# Untangling Fixed Effects and Constant Regressors

Franc Klaassen *
University of Amsterdam and Tinbergen Institute
&
Rutger Teulings
CPB Netherlands Bureau for Economic Policy Analysis
and University of Amsterdam

July 19, 2021

**Abstract**

Fixed effects (FE) in panel data models require normalization and prohibit identification of the impacts of "constant" regressors. Think of regressors that are constant across countries in a country-time panel with time FE. We show, however, there is identification if the normalized FE are zero, whatever the normalization. This gives a testable constraint for identification. If it holds, FE can be left out. If not, we propose "untangling normalization" to ease interpretation of the FE and find omitted regressors. In a gravity model for exports to the US, the constant regressors US GDP, world GDP, and US effective exchange rate explain 98% of the time FE, making these FE insignificant and redundant. We thus achieve identification.

*Keywords:* gravity model, identification, multicollinearity, normalization, unobserved heterogeneity.
*JEL classification:* C18; C23; F14.

---

*Corresponding author. Amsterdam School of Economics, PO Box 15867, 1001 NJ Amsterdam, The Netherlands; tel. +31-20-5254191; e-mail: f.klaassen@uva.nl.
An earlier version has appeared as Tinbergen Institute Discussion Paper 15-137/VI.
We would like to thank Peter Boswijk, Maurice Bun, Noud van Giersbergen, Artūras Juodis, Jan Kiviet, Frank Kleibergen, Maurice Koster, Naomi Leefmans, Jan Magnus, Andrey Vasnev, and seminar/conference participants at the University of Amsterdam, Panel Data Workshop in Amsterdam, 6th Economics & Finance Conference at the OECD Headquarters in Paris, University of Göttingen, 17th Annual Conference of the European Trade Study Group in Helsinki, CPB Netherlands Bureau for Economic Policy Analysis, Tinbergen Institute Amsterdam, and the 25th International Panel Data Conference in Vilnius for very useful and constructive comments.

## 1 Introduction

In panel data models, researchers often want to include fixed effects (FE) to control for unobserved heterogeneity. However, many important research questions concern regressors that only vary in the same dimension as the FE. Think of the impact of gender and ethnicity on wage inequality in panels with individual FE, and the effect of geographical characteristics or time-invariant institutions on the economic development of countries in case of country FE. Similarly, one would like to know the impact of global economic and environmental developments on countries, while controlling for time FE.

This paper studies the perfect multicollinearity due to FE and such "constant" regressors (constant, as they are invariant in a dimension of the panel). We focus on the (ceteris paribus) impact of constant regressors on the dependent variable. It is well known that this is not identified. Our approach, loosely speaking, is to exploit the explanatory power of the constant regressors and let the FE capture the remaining heterogeneity. This shrinks the FE compared to the usual approach, where FE contain the full heterogeneity, creating a more realistic possibility that the FE are zero. If they are, the impact of the constant regressors is identified, even though it was unidentifiable at first sight. Our identifying constraint — that all FE are zero — is testable, so we do not a priori *assume* identification. For the case where FE are nonzero, we introduce a way to better interpret and visualize them to find omitted constant regressors.

More precisely, the paper starts by recognizing that multicollinearity requires normalization of some parameters. The typical approach is to normalize the impact of the constant regressors to zero, thus leaving them out in estimation. One then gives up on estimating their impact. An alternative is to normalize some FE and get an estimate. However, this estimates a pseudo-true value, depending on the normalization, not the true value one is interested in. The true value is not identified.

Next, we realize the normalized FE contain information on the constant regressors. Our goal is to extract this, leading to two contributions. The first is our identifying constraint and its testability: if all normalized FE are zero (always testable), one estimates the true value of the impact of the constant regressors. This is an alternative to the leading existing approach by Hausman and Taylor (1981), denoted by HT.[1] That can deliver valuable insights if there are enough instruments, if they are sufficiently

