi_{y,ij})\, \widetilde{x}_{y,ij}(\pi_y)\, \widetilde{x}_{y,ij}'(\pi_y),
$$

and[20]

$$B_y^{(\beta)}(\pi_y) = -\frac{1}{2} W_y(\pi_y)^{-1} \left[\frac{1}{I} \sum_{i=1}^{I} \frac{J^{-1} \sum_{j \in \mathcal{D}_i} \Lambda^{(2)}(\pi_{y,ij})\, \widetilde{x}_{y,ij}(\pi_y)}{J^{-1} \sum_{j \in \mathcal{D}_i} \Lambda^{(1)}(\pi_{y,ij})} \right].$$

Then we can write $B_y^{(\beta)} = B_y^{(\beta)}(\pi_y^0)$ and $\widehat{B}_y^{(\beta)} = B_y^{(\beta)}(\widehat{\pi}_y)$. The consistency result for $\widehat{B}_y^{(\beta)} = \widehat{B}^{(\beta)}(y)$ follows from an expansion of $= B_y^{(\beta)}(\widehat{\pi}_y)$ in $\widehat{\pi}_y$ around π_y^0. $\Lambda^{(1)}(\pi_{y,ij})$ and $\Lambda^{(2)}(\pi_{y,ij})$ and $\left(\Lambda^{(1)}(\pi_{y,ij})\right)^{-1}$ are all uniformly bounded over $\pi_{y,ij} \in [\pi_{\min}, \pi_{\max}]$, for any bounded interval $[\pi_{\min}, \pi_{\max}]$. Using this one obtains

$$b_n := \sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \sup_{\pi_y \in [\pi_{\min}, \pi_{\max}]^n} \left\| \frac{\partial B_y^{(\beta)}(\pi_y)}{\partial \pi_{y,ij}} \right\| = O_P(n^{-1}),$$

because any individual $\pi_{y,ij}$ only enters via an appropriately normalized sample average into $B_y^{(\beta)}(\pi_y)$. Indeed for any function of the form

$$B(\pi_y) = \frac{1}{I} \sum_{i=1}^{I} \frac{J^{-1} \sum_{j \in \mathcal{D}_i} f_1(\pi_{y,ij})}{J^{-1} \sum_{j \in \mathcal{D}_i} f_2(\pi_{y,ij})},$$

where f_1 and f_2 are differentiable with bounded derivatives f_1' and f_2',

$$\frac{\partial B(\pi_y)}{\partial \pi_{y,ij}} = \frac{1}{IJ} \frac{f_1'(\pi_{y,ij}) J^{-1} \sum_{j' \in \mathcal{D}_i} f_2(\pi_{y,ij'}) - J^{-1} \sum_{j' \in \mathcal{D}_i} f_1(\pi_{y,ij'}) f_2'(\pi_{y,ij})}{\left[J^{-1} \sum_{j' \in \mathcal{D}_i} f_2(\pi_{y,ij'}) \right]^2} = O_P(n^{-1}).$$

Lemma 3 together with Lemma 5(ii) and Lemma 6(i) guarantee that

$$\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \left| \widehat{\pi}_{y,ij} - \pi_{y,ij}^0 \right| = o_P(n^{-1/6}) = o_P(1).$$

By a mean value of expansion in $\widehat{\pi}_y$ around π_y^0 we thus obtain

$$\sup_{y \in \mathcal{Y}} \left\| B_y^{(\beta)}(\widehat{\pi}_y) - B_y^{(\beta)}(\pi_y^0) \right\| \leq b_n \sum_{(i,j) \in \mathcal{D}} \left| \widehat{\pi}_{y,ij} - \pi_{y,ij}^0 \right| = o_P(1).$$

The proof of consistency for the other estimators in Theorem 2 is analogous. ■

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014). Finite population causal standard errors.

[20]Note that instead of $B_y^{(\beta)}(\pi_y)$ we could simply write $B^{(\beta)}(\pi_y)$ here, because all the dependence on y is through the parameter π_y. The only reason to write $B_y^{(\beta)}(\pi_y)$ is to avoid confusion with the notation $B^{(\beta)}(y)$ in the main text.

Abrevaya, J. and C. M. Dahl (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics 26*(4), 379–397.

Anderson, J. E. and E. van Wincoop (2003, March). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*(1), 170–192.

Arellano, M. and S. Bonhomme (2016). Nonlinear panel data estimationvia quantile regressions. *Unpublished manuscript*.

Arellano, M. and M. Weidner (2016). Instrumental variable quantile regressions in large panels with fixed effects. *Unpublished manuscript*.

Cameron, A. C. and D. L. Miller (2014). Robust inference for dyadic data.

Chernozhukov, V., D. Chetverikov, and K. Kato (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Process. Appl. 126*(12), 3632–3651.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*(3), 559–575.

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

Chernozhukov, V., I. Fernandez-Val, S. Hoderlein, H. Holzmann, and W. Newey (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics 188*(2), 378–392.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Chernozhukov, V., I. Fernández-Val, B. Melly, and K. Wüthrich (2016, August). Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes. *ArXiv e-prints*.

Cruz-Gonzalez, M., I. Fernandez-Val, and M. Weidner (2016, October). probitfe and logitfe: Bias corrections for probit and logit models with two-way fixed effects. *ArXiv e-prints*.

Dzemski, A. (2017). An empirical model of dyadic link formation in a network with unobserved heterogeneity. *Unpublished manuscript*.

Eaton, J. and S. Kortum (2001). Trade in capital goods. *European Economic Review 45*(7), 1195–1235.

Fernandez-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics 192*(1), 291–312.

Galvao, A. F., C. Lamarche, and L. R. Lima (2013). Estimation of censored quantile regression for panel data with fixed effects. *Journal of the American Statistical Association 108*(503), 1075–1089.

Giné, E. and J. Zinn (1984). Some limit theorems for empirical processes. *Ann. Probab. 12*(4), 929–998. With discussion.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation.

Graham, B. S. (2017). An econometric model of link formation with degree heterogeneity. *Econometrica 85*(4), 1033–1063.

Graham, B. S., J. Hahn, A. Poirier, and J. L. Powell (2015). Quantile regression with panel data.

Graham, B. S., J. Hahn, and J. L. Powell (2009). The incidental parameter problem in a non-differentiable panel data model. *Economics Letters 105*(2), 181–182.

Harrigan, J. (1994). Scale economies and the volume of trade. *The Review of Economics and Statistics*, 321–328.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. Volume 4, Chapter 3, pp. 131–195. Handbook of International Economics.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics 123*(2), 441–487.

Kato, K. and A. Galvao (2016). Smoothed quantile regression for panel data. Technical Report 1.

Kato, K., A. F. Galvao, and G. V. Montes-Rojas (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics 170*(1), 76–91.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics 157*(2), 396–408.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32.

Rosen, A. M. (2012). Set identification via quantile restrictions in short panels. *Journal of Econometrics 166*(1), 127–137.

Santos Silva, J. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

van der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak Convergence and Empirical Processes*, pp. 16–28. Springer.

Yan, T., B. Jiang, S. E. Fienberg, and C. Leng (2016). Statistical inference in a directed network model with covariates. *arXiv preprint arXiv:1609.04558*.

Supplementary Appendix

The following proof of Lemma 3 also relies on the results of Lemma 5 and Lemma 6(i), whose proof is presented afterwards, without using Lemma 3 of course.

Proof of Lemma 3. Define $Q_y^\perp := \mathbb{I}_n - Q_y$, which is the $n \times n$ symmetric idempotent matrix that projects onto the space orthogonal to the column span of $\left(\Lambda_y^{(1)}\right)^{1/2} w$, with $w = (w_{ij} : (i,j) \in \mathcal{D})$. In component notation we have $Q^\perp_{y,ij,i'j'} = \delta_{ii'}\delta_{jj'} - Q_{y,ij,i'j'}$, where $\delta_{..}$ refers to the Kronecker delta. We also define

$$\pi^*_{y,ij} := \left(\Lambda^{(1)}_{y,ij}\right)^{1/2} \pi_{y,ij}, \quad \pi^{*0}_{y,ij} := \left(\Lambda^{(1)}_{y,ij}\right)^{1/2} \pi^0_{y,ij}, \quad \ell^*_{y,ij}(\pi^*_{y,ij}) := \ell_{y,ij}\left[\left(\Lambda^{(1)}_{y,ij}\right)^{-1/2} \pi^*_{y,ij}\right],$$

which is simply a rescaling of $\pi_{y,ij}$ by $\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}$. The rescaling is infeasible, because $\left(\Lambda^{(1)}_{y,ij}\right)^{1/2}$ depends on the true parameter values, but for the analysis here it is more convenient to work with $\ell^*_{y,ij}(\pi^*_{y,ij})$ than with $\ell_{y,ij}(\pi_{y,ij})$. After the rescaling we have $s_{y,ij} = \partial_{\pi^*}\ell^*_{y,ij} := \partial_{\pi^*}\ell^*_{y,ij}\left(\pi^{*0}_{y,ij}\right)$ and $1 = \partial_{\pi^{*2}}\ell^*_{y,ij} := \partial_{\pi^{*2}}\ell^*_{y,ij}\left(\pi^{*0}_{y,ij}\right)$, that is, the variance of the score and the Hessian of $\ell^*_{y,ij}(\pi^*_{y,ij})$ evaluated at the true parameter values are normalized to one. Equation (A.2) can be rewritten as $Q_y^\perp \pi^*_y = 0$. where π^*_y is the n-vector with elements $\pi^*_{y,ij}$. Solving (A.1) is then equivalent to minimizing the function

$$\sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}(\pi^*_{y,ij}) + \sum_{(i,j)\in\mathcal{D}} \pi^*_{y,ij} \sum_{(i',j')\in\mathcal{D}} \left[Q^\perp_{y,ij,i'j'}\, \mu_{y,i'j'}\right]$$

over π^*_y and μ_y, where the μ_y are the Lagrange multipliers corresponding to the constraint $Q_y^\perp \pi^*_y = 0$, which is equivalent to existence of θ such that $\pi_y = w\ \theta_y$. The FOCs with respect to π^*_y read

$$\partial_{\pi^*}\ell^*_y(\widehat{\pi}^*_y) + Q_y^\perp\, \widehat{\mu}_y = 0,$$

where $\partial_{\pi^*}\ell^*_y(\widehat{\pi}^*_y)$ and $\widehat{\mu}_y$ are n-vectors obtained by stacking the elements of $\partial_{\pi^*}\ell^*_{y,ij}(\widehat{\pi}^*_y)$ and $\widehat{\mu}_{y,ij}$ for all $(i,j) \in \mathcal{D}$. Existence of $\widehat{\mu}_y$ that satisfy those FOCs is equivalent to

$$Q_y\partial_{\pi^*}\ell^*_y(\widehat{\pi}^*_y) + Q_yQ_y^\perp\, \widehat{\mu}_y = Q_y\partial_{\pi^*}\ell^*_y(\widehat{\pi}^*_y) = \sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\, \partial_{\pi^*}\ell^*_{y,i'j'}(\widehat{\pi}^*_{y,i'j'}) = 0. \tag{S.1}$$

In addition to this first order condition we have the constraint $Q_y^\perp \widehat{\pi}^*_y = 0$, which implies that $\widehat{\pi}^*_y = Q_y\xi_y$ for some $\xi_y \in \mathbb{R}^n$, that is, we only need to consider parameters π^*_y that can be represented as $Q_y\xi_y$. In the following we perform three expansion steps for the log-likelihood function (or for the corresponding score function), each time restricting $\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}$ further.

