Supplementary Appendix for "Bias and Consistency in Three-way Gravity Models"

by Martin Weidner & Thomas Zylkin

Journal of International Economics,
doi: 10.1016/j.jinteco.2021.103513.

This Appendix first describes an additional simulation exercise based on the trade data. We then introduce additional notation, definitions, and other technical details that supplement the formal theorems and remarks presented in Section 3 of the main text. Proofs of our formal results then follow, starting with a poof of our Proposition 3, which characterizes the asymptotic distribution of $\widehat{\beta}$ and its asymptotic bias. This proof naturally lends itself to further discussion of the "large T" results from Remark 2 as well as the consistency result from Proposition 1, which itself follows as a by-product of Proposition 3. We then demonstrate the uniqueness of this latter result as stated in Proposition 2 and highlight the general inconsistency of other three-way gravity estimators.

After the proofs, we include further supplementary discussions on the downward bias in the estimated standard errors, on allowing for conditional dependence across pairs in the trade data, and on how FE-PPML is affected by IPPs in more general settings beyond the gravity framework.

A.1 Simulation Based on Trade Data

As a additional simulation exercise, we revisit the BACI aggregate trade data we used in Section 5 and ask: if the estimated effect of an FTA and its standard error were indeed biased to the degrees implied by our bias corrections, would our corrections be successful at correctly identifying the bias and improving inference?

To answer this question, we start from the original aggregate trade data and FTA data but reconstruct the conditional mean λ_{ijt} as though $\beta = 0.086$. That is, we first adjust the estimated $\widehat{\lambda}_{ijt}$'s from the original estimation to account for the change in β and so that the FOC's for all fixed effects are consistent with $\beta = 0.086$. This gives us new "true" values of the conditional mean that we denote by $\lambda_{ijt}^{(1)}$. The original data is therefore assumed to have been generated by $y_{ijt} = \lambda_{ijt}^{(1)}\omega_{ijt}$, where the true disturbance ω_{ijt} is backed out using $\omega_{ijt} = y_{ijt}/\lambda_{ijt}^{(1)}$.

Next, we choose a DGP for ω_{ijt} that can reproduce the biases implied by our corrections. Taking our cues from DGP IV, which we found earlier to produce a downward bias in $\widehat{\beta}$, we consider a DGP where the conditional variance of y_{ijt} has the form $\text{Var}[y_{ijt}|.] = a\lambda_{ijt} + bFTA\lambda_{ijt}^2$. That is, we allow for some overdispersion that depends on the regressor of interest, as in DGP IV. We choose the two parameters a and b in order to come close to matching the following three values: (i) the bias in $\widehat{\beta}$, (ii) the standard deviation of $\widehat{\beta}$ (assumed to be 0.03), (iii) the bias of the standard error of $\widehat{\beta}$. To keep things simple, all of our simulations sample new values for ω_{ijt} only, holding $\lambda_{ijt}^{(1)}$ fixed. As in the main text, we use 5,000 replications and assume $\rho = 0.3$. For our chosen values of a and b—$a = 200,000$, $b = 0.08$—we obtain an average $\widehat{\beta}$ of 0.0823, an average standard error of 0.0267, and a standard deviation of 0.0307. The uncorrected coverage is 0.9078.

When our preferred bias corrections are applied, they do not completely solve the coverage problem but do induce across-the-board improvements. The average corrected $\widehat{\beta}$ using the analytical bias correction is 0.0842 and the corrected standard error is 0.0290. Coverage improves as well, but only to 0.9210. As discussed in the main text, one important factor that limits the improvement in coverage is the fact that applying bias corrections to the point estimates increases their variance. In this case, the standard deviation of the corrected estimate is 0.0321.

Turning to the jackknife bias correction, we find as before that it does a superior job of bias reduction than the analytical correction, producing a average corrected estimate of 0.0851. However, the standard deviation of the jackknife-corrected estimates is 0.0371, echoing our previous finding that the improved bias reduction performance of the jackknife comes at a steep penalty in terms of increased variance. As a result, the coverage we obtain when we combine the jackknife with the corrected standard errors is only 0.8756.

Interestingly, if we only use the correction to the standard errors, i.e., without applying any correction to the point estimates, we obtain a coverage ratio of 0.9312, which is better than if we also use the analytical correction. It nonetheless remains true that using corrections to both the point estimates and standard errors leads to an improvement in coverage, as we have consistently found throughout our results. In general, these simulations reinforce our earlier conclusion that bias corrections, though helpful, are not necessarily a panacea to the issues we raise in the paper.

A.2 Additional Notation and Definitions

It is convenient to define the log-likelihood as a function of the index vector π_{ij} as follows,

$$\ell_{ij}(\beta, \alpha_i, \gamma_j) =: \ell_{ij}(\beta, \pi_{ij}), \qquad \text{where} \qquad \pi_{ij} = \begin{pmatrix} \pi_{ij1} \\ \vdots \\ \pi_{ijT} \end{pmatrix} := \begin{pmatrix} \alpha_{i1} + \gamma_{j1} \\ \vdots \\ \alpha_{iT} + \gamma_{jT} \end{pmatrix}.$$

In this appendix, we will also be more explicit than in the main text in distinguishing true parameter values β^0, α_{it}^0, γ_{jt}^0, and the corresponding π_{ij}^0 and ϑ_{ijt}^0, from their generic equivalents. For example, the S_{ij}, H_{ij} and G_{ij} that were already defined in the main text can more formally be written as

$$S_{ij} := -\frac{\partial^2 \ell_{ij}(\beta^0, \pi_{ij}^0)}{\partial \pi_{ij}}, \qquad H_{ij} := -\frac{\partial^2 \ell_{ij}(\beta^0, \pi_{ij}^0)}{\partial \pi_{ij}\, \partial \pi_{ij}'},$$

and

$$G_{ij,tsr} = \frac{\partial^3 \ell_{ij}(\beta^0, \pi_{ij}^0)}{\partial \pi_{ijt}\, \partial \pi_{ijs}\, \partial \pi_{ijr}} = \begin{cases} -\vartheta_{ijt}^0 \left(1 - \vartheta_{ijt}^0\right) \left(1 - 2\vartheta_{ijt}^0\right) \sum_\tau y_{ij\tau} & \text{if } t = s = r, \\ -\vartheta_{ijs}^0 \left(1 - 2\vartheta_{ijs}^0\right) \vartheta_{ijt}^0 \sum_\tau y_{ij\tau} & \text{if } s = r \neq t, \\ -\vartheta_{ijs}^0 \left(1 - 2\vartheta_{ijs}^0\right) \vartheta_{ijr}^0 \sum_\tau y_{ij\tau} & \text{if } t = s \neq r, \\ -\vartheta_{ijt}^0 \left(1 - 2\vartheta_{ijt}^0\right) \vartheta_{ijs}^0 \sum_\tau y_{ij\tau} & \text{if } r = t \neq s, \\ -2\vartheta_{ijr}^0 \vartheta_{ijs}^0 \vartheta_{ijt}^0 \sum_\tau y_{ij\tau} & \text{if } r \neq s \neq t \neq r. \end{cases}$$

The $T \times K$ matrix $\widetilde{x}_{ij}$ that was informally introduced in the main text as a two-way within-transformation of x_{ij}, can be formally defined by $\widetilde{x}_{ij} = x_{ij} - \alpha_i^x - \gamma_j^x$, where α_i^x and γ_j^x are $T \times K$ matrices that minimize

$$\sum_{i=1}^{N} \sum_{j \in \mathfrak{N} \setminus \{i\}} \mathrm{Tr}\left[\left(x_{ij} - \alpha_i^x - \gamma_j^x\right)' \bar{H}_{ij} \left(x_{ij} - \alpha_i^x - \gamma_j^x\right) \right], \tag{16}$$

subject to appropriate normalizations on α_i^x and γ_j^x (e.g. $\iota_T' \alpha_i^x = \iota_T' \gamma_j^x = 0$, where $\iota_T = (1, \ldots, 1)'$ is a T-vector of ones). Each within-transformed regressor vector $\widetilde{x}_{ij,k}$ can be interpreted as containing the residuals left after partialing out $x_{ij,k}$ with respect to any i- and j-specific components and weighting by $\bar{H}_{ij}$.[1]

[1] While we present the computation of $\widetilde{x}_{ij}$ as a two-way within-transformation to preserve the analogy with Fernández-Val and Weidner (2016), each individual element $\widetilde{x}_{ijt,k}$ can also be shown to be equivalent (subject to a normalization) to a three-way within-transformation of $x_{ijt,k}$ with respect to it, jt, and ij and weighting by λ_{ijt}. Readers familiar with Larch, Wanner, Yotov, and Zylkin (2019) may find the latter presentation easier to digest.

A.2.1 Analytical Bias Correction Formulas

The analytical bias correction discussed in Section 3.4 orequires estimates of the expressions W_N, B_N, D_N defined in Proposition 3. For this, we first require plugin objects $\widehat{\widetilde{x}}_{ij}$, $\widehat{S}_{ij}$, $\widehat{H}_{ij}$, $\widehat{\widehat{H}}_{ij}$, and $\widehat{G}_{ij}$ — these objects are formed in the obvious way by replacing λ_{ijt} with $\widehat{\lambda}_{ijt}$ and ϑ_{ijt} with $\widehat{\vartheta}_{ijt} := \widehat{\lambda}_{ijt}/\sum_\tau \widehat{\lambda}_{ij\tau}$ where needed. Then, the $\widehat{B}_N$ and $\widehat{D}_N$ are K-vectors with elements given by

$$\widehat{B}_N^k = -\frac{1}{N-1}\sum_{i=1}^N \mathrm{Tr}\left[\left(\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{\widehat{H}}_{ij}\right)^\dagger \sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{H}_{ij}\,\widehat{\widetilde{x}}_{ij,k}\,\widehat{S}'_{ij}\right]$$
$$+\frac{1}{2\,(N-1)}\sum_{i=1}^N \mathrm{Tr}\left[\left(\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{G}_{ij}\,\widehat{\widetilde{x}}_{ij,k}\right)\left(\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{\widehat{H}}_{ij}\right)^\dagger\left[\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{S}_{ij}\,\widehat{S}'_{ij}\right]\left(\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{\widehat{H}}_{ij}\right)^\dagger\right],$$
$$\widehat{D}_N^k = -\frac{1}{N-1}\sum_{j=1}^N \mathrm{Tr}\left[\left(\sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{\widehat{H}}_{ij}\right)^\dagger \sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{H}_{ij}\,\widehat{\widetilde{x}}_{ij,k}\,\widehat{S}'_{ij}\right]$$
$$+\frac{1}{2\,(N-1)}\sum_{j=1}^N \mathrm{Tr}\left[\left(\sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{G}_{ij}\,\widehat{\widetilde{x}}_{ij,k}\right)\left(\sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{\widehat{H}}_{ij}\right)^\dagger\left[\sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{S}_{ij}\,\widehat{S}'_{ij}\right]\left(\sum_{i\in\mathfrak{N}\setminus\{j\}}\widehat{\widehat{H}}_{ij}\right)^\dagger\right],$$

and we have

$$\widehat{W} = \frac{1}{N\,(N-1)}\sum_{i=1}^N\sum_{j\in\mathfrak{N}\setminus\{i\}}\widehat{\widetilde{x}}'_{ij}\,\widehat{\widehat{H}}_{ij}\,\widehat{\widetilde{x}}_{ij},$$

The replacement of N with $N-1$ in $\widehat{B}_N^k$ and $\widehat{D}_N^k$ stems from a degrees-of-freedom correction. This correction is needed because creating plug-in values for the $\mathbb{E}\left(S'_{ij}H_{ij}\,\middle|x_{ij,k}\right)$ and $\mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij,k}\right)$ objects that appear in Proposition 3 requires computing terms of the form $\mathbb{E}[y_{ijt}^2]$ and $\mathbb{E}[y_{ijs}y_{ijt}]$, as illustrated in Remark 1.

A.2.2 Details on large T bias expansion

We also want to explain the result in Remark 2 of the main text in more detail here by rewriting the bias terms B_N and D_N to illuminate the role of the time dimension. Using generic definitions for S_{ij}, H_{ij}, G_{ij}, and $\widetilde{x}_{ij}$ (e.g., $S_{ij} := \partial\ell_{ij}/\partial\pi_{ij}$, $H_{ij} := \partial^2\ell_{ij}/\partial\pi_{ij}\partial\pi'_{ij}$, etc.), the formulas for the asymptotic distribution in Proposition 3 apply generally to M-estimators based on concave objective functions $\ell_{ij}(\beta,\alpha_{it},\gamma_{jt})$. Unlike in the two-way FE-PPML case, these formulas do not reduce to zero when we further specialize them to the profiled Poisson pseudo-likelihood shown in (11), but we still find it instructive to do

so (e.g., to discuss the large T limit from Remark 2). For that purpose, we define the $T \times T$ matrix $M_{ij} = \mathbf{I}_T - \vartheta_{ij}\iota_T'$. Furthermore, let Λ_{ij} be the $T \times T$ diagonal matrix with diagonal elements λ_{ijt}, and for $i, j \in \{1, \dots, N\}$ define the $T \times T$ matrices

$$Q_i = \frac{1}{N-1}\left(\sum_{j \in \mathfrak{N}\setminus\{i\}} M_{ij}\,\Lambda_{ij}\,M_{ij}'\right)^{\dagger}\left(\sum_{j \in \mathfrak{N}\setminus\{i\}} M_{ij}\,\mathbb{E}(y_{ij}y_{ij}')\,M_{ij}'\right)\left(\sum_{j \in \mathfrak{N}\setminus\{i\}} M_{ij}\,\Lambda_{ij}\,M_{ij}'\right)^{\dagger},$$

$$R_{ij} = \mathbb{E}(y_{ij}y_{ij}')\,M_{ij}'\left(\frac{1}{N-1}\sum_{j' \in \mathfrak{N}\setminus\{i\}} M_{ij}\,\Lambda_{ij}\,M_{ij}'\right)^{\dagger}\Lambda_{ij}M_{ij}'.$$

The bias term $B_N = (B_N^k)$ in Proposition 3 can then be expressed as

$$B_N^k = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j \in \mathfrak{N}\setminus\{i\}}\left[-\frac{\iota_T'\,R_{ij}\,\widetilde{x}_{ij,k}}{\iota_T'\lambda_{ij}} + \frac{\lambda_{ij}'\,Q_i\,\Lambda_{ij}\,M_{ij}'\,\widetilde{x}_{ij,k}}{\iota_T'\lambda_{ij}}\right], \tag{17}$$

and an analogous formula for D_N follows by interchanging i and j appropriately. As long as there is only weak time dependence between observations the matrix objects R_{ij} and $Q_i\Lambda_{ij}M_{ij}'$ above are both of order 1 as $T \to \infty$, such that both terms in brackets in (17) are likewise of order 1. This explains the result stated in Remark 2 of the main text.

A.3 Proof of Proposition 3

Known result for two-way fixed effect panel models

Our proof of Proposition 3 relies on results from Fernández-Val and Weidner (2016) – denoted FW in the following. That paper considers a standard panel setting where individuals i are observed over time periods t, and mixing conditions (as opposed to conditional independence assumptions) are imposed across time periods. By contrast, we consider a pseudo-panel setting, where the two panel dimensions are labelled by exporters i and importers j, and we impose conditional independence assumptions across both i and j here (see also Dzemski, 2019, who employs those results in a directed network setting where outcomes are binary, and Graham, 2017, for the undirected network case.) Given those differences—and before introducing any further complications—we briefly want to restate the main result in FW for the two-way pseudo-panel case. Outcomes Y_{ij}, $i, j = 1, \dots, N$, conditional on all the strictly exogenous regressors $X = (X_{ij})$, fixed effect N-vectors α and γ, and common parameters β are assumed to be generated as

$$Y_{ij} \mid X, \alpha, \gamma, \beta \sim f_Y(\cdot \mid X_{ij}, \alpha_i, \gamma_j, \beta),$$

where the conditional distribution f_Y is known, up to the unknown parameters $\alpha_i, \gamma_j \in \mathbb{R}$ and $\beta \in \mathbb{R}^K$. It is furthermore assumed that α_i and γ_j enter the distribution function only through the single index $\pi_{ij} = \alpha_i + \gamma_j$; that is, the log-likelihood can be defined by

$$\ell_{ij}(\beta,\, \pi_{ij}) = \log f_Y(Y_{ij} \mid X_{ij}, \alpha_i, \gamma_j, \beta).$$

The maximum likelihood estimator for β is given by

$$\widehat{\beta} = \operatorname*{argmax}_{\beta \in \mathbb{R}^K} \max_{\alpha,\gamma \in \mathbb{R}^N} \mathcal{L}(\beta, \alpha, \gamma), \qquad \mathcal{L}(\beta, \alpha, \gamma) = \sum_{i,j} \ell_{ij}(\beta,\, \alpha_i + \gamma_j).$$

Also, define the K-vector Ξ_{ij} with components, $k = 1, \ldots, K$,

$$\Xi_{ij,k} = \alpha^*_{i,k} + \gamma^*_{j,k}, \quad (\alpha^*_k,\, \gamma^*_k) = \operatorname*{argmin}_{\alpha_{i,k}, \gamma_{j,k}} \sum_{i,j} \mathbb{E}(-\partial_{\pi^2} \ell_{ij}) \left(\frac{\mathbb{E}(\partial_{\beta_k \alpha_i} \ell_{ij})}{\mathbb{E}(\partial_{\alpha_i^2} \ell_{ij})} - \alpha_{i,k} - \gamma_{j,k} \right)^2,$$

where here and in the following all expectations are conditional on regressors $X = (X_{ij})$, and on the parameters α, γ, β. For $q \in \{0, 1, 2\}$, the (within-transformation) differentiation operator $\mathcal{D}_{\beta\alpha_i^q} = \mathcal{D}_{\beta\gamma_j^q}$ is defined by

$$\mathcal{D}_{\beta\alpha_i^q} \ell_{ij} = \partial_{\beta\alpha_i^q} \ell_{ij} - \partial_{\alpha_i^{q+1}} \ell_{ij}\, \Xi_{ij}, \qquad \mathcal{D}_{\beta\gamma_j^q} \ell_{ij} = \partial_{\beta\gamma_j^q} \ell_{ij} - \partial_{\gamma_j^{q+1}} \ell_{ij}\, \Xi_{ij}. \tag{18}$$

Theorem 1. *Assume that*

(i) Conditional on X, α^0, γ^0, β^0 the outcomes Y_{ij} are distributed independently across i and j with

$$Y_{ij} \mid X, \alpha^0, \gamma^0, \beta^0 \sim \exp[\ell_{ij}(\beta^0, \pi^0_{ij})],$$

where $\pi^0_{ij} = \alpha^0_i + \gamma^0_j$.