[1] Hausman and Taylor (1981) treat the effect as random instead of fixed. Of course, they can restrict constant regressors to be uncorrelated with the random effect, a restriction that features many special cases, such as random effects (RE), fixed effects vector decomposition (FEVD) of Plümper and Troeger (2007), and the hybrid or between-within (BW) approach in Allison (2009, p. 23), which extends the correlated random effects approach in Mundlak (1978). But a key contribution is that HT can also allow for correlation, that is, endogeneity of constant regressors. They have an IV procedure where instruments are constructed from within the model by taking averages of variables that also vary over other panel dimensions, such as averages over time of country-time variables in case of country effects. If these instruments are exogenous and relevant, the true value of the (ceteris paribus) impact of the constant regressors is identified.

strong, and if the zero correlation typically assumed between some constant regressors and the HT random effect is warranted. These conditions can be hurdles in practice; Kripfganz and Schwarz (2019). Moreover, testing the zero correlation assumption is hard or can even be impossible; Ahn and Moon (2014). In contrast, our approach does not depend on instrument availability and strength, and we can always test our constraint by using the estimated normalized FE. Section 3.3 describes the trade-off between our method and HT more completely, and Section 5.5.3 illustrates the practical relevance of the HT conditions and how our method can also help out an HT analysis.

As indicated earlier, our approach consists of two parts, one using a normalization to estimate a pseudo-true value, and the next using a constraint to identify the true value. To avoid confusion between "normalization" and "constraint", let us fix the terminology used throughout the paper. Normalizations are irrelevant for the conditional distribution of the dependent variable (they set parameters in an observationally equivalent way), so in this sense they are without loss of generality. In contrast, we use constraint for a restrictive restriction. Hence, the terms are fundamentally different.

To illustrate both terms and what is exactly needed for identification, consider a simple linear country-time panel model. Ignore the constant term for now. The FE is $\alpha_i$ for country $i$, there is one constant regressor $v_i$, and $\nu$ is the true value of its impact.

One typically leaves out $v_i$ in estimation, that is, normalizes $\nu^0 = 0$. This exemplifies what we call "zero normalization", and we use the superscript 0 to distinguish zero-normalized parameters, which are pseudo-true values, from the true value. An alternative zero normalization is $\alpha_1^0 = 0$. This changes the normalized impact of the constant regressor into $\nu^0 = \nu + \alpha_1/v_1$.[2] Hence, both normalizations pin down their own value for $\nu^0$. Those values generally differ from $\nu$, reflecting that $\nu$ is not identified.

Still, both normalizations yield the same sum $\alpha_i^0 + v_i\nu^0$, equal to $\alpha_i + v_i\nu$, so that $\nu^0$ and $\alpha_i^0$ fully absorb the choice made for the normalization. We can thus safely take some zero normalization and estimate all normalized parameters and the sum $\alpha_i + v_i\nu$.

Now, under the constraint that *all* normalized FE $\alpha_i^0 = 0$, we show that all $\alpha_i = 0$. As the sum $\alpha_i + v_i\nu$ is identified, the constraint then implies the true value $\nu$ is identified. We also show that this idea generalizes to other normalizations, reflecting that the normalization itself does not identify the true value; it is the constraint.

The second contribution of the paper concerns the case where our constraint does not hold, so there exist nonzero normalized FE. We want to visualize them in a convenient way to extract information on omitted constant regressors. Let us maintain the example above, though no longer ignoring the constant term $\alpha$, and consider the zero normalization $\alpha^0 = \nu^0 = 0$. This implies that the normalized FE become $\alpha_i^0 = \alpha + \alpha_i + v_i\nu$. Hence, they capture not only the effect of country $i$, but also the

[2]The superscript 0 is used for all zero normalizations and its exact meaning appears from the context. The example assumes that the countries are ordered such that $v_1 \neq 0$.

overall intercept and the relevance of $v_i$, creating overlap. This blurs the signal from omitted determinants in the estimated $\alpha_i^0$.

To resolve overlap we introduce "untangling normalization", which makes the normalized FE orthogonal to each other and to constant regressors. Untangling is a normalization, so it is irrelevant and not used for identifying the true value; it is just for interpretation. Hence, it differs from the orthogonality relations in existing methods, such as RE, FEVD, BW, and HT, which are constraints assumed to get identification.