Step 1: We assume uniform boundedness of all parameters and variables that enter into the single index. Therefore, for all $y \in \mathcal{Y}$ and $(i,j) \in \mathcal{D}$ we have $\pi^0_{y,ij} \in [\pi_{\min}, \pi_{\max}]$, where $[\pi_{\min}, \pi_{\max}]$ is some bounded interval. By strict convexity of minus the logistic log-likelihood function there exist constants $c_{\min}$ and $c_{\max}$ such that $0 < c_{\min} \leq \Lambda^{(1)}_{y,ij} \leq c_{\max} < \infty$ for all $y \in \mathcal{Y}$ and $(i,j) \in \mathcal{D}$. Hence, $\pi^{*0}_{y,ij} \in [c^{1/2}_{\min}\pi_{\min}, c^{1/2}_{\max}\pi_{\max}]$, for all $y \in \mathcal{Y}$ and $(i,j) \in \mathcal{D}$. Define $\Pi_{\text{bnd}} := [c^{1/2}_{\min}\pi_{\min} - \epsilon, c^{1/2}_{\max}\pi_{\max} + \epsilon]$, where $\epsilon > 0$ is an arbitrary finite constant. In the following we only need to consider values of $\pi^*_{y,ij}$ inside Π_{bnd}.

Because Π_{bnd} is bounded and $\ell^*_{y,ij}(\pi^*_{y,ij})$ is smooth we know that all the derivatives of $\ell^*_{y,ij}(\pi^*_{y,ij})$ are uniformly bounded inside Π_{bnd}. In particular, there exists a finite constant b such that, for $k \in \{1,2,3\}$,

$$\sup_{\pi \in \Pi_{\text{bnd}}} \sup_{y \in \mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left|\partial_{\pi^{*k}} \ell^*_{y,ij}(\pi)\right| \leq b.$$

By a third order expansion of $\pi^*_{y,ij} \mapsto \ell^*_{y,ij}(\pi^*_{y,ij})$ around $\pi^{*0}_{y,ij}$ we find

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}(\pi^*_{y,ij}) - \sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}(\pi^{*0}_{y,ij}) \\
&= \sum_{(i,j)\in\mathcal{D}} s_{y,ij}\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right) + \frac{1}{2}\sum_{(i,j)\in\mathcal{D}} \left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right)^2 + \frac{1}{6}\sum_{(i,j)\in\mathcal{D}} \left(\partial_{\pi^{*3}} \ell^*_{y,ij}(\underline{\pi}^*_{y,ij})\right)\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right)^3 \\
&\geq \sum_{(i,j)\in\mathcal{D}} s_{y,ij}\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right) + \frac{1}{2}\sum_{(i,j)\in\mathcal{D}} \left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right)^2 - \frac{b}{6}\sum_{(i,j)\in\mathcal{D}} \left|\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right|^3, \qquad \text{(S.2)}
\end{aligned}$$

where $\underline{\pi}^*_{y,ij}$ is an intermediate values between $\pi^*_{y,ij}$ and $\pi^{*0}_{y,ij}$. Analogously,

$$\begin{aligned}
&\sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}(\pi^*_{y,ij}) - \sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}(\pi^{*0}_{y,ij}) \\
&\leq \sum_{(i,j)\in\mathcal{D}} s_{y,ij}\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right) + \frac{1}{2}\sum_{(i,j)\in\mathcal{D}} \left(\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right)^2 + \frac{b}{6}\sum_{(i,j)\in\mathcal{D}} \left|\pi^*_{y,ij} - \pi^{*0}_{y,ij}\right|^3, \qquad \text{(S.3)}
\end{aligned}$$

Evaluating (S.2) at $\pi^*_{y,ij} = \pi^{*0}_{y,ij} - (Q_y s_y)_{ij} + (Q_y \zeta_y)_{ij}$, and (S.3) at $\pi^*_{y,ij} = \pi^{*0}_{y,ij} - (Q_y s_y)_{ij}$ gives

$$
\begin{aligned}
&\sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}\left[\pi^{*0}_{y,ij} - (Q_y s_y)_{ij} + (Q_y\zeta_y)_{ij}\right] - \sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}\left[\pi^{*0}_{y,ij} - (Q_y s_y)_{ij}\right] \\
&\geq \sum_{(i,j)\in\mathcal{D}} s_{y,ij}\left[(Q_y\zeta_y)_{ij} - (Q_y s_y)_{ij}\right] + \frac{1}{2}\sum_{(i,j)\in\mathcal{D}}\left[(Q_y\zeta_y)_{ij} - (Q_y s_y)_{ij}\right]^2 - \frac{b}{6}\sum_{(i,j)\in\mathcal{D}}\left|(Q_y\zeta_y)_{ij} - (Q_y s_y)_{ij}\right|^3 \\
&\quad + \sum_{(i,j)\in\mathcal{D}} s_{y,ij}\,(Q_y s_y)_{ij} - \frac{1}{2}\sum_{(i,j)\in\mathcal{D}}(Q_y s_y)^2_{ij} - \frac{b}{6}\sum_{(i,j)\in\mathcal{D}}\left|(Q_y s_y)_{ij}\right|^3 \\
&= \frac{1}{2}\sum_{(i,j)\in\mathcal{D}}\left[(Q_y\zeta_y)^2_{ij} - \frac{b}{3}\left|(Q_y\zeta_y)_{ij} - (Q_y s_y)_{ij}\right|^3 - \frac{b}{3}\left|(Q_y s_y)_{ij}\right|^3\right] \\
&\geq \frac{1}{2}\sum_{(i,j)\in\mathcal{D}}\left[(Q_y\zeta_y)^2_{ij} - \frac{4b}{3}\left|(Q_y\zeta_y)_{ij}\right|^3 - \frac{5b}{3}\left|(Q_y s_y)_{ij}\right|^3\right] \\
&= \frac{1}{2}\sum_{(i,j)\in\mathcal{D}}\left\{(Q_y\zeta_y)^2_{ij}\left[1 - \frac{4b}{3}\left|(Q_y\zeta_y)_{ij}\right|\right] - \frac{5b}{3}\left|(Q_y s_y)_{ij}\right|^3\right\}, \qquad \text{(S.4)}
\end{aligned}
$$

where we also used that $Q_y Q_y = Q_y$ and $\left|(Q_y\zeta_y)_{ij} - (Q_y s_y)_{ij}\right|^3 \leq 4\left|(Q_y\zeta_y)_{ij}\right|^3 + 4\left|(Q_y s_y)_{ij}\right|^3$. By the result of Lemma 6(i) we know that there exists a sequence $\kappa_n = o(1)$ such that wpa1

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|(Q_y s_y)_{ij}\right| \leq \kappa_n\, n^{-1/6},$$

which implies that

$$\sup_{y\in\mathcal{Y}}\sum_{(i,j)\in\mathcal{D}}\left|(Q_y s_y)_{ij}\right|^3 \leq n^{1/2}\,\kappa_n^3.$$

Consider the sets

$$
\begin{aligned}
\Pi^*_{y,n} &:= \left\{\pi^*_y \in \mathbb{R}^n \,:\, Q_y^{\perp}\widehat{\pi}^*_y = 0 \;\text{ and }\; \sum_{(i,j)\in\mathcal{D}}\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_y s_y)_{ij}\right)^2 \leq n^{1/2}\,\kappa_n^2\right\}, \\
\overline{\Pi}^*_{y,n} &:= \left\{\pi^*_y \in \mathbb{R}^n \,:\, Q_y^{\perp}\widehat{\pi}^*_y = 0 \;\text{ and }\; \sum_{(i,j)\in\mathcal{D}}\left(\pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_y s_y)_{ij}\right)^2 = n^{1/2}\,\kappa_n^2\right\}.
\end{aligned}
$$

Here, $\overline{\Pi}^*_{y,n}$ is the boundary of $\Pi^*_{y,n}$ within the set of all π^*_y that satisfy the constraint $Q_y^{\perp}\widehat{\pi}^*_y = 0$. For any $\zeta \in \mathbb{R}^n$ with $Q_y^{\perp}\zeta = 0$ we have $Q_y\zeta = \zeta$, and by applying Cauchy-Schwarz inequality we thus find

$$\begin{aligned}\|\zeta\|_\infty &:= \max_{(i,j)\in\mathcal{D}} |\zeta_{ij}| = \max_{(i,j)\in\mathcal{D}} \left| \sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\zeta_{i'j'} \right| \\ &\leq \max_{(i,j)\in\mathcal{D}} \left(\sum_{(i',j')\in\mathcal{D}} Q^2_{y,ij,i'j'} \right)^{1/2} \left(\sum_{(i',j')\in\mathcal{D}} \zeta^2_{i'j'} \right)^{1/2} \\ &= \max_{(i,j)\in\mathcal{D}} \left(Q_{y,ij,ij}\right)^{1/2} \|\zeta\| = O_P(n^{-1/4})\|\zeta\|, \end{aligned} \tag{S.5}$$

where we also used that $Q_yQ_y = Q_y$ and employed Lemma 5(iii). By applying (S.5) to $\zeta_{ij} = \pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij}$ we find that for $\pi^*_y \in \Pi^*_{y,n}$ we have

$$\sup_{y\in\mathcal{Y}} \sup_{\pi^*_y\in\Pi^*_{y,n}} \max_{(i,j)\in\mathcal{D}} \left| \pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij} \right| = O_P(\kappa_n),$$

and also using Lemma 6(i) we thus have

$$\sup_{y\in\mathcal{Y}} \sup_{\pi^*_y\in\Pi^*_{y,n}} \max_{(i,j)\in\mathcal{D}} \left| \pi^*_{y,ij} - \pi^{*0}_{y,ij} \right| = O_P(\kappa_n) + o_P(n^{-1/6}) = o_P(1). \tag{S.6}$$

Hence, when applying (S.4) to $\pi^*_y \in \Pi^*_{y,n}$ with $(Q_y\zeta_y)_{ij} = \pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij}$, then the term $\frac{4b}{3}\left|(Q_y\zeta_y)_{ij}\right|$ is of order $o_P(1)$ and the term $\left|(Q_ys_y)_{ij}\right|^3$ is of smaller order than $(Q_y\zeta_y)^2_{ij}$. In addition, note that $\pi^{*0}_{y,ij} - (Q_ys_y)_{ij} \in \Pi^*_{y,n}$. Thus, by applying (S.4) with $(Q_y\zeta_y)_{ij} = \pi^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij}$, and using (S.6) we find that with probability approaching one we have

$$\sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}\left(\pi^*_{y,ij}\right) - \sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}\left[\pi^{*0}_{y,ij} - (Q_ys_y)_{ij}\right] > 0, \qquad \text{for all } \pi^*_y \in \overline{\Pi}^*_{y,n}. \tag{S.7}$$

Thus, we have a convex set $\Pi^*_{y,n}$ such that the convex function $\pi^*_y \mapsto \sum_{(i,j)\in\mathcal{D}} \ell^*_{y,ij}\left(\pi^*_{y,ij}\right)$ takes a smaller value inside the set $\Pi^*_{y,n}$ than on any point of its boundary $\overline{\Pi}^*_{y,n}$ (within the set of all π^*_y that satisfy the constraint $Q_y^{\perp}\widehat{\pi}^*_y = 0$). This guarantees that the minimizer of the objective function needs to be inside the set $\Pi^*_{y,n}$, that is, we have $\widehat{\pi}_y \in \Pi^*_{y,n}$, which implies

$$\sup_{y\in\mathcal{Y}} \sum_{(i,j)\in\mathcal{D}} \left(\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij}\right)^2 \leq \kappa_n^2\, n^{1/2} = o_P(n^{1/2}),$$

and by the inequality (S.5) with $\zeta_{ij} = \widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} + (Q_ys_y)_{ij}$, and Lemma 6($i$), we find

$$\sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left| \widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} \right| = O_P(\kappa_n) + o_P(n^{-1/6}) = o_P(1). \tag{S.8}$$

Step 2: An expansion of (S.1) in $\widehat{\pi}^*_{y,i'j'}$ around $\pi^{*0}_{y,i'j'}$ up to second order yields

$$\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\left[s_{y,i'j'} + \left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right) + \frac{1}{2}\left(\partial_{\pi^{*3}}\ell^*_{y,i'j'}(\widetilde{\pi}^*_{y,i'j'})\right)\left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right)^2\right] = 0,$$

where $\widetilde{\pi}^*_{y,i'j'}$ is a value between $\pi^{*0}_{y,i'j'}$ and $\widehat{\pi}^*_{y,i'j'}$. By combining this expansion with the constraint $Q_y^{\perp}(\widehat{\pi}^*_y - \pi^{*0}_y) = 0$, which implies that $\widehat{\pi}^*_y - \pi^{*0}_y = Q_y\xi_y$, for some ξ_y, we obtain

$$\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} = -\left(Q_y s_y\right)_{ij} + r^{(1)}_{y,ij},$$

where

$$r^{(1)}_{y,ij} = -\frac{1}{2}\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\left(\partial_{\pi^{*3}}\ell^*_{y,i'j'}(\widetilde{\pi}^*_{y,i'j'})\right)\left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right)^2$$

Using our initial convergence rate result (S.8) in part 1 of this proof, and Lemma 5, and also uniform boundedness of all the derivatives of $\ell^*_{y,i'j'}(\pi^*)$ within Π_{bnd}, we find

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|r^{(1)}_{y,ij}\right| = o_P(n^{-1/2+1/6}) + O_P(\kappa_n^2).$$

Hence, by Lemma 6(i),

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}\right| = o_P(n^{-1/6}) + O_P(\kappa_n^2). \tag{S.9}$$

Using (S.9) instead of (S.8), and reapplying the same argument a second time we obtain

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}\right| = o_P(n^{-1/6}) + O_P(\kappa_n^4).$$

And by iterating this argument q-times we obtain

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}\right| = o_P(n^{-1/6}) + O_P(\kappa_n^{2q}).$$

for any positive integer q. Since $\kappa_n = o(1)$ we can choose q large enough such that

$$\sup_{y\in\mathcal{Y}}\max_{(i,j)\in\mathcal{D}}\left|\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}\right| = o_P(n^{-1/6}). \tag{S.10}$$

Step 3: An expansion of (S.1) in $\widehat{\pi}^*_{y,i'j'}$ around $\pi^{*0}_{y,i'j'}$ up to third order yields

$$\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'}\Bigg[s_{y,i'j'} + \left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right) + \frac{1}{2}\left(\partial_{\pi^{*3}}\ell^*_{y,i'j'}\right)\left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right)^2$$
$$+ \frac{1}{6}\left(\partial_{\pi^{*4}}\ell^*_{y,i'j'}(\overline{\pi}^*_{y,i'j'})\right)\left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right)^3\Bigg] = 0,$$

where $\overline{\pi}^*_{y,i'j'}$ is a value between $\pi^{*0}_{y,i'j'}$ and $\widehat{\pi}^*_{y,i'j'}$. By again using the constraint $\widehat{\pi}^*_y - \pi^{*0}_y = Q_y \xi_y$, for some ξ_y, we obtain

$$\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} = -\left(Q_y s_y\right)_{ij} - \frac{1}{2} \sum_{(i',j')\in\mathcal{D}} \underbrace{\left(\partial_{\pi^{*3}} \ell^*_{y,i'j'}\right)}_{=\frac{\partial_{\pi^3}\ell_{y,i'j'}}{\left(\Lambda^{(1)}_{y,i'j'}\right)^{3/2}}} Q_{y,ij,i'j'} \left[\left(Q_y s_y\right)_{i'j'}\right]^2 + r_{y,ij},$$

where

$$r_{y,ij} = -\sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'} \Bigg\{ \frac{1}{2} \left(\partial_{\pi^{*3}} \ell^*_{y,i'j'}\right) r^{(1)}_{y,i'j'} \left[2\left(Q_y s_y\right)_{i'j'} + r^{(1)}_{y,i'j'}\right] \\ + \frac{1}{6} \left(\partial_{\pi^{*4}} \ell^*_{y,i'j'}(\overline{\pi}^*_{y,i'j'})\right) \left(\widehat{\pi}^*_{y,i'j'} - \pi^{*0}_{y,i'j'}\right)^3 \Bigg\},$$

and therefore

$$\max_{(i,j)\in\mathcal{D}} |r_{y,ij}| \leq \left(\max_{(i,j)\in\mathcal{D}} \sum_{(i',j')\in\mathcal{D}} Q_{y,ij,i'j'} \right) \max_{(i,j)\in\mathcal{D}} \Bigg| \frac{1}{2} \left(\partial_{\pi^{*3}} \ell^*_{y,ij}\right) r^{(1)}_{y,ij} \left[2\left(Q_y s_y\right)_{ij} + r^{(1)}_{y,ij}\right] \\ + \frac{1}{6} \left(\partial_{\pi^{*4}} \ell^*_{y,ij}(\overline{\pi}^*_{y,ij})\right) \left(\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij}\right)^3 \Bigg|.$$

Thus, using (S.10), Lemma 5, and Lemma 6(i), and also uniform boundedness of all the derivatives of $\ell^*_{y,i'j'}(\pi^*)$ within Π_{bnd}, we thus find $\sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} |r_{y,ij}| = o_P(n^{-1/2})$. This gives the result of the lemma, since $\widehat{\pi}^*_{y,ij} - \pi^{*0}_{y,ij} = \left(\Lambda^{(1)}_{y,ij}\right)^{1/2} \left(\widehat{\pi}_{y,ij} - \pi^0_{y,ij}\right)$. ■

Proof of Lemma 4. We prove this lemma by showing that the smallest eigenvalue of W_y is bounded from below, uniformly over $y \in \mathcal{Y}$. We know that there exists $b_{\min} > 0$ such

that $\Lambda^{(1)}_{y,ij} \geq b_{\min}$, uniformly over y, i, j. Then,

$$\begin{aligned}
\lambda_{\min}(W_y) &= \min_{\|\delta\|=1} \delta' W_y \delta \\
&= \min_{\|\delta\|=1} \min_{(\alpha,\gamma)\in\mathbb{R}^{I+J}} \left[\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} \Lambda^{(1)}_{y,ij} \; (x'_{ij}\delta - \alpha_i - \gamma_j)^2 \right] \\
&\geq \min_{\|\delta\|=1} \min_{(\alpha,\gamma)\in\mathbb{R}^{I+J}} \left[\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} b_{\min} \; (x'_{ij}\delta - \alpha_i - \gamma_j)^2 \right] \\
&= b_{\min} \min_{\|\delta\|=1} \min_{(\alpha,\gamma)\in\mathbb{R}^{I+J}} \left[\frac{1}{n} \sum_{(i,j)\in\mathcal{D}} (x'_{ij}\delta - \alpha_i - \gamma_j)^2 \right] \\
&\geq b_{\min} \, c_3 > 0,
\end{aligned}$$

where existence of $c_3 > 0$ is guaranteed by Assumption 1(vi). ■

Proof of Lemma 5. We showed that $Q_y = Q_y^{(1)} + Q_y^{(\mathrm{FE})}$ in equation (A.3). We now want to find the bound on $Q_y^{(\mathrm{rem})} = Q_y^{(\mathrm{FE})} - Q_y^{(2)} - Q_y^{(3)}$ in part (i) of the lemma. Let

$$\mathcal{H}_y := \left[w^{(2)}, w^{(3)} \right]' \Lambda^{(1)}_y \left[w^{(2)}, w^{(3)} \right].$$

Then,

$$Q_y^{(\mathrm{FE})} := \left(\Lambda^{(1)}_y \right)^{1/2} \left[w^{(2)}, w^{(3)} \right] \mathcal{H}_y^{\dagger} \left[w^{(2)}, w^{(3)} \right]' \left(\Lambda^{(1)}_y \right)^{1/2},$$

where we use the Moore-Penrose pseudo-inverse †, because $\mathcal{H}_y$ has one zero-eigenvalue with corresponding eigenvector $v = (1_I, -1_J)$, that is, v is a column vector with I ones follows by J minus ones.[21] We can therefore write

$$\mathcal{H}_y^{\dagger} = \left[\mathcal{H}_y + vv'/(I+J) \right]^{-1} - vv'/(I+J).$$

The matrix $\mathcal{H}_y = \partial_{\phi_y \phi'_y} \sum_{(i,j)\in\mathcal{D}} \ell_{y,ij}(\pi^0_{y,ij})$ is simply the $(I+J)\times(I+J)$ Hessian matrix of minus the log-likelihood function with respect to all the fixed effects $\phi_y = (\alpha'_y, \gamma'_y)'$. We decompose $\mathcal{H}_y = \mathcal{D}_y + \mathcal{R}_y$, where

$$\mathcal{D}_y := w^{(2)\prime} \Lambda^{(1)}_y w^{(2)} + w^{(3)\prime} \Lambda^{(1)}_y w^{(3)} = \begin{pmatrix} \mathrm{diag}\left(\sum_{j\in\mathcal{D}_i} \Lambda^{(1)}_{y,ij} \right)_{i=1\ldots,I} & 0_{I\times J} \\ 0_{J\times I} & \mathrm{diag}\left(\sum_{i\in\mathcal{D}_j} \Lambda^{(1)}_{y,ij} \right)_{j=1\ldots,J} \end{pmatrix}$$

$$\mathcal{R}_y := w^{(2)\prime} \Lambda^{(1)}_y w^{(3)} + w^{(3)\prime} \Lambda^{(1)}_y w^{(2)} = \begin{pmatrix} 0_{I\times I} & A_y \\ A'_y & 0_{J\times J} \end{pmatrix},$$

[21] Note that the additively separable structure $\alpha_i(y) + \gamma_j(y)$ is invariant to adding a constant to all the $\alpha_i(y)$ and subtracting the same constant to all the $\gamma_j(y)$.

where A_y is the $I \times J$ matrix with entries $A_{y,ij} = \Lambda^{(1)}_{y,ij}$ if $(i,j) \in \mathcal{D}$, and zero otherwise. Because $\infty > b_{\max} \geq \Lambda^{(1)}_{y,ij} \geq b_{\min} > 0$, Lemma D.1 in Fernandez-Val and Weidner (2016) shows that this incidental parameter Hessian satisfies

$$\sup_{y \in \mathcal{Y}} \left\| \mathcal{H}^{\dagger}_y - \mathcal{D}^{-1}_y \right\|_{\max} = O_P(n^{-1}), \tag{S.11}$$

where $\|A\|_{\max}$ refers to the maximum over the absolute values of all the elements of the matrix A. Note that Lemma D.1 in Fernandez-Val and Weidner (2016) is for the "expected Hessian", but for our logit model we have $\mathcal{H}_y = \mathbb{E}\mathcal{H}_y$, conditional on regressors and fixed effects, so the distinction between Hessian and expected Hessian is irrelevant here. Also, in Fernandez-Val and Weidner (2016) the Hessian is no indexed by y, but the derivation of the bound there is in terms of global constants $b_{\min}$, $b_{\max}$ and thus holds uniformly over y. Finally, Fernandez-Val and Weidner (2016) does not allow for missing observations, but since we only allow for a finite number of missing observations for every i and j that can only have a negligible effect on the Hessian matrix.

We thus have

$$\begin{aligned} Q^{(\mathrm{FE})}_y &:= \left(\Lambda^{(1)}_y\right)^{1/2} \left[w^{(2)}, w^{(3)}\right] \mathcal{D}^{-1}_y \left[w^{(2)}, w^{(3)}\right]' \left(\Lambda^{(1)}_y\right)^{1/2} \\ &\quad + \left(\Lambda^{(1)}_y\right)^{1/2} \left[w^{(2)}, w^{(3)}\right] \left[\mathcal{H}^{\dagger}_y - \mathcal{D}^{-1}_y\right] \left[w^{(2)}, w^{(3)}\right]' \left(\Lambda^{(1)}_y\right)^{1/2} \\ &= Q^{(2)}_y + Q^{(3)}_y + Q^{(\mathrm{rem})}_y, \end{aligned}$$

where

$$Q^{(\mathrm{rem})}_y = \left(\Lambda^{(1)}_y\right)^{1/2} \left[w^{(2)}, w^{(3)}\right] \left[\mathcal{H}^{\dagger}_y - \mathcal{D}^{-1}_y\right] \left[w^{(2)}, w^{(3)}\right]' \left(\Lambda^{(1)}_y\right)^{1/2},$$

and therefore

$$\begin{aligned} \sup_{y \in \mathcal{Y}} \left\| Q^{(\mathrm{rem})}_y \right\|_{\max} &\leq \sup_{y \in \mathcal{Y}} \left\| \left(\Lambda^{(1)}_y\right)^{1/2} \right\|_{\max} \left\| \mathcal{H}^{\dagger}_y - \mathcal{D}^{-1}_y \right\|_{\max} \left\| \left(\Lambda^{(1)}_y\right)^{1/2} \right\|_{\max} \\ &= \sup_{y \in \mathcal{Y}} \left(\max_{(i,j) \in \mathcal{D}} \Lambda^{(1)}_{y,ij} \right) \left\| \mathcal{H}^{\dagger}_y - \mathcal{D}^{-1}_y \right\|_{\max} = O_P(n^{-1}), \end{aligned}$$

which can equivalently be written as $\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \max_{(i',j') \in \mathcal{D}} \left| Q^{(\mathrm{rem})}_{y,ij,i'j'} \right| = O_P(n^{-1})$.

Part (ii) and (iii) follow immediately from part (i) and the explicit formulas for the elements of $Q^{(1)}_y$, $Q^{(2)}_y$ and $Q^{(3)}_y$ in (A.5) and (A.6) above. ■

Intermediate results for the proof of of Lemma 6

The proof of Lemma 6 requires several intermediate results from the theory of stochastic processes, which are presented in the following. The notation $a_n \lesssim b_n$ means that $a_n \leq C\, b_n$

for some constant C that is independent of the sample size n. It is also convenient to define $\mathbf{I} := \{1, 2, \ldots, I\}$ and $\mathbf{J} := \{1, 2, \ldots, J\}$. In this section we assume that $\mathcal{Y}$ is a bounded interval, and that y_{ij} is continuously distributed with density bounded away from zero. The results for the case where y_{ij} is discrete follow directly from Fernandez-Val and Weidner (2016). Results for a mixed distribution of y_{ij} follow by combining the results for the continuous and discrete cases.

Bounds on sample averages over the score $\partial_\pi \ell_{y,ij}$. For every $i \in \mathbf{I}$ we define the empirical process

$$\mathbb{G}_{J,i} f := \frac{1}{\sqrt{|\mathcal{D}_i|}} \sum_{j \in \mathcal{D}_i} \left[f(y_{ij}) - \mathbb{E} f(y_{ij}) \right], \qquad f \in \mathcal{F} := \{\widetilde{y} \mapsto 1(\widetilde{y} \leq y) \ : \ y \in \mathcal{Y}\}. \quad \text{(S.12)}$$

Here, for ease of notation, we use the subscript J to denote the sample size, corresponding to the balanced panel case where $|\mathcal{D}_i| = J$. Following standard notation we write $\|\mathbb{G}_{J,i}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_{J,i} f|$. Every element of $\mathcal{F}$ correspond to exactly one $y \in \mathcal{Y}$, and in the following we write $f_y : \widetilde{y} \mapsto 1(\widetilde{y} \leq y)$ for that element. Since $\mathbb{E}\, 1\{y_{ij} \leq y\} = \Lambda_{y,ij}$,

$$\mathbb{G}_{J,i} f_y = -\frac{1}{\sqrt{|\mathcal{D}_i|}} \sum_{j \in \mathcal{D}_i} \partial_\pi \ell_{y,ij}, \quad \text{(S.13)}$$

where

$$\partial_\pi \ell_{y,ij} := \partial_\pi \ell_{y,ij}(\pi^0_{y,ij}) = \Lambda_{y,ij} - 1\{y_{ij} \leq y\}.$$

Our goal is to show that $\max_{i \in \mathbf{I}} \|\mathbb{G}_{J,i}\|_{\mathcal{F}} = o_P\left(I^{1/6}\right)$ by using the following theorem. The theorem uses standard notation $\mathbb{G}_n$ for the empirical process, denoting the sample size by n (not J or $|\mathcal{D}_i|$), and without an extra index i. The definition of $J(\delta, \mathcal{F})$ is given on p.239 of van der Vaart and Wellner (1996). All that matters to us is that $J(1, \mathcal{F})$ only depends on $\mathcal{F}$ (not on the probability measure or on the empirical process) and that for the $\mathcal{F}$ defined in (S.12) we have $J(1, \mathcal{F}) < \infty$, because $\mathcal{F}$ is VC class. An obvious envelope function for that $\mathcal{F}$ is $F : \tilde{y} \mapsto 1$, which satisfies $\|F\|_n = 1$. The minimal measurable majorant of $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ is denoted by $\|\mathbb{G}_n\|^*_{\mathcal{F}}$, and is identical to $\|\mathbb{G}_n\|_{\mathcal{F}}$ for our purposes.

Lemma S.1 (Restatement of Theorem 2.14.1 in van der Vaart and Wellner, 1996, for INID case). *Let $\mathcal{F}$ be a P-measurable class of measurable functions with measurable envelope F. Then, for $p \geq 2$,*

$$\mathbb{E}\left[(\|\mathbb{G}_n\|^*_{\mathcal{F}})^p\right] \lesssim J(1, \mathcal{F})^p \, \mathbb{E}\|F\|^p_n,$$

where $\|F\|_n$ is the $L_2(\mathbb{P}_n)$-seminorm and the inequality is valid up to a constant depending only on the p involved in the statement.

Proof. In van der Vaart and Wellner (1996) the theorem is stated for empirical processes from iid samples , but their proof relies only on symmetrization arguments (their Lemma 2.3.1) and sub-Gaussianity of the symmetrized process, which continue to hold for INID

samples that we consider here (our y_{ij} are conditionally independent, but not identically distributed). ■

Corollary S.1. *Under Assumption 1 we have*

$$\sup_{y\in\mathcal{Y}}\max_{i\in\mathbf{I}}\left|\frac{1}{\sqrt{|\mathcal{D}_i|}}\sum_{j\in\mathcal{D}_i}\partial_\pi\ell_{y,ij}\right| = o_P\left(n^{1/12}\right),$$

and

$$\sup_{y\in\mathcal{Y}}\max_{j\in\mathbf{J}}\left|\frac{1}{\sqrt{|\mathcal{D}_j|}}\sum_{i\in\mathcal{D}_j}\partial_\pi\ell_{y,ij}\right| = o_P\left(n^{1/12}\right).$$

Proof. The definition (S.12) implies (S.13), so we want to show $\max_{i\in\mathbf{I}}\|\mathbb{G}_{J,i}\|_{\mathcal{F}} = o_P\left(I^{1/6}\right)$. Applying Lemma S.1 for the function class $\mathcal{F}$with the envelope function $F:\tilde{y}\mapsto 1$ we find for $p\geq 1$,

$$\begin{aligned}\mathbb{E}\left(\max_{i\in\mathbf{I}}\|\mathbb{G}_{J,i}\|_{\mathcal{F}}\right)^p &= \mathbb{E}\max_{i\in\mathbf{I}}\left(\|\mathbb{G}_{J,i}\|_{\mathcal{F}}\right)^p \leq \mathbb{E}\sum_{i\in\mathbf{I}}\left(\|\mathbb{G}_{J,i}\|_{\mathcal{F}}\right)^p = \sum_{i\in\mathbf{I}}\mathbb{E}\left(\|\mathbb{G}_{J,i}\|_{\mathcal{F}}\right)^p \\ &\leq I\,J(1,\mathcal{F})^p = O(I),\end{aligned}$$

where $J(1,\mathcal{F})$ is a finite constant, independent of i, as noted earlier above. By Markov's inequality we thus find $\max_{i\in\mathbf{I}}\|\mathbb{G}_{J,i}\|_{\mathcal{F}} = O_P(I^{1/p})$. Choosing $p>6$ gives the desired result.

The second statement $\sup_{y\in\mathcal{Y}}\max_{j\in\mathbf{J}}\left|\frac{1}{\sqrt{|\mathcal{D}_j|}}\sum_{i\in\mathcal{D}_j}\partial_\pi\ell_{y,ij}\right| = o_P\left(n^{1/12}\right)$ can be shown analogously. ■

Bounds on weighted sample averages over the score $\partial_\pi\ell_{y,ij}$. We also need results on sample averages of the form e.g. $\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}} b^{(n)}_{y,hij}\,\partial_\pi\ell_{y,ij}$, where $b^{(n)}_{y,hij}$ are weights that also depend on the index y. The following lemma is useful for that purpose.

Lemma S.2. *Suppose $Z_1(t)$, …, $Z_n(t)$ are independent, stochastic processes indexed by $t\in T$ which are suitably measurable. Let B_i denote a measurable envelope of $\{Z_i(t), t\in T\}$. such that $\mathbb{E}B_i^p<\infty$ for $p\geq 1$. Let*

$$X_n(t) = \frac{1}{n}\sum_{i=1}^n Z_i(t), \quad \mathbb{E}X_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}Z_i(t) \quad t\in T.$$

Let B_i denote a measurable envelope of $\{Z_i(t), t\in T\}$. Let T be equipped with the pseudo-metric

$$d_n(t,t') = \sqrt{\frac{1}{n}\sum_{i=1}^n\left(Z_i(t)-Z_i(t')\right)^2}.$$

Let $N(\epsilon, T, d_n)$ denote the covering number of T under d_n balls of radius ϵ. Let

$$J_n(\delta, T) = \int_0^\delta \sqrt{1 + \log N(\epsilon \|B\|_n, T, d_n)} d\epsilon,$$

where

$$\|B\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n |B_i|^2}.$$

Then

$$\left\| \|X_n - \mathbb{E}X_n\|_T^* \right\|_{P,p} \lesssim \|J_n(1, T) \|B\|_n\|_{P,p}.$$

Proof. The proof is analogous to the proof of Theorem 2.14.1 in van der Vaart and Wellner (1996), p.239, with a few notational adjustments.

Let

$$X_n^o(t) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i Z_i(t), \quad t \in T,$$

denote the symmetrized version of X_n, where $\varepsilon = (\varepsilon_i)_{i=1}^n$ are independent Rademacher. By Lemma 2.3.6 in van der Vaart and Wellner (1996) the $L^p(P)$ norm of $\|X_n - \mathbb{E}X_n\|_T^*$ is bounded by the $L^p(P)$ norm of $2\|X_n^o\|_T^*$. Let P_ε denote the distribution of ε. Then by the standard argument, conditional on $(Z_i)_{i=1}^n$, X_n^o is sub-Gaussian with respect to d_n:

$$\left\| X_n^o(t) - X_n^0(t') \right\|_{\Psi_2(P_\varepsilon)} \lesssim d_n(t, t').$$

Hence by Corollary 2.2.5 in van der Vaart and Wellner (1996), we conclude

$$\left\| \|X_n^o\|_T \right\|_{\Psi_2(P_\varepsilon)} \leq \int_0^{\mathrm{diam}(T, d_n)} \sqrt{1 + \log N(\epsilon, T, d_n)} d\epsilon.$$

By a change of variables the right side is bounded by

$$\|B\|_n \int_0^{\mathrm{diam}(T, d_n)/\|B\|_n} \sqrt{1 + \log N(\epsilon \|B\|_n, T, d_n)} d\epsilon,$$

which is further bounded by

$$\|B\|_n \, J_n(1, T).$$

Every L_p-norm is bounded by a multiple of the Ψ_2-Orliczs norm. Hence

$$\mathbb{E}_\varepsilon \|X_n^o\|_T^p \lesssim (J_n(1, T) \|B\|_n)^p,$$

where $\mathbb{E}_\varepsilon$ is the expectation conditional on $(Z_i)_{i=1}^n$. Take expectations over $(Z_i)_{i=1}^n$ to obtain the lemma. ■

Using Lemma S.2 we obtain the following corollary.

Corollary S.2. *Let Assumption 1 hold. For $y \in \mathcal{Y}$, $h \in \{1, \dots, I+J\}$, $i \in \mathbf{I}$ and $j \in \mathbf{J}$, let $b_{y,hij}^{(n)}, c_{y,i}^{(n)}$, $d_{y,j}^{(n)}$ be real numbers, which can depend on the sample size n, and on the regressors and fixed effects, but not on the outcome variable, and assume that $\sup_{y\in\mathcal{Y}} \max_{h\in\{1,\dots,I+J\}} \max_{i\in\mathbf{I}} \max_{j\in\mathbf{I}} \max\left(\left|b_{y,hij}^{(n)}\right|, \left|\frac{\partial b_{y,hij}^{(n)}}{\partial y}\right|\right) = O_P(1)$, and also that $\sup_{y\in\mathcal{Y}} \max_{i\in\mathbf{I}} \max\left(\left|c_{y,i}^{(n)}\right|, \left|\frac{\partial c_{y,i}^{(n)}}{\partial y}\right|\right) = O_P(1)$, and $\sup_{y\in\mathcal{Y}} \max_{j\in\mathbf{I}} \max\left(\left|d_{y,j}^{(n)}\right|, \left|\frac{\partial d_{y,j}^{(n)}}{\partial y}\right|\right) = O_P(1)$. Then,*

$$(i) \quad \sup_{y\in\mathcal{Y}} \max_{h\in\{1,\dots,I+J\}} \left| \frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} b_{y,hij}^{(n)} \, \partial_\pi \ell_{y,ij} \right| = O_P\left(n^{1/6}\right),$$

and

$$(ii) \quad \sup_{y\in\mathcal{Y}} \left| \frac{1}{I} \sum_{i=1}^{I} c_{y,i}^{(n)} \left\{ \left(\frac{1}{\sqrt{|\mathcal{D}_i|}} \sum_{j\in\mathcal{D}_i} \partial_\pi \ell_{y,ij} \right)^2 - \mathbb{E}\left[\left(\frac{1}{\sqrt{|\mathcal{D}_i|}} \sum_{j\in\mathcal{D}_i} \partial_\pi \ell_{y,ij} \right)^2 \right] \right\} \right| = o_P(1),$$

$$\sup_{y\in\mathcal{Y}} \left| \frac{1}{J} \sum_{j=1}^{J} d_{y,j}^{(n)} \left\{ \left(\frac{1}{\sqrt{|\mathcal{D}_j|}} \sum_{i\in\mathcal{D}_j} \partial_\pi \ell_{y,ij} \right)^2 - \mathbb{E}\left[\left(\frac{1}{\sqrt{|\mathcal{D}_j|}} \sum_{i\in\mathcal{D}_j} \partial_\pi \ell_{y,ij} \right)^2 \right] \right\} \right| = o_P(1).$$

Proof. For part (i) we apply Lemma S.2 with $T = \mathcal{Y}$ and

$$Z_i(y) = b_{y,hij}^{(n)} \, \partial_\pi \ell_{y,ij},$$

for given $h \in \{1, \dots, I+J\}$, we can use constant envelope $B_i = B$ and the bound $J_n(1,T) \leq C$, which can be established using standard arguments, we find that $A_h = \sup_{y\in\mathcal{Y}} \left| \frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} b_{y,hij}^{(n)} \, \partial_\pi \ell_{y,ij} \right|$ satisfies $\max_{h\in\{1,\dots,I+J\}} \mathbb{E} A_h^4 = O_P(1)$. We therefore have

$$\mathbb{E}\left(\max_{h\in\{1,\dots,I+J\}} A_h \right)^4 \leq \mathbb{E}\left(\sum_{h=1}^{I+J} A_h^4 \right) \leq (I+J) \max_h \mathbb{E}\left(A_h^4\right) = O_P(n^{1/2}),$$

and therefore $\max_{h\in\{1,\dots,I+J\}} A_h = O_P(n^{1/6})$ as desired.

For the first result in part (ii), we apply Lemma S.2 with $T = \mathcal{Y}$ and

$$Z_i(y) := c_{y,i}^{(n)} \left[\left(\frac{1}{\sqrt{|\mathcal{D}_i|}} \sum_{j\in\mathcal{D}_i} \partial_\pi \ell_{y,ij} \right)^2 - \mathbb{E}\left(\frac{1}{\sqrt{|\mathcal{D}_j|}} \sum_{i\in\mathcal{D}_j} \partial_\pi \ell_{y,ij} \right)^2 \right].$$

Verification of the conditions of the lemma implies gives the desired result. We omit details in this version of the paper. The claim follows analogously for the second result in part (ii). ∎

FCLT for weighted sample averages over the score $\partial_\pi \ell_{y,ij}$. The following theorem will be used in the proof of part (ii) and (iii) of Lemma 6.

Lemma S.3 (**Theorem 2.11.11 in van der Vaart and Wellner, 1996**). *For each n, let $Z_{n1}, \ldots, Z_{n,m_n}$ be independent stochastic processes indexed by an arbitrary index set $\mathcal{F}$. Suppose that there exists a Gaussian-dominated semimetric ρ on $\mathcal{F}$ such that*

$$(i) \qquad \sum_{\ell=1}^{m_n} \mathbb{E}^* \left[\|Z_{n\ell}\|_{\mathcal{F}} \ 1\left\{ \|Z_{n\ell}\|_{\mathcal{F}} > \eta \right\} \right] \to 0, \qquad \text{for every } \eta > 0,$$

$$(ii) \qquad \sum_{\ell=1}^{m_n} \mathbb{E}\left(Z_{n\ell}(f) - Z_{n\ell}(g) \right)^2 \leq \rho^2(f,g), \qquad \text{for every } f, g \in \mathcal{F},$$

$$(iii) \qquad \sup_{t>0} \sum_{\ell=1}^{m_n} t^2 \, \mathbb{P}^* \left(\sup_{f,g \in \mathcal{B}(\varepsilon)} |Z_{n\ell}(f) - Z_{n\ell}(g)| > t \right) \leq \varepsilon^2,$$

for every ρ-ball $\mathcal{B}(\varepsilon) \subset \mathcal{F}$ of radius less than ε and for every n. Then the sequence $\sum_{\ell=1}^{m_n} (Z_{\ell,n} - \mathbb{E}\, Z_{\ell,n})$ is asymptotically tight in $\ell^\infty(\mathcal{F})$. It converges in distribution provided it converges marginally.

A semi-metric ρ is Gaussian-dominated if it is bounded above by a Gaussian semi-metric. Any semi-metric such that $\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, \rho)} d\epsilon < \infty$ is Gaussian dominated.

Proof of Lemma 6

Proof of Lemma 6, Part (i)**.** We have

$$\left(Q_y^{(2)} s_y \right)_{ij} = |\mathcal{D}_i|^{-1/2} \frac{\left(\Lambda_{y,ij}^{(1)} \right)^{1/2}}{|\mathcal{D}_i|^{-1} \sum_{j' \in \mathcal{D}_i} \Lambda_{y,ij'}^{(1)}} \left[|\mathcal{D}_i|^{-1/2} \sum_{j' \in \mathcal{D}_i} \partial_\pi \ell_{y,ij'} \right],$$

$$\left(Q_y^{(3)} s_y \right)_{ij} = |\mathcal{D}_j|^{-1/2} \frac{\left(\Lambda_{y,ij}^{(1)} \right)^{1/2}}{|\mathcal{D}_j|^{-1} \sum_{i' \in \mathcal{D}_j} \Lambda_{y,i'j}^{(1)}} \left[|\mathcal{D}_j|^{-1/2} \sum_{i' \in \mathcal{D}_j} \partial_\pi \ell_{y,i'j} \right]. \tag{S.14}$$

With $\max_i |\mathcal{D}_i|^{-1/2} = O_P(n^{-1/4})$, $\max_j |\mathcal{D}_j|^{-1/2} = O_P(n^{-1/4})$,

$$\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \left| \frac{\left(\Lambda_{y,ij}^{(1)} \right)^{1/2}}{|\mathcal{D}_i|^{-1} \sum_{j' \in \mathcal{D}_i} \Lambda_{y,ij'}^{(1)}} \right| = O_P(1), \quad \sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \left| \frac{\left(\Lambda_{y,ij}^{(1)} \right)^{1/2}}{|\mathcal{D}_j|^{-1} \sum_{i' \in \mathcal{D}_j} \Lambda_{y,i'j}^{(1)}} \right| = O_P(1),$$

we obtain by Corollary S.1 that

$$\sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \left| \left(Q_y^{(2)} s_y \right)_{ij} \right| = o_P(n^{-1/6}) \qquad \sup_{y \in \mathcal{Y}} \max_{(i,j) \in \mathcal{D}} \left| \left(Q_y^{(3)} s_y \right)_{ij} \right| = o_P(n^{-1/6}). \tag{S.15}$$

Next,

$$\left(Q_y^{(1)} s_y\right)_{ij} = \left(\Lambda_{y,ij}^{(1)}\right)^{1/2} \widetilde{x}'_{y,ij} W_y^{-1} \left[\frac{1}{n} \sum_{(i',j')\in\mathcal{D}} \widetilde{x}_{y,i'j'} \left(\partial_\pi \ell_{y,i'j'}\right)\right],$$

$$\begin{aligned}\left(Q_y^{(\mathrm{rem})} s_y\right)_{ij} &= \left\{\left(\Lambda_y^{(1)}\right)^{1/2} \left[w^{(2)}, w^{(3)}\right] \left(\mathcal{H}_y^\dagger - \mathcal{D}_y^{-1}\right) \left[w^{(2)}, w^{(3)}\right]' \left(\Lambda_y^{(1)}\right)^{1/2} s_y\right\}_{ij} \\ &= \left(\Lambda_{y,ij}^{(1)}\right)^{1/2} \sum_{(i',j')\in\mathcal{D}} \left[\mathcal{G}_{y,ii'}^{(I\times I)} + \mathcal{G}_{y,ij'}^{(I\times J)} + \mathcal{G}_{y,ji'}^{(J\times I)} + \mathcal{G}_{y,jj'}^{(J\times J)}\right] \partial_\pi \ell_{y,i'j'}\end{aligned}$$

where $\mathcal{H}_y$ and $\mathcal{D}_y$ are $(I+J)\times(I+J)$ matrices introduced in the proof of Lemma 5, and $\mathcal{G}_y := \mathcal{H}_y^\dagger - \mathcal{D}_y^{-1}$, and $\mathcal{G}_y^{(I\times I)}$, $\mathcal{G}_y^{(I\times J)}$, $\mathcal{G}_y^{(J\times I)}$, $\mathcal{G}_y^{(J\times J)}$ denotes the various blocks of this $(I+J)\times(I+J)$ matrix. Remember that according to (S.11) all the elements of $\mathcal{G}_y$ are uniformly bounded of order n^{-1}. Thus, by applying Corollary S.2(i) with $b_{y,hij}^{(n)}$ equal to $\widetilde{x}_{y,ij}^h$, for $h = 1,\dots,d_x$, and also with $b_{y,hij}^{(n)}$ equal to $n\left(\mathcal{G}_{y,hi}^{(I\times I)} + \mathcal{G}_{y,hj}^{(I\times J)}\right)$, for $h = 1,\dots,I$, and equal to $n\left(\mathcal{G}_{y,hi}^{(J\times I)} + \mathcal{G}_{y,hj}^{(J\times J)}\right)$, for $h = 1,\dots,J$, we find that

$$\sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left|\left(Q_y^{(1)} s_y\right)_{ij}\right| = O_P(n^{-1/2+1/6}) \quad \sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left|\left(Q_y^{(\mathrm{rem})} s_y\right)_{ij}\right| = O_P(n^{-1/2+1/6}). \tag{S.16}$$

Combining the above we find that $Q_y s_y = Q_y^{(1)} s_y + Q_y^{(2)} s_y + Q_y^{(3)} s_y + Q_y^{(\mathrm{rem})} s_y$ indeed satisfies $\sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left|\left(Q_y s_y\right)_{ij}\right| = o_P(n^{-1/6})$. ■

Proof of Lemma 6, Part (ii) **and** (iii)**.** Here, we use Theorem 2.11.11 in van der Vaart and Wellner (1996), which is restated above as Lemma S.3. To relate this to our model we define

$$Z_{\ell,n}(y,k) = \frac{\left[W_y^{-1}\, \widetilde{x}_{y,i_\ell j_\ell}\right]_k}{\sqrt{n}}\, 1\{y_{i_\ell j_\ell} \le y\}, \tag{S.17}$$

where $\ell \in \{1,\dots,n\}$, and $i_\ell \in \mathbf{I}$, $j_\ell \in \mathbf{J}$ are chosen such that $B = \{(i_\ell, j_\ell) \,:\, \ell = 1,\dots,n\}$. $Z_{\ell,n}$ defines a stochastic process with index set $\mathcal{F} = \mathcal{Y}\times\{1,\dots,\dim\beta\}$. For $f = (y,k) \in \mathcal{F}$ we write $Z_{\ell,n}(f)$. Part (ii) of Lemma 6 can then be written as

$$\sum_{\ell=1}^n \left(Z_{\ell,n} - \mathbb{E}\, Z_{\ell,n}\right) \rightsquigarrow \mathcal{Z}^{(\beta)},$$

where the limiting process $\mathcal{Z}^{(\beta)}$ is also indexed by $f \in \mathcal{F}$. We also define the following metric on $\mathcal{F}$,

$$\rho(f_1, f_2) := C\left[|y_1 - y_2|^{1/2} + 1(k_1 \ne k_2)\right], \tag{S.18}$$

for some sufficiently large constant $C > 0$. For a general index set $\mathcal{F}$, a sufficient condition for a metric ρ on $\mathcal{F}$ to be "Gaussian dominated" is given by (see van der Vaart and Wellner 1996, p.212)

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, \rho)}\, d\varepsilon \;<\; \infty, \tag{S.19}$$

where $N(\varepsilon, \mathcal{F}, \rho)$ denotes the covering number.

$Z_{\ell,n}$ is a triangular array, because W_y and $\widetilde{x}_{y,ij}$ both implicitly depend on n, implying that $Z_{\ell,n_1} \neq Z_{\ell,n_2}$ for $n_1 \neq n_2$. Remember that the probability measure we use throughout is conditional on x, α^0, γ^0, implying that the $Z_{\ell,n}$ are independent (but not identically distributed) across ℓ, according to our assumptions.

Using the model and the definition (S.17) we have

$$W_y^{-1}\left[-\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}} \partial_\pi \ell_{y,ij}\, \widetilde{x}_{y,ij}\right] = \sum_{\ell=1}^n \left(Z_{\ell,n} - \mathbb{E}\, Z_{\ell,n}\right) =: Z_n.$$

Using the Lyapunov CLT it is easy to verify that all the finite dimensional marginals $(Z_n(f_1), Z_n(f_2), \ldots, Z_n(f_p))$ of the stochastic process Z_n converge weakly to a zero mean Gaussian limit process $(\mathcal{Z}^{(\beta)}(f_1), \mathcal{Z}^{(\beta)}(f_2), \ldots, \mathcal{Z}^{(\beta)}(f_p))$. It is also easy to show that the second moments of the limit process are given by $\mathbb{E}\mathcal{Z}^{(\beta)}(f_1)\mathcal{Z}^{(\beta)}(f_2) = \left[\overline{W}_{y_1}^{-1}\, \overline{V}_{y_1,y_2}\, \overline{W}_{y_2}^{-1}\right]_{k_1 k_2}$.

In order to conclude that the process Z_n is weakly convergent we also need to show that Z_n is tight. For this we employ Lemma S.3 above with $m_n = n$ and metric ρ given in (S.18). This ρ is Gaussian dominated on $\mathcal{F}$, because we have

$$\log N(\varepsilon, \mathcal{F}, \rho) \lesssim \begin{cases} \log(K/\varepsilon), & \text{for } 0 < \varepsilon < K, \\ 0, & \text{for } \varepsilon \geq K, \end{cases}$$

for some constant $K > 0$, implying that (S.19) is satisfied.

To verify condition (i) of Lemma S.3, we calculate

$$\begin{aligned}
\sum_{\ell=1}^n \mathbb{E}^*\left[\|Z_{n\ell}\|_{\mathcal{F}}\;\; 1\left\{\|Z_{n\ell}\|_{\mathcal{F}} > \eta\right\}\right] &\leq n \max_\ell \mathbb{E}^*\left[\|Z_{n\ell}\|_{\mathcal{F}}\;\; 1\left\{\|Z_{n\ell}\|_{\mathcal{F}} > \eta\right\}\right] \\
&\leq n \max_\ell \mathbb{E}^*\left[\frac{\|Z_{n\ell}\|_{\mathcal{F}}^2}{\eta}\;\; 1\left\{\|Z_{n\ell}\|_{\mathcal{F}} > \eta\right\}\right] \\
&\leq \max_{i,j} \mathbb{E}\left[\frac{\sup_y \left\|W_y^{-1}\, \widetilde{x}_{y,ij}\right\|_\infty^2}{\eta}\;\; 1\left\{\frac{\sup_y \|W_y^{-1}\, \widetilde{x}_{y,ij}\|_\infty}{\sqrt{n}} > \eta\right\}\right], \\
&\to 0.
\end{aligned}$$

where for the second inequality we multiplied with $\|Z_{n\ell}\|_{\mathcal{F}}/\eta$ inside the expectation, which is larger than one for $\|Z_{n\ell}\|_{\mathcal{F}} > \eta$; for the third inequality we used that $\|Z_{n\ell}\|_{\mathcal{F}} \leq \sup_y \|W_y^{-1}\, \widetilde{x}_{y,i_\ell j_\ell}\|_\infty/\sqrt{n}$; and for the final conclusion we used that $\sup_{i,j} E \sup_y \left\|W_y^{-1}\, \widetilde{x}_{y,ij}\right\|^{2+\delta}$ is uniformly bounded.

Next, for $y_1 \leq y_2$ we have

$$\begin{aligned}
\sqrt{n}\,|Z_{n\ell}(f_1) - Z_{n\ell}(f_2)| &= \left|[W_{y_1}^{-1}\, \widetilde{x}_{y_1,i(\ell)j(\ell)}]_{k_1}\, 1\{y_{i(\ell)j(\ell)} \leq y_1\} - [W_{y_2}^{-1}\, \widetilde{x}_{y_2,i(\ell)j(\ell)}]_{k_2}\, 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| \\
&\leq \left|[W_{y_2}^{-1}\, \widetilde{x}_{y_2,i(\ell)j(\ell)}]_{k_2}\right| 1\{y_1 < y_{i(\ell)j(\ell)} \leq y_2\} \\
&\qquad + \left|[W_{y_1}^{-1}\, \widetilde{x}_{y_1,i(\ell)j(\ell)}]_{k_1} - [W_{y_2}^{-1}\, \widetilde{x}_{y_2,i(\ell)j(\ell)}]_{k_2}\right| \\
&\lesssim 1\{y_1 < y_{i(\ell)j(\ell)} \leq y_2\} \\
&\qquad + \left\|W_{y_1}^{-1}\, \widetilde{x}_{y_1,i(\ell)j(\ell)} - W_{y_2}^{-1}\, \widetilde{x}_{y_2,i(\ell)j(\ell)}\right\|_\infty + 1(k_1 \neq k_2) \\
&\lesssim \left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| + |y_1 - y_2| + 1(k_1 \neq k_2).
\end{aligned} \tag{S.20}$$

where we used uniform boundedness of $W_y^{-1}\, \widetilde{x}_{y,ij}$ and of its derivative wrt y. The final result in (S.20) is written such that the bound is also applicable for $y_1 > y_2$.

Using the bound (S.20) we now verify condition (ii) of Lemma S.3,

$$\begin{aligned}
\sum_{\ell=1}^{n} \mathbb{E}\left(Z_{n\ell}(f_1) - Z_{n\ell}(f_2)\right)^2 &\leq n \max_\ell \mathbb{E}\left(Z_{n\ell}(f_1) - Z_{n\ell}(f_2)\right)^2 \\
&= \max_\ell \mathbb{E}\left[\sqrt{n}\left(Z_{n\ell}(f_1) - Z_{n\ell}(f_2)\right)\right]^2 \\
&\lesssim \max_{i,j} \mathbb{E}\left\{|1\{y_{ij} \leq y_1\} - 1\{y_{ij} \leq y_2\}| + |y_1 - y_2| + 1(k_1 \neq k_2)\right\}^2 \\
&\lesssim \max_{i,j} \mathbb{E}\,|1\{y_{ij} \leq y_1\} - 1\{y_{ij} \leq y_2\}|^2 + |y_1 - y_2|^2 + [1(k_1 \neq k_2)]^2 \\
&\lesssim \max_{i,j} \left|\Lambda(\pi^0_{y_2,ij}) - \Lambda(\pi^0_{y_1,ij})\right| + |y_1 - y_2|^2 + 1(k_1 \neq k_2) \\
&\lesssim |y_1 - y_2| + |y_1 - y_2|^2 + 1(k_1 \neq k_2) \\
&\lesssim \left[|y_1 - y_2|^{1/2} + 1(k_1 \neq k_2)\right]^2.
\end{aligned}$$

where, we used that $\mathbb{E}\,|1\{y_{ij} \leq y_1\} - 1\{y_{ij} \leq y_2\}| = \left|\Lambda(\pi^0_{y_2,ij}) - \Lambda(\pi^0_{y_1,ij})\right| \lesssim |y_1 - y_2|$; and we also used that $\mathcal{Y}$ is bounded, implying that $|y_1 - y_2|^2 \lesssim |y_1 - y_2|$. Thus, condition (ii) of Lemma S.3 holds for sufficiently large C in the definition of ρ in (S.18).

To verify condition (iii) of Lemma S.3, let $C_1 > 0$ be the omitted constant that makes the result in (S.20) a regular inequality. We then have

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |Z_{n\ell}(f_1) - Z_{n\ell}(f_2)| > \frac{t}{\sqrt{n}}\right) \\
&\leq \mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} C_1\left[\left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| + |y_1 - y_2| + 1(k_1 \neq k_2)\right] > t\right) \\
&\leq \mathbb{P}^*\Bigg(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} \left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| \\
&\qquad\qquad + \sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |y_1 - y_2| + \sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} 1(k_1 \neq k_2) > \frac{t}{C_1}\Bigg) \\
&\leq \mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} \left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| > \frac{t}{3\,C_1}\right) \\
&\quad + \mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |y_1 - y_2| > \frac{t}{3\,C_1}\right) + \mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} 1(k_1 \neq k_2) > \frac{t}{3\,C_1}\right).
\end{aligned}$$

Any given ρ-ball $\mathcal{B}(\varepsilon)$ of radius less then ε also corresponds to a given ball in $\mathcal{Y}$ of radius less than $(\varepsilon/C)^2$. The event $\left[\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} \left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| > \frac{t}{3\,C_1}\right]$ can only occurs if $\frac{t}{3\,C_1} \leq 1$ and if $y_{i(\ell)j(\ell)}$ is realized in that particular ball in $\mathcal{Y}$ of radius less than $(\varepsilon/C)^2$. Since our assumptions guarantee that the pdf of $y_{i(\ell)j(\ell)}$ is uniformly bounded from below by a constant $C_2 > 0$ we thus find that

$$\mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} \left|1\{y_{i(\ell)j(\ell)} \leq y_1\} - 1\{y_{i(\ell)j(\ell)} \leq y_2\}\right| > \frac{t}{3\,C_1}\right) \leq 2\,C_2\left(\frac{\varepsilon}{C}\right)^2\ 1\left(\frac{t}{3\,C_1} \leq 1\right).$$

Similarly we find

$$\begin{aligned}
\mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |y_1 - y_2| > \frac{t}{3\,C_1}\right) &\leq 1\left(\frac{t}{3\,C_1} \leq 2\left(\frac{\varepsilon}{C}\right)^2\right), \\
\mathbb{P}^*\left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} 1(k_1 \neq k_2) > \frac{t}{3\,C_1}\right) &\leq 1\left(\frac{t}{3\,C_1} \leq 1\ \ \&\ \ C \leq \varepsilon\right).
\end{aligned}$$

We thus calculate

$$\begin{aligned}
&\sup_{t>0} \sum_{\ell=1}^{n} t^2\, \mathbb{P}^* \left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |Z_{n\ell}(f_1) - Z_{n\ell}(f_2)| > t \right) \\
&\leq \sup_{t>0} \max_{\ell}\, n\, t^2\, \mathbb{P}^* \left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |Z_{n\ell}(f_1) - Z_{n\ell}(f_2)| > t \right) \\
&= \sup_{t>0} \max_{\ell}\, t^2\, \mathbb{P}^* \left(\sup_{f_1,f_2\in\mathcal{B}(\varepsilon)} |Z_{n\ell}(f_1) - Z_{n\ell}(f_2)| > \frac{t}{\sqrt{n}} \right) \\
&\leq \sup_{t>0} t^2 \left\{ 2\, C_2 \left(\frac{\varepsilon}{C}\right)^2 1\left(\frac{t}{3\, C_1} \leq 1\right) + 1\left(\frac{t}{3\, C_1} \leq 2\left(\frac{\varepsilon}{C}\right)^2\right) + 1\left(\frac{t}{3\, C_1} \leq 1 \;\;\&\;\; C \leq \varepsilon\right) \right\} \\
&\leq \varepsilon^2,
\end{aligned}$$

for sufficiently large choice of C. In the last step we also use that $\mathcal{Y}$ is bounded, which together with $\mathcal{B}(\varepsilon) \subset \mathcal{F}$ implies that the possible values of ε are bounded, so that we can always choose C sufficiently large to guarantee that $1\left(\frac{t}{3\, C_1} \leq 1 \;\;\&\;\; C \leq \varepsilon\right) = 0$.

Thus, we can apply Lemma S.3 to find that $\sum_{\ell=1}^{n} (Z_{\ell,n} - \mathbb{E}\, Z_{\ell,n}) \rightsquigarrow \mathcal{Z}^{(\beta)}$, where $\mathcal{Z}^{(\beta)}$ is a tight zero mean Gaussian process with second moments given above.

The proof of part (iii) of Lemma 6 is analogous. ■

Proof of Lemma 6, Part (iv) **and** (v)**.** Decomposing $Q_y s_y = Q_y^{(1)} s_y + Q_y^{(2)} s_y + Q_y^{(3)} s_y + Q_y^{(\text{rem})} s_y$ and using (S.15) and (S.16) we find that

$$\sup_{y\in\mathcal{Y}} \max_{(i,j)\in\mathcal{D}} \left\{ \left[(Q_y s_y)_{ij} \right]^2 - \left[\left(Q_y^{(2)} s_y \right)_{ij} + \left(Q_y^{(3)} s_y \right)_{ij} \right]^2 \right\} = o_P(n^{-1}),$$

and therefore

$$-\frac{1}{2} W_y^{-1} \frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij} \left(\Lambda_{y,ij}^{(1)} \right)^{-1} \Lambda_{y,ij}^{(2)} \left[(Q_y s_y)_{ij} \right]^2 = \frac{I}{\sqrt{n}} C_y^{(1,\beta)} + \frac{J}{\sqrt{n}} C_y^{(2,\beta)} + C_y^{(3,\beta)} + o_P(1),$$

where

$$\begin{aligned}
C_y^{(1,\beta)} &:= -\frac{1}{2} W_y^{-1} \frac{1}{I} \sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij} \left(\Lambda_{y,ij}^{(1)} \right)^{-1} \Lambda_{y,ij}^{(2)} \left[\left(Q_y^{(2)} s_y \right)_{ij} \right]^2, \\
C_y^{(2,\beta)} &:= -\frac{1}{2} W_y^{-1} \frac{1}{I} \sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij} \left(\Lambda_{y,ij}^{(1)} \right)^{-1} \Lambda_{y,ij}^{(2)} \left[\left(Q_y^{(3)} s_y \right)_{ij} \right]^2, \\
C_y^{(3,\beta)} &:= -W_y^{-1} \frac{1}{\sqrt{n}} \sum_{(i,j)\in\mathcal{D}} \widetilde{x}_{y,ij} \left(\Lambda_{y,ij}^{(1)} \right)^{-1} \Lambda_{y,ij}^{(2)} \left(Q_y^{(2)} s_y \right)_{ij} \left(Q_y^{(3)} s_y \right)_{ij}.
\end{aligned}$$

Using that $\mathbb{E}\left[\left(Q_y^{(2)}s_y\right)_{ij}\right]^2 = Q^{(2)}_{y,ij,ij} = \Lambda^{(1)}_{y,ij}\left(\sum_{j'\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij'}\right)^{-1}$ and $\mathbb{E}\left[\left(Q_y^{(3)}s_y\right)_{ij}\right]^2 = Q^{(3)}_{y,ij,ij} = \Lambda^{(1)}_{y,ij}\left(\sum_{i'\in\mathcal{D}_j}\Lambda^{(1)}_{y,i'j}\right)^{-1}$ we find that

$$\mathbb{E}C_y^{(1,\beta)} = B_y^{(\beta)}, \qquad \mathbb{E}C_y^{(2,\beta)} = D_y^{(\beta)}.$$

Furthermore, using the expressions for $\left(Q_y^{(2)}s_y\right)_{ij}$ and $\left(Q_y^{(3)}s_y\right)_{ij}$ in (S.14) above we can write, for given $\ell \in \{1,\ldots,d_x\}$,

$$\left[W_y\, C_y^{(1,\beta)}\right]_\ell = -\frac{1}{2}\frac{1}{I}\sum_{i=1}^{I} c_{y,i}^{(n)}\left(\frac{1}{\sqrt{|\mathcal{D}_i|}}\sum_{j\in\mathcal{D}_i}\partial_\pi \ell_{y,ij}\right)^2,$$

$$\left[W_y\, C_y^{(2,\beta)}\right]_\ell = -\frac{1}{2}\frac{1}{J}\sum_{j=1}^{J} d_{y,j}^{(n)}\left(\frac{1}{\sqrt{|\mathcal{D}_j|}}\sum_{i\in\mathcal{D}_j}\partial_\pi \ell_{y,ij}\right)^2,$$

where

$$c_{y,i}^{(n)} = \frac{|\mathcal{D}_i|^{-1}\sum_{j\in\mathcal{D}_i}\widetilde{x}_{y,ij}\Lambda^{(2)}_{y,ij}}{\left(|\mathcal{D}_i|^{-1}\sum_{j\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij}\right)^2}, \qquad d_{y,j}^{(n)} = \frac{|\mathcal{D}_j|^{-1}\sum_{i\in\mathcal{D}_j}\widetilde{x}_{y,ij}\Lambda^{(2)}_{y,ij}}{\left(|\mathcal{D}_j|^{-1}\sum_{i\in\mathcal{D}_j}\Lambda^{(1)}_{y,ij}\right)^2},$$

which are of order $O_P(1)$, uniformly over y and i and j. By employing part (ii) of Corollary S.2 we thus find that

$$C_y^{(1,\beta)} - \mathbb{E}C_y^{(1,\beta)} = o_P(1), \qquad C_y^{(2,\beta)} - \mathbb{E}C_y^{(2,\beta)} = o_P(1).$$

Finally, again using (S.14) we can write

$$C_y^{(3,\beta)} = -W_y^{-1}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\frac{\widetilde{x}_{y,ij}\Lambda^{(2)}_{y,ij}}{|\mathcal{D}_i|^{1/2}|\mathcal{D}_j|^{1/2}}\frac{\left(|\mathcal{D}_i|^{-1/2}\sum_{j'\in\mathcal{D}_i}\partial_\pi\ell_{y,ij'}\right)\left(|\mathcal{D}_j|^{-1/2}\sum_{i'\in\mathcal{D}_j}\partial_\pi\ell_{y,i'j}\right)}{\left(|\mathcal{D}_i|^{-1}\sum_{j'\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij'}\right)\left(|\mathcal{D}_j|^{-1}\sum_{i'\in\mathcal{D}_j}\Lambda^{(1)}_{y,i'j}\right)},$$

and therefore

$$\begin{aligned}\left\|C_y^{(3,\beta)}\right\| &\le \max_{j\in\mathbf{J}}\left|\,|\mathcal{D}_j|^{-1/2}\sum_{i\in\mathcal{D}_j}\partial_\pi\ell_{y,ij}\right| \\ &\quad\times\left\|W_y^{-1}\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\frac{\widetilde{x}_{y,ij}\Lambda^{(2)}_{y,ij}}{|\mathcal{D}_i|^{1/2}|\mathcal{D}_j|^{1/2}}\frac{\left(|\mathcal{D}_i|^{-1/2}\sum_{j'\in\mathcal{D}_i}\partial_\pi\ell_{y,ij'}\right)}{\left(|\mathcal{D}_i|^{-1}\sum_{j'\in\mathcal{D}_i}\Lambda^{(1)}_{y,ij'}\right)\left(|\mathcal{D}_j|^{-1}\sum_{i'\in\mathcal{D}_j}\Lambda^{(1)}_{y,i'j}\right)}\right\| \\ &= \left(\max_{i\in I}|\mathcal{D}_i|^{-1/2}\right)\left(\max_{j\in\mathbf{J}}\left|\,|\mathcal{D}_j|^{-1/2}\sum_{i\in\mathcal{D}_j}\partial_\pi\ell_{y,ij}\right|\right)\left\|\frac{1}{\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}e_{y,i}^{(n)}\,\partial_\pi\ell_{y,ij}\right\|,\end{aligned}$$

where

$$e_{y,i}^{(n)} = W_y^{-1}|\mathcal{D}_i|^{-1/2}\sum_{j\in\mathcal{D}_i}|\mathcal{D}_j|^{-1/2}\frac{\widetilde{x}_{y,ij}\Lambda_{y,ij}^{(2)}}{\left(|\mathcal{D}_i|^{-1}\sum_{j'\in\mathcal{D}_i}\Lambda_{y,ij'}^{(1)}\right)\left(|\mathcal{D}_j|^{-1}\sum_{i'\in\mathcal{D}_j}\Lambda_{y,i'j}^{(1)}\right)},$$

which is of order one, uniformly over y and i. Thus, by applying Corollary S.1, and Corollary S.2 with $b_{y,hij}^{(n)}$ equal to the elements of the d_x-vector $e_{y,i}^{(n)}$ (i.e. no j-dependence), we find that $\left\|C_y^{(3,\beta)}\right\| = O_P(n^{-1/4})o_P\left(n^{1/12}\right)O_P\left(n^{1/6}\right) = o_P(1)$. Combining the above we conclude

$$-\frac{1}{2}W_y^{-1}\,n^{-1/2}\sum_{(i,j)\in\mathcal{D}}\widetilde{x}_{y,ij}\left(\Lambda_{y,ij}^{(1)}\right)^{-1}\Lambda_{y,ij}^{(2)}\left[(Q_ys_y)_{ij}\right]^2-\left(n^{-1/2}IB_y^{(\beta)}+n^{-1/2}JD_y^{(\beta)}\right)\to_P 0.$$

The proof for

$$\frac{1}{2\sqrt{n}}\sum_{(i,j)\in\mathcal{D}}\left(\Lambda_{y,ij}^{(1)}\right)^{-1}\left(\Lambda_{y,ij,k}^{(2)}-\Lambda_{y,ij}^{(2)}\Psi_{y,ij,k}\right)\left[(Q_ys_y)_{ij}\right]^2-\left(n^{-1/2}IB_{y,k}^{(\Lambda)}+n^{-1/2}JD_{y,k}^{(\Lambda)}\right)\to_P 0$$

is analogous. ■