(ii) The map $(\beta, \pi) \mapsto \ell_{ij}(\beta, \pi)$ is four times continuously differentiable, almost surely. All partial derivatives of $\ell_{ij}(\beta, \pi)$ up to fourth order are bounded in absolute value by a function $m(Y_{it}, X_{it}) > 0$, almost surely, uniformly over a convex compact set $\mathcal{B} \subset \mathbb{R}^{\dim \beta + 1}$, which contains an ε-neighbourhood of (β^0, π^0_{ij}) for all i, j, N, and some $\varepsilon > 0$. Furthermore, $\max_{i,j} \mathbb{E}[m(Y_{ij}, X_{ij})]^{8+\nu}$ is uniformly bounded over N, almost surely, for some $\nu > 0$.

(iii) For all N, the function $(\beta, \alpha, \gamma) \mapsto \mathcal{L}(\beta, \alpha, \gamma)$ is almost surely strictly concave over $\mathbb{R}^{K+2N}$, apart from one "flat direction" described by the transformation $\alpha_i \mapsto \alpha_i + c$, $\gamma_j \mapsto \gamma_j - c$, which leaves $\mathcal{L}(\beta, \alpha, \gamma)$ unchanged for all $c \in \mathbb{R}$. Furthermore, there exist constants $b_{\min}$ and $b_{\max}$ such that for all $(\beta, \pi) \in \mathcal{B}$, $0 < b_{\min} \leq -\mathbb{E}\left[\partial_{\alpha_i^2} \ell_{ij}(\beta, \pi)\right] \leq b_{\max}$, almost surely, uniformly over i, j, N.

In addition, assume that the following limits exist

$$\overline{B} = \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{i,j}\frac{\mathbb{E}\left(\partial_{\alpha_i}\ell_{ij}\mathcal{D}_{\beta\alpha_i}\ell_{ij} + \frac{1}{2}\mathcal{D}_{\beta\alpha_i^2}\ell_{ij}\right)}{\sum_{j'}\mathbb{E}\left(\partial_{\alpha_i^2}\ell_{ij'}\right)}\right],$$

$$\overline{D} = \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{i,j}\frac{\mathbb{E}\left(\partial_{\gamma_j}\ell_{ij}\mathcal{D}_{\beta\gamma_j}\ell_{ij} + \frac{1}{2}\mathcal{D}_{\beta\gamma_j^2}\ell_{ij}\right)}{\sum_{i'}\mathbb{E}\left(\partial_{\gamma_j^2}\ell_{i'j}\right)}\right],$$

$$\overline{W} = \lim_{N\to\infty}\left[-\frac{1}{N^2}\sum_{i,j}\mathbb{E}\left(\partial_{\beta\beta'}\ell_{ij} - \partial_{\alpha_i^2}\ell_{ij}\Xi_{ij}\Xi'_{ij}\right)\right],$$

where expectations are conditional on X, α, γ, β. Finally, assume that $\overline{W} > 0$. Then, as $N \to \infty$, we have

$$N\left(\widehat{\beta} - \beta^0\right) \to_d \overline{W}^{-1}\mathcal{N}(\overline{B} + \overline{D},\ \overline{W}),$$

Remarks:

(a) This is just a reformulation of Theorem 4.1 in FW to the case of pseudo-panels, and the proof is provided in that paper. Since we consider only strictly exogenous regressors, all the analysis is conditional on X; and the bias term $\overline{B}$ simplifies here, since conditional on X (and the other parameters), we assume independence across both i and j. Thus, no Nickell-type bias (Nickell, 1981; Hahn and Kuersteiner, 2002) appears here, but we still have incidental parameter biases because the model is nonlinear (Neyman and Scott, 1948; Hahn and Newey, 2004).

(b) In the original version of this theorem, the sums in the definitions of $\mathcal{L}(\beta, \alpha, \gamma)$, $\overline{B}$, $\overline{D}$, and $\overline{W}$ run over all possible pairs $(i, j) \in \{1, \ldots, N\}^2$. However, for the trade application in the current paper we assume we only have observations for $i \neq j$; that is, those sums over i and j only run over the set $\{(i, j) \in \{1, \ldots, N\}^2 : i \neq j\}$ of $N(N-1)$ observed country pairs. The sum over j' (in $\overline{B}$) then also only runs over $j' \neq i$, and the sum over i' (in $\overline{D}$) only runs over $i' \neq j$. It turns out that those changes make no difference to the proof of the theorem, because the proportion of missing observations for each i and j is asymptotically vanishing. For that reason it also does not matter whether we change the $1/N^2$ in $\overline{W}$ to $1/[N(N-1)]$, or whether we change $N\left(\widehat{\beta} - \beta^0\right)$ to $\sqrt{N(N-1)}\left(\widehat{\beta} - \beta^0\right)$. The same equivalence holds throughout our own results for applications in which researchers wish to use observations for which $i = j$ (simply replace $N - 1$ with N where appropriate.) It

also does not matter for the proof that the number of exporters and importers is the same, since this is already allowed for in FW's existing results. If we let I be the number of exporters and J be the number of importers, FW's results apply so long as I and J grow large at the same rate.

(c) More generally, careful examination of these proofs and results reveals that all explicit appearances of N and $N-1$ in the definitions of $\overline{W}$, $\overline{B}$, and $\overline{D}$ actually play no role in the fully expressed formula for the asymptotic bias, i.e., $N^{-1}\overline{W}(\overline{B}+\overline{D})$. Thus, there is no need to adjust the terms that explicitly depend on N if some of the data are missing. So long as the missing data occur at random, applying the formulas as written should still generally be expected to deliver an asymptotically valid bias correction. A similar observation applies for missing values in the three-way model. That said, if the missing values occur in such a way that some of the α_i's or γ_j's appear only a small number of times in the data, they will tend to be estimated with a larger degree of estimation noise than the other fixed effects, which could affect the performance of bias corrections based on these formulas in practice.

(d) The above theorem assumes that the log-likelihood $\ell_{ij}(\beta,\, \alpha_i+\gamma_j)$ for $Y_{ij} \mid X, \alpha, \gamma, \beta$ is correctly specified. This is an unrealistic assumption for the PPML estimators in this paper, where we only want to assume that the score of the pseudo-log-likelihood has zero mean at the true parameters, that is, $\mathbb{E}\Big[\partial_\beta \ell_{ij}(\beta^0,\, \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\Big] = 0$ and $\mathbb{E}\Big[\partial_{\alpha_i} \ell_{ij}(\beta^0,\, \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\Big] = 0$ and $\mathbb{E}\Big[\partial_{\gamma_j} \ell_{ij}(\beta^0,\, \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\Big] = 0$. This extension to "conditional moment models" is discussed in Remark 3 of FW. The statement of the theorem then needs to be changed as follows:

$$N\left(\widehat{\beta}-\beta^0\right) \;\to_d\; \overline{W}^{-1}\,\mathcal{N}(\overline{B}+\overline{D},\, \overline{\Omega}), \tag{19}$$

where the definition of $\overline{W}$ is unchanged, but the expression of $\overline{B} = \overline{B}_1 + \overline{B}_2$, $\overline{D} =$

$\overline{D}_1 + \overline{D}_2$ and $\overline{\Omega}$ now read

$$\begin{aligned}
\overline{B}_1 &= \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{i,j}\frac{\mathbb{E}\left(\partial_{\alpha_i}\ell_{ij}\mathcal{D}_{\beta\alpha_i}\ell_{ij}\right)}{\sum_{j'}\mathbb{E}\left(\partial_{{\alpha_i}^2}\ell_{ij'}\right)}\right], \\
\overline{B}_2 &= \lim_{N\to\infty}\left[\frac{1}{2}\frac{1}{N}\sum_{i}\frac{\left[\sum_j \mathbb{E}(\partial_{\alpha_i}\ell_{ij})^2\right]\sum_j \mathbb{E}(\mathcal{D}_{\beta{\alpha_i}^2}\ell_{ij})}{\left[\sum_j \mathbb{E}\left(\partial_{{\alpha_i}^2}\ell_{ij}\right)\right]^2}\right], \\
\overline{D}_1 &= \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{j}\frac{\sum_i \mathbb{E}\left[\partial_{\gamma_j}\ell_{ij}\mathcal{D}_{\beta\gamma_j}\ell_{ij}\right]}{\sum_i \mathbb{E}\left(\partial_{{\gamma_j}^2}\ell_{ij}\right)}\right], \\
\overline{D}_2 &= \lim_{N\to\infty}\left[\frac{1}{2}\frac{1}{N}\sum_{j}\frac{\sum_i \left[\mathbb{E}(\partial_{\gamma_j}\ell_{ij})^2\right]\sum_i \mathbb{E}(\mathcal{D}_{\beta{\gamma_j}^2}\ell_{ij})}{\left[\sum_i \mathbb{E}\left(\partial_{{\gamma_j}^2}\ell_{ij}\right)\right]^2}\right], \\
\overline{\Omega} &= \lim_{N\to\infty}\left[\frac{1}{N^2}\sum_{i,j}\mathbb{E}\left[\mathcal{D}_\beta\ell_{ij}(\mathcal{D}_\beta\ell_{ij})'\right]\right]. \qquad (20)
\end{aligned}$$

These are the formulas that we have to use as a starting point for the bias results derived in this paper.

Our task in the following is to translate and generalize the conditions, statement, and proof of Theorem 1, as extended in (19) and (20), to the case of the three-way PPML estimator discussed in the main text.

Regularity conditions for Proposition 3

The following regularity conditions are required for the statement of Proposition 3 to hold.

Assumption A. *(i) Conditional on* $x = (x_{ijt})$, $\alpha^0 = (\alpha_{it}^0)$, $\gamma^0 = (\gamma_{jt}^0)$, $\eta^0 = (\eta_{ij}^0)$ *and* β^0, *the outcomes* $y_{ij} = (y_{ij,1}, \ldots, y_{ij,T})'$ *are distributed independently across* i *and* j, *and the conditional mean of* y_{ijt} *is given by equation (8) for all* i, j, t.

(ii) The range of x_{ijt}, α_{it}^0 *and* γ_{jt}^0 *is uniformly bounded, and there exists* $\nu > 0$ *such that* $\mathbb{E}(y_{ijt}^{8+\nu}|x_{ijt}, \alpha_{it}, \gamma_{jt}, \eta_{ij})$ *is uniformly bounded over* i, j, t, N.

(iii) $\lim_{N\to\infty} W_N > 0$, *with* W_N *as defined in Proposition 3.*

Those assumptions are very similar to those in Theorem 1 above: Assumption A(i) is analogous to condition (i) in the theorem, except that we only impose the conditional mean of y_{ijt} to be correctly specified, as already discussed in remark (c) above. Notice

also that this assumption requires conditional independence across i and j, but we do not have to restrict the dependence of y_{ijt} over t for our results.

We consider the Poisson log-likelihood in this paper, which after profiling out η_{ij} gives the pseudo-log-likelihood function $\ell_{ij}(\beta, \alpha_{it}, \gamma_{jt})$ defined in equation (11). This log-likelihood is strictly concave and arbitrarily often differentiable in the parameters, so corresponding assumptions in Theorem 1 are automatically satisfied. Assumption A(ii) is therefore already sufficient for the corresponding assumptions (ii) and (iii) in Theorem 1. Finally, Assumption A(iii) simply corresponds to the condition $\overline{W} > 0$, which is just an appropriate non-collinearity condition on the regressors x_{ijt}.

Translation to our main text notation

The main difference between Theorem 1 in the Appendix and Proposition 3 in the main text is that Theorem 1 only covers the case where $\pi_{ij} = \alpha_i + \gamma_j$ is a scalar, while in our model in the main text α_i, γ_j and $\pi_{ij} = \alpha_i + \gamma_j$ are all T-vectors. We can impose additional normalizations on those T-vectors, because the profile likelihood $\mathcal{L}(\beta, \alpha, \gamma)$ in (10) is invariant under parameter transformations $\alpha_i \mapsto \alpha_i + c_i\, \iota_T$ and $\gamma_j \mapsto \gamma_j + d_j\, \iota_T$ for arbitrary $c_i, d_j \in \mathbb{R}$, where $\iota_T = (1, \ldots, 1)'$ is the T-vector of ones.[2] In the following we choose the normalizations $\iota_T' \alpha_i = 0$ and $\iota_T' \gamma_j = 0$, implying $\iota_T' \pi_{ij} = 0$ for all i, j. Accounting for this normalization we actually only have $(T-1)$ fixed effects α_i and γ_j for each i, j here. Theorem 1 is therefore directly applicable to the case $T = 2$, but for $T > 2$ we need to provide an appropriate extension.

The appropriate generalization of the operator $\mathcal{D}_{\beta\alpha_i^q} = \mathcal{D}_{\beta\gamma_j^q}$ in (18) to the case of vector-valued α_i and γ_j was described in Section 4.2 of Fernández-Val and Weidner (2018). Remember the definition of $\ell_{ij}(\beta, \pi_{ij}) = \ell_{ij}(\beta, \alpha_i, \gamma_j)$ and $\widetilde{x}_{ij} := x_{ij} - \alpha_i^x - \gamma_j^x$. Then, by reparameterizing the pseudo-log-likelihood $\ell_{ij}(\beta, \alpha_i, \gamma_j)$ as follows

$$\ell_{ij}^*(\beta, \alpha_i, \gamma_j) := \ell_{ij}(\beta, \pi_{ij} - \beta'(\alpha_i^x + \gamma_j^x)) = \ell_{ij}(\beta, \alpha_i - \beta' \alpha_i^x, \gamma_j - \beta' \gamma_j^x) \tag{21}$$

one achieves that the expected Hessian of $\mathcal{L}^*(\beta, \alpha, \gamma) = \sum_{i,j} \ell_{ij}^*(\beta, \alpha_i, \gamma_j)$ is block-diagonal, in the sense that $\mathbb{E}\, \partial_{\beta\alpha_i} \mathcal{L}^*(\beta_0, \alpha_0, \gamma_0) = 0$ and $\mathbb{E}\, \partial_{\beta\gamma_j} \mathcal{L}^*(\beta_0, \alpha_0, \gamma_0) = 0$ — the definition of α_i^x and γ_j^x by equation (16) earlier in this Appendix exactly corresponds to those

[2] Those invariances $\alpha_i \mapsto \alpha_i + c_i\, \iota_T$ and $\gamma_j \mapsto \gamma_j + d_j\, \iota_T$ correspond to parameter transformations that in the original model could be absorbed by the parameters η_{ij}.

block-diagonality conditions. With those definitions, we then have that

$$\mathcal{D}_{\beta\alpha_i^q}\ell_{ij} = \partial_{\beta\alpha_i^q}\ell_{ij}^* = \widetilde{x}_{ij}\,\partial_{\alpha_i^{q+1}}\ell_{ij}.$$

In particular, we find that our definitions of

$$W_N = \frac{1}{N\,(N-1)}\sum_{i=1}^{N}\sum_{j\in\mathfrak{N}\setminus\{i\}}\widetilde{x}_{ij}'\,\bar{H}_{ij}\,\widetilde{x}_{ij},$$
$$\Omega_N = \frac{1}{N\,(N-1)}\sum_{i=1}^{N}\sum_{j\in\mathfrak{N}\setminus\{i\}}\widetilde{x}_{ij}'\,\left[\mathrm{Var}\left(S_{ij}\,\middle|\,x_{ij}\right)\right]\,\widetilde{x}_{ij},$$

in Proposition 3 correspond to $-\frac{1}{N(N-1)}\sum_{i,j}\mathbb{E}\left(\partial_{\beta\beta'}\ell_{ij} - \partial_{\alpha_i^2}\ell_{ij}\Xi_{ij}\Xi_{ij}'\right)$ and $\frac{1}{N(N-1)}\sum_{i,j}\mathbb{E}\Big[\mathcal{D}_\beta\ell_{ij}(\mathcal{D}_\beta\ell_{ij})'\Big]$ in the notation of Theorem 1 and equation (20). Thus, the asymptotic variance in (19) indeed corresponds to the asymptotic variance formula in Proposition 3.

Inverse expected incidental parameter Hessian

The asymptotic bias results that follow require that we first derive some key properties of the expected Hessian with respect to the incidental parameters. Remember the definitions of the $2NT$-vector $\phi = \mathrm{vec}(\alpha,\gamma)$ from the main text. The expected incidental parameter Hessian is the $2NT\times 2NT$ matrix given by

$$\bar{\mathcal{H}} := \mathbb{E}\left[-\partial_{\phi\phi'}\mathcal{L}(\beta_0,\phi_0)\right] = \begin{pmatrix} \bar{\mathcal{H}}_{(\alpha\alpha)} & \bar{\mathcal{H}}_{(\alpha\gamma)} \\ [\bar{\mathcal{H}}_{(\alpha\gamma)}]' & \bar{\mathcal{H}}_{(\gamma\gamma)} \end{pmatrix},$$

where $\mathcal{L}(\beta,\phi) = \mathcal{L}(\beta,\alpha,\gamma)$ is defined in (10), and $\bar{\mathcal{H}}_{(\alpha\alpha)}$, $\bar{\mathcal{H}}_{(\alpha\gamma)}$ and $\bar{\mathcal{H}}_{(\gamma\gamma)}$ are $NT\times NT$ submatrices. Here and in the following all expectations are conditional on all the regressor realizations. The matrix $\bar{\mathcal{H}}_{(\alpha\alpha)} = \mathbb{E}\left[-\partial_{\alpha\alpha'}\mathcal{L}(\beta_0,\phi_0)\right]$ is block-diagonal with N non-zero diagonal $T\times T$ blocks given by $\mathbb{E}\left[-\partial_{\alpha_i\alpha_i'}\mathcal{L}(\beta_0,\phi_0)\right] = \sum_{j\in\mathfrak{N}\setminus\{i\}}\bar{H}_{ij}$, because for $i\neq j$ we have $\mathbb{E}\left[-\partial_{\alpha_i\alpha_j'}\mathcal{L}(\beta_0,\phi_0)\right] = 0$, since the parameters α_i and α_j never enter into the same observation. Analogously, the matrix $\bar{\mathcal{H}}_{(\gamma\gamma)} = \mathbb{E}\left[-\partial_{\gamma\gamma'}\mathcal{L}(\beta_0,\phi_0)\right]$ is block-diagonal with N non-zero diagonal $T\times T$ blocks given by $\sum_{i\in\mathfrak{N}\setminus\{j\}}\bar{H}_{ij}$. By contrast, the matrix $\bar{\mathcal{H}}_{(\alpha\gamma)}$ consistents of blocks $\mathbb{E}\left[-\partial_{\alpha_i\gamma_j'}\mathcal{L}(\beta_0,\phi_0)\right] = \bar{H}_{ij}$ that are non-zero for $i\neq j$, because any two parameters α_i and γ_j jointly enter into T observations. The incidental parameter Hessian matrix $\bar{\mathcal{H}}$ therefore has strong diagonal $T\times T$ blocks of order N, but also many off-diagonal elements of order one.

The pseudoinverse of $\bar{\mathcal{H}}$ crucially enters in the stochastic expansion for $\widehat{\beta}$ below. It is therefore necessary to understand the asymptotic properties of this pseudoinverse $\bar{\mathcal{H}}^\dagger$. The

following lemma shows that $\bar{\mathcal{H}}^{\dagger}$ has a structure analogous to $\bar{\mathcal{H}}$, namely, strong diagonal $T \times T$ blocks of order $1/N$, and much smaller off-diagonal elements of order $1/N^2$. We can write $\bar{\mathcal{H}} = \mathfrak{D} + \mathcal{G}$, where

$$\mathfrak{D} := \begin{pmatrix} \bar{\mathcal{H}}_{(\alpha\alpha)} & 0_{NT\times NT} \\ 0_{NT\times NT} & \bar{\mathcal{H}}_{(\gamma\gamma)} \end{pmatrix}, \qquad \mathcal{G} := \begin{pmatrix} 0_{NT\times NT} & \bar{\mathcal{H}}_{(\alpha\gamma)} \\ [\bar{\mathcal{H}}_{(\alpha\gamma)}]' & 0_{NT\times NT} \end{pmatrix}.$$

The matrix $\mathfrak{D}$ is block-diagonal, and its pseudoinverse $\mathfrak{D}^{\dagger}$ is therefore also block-diagonal with $T\times T$ blocks on its diagonal given by $\left(\sum_{j\in\mathfrak{N}\setminus\{i\}} \bar{H}_{ij}\right)^{\dagger}$, $i = 1, \ldots, N$ and $\left(\sum_{i\in\mathfrak{N}\setminus\{j\}} \bar{H}_{ij}\right)^{\dagger}$, $j = 1, \ldots, N$. Thus, $\mathfrak{D}^{\dagger}$ has diagonal elements of order N^{-1}. For any matrix A we denote by $\|A\|_{\max}$ the maximum over the absolute values of all elements of A.

Lemma 1. *Under Assumption A we have, as $N \to \infty$,*

$$\left\|\bar{\mathcal{H}}^{\dagger} - \mathfrak{D}^{\dagger}\right\|_{\max} = O_P\left(N^{-2}\right).$$

This result is crucial in order to derive the stochastic expansion of $\widehat{\beta}$. Indeed, as we will see below, once Lemma 1 is available, then the proof of Proposition 3 is a straightforward extension of the proof of Theorem 4.1 in FW. Lemma 1 is analogous to Lemma D.1 in FW, but our proof strategy for Lemma 1 is different here, because we need to account for the vector-valued nature of α_i and γ_j, which requires new arguments.

Proof of Lemma 1. # Expansion of $\bar{\mathcal{H}}^{\dagger}$ in powers of $\mathcal{G}$: The matrix $\bar{\mathcal{H}}$ is (minus) the expected Hessian of the profile log-likelihood $\mathcal{L} = \sum_{i,j} \ell_{ij}$. Because in that objective function we have already profiled out the fixed effect parameters η_{ij} we have $\bar{\mathcal{H}}_{ij}\iota_T = 0$ for all i, j, where $\iota_T = (1, \ldots, 1)'$ is the T-vector of ones. This implies that

$$\bar{\mathcal{H}}\left(\mathbb{I}_{2N} \otimes \iota_T\right) = 0. \tag{22}$$

The last equation describes $2N$ zero-eigenvectors of $\bar{\mathcal{H}}$ (i.e. the eigenvalue zero of $\bar{\mathcal{H}}$ has multiplicity at least $2N$). Because the original log-likelihood function of the Poisson model was strictly concave in the single index $x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}$ it must be the case that any additional zero-eigenvalue of $\bar{\mathcal{H}}$ is due to linear transformations of the parameters α and γ that leave $\alpha_{it} + \gamma_{jt}$ unchanged for all i, j, t.[3] There is exactly one such transformation for every $t \in \{1, \ldots, T\}$, namely the likelihood is invariant under $\alpha_{it} \mapsto \alpha_{it} + c_t$ and

[3]Notice that any collinearity problem in the likelihood involving the regression parameters β is ruled out for sufficiently large sample sizes by our assumption that $\lim_{N\to\infty} W_N > 0$, which guarantees that the expected Hessian wrt β is positive definite asymptotically.

$\gamma_{jt} \mapsto \gamma_{jt} - c_t$ for any $c_t \in \mathbb{R}$. The expected Hessian $\bar{\mathcal{H}}$ therefore has additional zero-eigenvectors, which are given by the columns of the $2NT \times T$ matrix

$$v := (\iota_N', -\iota_N')' \otimes M_{\iota_T}, \tag{23}$$

where $M_{\iota_T} := \mathbb{I}_T - P_{\iota_T}$ and $P_{\iota_T} := T^{-1}\iota_T\iota_T'$. In the last display we could have used the identity matrix $\mathbb{I}_T$ instead of M_{ι_T}, but we want the columns of v to be orthogonal to the zero-eigenvectors already given by (22), which is achieved by using M_{ι_T}. As a consequence of this, we have $\text{rank}(v) = T - 1$; that is, since we already have (22) we only find $T - 1$ additional zero-eigenvectors here. Thus, the total number of zero eigenvalues of $\bar{\mathcal{H}}$ (i.e. the multiplicity of the eigenvalue zero) is equal to $2N + T - 1$. It is easy to verify that indeed

$$\bar{\mathcal{H}} v = 0. \tag{24}$$

Equations (22) and (24) describe all the zero-eigenvectors of $\bar{\mathcal{H}}$. The projector onto the null-space of $\bar{\mathcal{H}}$ is therefore given by

$$P_{\text{null}} := \mathbb{I}_{2N} \otimes P_{\iota_T} + P_v, \tag{25}$$

where $P_v = v(v'v)^{\dagger}v'$. The Moore-Penrose pseudoinverse of $\bar{\mathcal{H}}$ therefore satisfies

$$\bar{\mathcal{H}}\,\bar{\mathcal{H}}^{\dagger} = \bar{\mathcal{H}}^{\dagger}\,\bar{\mathcal{H}} = \mathbb{I}_{2NT} - P_{\text{null}} = M_{(\iota_N', -\iota_N')'} \otimes M_{\iota_T}, \tag{26}$$

where the RHS is the projector orthogonal to the null-space of $\bar{\mathcal{H}}$ (i.e. the projector onto the span of $\bar{\mathcal{H}}$). The definition of the Moore-Penrose pseudoinverse guarantees that $\bar{\mathcal{H}}^{\dagger}$ has the same zero-eigenvectors as $\bar{\mathcal{H}}$; that is, we also have $\bar{\mathcal{H}}^{\dagger} v = 0$ and $\bar{\mathcal{H}}^{\dagger}\,(\mathbb{I}_{2N} \otimes \iota_T) = 0$. The last equation together with the symmetry of $\bar{\mathcal{H}}^{\dagger}$ implies that

$$(\mathbb{I}_{2N} \otimes P_{\iota_T})\,\bar{\mathcal{H}}^{\dagger} = 0. \tag{27}$$

Next, similar to the above argument for $\bar{\mathcal{H}}$ we have that the only zero-eigenvector of the $T \times T$ matrices $\sum_{j \in \mathfrak{N}\setminus\{i\}} \bar{H}_{ij}$ and $\sum_{i \in \mathfrak{N}\setminus\{j\}} \bar{H}_{ij}$ is given by ι_T, and therefore we have

$$\left(\sum_{j \in \mathfrak{N}\setminus\{i\}} \bar{H}_{ij}\right)\left(\sum_{j \in \mathfrak{N}\setminus\{i\}} \bar{H}_{ij}\right)^{\dagger} = M_{\iota_T}, \qquad \left(\sum_{i \in \mathfrak{N}\setminus\{j\}} \bar{H}_{ij}\right)\left(\sum_{i \in \mathfrak{N}\setminus\{j\}} \bar{H}_{ij}\right)^{\dagger} = M_{\iota_T},$$

which can equivalently be written as

$$\mathfrak{D}^{\dagger}\,\mathfrak{D} = \mathfrak{D}\,\mathfrak{D}^{\dagger} = \mathbb{I}_{2N} \otimes M_{\iota_T} = \mathbb{I}_{2NT} - \mathbb{I}_{2N} \otimes P_{\iota_T}, \tag{28}$$

where $P_{\iota_T} := T^{-1}\iota_T\iota_T'$. Now, using (26) and $\bar{\mathcal{H}} = \mathfrak{D} + \mathcal{G}$ we have

$$\bar{\mathcal{H}}^\dagger (\mathfrak{D} + \mathcal{G}) = \mathbb{I}_{2NT} - P_{\text{null}}.$$

Multiplying this with $\mathfrak{D}^\dagger$ from the right, using (28) and (27), and bringing $\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger$ to the RHS gives

$$\bar{\mathcal{H}}^\dagger = \mathfrak{D}^\dagger - P_{\text{null}}\mathfrak{D}^\dagger - \bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger. \tag{29}$$

By transposing this last equation we obtain

$$\bar{\mathcal{H}}^\dagger = \mathfrak{D}^\dagger - \mathfrak{D}^\dagger P_{\text{null}} - \mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger, \tag{30}$$

and now plugging (29) into the RHS of (30) gives

$$\begin{aligned}\bar{\mathcal{H}}^\dagger &= \mathfrak{D}^\dagger - \mathfrak{D}^\dagger P_{\text{null}} - \mathfrak{D}^\dagger \mathcal{G}\mathfrak{D}^\dagger + \mathfrak{D}^\dagger \mathcal{G} P_{\text{null}}\mathfrak{D}^\dagger - \mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger \\ &= \mathfrak{D}^\dagger - \mathfrak{D}^\dagger \mathcal{G}\mathfrak{D}^\dagger - \mathfrak{D}^\dagger P_{\text{null}} - P_{\text{null}}\mathfrak{D}^\dagger + \mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger,\end{aligned}$$

where in the second step we used that $\mathfrak{D}^\dagger \mathcal{G} P_{\text{null}} = -P_{\text{null}}$, which follows from $0 = \bar{\mathcal{H}}\, P_{\text{null}} = \mathfrak{D} P_{\text{null}} + \mathcal{G} P_{\text{null}}$ by left-multiplication with $\mathfrak{D}^\dagger$ and using that $\mathfrak{D}^\dagger \mathfrak{D} P_{\text{null}} = 0$. This expansion argument for $\bar{\mathcal{H}}^\dagger$ so far has followed the proof of Theorem 2 in Jochmans and Weidner (2019). We furthermore have here that $\mathfrak{D}^\dagger (\mathbb{I}_{2N} \otimes P_{\iota_T}) = 0$, because $\bar{H}_{ij}\iota_T = 0$, implying that $\mathfrak{D}^\dagger P_{\text{null}} = \mathfrak{D}^\dagger P_v$. The expansion in the last display therefore becomes

$$\bar{\mathcal{H}}^\dagger - \mathfrak{D}^\dagger = -\mathfrak{D}^\dagger \mathcal{G}\mathfrak{D}^\dagger - \mathfrak{D}^\dagger P_v - P_v \mathcal{D}^\dagger + \mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger, \tag{31}$$

with $2NT \times T$ matrix v defined in (23). This expansion is the first key step in the proof of the lemma.

Bound on the spectral norm of $\bar{\mathcal{H}}^\dagger$: The term $\mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger$ in the expansion (31) still contains $\bar{\mathcal{H}}^\dagger$ itself. In order to bound contributions from this term we therefore need a preliminary bound on the spectral norm of $\bar{\mathcal{H}}^\dagger$.

The objective function $\ell_{ij}(\beta, \pi_{ij}) := \ell_{ij}(\beta, \alpha_{it}, \gamma_{jt})$ in (11) is strictly convex in π_{ij}, apart from the flat direction given by the invariance $\pi_{ij} \mapsto \pi_{ij} + c_{ij}\,\iota_T$, $c_{it} \in \mathbb{R}$. This strict convexity together with our Assumption A(ii) that all regressors and parameters are uniformly bounded over i, j, N, T implies that for the $T \times T$ expected Hessian $\bar{H}_{ij} := \mathbb{E}\left[-\partial^2 \ell_{ij}/\partial\pi_{ij}\partial\pi_{ij}'(\beta_0, \alpha_0, \gamma_0)\right]$ there exists a constant $b > 0$ that is independent of i, j, N, T such that

$$\min_{\{v \in \mathbb{R} : \iota_T' v = 0\}} v' \bar{H}_{ij} v \geq b > 0.$$

The last display states that $\bar{H}_{ij}$ is positive definite in all directions orthogonal to ι_T. Again, the lower bound $b > 0$ holds uniformly due to Assumption A(ii). The last display result can equivalently be written as

$$\bar{H}_{ij} \geq b\, M_{\iota_T}, \tag{32}$$

where $\geq$ means that the difference between the matrices is positive definite.

Next, let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ be the i'th standard unit vector of dimension N. For all $i, j \in \mathfrak{N} := \{1, \ldots, N\}$ we then have

$$\partial_\phi \pi'_{ij} = \begin{pmatrix} e_i \\ e_j \end{pmatrix} \otimes \mathbb{I}_T,$$

which are $2NT \times T$ matrices. Because $\mathcal{L}(\beta, \phi) = \sum_{i=1}^N \sum_{j \in \mathfrak{N}\setminus\{i\}} \ell_{ij}(\beta, \pi_{ij})$ we thus find that

$$\begin{aligned}
\bar{\mathcal{H}} = \mathbb{E}\left[-\partial_{\phi\phi'}\mathcal{L}\right] &= \sum_{i=1}^N \sum_{j \in \mathfrak{N}\setminus\{i\}} \left(\partial_\phi \pi'_{ij}\right) \mathbb{E}\left[-\partial_{\pi_{ij}\pi'_{ij}} \ell_{ij}\right] \left(\partial_\phi \pi'_{ij}\right)' \\
&= \sum_{i=1}^N \sum_{j \in \mathfrak{N}\setminus\{i\}} \left[\begin{pmatrix} e_i \\ e_j \end{pmatrix} \otimes \mathbb{I}_T\right] \bar{H}_{ij} \left[\begin{pmatrix} e_i \\ e_j \end{pmatrix} \otimes \mathbb{I}_T\right]' \\
&\geq b \sum_{i=1}^N \sum_{j \in \mathfrak{N}\setminus\{i\}} \left[\begin{pmatrix} e_i \\ e_j \end{pmatrix} \otimes \mathbb{I}_T\right] M_{\iota_T} \left[\begin{pmatrix} e_i \\ e_j \end{pmatrix} \otimes \mathbb{I}_T\right]' \\
&= b \left[\sum_{i=1}^N \sum_{j \in \mathfrak{N}\setminus\{i\}} \begin{pmatrix} e_i \\ e_j \end{pmatrix} \begin{pmatrix} e_i \\ e_j \end{pmatrix}'\right] \otimes M_{\iota_T} \\
&= b \underbrace{\begin{pmatrix} (N-1)\mathbb{I}_N & \iota_N \iota'_N - \mathbb{I}_N \\ \iota_N \iota'_N - \mathbb{I}_N & (N-1)\mathbb{I}_N \end{pmatrix}}_{=:Q_N} \otimes M_{\iota_T}
\end{aligned}$$

where we also used (32). It is easy to show that for $N > 2$ the $2N \times 2N$ matrix Q_N has an eigenvalue zero with multiplicity one, an eigenvalue $N-2$ with multiplicity $N-1$, an eigenvalue N with multiplicity $N-1$, and an eigenvalue $2(N-1)$ with multiplicity one. Thus, the smallest non-zero eigenvalue of Q_N is $(N-2)$. Also, the zero-eigenvector of Q_N is given by $v_0 := (\iota'_N, -\iota'_N)'$, and therefore we have $Q_N \geq (N-2)\, M_{v_0}$, where $M_{v_0} = \mathbb{I}_{2N} - (2N)^{-1} v_0 v'_0$ is the projector orthogonal to v_0. We therefore have

$$\begin{aligned}
\bar{\mathcal{H}} &\geq b\,(N-2)\, M_{(\iota'_N, -\iota'_N)'} \otimes M_{\iota_T} \\
&= b\,(N-2)\, \left(\mathbb{I}_{2NT} - P_{\text{null}}\right),
\end{aligned}$$

where P_{null} is the projector onto the null-space of $\bar{\mathcal{H}}$, as already defined above. From this it follows that

$$\bar{\mathcal{H}}^{\dagger} \leq \frac{1}{b\,(N-2)}\left(\mathbb{I}_{2NT} - P_{\text{null}}\right),$$

and therefore for the spectral norm

$$\left\|\bar{\mathcal{H}}^{\dagger}\right\| \leq \frac{1}{b\,(N-2)} = O(1/N). \tag{33}$$

Final bound on $\left\|\bar{\mathcal{H}}^{\dagger} - \mathfrak{D}^{\dagger}\right\|_{\max}$: Using (32) we find

$$\max_{i\in\mathfrak{N}}\left(\frac{1}{N-1}\sum_{j\in\mathfrak{N}\setminus\{i\}}\bar{H}_{ij}\right)^{\dagger} = O_P(1), \qquad \max_{j\in\mathfrak{N}}\left(\frac{1}{N-1}\sum_{i\in\mathfrak{N}\setminus\{j\}}\bar{H}_{ij}\right)^{\dagger} = O_P(1).$$

This together with our boundedness Assumption A(ii) implies that

$$\left\|\mathfrak{D}^{\dagger}\right\|_{\max} = O_P(1/N), \qquad \|\mathcal{G}\|_{\max} = O_P(1). \tag{34}$$

The definition of the $2NT \times T$ matrix v in (23) implies that

$$\begin{aligned}\|P_v\|_{\max} &= \left\|P_{(\iota_N',-\iota_N')'}\otimes M_{\iota_T}\right\|_{\max} \leq \left\|P_{(\iota_N',-\iota_N')'}\right\|_{\max} = (2N)^{-1}\left\|(\iota_N',-\iota_N')'(\iota_N',-\iota_N')\right\|_{\max} \\ &= (2N)^{-1} = O(1/N),\end{aligned} \tag{35}$$

where we also used that $\|M_{\iota_T}\|_{\max} \leq 1$. In the following display, let $e_k = (0,\ldots,0,1,0,\ldots,0)'$ be the k'th standard unit vector of dimension $2NT$. We find that

$$\begin{aligned}\left\|\mathcal{G}\bar{\mathcal{H}}^{\dagger}\mathcal{G}\right\|_{\max} &= \max_{k,\ell\in\{1,\ldots,2NT\}}\left|e_k'\,\mathcal{G}\bar{\mathcal{H}}^{\dagger}\mathcal{G}e_\ell\right| \\ &\leq \left(\max_{k\in\{1,\ldots,2NT\}}\|\mathcal{G}e_k\|\right)\left\|\bar{\mathcal{H}}^{\dagger}\right\|\left(\max_{\ell\in\{1,\ldots,2NT\}}\|\mathcal{G}e_\ell\|\right) \\ &= \left(\max_{k\in\{1,\ldots,2NT\}}\|\mathcal{G}e_k\|\right)^2\left\|\bar{\mathcal{H}}^{\dagger}\right\| \\ &\leq \left(\sqrt{2NT}\,\|\mathcal{G}\|_{\max}\right)^2\left\|\bar{\mathcal{H}}^{\dagger}\right\| \\ &= O_P(1),\end{aligned} \tag{36}$$

where the first line is just the definition of $\|\cdot\|_{\max}$, the second step uses definition of the spectral norm $\left\|\bar{\mathcal{H}}^{\dagger}\right\|$, the third step is an obvious rewriting, the fourth step uses that the norm of $2NT$-vector $\mathcal{G}e_k$ can at most be $\sqrt{2NT}$ times the maximal absolute element of

the vector, and the final step uses that T is fixed in our asymptotic and $\|\mathcal{G}\|_{\max} = O_P(1)$ and also (33).

Next, for general $2NT \times 2NT$ matrices A and B we have the bound (notice that $\|\cdot\|_{\max}$ is not a matrix norm)

$$\|AB\|_{\max} \leq 2NT \, \|A\|_{\max} \, \|B\|_{\max},$$

but because $\mathfrak{D}$ is block-diagonal (with non-zero $T \times T$ blocks on the diagonal) we have for any $2NT \times 2NT$ matrix A the much improved bound

$$\|\mathfrak{D}A\|_{\max} \leq T \, \|\mathfrak{D}\|_{\max} \, \|A\|_{\max}.$$

Applying those inequalities to the expansion of $\bar{\mathcal{H}}^{\dagger} - \mathfrak{D}^{\dagger}$ obtained from (31), and also using (34) and (35) and (36), we find that

$$\begin{aligned}\left\|\bar{\mathcal{H}}^{\dagger} - \mathfrak{D}^{\dagger}\right\|_{\max} &\leq T^2 \left\|\mathfrak{D}^{\dagger}\right\|_{\max}^2 \|\mathcal{G}\|_{\max} + 2T \left\|\mathfrak{D}^{\dagger}\right\|_{\max} \|P_v\|_{\max} + T^2 \left\|\mathfrak{D}^{\dagger}\right\|_{\max}^2 \left\|\mathcal{G}\bar{\mathcal{H}}^{\dagger}\mathcal{G}\right\|_{\max} \\ &= O_P(1/N^2),\end{aligned}$$

as $N \to \infty$ (remember that T is fixed in our asymptotic.) This is what we wanted to show. ■

Proof of Proposition 3

The pseudo-likelihood function of the Poisson model is strictly concave in the single index. Therefore, Assumption A together with Lemma 1 guarantee that the conditions of Theorem B.1 in Fernández-Val and Weidner (2016) are satisfied for the rescaled and penalized objective function[4]

$$\frac{1}{\sqrt{N(N-1)}} \mathcal{L}(\beta, \phi) - \frac{1}{2} \phi' \, P_{\text{null}} \, \phi,$$

with P_{null} defined in (25). Here, the penalty term $\phi' \, P_{\text{null}} \, \phi$ guarantees *strict* concavity in (β, ϕ). However, in the following all derivatives of $\mathcal{L}(\beta, \phi)$ are evaluated at the true parameters, and since we impose the normalization $P_{\text{null}} \, \phi_0 = 0$ the only derivative of

[4]Since we have a concave objective function, we can apply Theorem B.3 in FW to obtain preliminary convergence results for both $\widehat{\beta}$ and $\widehat{\phi}$. That theorem guarantees that that the consistency condition on $\widehat{\phi}(\beta)$ in Assumption (iii) of Theorem B.1 in FW is satisfied under our Assumption A, and it also guarantees $\left\|\widehat{\beta} - \beta^0\right\| = O_P(N^{-1/2})$, which is important to apply Corollary B.2 in FW to obtain the expansion result in our equation (37).

$\mathcal{L}(\beta,\phi)$ where the penalty term gives a non-zero contribution is the incidental parameter Hessian matrix $\bar{\mathcal{H}} = \mathbb{E}\left[-\partial_{\phi\phi'}\mathcal{L}(\beta_0,\phi_0)\right]$ for which the penalty term provides exactly the correct regularization. However, instead of that regularization, we can equivalently use the pseudoinverse; namely we have

$$\left(\bar{\mathcal{H}} + c\, P_{\text{null}}\right)^{-1} = \bar{\mathcal{H}}^{\dagger} + \frac{1}{c}\, P_{\text{null}},$$

for any $c > 0$. In all expressions below where $\bar{\mathcal{H}}^{\dagger}$ appears we could equivalently write $\bar{\mathcal{H}}^{\dagger} + \frac{1}{N} P_{\text{null}}$, but the additional contributions from $\frac{1}{N} P_{\text{null}}$ will always vanish because the gradient of $\mathcal{L}(\beta,\phi)$ with respect to ϕ is orthogonal to P_{null}.

By applying Theorem B.1 and its Corollary B.2 in FW we thus obtain

$$\sqrt{N(N-1)}(\widehat{\beta} - \beta^0) = W_N^{-1} U_N + o_P(1), \tag{37}$$

where

$$\begin{aligned} W_N &= -\frac{1}{N(N-1)} \left(\partial_{\beta\beta'}\bar{\mathcal{L}} + [\partial_{\beta\phi'}\bar{\mathcal{L}}]\, \bar{\mathcal{H}}^{\dagger}\, [\partial_{\phi\beta'}\bar{\mathcal{L}}]\right) \\ &= -\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in \mathfrak{N}\setminus\{i\}} \partial_{\beta\beta'} \bar{\ell}^*_{ij} \end{aligned}$$

was already defined in Proposition 3, and we have $U_N := U_N^{(0)} + U_N^{(1)}$, with

$$\begin{aligned} U_N^{(0)} &= \frac{1}{\sqrt{N(N-1)}} \left[\partial_{\beta}\mathcal{L} + [\partial_{\beta\phi'}\bar{\mathcal{L}}]\, \bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}\right] = \frac{1}{\sqrt{N(N-1)}}\, \partial_{\beta}\mathcal{L}^* \\ &= \frac{1}{\sqrt{N(N-1)}} \sum_{i=1}^{N} \sum_{j \in \mathfrak{N}\setminus\{i\}} \partial_{\beta} \ell^*_{ij}, \end{aligned}$$

$$\begin{aligned} \sqrt{N(N-1)}\, U_N^{(1)} &= [\partial_{\beta\phi'}\mathcal{L} - \partial_{\beta\phi'}\bar{\mathcal{L}}]\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} - [\partial_{\beta\phi'}\bar{\mathcal{L}}]\, \bar{\mathcal{H}}^{\dagger}\, \left[\mathcal{H} - \bar{\mathcal{H}}\right]\, \bar{\mathcal{H}}^{\dagger}\, \partial_{\phi}\mathcal{L} \\ &\quad + \frac{1}{2} \sum_{g=1}^{\dim\phi} \left(\partial_{\beta\phi'\phi_g}\bar{\mathcal{L}} + [\partial_{\beta\phi'}\bar{\mathcal{L}}]\, \bar{\mathcal{H}}^{\dagger}[\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}]\right) [\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}]_g \bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} \\ &= [\partial_{\beta\phi'}\mathcal{L}^* - \partial_{\beta\phi'}\bar{\mathcal{L}}^*]\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} + \frac{1}{2} \sum_{g=1}^{\dim\phi} \partial_{\beta\phi'\phi_g}\bar{\mathcal{L}}^* [\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}]_g \bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}. \end{aligned}$$

Here, ℓ^*_{ij} was defined in (21), all "bars" denote expectations conditional on X and ϕ, and all the partial derivatives are evaluated at the true parameters. We also defined $\mathcal{L}^*(\beta,\phi) := \sum_{i=1}^{N} \sum_{j\in\mathfrak{N}\setminus\{i\}} \ell^*_{ij}(\beta, \alpha_{it}, \gamma_{jt})$. Remember that we use a different scaling of the (profile) likelihood function than FW; namely in (10) we define $\mathcal{L}(\beta,\phi) = \sum_{i=1}^{N} \sum_{j\in\mathfrak{N}\setminus\{i\}} \ell_{ij}(\beta, \alpha_{it}, \gamma_{jt})$, while in FW this function would have an additional factor $1/\sqrt{N(N-1)}$. This explains

the additional $\sqrt{N(N-1)}$ factors in W_N, $U_N^{(0)}$ and $U_N^{(1)}$ as compared to Theorem B.1 in FW.

The score term $\partial_\beta \ell_{ij}^* = \widetilde{x}_{ij}' S_{ij}$ has zero mean and finite variance and is independent across i and j, conditional on X and ϕ. By the central limit theorem we thus find

$$U_N^{(0)} \Rightarrow \mathcal{N}(0, \Omega_N),$$

where

$$\begin{aligned}\Omega_N &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in \mathfrak{N} \setminus \{i\}} \mathrm{Var}\left(\partial_\beta \ell_{ij}^* \,\middle|\, x_{ij}\right) \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in \mathfrak{N} \setminus \{i\}} \widetilde{x}_{ij}' \left[\mathrm{Var}\left(S_{ij} \,\middle|\, x_{ij}\right)\right] \widetilde{x}_{ij}.\end{aligned}$$

Thus, the term $U_N^{(0)}$ only contributes variance to the asymptotic distribution of $\widehat{\beta}$, but no asymptotic bias. By contrast, the term $U_N^{(1)}$ only contributes bias to the asymptotic distribution of $\widehat{\beta}$, but no variance. Namely, one finds that

$$U_N^{(1)} \;\to_p\; B_N + D_N, \tag{38}$$

with B_N and D_N as given in the proposition. The proof of (38) is exactly analogous to the corresponding discussion of those terms in the proof of Theorem 4.1 in FW, which we restated above as Theorem 1 (remember that for $T = 2$ our result here is indeed just a special case of Theorem 4.1 in FW.) Therefore, instead of repeating those derivations here, we provide in the following a slightly less rigorous, but much easier to follow, derivation of those bias terms.

Derivation of the asymptotic bias in Proposition 3

Remember that the main difference between Theorem 1 and our case here is that for us the incidental parameters α_i and γ_j are T-vectors, while in Theorem 1 the index $\pi_{ij} = \alpha_i + \gamma_j$ is just a scalar. An easy way to generalize the asymptotic bias formulas in Theorem 1 and display (20) to vector-valued incidental parameters is to use a suitable parameterization for the incidental parameters α_i and γ_j. The formulas for $\overline{B}_1$ and $\overline{D}_1$ can most easily be generalized by parameterizing the incidental parameters as follows

$$\alpha_i = A_i \, \widetilde{\alpha}_i, \qquad\qquad \gamma_j = C_j \, \widetilde{\gamma}_j, \tag{39}$$

where $\widetilde{\alpha}_i$ and $\widetilde{\gamma}_j$ are $T-1$ vectors, and A_i and C_j are $T \times (T-1)$ matrices that satisfy

$$A_i A_i' = \left(\sum_j \bar{H}_{ij} \right)^\dagger, \qquad C_j C_j' = \left(\sum_i \bar{H}_{ij} \right)^\dagger. \tag{40}$$

Let $\widetilde{\mathcal{L}}(\beta, \widetilde{\alpha}, \widetilde{\gamma}) = \mathcal{L}(\beta, (A_i\, \widetilde{\alpha}_i), (C_j\, \widetilde{\gamma}_j))$. This reparameterization guarantees that

$$\begin{aligned} \frac{\partial^2 \widetilde{\mathcal{L}}(\beta^0, \widetilde{\alpha}^0, \widetilde{\gamma}^0)}{(\partial \widetilde{\alpha}_i)(\partial \widetilde{\alpha}_i)'} &= A_i' \left(\sum_j \bar{H}_{ij} \right) A_i = \mathbb{I}_{T-1}, \\ \frac{\partial^2 \widetilde{\mathcal{L}}(\beta^0, \widetilde{\alpha}^0, \widetilde{\gamma}^0)}{(\partial \widetilde{\gamma}_j)(\partial \widetilde{\gamma}_j)'} &= C_j' \left(\sum_i \bar{H}_{ij} \right) C_j = \mathbb{I}_{T-1}. \end{aligned} \tag{41}$$

That is, the Hessian matrix with respect to the incidental parameters $\widetilde{\alpha}_i$ and $\widetilde{\gamma}_j$ is normalized to be an identity matrix under that normalization. It can be shown that this implies that the incidental parameter biases $\overline{B}_1$ and $\overline{D}_1$ "decouple" across the $T-1$ components of $\widetilde{\alpha}_i$ and $\widetilde{\gamma}_j$; that is, the total contribution to the incidental parameter bias of $\widehat{\beta}$ just becomes a sum over $T-1$ contributions of the form $\overline{B}_1$ and $\overline{D}_1$ in (20). Thus, for $k \in \{1, \ldots, K\}$ we have

$$\begin{aligned} B_{1,k} &= \sum_{q=1}^{T-1} \left[-\frac{1}{N} \sum_{i,j} \frac{\mathbb{E}\left(\partial_{\tilde{\alpha}_{i,q}} \ell_{ij} \mathcal{D}_{\beta_k \tilde{\alpha}_{i,q}} \ell_{ij} \right)}{\sum_{j'} \mathbb{E}\left(\partial^2_{\tilde{\alpha}_{i,q}} \ell_{ij'} \right)} \right] = \sum_{q=1}^{T-1} \left[-\frac{1}{N} \sum_{i,j} \mathbb{E}\left(\partial_{\tilde{\alpha}_{i,q}} \ell_{ij} \mathcal{D}_{\beta_k \tilde{\alpha}_{i,q}} \ell_{ij} \right) \right] \\ &= -\frac{1}{N} \sum_{i,j} \mathbb{E}\left[(\partial_{\tilde{\alpha}_i} \ell_{ij})' \left(\mathcal{D}_{\beta_k \tilde{\alpha}_i} \ell_{ij} \right) \right] = -\frac{1}{N} \sum_{i,j} \mathbb{E}\left[(\partial_{\alpha_i} \ell_{ij})' A_i A_i' \left(\mathcal{D}_{\beta_k \alpha_i} \ell_{ij} \right) \right] \\ &= -\frac{1}{N} \sum_{i,j} \mathbb{E}\left[S_{ij}' \left(\sum_{j'} \bar{H}_{ij'} \right)^\dagger H_{ij}\, \widetilde{x}_{ij,k} \right], \end{aligned}$$

where in the second step we used the fact that $\sum_{j'} \mathbb{E}\left(\partial^2_{\tilde{\alpha}_{i,q}} \ell_{ij'} \right) = 1$ according to (41), in the third step we rewrote the sum over $q \in \{1, \ldots, T-1\}$ in terms of the vector product of the $T-1$ vectors $\partial_{\tilde{\alpha}_i} \ell_{ij}$ and $\mathcal{D}_{\beta_k \tilde{\alpha}_i} \ell_{ij}$, in the fourth step we used that $\alpha_i = A_i\, \widetilde{\alpha}_i$, and in the final step we used (40) and the definitions of S_{ij}, H_{ij} and $\widetilde{x}_{ij,k}$. All expectations here are conditional on X (in the main text we always make that conditioning explicit), and $\bar{H}_{ij'}$ and $\widetilde{x}_{ij,k}$ are non-random conditional on X; that is, we can also write this last expression as

$$B_{1,k} = -\frac{1}{N} \sum_i \mathrm{Tr} \left[\left(\sum_{j'} \bar{H}_{ij'} \right)^\dagger \sum_j \mathbb{E}\left(H_{ij}\, \widetilde{x}_{ij,k}\, S_{ij}' \right) \right].$$

Analogously we find

$$D_{1,k} = -\frac{1}{N}\sum_{i,j}\mathbb{E}\left[S'_{ij}\left(\sum_{i'}\bar{H}_{i'j}\right)^{\dagger}H_{ij}\,\widetilde{x}_{ij,k}\right].$$

Next, to generalize the incidental parameter biases $\overline{B}_2$ and $\overline{D}_2$ in (20) to vector-values α_i and γ_j we again make a transformation (39), but this time we choose

$$\begin{aligned} A_iA'_i &= \left(\sum_j \bar{H}_{ij}\right)^{\dagger}\left[\sum_j \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij}\right)\right]\left(\sum_j \bar{H}_{ij}\right)^{\dagger}. \\ C_jC'_j &= \left(\sum_i \bar{H}_{ij}\right)^{\dagger}\left[\sum_i \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij}\right)\right]\left(\sum_i \bar{H}_{ij}\right)^{\dagger}. \end{aligned} \tag{42}$$

Notice that for a correctly specified likelihood we have the Bartlett identities $\bar{H}_{ij} = \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij}\right)$, implying that (40) and (42) are identical for correctly specified likelihoods. In general, however, the transformation now is different. Instead of normalizing the Hessian matrices to be identities, as in (41), the new transformation defined by (42) guarantees that

$$\begin{aligned} \text{AsyVar}\left(\widehat{\widetilde{\alpha}}_i\right) &= \left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\alpha}_i)(\partial\widetilde{\alpha}_i)'}\right]^{\dagger}\text{Var}\left[\frac{\partial\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{\partial\widetilde{\alpha}_i}\middle|X\right]\left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\alpha}_i)(\partial\widetilde{\alpha}_i)'}\right]^{\dagger} = \mathbb{I}_{T-1}, \\ \text{AsyVar}\left(\widehat{\widetilde{\gamma}}_j\right) &= \left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\gamma}_j)(\partial\widetilde{\gamma}_j)'}\right]^{\dagger}\text{Var}\left[\frac{\partial\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{\partial\widetilde{\gamma}_j}\middle|X\right]\left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\gamma}_j)(\partial\widetilde{\gamma}_j)'}\right]^{\dagger} = \mathbb{I}_{T-1}. \end{aligned} \tag{43}$$

Again, it can be shown that with this normalization the incidental parameter bias contributions $\overline{B}_2$ and $\overline{D}_2$ "decouple"; that is, each component of $\widehat{\widetilde{\alpha}}_i$ contributes an incidental parameter bias of the form $\overline{B}_2$ in (20) to $\widehat{\beta}$, and each component of $\widehat{\widetilde{\gamma}}_i$ contributes an incidental parameter bias of the form $\overline{D}_2$ in (20) to $\widehat{\beta}$. The total contribution thus reads, for $k \in \{1,\ldots,K\}$,

$$\begin{aligned} B_{2,k} &= \sum_{q=1}^{T-1}\left[\frac{1}{2}\frac{1}{N}\sum_i \frac{\left[\sum_j \mathbb{E}(\partial_{\tilde{\alpha}_{i,q}}\ell_{ij})^2\right]\sum_j \mathbb{E}(\mathcal{D}_{\beta_k\tilde{\alpha}^2_{i,q}}\ell_{ij})}{\left[\sum_j \mathbb{E}\left(\partial_{\tilde{\alpha}^2_{i,q}}\ell_{ij}\right)\right]^2}\right] \\ &= \sum_{q=1}^{T-1}\frac{1}{2}\frac{1}{N}\sum_{i,j}\mathbb{E}(\mathcal{D}_{\beta_k\tilde{\alpha}^2_{i,q}}\ell_{ij}) = \frac{1}{2}\frac{1}{N}\sum_{i,j}\text{Tr}\left[\mathbb{E}(\mathcal{D}_{\beta_k\,\tilde{\alpha}_i\tilde{\alpha}'_i}\ell_{ij})\right] \\ &= \frac{1}{2}\frac{1}{N}\sum_{i,j}\text{Tr}\left[A'_i\,\mathbb{E}(\mathcal{D}_{\beta_k\,\alpha_i\alpha'_i}\ell_{ij})\,A_i\right] \\ &= \frac{1}{2\,N}\sum_i \text{Tr}\left[\left(\sum_j \bar{G}_{ij}\,\widetilde{x}_{ij,k}\right)\left(\sum_j \bar{H}_{ij}\right)^{\dagger}\left[\sum_j \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij,k}\right)\right]\left(\sum_j \bar{H}_{ij}\right)^{\dagger}\right], \end{aligned}$$

where in the second step we used that $\left[\sum_j \mathbb{E}(\partial_{\tilde{\alpha}_{i,q}} \ell_{ij})^2\right] / \left[\sum_j \mathbb{E}\left(\partial_{\tilde{\alpha}^2_{i,q}} \ell_{ij}\right)\right]^2 = 1$ according to (43), in the third step we rewrote the sum over $q \in \{1, \dots, T-1\}$ as a trace over the $(T-1) \times (T-1)$ matrix of third-order partial derivatives $\mathbb{E}(\mathcal{D}_{\beta_k \tilde{\alpha}_i \tilde{\alpha}'_i} \ell_{ij})$, in the fourth step we used that $\alpha_i = A_i \, \widetilde{\alpha}_i$, and in the final step we used the cyclicity of the trace and (42) and the definitions of $\bar{G}_{ij}$, $\widetilde{x}_{ij,k}$, and the tensor-vector product $\bar{G}_{ij} \widetilde{x}_{ij,k}$ (which, recall, is a $T \times T$ matrix).

Analogously we find

$$
\begin{aligned}
D_{2,k} &= \sum_{q=1}^{T-1} \left[\frac{1}{2} \frac{1}{N} \sum_j \frac{\left[\sum_i \mathbb{E}(\partial_{\tilde{\gamma}_{j,q}} \ell_{ij})^2\right] \sum_i \mathbb{E}(\mathcal{D}_{\beta_k \tilde{\gamma}^2_{j,q}} \ell_{ij})}{\left[\sum_i \mathbb{E}\left(\partial_{\tilde{\gamma}^2_{j,q}} \ell_{ij}\right)\right]^2} \right] \\
&= \frac{1}{2\,N} \sum_j \mathrm{Tr} \left[\left(\sum_i \bar{G}_{ij} \, \widetilde{x}_{ij,k} \right) \left(\sum_i \bar{H}_{ij} \right)^{\dagger} \left[\sum_i \mathbb{E}\left(S_{ij} \, S'_{ij} \middle| x_{ij,k} \right) \right] \left(\sum_i \bar{H}_{ij} \right)^{\dagger} \right].
\end{aligned}
$$

We have thus translated all the formulas in Theorem 1 and in display (20) to the case of vector-valued α_i and γ_j to find exactly the expression for the asymptotic biases $B^k_N = B_{1,k} + B_{2,k}$ and $D^k_N = D_{1,k} + D_{2,k}$ in Proposition 3.

Rewriting the bias expressions as in Appendix A.2.2

In the following, we unpack the formulas provided in Appendix A.2.2 in order to provide additional detail on why the leading bias term is of order $1/(NT)$ as both N and $T \to \infty$ simultaneously. Remember that $\mathbb{E}(y_{ijt} | x_{ijt}, \alpha_{it}, \gamma_{ij}) = \lambda_{ijt} := \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{ij})$ and $\vartheta_{ijt} := \frac{\lambda_{ijt}}{\sum_\tau \lambda_{ij\tau}}$, and denote the corresponding T-vectors by y_{ij}, λ_{ij} and ϑ_{ij}. It is convenient to define the $T \times T$ matrices

$$
\Lambda_{ij} := \mathrm{diag}\left(\lambda_{ij}\right),
$$

and

$$
M_{ij} := \mathbf{I}_T - \frac{\lambda_{ij} \, \iota'_T}{\iota'_T \lambda_{ij}} = \mathbf{I}_T - \vartheta_{ij} \iota'_T,
$$

which is the unique idempotent $T \times T$ matrix (i.e. $M_{ij} M_{ij} = M_{ij}$) that satisfies $\mathrm{rank}(M_{ij}) = T - 1$, $M_{ij} \lambda_{ij} = 0$, and $\iota'_T M_{ij} = 0$. Notice also that $\lambda_{ij} = \Lambda_{ij} \iota_T$, and therefore

$M_{ij}\Lambda_{ij} = \Lambda_{ij}M_{ij}'$. We then have

$$\begin{aligned}
S_{ij} &= M_{ij}' y_{ij}, \\
\bar{H}_{ij} &= M_{ij}\,\Lambda_{ij}\,M_{ij}' = M_{ij}\,\Lambda_{ij} = \Lambda_{ij}M_{ij}' = \Lambda_{ij} - \frac{\lambda_{ij}\lambda_{ij}'}{\iota_T'\lambda_{ij}}, \\
H_{ij} &= \bar{H}_{ij}\left(\frac{\iota_T' y_{ij}}{\iota_T'\lambda_{ij}}\right),
\end{aligned}$$

and

$$\bar{G}_{ij,tsr} = -\sum_{u=1}^{T} \lambda_{ij,u}\, M_{ij,tu}\, M_{ij,su}\, M_{ij,ru},$$

where $t, s, r \in \{1, \ldots, T\}$.

Next, define $\widetilde{x}_{ij,k}^* := M_{ij}'\widetilde{x}_{ij,k}$. Noting that $\lambda_{ij}'\widetilde{x}_{ij,k}^* = 0$, we find

$$\begin{aligned}
W_{N,k\ell} &= \frac{1}{N\,(N-1)} \sum_{i,j} \widetilde{x}_{ij,k}^{*\prime}\,\Lambda_{ij}\,\widetilde{x}_{ij,\ell}^* \\
&= \frac{1}{N\,(N-1)} \sum_{i,j,t} \lambda_{ijt}\,\widetilde{x}_{ijt,k}^*\,\widetilde{x}_{ijt,\ell}^*.
\end{aligned}$$

This shows that W_N has an additional sum over t, so W_N increases linearly in T, and $W_N^{-1} = O(T^{-1})$, for $T \to \infty$.

Now, also define $D_{ij,k} := \text{diag}\left[\left(\lambda_{ijt}\,\widetilde{x}_{ijt,k}^*\right)_{t=1,\ldots,T}\right]$, which is the diagonal $T \times T$ matrix with diagonal entries $\lambda_{ijt}\,\widetilde{x}_{ijt,k}^*$. The first-order conditions of the optimization problem that defines $\widetilde{x}_{ij,k}$ read

$$\sum_i \bar{H}_{ij}\,\widetilde{x}_{ij,k} = 0, \qquad\qquad \sum_j \bar{H}_{ij}\,\widetilde{x}_{ij,k} = 0,$$

or equivalently

$$\sum_i \Lambda_{ij}\,\widetilde{x}_{ij,k}^* = 0, \qquad\qquad \sum_j \Lambda_{ij}\,\widetilde{x}_{ij,k}^* = 0,$$

which can also be written as

$$\sum_i D_{ij,k} = 0, \qquad\qquad \sum_j D_{ij,k} = 0. \tag{44}$$

These FOC's are only important to simplify the term $B_{2,k}$ in what follows. We have

$$\begin{aligned}
B_{1,k} &= -\frac{1}{N}\sum_{i,j}\frac{\mathbb{E}\left[(\iota_T' y_{ij})\, S_{ij}'\right]}{\iota_T'\lambda_{ij}}\left(\sum_{j'}\bar{H}_{ij'}\right)^{\dagger}\Lambda_{ij}\,\widetilde{x}_{ij,k}^{*} \\
&= -\frac{1}{N(N-1)}\sum_{i,j}\frac{\iota_T'}{\iota_T'\lambda_{ij}}\mathrm{Var}(y_{ij})M_{ij}'\left(\frac{1}{N}\sum_{j'}\bar{H}_{ij'}\right)^{\dagger}\Lambda_{ij}M_{ij}'\,\widetilde{x}_{ij,k}, \\
B_{2,k} &= -\frac{1}{2\,N}\sum_{i}\mathrm{Tr}\left\{\left[\sum_{j}M_{ij}\,D_{ij,k}\,M_{ij}'\right]\left(\sum_{j}\bar{H}_{ij}\right)^{\dagger}\left[\sum_{j}M_{ij}\mathrm{Var}(y_{ij})M_{ij}'\right]\left(\sum_{j}\bar{H}_{ij}\right)^{\dagger}\right\} \\
&= \frac{1}{N(N-1)}\sum_{i,j}\left\{\frac{\lambda_{ij}'\,Q_i\,\Lambda_{ij}\,\widetilde{x}_{ij,k}^{*}}{\iota_T'\lambda_{ij}} - \frac{\left(\lambda_{ij}'\;\widetilde{x}_{ij,k}^{*}\right)\left(\lambda_{ij}'Q_i\lambda_{ij}\right)}{\left(\iota_T'\lambda_{ij}\right)^2}\right\} \\
&= \frac{1}{N(N-1)}\sum_{i,j}\frac{\lambda_{ij}'\,Q_i\,\Lambda_{ij}M_{ij}'\,\widetilde{x}_{ij,k}}{\iota_T'\lambda_{ij}},
\end{aligned}$$

where, in the second-to-last step, we used the definition of M_{ij}, (44), that $\iota_T' D_{ij,k}\iota_T = \lambda_{ij}'\;\widetilde{x}_{ij,k}^{*}$, and that $D_{ij,k}\iota_T = \Lambda_{ij}\,\widetilde{x}_{ij,k}^{*}$; and in the last step, we used that $\Lambda_{ij}\,\widetilde{x}_{ij,k}^{*} = \Lambda_{ij}\,M_{ij}'\widetilde{x}_{ij,k}$ and $\lambda_{ij}'\;\widetilde{x}_{ij,k}^{*} = 0$. We also used the definition of Q_i given in Appendix A.2.2. We then have for $B_N^k = B_{1,k} + B_{2,k}$ that

$$B_N^k = -\frac{1}{N(N-1)}\sum_{i,j}\frac{\frac{1}{T}\,\iota_T'\,R_{ij}\,\widetilde{x}_{ij,k}}{\frac{1}{T}\,\iota_T'\lambda_{ij}} + \frac{1}{N(N-1)}\sum_{i,j}\frac{\frac{1}{T}\,\lambda_{ij}'\,Q_i\,\Lambda_{ij}M_{ij}'\,\widetilde{x}_{ij,k}}{\frac{1}{T}\,\iota_T'\lambda_{ij}},$$

where we have now also used the definition of R_{ij} from Appendix A.2.2 in order to simplify $B_{1,k}$. Under appropriate regularity conditions, the $T \times T$ matrices Q_i and R_{ij} each maintain diagonal elements of order one and off-diagonal elements of order $1/T^2$ through their dependence on $\mathrm{Var}(y_{ij})$. Therefore, all the numerators and denominators in the last expression for B_N^k remain of order one as $T \to \infty$, such that $B_N^k = O(1)$ as $T \to \infty$, with an analogous result also following for D_N^k. Recalling that W_N increases linearly with T, we thus conclude that the bias term

$$\frac{W_N^{-1}(B_N + D_N)}{N-1},$$

is of order $1/(NT)$ as both N and T grow large.

Comment on Proposition 1

We note that the consistency result from Proposition 1 also follows from the above proof of Proposition 3:

Remark 3. *If the asymptotic bias in $\widehat{\beta}$ is characterized by Proposition 3, then $\widehat{\beta}$ is consistently estimated as $N \to \infty$.*

As we have noted in the text, for this consistency result to hold, we need for the score of the profile log-likelihood $\ell_{ij}(\beta, \alpha_{it}, \gamma_{jt})$ from (11) to be unbiased when evaluated at the true parameters $(\beta^0, \alpha^0, \gamma^0)$. In particular, we need for there to be no incidental parameter bias term of order $1/T$ associated with the pair fixed effect η_{ij}. As the following proof and subsequent discussion clarify, the FE-PPML estimator is quite special in this regard.

A.4 Proof of Proposition 2

To prove Proposition 2, it will first be useful to prove the following lemma:

Lemma 2. *Assume a "one way" panel data model with $\lambda_{it} = \exp(x_{it}'\beta + \alpha_i)$ and consider the class of FE-PML panel estimators with FOC's given by*

$$\widehat{\beta}\text{:}\ \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\Big(y_{it} - \widehat{\lambda}_{it}\Big)g(\widehat{\lambda}_{it}) = 0, \qquad\qquad \widehat{\alpha}_i\text{:}\ \sum_{t=1}^{T}\Big(y_{it} - \widehat{\lambda}_{it}\Big)g(\widehat{\lambda}_{it}) = 0,$$

where $i = 1, \ldots, N$, $t = 1, ..., T$, and $g(\widehat{\lambda}_{it})$ is an arbitrary positive function of $\widehat{\lambda}_{it}$. If T is fixed, $\widehat{\beta}$ is only consistent under general assumptions about $\text{Var}(y|x, \alpha)$ if $g(\lambda)$ is constant over the range of λ's that are realized in the data-generating process.

Put simply, if Lemma 2 holds, then no other FE-PML estimator of the form described in Proposition 2 aside from FE-PPML can be consistent under general assumptions about the conditional variance $\text{Var}(y|x, \alpha, \gamma, \eta)$. We have already shown that the three-way FE-PPML estimator is generally consistent regardless of the conditional variance. Thus, if we can prove Lemma 2, Proposition 2 follows directly.

Proof of Lemma 2. Our strategy here will be to adopt a specific parameterization for the conditional variance $\text{Var}(y|x, \alpha)$ and then examine the conditions under which $\widehat{\beta}$ is sensitive to small changes in the conditional variance. If $\widehat{\beta}$ depends on $\text{Var}(y|x, \alpha)$ even for large N, then it is not possible for $\widehat{\beta}$ to be consistent under general assumptions about $\text{Var}(y|x, \alpha)$.

To proceed, let the true data generating process be given by

$$y_{it} = \lambda_{it}\omega_{it},$$

where λ_{it} is the true conditional mean and

$$\omega_{it} := \exp\left[-\frac{1}{2}\ln\left(1+\lambda_{it}^{\rho}\right)+\sqrt{\ln\left(1+\lambda_{it}^{\rho}\right)}z_{it}\right] \tag{45}$$

with z_{it} a randomly-generated variable distributed $\mathcal{N}(0,1)$. ω_{it} is therefore a heteroskedastic multiplicative disturbance that follows a log-normal distribution with $\mathbb{E}[\omega_{it}]=1$ and $\text{Var}(\omega_{it})=\lambda_{it}^{\rho}$. The conditional mean of y_{it} is in turn given by $\mathbb{E}[y_{it}|x,\alpha]=\lambda_{it}$ and the conditional variance is given by $\text{Var}(y_{it}|x,\alpha)=\text{Var}(y_{it}|\lambda_{it})=\lambda_{it}^{2}\text{Var}(\omega_{it})=\lambda_{it}^{\rho+2}$. Our focus is the exponent ρ, which governs the nature of the heteroskedasticity and can be any real number. With this in mind, it is useful to document the following results,

$$\mathbb{E}\left[\frac{\partial\omega_{it}}{\partial\rho}\right]=\frac{\partial\mathbb{E}\left[\omega_{it}\right]}{\partial\rho}=0 \tag{46}$$

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\left(\omega_{it}^{2}\right)}{\partial\rho}\right]&=\mathbb{E}\left[2\omega_{it}\frac{\partial\omega_{it}}{\partial\rho}\right]=\frac{\partial\mathbb{E}\left(\omega_{it}^{2}\right)}{\partial\rho}\\&=\frac{\partial V\left[\omega_{it}\right]}{\partial\rho}=\lambda_{it}^{\rho}\ln\lambda_{it}\neq 0.\end{aligned} \tag{47}$$

Put another way, the expected value of the change in ω_{it} with respect to ρ must always be zero because $\mathbb{E}[\omega_{it}]=1$ regardless of ρ. Similarly, the expected change in the second moment of ω_{it} must be $\lambda_{it}^{\rho}\ln\lambda_{it}$ because this gives the change in the variance of ω_{it}.[5]

To facilitate the rest of the proof, we invoke the following conceit: the random disturbance term z_{it}, once drawn from $\mathcal{N}(0,1)$, is *known* and *fixed*, such that each ω_{it} may be treated as a known transformation of the underlying value for z_{it} given by (45). Among other things, this means we can always treat the partial derivatives $\frac{\partial\omega_{it}}{\partial\rho}$ and $\frac{\partial y_{it}}{\partial\rho}=\lambda_{it}\frac{\partial\omega_{it}}{\partial\rho}$ as well-defined; similarly, we can treat the estimated parameters $\widehat{\beta}$ and $\widehat{\alpha}_i$ as deterministic functions of the variance parameter ρ with well-defined total derivatives $\frac{d\widehat{\beta}}{d\rho}$ and $\frac{d\widehat{\alpha}_i}{d\rho}$. That is, for a given draw of z_{it}'s, we can perturb how the corresponding ω_{it}'s are generated and consider comparative statics for how estimates are affected. If $\widehat{\beta}$ is consistent regardless of the variance assumption used to generate ω_{it}, then small changes in ρ should have no effect on $\widehat{\beta}$ asymptotically. Thus, our goal in the following is to determine if there are any estimators in this class other than FE-PPML under which $\lim_{N\to\infty}\frac{d\widehat{\beta}}{d\rho}=0$ in this experiment.

The next step is to totally differentiate the FOC's for $\widehat{\beta}$ and $\widehat{\alpha}_i$ with respect to a change in ρ. Let $\mathcal{L}$ denote the pseudo-likelihood function to be maximized.[6] For notational

[5] Note here that $\frac{\partial(\omega_{it}^{2})}{\partial\rho}=2\omega_{it}\frac{\partial\omega_{it}}{\partial\rho}$.

[6] The implied pseudo-likelihood function is given here by $\mathcal{L} := \sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}\int\frac{g(\lambda_{it})}{\lambda_{it}}d\lambda_{it}-\sum_{i=1}^{N}\sum_{t=1}^{T}\int g(\lambda_{it})d\lambda_{it}$.

convenience, we can express the scores for $\widehat{\beta}$ and $\widehat{\alpha}_i$ as $\mathcal{L}_\beta$ and $\mathcal{L}_{\alpha_i}$, such that their FOCs can respectively be written as $\mathcal{L}_\beta = 0$ and $\mathcal{L}_{\alpha_i} = 0$. Differentiating the FOC for $\widehat{\beta}$, we obtain

$$\frac{d\widehat{\beta}}{d\rho} = -\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho} - \mathcal{L}_{\beta\beta}^{-1}\sum_i \mathcal{L}_{\beta\alpha_i}\frac{d\widehat{\alpha}_i}{d\rho}, \tag{48}$$

where $\mathcal{L}_{\beta\beta}$ is the matrix obtained from partially differentiating the score for $\widehat{\beta}$ with respect to $\widehat{\beta}$, $\mathcal{L}_{\beta\rho}$ (a vector) is the partial derivative of $\mathcal{L}_\beta$ with respect to ρ, and $\mathcal{L}_{\beta\alpha_i}$ (also a vector) is its partial derivative with respect to $\widehat{\alpha}_i$. Applying a similar set of operations to the FOC for $\widehat{\alpha}_i$ then gives

$$\frac{d\widehat{\alpha}_i}{d\rho} = -\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\alpha_i\rho} - \mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}'_{\beta\alpha_i}\frac{d\widehat{\beta}}{d\rho}, \tag{49}$$

where $\mathcal{L}_{\alpha_i\alpha_i}$ and $\mathcal{L}_{\alpha_i\rho}$ are scalars that respectively contain the partial derivatives of $\mathcal{L}_{\alpha_i}$ with respect to $\widehat{\alpha}_i$ and ρ. Plugging (49) into (48), we have

$$\begin{aligned}
\frac{d\widehat{\beta}}{d\rho} &= -\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho} + \mathcal{L}_{\beta\beta}^{-1}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho} + \mathcal{L}_{\beta\beta}^{-1}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\frac{d\widehat{\beta}}{d\rho} \\
&= -\left(\mathbf{I} - \mathcal{L}_{\beta\beta}^{-1}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\right)^{-1}\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho} \qquad (50) \\
&\quad + \left(\mathbf{I} - \mathcal{L}_{\beta\beta}^{-1}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\right)^{-1}\mathcal{L}_{\beta\beta}^{-1}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}, \qquad (51)
\end{aligned}$$

where $\mathbf{I}$ is an identity matrix whose dimensions equal the size of β.

Let $\mathbf{P}$ henceforth denote the combined matrix object $\mathbf{I} - \mathcal{L}_{\beta\beta}^{-1}\sum_i \mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}$. It is straightforward to show that that first term in (51), $-\mathbf{P}^{-1}\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho}$, converges in probability to a zero vector when $N \to \infty$. To see this, note first that $\mathbf{P}$ and $\mathcal{L}_{\beta\beta}$ must be non-singular and finite for $\widehat{\beta}$ to be at a maximum point of $\mathcal{L}$ and for $\frac{d\widehat{\beta}}{d\rho}$ to exist. Furthermore, $\lim_{N\to\infty} NT\mathcal{L}_{\beta\beta}^{-1} = -\mathbb{E}[x_{it}\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})x'_{it}]^{-1}$ must also be non-singular and finite. Slutsky's theorem then implies $\lim_{N\to\infty} -\mathbf{P}^{-1}\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho} \to_p 0$ if $\lim_{N\to\infty} N^{-1}T^{-1}\mathcal{L}_{\beta\rho} \to_p 0$. Examining the vector $\mathcal{L}_{\beta\rho}$ more closely, we have

$$\mathcal{L}_{\beta\rho} = \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\frac{\partial y_{it}}{\partial\rho}g(\widehat{\lambda}_{it}) = \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\lambda_{it}\frac{\partial\omega_{it}}{\partial\rho}g(\widehat{\lambda}_{it}).$$

$\lim_{N\to\infty} N^{-1}T^{-1}\mathcal{L}_{\beta\rho} \to_p 0$ then follows via standard arguments because $\mathbb{E}\left[\frac{\partial\omega_{it}}{\partial\rho}\right] = 0$ (by (46)). We may therefore focus our attention on the second term on the RHS in (51),

$\mathbf{P}^{-1}\mathcal{L}_{\beta\beta}^{-1}\sum_i \mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}$. Noting that $\mathcal{L}_{\alpha_i\alpha_i}^{-1}$must be < 0, in this case we consider the conditions under which $\lim_{N\to\infty} N^{-1}T^{-1}\sum_i \mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}$ similarly converges in probability to zero. The summation in this latter term may be expressed as

$$\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho} = \sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\left[\sum_{t=1}^{T}x_{it}\Big(y_{it}-\widehat{\lambda}_{it}\Big)g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it} - \sum_{t=1}^{T}x_{it}\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})\right]\sum_{t=1}^{T}\frac{\partial y_{it}}{\partial\rho}g(\widehat{\lambda}_{it}).$$

Re-arranging this expression, we have that

$$\begin{aligned}\sum_{i=1}^{N}\mathcal{L}_{\alpha_i\alpha_i}^{-1}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho} = &\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}y_{it}g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\frac{\partial y_{is}}{\partial\rho}\\ &-\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}\Big(\widehat{\lambda}_{it}g'(\widehat{\lambda}_{it})+g(\widehat{\lambda}_{it})\Big)\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\frac{\partial y_{is}}{\partial\rho}. \qquad (52)\end{aligned}$$

Focusing first on the second of the two summation terms in (52), we again apply $y_{it} = \lambda_{it}\omega_{it}$, $\frac{\partial y_{is}}{\partial\rho} = \lambda_{it}\frac{\partial\omega_{is}}{\partial\rho}$, and $\mathbb{E}\left[\frac{\partial\omega_{it}}{\partial\rho}\right] = 0$. We have that

$$\lim_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}\Big(\widehat{\lambda}_{it}g'(\widehat{\lambda}_{it})+g(\widehat{\lambda}_{it})\Big)\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\lambda_{is}\frac{\partial\omega_{is}}{\partial\rho}\to_p 0.$$

This follows for the same reason $\lim_{N\to\infty}N^{-1}T^{-1}\mathcal{L}_{\beta\rho}\to_p 0$ above. The first summation term in (52) obviously $\to_p 0$ as well if the estimator is FE-PPML, in which case $g'(\widehat{\lambda}_{it}) = 0$. To complete the proof, we just need to show that this term does not reduce to 0 if $g'(\widehat{\lambda}_{it})\neq 0$. A final step gives us

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}y_{it}g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\frac{\partial y_{is}}{\partial\rho} &= \lim_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})y_{it}\frac{\partial y_{it}}{\partial\rho}\\ &= \lim_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{L}_{\alpha_i\alpha_i}^{-1}x_{it}g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})\lambda_{it}^2\omega_{it}\frac{\partial\omega_{it}}{\partial\rho}\\ &\neq 0.\end{aligned}$$

To elaborate, the terms where $s\neq t$ vanish as $N\to\infty$ because disturbances are assumed to be independently distributed ($\mathbb{E}[\omega_{it}\frac{\partial\omega_{is}}{\partial\rho}] = 0$ if $s\neq t$.)[7] The remaining details follow from (47).[8] We have now shown $\lim_{N\to\infty}\frac{d\widehat{\beta}}{d\rho} = 0$ if and only if $g'(\widehat{\lambda}_{it}) = 0$. In other words, the estimator must be FE-PPML, which assumes $g(\widehat{\lambda}_{it})$ is a constant. For other FE-PML estimators, even if $\widehat{\beta}$ is consistent for a particular ρ, it cannot be consistent for

[7]Note that under FE-PPML, where $g'(\widehat{\lambda}_{it}) = 0$, the estimator is consistent even if disturbances are correlated. This is yet another reason why FE-PPML is an especially robust estimator.

[8]Notice that if $T\to\infty$ also, we have that $\lim_{T\to\infty}T\mathcal{L}_{\alpha_i\alpha_i}^{-1} = -\mathbb{E}\left[\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})\right]^{-1}$ must be finite. We

all ρ because $\widehat{\beta}$ does not converge to the same value for $N \to \infty$ when we vary ρ. As we discuss below, this is what happens for FE-Gamma PML (where $g(\widehat{\lambda}_{it}) = \widehat{\lambda}_{it}^{-1}$) and some other similar estimators. ■

To be clear, the robustness of the FE-PPML estimator to misspecification is a known result established by Wooldridge (1999). However, to our knowledge, it has not previously been shown that FE-PPML is the only estimator in the class we consider that has this property.[9] At the same time, it is worth clarifying that FE-PPML is not the only estimator that is capable of producing consistent estimates of three-way gravity models. Rather, it is the only estimator in the class we consider that only requires correct specification of the conditional mean and for the covariates to be conditionally exogenous in order to be consistent. The following discussion describes some known cases in which other estimators will be consistent.

A.5 Results for Other Three-way Estimators

Depending on the distribution of the data, there may be some other consistent estimator available aside from FE-PPML. In particular, if $g(\widehat{\lambda}_{ijt})$ is of the form $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}^{q}$, with q an arbitrary real number, the FOC for $\widehat{\eta}_{ij}$ has a solution of the form $\widehat{\eta}_{ij} = [\sum_{t=1}^{T} \widehat{\mu}_{ijt}^{q+1}]^{-1} \sum_{t=1}^{T} y_{ijt}\widehat{\mu}_{ijt}^{q}$. It is therefore possible to "profile out" $\widehat{\eta}_{ij}$ from the FOC for $\widehat{\beta}$, just as in the FE-PPML case. As such, it is possible for the estimator to be consistently estimated, but only if the conditional variance is correctly specified (more precisely, we must have $\text{Var}(y|x, \alpha, \gamma, \eta) \propto \widehat{\lambda}_{it}^{1-q}$, the equivalent of $\rho = -1 - q$.) In this case, the estimator is not only consistent, but should be more efficient as well.

An interesting example to consider in the gravity context is the Gamma PML (GPML) estimator, which imposes $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}^{-1}$. Generally speaking, GPML is considered the

would therefore have

$$\lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[T\mathcal{L}_{\alpha_i\alpha_i}^{-1}\right] x_{it} g'(\widehat{\lambda}_{it}) \widehat{\lambda}_{it} g(\widehat{\lambda}_{it}) \lambda_{it}^{2} \left[T^{-1}\omega_{it} \frac{\partial \omega_{it}}{\partial \rho}\right] = 0,$$

ensuring that $\widehat{\beta}$ does not depend on ρ for the large N, large T case. This follows because $\lim_{T\to\infty} T^{-1} V\left[\omega_{it}\right] = 0 \implies \lim_{T\to\infty} T^{-1}\mathbb{E}\left[\omega_{it}\frac{\partial \omega_{it}}{\partial \rho}\right] = 0$.

[9] Alternatively, it is possible to extend the above result to an even more general class of estimators by considering estimators that depend on $g(\widehat{\alpha}_i)$ rather than $g(\widehat{\lambda}_{it})$. The same type of proof may be used to show that $\widehat{\beta}$ depends on the variance assumption if $g'(\widehat{\alpha}_i) \neq 0$. Furthermore, the estimator can be shown to be consistent if $g'(\widehat{\alpha}_i) = 0$.

primary alternative to PPML and OLS as an estimator for use with gravity equations (see Head and Mayer, 2014; Bosquet and Boulhol, 2015.) However, to our knowledge, no references to date on gravity estimation make it clear that, unlike in a two-way setting, the three-way FE-GPML estimator is only consistent when the conditional variance is correctly specified.[10] Thus, it is possible that researchers could mistakenly infer that the appeal of FE-GPML as an alternative to FE-PPML in the two-way gravity setting carries over to the three-way setting.[11] This is especially a concern now that recent computational advances have made estimation of FE-GLM models significantly more feasible.

To illuminate the unique IPP-robustness properties of FE-PPML in the three-way context, Fig. 3 shows a comparison of simulation results for FE-PPML versus log-OLS and Gamma PML.[12] The displayed kernel densities are computed using 500 replications of a three-way panel structure with $N = 50$ and $T = 5$.[13] The i and j dimensions of the panel both have size $N = 50$ and the size of the time dimension is $T = 5$. The fixed effects are generated according to the same procedures described in the text and we again model four different scenarios for the distribution of the error term (Gaussian, Poisson, Log-homoskedastic, and Quadratic).

As we would expect based on Proposition 2 , FE-PPML is relatively unbiased across all four different assumptions considered for the distribution of the error term. The general inconsistency of the three-way OLS estimator—which is only unbiased for DGP III where the error term is log-homoskedastic—is also as expected. However, the reasons behind the bias in the OLS estimate are well-documented (see Santos Silva and Tenreyro, 2006) and

[10] As discussed in Greene (2004), the fixed effects Gamma model is generally known not to suffer from an incidental parameter problem, similar to FE-Poisson. However, the result stated in Greene (2004) is for the Gamma MLE estimator, which restricts the conditional variance to be equal to the square of the conditional mean. The FE-Gamma PML estimator is consistent under the slightly more general assumption that the conditional variance is proportional to the square of the conditional mean.

[11] For example, Head and Mayer (2014), arguably the leading reference to date on gravity estimation, suggest comparing PPML estimates with GPML estimates to determine if the RHS of the model is potentially misspecified. Such a comparison is not straightforward in a three-way setting because the GPML estimator is likely to be inconsistent. Their other suggestion to compare GPML and OLS estimates still seems sensible, however. As we show below, both estimators give similar results when the Gamma variance assumption is satisfied and give different results otherwise.

[12] We were able to compute three-way FE-Gamma PML estimates using a modified version of the HDFE-IRLS algorithm used in Correia, Guimarães, and Zylkin (2020). To our knowledge, these are the first results presented anywhere documenting the inconsistency of the three-way Gamma PML estimator.

[13] Simulations with larger N are more narrowly distributed, but otherwise are very similar.

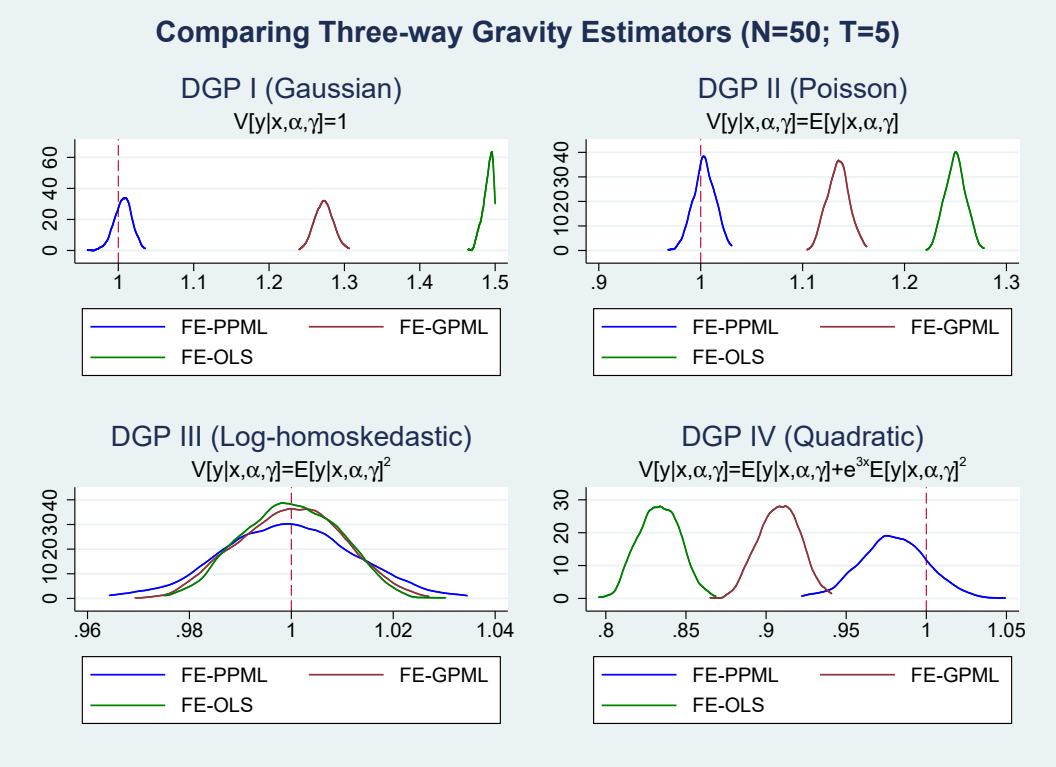

Figure 3: Kernel density plots of three-way gravity model estimates using different FE estimators, based on 500 replications. The model being estimated is $y_{ijt} = \exp[\alpha_{it} + \gamma_{jt} + \eta_{ij} + x_{ijt}\beta]\omega_{ijt}$, where the distribution of ω_{ijt} depends on the DGP and the true value of β is 1 (indicated by the vertical dotted lines). The size of the i and j dimensions is given by $N = 50$ and the t dimension has size $T = 5$. See text for further details.

do not have to do with the incidental parameters included in the model. The three-way FE-GPML estimator is also consistent under DGP III because it assumes the error term has a variance equal to the square of the conditional mean. Both OLS and GPML are also more efficient than PPML in this case. However, as the other three panels show, when this variance assumption is relaxed, three-way FE-GPML clearly suffers from an IPP, exhibiting an average bias equal to roughly half that of OLS in all three cases.

We have also performed some simulations with three-way FE-Gaussian PML, which imposes $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}$. We do not show results for this other estimator because the HDFE-IRLS algorithm we used to produce the FE-PPML and FE-Gamma PML estimates frequently did not converge for the FE-Gaussian PML estimator. However, the results we did obtain were in line with our results for FE-GPML and with our discussion of Proposition 2 above: the FE-Gaussian PML estimates were consistent when the DGP for ω_{ijt} was itself Gaussian (as in DGP I), but were inconsistent otherwise.

A.6 Allowing for Conditional Dependence across Pairs

The bias expansion in Proposition 3 allows for errors to be clustered within each pair (i, j), but assumes conditional independence of y_{ij} and $y_{i'j'}$ for all $(i, j) \neq (i', j')$. This assumption is consistent with the standard practice in the literature of assuming that errors are clustered within pairs when computing standard errors (see Yotov, Piermartini, Monteiro, and Larch, 2016.) However, it is important to clarify that the results in Proposition 3 may change when other assumptions are used. For example, if we want to allow y_{ij} and y_{ji} (i.e., both directions of trade) to be correlated, then the bias results would not actually change, but we would need to modify the definition of Ω_N to allow for the additional clustering; namely, we would need

$$\begin{aligned}\Omega_N &= \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \operatorname{Var}\left(\widetilde{x}'_{ij} S_{ij} + \widetilde{x}'_{ji} S_{ji} \,\middle|\, x\right) \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Big\{ \widetilde{x}'_{ij} \left[\operatorname{Var}\left(S_{ij} \,\middle|\, x_{ij}\right)\right] \widetilde{x}_{ij} + \widetilde{x}'_{ji} \left[\operatorname{Var}\left(S_{ji} \,\middle|\, x_{ji}\right)\right] \widetilde{x}_{ji} \\ &\qquad + \widetilde{x}'_{ij} \left[\operatorname{Cov}\left(S_{ij}, S_{ji} \,\middle|\, x_{ij}\right)\right] \widetilde{x}_{ji} + \widetilde{x}'_{ji} \left[\operatorname{Cov}\left(S_{ji}, S_{ij} \,\middle|\, x_{ji}\right)\right] \widetilde{x}_{ij} \Big\}.\end{aligned} \tag{53}$$

Notice, however, this is just one possibility. Similar adjustments could be made to allow for clustering by exporter or importer, for example, or even for multi-way clustering à la Cameron, Gelbach, and Miller (2011). In these cases, the bias would also need to be modified; specifically, one would have to modify the portions of D_N^k that B_N^k that depend on the variance of S_{ij} to allow for correlations across i and/or j.

A.7 Showing Bias in the Cluster-robust Sandwich Estimator

For convenience, let $\mathbf{x}_{ij} := (x_{ij},\, d_{ij})$ be the matrix of covariates associated with pair ij, inclusive of the it- and jt-specific dummy variables needed to estimate α_i and γ_j. Similarly, let $b := (\beta',\, \phi')'$ be the vector of coefficients to be estimated and let $\widehat{b}$ be the vector of coefficient estimates. Note that we can write a first-order approximation for $\widehat{S}_{ij}$ as

$$\widehat{S}_{ij} \approx S_{ij} - \bar{H}_{ij} \mathbf{x}_{ij} (\widehat{b} - b),$$

which is consistent with the approximation provided in (13). We can then replace $\widehat{b} - b$ with the standard first-order expansion $\widehat{b} - b \approx -\bar{\mathcal{L}}_{bb}^{-1} \mathcal{L}_b^0$, where $\mathcal{L} = \sum_{i,j} \ell_{ij}$ is the profile

likelihood. This expansion in turn can be written out as

$$\widehat{b} - b \approx -\bar{\mathcal{L}}_{bb}^{-1}\left[\sum_{m,n}\mathbf{x}'_{mn}S_{mn}\right].$$

Now we turn our attention to the outer product $\widehat{S}_{ij}\widehat{S}'_{ij}$:

$$\begin{aligned}\widehat{S}_{ij}\widehat{S}'_{ij} &\approx S_{ij}S'_{ij} + \bar{H}_{ij}\mathbf{x}_{ij}(\widehat{b}-b)^2\mathbf{x}'_{ij}\bar{H}_{ij} - 2\bar{H}_{ij}\left[\mathbf{x}_{ij}(\widehat{b}-b)\right]S'_{ij} \\ &= S_{ij}S'_{ij} + \bar{H}_{ij}\mathbf{x}_{ij}(\widehat{b}-b)^2\mathbf{x}'_{ij}\bar{H}_{ij} + 2\bar{H}_{ij}\mathbf{x}_{ij}\bar{\mathcal{L}}_{bb}^{-1}\left[\sum_{m,n}\mathbf{x}'_{mn}S_{mn}\right]S'_{ij}\end{aligned}$$

Because we assume we are in the special case where FE-PPML is correctly specified, we have that $\mathbb{E}[(\widehat{b}-b)^2] = -\kappa\bar{\mathcal{L}}_{bb}^{-1}$, where $\bar{\mathcal{L}}_{bb} := \mathbb{E}[\mathcal{L}_{bb}]$. We also have that $\mathbb{E}[S_{ij}S'_{ij}] = \kappa\bar{H}_{ij}$. Therefore, after applying expectations where appropriate, we have that

$$\mathbb{E}[\widehat{S}_{ij}\widehat{S}'_{ij}] \approx S_{ij}S'_{ij} + \kappa\bar{H}_{ij}\mathbf{x}_{ij}\bar{\mathcal{L}}_{bb}^{-1}\mathbf{x}'_{ij}\bar{H}_{ij},$$

which can be seen as extending Kauermann and Carroll (2001)'s results to the case of a panel data pseudo-likelihood model with within-panel clustering. We are not done, however, as we have not yet isolated the influence of the incidental parameters. To complete the derivation of the bias, we must more carefully consider the full inverse Hessian term $\bar{\mathcal{L}}_{bb}^{-1}$. Using standard matrix algebra, this inverse can be written as:

$$\bar{\mathcal{L}}_{bb}^{-1} = \begin{pmatrix} \left(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\right)^{-1} & -\left(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\right)^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1} \\ -\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\left(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\right)^{-1} & \bar{\mathcal{L}}_{\phi\phi}^{-1} + \bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\left(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}\right)^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1} \end{pmatrix},$$

where we have used $\bar{\mathcal{L}}_{\phi\phi}$ in place of $\bar{\mathcal{H}}$ in order to add clarity. Making use of some already-established definitions, we have that the top-left term $(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}^{*\prime}_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta})^{-1} = -[N(N-1)]^{-1}W_N^{-1}$ and, similarly, that $\bar{\mathcal{L}}_{\phi\phi}^{-1} = -[N(N-1)]^{-1}W_N^{(\phi)-1}$. If we again consider $\mathbb{E}[\widehat{S}_{ij}\widehat{S}'_{ij}]$, we can now write

$$\begin{aligned}\mathbb{E}[\widehat{S}_{ij}\widehat{S}'_{ij} - S_{ij}S'_{ij}] \approx &-\frac{\kappa}{N(N-1)}\bar{H}_{ij}(x_{ij}\, d_{ij}) \times \\ &\begin{pmatrix} W_N^{-1} & -W_N^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1} \\ -\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}W_N^{-1} & W_N^{(\phi)-1} + \bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}^{*}_{\phi\beta}W_N^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}^{*-1}_{\phi\phi} \end{pmatrix}(x_{ij}\, d_{ij})'\bar{H}_{ij} \\ &= -\frac{\kappa}{N(N-1)}\bar{H}_{ij}\left\{x_{ij}W_N^{-1}x'_{ij} - x_{ij}W_N^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}d'_{ij} - d_{ij}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}^{*}_{\phi\beta}W_N^{-1}x'_{ij}\right. \\ &\left. + d_{ij}\bar{\mathcal{L}}_{\phi\phi}^{-1}\bar{\mathcal{L}}_{\phi\beta}W_N^{-1}\bar{\mathcal{L}}'_{\phi\beta}\bar{\mathcal{L}}_{\phi\phi}^{-1}d'_{ij} + d_{ij}W_N^{(\phi)-1}d'_{ij}\right\}\bar{H}_{ij},\end{aligned}$$

which simplifies to the expression shown in (13).

Results for the two-way model. Though we have focused on the downward bias of the sandwich estimator for the three-way gravity model, it is also known to be biased for the standard two-way gravity model without pair fixed effects (Egger and Staub, 2015; Jochmans, 2017; Pfaffermayr, 2019). As it turns out, the analytics for the two-way and three-way models are very similar here, and we can easily adapt our results to the simpler two-way setting. The main change we would need to make is to replace H_{ij} everywhere it appears with Λ_{ij}, including in the definitions of $\widetilde{x}_{ij}$, W_N, and $W_N^{(\phi)}$. The rest of the derivations then follow in the same manner as for the three-way model. The resulting correction has been included in our `ppml_fe_bias` Stata package for users working with two-way gravity models. A version of this correction has been studied alongside other methods in a recent paper by Pfaffermayr (2021).

A.8 More Discussion of IPPs in FE-PPML Models

In this part of the Appendix, we wish to give a more expansive discussion of when IPPs may arise in case of an FE-PPML estimator. We have already reviewed the two-way and three-way gravity models in the main text; thus, here, we will focus first on the classic "one-way" FE panel setting where no IPP occurs. Doing so will allow us to draw a contrast with other, more complex models where IPPs could be a problem. As part of this discussion, we give some examples of panel models where FE-PPML is actually inconsistent, unlike the models covered in the main text.

The Classic (One-way) Setting

Consider a static panel data model with individuals $i = 1, \ldots, N$, time periods $t = 1, \ldots, T$, outcomes y_{it}, and strictly exogenous regressors x_{it} satisfying

$$\mathbb{E}(y_{it}|x_{it}, \alpha_i) = \lambda_{it} := \exp(x_{it}'\beta + \alpha_i). \tag{54}$$

The FE-PPML estimator maximizes $\sum_{i,t} (y_{it} \log \lambda_{it} + \lambda_{it})$ over β and α. The corresponding FOC's may be written as

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it} \left(y_{it} - \widehat{\lambda}_{it}\right) = 0, \qquad \forall i: \ \sum_{t=1}^{T} \left(y_{it} - \widehat{\lambda}_{it}\right) = 0, \tag{55}$$

where $\widehat{\lambda}_{it} := \exp(x_{it}'\widehat{\beta} + \widehat{\alpha}_i)$. Solving for $\widehat{\alpha}_i$ and plugging the expression back into the FOC for $\widehat{\beta}$ we find

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left[y_{it} - \frac{\exp(x_{it}'\widehat{\beta})}{\sum_{\tau=1}^{T}\exp(x_{i\tau}'\widehat{\beta})}\sum_{\tau=1}^{T} y_{i\tau}\right] = 0, \tag{56}$$

which, as long as (54) holds, are valid (sample) moments to estimate β. Thus, under standard regularity conditions, we have that $\sqrt{N}(\widehat{\beta} - \beta^0) \to_d \mathcal{N}(0, V)$ as $N \to \infty$, where V is the asymptotic variance. The FE-PPML estimator therefore does not suffer from an IPP: even though $\widehat{\alpha}_i$ is an inconsistent estimate of α_i, the FE-PPML score for β has zero mean when evaluated at the true parameter β^0, and $\widehat{\beta}$ therefore converges in probability to β^0 without any asymptotic bias. This is a well known result that can also be obtained in the Poisson-MLE case by conditioning on $\sum_t y_{it}$; see Cameron and Trivedi (2015). For our purposes, it gives us a benchmark against which other, more complex models may be compared.

Examples where FE-PPML is Inconsistent

In the above "classic" setting, every observation is affected by exactly one fixed effect. In current applied work, it is common to specify models with what we will call "overlapping" fixed effects, where each observation may be affected by more than one fixed effects. Some standard examples include the gravity model from international trade (as is our focus in the main text) as well as other settings where researchers may wish to control for multiple sources of heterogeneity (e.g., firm and employee, teacher and student). Thus, it is important to clarify that the presence of overlapping fixed effects can easily lead to an IPP, even when the underlying estimator is Poisson or PPML. We give the following simple example:

Example 1. *Consider a model with three time periods $T = 3$ and two fixed effects α_i and γ_i for each individual:*

$$\begin{aligned}
t = 1: &\qquad \mathbb{E}(y_{i1}|x_{i1}, \alpha_i, \gamma_i) = \lambda_{i1} := \exp(x_{i1}'\beta + \alpha_i),\\
t = 2: &\qquad \mathbb{E}(y_{i2}|x_{i2}, \alpha_i, \gamma_i) = \lambda_{i2} := \exp(x_{i2}'\beta + \alpha_i + \gamma_i),\\
t = 3: &\qquad \mathbb{E}(y_{i3}|x_{i3}, \alpha_i, \gamma_i) = \lambda_{i3} := \exp(x_{i3}'\beta + \gamma_i).
\end{aligned}$$

The FE-PPML estimator maximizes $\sum_{i=1}^{N}\sum_{t=1}^{3}(y_{it}\log\lambda_{it} + \lambda_{it})$ over β, α and γ. $T = 3$ is fixed as $N \to \infty$.

Example 2. *In addition to $i = 1, \ldots, N$ and $t = 1, \ldots, T$ we re-introduce another panel dimension $j = 1, \ldots, J$ and consider*

$$\mathbb{E}(y_{ijt}|x_{ijt}, \alpha_{it}, \gamma_{ij}) = \lambda_{ijt} := \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{ij}),$$

where α_{it} is now indexed by both i and t and our second fixed effect is similarly indexed by i and j. The FE-PPML estimator in this case maximizes $\sum_{i,j,t}(y_{ijt} \log \lambda_{ijt} + \lambda_{ijt})$ over β, α and γ. We consider $N \to \infty$ with both J and T fixed, e.g. $J = T = 2$.

In these examples, because the fixed effects are overlapping, we have that $\widehat{\alpha}$ enters into the FOC for $\widehat{\gamma}$, and vice versa. Therefore, when for a given value $\widehat{\beta}$ we want to solve the FOC for $\widehat{\alpha}$ and $\widehat{\gamma}$ we have to solve a system of equations, and the solutions become much more complicated functions of the outcome variable than in the one-way model. While having this type of co-dependence between the FOCs for the various fixed effects need not necessarily lead to an IPP, as we discuss next, it does create one in models where more than one fixed effect dimension grows at the same rate as the panel size, as is the case with α and γ in both of these examples.

In gravity settings, by contrast, the crucial distinction is that the dimensions of each fixed effect grow only with the square root of the sample size as the number of countries increases. As mentioned in the text, this ensures that the IPPs associated with each fixed effect "decouple" from one another, in the sense described by Fernández-Val and Weidner (2016). However, in the inconsistency examples just given, the estimation noise in the estimated α parameters will always depend on the estimation noise in the estimated γ parameters, and vice versa, even as $N \to \infty$. Thus, decoupling cannot occur. For illustration, we have used the model from Example 1 as our example of inconsistency in our earlier Figure 1. We have also performed simulations for the model in Example 2 and found similar results.

A Suggested Heuristic for IPPs when the Estimator is FE-PPML

As has been made clear from our discussion and results, the usual generic bias heuristic shown in our equation (6) from Fernández-Val and Weidner (2018) is generally not appropriate for FE-PPML. Indeed, because FE-PPML can be asymptotically unbiased in special cases, it may not be productive to try to boil down how their biases are likely to behave to a single formula. Instead, we propose the following approach:

1. If there are no fixed effect dimensions that grow proportionately with the sample, we expect FE-PPML to be unbiased asymptotically.

2. Otherwise, the likely order of the bias can be derived as follows:

   (a) Construct the equivalent multinomial model by profiling out the largest fixed effect dimension.

   (b) Infer what the order of the asymptotic bias would be for the equivalent multinomial model by calculating p/n (i.e., as in (6)).

For example, in the three-way gravity model, the number of observations is on the order of N^2T and the number of parameters is on the order of N^2 pair fixed effects plus $2NT$ exporter-time and importer-time fixed effects. However, after profiling out the pair fixed effects, we only have the $2NT$ exporter-time and importer-time fixed effects. Thus, we take p/n to be proportional to $1/N$ as $N \to \infty$, implying an asymptotic bias of order $1/N$.

For further illustration, consider Examples 1 and 2 above. In these cases, even after profiling out α, one still finds that p is proportional to n as $n \to \infty$, implying inconsistency. As a contrast, consider the two-way gravity model. As we have just discussed, all of the fixed effects grow only with the square root of n in that case, implying it is asymptotically unbiased.

"Four-way" gravity models. As a more complicated example, consider the following "four-way" gravity model:

$$y_{ijlt} = \exp\left[\alpha_{ilt} + \alpha_{jlt} + \eta_{ijl} + \zeta_{ijt} + x_{ijlt}\beta\right]\omega_{ijlt}. \tag{57}$$

This type of model may be used for trade data that is observed separately for different industries or commodities, which here are indexed by $l = 1...L$. α_{ilt}, α_{jlt}, and η_{ijl} respectively are industry-level analogs of α_{it}, α_{jt}, and η_{ij} from the three-way model. Thus, they allow multilateral resistance effects and cross-sectional heterogeneity in trade costs to vary by industry. The fourth fixed effect, ζ_{ijt}, captures general changes in trade across all industries for a given pair. x_{ijlt} is assumed to be an industry-specific policy variable of interest (e.g., tariffs). We assume that the error term ω_{ijlt} exhibits correlation over time within the same exporter-importer-industry triplet but is independent across trade partners and across industries within the same exporter-importer pair.

The four-way model does not conform to the framework from our main analysis, but we can nonetheless use the above heuristic to infer the order of the bias and propose a correction. After profiling out the order-N^2L exporter-importer-industry fixed effects, the model has on the order of $2NLT$ exporter-industry-time and importer-industry-time fixed effects and N^2T exporter-importer-time fixed effects. The number of observations is on the order of N^2LT. Following our discussion from Section 2.2, the bias is thus expected to be of the form

$$\frac{1}{N}b^{(\alpha)} + \frac{1}{N}b^{(\gamma)} + \frac{1}{L}b^{(\zeta)}. \tag{58}$$

Where $b^{(\alpha)}$, $b^{(\gamma)}$, and $b^{(\zeta)}$ are unknown constants. Two observations stand out. First, for consistency, we require both N and L to be large. For data sets where the number of industries is relatively small, the $1/L$ bias term associated with the ζ_{ijt} fixed effect is likely to induce substantial bias. We will thus consider the implications for asymptotic bias as N and L grow large at the same rate. Second, the order of the standard error as N and L both $\to \infty$ while T is fixed is $1/(N\sqrt{L})$. The ratio of the asymptotic bias to the standard error as N and L both $\to \infty$ is expected to be of the form

$$\frac{\sqrt{L}b^{(\alpha)} + \sqrt{L}b^{(\gamma)} + \frac{N}{\sqrt{L}}b^{(\zeta)}}{c},$$

where $c > 0$ is a positive constant. This ratio diverges to infinity as $N, L \to \infty$, implying that the standard error shrinks to zero faster than the bias does. This is a more severe form of the asymptotic bias problem than the one we found for the three-way model, where the bias and standard error both decreased at the same rate asymptotically. It is therefore advised that a jackknife correction should be used to reduce the bias. This can be done by holding out industries to inflate the $1/L$ bias while simultaneuously holding out countries to inflate the $1/N$ bias. For example, a jackknife sample with half the number of countries and half the number of industries will have an asymptotic bias of order $2/N + 2/L$.

A closely related model that we can also discuss is the case when the fourth fixed effect, ζ_{ijt}, is not included in the model shown in (57). An example of when this might be desirable is when the policy variable of interest does not vary across industries (e.g., a trade agreement dummy). In this case, one can infer from the formula in (58) that we would expect an asymptotic bias of order $1/N$, like what we found in the case of the three-way model. Furthermore, if only the number of countries is allowed to become large, while both L and T are held fixed, the standard error is also of order $1/N$, and the behavior

of the asymptotic bias is exactly like what we found for the three-way case. However, interestingly, if both N and $L \to \infty$, we are back in the case where the bias-to-standard error ratio heads to infinity. Thus, the severity of the problem will depend importantly on the number of industries in both of these models.

Finally, note that if T is no longer fixed, so that N, L, and T all $\to \infty$ jointly, we expect the special properties of PPML to cause the IPP bias to become more benign. We know from Fernández-Val and Weidner (2016)'s earlier results for the two-way model and from our own results for the three-way model that the decoupling of the IPPs as all dimensions of the panel become large at the same rate eliminates the asymptotic bias in these settings. Future work can investigate to what extent this holds true for four-way panel models.

References

Bosquet, C., and H. Boulhol (2015): "What is really puzzling about the "distance puzzle"," *Review of World Economics*, 151(1), 1–21.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 29(2), 238–249.

Cameron, A. C., and P. K. Trivedi (2015): "Count Panel Data," *Oxford Handbook of Panel Data Econometrics (Oxford: Oxford University Press, 2013)*.

Correia, S., P. Guimarães, and T. Zylkin (2020): "Fast Poisson estimation with high-dimensional fixed effects," *The Stata Journal*, 20(1), 95–115.

Dzemski, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101(5), 763–776.

Egger, P. H., and K. E. Staub (2015): "GLM estimation of trade gravity models with fixed effects," *Empirical Economics*, 50(1), 137–175.

Fernández-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.

Fernández-Val, I., and M. Weidner (2018): "Fixed effect estimation of large T panel data models," *Annual Review of Economics*, 10, 109–138.

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

Greene, W. (2004): "Fixed effects and bias due to the incidental parameters problem in the tobit model," *Econometric Reviews*, 23(2), 125–147.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., and W. Newey (2004): "Jackknife and analytical bias reduction for nonlinear panel models," *Econometrica*, 72(4), 1295–1319.

Head, K., and T. Mayer (2014): "Gravity equations: workhorse, toolkit, and cookbook," *Handbook of International Economics*, 4, 131–196.

Jochmans, K. (2017): "Two-way models for gravity," *Review of Economics and Statistics*, 99(3), 478–485.

Jochmans, K., and M. Weidner (2019): "Fixed-effect regressions on network data," *Econometrica*, 87(5), 1543–1560.

Kauermann, G., and R. J. Carroll (2001): "A note on the efficiency of sandwich covariance matrix estimation," *Journal of the American Statistical Association*, 96(456), 1387–1396.

Larch, M., J. Wanner, Y. V. Yotov, and T. Zylkin (2019): "Currency unions and trade: a PPML re-assessment with high-dimensional fixed effects," *Oxford Bulletin of Economics and Statistics*, 81(3), 487–510.

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

Nickell, S. (1981): "Biases in dynamic models with fixed effects," *Econometrica*, pp. 1417–1426.

Pfaffermayr, M. (2019): "Gravity models, PPML estimation and the bias of the robust standard errors," *Applied Economics Letters*, pp. 1–5.

——— (2021): "Confidence intervals for the trade cost parameters of cross-section gravity models," *Economics Letters*, 201, 109787.

Santos Silva, J. M. C., and S. Tenreyro (2006): "The log of gravity," *Review of Economics and Statistics*, 88(4), 641–658.

Wooldridge, J. M. (1999): "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics*, 90(1), 77–97.

Yotov, Y. V., R. Piermartini, J.-A. Monteiro, and M. Larch (2016): "An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model," *World Trade Organization, Geneva*.