Untangling offers several advantages. First, it eases interpretation and is unique. For example, untangling sets the mean of the country FE to zero, so that the untangled constant $\alpha^u$ captures the overall intercept, and the untangled country FE $\alpha_i^u$ is the country deviation from the overall intercept. There is no overlap between them, easing interpretation. This specific example is not new; see Suits (1984). But we use a richer setting, with country FE, time FE, country-specific trends, and regressors that are constant across countries or time, and untangling can be generalized further, for example, to three-dimensional panels. We also show how to estimate such FE.

The second advantage of untangling is that the $\alpha_i^u$ capture what is left after accounting for $v_i$. This facilitates judging how large the deviation from the null hypothesis is, and it helps to find omitted variables, as we will show in our application. Adding newly discovered variables to the model can then shrink the remaining FE.

Our approach is broadly applicable. We illustrate it in a gravity model for exports from OECD countries to the US; see Head and Mayer (2014) for a gravity review. We focus on time FE and have three constant regressors, namely US GDP, world GDP, and the US real effective exchange rate (REER). Our identifying constraint is not rejected, and the constant regressors explain 98% of the time FE, so we leave them out. We thus identify the true value of the constant regressor impacts. This is typically considered beyond reach. The REER is a key determinant, accounting for 20% of the time FE, calling for extension of gravity theory, as Klaassen and Teulings (2017) do.

Another successful application concerns pension reforms in Beetsma et al. (2020). In general, our method can help researchers to motivate leaving out FE to get an estimate for constant regressors and, if the constraint is rejected, find omitted variables.

The paper is organized as follows. In Section 2 we describe the model and discuss normalization. Section 3 introduces the identifying constraint and shows the differences with Hausman and Taylor (1981). Section 4 sets out untangling normalization. In Section 5 we apply our approach and HT to the gravity model. Section 6 concludes.

## 2 Model specification

Throughout the paper we consider a two-dimensional balanced panel model with dimensions $i$ and $t$, representing country and time, say. There are $N$ countries and $T$

time periods. This section discusses the model, the perfect multicollinearity involved, and the required normalization, leading to the distinction between pseudo-true and true value of the impact of the constant regressors. This is well known, but we present it in a general way that we exploit later. Identification of the true value is a separate issue and will be addressed in Section 3.

## 2.1 Model with the full set of FE and constant regressors

The dependent variable is

$$y_{it} = \alpha + \alpha_i + \tau \cdot t + \tau_i \cdot t + \theta_t + v_i'\nu + w_t'\omega + x_{it}'\beta + \varepsilon_{it}. \tag{1}$$

Overparameterization, such as having $\alpha$ and $\alpha_i$, will be resolved in Section 2.2.

The vector $v_i$ contains $K_v$ $i$-regressors, all variables that only vary over countries. Likewise, $w_t$ contains $K_w$ $t$-regressors. Hence, $v_i$ and $w_t$ are the constant regressors. Their ceteris paribus impacts are the parameters of interest and denoted by $\nu$ and $\omega$.[3] Variables that vary over both dimensions, the $it$-regressors, are in $x_{it}$ of length $K_x$. All vectors in the paper are column vectors.

To control for potential omitted regressors, we add unobserved effects. They are grouped in three FE-families. The $\alpha$-family targets the level variation across countries. It has a homogeneous type, $\alpha$, and a heterogeneous type, $\alpha_i$. The $\tau$-family targets linear trends across countries and consists of $\tau$ and $\tau_i$. The $\theta$-family targets the time variation, consisting of $\theta_t$. This adds $K_d$ parameters in total.

The assumptions involving the error term $\varepsilon_{it}$ are those that the user prefers, as long as they deliver an estimator of $\alpha + \alpha_i + \ldots + w_t'\omega$ in (1), that is, the *sum* of the deterministic and constant regressors parts, not the parts themselves. Hence, our method can be combined with various assumptions and estimators. A typical example is to assume a zero mean of $\varepsilon_{it}$ conditional on the regressors in all times and no cross-sectional correlation, while allowing for heteroscedasticity and serial correlation.

It is convenient to write (1) in matrix form, stacking the time series of the countries:

$$y = D\delta + Z\gamma + X\beta + \varepsilon, \tag{2}$$

[3]Defining $\nu$ (and $\omega$) as the ceteris paribus impact is in line with the literature. First, in HT $\nu$ is also the impact of $v_i$ when keeping omitted $i$-regressors constant, where the latter are in the HT composite error term. