Weidner, Martin; Zylkin, Thomas

Working Paper

Bias and consistency in three-way gravity models

cemmap working paper, No. CWP1/20

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Weidner, Martin; Zylkin, Thomas (2020) : Bias and consistency in three-way gravity models, cemmap working paper, No. CWP1/20, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2020.120

This Version is available at:
https://hdl.handle.net/10419/241876

cemmap
centre for microdata methods and practice

Bias and Consistency in Three-way Gravity Models

Martin Weidner
Thomas Zylkin

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP1/20

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

Bias and Consistency in Three-way Gravity Models*

Martin Weidner
UCL

Thomas Zylkin†
Richmond

December 20, 2019

We study the incidental parameter problem in "three-way" Poisson Pseudo-Maximum Likelihood ("PPML") gravity models recently recommended for identifying the effects of trade policies. Despite the number and variety of fixed effects this model entails, we confirm it is consistent for small T and we show it is in fact the only estimator among a wide range of PML gravity estimators that is generally consistent in this context when T is small. At the same time, asymptotic confidence intervals in fixed-T panels are not correctly centered at the true point estimates, and cluster-robust variance estimates used to construct standard errors are generally biased as well. We characterize each of these biases analytically and show both numerically and empirically that they are salient even for real-data settings with a large number of countries. We also offer practical remedies that can be used to obtain more reliable inferences of the effects of trade policies and other time-varying gravity variables.

JEL Classification Codes: C13; C50; F10

Keywords: Structural Gravity, Trade Agreements; Asymptotic Bias Correction

*Thomas Zylkin is grateful for support from NUS Strategic Research Grant WBS: R-109-000-183-646 awarded to the Global Production Networks Centre (GPN@NUS) for the project titled "Global Production Networks, Global Value Chains, and East Asian Development". Martin Weidner acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA. We also thank Valentina Corradi, Riccardo D'adamo, Ivan Fernández-Val, Koen Jochmans, Michael Pfaffermayr, Amrei Stammann, and Yoto Yotov.

†*Contact information*: Weidner: Department of Economics, University College London, London WC1H 0AX. Email: m.weidner@ucl.ac.uk. Zylkin: Robins School of Business, University of Richmond, Richmond, VA, USA 23217. E-mail: tzylkin@richmond.edu.

1 Introduction

Despite intense and longstanding empirical interest, the effects of bilateral trade agreements on trade are still considered highly difficult to assess. As emphasized in a recent practitioner's guide put out by the WTO (Yotov, Piermartini, Monteiro, and Larch, 2016), many current estimates in the literature suffer from easily identifiable sources of bias (or "estimation challenges"). This is not for a lack of awareness. Papers showing leading causes of bias in the gravity equation are often among the most widely celebrated and cited in the trade field, if not in all of Economics.[1] In particular, it is now generally accepted that trade flows across different partners are interdependent via "multilateral resistance" (the main contribution of Anderson and van Wincoop, 2003), that log-transforming the dependent variable is not innocuous (as argued by Santos Silva and Tenreyro, 2006), and—most relevant to the context of trade agreements—that earlier, puzzlingly small estimates of the effects of free trade agreements were almost certainly biased downwards by treating them as exogenous (Baier and Bergstrand, 2007).

As a consequence—and aided by some recent computational developments—researchers seeking to identify the effects of trade agreements have naturally moved towards more advanced estimation strategies that take on board all of the above concerns.[2] In particular, a "three-way" fixed effects Poisson Pseudo-Maximum Likelihood ("FE-PPML") model with time-varying exporter and importer fixed effects to account for multilateral resistance and time-invariant exporter-importer ("pair") fixed effects to address endogeneity has recently emerged as a logical workhorse model for empirical trade policy analysis.[3] A

[1] For some context, if we start citation counts in 2003, Anderson and van Wincoop (2003) and Santos Silva and Tenreyro (2006) are, respectively, the most cited articles in the *American Economic Review* and in the *Review of Economics and Statistics*. Paling only slightly in this exclusive company, Baier and Bergstrand (2007) is the 5th most-cited article in the *Journal of International Economics*, having gathered "only" 2,000 citations. Readers familiar with these other papers will also likely be familiar with Helpman, Melitz, and Rubinstein (2008)'s work on the selection process underlying zero trade flows, an issue we do not take up here.

[2] Larch, Wanner, Yotov, and Zylkin (2019), Correia, Guimarães, and Zylkin (2019), and Stammann (2018) describe algorithms that enable estimation of the three-way PPML models considered here.

[3] Pair fixed effects are of course no substitute for good instruments. However, instruments for trade policy changes which are also exogenous to trade are understandably hard to come by. As discussed in Head and Mayer (2014)'s essential handbook chapter on gravity estimation, pair fixed effects have the advantage that the effects of trade agreements and other trade policies are identified from time-variation in trade within pairs. Causal interpretations follow if standard "parallel trend" assumptions are satisfied.

clear conceptual obstacle, however, is the current lack of clarity regarding the asymptotic properties of a nonlinear estimator with more than two levels of fixed effects, especially in the standard "small T" case where the number of time periods is small relative to the number of countries. Even though FE-PPML models can be shown to be asymptotically unbiased with a single fixed effect (a well-known result) as well as in a two-way setting where both dimensions of the panel become large (Fernández-Val and Weidner, 2016), the latter result does not come strictly as a generalization of the former one, leaving it potentially unclear whether a three-way model with a fixed time dimension should be expected to inherit the nice asymptotic properties of these other models.

Accordingly, the question we investigate in this paper might simply be phrased as: "*Do three-way FE-PPML gravity models suffer from an incidental parameter problem (IPP)?*" As it turns out, there are two answers to this question: "*no... but also yes.*" From a traditional (i.e., small-T inconsistency) perspective, there is no IPP: because the first-order conditions of FE-PPML allow us to "profile out" the pair fixed effect terms from the first-order conditions of the other parameters, we can re-express the model as a two-way profile likelihood that we can then deconstruct using the basic approach established by Fernández-Val and Weidner (2016) for two-way asymptotic analysis. The three-way model is therefore consistent in fixed-T settings for largely the same reasons the two-way models considered in Fernández-Val and Weidner (2016) are consistent, and we provide suitably modified versions of the regularity conditions and consistency results established by Fernández-Val and Weidner (2016) for the simpler two-way case. Importantly, this consistency property turns out to be very specific to the FE-PPML estimator. As we are able to show, FE-PPML is in fact the only estimator among a wide range of related FE-PML gravity estimators that is generally consistent in this context when T is small.

At the same time, it does not also follow that Fernández-Val and Weidner (2016)'s earlier results for the *asymptotic unbiased*-ness of the two-way FE-PPML model similarly carry over to the three-way case. This is where the "*...but also yes*" part of our answer comes in. There is, in fact, a unique type of IPP in the three-way FE-PPML model that, to our knowledge, can only arise in models where there are different levels of fixed effects that grow large at different rates. Specifically, if N is the number of countries, profiling out the large (on the order of N^2) number of pair fixed effects eliminates any "$1/T$"-specific bias term that would normally be associated with a short time series. Using the heuristic suggested by Fernández-Val and Weidner (2018), we would then expect an asymptotic bias with an order given by the ratio between the order of the number of remaining parameters

(NT) and that of number of observations (N^2T), or $1/N$. However, due to the special properties of FE-PPML, the asymptotic bias in our setting behaves more like a $1/(NT)$ bias as N and T grow large at the same rate. The bias thus vanishes at a rate of $1/N$ as $N \to \infty$, ensuring consistency even for fixed T, and the estimator is actually unbiased as both N and $T \to \infty$, exactly like in the two-way FE-PPML model.[4]

What makes this bias a concern in fixed-T settings then is that the asymptotic standard deviation is of order $1/(N\sqrt{T})$; thus, the asymptotic bias in point estimates will always be of comparable magnitude to their standard errors when T is fixed. Put another way, without a bias correction, asymptotic confidence intervals will be incorrectly centered and will therefore produce misleading inferences, even as $N \to \infty$. This is effectively a version of the so-called "large T" IPP, so-named because this type of result typically only arises when taking asymptotics on the time dimension (e.g., Arellano and Hahn, 2007), usually for the purposes of deriving bias corrections for an estimator known to be inconsistent in short panels (e.g., Hahn and Newey, 2004).[5] Unlike in most other settings explored in this literature, and even though the size of the time dimension does play a role in conditioning the bias, the panel estimator we consider is consistent regardless of T. Nonetheless, the leading remedies recommended by the "large T" literature can still be adapted to reduce the bias and correct inferences.

Aside from the bias in point estimates, another (not unrelated) issue that affects the three-way model is a general downward bias in the cluster-robust sandwich estimator typically used to compute standard errors. This latter bias is similar to one that has been found in the simpler two-way gravity model by several recent studies (Egger and Staub, 2015; Jochmans, 2016; Pfaffermayr, 2019) and arises for the same reason: because the origin-time and destination-time fixed effects in the model each converge to their true values at a rate of only $1/\sqrt{N}$ (not $1/N$), the cluster-robust sandwich estimator for the variance has a leading bias of order $1/N$ (not $1/N^2$), and standard errors in turn have a bias of order $1/\sqrt{N}$. This latter type of bias is related to the general result that standard

[4] A similar IPP can arise for certain other three-way PML estimators aside from three-way FE-PPML. However, because these other estimators are generally inconsistent for fixed T, they will typically have an additional bias term of order $1/T$ that only disappears if the model is correctly specified.

[5] The new literature on "large T" asymptotic bias in nonlinear FE models has emerged as a recent response to the well-known "small T" consistency problem first described in Neyman and Scott (1948). Other examples include Phillips and Moon (1999), Hahn and Kuersteiner (2002), Lancaster (2002), Woutersen (2002), Alvarez and Arellano (2003), Carro (2007), Arellano and Bonhomme (2009), Fernández-Val and Vella (2011), and Kato, F. Galvao Jr., and Montes-Rojas (2012).

"heteroskedasticity-robust" variance estimators are downward-biased in small samples (see, e.g., MacKinnon and White, 1985; Imbens and Kolesar, 2016), including for PML models (Kauermann and Carroll, 2001). The fact that the bias in the sandwich estimator converges at a slower rate due to the incidental parameters merits special consideration on top of these already-known issues. We should therefore be concerned that estimated confidence intervals may be too narrow in addition to being off-center.

Our analysis provides theoretical characterizations of both of these issues as well as a series of possible bias corrections, which we evaluate using simulations and a real-data application. For the bias in point estimates, we construct two-way analytical and jackknife bias corrections inspired by the corrections proposed in Fernández-Val and Weidner (2016; 2018). For the bias in standard errors, we show how Kauermann and Carroll (2001)'s method for correcting the PML sandwich estimator may be adapted to the case of a conditional estimator with multi-way fixed effects and cluster-robust standard errors. Our simulations confirm that these methods are usually effective at improving inferences. The jackknife correction reduces more of the bias in point estimates than the analytical correction in smaller samples, but the analytical correction does a better job at improving coverage, especially when also paired with corrected standard errors.

For our empirical application, we estimate the average effects of a free trade agreement (FTA) on trade for a range of different industries using what would typically be considered a large trade data set, with 169 countries and 5 time periods. The biases we uncover vary in size across the different industries, but are generally large enough to indicate that our bias corrections should be worthwhile in most three-way gravity settings. For aggregate trade data (which yields results that are fairly representative), the estimated coefficient for FTA has an implied downward bias about 15%-18% of the estimated standard error, and the implied downward bias in the standard error itself is about 10% of the original standard error.

The literature on large-T IPPs with more than one fixed effect is small but growing. Aside from Fernández-Val and Weidner (2016)'s work on bias corrections for two-way nonlinear models, Pesaran (2006), Bai (2009), Hahn and Moon (2006), and Moon and Weidner (2017) have each conducted similar analyses for two-way linear models with interacted individual and time fixed effects. Turning to three-way models, Hinz, Stammann, and Wanner (2019) have recently developed bias corrections for dynamic three-way probit and logit models based on asymptotics suggested by Fernández-Val and Weidner (2018) where all three panel dimensions grow at the same rate. Though widely applica-

ble, this approach is not appropriate for our setting because of the different role played by the time dimension when the estimator is FE-PPML.[6] In the network context, Graham (2017), Dzemski (2018), and Chen, Fernández-Val, and Weidner (2019) have studied large-T IPPs in dyadic models where the different nodes in the network are characterized by node-specific (possibly sender- and receiver-specific) fixed effects. The analysis of Chen, Fernández-Val, and Weidner (2019) bears some especial similarity to our own in that they allow these node-specific effects to be vectors rather than scalars, similar to the exporter-time and importer-time fixed effects that feature in gravity models. Our bias expansions mainly differ from those of Chen, Fernández-Val, and Weidner (2019) because the equivalent outcome variable in our setting (trade flows observed over time for a given pair) is also a vector rather than a scalar and because we work with a conditional moment model where the distribution of the outcome may be misspecified. These distinctions are important because they together imply that the asymptotic bias is necessarily a function of the joint distribution of the outcome vector, a complication that does not arise in these other settings.

In what follows, Section 2 first provides a general overview of the no-IPP properties of the FE-PPML model (including the limits thereof). Section 3 then establishes bias and consistency results for the three-way gravity model specifically and discusses how to implement bias corrections. Sections 4 and 5 respectively present simulation evidence and an empirical application. Section 6 concludes, and an Appendix adds proofs and further simulation results.

2 FE-PPML Models and Incidental Parameters

In this section, we consider scenarios under which PPML models with various combinations of fixed effects may or may not suffer from an IPP. Our focus for now will be general; while our sights are ultimately set on three-way gravity models, it will first prove useful to present some other models that illustrate both what sets FE-PPML apart from other nonlinear FE models as well as its limitations in this context. As we will show, while FE-PPML is sometimes free from incidental parameter bias, even in settings with

[6] Also related are the GMM-based differencing strategies for two-way FE models proposed by Charbonneau (2017) and Jochmans (2016). These strategies rely on differencing the data in such as way that the resulting GMM moments do not depend on any of the incidental parameters. In principle, these methods could be extended to allow for differencing across a time dimension as well in a three-way panel.

multiple fixed effects, it is by no means immune to IPPs in general cases.

2.1 The Classic (One-way) Setting

The classic "one-way" FE setting is a natural way of demonstrating why FE-PPML models sometimes do not suffer from incidental parameter bias when other nonlinear FE models normally would. 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

$$\mathbb{E}(y_{it}|x_{it}, \alpha_i) = \lambda_{it} := \exp(x_{it}'\beta + \alpha_i). \tag{1}$$

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

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it} \left(y_{it} - \widehat{\lambda}_{it}\right) = 0, \qquad \forall i: \ \sum_{t=1}^{T} \left(y_{it} - \widehat{\lambda}_{it}\right) = 0, \tag{2}$$

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

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it} \left[y_{it} - \frac{\exp(x_{it}'\widehat{\beta})}{\sum_{\tau=1}^{T}\exp(x_{i\tau}'\widehat{\beta})} \sum_{\tau=1}^{T} y_{i\tau}\right] = 0, \tag{3}$$

which, as long as (1) 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).[7]

Of course, with a doubly-indexed panel indexed by individuals and time, a standard approach here would be to also include a time fixed effect for each period t. For small T, the addition of time fixed effects has little effect on the above example: the small number of time dummies needed for the fixed effects can be thought of as components of

[7] The earliest references to present versions of this result include Andersen (1970), Palmgren (1981), and Hausman, Hall, and Griliches (1984). Another important contribution is Wooldridge (1999), who shows that FE-PPML is consistent even when the assumed distribution of the data is misspecified. Our Lemma 2 in the Appendix clarifies that FE-PPML is relatively unique in this regard versus similar models.

x_{it} without loss of generality and are therefore consistently estimated for the same reasons the other components of x_{it} are consistently estimated. A more interesting case is where T is large, such that we have a more complex, "two-way" estimator where both dimensions of the panel—individual and time—grow with the sample. As shown by Fernández-Val and Weidner (2016)—and as we ourselves will show shortly—a two-way FE-PPML model of this type is again consistent and exhibits no asymptotic bias. Thus, this series of results may create the impression that Poisson models are immune to IPPs, regardless of how many fixed effects are included or which dimensions of the panel grow with the sample. The following discussion makes it clear this is not generally the case.

2.2 Overlapping Fixed Effects

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 effect. Some standard examples include the gravity model from international trade (which we discuss next) 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 $N \to \infty$.

In this example, 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 our gravity examples will show), 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 Example 1.

The easiest way to demonstrate that this type of model suffers from an IPP is by way of simulations. The top-left panel of Fig. 1 presents simulated FE-PPML estimates of β based on Example 1 using panel sizes of $N = 100$, $N = 1,000$, and $10,000$. For ease of exposition, the conditional distribution of y_{ijt} is assumed to be log-normal with variance equal to λ_{ijt} (as in a Poisson distribution), but we have found similar results for other data-generating assumptions such as those described in Section 4. The true value for β is 1, and values for x, α, and γ are constructed using the same methods as Fernández-Val and Weidner (2016). The results show that FE-PPML clearly suffers from an IPP in this example. Even for the largest panel size where $N = 10,000$, the mass point of the simulated distribution for $\widehat{\beta}$ is about 1.1-1.15, and the estimates do not show signs of converging to the true estimate of $\beta = 1$ as the panel size increases.

Gravity models, by contrast, also feature multiple levels of overlapping fixed effects, but generally either the number of fixed effects grows at a slower rate than the size of the panel—as in the two-way gravity model—or there is only one fixed effect dimension that grows at the same rate as the panel size—as in the three-way gravity model usually recommended for trade policy analysis. Determining whether an IPP is present (and what type) for FE-PPML applied to gravity models therefore requires a closer examination of these models, which we now turn to.

2.3 Two-way Gravity Models

We introduce the concept of a "gravity model" as follows. Countries are indexed by $i, j \in \mathfrak{N} := \{1, \ldots, N\}$, with $i \neq j$, and y_{ij} is the volume of trade between i and j.[8] In general, we allow there to be $T > 1$ time periods, such that a time subscript will also be needed, but for the time being we will suppose $T = 1$. Exporter- and importer- specific fixed effects are in this setting denoted α_i and γ_j. The model reads

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

[8]This panel structure can be easily relaxed to allow the number of exporters and the number of importers to be different; the real key here is that we assume both dimensions of the panel grow at the same rate asymptotically.

The FE-PPML estimator maximizes $\sum_{i=1}^{n}\sum_{j\neq i}\left(y_{ij}\log\lambda_{ij}+\lambda_{ij}\right)$ over β, α, and γ, where x_{ij} would normally contain a set of exogenous bilateral regressors (e.g., the log of geographic distance, the sharing of a common border, and so on).

From Fernández-Val and Weidner (2016), we know that $\sqrt{N\left(N-1\right)}(\widehat{\beta}-\beta^{0})\rightarrow_{d}\mathcal{N}(0,V)$ as $N\rightarrow\infty$. That is, in contrast to what we found above for the model in Example 1, we have no IPP here (neither an inconsistency nor an asymptotic bias problem).[9] The reasons behind this result are twofold. First, although we consider an asymptotic setting where both fixed effect dimensions (the number of exporters and the number of importers) grow with N, the sample size grows with N^2; all α_i's and γ_j's are therefore consistently estimated as $N\rightarrow\infty$, and β is in turn consistently estimated as well. Second, for the FE-PPML model specifically, we can either solve for $\widehat{\alpha}_i$ or solve for $\widehat{\gamma}_j$ to obtain a profile score for the remaining parameters (including $\widehat{\beta}$) that is asymptotically unbiased as $N\rightarrow\infty$.[10] The simulations presented in the top-right panel of Fig. 1 provide a visual illustration of this property, confirming that estimates are correctly centered regardless of N.

These results might perhaps create the impression that FE-PPML gravity models generally inherit all the same no-IPP properties as the classic one-way panel data model. As we will now discuss in detail, the three-way FE-PPML gravity model only inherits some, not all, of these nice properties. As we will also see, this impression is misleading for other reasons as well: even for the two-way model, while the α_i and γ_j parameters do not affect the score for $\widehat{\beta}$, they nonetheless have implications for the estimated variance of $\widehat{\beta}$ that are not innocuous; we thus will also devote some attention to whether the three-way model suffers from a similar issue.

[9] Note that Theorem 4.1 in Fernández-Val and Weidner (2016) is written for the correctly specified case, where y_{ij} is actually Poisson distributed. However, Remark 3 in the paper gives the extension to conditional moment models, where for the FE-PPML case only $\mathbb{E}(y_{ij}|x_{ij},\alpha_i,\gamma_j)=\exp(x_{ij}'\beta+\alpha_i+\gamma_j)$ needs to hold. That remark also states that the asymptotic bias of the FE-PPML estimator $\widehat{\beta}$ is zero; that is, no bias correction is necessary for valid asymptotic inference here. Their paper considers standard panel models, as opposed to trade models, but the only technical difference is that y_{ij} is often not observed for the trade model when $i\neq j$. This missing diagonal has no effect on any of the results we discuss.

[10] Egger, Larch, Staub, and Winkelmann (2011) have previously observed that the two-way FE-PPML estimator is consistent in this setting, as is any two-way FE-PML estimator where both dimensions of the panel increase with the square root of the sample size. However, as shown by Fernández-Val and Weidner (2016), the no-bias result for FE-PPML does not extend to other similar estimators in this context.

3 Results for the Three-way Gravity Model

To recap the sequence of results just described, we know that FE-PPML estimates with one fixed effect do not suffer from an IPP. We also know that FE-PPML may have an IPP in models with more than one fixed effect, but it is both consistent and asymptotically unbiased in two-way gravity settings when neither fixed effect dimension grows at the same rate as the size of the panel. As we will now show, each of these earlier results will be useful for understanding the more complex case of a three-way gravity model where we add a time dimension and a third set of fixed effects to the above two-way model. We also describe a series of bias corrections for the three-way model, including for the possible downward bias of the estimated standard errors.

3.1 Consistency

To formally introduce the three-way model, we now add an explicit time subscript $t \in \{1, \ldots, T\}$ to y_{ij}, x_{ij}, α_i, and γ_j to the prior model and also add a bilateral (or "country-pair")-specific fixed effect η_{ij}. The model now reads as

$$\mathbb{E}(y_{ijt}|x_{ijt}, \alpha_{it}, \gamma_{jt}, \eta_{ij}) = \lambda_{ijt} := \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij}), \tag{4}$$

where the three fixed effects now respectively index exporter-time, importer-time, and country-pair.[11] To append an error term, we further assume $y_{ijt} = \lambda_{ijt}\omega_{ijt} \geq 0$, with $\omega_{ijt} \geq 0$ serving as a residual. For the asymptotics using the three-way model, we consider T fixed, while $N \to \infty$. The FE-PPML estimator maximizes

$$\mathcal{L}(\beta, \alpha, \gamma, \eta) := \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{t=1}^{T} (y_{ijt} \log \lambda_{ijt} + \lambda_{ijt})$$

over β, α, γ and η.

With the added country-pair fixed effect η, notice that not all of the fixed effect dimensions grow at the same rate as N increases. The numbers of exporter-time and importer-time fixed effects each increase with N (as before), but the dimension of η increases with N^2, since adding another country to the data adds another $N - 1$ trade flows to the estimation. It therefore makes sense to first "profile out" η (as we did with α

[11] For discussion of this model, see Yotov, Piermartini, Monteiro, and Larch (2016) or Larch, Wanner, Yotov, and Zylkin (2019).

in (3)), so that we may deal with the remaining two fixed effects in turn. For given values of β, α, γ the maximizer over η satisfies

$$\exp\left[\widehat{\eta}_{ij}(\beta,\alpha,\gamma)\right] = \frac{\sum_{t=1}^{T} y_{ijt}}{\sum_{t=1}^{T} \mu_{ijt}}, \qquad \mu_{ijt} := \exp(x_{ijt}'\beta + \alpha_{it} + \gamma_{jt}). \tag{5}$$

We therefore have

$$\mathcal{L}(\beta,\alpha,\gamma) := \max_{\eta} \mathcal{L}(\beta,\alpha,\gamma,\eta) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \ell_{ij}(\beta,\alpha_{it},\gamma_{jt}), \tag{6}$$

with

$$\begin{aligned}\ell_{ij}(\beta,\alpha_{it},\gamma_{jt}) &:= \sum_{t=1}^{T}\left[y_{ijt}\log\left(\frac{\mu_{ijt}}{\sum_{s=1}^{T}\mu_{ijs}}\right) + \frac{\mu_{ijt}}{\sum_{s=1}^{T}\mu_{ijs}}\sum_{s=1}^{T} y_{ijs}\right] + \sum_{t=1}^{T} y_{ijt}\log\left(\sum_{s=1}^{T} y_{ijs}\right). \\ &= \sum_{t=1}^{T} y_{ijt}\log\left(\frac{\mu_{ijt}}{\sum_{s=1}^{T}\mu_{ijs}}\right) + \text{terms not depending on any parameters.}\end{aligned} \tag{7}$$

Thus, after profiling out the η_{ij} parameters, we are left with the likelihood of a multinomial model where the only incidental parameters are α_{it} and γ_{jt}. The FE-PPML estimators for β, α_{it} and γ_{jt} are given by

$$(\widehat{\beta},\widehat{\alpha},\widehat{\gamma}) = \underset{\beta,\alpha,\gamma}{\operatorname{argmax}}\, \mathcal{L}(\beta,\alpha,\gamma). \tag{8}$$

Using (4) one can easily verify that

$$\mathbb{E}\left[\frac{\partial \ell_{ij}(\beta,\alpha_{it},\gamma_{jt})}{\partial\beta}\right] = 0, \qquad \mathbb{E}\left[\frac{\partial \ell_{ij}(\beta,\alpha_{it},\gamma_{jt})}{\partial\alpha_{it}}\right] = 0, \qquad \mathbb{E}\left[\frac{\partial \ell_{ij}(\beta,\alpha_{it},\gamma_{jt})}{\partial\gamma_{jt}}\right] = 0. \tag{9}$$

Thus, after profiling out η_{ij}, there is no bias in the score of the profile log-likelihood $\ell_{ij}(\beta,\alpha_{it},\gamma_{jt})$. The reason for this is exactly the same as for the no-IPP result in the classic panel setting above. Furthermore, note that the only fixed effects that need to be estimated in $\ell_{ij}(\beta,\alpha_{it},\gamma_{jt})$ are α_{it} and γ_{jt}, which only grow with the square root of the sample size as $N\to\infty$, implying that they are consistently estimated. Thus, we can state the following result:

Proposition 1. *So long as the set of non-fixed effect regressors x_{ijt} is exogenous to the residual disturbance ω_{ijt} after conditioning on the fixed effects α_{it}, γ_{jt}, and η_{ij}, FE-PPML estimates of β from the three-way gravity model are consistent for $N\to\infty$.*[12]

[12]This consistency result can be seen as a corollary of the asymptotic normality result in Proposition 3 below, for which formal regularity conditions are stated in Assumption A of the Appendix.

This result follows because we can re-write the three-way FE-PPML estimator as a two-way estimator without introducing a $1/T$-bias, such that the earlier consistency result from Fernández-Val and Weidner (2016) for two-way estimators can again be applied. In other words, the three-way FE-PPML model is consistent as $N \to \infty$ largely for the same reason two-way FE-PPML and other two-way nonlinear gravity estimators are generally consistent. However, in the context of *three-way* estimators, we can also state a stronger result that applies more narrowly to FE-PPML in particular:

Proposition 2. *Consider the class of "three-way" FE-PML gravity estimators with conditional means given by* $\lambda_{ijt} := \exp(x'_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \eta_{ij})$ *and FOC's given by*

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

$$\widehat{\gamma}_{jt}\text{:} \sum_{i=1}^{N}\Big(y_{ijt} - \widehat{\lambda}_{ijt}\Big)g(\widehat{\lambda}_{ijt}) = 0, \qquad \widehat{\eta}_{ij}\text{:} \sum_{t=1}^{T}\Big(y_{ijt} - \widehat{\lambda}_{ijt}\Big)g(\widehat{\lambda}_{ijt}) = 0,$$

where $i, j = 1, \ldots, N$, $t = 1, ..., T$, *and* $g(\widehat{\lambda}_{ijt})$ *is an arbitrary function of* $\widehat{\lambda}_{ijt}$. *If* T *is small, then for* $\widehat{\beta}$ *to be consistent under general assumptions about* $\mathrm{Var}(y|x, \alpha, \gamma, \eta)$, *we must have that* $g(\lambda_{ijt})$ *is constant over the range of* λ*'s that are realized in the data-generating process. That is, the estimator must be equivalent to FE-PPML.*

The details behind this latter result are somewhat subtle. Clearly, for arbitrary $g(\widehat{\lambda}_{ijt})$, it is generally not possible to write down a closed form solution $\widehat{\eta}_{ij} = \ln\sum_{t=1}^{T} y_{ijt}g(\widehat{\lambda}_{ijt}) - \ln\sum_{t=1}^{T}\mu_{ijt}g(\widehat{\lambda}_{ijt})$ that would allow us to derive a two-way profile likelihood that does not depend on $\widehat{\eta}_{ij}$. However, as we discuss in the Appendix, it is still possible to obtain a two-way profile likelihood if $g(\widehat{\lambda}_{ijt})$ is of the form $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}^{q}$, where q can be any real number. Notably, this latter class of models not only includes FE-PPML (for which $q = 0$), but also includes other popular gravity estimators such as Gamma PML ($q = -1$) and Gaussian PML ($q = 1$). And yet, the existence of equivalent profile likelihood expressions for these other estimators does not guarantee that they are consistent. Actually, the three-way gravity estimators associated with $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}^{q}$ can be shown to suffer from a $1/T$-bias that only disappears if either $q = 0$ (in which case the estimator is FE-PPML) or if the conditional variance is proportional to λ_{ijt}^{1-q} (in which case the estimator inherits the properties of the MLE estimator).

3.2 Asymptotic Bias

Because the three-way FE-PPML model inherits the consistency properties of the two-way estimator, one might expect that it also inherits its *unbiased*-ness properties as well. However, this is where the limitations of PPML's no-IPP properties become apparent. While the profile log-likelihood in (6) is now of a similar form to the two-way models considered in Fernández-Val and Weidner (2016), notice that it no longer resembles the original FE-PPML log-likelihood. The no-bias result for two-way FE-PPML from Fernández-Val and Weidner (2016) therefore does not carry over to the profile log-likelihood and it is possible to show that FE-PPML has an asymptotic bias in this setting.

Preliminaries

As with the models considered in Fernández-Val and Weidner (2016), the origins of this bias have to do with the rate at which the estimated incidental parameters $\widehat{\alpha}_i$ and $\widehat{\gamma}_j$ converge to their true values α_i^0 and γ_j^0. As such, it will be useful to pause here to establish to some additional notation, mostly to provide some shorthand for the higher-order partial derivatives of ℓ_{ij} with respect to $\widehat{\alpha}_i$ and $\widehat{\gamma}_j$. To this end, let

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

It will also be convenient to let $\vartheta_{ijt} := \lambda_{ijt} / \sum_\tau \lambda_{ij\tau}$. With ℓ_{ij} now expressed in similar form to the objective function considered in Fernández-Val and Weidner (2016), we can now define the following objects:

- $S_{ij} := \partial \ell_{ij} / \partial \pi_{ij}$ is a $T \times 1$ vector with elements $y_{ijt} - \vartheta_{ijt} \sum_\tau y_{ij\tau}$.
- $H_{ij} := -\partial^2 \ell_{ij} / \partial \pi_{ij} \partial \pi_{ij}'$ gives us a $T \times T$ matrix with diagonal elements $\vartheta_{ijt}(1 - \vartheta_{ijt}) \sum_\tau y_{ij\tau}$ and off-diagonal ($s \neq t$) elements given by $-\vartheta_{ijs}\vartheta_{ijt} \sum_\tau y_{ij\tau}$.
- $G_{ij} := \partial^3 \ell_{ij} / \partial \pi_{ij} \partial \pi_{ij}' \partial \pi_{ijt}$ is a $T \times T \times T$ cubic tensor. The elements on the main diagonal of G_{ij} are given by $-\vartheta_{ijt}(1 - \vartheta_{ijt})(1 - 2\vartheta_{ijt}) \sum_\tau y_{ij\tau}$. The elements of the 3 planar diagonals with $r = s \neq t$ are given by $\vartheta_{ijs}(1 - 2\vartheta_{ijs})\vartheta_{ijt} \sum_\tau y_{ij\tau}$. All other elements with $r \neq s \neq t$ are given by $-2\vartheta_{ijr}\vartheta_{ijs}\vartheta_{ijt} \sum_\tau y_{ij\tau}$.

The value of presenting these objects is that they allow us to easily form other terms we need that help define how $\widehat{\beta}$ depends on $\widehat{\alpha}_i$ and $\widehat{\gamma}_j$. For example, S_{ij} not only doubles for both $\partial \ell_{ij}/\partial \alpha_i$ as well as for $\partial \ell_{ij}/\partial \gamma_j$, but also allows us to obtain $\partial \ell_{ij}/\partial \beta^k = x'_{ij,k} S_{ij}$. Likewise, we also have that $\partial^2 \ell_{ij}/\partial \alpha_i \partial \alpha_i' = \partial^2 \ell_{ij}/\partial \alpha_i \partial \gamma_j' = \partial^2 \ell_{ij}/\partial \gamma_j \partial \gamma_j' = -H_{ij}$, $\partial^2 \ell_{ij}/\partial \alpha_i \partial \beta^k = \partial^2 \ell_{ij}/\partial \gamma_j \partial \beta^k = -H_{ij} x_{ij,k}$ and that

$$\frac{\partial^3 \ell_{ij}}{\partial \alpha_i \partial \alpha_i' \partial \beta^k} = \frac{\partial^3 \ell_{ij}}{\partial \alpha_i \partial \gamma_j' \partial \beta^k} = \frac{\partial^3 \ell_{ij}}{\partial \gamma_j \partial \alpha_i' \partial \beta^k} = \frac{\partial^3 \ell_{ij}}{\partial \gamma_j \partial \gamma_j' \partial \beta^k} = G_{ij} x_{ij,k},$$

where it is important to note that the product $G_{ij} x_{ij,k}$ is a $T \times T$ matrix with individual elements $[G_{ij} x_{ij,k}]_{st} = \sum_r G_{ijrst} x_{ijr,k}$. In addition, we will for the most part assume that score vectors are conditionally independent of one another—i.e., $\text{Cov}\left(S_{ij}, S_{i'j'} \,\middle|\, x_{ij}\right) = 0$ if $(i,j) \neq (i',j')$—though this assumption can be relaxed, as we explain later on.

The remaining preliminaries then require that we also define the expected Hessian $\bar{H}_{ij} = \mathbb{E}\left(H_{ij} \,\middle|\, x_{ij}\right)$. Because we have not chosen a normalization for α_i and γ_j, $\bar{H}_{ij}$ is only positive semi-definite (not positive definite). Therefore, we will use a Moore-Penrose pseudoinverse, to be denoted with a $\dagger$, whenever the analysis requires we work with an inverse of $\bar{H}_{ij}$ or similar matrices.[13] We likewise find it useful to define $\bar{G}_{ij} = \mathbb{E}(G_{ij})$.

Finally, with $\bar{H}_{ij}$ in hand, we can define the within-transformed regressor matrix $\widetilde{x}_{ij} := x_{ij} - \alpha_i^x - \gamma_j^x$, where α_i^x and γ_j^x are $T \times K$ matrices that minimize

$$\sum_{i=1}^{N} \sum_{j \in \mathfrak{N} \setminus \{i\}} \text{Tr}\left[\left(x_{ij} - \alpha_i^x - \gamma_j^x\right)' \bar{H}_{ij} \left(x_{ij} - \alpha_i^x - \gamma_j^x\right)\right], \tag{10}$$

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 $\bar{H}_{ij}$.[14]

[13] Specifically, we have that $\bar{H}_{ij}\, \iota_T = 0$, where $\iota_T = (1, \ldots, 1)'$ is a T-vector of ones. Thus, $\bar{H}_{ij}$ is only of rank $T-1$ rather than of rank T. The Moore-Penrose inverse allows us to avoid the problem of choosing what normalizations to use for α_i and γ_j while still leading to the same end results.

[14] 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.

Bias Expansion

As in Fernández-Val and Weidner (2016), we can characterize the asymptotic bias in $\widehat{\beta}$ by examining how the estimated fixed effects $\widehat{\alpha}_i$ and $\widehat{\gamma}_j$ enter the score for $\widehat{\beta}$. The full details behind this derivation are left for the Appendix, but the following second-order expansion provides a general basis. Let $\phi := \text{vec}(\alpha, \gamma)$ be a vector that collects all of the two-way incidental parameters, such that we can again re-express ℓ_{ij} slightly as $\ell_{ij} = \ell_{ij}(\beta, \phi)$. We can then define the function $\widehat{\phi}(\beta)$ as

$$\widehat{\phi}(\beta) := \arg\max_{\phi} \frac{1}{N(N-1)} \sum_{i,j} \ell_{ij}(\beta, \phi),$$

which allows us to succinctly characterize the estimated values for $\widehat{\alpha}$ and $\widehat{\gamma}$ as a function of β. Next, we construct a second-order expansion of the expected score for $\widehat{\beta}$ around the true incidental parameter vector ϕ^0 and evaluated at the true parameter β^0:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \ell_{ij}(\beta^0, \widehat{\phi}(\beta^0))}{\partial \beta}\right] \approx \mathbb{E}\left[\frac{\partial \ell_{ij}(\beta^0, \phi^0)}{\partial \beta}\right] + \mathbb{E}\left[\frac{\partial^2 \ell_{ij}(\beta^0, \phi^0)}{\partial \beta \partial \phi'}\left(\widehat{\phi}(\beta^0) - \phi^0\right)\right] \\ + \frac{1}{2}\sum_{f,g}^{\dim \phi} \mathbb{E}\left[\frac{\partial^3 \ell_{ij}(\beta^0, \phi^0)}{\partial \beta \partial \phi_f \partial \phi_g}\left(\widehat{\phi}_f(\beta^0) - \phi_f^0\right)\left(\widehat{\phi}_g(\beta^0) - \phi_g^0\right)\right]. \qquad (11)\end{aligned}$$

This expression is near-identical to a similar expansion that appears in Fernández-Val and Weidner (2016)—differing mainly in that ℓ_{ij} is a vector rather than a scalar—and communicates the same essential insights: because the latter two terms in (11) are generally $\neq 0$, the score for $\widehat{\beta}$ is biased, with the bias depending on the interaction between the higher-order partial derivatives of ℓ_{ij} and the estimation error in the incidental parameters as well as their variances and covariances.

After dropping terms that are asymptotically small[15] and plugging in the just-defined expressions S_{ij}, H_{ij}, G_{ij}, and $\widetilde{x}_{ij}$ where appropriate, we can use (11) to obtain a tractable expression for the bias that serves as the centerpiece of the following proposition.

Proposition 3. *Under appropriate regularity conditions (Assumption A in the Appendix), for T fixed and $N \to \infty$ we have*

$$\sqrt{N(N-1)}\left(\widehat{\beta} - \beta^0 - \frac{W_N^{-1}(B_N + D_N)}{N-1}\right) \to_d \mathcal{N}\left(0, W_N^{-1}\,\Omega_N\,W_N^{-1}\right),$$

[15] In particular, all elements of the cross-partial objects $\mathbb{E}[\partial^2 \ell_{ij}/\partial\alpha_i\partial\gamma_j]$, $\mathbb{E}[\partial^3 \ell_{ij}/\partial\alpha_i\alpha_i'\partial\gamma_j]$, etc. can be shown to be asymptotically small. Thus, in what follows, B_N reflects the contribution of the α_i parameters to the bias and D_N reflects the contribution of the γ_j parameters.

where W_N and Ω_N are $K \times K$ matrices given by

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

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

and B_N and D_N are K-vectors with elements given by

$$B_N^k = -\frac{1}{N} \sum_{i=1}^{N} \mathrm{Tr}\left[\left(\sum_{j \in \mathfrak{N} \setminus \{i\}} \bar{H}_{ij}\right)^{\dagger} \sum_{j \in \mathfrak{N} \setminus \{i\}} \mathbb{E}\left(H_{ij}\, \widetilde{x}_{ij,k}\, S'_{ij} \middle| x_{ij,k}\right)\right]$$
$$+ \frac{1}{2\,N} \sum_{i=1}^{N} \mathrm{Tr}\left[\left(\sum_{j \in \mathfrak{N} \setminus \{i\}} \bar{G}_{ij}\, \widetilde{x}_{ij,k}\right)\left(\sum_{j \in \mathfrak{N} \setminus \{i\}} \bar{H}_{ij}\right)^{\dagger}\left[\sum_{j \in \mathfrak{N} \setminus \{i\}} \mathbb{E}\left(S_{ij}\, S'_{ij} \middle| x_{ij,k}\right)\right]\left(\sum_{j \in \mathfrak{N} \setminus \{i\}} \bar{H}_{ij}\right)^{\dagger}\right],$$

$$D_N^k = -\frac{1}{N} \sum_{j=1}^{N} \mathrm{Tr}\left[\left(\sum_{i \in \mathfrak{N} \setminus \{j\}} \bar{H}_{ij}\right)^{\dagger} \sum_{i \in \mathfrak{N} \setminus \{j\}} \mathbb{E}\left(H_{ij}\, \widetilde{x}_{ij,k}\, S'_{ij} \middle| x_{ij,k}\right)\right]$$
$$+ \frac{1}{2\,N} \sum_{j=1}^{N} \mathrm{Tr}\left[\left(\sum_{i \in \mathfrak{N} \setminus \{j\}} \bar{G}_{ij}\, \widetilde{x}_{ij,k}\right)\left(\sum_{i \in \mathfrak{N} \setminus \{j\}} \bar{H}_{ij}\right)^{\dagger}\left[\sum_{i \in \mathfrak{N} \setminus \{j\}} \mathbb{E}\left(S_{ij}\, S'_{ij} \middle| x_{ij,k}\right)\right]\left(\sum_{i \in \mathfrak{N} \setminus \{j\}} \bar{H}_{ij}\right)^{\dagger}\right].$$

The above proposition establishes the asymptotic distribution of the three-way gravity estimator as $N \to \infty$, including the asymptotic bias $(N-1)^{-1} W_N^{-1}(B_N + D_N)$. Intuitively, this bias can be decomposed as the product of the inverse expected Hessian with respect to β (i.e. W_N^{-1}) and the bias of the score in (11), which in turn is captured by the two-way bias terms B_N and D_N and the rate of asymptotic convergence (essentially $1/N$). In the two-way FE-PPML setting considered in Fernández-Val and Weidner (2016), we would have that $B_N = D_N = 0$, such that $\widehat{\beta}$ is unbiased. Importantly, and unlike in the two-way FE-PPML setting, the three-way model does not give us the no-bias result that $B_N = D_N = 0$, as we will illustrate in more detail momentarily.

What if T is Large?

While Proposition 3 only focuses on asymptotics where $N \to \infty$, the three-way gravity panel also features a time dimension (T), and it is interesting to wonder how the above results may depend on changes in T. The following remark clarifies how the bias terms B_N and D_N can be re-written to illuminate the role of the time dimension.

Remark 1. *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 of the form (8) based on concave objective functions $\ell_{ij}(\beta, \alpha_{it}, \gamma_{jt})$. Unlike with two-way FE-PPML models, these formulas do not reduce to zero when we further specialize them to the profiled Poisson pseudo-likelihood in (7), but we still find it instructive to do so (e.g. to discuss the large T limit below). 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 $i, j \in \{1, \ldots, N\}$ 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

$$B_N^k = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in \mathfrak{N} \setminus \{i\}} \left[-\frac{\iota_T' \, R_{ij} \, \widetilde{x}_{ij,k}}{\iota_T' \lambda_{ij}} + \frac{\lambda_{ij}' \, Q_i \, \Lambda_{ij} \, M_{ij}' \, \widetilde{x}_{ij,k}}{\iota_T' \lambda_{ij}} \right], \tag{12}$$

and an analogous formula for D_N follows by interchanging i and j appropriately.

So long as there is only weak time dependence between observations belonging to the same pair (in the sense described by Hansen, 2007), the matrix objects R_{ij} and $Q_i \Lambda_{ij} M_{ij}'$ in Remark 1 are both of order 1 as $T \to \infty$, such that both terms in brackets in (12) are likewise of order 1.[16] We will henceforth assume any time dependence is weak. Remark 2 then describes some additional asymptotic results for when T is large.

Remark 2. *Under asymptotics where $T \to \infty$, we have the following:*

(i) If N is fixed and $T \to \infty$, then $\widehat{\beta}$ is generally inconsistent.

(ii) As $T \to \infty$, the combined bias term $W_N^{-1}(B_N + D_N)$ goes to zero at a rate of $1/T$. Therefore, because the standard error is of order $1/(N\sqrt{T})$, there is no bias in the asymptotic distribution of $\widehat{\beta}$ as N and T both $\to \infty$.

[16] By "weak" time dependence, we mean that any such dependence dissipates as the temporal distance between observations increases. Alternatively, if observations are correlated regardless of how far apart they are in time, the standard error is always of order $1/N$ (see Hansen, 2007), and the same will also be true for the asymptotic bias. The latter is arguably a less natural assumption in this context, however.

To elaborate further, letting $T \to \infty$ is obviously not sufficient for either α or γ to be consistently estimated and does not solve the IPP, as stated in part (i). However, as part (ii) tells us, T still plays an interesting role in conditioning the bias when both N and T jointly become large. Intuitively, because W_N^{-1} is of order $1/T$, and because B_N and D_N are bounded as $T \to \infty$, the bias in $\widehat{\beta}$ effectively vanishes at a rate of $1/(NT)$ as both $N, T \to \infty$, such that it increasingly shrinks in relation to the order-$1/(N\sqrt{T})$ standard error. This is what we mean when we say the IPP the three-way PPML model suffers from is rather unique: it can be resolved by large enough T (like most IPPs), yet large T is actually neither necessary nor sufficient to ensure consistency.

Illustrating the Bias using the $T = 2$ Case

Admittedly, the complexity of the objects that appear in Proposition 3 may make it difficult to appreciate the general point that the three-way estimator is not unbiased. One way to make these details more transparent is to focus our attention on the simplest possible panel model where $T = 2$. The convenient thing about this simplified setting is the likelihood function ℓ_{ij} can be reduced to just a scalar: $\ell_{ij} = y_{ij1} \log \vartheta_{ij1} + y_{ij2} \log (1 - \vartheta_{ij1})$, where

$$\vartheta_{ij1} = \frac{\exp(\Delta x_{ij}\beta + \pi_{ij})}{\exp(\Delta x_{ij}\beta + \pi_{ij}) + 1}$$

and where $\Delta x_{ij} = x_{ij1} - x_{ij2}$ and $\pi_{ij} = \pi_{ij1} - \pi_{ij2}$. Importantly, these normalizations allow us to express $\partial \ell_{ij}/\partial \pi_{ij}$, $\partial^2 \ell_{ij}/\partial \pi_{ij}^2$, etc. as also just scalars, and we can therefore easily derive the following result:

Remark 3. *For $T = 2$, we calculate $S_{ij} = \vartheta_{ij2} y_{ij1} - \vartheta_{ij1} y_{ij2}$, $H_{ij} = \vartheta_{ij1}\vartheta_{ij2}(y_{ij1} + y_{ij2})$, $\bar{H}_{ij} = \vartheta_{ij1}\lambda_{ij2}$, $G_{ij} = \vartheta_{ij1}\vartheta_{ij2}(\vartheta_{ij1} - \vartheta_{ij2})(y_{ij1} + y_{ij2})$, $\bar{G}_{ij} = \vartheta_{ij1}(\vartheta_{ij1} - \vartheta_{ij2})\lambda_{ij2}$, and $\Delta\widetilde{x}_{ij} = \widetilde{x}_{ij1} - \widetilde{x}_{ij2}$. The bias term B_N^k in Proposition 3 can then be written as*

$$B_N^k = \operatorname*{plim}_{N\to\infty} \left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j\neq i} \Delta\widetilde{x}_{ij}\vartheta_{ij1}\vartheta_{ij2}\left[\vartheta_{ij2}\mathbb{E}(y_{ij1}^2) - \vartheta_{ij1}\mathbb{E}(y_{ij2}^2) + (\vartheta_{ij2} - \vartheta_{ij1})\mathbb{E}(y_{ij1}y_{ij2})\right]}{\sum_{j\neq i}\vartheta_{ij1}\lambda_{ij2}} \right.$$
$$\left. + \frac{1}{2N} \sum_{i=1}^{N} \frac{\left\{\sum_{j\neq i}\Delta\widetilde{x}_{ij}\vartheta_{ij1}(\vartheta_{ij1} - \vartheta_{ij2})\lambda_{ij2}\right\}\left\{\sum_{j=1}^{N}\vartheta_{ij2}^2\mathbb{E}(y_{ij1}^2) + \vartheta_{ij1}^2\mathbb{E}(y_{ij2}^2) - 2\vartheta_{ij1}\vartheta_{ij2}\mathbb{E}(y_{ij1}y_{ij2})\right\}}{\left[\sum_{j\neq i}\vartheta_{ij1}\lambda_{ij2}\right]^2} \right],$$

with an analogous expression also following for D_N^k.

Two points then stand out based on the above expression. First, unlike in the two-way FE-PPML case, neither of the two terms in B_N^k generally equals 0. Even in the correctly

specified case (where $\mathbb{E}(y_{ij1}^2) = \lambda_{ij1}^2 + \lambda_{ij1}$ and $\mathbb{E}(y_{ij1}y_{ij2}) = \lambda_{ij1}\lambda_{ij2}$), the first term can be shown to cancel, but the second term does not, because $\sum_j \bar{G}_{ij}\Delta\widetilde{x}_{ij} \neq 0$. This is very different from the two-way case where $\bar{H}_{ij} = \bar{G}_{ij} = -\lambda_{ij}$. In that case, both terms in B_N^k and D_N^k always cancel, regardless of whether the PPML model is correctly specified. The difference can be appreciated by comparing simulation results from the top-right panel of Fig. 1, which are based on the two-way model and are therefore unbiased, with those from the bottom-left panel, which are based on the three-way model with $T = 2$ and show a clear asymptotic bias.

Second, it is plain from Remark 3 that both terms in the bias generally depend on the expected second moments of y_{ij} (e.g., $\mathbb{E}(y_{ij1}^2)$, $\mathbb{E}(y_{ij1}y_{ij2})$, etc.). This is again different from the models that were previously considered in Fernández-Val and Weidner (2016).[17] Among other things, the difficulty associated with estimating these second moments means that analytical bias corrections may not necessarily offer superior performance relative to distribution-free methods such as the jackknife. It also means that allowing for conditional dependence between pairs may change the expression of bias, as we discuss next.

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

[17] The specific examples used in Fernández-Val and Weidner (2016) are the Poisson model, which is unbiased, and the probit model, which requires the distribution of y_{ij} to be correctly specified. They also provide a bias expansion for "conditional moment" models that allow the distribution of y_{ij} to be misspecified. Beyond this theoretical discussion, bias corrections for misspecified models have yet to receive much attention, however. As can be seen above, an important complication that arises for these models is that the bias depends on the distribution of the data, which is typically treated as unknown.

additional clustering; namely, we would need

$$\begin{aligned}\Omega_N &= \frac{1}{N\left(N-1\right)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\text{Var}\left(\widetilde{x}_{ij}'S_{ij}+\widetilde{x}_{ji}'S_{ji}\,\middle|\,x\right)\\ &= \frac{1}{N\left(N-1\right)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Big\{\widetilde{x}_{ij}'\left[\text{Var}\left(S_{ij}\,\middle|\,x_{ij}\right)\right]\widetilde{x}_{ij}+\widetilde{x}_{ji}'\left[\text{Var}\left(S_{ji}\,\middle|\,x_{ji}\right)\right]\widetilde{x}_{ji}\\ &\qquad +\widetilde{x}_{ij}'\left[\text{Cov}\left(S_{ij},S_{ji}\,\middle|\,x_{ij}\right)\right]\widetilde{x}_{ji}+\widetilde{x}_{ji}'\left[\text{Cov}\left(S_{ji},S_{ij}\,\middle|\,x_{ji}\right)\right]\widetilde{x}_{ij}\Big\}.\end{aligned} \tag{13}$$

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 á 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.

3.3 Downward Bias in Robust Standard Errors

Of course, even if the point estimates are correctly centered, inferences will still be unreliable if the estimates of the variance used to construct confidence intervals are not themselves unbiased. For PPML models, confidence intervals are typically obtained using a "sandwich" estimator for the variance that accounts for the possible misspecification of the model. However, as shown by Kauermann and Carroll (2001), the PPML sandwich estimator is generally downward-biased in finite samples. Furthermore, for gravity models (both two-way and three-way), the bias in the sandwich estimator can itself be formalized as a kind of IPP.[18]

To illustrate the bias of the sandwich estimator in our three-way setting, recall that we can express the variance of $\widehat{\beta}$ as $\text{Var}(\widehat{\beta}-\beta) = N^{-1}(N-1)^{-1}W_N^{-1}\Omega_N W_N^{-1}$. As is also true for the linear model (cf., MacKinnon and White, 1985; Imbens and Kolesar, 2016), the bias arises because plugin estimates for Ω_N depend on the estimated variance $\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}') = E[(y_{ij}-\widehat{\lambda}_{ij})(y_{ij}-\widehat{\lambda}_{ij})']$ rather than on the true variance $\mathbb{E}(S_{ij}S_{ij}') = \mathbb{E}[(y_{ijt}-\lambda_{ijt})(y_{ijt}-\lambda_{ijt})']$. Even though $\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}')$ is a consistent estimate for $\mathbb{E}(S_{ij}S_{ij}')$, it will generally be downward-biased in finite samples. Notably, this bias may be especially slow to vanish for models with gravity-like fixed effects.

[18] This type of IPP has similar origins to the one described in Verdier (2018), who considers a dyadic linear model with two-way FEs and sparse matching between the two panel dimensions.

To see this, continue to let $\phi := vec(\alpha, \gamma)$ and now let d_{ij} be a $T \times dim(\phi)$ matrix of dummies such that each row of d_{ij} satisfies $d_{ijt}\phi = \alpha_{it} + \gamma_{jt}$. Using the same approach as Kauermann and Carroll (2001), we can then use the special case where $\mathbb{E}(S_{ij}S_{ij}') = \kappa\bar{H}_{ij}$ (such that $\Omega_N = \kappa W_N$, meaning the model is correctly specified) to demonstrate that $\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}')$ generally has a downward bias. Specifically, let the fitted score vector $\widehat{S}_{ij}$ be approximated by the first-order expansion $\widehat{S}_{ij} = S_{ij} - \bar{H}_{ij}x_{ij}(\widehat{\beta} - \beta) - \bar{H}_{ij}d_{ij}(\widehat{\phi} - \phi)$. Also assume that $\mathbb{E}(S_{ij}S_{ij}') = \kappa\bar{H}_{ij}$, such that the FE-PPML model is correctly specified. Then the expected outer product of the fitted score $\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}')$ has a first-order bias of

$$\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}' - S_{ij}S_{ij}') \approx \; -\tfrac{\kappa}{N(N-1)}\bar{H}_{ij}\widetilde{x}_{ij}W_N^{-1}\widetilde{x}_{ij}'\bar{H}_{ij} - \tfrac{\kappa}{N(N-1)}\bar{H}_{ij}d_{ij}W_N^{(\phi)-1}d_{ij}'\bar{H}_{ij} \tag{14}$$

where $W_N^{(\phi)} := \mathbb{E}_N[-\partial^2\ell_{ij}/\partial\phi\partial\phi'] = -[N(N-1)]^{-1}\sum_{i,j} d_{ij}'\bar{H}_{ij}d_{ij}$ captures the expected Hessian of the concentrated likelihood with respect to ϕ.[19]

The two terms on the right-hand side of (14) are both negative definite, implying that the sandwich estimator is generally downward-biased—and definitively so if the model is correctly specified. The most meaningful difference with the earlier results of Kauermann and Carroll (2001) is how we can use these two terms to decompose the bias in $\mathbb{E}(\widehat{S}_{ij}\widehat{S}_{ij}')$ into two distinct sources. The first term in (14), which depends on $[N(N-1)]^{-1}W_N^{-1}$, captures how the bias depends on the variance of $\widehat{\beta}$. The second term, which depends on $[N(N-1)]^{-1}W_N^{(\phi)-1}$, captures how much of the bias is due to the variance in the estimated incidental parameter vector $\widehat{\phi}$. The former term decreases with N^2, but the latter term only decreases with N, since increasing N by 1 only adds 1 additional observation of each element of $\widehat{\phi}$.[20]

All together, this analysis implies that the estimated standard error for $\widehat{\beta}$ will exhibit a bias that only disappears at the relatively slow rate of $1/\sqrt{N}$. We should therefore be concerned that asymptotic confidence intervals for $\widehat{\beta}$ may exhibit inadequate coverage even in moderately large samples, similar to what has been found for the two-way FE-PPML model in recent simulation studies by Egger and Staub (2015), Jochmans (2016), and Pfaffermayr (2019). Indeed, the bias approximation we have derived in (14) can be readily adapted to the two-way setting or even to more general settings with k-way fixed effects.

[19] A detailed derivation of (14) is provided in the Appendix.

[20] Pfaffermayr (2019) makes a similar point about the order of the bias of the standard errors for the two-way FE-PPML model, albeit using a slightly different analysis.

3.4 Bias Corrections for the Three-way Gravity Model

We now present two methods for correcting the bias in estimates: a jackknife method based on the split-panel jackknife of Dhaene and Jochmans (2015) and an analytical correction based on the expansion shown in Proposition 3. We also provide an analytical correction for the downward bias in standard errors.

Jackknife Bias Correction

The advantage of the jackknife correction is that it does not require explicit estimation of the bias yet still has a simple and powerful applicability. To see this, note first that the asymptotic bias we characterize can be written as

$$\frac{1}{N}B^{\beta} + o_p(N^{-1}),$$

where B^{β} is a combined term that captures any suspected asymptotic bias contributions of order $1/N$. The specific jacknife we will apply for our current purposes is a split-panel jackknife based on Dhaene and Jochmans (2015). As in Dhaene and Jochmans (2015), we want to divide the overall data set into subpanels of roughly even size and then estimate $\widehat{\beta}_{(p)}$ for each subpanel p. Given the gravity structure of the model, we first divide the set of countries into evenly-sized groups a and b. We then consider 4 subpanels of the form "(a, b)", where "(a, b)" denotes a subpanel where exporters from group a are matched with importers from group b. The other three subpanels are (a, a), (b, a), and (b, b). For randomly-generated data, we can define a and b based on their ordering in the data (i.e., $a := i : i \leq N/2$; $b := i : i > N/2$). For actual data, it would be more sensible to draw these subpanels randomly and repeatedly.[21]

The split-panel jackknife estimator for β, $\widetilde{\beta}_N^J$, is then defined as

$$\widetilde{\beta}_N^J := 2\widehat{\beta} - \sum_p \frac{\widehat{\beta}_{(p)}}{4}. \tag{15}$$

This correction works to reduce the bias because, so long as the distribution of y_{ij} and x_{ij} is homogeneous across the different partitions of the data, each $\widehat{\beta}_{(p)}$ has a leading bias term equal to $2B^{\beta}/N$. The average $\widehat{\beta}_{(p)}$ across these four subpanels thus also has

[21] This is just one possible way to construct a jackknife correction for two-way panels. We have also experimented with splitting the panel one dimension at a time as in Fernández-Val and Weidner (2016), but we find the present method performs noticeably better at reducing the bias.

a leading bias of $2B^\beta/N$ and any terms depending on B^β/N cancel out of (15). Thus, the bias-corrected estimate $\widetilde{\beta}_N^J$ only has a bias of order $o_p(N^{-1})$, which is obtained by combining the second-order bias from $\widehat{\beta}$ with that of the average subpanel estimate. This latter bias can be shown to be larger than the original second-order bias in (3.4), but the overall bias should still be smaller because of the elimination of the leading bias term.

Analytical Bias Correction

Our anaytical correction for the bias is based on the bias expression in Proposition 3 and uses the plugin objects $\widehat{\tilde{x}}_{ij}$, $\widehat{S}_{ij}$, $\widehat{H}_{ij}$, $\widehat{\overline{H}}_{ij}$, and $\widehat{G}_{ij}$. For the most part, 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. The resulting bias correction is given by $(N-1)^{-1}\widehat{W}_N^{-1}(\widehat{B}_N + \widehat{D}_N)$, where $\widehat{B}_N$ and $\widehat{D}_N$ are K-vectors with elements given by

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

and where

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

As in Fernández-Val and Weidner (2016), it is possible to show that these plug-in corrections lead to estimates that are asymptotically unbiased as $N \to \infty$.[22] Still, for finite samples, it is evident that the bias in some of these plug-in objects—the $\widehat{S}_{ij}\,\widehat{S}'_{ij}$ outer product terms, for example—could cause the analytical bias correction to itself exhibit

[22]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 $\mathbb{E}\left(S'_{ij}H_{ij}\,|x_{ij,k}\right)$ and $\mathbb{E}\left(S_{ij}\,S'_{ij}|x_{ij,k}\right)$ 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 3.

some bias. For this reason, it is not obvious *a priori* whether the analytical correction will outperform the jackknife at reducing the bias in $\widehat{\beta}$. One clear advantage the analytical correction has over the jackknife is that it does not require the distribution of y_{ij} and x_{ij} to be homogeneous over the different partitions of the data in order to be valid.

Bias-corrected Standard Errors

Under the assumption of clustered errors within pairs, a natural correction for the variance estimate is available based on (14). Specifically, let

$$\widehat{\Omega}^U := \frac{1}{N(N-1)} \sum_{i,j} \widehat{\tilde{x}}_{ij} \left[\mathbf{I}_T - \frac{1}{N(N-1)} \bar{H}_{ij} \widehat{\tilde{x}}_{ij} \widehat{W}_N^{-1} \widehat{\tilde{x}}' - \frac{1}{N(N-1)} \bar{H}_{ij} d_{ij} \widehat{W}_N^{(\phi)-1} d'_{ij} \right]^{-1} \widehat{S}_{ij} \widehat{S}'_{ij} \widehat{\tilde{x}}_{ij},$$

where $\mathbf{I}_T$ is a $T \times T$ identity matrix and $\widehat{W}_N^{(\phi)}$ is a plugin estimate for $W_N^{(\phi)}$. The corrected variance estimate is then given by

$$\widehat{V}^U = \frac{1}{N(N-1)-1} \widehat{W}^{-1} \widehat{\Omega}^U \widehat{W}^{-1}.$$

The logic of this adjusted variance estimate follows directly from Kauermann and Carroll (2001): if the PPML estimator is correctly specified (such that $E(S_{ij} S'_{ij}) = \kappa \bar{H}_{ij}$), then $\widehat{V}^U$ can be shown to eliminate the first-order bias in $\widehat{V}(\widehat{\beta} - \beta^0)$ shown in (14). It is not generally unbiased otherwise, but it is plausible that it should eliminate a significant portion of any downward bias under other variance assumptions as well.

4 Simulation Evidence

For our simulation analysis, we assume the following: (i) the data generating process (DGP) for the dependent variable is of the form $y_{ijt} = \lambda_{ijt} \omega_{ijt}$, where ω_{ijt} is a log-normal disturbance with mean 1 and variance σ^2_{ijt}. (ii) $\beta = 1$. (iii) The model-relevant fixed effects α, γ, and η are each $\sim \mathcal{N}(0, 1/16)$. (iv) $x_{ijt} = x_{ijt-1}/2 + \alpha + \gamma + \nu_{ijt}$, where $\nu_{ijt} \sim \mathcal{N}(0, 1/16)$.[23] (v) Taking our cue from Santos Silva and Tenreyro (2006), we

[23] These assumptions on α, γ, η, x_{ijt}, and ν_{ijt} are taken from Fernández-Val and Weidner (2016). Notice that x_{ijt} is strictly exogenous with respect to ω_{ijt} conditional on α, γ, and η.

consider 4 different assumptions about the residual disturbance ω_{ijt}:

$$\begin{aligned}
&\textbf{DGP I}: \ \sigma^2_{ijt} = \lambda^{-2}_{ijt}; && \text{Var}(y_{ijt}|x_{it},\alpha,\gamma,\eta) = 1. \\
&\textbf{DGP II}: \ \sigma^2_{ijt} = \lambda^{-1}_{ijt}; && \text{Var}(y_{ijt}|x_{it},\alpha,\gamma,\eta) = \lambda_{ijt}. \\
&\textbf{DGP III}: \ \sigma^2_{ijt} = 1; && \text{Var}(y_{ijt}|x_{it},\alpha,\gamma,\eta) = \lambda^2_{ijt}. \\
&\textbf{DGP IV}: \ \sigma^2_{ijt} = 0.5\lambda^{-1}_{ijt} + 0.5e^{2x_{ijt}}; && \text{Var}(y_{ijt}|x_{it},\alpha,\gamma,\eta) = 0.5\lambda_{ijt} + 0.5e^{2x}\lambda^2_{ijt},
\end{aligned}$$

where we also allow for serial correlation within pairs by imposing

$$\text{Cov}[\omega_{ijs},\omega_{ijt}] = \exp\left[0.3^{|s-t|} \times \sqrt{\ln(1+\sigma^2_{ijs})}\sqrt{\ln(1+\sigma^2_{ijt})}\right] - 1,$$

such that the degree of correlation weakens for observations further apart in time.[24]

The relevance of these various assumptions to commonly used error distributions is best described by considering the conditional variance $\text{Var}(y_{ijt}|x_{it},\alpha,\gamma,\eta)$. For example, DGP I assumes that the conditional variance is constant, as in a Gaussian process with i.i.d disturbances. In DGP II, the conditional variance equals the conditional mean, as in a Poisson distribution. DGP III—which we will also refer to as the case of "log-homoscedastic" data—is the unique case highlighted in Santos Silva and Tenreyro (2006) where the assumption that the conditional variance is proportional to the square of the conditional mean leads to a homoscedastic error when the model is estimated in logs using a linear model. Finally, DGP IV provides a "quadratic" error distribution that mixes DGP II and DGP III and also models a more complex dependence between x_{ijt} and the variance of the error term.

Tables 1 and 2 present simulation evidence comparing the uncorrected three-way FE-PPML estimator with results computed using the analytical and jackknife corrections described in Section 3.4. As in the prior simulations, we again compute results for a variety of different panel sizes—in this case for $N = 20, 50, 100$ and $T = 2, 5, 10$.[25] In order to validate our analytical predictions regarding these estimates, we compute the average bias of each estimator, the ratios of the average bias to the average standard error

[24]The 0.3 that appears here serves as a quasi-correlation parameter. Replacing 0.3 with 1 would be analogous to assuming disturbances are perfectly correlated within pairs. Replacing it with 0 removes any serial correlation. Choosing other values for this parameter produces similar results.

[25]Note that the trade literature currently recommends using wide intervals of 4-5 years between time periods so as to allow trade flows time to adjust to changes in trade costs (see Cheng and Wall, 2005.) Thus, for practical purposes, $T = 10$ may be thought of as a relatively "long" panel in this context that might span 40+ years.

and of the average standard error to the standard deviation of the simulated estimates, and the probability that the estimated 95% confidence interval covers the true estimate of $\beta = 1$. In particular, we expect that the bias in $\widehat{\beta}$ should be decreasing in either N or T but should remain large relative to the estimated standard error and induce inadequate coverage for small T. We are also interested in whether the usual cluster-robust standard errors accurately reflect the true dispersion of estimates. Results for DGPs I and II are shown in Table 1, whereas Table 2 shows results for DGPs III and IV.

The results in both tables collectively confirm the presence of bias and the viability of the analytical and jackknife bias corrections. The average bias is generally larger for DGPs I and IV than II and III. As expected, it generally falls with both N and T across all the different DGPs, though only weakly so for DGP III (the log-homoscedastic case), which generally only has a small bias.[26] To use DGP II—the Poisson case, where PPML should otherwise be an optimal estimator—as a representative example, we see that the average bias falls from 3.774% for the smallest sample where $N = 20$, $T = 2$ to a low of 0.234% at the other extreme where $N = 100$, $T = 10$. For DGP IV, the least favorable of these cases, the average bias ranges from -6.544% down to -1.899%. On the whole, these results support our main theoretical findings that β should be consistently estimated even for small T but has an asymptotic bias that depends on the number of countries and on the number of time periods.

Interestingly, while the average bias almost always decreases with T, the ratio of the bias to standard error usually does not, seemingly contrary to the expectations laid out in Remark 2. Evidently, when T is sufficiently small, the rate at which the bias decreases with T may be slower than $1/T$. Researchers should thus be careful to note that the implications of Remark 2 do not necessarily apply to settings with small T or even moderately large T.[27] Instead, it seems reasonable to expect that the bias will generally be non-negligible relative to the standard error except for very large T. Furthermore, the estimated cluster-robust standard errors themselves clearly exhibit a bias in all cases as well. Even when $N = 100$, SE/SD ratios are uniformly below 1; generally they are

[26] Numerically, what we have found is that the two terms that appear in both B and D in Proposition 3 tend to have opposite signs when the DGP is log-homoscedastic. Thus, they tend to mitigate one another, leading to a somewhat muted bias in this case.

[27] We have also simulated the bias for larger values of T beyond $T = 10$. What we find is that the bias decreases somewhat slowly with T for small values of T (consistent with the results in these tables), but does indeed start to decrease with $1/T$ as T becomes increasingly large.

closer to 0.9 or 0.95, and for DGP IV, they are closer to 0.85 or even 0.8. Because of these biases, the simulated FE-PPML coverage ratios are unsurprisingly below the 0.95 we would expect for an unbiased estimator.

Bias corrections to the point estimates do help with addressing some, but not all, of these issues. The jackknife generally performs more reliably than the analytical correction at reducing the average bias when compared across all values of N and T—notice how, for the Poisson case, for example, the average bias left by the jackknife correction is never greater than 0.1%, whereas the analytical-corrected estimates still have average biases ranging between 0.08% and 1.12%. However, when $N = 100$, the analytical correction often dominates, especially when T is at least 5. All the same, both corrections generally have a positive effect, and the better across-the-board bias-reduction performance of the jackknife comes at the important cost of a relatively large increase in the variance. Thus, the analytical correction generally performs as well as or better than the jackknife in terms of improving coverage even in the smaller samples. Neither correction is sufficient to bring coverage ratios to the immediate vicinity of 0.95, however, though corrected Gaussian-DGP estimates and Poisson-DGP estimates both reach 0.93-0.94 using the analytical correction when $N = 100$, and coverage for the analytical-corrected Poisson-DGP estimates reaches 0.94-0.96 when $N = 50$.

Table 3 then evaluates the efficacy of our bias correction for the estimated variance. Keeping in mind that this correction is calibrated for the case of a correctly specified variance (which corresponds to DGP II), it is unsurprising that the effect of this correction varies depending on the conditional distribution of the data. The best results by far are for the Gaussian, Poisson, and Log-homoscedastic DGPs (DGPs I, II, and III, respectively), where combining the analytical bias correction for the point estimates with the correction for the variance yields coverage ratios that fall within an acceptable range between 0.932 and 0.962 when N is either 50 or 100 and are often close to the target value of 0.95 in these cases. These corrections lead to dramatic improvements in coverage for DGP IV as well, but there the remaining biases in both the point estimate and the standard error remain large even for $N = 100$ and $T = 10$.

Overall, these simulations suggest that combining an analytical bias correction for $\widehat{\beta}$ with a further correction for the variance based on (14) should be a reliable way of reducing bias and improving coverage. At the same time, it should be noted that neither offers a complete bias removal. For smaller samples, if reducing bias on average is heavily favored, and if the distribution of y_{ij} and x_{ij} can be reasonably assumed to be homogeneous, then

the split-panel jackknife method might be preferable to the analytical correction method. We should also be careful to point out that the results produced here are based on the particular assumptions we have chosen to generate the data. To determine the practical implications of these corrections, a more meaningful test will be to apply them to estimates produced using real data.

5 Empirical Application

For our empirical application, we estimate the average effects of an FTA using a panel with what would typically be considered a relatively large number of countries. Our trade data is from the BACI database of Gaulier and Zignago (2010), from which we extract data on trade flows between 169 countries for the years 1995, 2000, 2005, 2010, and 2015. Countries are chosen so that the same 169 countries always appear as both exporters and importers in every period; hence, the data readily maps to the setting just described with $N = 169$ and $T = 5$. We combine this trade data with data on FTAs from the NSF-Kellogg database maintained by Scott Baier and Jeff Bergstrand, which we crosscheck against data from the WTO in order to incorporate agreements from more recent years.[28] The specification we estimate is

$$y_{ijt} = \exp[\alpha_{it} + \gamma_{jt} + \eta_{ij} + \beta FTA_{ijt}]\omega_{ijt}, \tag{16}$$

where y_{ijt} is trade flows (measured in current USD), FTA_{ijt} is a 0/1 dummy for whether or not i and j have an FTA at time t, and ω_{ijt} is an error term. As we have noted, estimation of specifications such as (16) via PPML has become an increasingly standard method for estimating the effects of trade agreements and other trade policies and is currently recommended as such by the WTO (see Yotov, Piermartini, Monteiro, and Larch, 2016.)

Table 4 presents results from FE-PPML estimation of (16), including results obtained using our bias corrections. Because biases may vary depending on the specific heteroscedasticity patterns native to each industry, we show results for industry-specific regressions at the 2 digit ISIC (rev. 3) industry level as well as for aggregate trade. The results for aggregate trade flows, shown in the bottom row of Table 4, are nonetheless

[28] This database is available for download on Jeff Bergstrand's website: https://www3.nd.edu/~jbergstr/. The most recent version runs from 1950-2012. The additional data from the WTO is needed to capture agreements that entered into force between 2012 and 2015.

fairly representative. To provide some basic interpretation, the coefficient on FTA_{ijt} for aggregate trade is initially estimated to be 0.082, which equates to an $e^{0.082} - 1 = 8.5\%$ average "partial" effect of an FTA on trade.[29] The estimated standard error is 0.027, implying that this effect is statistically different from zero at the $p < 0.01$ significance level. Our bias-corrected estimates do not paint an altogether different picture, but do highlight the potential for meaningful refinement. Both the analytical and jackknife bias corrections for β suggest a downward bias of 0.04-0.05, or about 15%-18% of the estimated standard error. As our bias-corrected standard errors show (in the last column of Table (4)), the initially estimated standard error itself has an implied downward bias of 10% (i.e., 0.027 versus 0.030).

Turning to the industry-level estimates, the analytical bias correction more often than not indicates a downward bias ranging between 5%-20% of the estimated standard error. Exceptions are present on both sides of this range. Estimates for the Chemical and Furniture industries appear to be unbiased, for example, and some (such as Tobacco) are associated with an upward bias. On the other end of the spectrum, implied downward biases can also be larger than 20% of the standard error, as is seen for Petroleum (46%), Fabricated Metal Products (28%), and Electrical Equipment (26%). The biases implied by the jackknife are often even larger (see Fabricated Machinery Products, for example), consistent with what we found in our simulations for smaller panel sizes. One possible interpretation is that the jackknife-corrected estimates are giving us a less conservative alternative to the analytical corrections in these cases. However, as we have noted, these jackknife estimates could be reflecting non-homogeneity across the different subpanels and/or the higher variance introduced by the jackknife. Implied biases in the standard error, meanwhile, tend to range between 10%-20% of the original standard error, again with some exceptions.

[29] The term "partial effect" is conventionally used to distinguish this type of estimate from the "general equilibrium" effects of an FTA, which would typically be calculated solving a general equilibrium trade model where prices, incomes, and output levels (which are otherwise absorbed by the α_{it} and γ_{jt} fixed effects) are allowed to evolve endogenously in response to the FTA. In the context of such models, β can usually be interpreted as capturing the average effect of an FTA on bilateral trade frictions specifically, holding fixed all other determinants of trade.

6 Conclusion

Thanks to recent methodological and computational advances, nonlinear estimation with three-way fixed effects has become increasingly popular for investigating the effects of trade policies on trade flows. However, the asymptotic and finite-sample properties of such an estimator have not been rigorously studied, especially with regards to potential IPPs. The performance of the FE-PPML estimator in particular is of natural interest in this context, both because FE-PPML is known to be relatively robust to IPPs as well as because it is likely to be a researcher's first choice for estimating three-way gravity models. Our results regarding the consistency of PPML in this setting reflect these unique properties of PPML and support its current status as a workhorse estimator for estimating the effects of trade polices.

Given the consistency of PPML in this setting, and given the nice IPP-robustness properties of PPML in general, it may come as a surprise that three-way PPML nonetheless suffers from an IPP bias. We show that the leading component of this bias is decreasing in the number of countries in the panel as well as in the number of time periods. Thus, the bias is likely to be of comparable magnitude to the standard error when the time dimension of the panel is small, even for large panels with many countries. Typical cluster-robust estimates of the standard error are also biased, implying asymptotic confidence intervals not only off-center but also too narrow.

These issues are not so severe that they leave researchers in the wilderness, but we do recommend taking advantage of the corrective measures described in the paper when estimating three-way gravity models. In particular, we find that analytical bias corrections based on Taylor expansions to both the point estimates and standard errors generally lead to improved inferences when applied simultaneously. These corrections are not a panacea, however, and several avenues remain open for future work. For example, confidence interval estimates could be adjusted further to account for the uncertainty in the estimated variance—Kauermann and Carroll (2001) describe such a correction for the standard PPML model. A quasi-differencing approach similar to Jochmans (2016) could also provide another angle of attack. Turning to broader applications, the essential dyadic structure of our bias corrections could be easily extended to network models that study changes in network behavior over time, especially settings that involve studying the number of interactions between network members.

References

Alvarez, J., and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71(4), 1121–1159.

Andersen, E. B. (1970): "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2), 283–301.

Anderson, J. E., and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93(1), 170–192.

Arellano, M., and S. Bonhomme (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

Arellano, M., and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," *Econometric Society Monographs*, 43, 381.

Bai, J. (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77(4), 1229–1279.

Baier, S. L., and J. H. Bergstrand (2007): "Do Free Trade Agreements actually Increase Members' International Trade?," *Journal of International Economics*, 71(1), 72–95.

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)*.

Carro, J. M. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," *Journal of Econometrics*, 140(2), 503–528.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20(3), S1–S13.

Chen, M., I. Fernández-Val, and M. Weidner (2019): "Nonlinear factor models for network and panel data," *arXiv preprint arXiv:1412.5647*.

Cheng, I. H., and H. J. Wall (2005): "Controlling for Heterogeneity in gravity models of trade," *Federal Reserve Bank of St. Louis Review*, 87(1), 49–63.

Correia, S., P. Guimarães, and T. Zylkin (2019): "PPMLHDFE: Fast Poisson Estimation with High-dimensional Fixed Effects," *Unpublished manuscript*.

Dhaene, G., and K. Jochmans (2015): "Split-panel Jackknife Estimation of Fixed-effect Models," *The Review of Economic Studies*, 82(3), 991–1030.

Dzemski, A. (2018): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*.

Egger, P., M. Larch, K. E. Staub, and R. Winkelmann (2011): "The Trade Effects of Endogenous Preferential Trade Agreements," *American Economic Journal: Economic Policy*, 3(3), 113–143.

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 F. Vella (2011): "Bias Corrections for Two-step Fixed Effects Panel Data Estimators," *Journal of Econometrics*, 163(2), 144–162.

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*.

Gaulier, G., and S. Zignago (2010): "BACI: International Trade Database at the Product-Level. The 1994-2007 Version," Working Paper 2010-23, CEPII research center.

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 H. R. Moon (2006): "Reducing Bias of MLE in a Dynamic Panel Model," *Econometric Theory*, 22(3), 499–512.

Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141(2), 597–620.

Hausman, J., B. H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52(4), 909–938.

Head, K., and T. Mayer (2014): "Gravity equations: Workhorse, Toolkit, and Cookbook," *Handbook of International Economics*, 4, 131–196.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," *The Quarterly Journal of Economics*, 123(2), 441–487.

Hinz, J., A. Stammann, and J. Wanner (2019): "Persistent Zeros: The Extensive Margin of Trade," *Unpublished manuscript*.

Imbens, G. W., and M. Kolesar (2016): "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, 98(4), 701–712.

Jochmans, K. (2016): "Two-Way Models for Gravity," *Review of Economics and Statistics*.

Jochmans, K., and M. Weidner (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87(5), 1543–1560.

Kato, K., A. F. Galvao Jr., and G. V. Montes-Rojas (2012): "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics*, 170(1), 76–91.

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.

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *The Review of Economic Studies*, 69(3), 647–666.

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.

MacKinnon, J. G., and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of econometrics*, 29(3), 305–325.

Moon, H. R., and M. Weidner (2017): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," *Econometric Theory*, 33(1), 158–195.

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.

Palmgren, J. (1981): "The Fisher Information Matrix for Log Linear Models Arguing Conditionally on Observed Explanatory Variables," *Biometrika*, 68(2), 563–566.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

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

Phillips, P. C. B., and H. R. Moon (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67(5), 1057–1111.

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

Stammann, A. (2018): "Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects," *arXiv preprint arXiv:1707.01815*.

Verdier, V. (2018): "Estimation and inference for linear models with two-way fixed effects and sparsely matched data," *Review of Economics and Statistics*, (forthcoming).

Wooldridge, J. M. (1999): "Distribution-free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics*, 90(1), 77–97.

Woutersen, T. (2002): "Robustness against Incidental Parameters," Discussion paper, University of Western Ontario, Department of Economics.

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*.

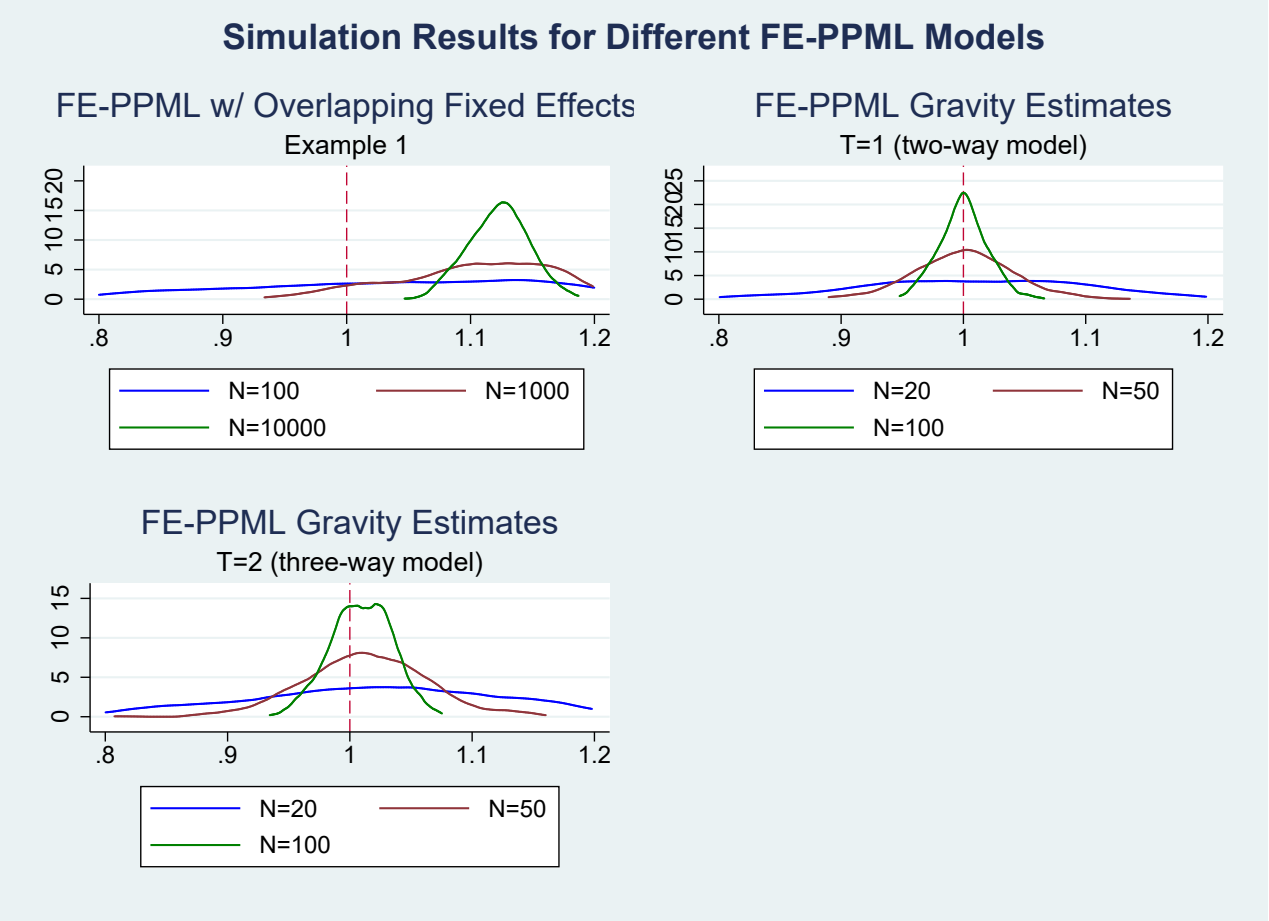

Figure 1: Kernel density plots of FE PPML estimates for 3 different models, using 500 replications. Clockwise from top left, the 3 models are: $y_{it} = \exp[\alpha_i \times 1(t \leq 2) + \gamma_i \times 1(t \geq 2) + x_{it}\beta]\omega_{it}$, with the t dimension of the panel fixed at $T = 3$; a two-way gravity model with $y_{ij} = \exp[\alpha_i + \gamma_j + x_{ij}\beta]\omega_{ij}$; a three-way gravity model with $y_{ijt} = \exp[\alpha_{it} + \gamma_{jt} + \eta_{ij} + x_{ijt}\beta]\omega_{ijt}$ and $T = 2$. The i and j dimensions of the panel both have size N in the latter two models. The true value of β is 1 (indicated by the vertical dotted lines) and the data is generated using $\text{Var}(y|\cdot) = \mathbb{E}(y|\cdot)$. See text for further details.

Table 1: Finite-sample Properties of the Three-way FE-PPML Gravity Model

	N=20			N=50			N=100		
	T=2	T=5	T=10	T=2	T=5	T=10	T=2	T=5	T=10
A. Gaussian DGP ("DGP I")									
Average bias (×100)									
FE-PPML	6.588	3.645	2.297	2.659	1.446	0.904	1.376	0.702	0.410
Analytical	2.193	0.896	0.464	0.326	0.120	0.101	0.071	-0.009	-0.006
Jackknife	0.233	-0.043	-0.334	0.031	-0.038	-0.048	-0.012	-0.058	-0.043
Bias / SE ratio									
FE-PPML	0.683	0.778	0.736	0.620	0.709	0.684	0.606	0.659	0.601
Analytical	0.227	0.191	0.149	0.076	0.059	0.076	0.031	-0.008	-0.009
Jackknife	0.024	-0.009	-0.107	0.007	-0.019	-0.036	-0.005	-0.054	-0.063
SE / SD ratio									
FE-PPML	0.837	0.826	0.883	0.926	0.936	0.965	0.932	0.921	0.943
Analytical	0.795	0.805	0.863	0.882	0.905	0.954	0.899	0.901	0.936
Jackknife	0.716	0.749	0.802	0.844	0.870	0.923	0.892	0.907	0.937
Coverage probability (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.828	0.798	0.818	0.878	0.856	0.884	0.888	0.870	0.884
Analytical	0.864	0.876	0.904	0.916	0.928	0.928	0.926	0.944	0.938
Jackknife	0.836	0.858	0.890	0.910	0.912	0.928	0.924	0.936	0.944
B. Poisson DGP ("DGP II")									
Average bias (×100)									
FE-PPML	3.774	2.017	1.314	1.517	0.839	0.563	0.721	0.395	0.234
Analytical	1.121	0.459	0.355	0.143	0.079	0.117	-0.038	-0.015	-0.004
Jackknife	-0.090	-0.096	-0.165	-0.012	-0.010	0.024	-0.093	-0.050	-0.031
Bias / SE ratio									
FE-PPML	0.683	0.778	0.736	0.620	0.709	0.684	0.606	0.659	0.601
Analytical	0.227	0.191	0.149	0.076	0.059	0.076	0.031	-0.008	-0.009
Jackknife	0.024	-0.009	-0.107	0.007	-0.019	-0.036	-0.005	-0.054	-0.063
SE / SD ratio									
FE-PPML	0.875	0.828	0.918	0.959	0.977	0.988	0.962	0.950	0.930
Analytical	0.835	0.806	0.899	0.931	0.959	0.981	0.946	0.944	0.925
Jackknife	0.749	0.750	0.830	0.894	0.920	0.954	0.942	0.938	0.921
Coverage probability (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.884	0.870	0.902	0.928	0.908	0.918	0.920	0.928	0.904
Analytical	0.884	0.880	0.922	0.940	0.940	0.956	0.936	0.944	0.930
Jackknife	0.856	0.868	0.892	0.922	0.926	0.944	0.934	0.936	0.932

Notes: Results computed using 500 replications. The model being estimated is $y_{ijt} = \lambda_{ijt}\omega_{ijt}$, where $\lambda_{ijt} = \exp(\alpha_{it} + \gamma_{jt} + \eta_{ij} + \beta x_{ijt})$. The data is generated using $\alpha_{it} \sim \mathcal{N}(0, 1/16)$, $\gamma_{jt} \sim \mathcal{N}(0, 1/16)$, $\eta_{ij} \sim \mathcal{N}(0, 1/16)$ and $\beta = 1$. $x_{ijt} = x_{ijt-1}/2 + \alpha_{it} + \gamma_{jt} + \eta_{ij} + \nu_{ijt}$, with $x_{ij0} = \eta_{ij} + \nu_{ij0}$ and $\nu_{ijt} \sim \mathcal{N}(0, 1/2)$. Results are shown for two different assumptions about $\text{Var}(y_{ijt})$. The "Gaussian" DGP (panel A) assumes $\text{Var}(\omega_{ijt}) = \lambda_{ijt}^{-2}$. The "Poisson" DGP (panel B) assumes $\text{Var}(\omega_{ijt}) = \lambda_{ijt}^{-1}$. SE/SD refers to the ratio of the average standard error of of $\widehat{\beta}$ relative to the standard deviation of $\widehat{\beta}$ across simulations. Coverage probability refers to the probability β^0 is covered in the 95% confidence interval for $\widehat{\beta}_{NT}$. "Analytical" and "Jackknife" respectively indicate Analytical and Jackknife bias-corrected FE-PPML estimates. "FE-PPML" indicates uncorrected estimates. SEs allow for within-ij clustering.

Table 2: Finite-sample Properties of the Three-way FE-PPML Gravity Model

	N=20			N=50			N=100		
	T=2	T=5	T=10	T=2	T=5	T=10	T=2	T=5	T=10
A. Log-homoscedastic DGP ("DGP III")									
Average bias (×100)									
FE-PPML	0.223	-0.291	-0.292	0.161	-0.070	-0.048	-0.033	-0.083	-0.103
Analytical	-0.264	-0.356	-0.126	-0.022	-0.049	0.068	-0.126	-0.057	-0.033
Jackknife	-0.580	-0.440	-0.320	0.016	-0.046	0.046	-0.146	-0.076	-0.044
Bias / SE ratio									
FE-PPML	0.022	-0.058	-0.087	0.036	-0.032	-0.032	-0.014	-0.072	-0.133
Analytical	-0.026	-0.071	-0.038	-0.005	-0.022	0.046	-0.054	-0.049	-0.043
Jackknife	-0.057	-0.088	-0.095	0.004	-0.021	0.031	-0.063	-0.066	-0.057
SE / SD ratio									
FE-PPML	0.869	0.797	0.887	0.940	0.953	0.954	0.962	0.934	0.897
Analytical	0.816	0.756	0.837	0.902	0.915	0.920	0.936	0.913	0.872
Jackknife	0.731	0.705	0.778	0.870	0.881	0.903	0.935	0.895	0.862
Coverage probability (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.902	0.886	0.920	0.948	0.934	0.938	0.942	0.934	0.926
Analytical	0.880	0.864	0.912	0.932	0.924	0.932	0.938	0.932	0.912
Jackknife	0.838	0.838	0.878	0.930	0.908	0.914	0.940	0.916	0.924
B. Quadratic DGP ("DGP IV")									
Average bias (×100)									
FE-PPML	-6.544	-6.024	-5.210	-3.341	-3.275	-2.798	-2.305	-2.144	-1.899
Analytical	-5.051	-4.412	-3.586	-1.810	-1.831	-1.448	-1.110	-1.025	-0.883
Jackknife	-4.441	-3.836	-3.228	-1.429	-1.574	-1.249	-1.042	-0.961	-0.812
Bias / SE ratio									
FE-PPML	-0.544	-0.960	-1.203	-0.597	-1.089	-1.319	-0.750	-1.260	-1.562
Analytical	-0.420	-0.703	-0.828	-0.324	-0.609	-0.682	-0.361	-0.603	-0.726
Jackknife	-0.369	-0.612	-0.746	-0.256	-0.523	-0.589	-0.339	-0.565	-0.668
SE / SD ratio									
FE-PPML	0.817	0.734	0.787	0.855	0.845	0.842	0.898	0.845	0.805
Analytical	0.753	0.676	0.715	0.788	0.771	0.768	0.830	0.780	0.739
Jackknife	0.670	0.621	0.665	0.751	0.735	0.743	0.823	0.758	0.720
Coverage probability (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.860	0.750	0.732	0.852	0.756	0.694	0.868	0.698	0.588
Analytical	0.834	0.770	0.786	0.880	0.808	0.810	0.888	0.818	0.782
Jackknife	0.784	0.742	0.744	0.850	0.816	0.820	0.894	0.820	0.770

Notes: Results computed using 500 replications. The model being estimated is $y_{ijt} = \lambda_{ijt}\omega_{ijt}$, where $\lambda_{ijt} = \exp(\alpha_{it} + \gamma_{jt} + \eta_{ij} + \beta x_{ijt})$. The data is generated using $\alpha_{it} \sim \mathcal{N}(0, 1/16)$, $\gamma_{jt} \sim \mathcal{N}(0, 1/16)$, $\eta_{ij} \sim \mathcal{N}(0, 1/16)$ and $\beta = 1$. $x_{ijt} = x_{ijt-1}/2 + \alpha_{it} + \gamma_{jt} + \eta_{ij} + \nu_{ijt}$, with $x_{ij0} = \eta_{ij} + \nu_{ij0}$ and $\nu_{ijt} \sim \mathcal{N}(0, 1/2)$. Results are shown for two different assumptions about $\text{Var}(y_{ijt})$. The "Log-homoscedastic" DGP (panel A) assumes $\text{Var}(\omega_{ijt}) = 1$. The "Quadratic" DGP (Panel B) assumes ω_{ijt} is log-normal with variance equal to $0.5\lambda_{ijt}^{-1} + 0.5\exp(2x_{ijt})$. SE/SD refers to the ratio of the average standard error of of $\widehat{\beta}$ relative to the standard deviation of $\widehat{\beta}$ across simulations. Coverage probability refers to the probability β^0 is covered in the 95% confidence interval for $\widehat{\beta}_{NT}$. "Analytical" and "Jackknife" respectively indicate Analytical and Jackknife bias-corrected FE-PPML estimates. "FE-PPML" indicates uncorrected estimates. SEs allow for within-ij clustering.

Table 3: Improving Coverage in the Three-way FE-PPML Gravity Model

	N=20			N=50			N=100		
	T=2	T=5	T=10	T=2	T=5	T=10	T=2	T=5	T=10
A. Gaussian DGP ("DGP I")									
SE / SD ratio with corrected SEs									
FE-PPML	0.938	0.907	0.962	0.980	0.978	1.001	0.963	0.944	0.961
Analytical	0.891	0.884	0.940	0.933	0.946	0.990	0.928	0.923	0.954
Jackknife	0.803	0.823	0.873	0.894	0.909	0.957	0.921	0.929	0.955
Coverage probability with corrected SEs (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.864	0.830	0.848	0.900	0.876	0.888	0.896	0.876	0.890
Analytical	0.912	0.906	0.918	0.936	0.934	0.950	0.934	0.948	0.948
Jackknife	0.880	0.904	0.908	0.922	0.924	0.936	0.936	0.942	0.952
B. Poisson DGP ("DGP II")									
SE / SD ratio with corrected SEs									
FE-PPML	0.977	0.910	1.003	1.009	1.018	1.025	0.988	0.971	0.949
Analytical	0.933	0.886	0.982	0.979	1.000	1.019	0.971	0.965	0.943
Jackknife	0.836	0.825	0.907	0.940	0.959	0.991	0.968	0.958	0.939
Coverage probability with corrected SEs (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.922	0.892	0.930	0.946	0.922	0.924	0.926	0.940	0.908
Analytical	0.916	0.918	0.938	0.956	0.952	0.962	0.940	0.946	0.934
Jackknife	0.896	0.904	0.916	0.936	0.940	0.946	0.948	0.936	0.934
C. Log-homoscedastic DGP ("DGP III")									
SE / SD ratio with corrected SEs									
FE-PPML	0.984	0.896	0.998	1.001	1.013	1.012	0.998	0.967	0.928
Analytical	0.924	0.850	0.940	0.960	0.972	0.977	0.970	0.945	0.902
Jackknife	0.827	0.793	0.875	0.926	0.936	0.959	0.969	0.927	0.891
Coverage probability with corrected SEs (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.946	0.926	0.946	0.954	0.952	0.956	0.948	0.942	0.938
Analytical	0.920	0.908	0.938	0.948	0.942	0.946	0.944	0.934	0.932
Jackknife	0.896	0.878	0.914	0.950	0.930	0.942	0.948	0.932	0.926
D. Quadratic DGP ("DGP IV")									
SE / SD ratio with corrected SEs									
FE-PPML	0.952	0.861	0.929	0.947	0.948	0.949	0.970	0.924	0.882
Analytical	0.877	0.793	0.845	0.873	0.865	0.865	0.897	0.853	0.810
Jackknife	0.781	0.729	0.786	0.833	0.824	0.837	0.889	0.828	0.789
Coverage probability with corrected SEs (should be 0.95 *for an unbiased estimator)*									
FE-PPML	0.900	0.820	0.808	0.894	0.806	0.762	0.892	0.752	0.652
Analytical	0.894	0.830	0.838	0.908	0.864	0.862	0.918	0.856	0.818
Jackknife	0.844	0.808	0.814	0.894	0.860	0.866	0.908	0.848	0.822

Notes: Results computed using 500 replications. The model being estimated is $y_{ijt} = \lambda_{ijt}\omega_{ijt}$, where $\lambda_{ijt} = \exp(\alpha_{it} + \gamma_{jt} + \eta_{ij} + \beta x_{ijt})$. The data is generated using $\alpha_{it} \sim \mathcal{N}(0, 1/16)$, $\gamma_{jt} \sim \mathcal{N}(0, 1/16)$, $\eta_{ij} \sim \mathcal{N}(0, 1/16)$ and $\beta = 1$. $x_{ijt} = x_{ijt-1}/2 + \alpha_{it} + \gamma_{jt} + \eta_{ij} + \nu_{ijt}$, with $x_{ij0} = \eta_{ij} + \nu_{ij0}$ and $\nu_{ijt} \sim \mathcal{N}(0, 1/2)$. Results are shown for four different assumptions about ω_{ijt}. The "Gaussian" DGP (panel A) assumes $\text{Var}(\omega_{ijt}) = \lambda_{ijt}^{-2}$. The "Poisson" DGP (panel B) assumes $\text{Var}(\omega_{ijt}) = \lambda_{ijt}^{-1}$. The "Log-homoscedastic" DGP (panel C) assumes $\text{Var}(\omega_{ijt}) = 1$. The "Quadratic" DGP (Panel D) assumes ω_{ijt} is log-normal with variance equal to $0.5\lambda_{ijt}^{-1} + 0.5\exp(2x_{ijt})$. SE/SD refers to the ratio of the average standard error of of $\widehat{\beta}$ relative to the standard deviation of $\widehat{\beta}$ across simulations. Coverage probability refers to the probability β^0 is covered in the 95% confidence interval for $\widehat{\beta}$. "Analytical" and "Jackknife" respectively indicate Analytical and Jackknife bias-corrected FE-PPML estimates. "FE-PPML" indicates uncorrected estimates. SEs allow for within-ij clustering. The corrected SEs correct for first-order finite sample bias in the estimated variance.

Table 4: Bias Correction Results Using BACI Trade Data ($N = 169$)

		Original estimates		Bias-corrected estimates		
Industry	Code	$\widehat{\beta}$	SE	Analytical	Jackknife	SE
Agriculture	1	0.100	(0.046)	0.110	0.115	(0.051)
Forestry	2	-0.205	(0.125)	-0.199	-0.189	(0.155)
Fishing	5	0.128	(0.141)	0.140	0.182	(0.164)
Coal	10	0.025	(0.131)	-0.039	-0.063	(0.164)
Metal Ores	13	0.040	(0.100)	0.033	-0.025	(0.123)
Other Mining & Quarrying n.e.c.	14	0.048	(0.096)	0.079	0.097	(0.107)
Food & Beverages	15	0.019	(0.043)	0.026	0.031	(0.048)
Tobacco	16	0.535	(0.139)	0.525	0.571	(0.162)
Textiles	17	0.228	(0.045)	0.226	0.234	(0.055)
Apparel	18	0.092	(0.092)	0.094	0.127	(0.122)
Leather Products	19	0.224	(0.067)	0.220	0.240	(0.079)
Wood & Cork Products	20	0.078	(0.109)	0.098	0.101	(0.127)
Paper & Paper Products	21	-0.002	(0.062)	-0.004	-0.018	(0.071)
Printed & Recorded Media	22	-0.115	(0.065)	-0.144	-0.180	(0.076)
Coke & Refined Petroleum	23	0.256	(0.076)	0.291	0.319	(0.090)
Chemicals & Chemical Products	24	0.073	(0.035)	0.073	0.078	(0.040)
Rubber & Plastic Products	25	0.141	(0.030)	0.146	0.157	(0.035)
Non-metallic Mineral Products	26	0.217	(0.049)	0.223	0.225	(0.058)
Basic Metal Products	27	0.268	(0.100)	0.273	0.301	(0.115)
Fabricated Metal Products (excl. Machinery)	28	0.196	(0.036)	0.206	0.225	(0.041)
Machinery & Equipment n.e.c.	29	0.049	(0.035)	0.052	0.056	(0.041)
Office, Accounting, and Computer Equipment	30	-0.036	(0.062)	-0.044	-0.045	(0.074)
Electrical Equipment	31	0.213	(0.045)	0.225	0.240	(0.052)
Communications Equipment	32	-0.127	(0.067)	-0.143	-0.173	(0.081)
Medical & Scientific Equipment	33	0.063	(0.039)	0.069	0.082	(0.044)
Motor Vehicles, Trailers & Semi-trailers	34	0.157	(0.064)	0.169	0.194	(0.077)
Other Transport Equipment	35	0.208	(0.124)	0.231	0.267	(0.137)
Furniture & Other Manufacturing n.e.c.	36	0.224	(0.073)	0.224	0.227	(0.082)
Total	All	0.082	(0.027)	0.086	0.087	(0.030)

Notes: These results are computed using ISIC Rev. 3 industry-level trade data for trade between 169 countries during years 1995, 2000, 2005, 2010, & 2015. The original data is from BACI. The model being estimated is $y_{ijt} = \exp(\alpha_{it} + \gamma_{jt} + \eta_{ij} + \beta FTA_{ijt})\omega_{ijt}$, where y_{ijt} is the trade volume and FTA_{ijt} is a dummy for the presence of an FTA. α_{it}, γ_{jt}, & η_{ij} respectively denote exporter-time, importer-time, & exporter-importer fixed effects. We estimate each industry separately. The jackknife corrections use the average of 200 randomly-assigned split-panel partitions. SEs are clustered by exporter-importer.

A Appendix with proofs

In what follows, we find it convenient to first provide a proof of 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 Remarks 1 and 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. We also include more details behind the downward bias in the estimated variance.

A.1 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, 2018, 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, $k = 1, \ldots, K$,

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

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

$$\mathcal{D}_{\beta\alpha_i^q}\ell_{ij} = \partial_{\beta\alpha_i^q}\ell_{ij} - \partial_{\alpha_i^{q+1}}\ell_{ij}\,\Xi_{ij}, \qquad \mathcal{D}_{\beta\gamma_j^q}\ell_{ij} = \partial_{\beta\gamma_j^q}\ell_{ij} - \partial_{\gamma_j^{q+1}}\ell_{ij}\,\Xi_{ij}. \tag{17}$$

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.)

(c) 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, $\mathbb{E}\left[\partial_\beta \ell_{ij}(\beta^0,\ \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\right] = 0$ and $\mathbb{E}\left[\partial_{\alpha_i} \ell_{ij}(\beta^0,\ \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\right] = 0$ and $\mathbb{E}\left[\partial_{\gamma_j} \ell_{ij}(\beta^0,\ \alpha_i^0 + \gamma_j^0) \mid X_{ij}, \alpha_i^0, \gamma_j^0, \beta^0\right] = 0$. 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:

$$N\left(\widehat{\beta}-\beta^{0}\right) \;\rightarrow_{d}\; \overline{W}^{-1}\mathcal{N}(\overline{B}+\overline{D},\,\overline{\Omega}), \tag{18}$$

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

$$\begin{aligned}
\overline{B}_1 &= \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{i,j}\frac{\mathbb{E}\left(\partial_{\alpha_i}\ell_{ij}\mathcal{D}_{\beta\alpha_i}\ell_{ij}\right)}{\sum_{j'}\mathbb{E}\left(\partial_{\alpha_i{}^2}\ell_{ij'}\right)}\right],\\
\overline{B}_2 &= \lim_{N\to\infty}\left[\frac{1}{2}\frac{1}{N}\sum_{i}\frac{\left[\sum_j\mathbb{E}(\partial_{\alpha_i}\ell_{ij})^2\right]\sum_j\mathbb{E}(\mathcal{D}_{\beta\alpha_i{}^2}\ell_{ij})}{\left[\sum_j\mathbb{E}\left(\partial_{\alpha_i{}^2}\ell_{ij}\right)\right]^2}\right],\\
\overline{D}_1 &= \lim_{N\to\infty}\left[-\frac{1}{N}\sum_{j}\frac{\sum_i\mathbb{E}\left[\partial_{\gamma_j}\ell_{ij}\mathcal{D}_{\beta\gamma_j}\ell_{ij}\right]}{\sum_i\mathbb{E}\left(\partial_{\gamma_j{}^2}\ell_{ij}\right)}\right],\\
\overline{D}_2 &= \lim_{N\to\infty}\left[\frac{1}{2}\frac{1}{N}\sum_{j}\frac{\sum_i\left[\mathbb{E}(\partial_{\gamma_j}\ell_{ij})^2\right]\sum_i\mathbb{E}(\mathcal{D}_{\beta\gamma_j{}^2}\ell_{ij})}{\left[\sum_i\mathbb{E}\left(\partial_{\gamma_j{}^2}\ell_{ij}\right)\right]^2}\right],\\
\overline{\Omega} &= \lim_{N\to\infty}\left[\frac{1}{N^2}\sum_{i,j}\mathbb{E}\left[\mathcal{D}_\beta\ell_{ij}(\mathcal{D}_\beta\ell_{ij})'\right]\right].
\end{aligned} \tag{19}$$

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 (18) and (19), 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* (4) *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 *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 (7). 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 (6) 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.[30] 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 therfore 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 (17) to the case of vector-value 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

$$\ell_{ij}^*(\beta, \alpha_i, \gamma_j) := \ell_{ij}(\beta, \pi_{ij} - \beta'(\alpha_i^x + \gamma_j^x)) = \ell_{ij}(\beta, \alpha_i - \beta' \alpha_i^x, \gamma_j - \beta' \gamma_j^x) \tag{20}$$

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

[30] 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}.

of α_i^x and γ_j^x by (10) in the main text exactly corresponds to those 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 (19). Thus, the asymptotic variance in (18) 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 (6), 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

$$\bar{\mathcal{H}}\left(\mathbb{I}_{2N} \otimes \iota_T\right) = 0. \tag{21}$$

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.[31] There is exactly one such transformation

[31] 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.

for every $t \in \{1, \dots, T\}$, namely the likelihood is invariant under $\alpha_{it} \mapsto \alpha_{it} + c_t$ and $\gamma_{jt} \mapsto \gamma_{jt} - c_t$ 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

$$v := (\iota_N', -\iota_N')' \otimes M_{\iota_T}, \tag{22}$$

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 (21), 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 (21) 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

$$\bar{\mathcal{H}} v = 0. \tag{23}$$

Equations (21) and (23) describe all the zero-eigenvectors of $\bar{\mathcal{H}}$. The projector onto the null-space of $\bar{\mathcal{H}}$ is therefore given by

$$P_{\text{null}} := \mathbb{I}_{2N} \otimes P_{\iota_T} + P_v, \tag{24}$$

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

$$\bar{\mathcal{H}}\,\bar{\mathcal{H}}^{\dagger} = \bar{\mathcal{H}}^{\dagger}\,\bar{\mathcal{H}} = \mathbb{I}_{2NT} - P_{\text{null}} = M_{(\iota_N', -\iota_N')'} \otimes M_{\iota_T}, \tag{25}$$

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

$$(\mathbb{I}_{2N} \otimes P_{\iota_T})\,\bar{\mathcal{H}}^{\dagger} = 0. \tag{26}$$

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

$$\mathfrak{D}^{\dagger}\,\mathfrak{D} = \mathfrak{D}\,\mathfrak{D}^{\dagger} = \mathbb{I}_{2N} \otimes M_{\iota_T} = \mathbb{I}_{2NT} - \mathbb{I}_{2N} \otimes P_{\iota_T}, \tag{27}$$

where $P_{\iota_T} := T^{-1}\iota_T\iota_T'$. Now, using (25) 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 (27) and (26), and bringing $\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger$ to the rhs gives

$$\bar{\mathcal{H}}^\dagger = \mathfrak{D}^\dagger - P_{\text{null}}\mathfrak{D}^\dagger - \bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger. \tag{28}$$

By transposing this last equation we obtain

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

and now plugging (28) into the rhs of (29) 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

$$\bar{\mathcal{H}}^\dagger - \mathfrak{D}^\dagger = -\mathfrak{D}^\dagger \mathcal{G}\mathfrak{D}^\dagger - \mathfrak{D}^\dagger P_v - P_v \mathcal{D}^\dagger + \mathfrak{D}^\dagger \mathcal{G}\bar{\mathcal{H}}^\dagger \mathcal{G}\mathfrak{D}^\dagger, \tag{30}$$

with $2NT \times T$ matrix v defined in (22). 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 (30) 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 (7) 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

$$\bar{H}_{ij} \geq b\, M_{\iota_T}, \tag{31}$$

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 (31). 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

$$\left\|\bar{\mathcal{H}}^{\dagger}\right\| \leq \frac{1}{b\,(N-2)} = O(1/N). \tag{32}$$

Final bound on $\left\|\bar{\mathcal{H}}^{\dagger} - \mathfrak{D}^{\dagger}\right\|_{\max}$: Using (31) 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

$$\left\|\mathfrak{D}^{\dagger}\right\|_{\max} = O_P(1/N), \qquad \|\mathcal{G}\|_{\max} = O_P(1). \tag{33}$$

The definition of the $2NT \times T$ matrix v in (22) implies that

$$\begin{aligned}\|P_v\|_{\max} &= \left\|P_{(\iota_N',-\iota_N')'}\otimes M_{\iota_T}\right\|_{\max} \leq \left\|P_{(\iota_N',-\iota_N')'}\right\|_{\max} = (2N)^{-1}\left\|(\iota_N',-\iota_N')'(\iota_N',-\iota_N')\right\|_{\max} \\ &= (2N)^{-1} = O(1/N),\end{aligned} \tag{34}$$

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

$$\begin{aligned}\left\|\mathcal{G}\bar{\mathcal{H}}^{\dagger}\mathcal{G}\right\|_{\max} &= \max_{k,\ell\in\{1,\ldots,2NT\}}\left|e_k'\,\mathcal{G}\bar{\mathcal{H}}^{\dagger}\mathcal{G}e_\ell\right| \\ &\leq \left(\max_{k\in\{1,\ldots,2NT\}}\|\mathcal{G}e_k\|\right)\left\|\bar{\mathcal{H}}^{\dagger}\right\|\left(\max_{\ell\in\{1,\ldots,2NT\}}\|\mathcal{G}e_\ell\|\right) \\ &= \left(\max_{k\in\{1,\ldots,2NT\}}\|\mathcal{G}e_k\|\right)^2\left\|\bar{\mathcal{H}}^{\dagger}\right\| \\ &\leq \left(\sqrt{2NT}\,\|\mathcal{G}\|_{\max}\right)^2\left\|\bar{\mathcal{H}}^{\dagger}\right\| \\ &= O_P(1),\end{aligned} \tag{35}$$

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 (32).

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 (30), and also using (33) and (34) and (35), 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[32]

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

with P_{null} defined in (24). 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

[32]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 (36).

$\mathcal{L}(\beta,\phi)$ 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

$$\sqrt{N(N-1)}(\widehat{\beta} - \beta^0) = W_N^{-1}U_N + o_P(1), \tag{36}$$

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

$$\begin{aligned} U_N^{(0)} &= \frac{1}{\sqrt{N(N-1)}}\left[\partial_{\beta}\mathcal{L} + [\partial_{\beta\phi'}\bar{\mathcal{L}}]\,\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}\right] = \frac{1}{\sqrt{N(N-1)}}\,\partial_{\beta}\mathcal{L}^* \\ &= \frac{1}{\sqrt{N(N-1)}}\sum_{i=1}^{N}\sum_{j\in\mathfrak{N}\setminus\{i\}}\partial_{\beta}\ell^*_{ij}, \\ \sqrt{N(N-1)}U_N^{(1)} &= [\partial_{\beta\phi'}\mathcal{L} - \partial_{\beta\phi'}\bar{\mathcal{L}}]\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} - [\partial_{\beta\phi'}\bar{\mathcal{L}}]\,\bar{\mathcal{H}}^{\dagger}\,\left[\mathcal{H} - \bar{\mathcal{H}}\right]\,\bar{\mathcal{H}}^{\dagger}\,\partial_{\phi}\mathcal{L} \\ &\qquad + \frac{1}{2}\sum_{g=1}^{\dim\phi}\left(\partial_{\beta\phi'\phi_g}\bar{\mathcal{L}} + [\partial_{\beta\phi'}\bar{\mathcal{L}}]\,\bar{\mathcal{H}}^{\dagger}[\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}]\right)[\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}]_g\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} \\ &= [\partial_{\beta\phi'}\mathcal{L}^* - \partial_{\beta\phi'}\bar{\mathcal{L}}^*]\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L} + \frac{1}{2}\sum_{g=1}^{\dim\phi}\partial_{\beta\phi'\phi_g}\bar{\mathcal{L}}^*[\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}]_g\bar{\mathcal{H}}^{\dagger}\partial_{\phi}\mathcal{L}. \end{aligned}$$

Here, ℓ^*_{ij} was defined in (20), 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 (6) 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

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

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

$$U_N^{(1)} \;\to_p\; B_N + D_N, \tag{37}$$

with B_N and D_N as given in the proposition. The proof of (37) 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 (19) 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

$$\alpha_i = A_i \, \widetilde{\alpha}_i, \qquad\qquad \gamma_j = C_j \, \widetilde{\gamma}_j, \tag{38}$$

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

$$A_i A_i' = \left(\sum_j \bar{H}_{ij} \right)^\dagger, \qquad C_j C_j' = \left(\sum_i \bar{H}_{ij} \right)^\dagger. \tag{39}$$

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

$$\begin{aligned} \frac{\partial^2 \widetilde{\mathcal{L}}(\beta^0, \widetilde{\alpha}^0, \widetilde{\gamma}^0)}{(\partial \widetilde{\alpha}_i)(\partial \widetilde{\alpha}_i)'} &= A_i' \left(\sum_j \bar{H}_{ij} \right) A_i = \mathbb{I}_{T-1}, \\ \frac{\partial^2 \widetilde{\mathcal{L}}(\beta^0, \widetilde{\alpha}^0, \widetilde{\gamma}^0)}{(\partial \widetilde{\gamma}_j)(\partial \widetilde{\gamma}_j)'} &= C_j' \left(\sum_i \bar{H}_{ij} \right) C_j = \mathbb{I}_{T-1}. \end{aligned} \tag{40}$$

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 (19). 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 (40), 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 (39) 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 (19) to vector-values α_i and γ_j we again make a transformation (38), but this time we choose

$$\begin{aligned} A_iA'_i &= \left(\sum_j \bar{H}_{ij}\right)^{\dagger}\left[\sum_j \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij}\right)\right]\left(\sum_j \bar{H}_{ij}\right)^{\dagger}. \\ C_jC'_j &= \left(\sum_i \bar{H}_{ij}\right)^{\dagger}\left[\sum_i \mathbb{E}\left(S_{ij}\,S'_{ij}\middle|x_{ij}\right)\right]\left(\sum_i \bar{H}_{ij}\right)^{\dagger}. \end{aligned} \tag{41}$$

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 (39) and (41) 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 (40), the new transformation defined by (41) guarantees that

$$\begin{aligned} \text{AsyVar}\left(\widehat{\widetilde{\alpha}}_i\right) &= \left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\alpha}_i)(\partial\widetilde{\alpha}_i)'}\right]^{\dagger}\text{Var}\left[\frac{\partial\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{\partial\widetilde{\alpha}_i}\middle|X\right]\left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\alpha}_i)(\partial\widetilde{\alpha}_i)'}\right]^{\dagger} = \mathbb{I}_{T-1}, \\ \text{AsyVar}\left(\widehat{\widetilde{\gamma}}_j\right) &= \left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\gamma}_j)(\partial\widetilde{\gamma}_j)'}\right]^{\dagger}\text{Var}\left[\frac{\partial\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{\partial\widetilde{\gamma}_j}\middle|X\right]\left[\frac{\partial^2\widetilde{\mathcal{L}}(\beta^0,\widetilde{\alpha}^0,\widetilde{\gamma}^0)}{(\partial\widetilde{\gamma}_j)(\partial\widetilde{\gamma}_j)'}\right]^{\dagger} = \mathbb{I}_{T-1}. \end{aligned} \tag{42}$$

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 (19) to $\widehat{\beta}$, and each component of $\widehat{\widetilde{\gamma}}_i$ contributes an incidental parameter bias of the form $\overline{D}_2$ in (19) 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 (42), 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 (41) 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 (19) 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 Remarks 1 and 2

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

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

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 $D_{ij,k} := \operatorname{diag}\left[\left(\lambda_{ijt}\, \widetilde{x}^*_{ijt,k}\right)_{t=1,\ldots,T}\right]$, 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

$$\sum_i D_{ij,k} = 0, \qquad\qquad \sum_j D_{ij,k} = 0. \tag{43}$$

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}, (43), 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 Remark 1. 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 Remark 1 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 $\text{Var}(y_{ij})$. 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 4. *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 two-way profile score in (9) to be unbiased at the true parameters (β, α, γ). 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 discussion clarifies, the FE-PPML model is quite special in this regard.

A.2 Proof of Proposition 2

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

Lemma 2. *Consider the class of "one-way" FE-PML panel estimators with conditional means given by* $\lambda_{it} := \exp(x_{it}'\beta + \alpha_i)$ *and 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 small,* $\widehat{\beta}$ *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

$$\omega_{it} := \exp\left[-\frac{1}{2}\ln\left(1 + \lambda_{it}^{\rho}\right) + \sqrt{\ln\left(1 + \lambda_{it}^{\rho}\right)}z_{it}\right] \tag{44}$$

with z_{it} a randomly-generated variable distributed $\mathcal{N}(0, 1)$. ω_{it} is therefore a heteroscedastic 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 heteroscedasticity and can be

any real number. With this in mind, it is useful to document the following results,

$$\mathbb{E}\left[\frac{\partial \omega_{it}}{\partial \rho}\right] = \frac{\partial \mathbb{E}\left[\omega_{it}\right]}{\partial \rho} = 0 \tag{45}$$

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\left(\omega_{it}^2\right)}{\partial \rho}\right] &= \mathbb{E}\left[2\omega_{it}\frac{\partial \omega_{it}}{\partial \rho}\right] = \frac{\partial \mathbb{E}\left(\omega_{it}^2\right)}{\partial \rho} \\ &= \frac{\partial V\left[\omega_{it}\right]}{\partial \rho} = \lambda_{it}^{\rho}\ln\lambda_{it} \neq 0.\end{aligned} \tag{46}$$

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}.[33]

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 (44). 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.[34] For notational 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

$$\frac{d\widehat{\beta}}{d\rho} = -\mathcal{L}_{\beta\beta}^{-1}\mathcal{L}_{\beta\rho} - \mathcal{L}_{\beta\beta}^{-1}\sum_i \mathcal{L}_{\beta\alpha_i}\frac{d\widehat{\alpha}_i}{d\rho}, \tag{47}$$

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

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

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

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

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

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 (48) into (47), we have

$$\begin{aligned}\frac{d\widehat{\beta}}{d\rho} &= -\mathcal{L}^{-1}_{\beta\beta}\mathcal{L}_{\beta\rho} + \mathcal{L}^{-1}_{\beta\beta}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho} + \mathcal{L}^{-1}_{\beta\beta}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\frac{d\widehat{\beta}}{d\rho} \\ &= -\left(\mathbf{I} - \mathcal{L}^{-1}_{\beta\beta}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\right)^{-1}\mathcal{L}^{-1}_{\beta\beta}\mathcal{L}_{\beta\rho} \qquad (49) \\ &\quad + \left(\mathbf{I} - \mathcal{L}^{-1}_{\beta\beta}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}\right)^{-1}\mathcal{L}^{-1}_{\beta\beta}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}, \qquad (50)\end{aligned}$$

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

Let $\mathbf{P}$ henceforth denote the combined matrix object $\mathbf{I} - \mathcal{L}^{-1}_{\beta\beta}\sum_i \mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}'_{\beta\alpha_i}$. It is straightforward to show that that first term in (50), $-\mathbf{P}^{-1}\mathcal{L}^{-1}_{\beta\beta}\mathcal{L}_{\beta\rho}$, 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, $\lim_{N\to\infty} NT\mathcal{L}^{-1}_{\beta\beta} = -\mathbb{E}[x_{it}\widehat{\lambda}_{it}g(\widehat{\lambda}_{it})x'_{it}]^{-1}$ must also be non-singular and finite. Slutsky's theorem then implies $\lim_{N\to\infty} -\mathbf{P}^{-1}\mathcal{L}^{-1}_{\beta\beta}\mathcal{L}_{\beta\rho} \to_p 0$ 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 (45)). We may therefore focus our attention on the second term on the RHS in (50), $\mathbf{P}^{-1}\mathcal{L}^{-1}_{\beta\beta}\sum_i \mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}$. Noting that $\mathcal{L}^{-1}_{\alpha_i\alpha_i}$must be < 0, in this case we consider the conditions under which $\lim_{N\to\infty} N^{-1}T^{-1}\sum_i \mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho}$ similarly converges in probability to zero. The summation in this latter term may be expressed as

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

Re-arranging this expression, we have that

$$\begin{aligned}\sum_{i=1}^{N}\mathcal{L}^{-1}_{\alpha_i\alpha_i}\mathcal{L}_{\beta\alpha_i}\mathcal{L}_{\alpha_i\rho} &= \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}^{-1}_{\alpha_i\alpha_i}x_{it}y_{it}g'(\widehat{\lambda}_{it})\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\frac{\partial y_{is}}{\partial\rho} \\ &\quad - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{L}^{-1}_{\alpha_i\alpha_i}x_{it}\Big(\widehat{\lambda}_{it}g'(\widehat{\lambda}_{it}) + g(\widehat{\lambda}_{it})\Big)\widehat{\lambda}_{it}g(\widehat{\lambda}_{is})\frac{\partial y_{is}}{\partial\rho}. \qquad (51)\end{aligned}$$

Focusing first on the second of the two summation terms in (51), 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

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

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 (51) 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

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

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 $s \neq t$.)[35] The remaining details follow from (46).[36] 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 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 models. ■

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

[35] 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.

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

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

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

this property.[37] 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.3 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) model, which imposes $g(\widehat{\lambda}_{ijt}) = \widehat{\lambda}_{ijt}^{-1}$. Generally speaking, GPML is considered the 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.[38] 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.[39] This is especially a concern now that recent computational

[37] Alternatively, it is possible to extend the above result to an even more general class of models 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 $g'(\widehat{\alpha}_i) = 0$.

[38] 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 model is consistent under the slightly more general assumption that the conditional variance is proportional to the square of the conditional mean.

[39] For example, Head and Mayer (2014), arguably the leading reference to date on gravity estimation,

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. 2 shows a comparison of simulation results for FE-PPML versus log-OLS and Gamma PML.[40] The displayed kernel densities are computed using 500 replications of a three-way panel structure with $N = 50$ and $T = 5$.[41] 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-heteroscedastic, 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 linear model—which is only unbiased for DGP III where the error term is log-homoscedastic—is also as expected. However, the reasons behind the bias in the OLS estimate are well-documented (see Santos Silva and Tenreyro, 2006) and do not have to do with the incidental parameters included in the model. The three-way FE-GPML 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, the three-way FE-GPML model 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 model. 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 unbiased when the DGP for ω_{ijt} was itself

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.

[40] 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 (2019). To our knowledge, these are the first results presented anywhere documenting the inconsistency of the three-way Gamma PML estimator.

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

Gaussian (as in DGP I), but were biased and inconsistent otherwise.

A.4 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 (14). 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:

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

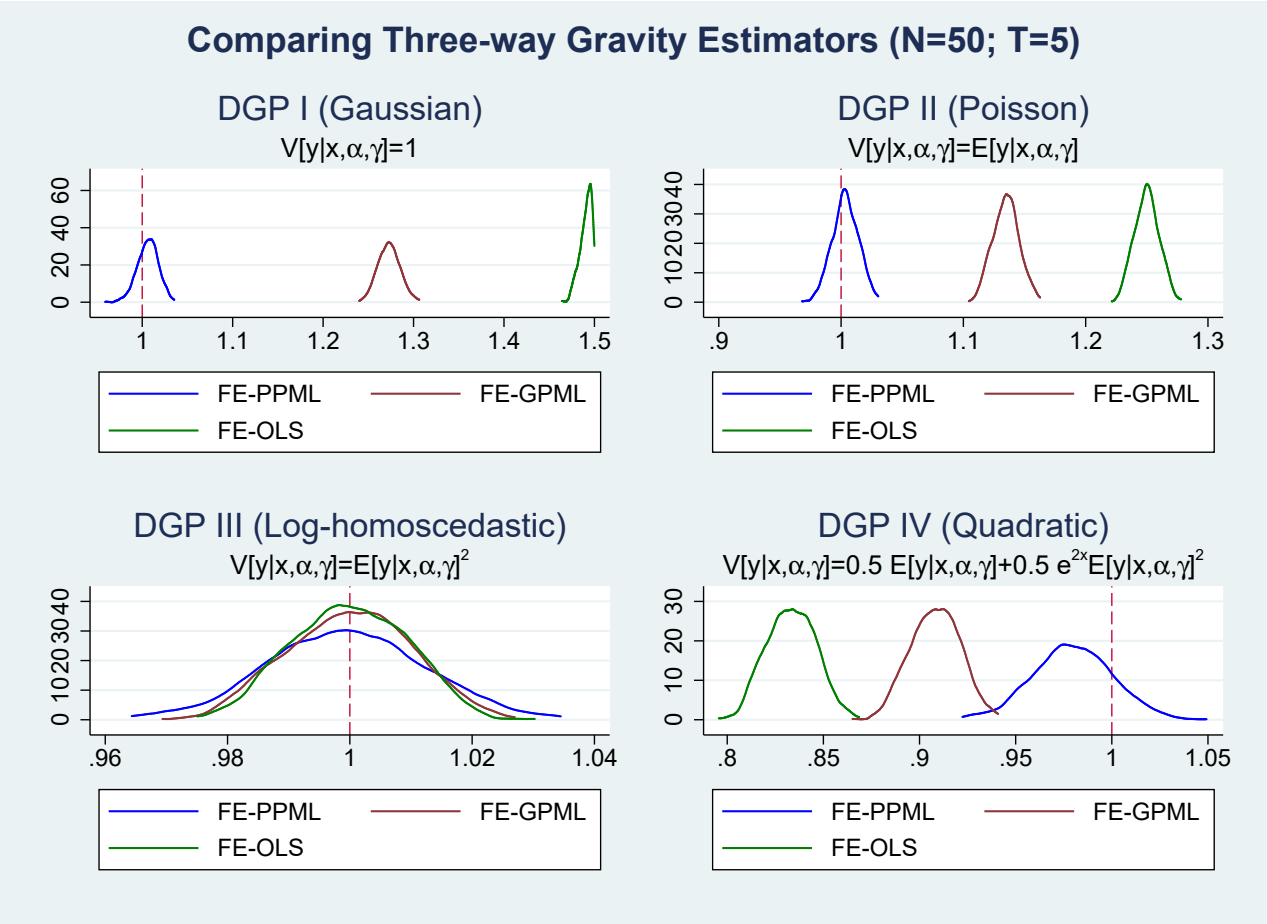

Figure 2: Kernel density plots of three-way gravity model estimates using different FE estimators, based on 500 replications. The model being estimated is $y_{ijt} = \exp[\alpha_{it} + \gamma_{jt} + \eta_{ij} + x_{ijt}\beta]\omega_{ijt}$, where the distribution of ω_{ijt} depends on the DGP and the true 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.

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 $(\bar{\mathcal{L}}_{\beta\beta} - \bar{\mathcal{L}}^{*\prime}_{\phi\beta}\bar{\mathcal{L}}^{-1}_{\phi\phi}\bar{\mathcal{L}}_{\phi\beta})^{-1} = -[N(N-1)]^{-1}W_N^{-1}$ and, similarly, that $\bar{\mathcal{L}}^{-1}_{\phi\phi} = -[N(N-1)]^{-1}W_N^{(\phi)-1}$. If we again consider $\mathbb{E}[\widehat{S}_{ij}\widehat{S}'_{ij}]$, we can now write

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

which simplifies to the expression shown in (14).

Results for the two-way model. The sandwich estimator is also known to be biased for the standard two-way gravity model without pair fixed effects. This bias has been documented in numerous places (Egger and Staub, 2015; Jochmans, 2016; Pfaffermayr, 2019) but the literature has yet to offer a bias correction that may be used to obtain improved inferences for this very popular model. 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.