Monte Carlo Stress Test of High-FE Dynamic Panels

The "Gold Standard" Trap in Applied Gravity Models

Gabriel Dunin-Borkowski Farid Salmanov

February 27, 2026

1. Applied practice and the issue of identification

The Problem: In applied trade and migration research, 3-way fixed effects (exporter-time, importer-time, dyad) are considered the "gold standard." But does this standard actually reliably identify effects?

Core claims from this Monte Carlo:

- **The variance is wiped out:** 3-way FEs strip out $\sim 90\%$ of the signal, leaving estimators to fit to undersized, idiosyncratic noise.
- **Spurious significance:** Estimators heavily over-reject under a structured null (H1: persistent noise) and report standard errors which are biased downwards.
- **Severe attenuation:** Statistical significance coexists with massive bias and 0% interval coverage.

Context: `ppml_3fe` vs `ppml_2fe`, `diff_gmm`, `system_gmm`; small and medium panels; **1,000 replications** per cell.

2. Study design

To prove this, we simulate two known Data Generating Processes (DGPs):

- `gravity_true`**:** True structural effect $\beta = -0.55$.
- `persistent_random`**:** $\beta = 0$ by construction, high persistence.
- **Estimators:** `ppml_2fe`, `ppml_3fe`, `diff_gmm`, `system_gmm`.
- **Robustness:** Baseline GMM vs. Restricted GMM (lag window 2:3, collapsed).

Design	Countries (directed dyads)	Periods
Small	14 ($14 \times 13 = 182$)	10
Medium	20 ($20 \times 19 = 380$)	12

Scaled up to 1,000 replications per design-scenario cell.

3. DGP: Structural equations

We simulate a dynamic panel where both the regressor and the outcome are driven by AR(1) persistence and 3-way unobserved heterogeneity.

1. Regressor (x_{ijt}): Highly persistent, embedded with FEs.

$$x_{ijt} = \underbrace{\rho_x x_{ij,t-1}}_{\text{AR(1)}} + \underbrace{\eta^x_{ij}}_{\text{Pair FE}} + \underbrace{\theta^x_{it}}_{\text{Exp-Year FE}} + \underbrace{\psi^x_{jt}}_{\text{Imp-Year FE}} + \varepsilon^x_{ijt}$$

2. Trade Outcome (Y_{ijt}): Dynamic Poisson gravity process.

$$Y_{ijt} \sim \text{Poisson}(\mu_{ijt}) \quad \text{where:}$$

$$\ln(\mu_{ijt}) = \alpha + \underbrace{\phi \ln(1 + Y_{ij,t-1})}_{\text{Trade AR(1)}} + \beta \mathbf{x_{ijt}} + \underbrace{\eta^y_{ij} + \theta^y_{it} + \psi^y_{jt}}_{\text{Outcome FEs}} + \varepsilon^y_{ijt}$$

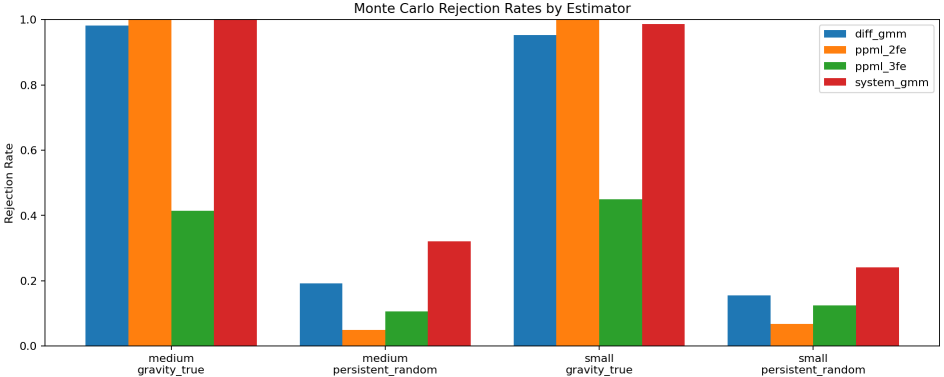
3. Scenarios:

- `gravity_true` **(Signal):** $\beta = -0.55$. The variable x has a true causal effect.
- `persistent_random` **(Noise):** $\beta = 0$. x and Y simply co-evolve as persistent series.

4. Baseline size and power (1,000 reps)

Size under Null ($\beta = 0$): Diff-GMM falsely rejects the null 15.5% (small) and 19.2% (medium) of the time. System-GMM falsely rejects 24.1% (small) and 32.1% (medium).

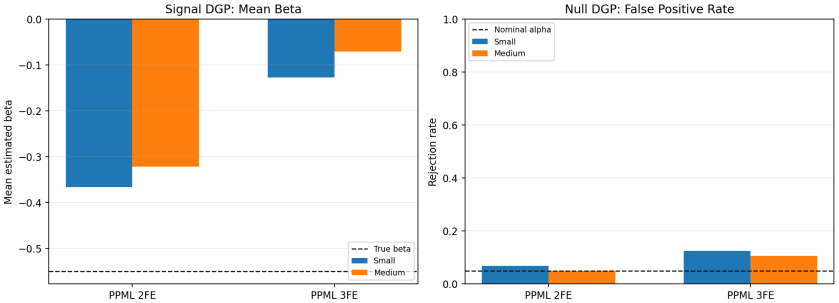
Takeaway: The models heavily over-reject the null. Significant findings under such a scenario can be false positives fitting to persistent noise.



5. PPML FE Sweep: The 3FE attenuation trap

By comparing 2-way FE (Exp-Time, Imp-Time) against 3-way FE (+ Dyad), we show the source of the problem.

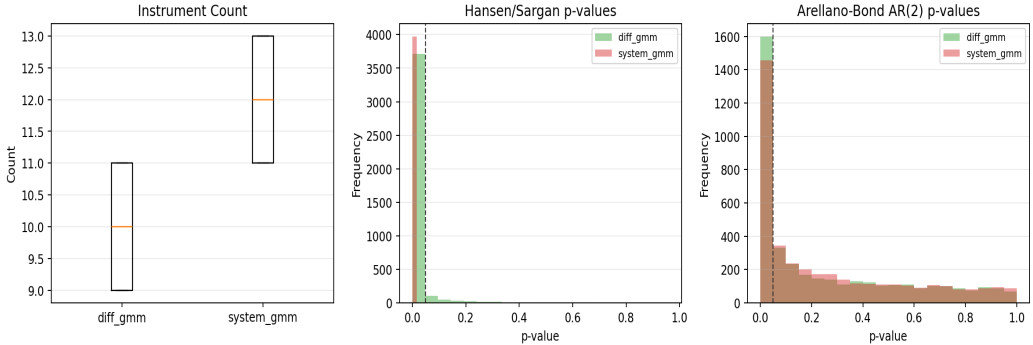
- **Signal mean $\hat{\beta}$ (True: -0.55):** Small 2FE recovers -0.366, 3FE attenuates to -0.127. Medium 2FE recovers -0.322, 3FE attenuates to -0.071.
- **Takeaway:** The 3FE "gold standard" wipes out so much variance that measurement error and idiosyncratic noise completely dominate the remaining signal.



6. GMM diagnostics reveal severe moment stress

Are the GMM moments actually valid? We logged AR(2) and Hansen/Sargan p-values per replication.

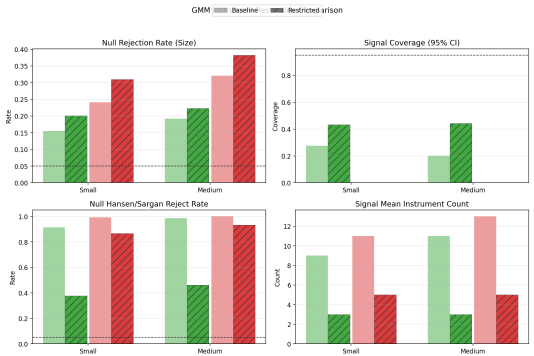
- **System-GMM Hansen rejection:** 99.2% (small/null) and 100.0% (medium/null).
- **Diff-GMM Hansen rejection:** 91.4% (small/null) and 98.7% (medium/null).
- AR(2) rejections also sit uncomfortably high (30% - 47%).



7. Fix?: GMM instrument restriction

We attempted to fix the GMM estimators by limiting instruments (lag window 2:3, collapsed matrix).

- **Instrument counts dropped:** Diff (9/11 to 3); System (11/13 to 5).
- **Coverage improved:** Diff-GMM signal coverage rose +24.0 pp (medium).
- **Size under null did not improve:** False positive rates actually *worsened*. System-GMM null rejection rose to 38.2% (medium).



8. Causes of failure

PPML 3-Way FE

Attenuation is strongly tied to the properties of FE. When 90% of the regressor's variation is thrown out, signal-to-noise ratios collapse. 3FE forces the model to fit to the noise.

Difference GMM

Restricting instruments improves some calibration/coverage under a true signal, but moment invalidity remains high. High null over-rejection limits its reliability.

System GMM

Routinely used in applied macro/regional panels, yet it is catastrophically unstable here. It suffers from 100% Hansen rejection, massive attenuation, and terrible size. Instrument restriction does not save it.

9. What now?

Takeaway: In high-FE settings, significant findings without diagnostic context are practically meaningless.

A proposal for how to avoid fitting data to noise:

1. **Report Retained-Variation:** You must report the auxiliary R^2 of your variable of interest against the FEs. (If $\rho_x < 0.10$, beware).
2. **Compare FE:** Always report 2-way FE alongside 3-way FE to benchmark attenuation.
3. **GMM Diagnostics:** Explicitly report instrument counts vs. N, and over-identification diagnostics (Hansen J, AR(2)). A high number of instruments severely compromises System-GMM.

Applying the 3-way FE is a recipe for fitting your data to the noise.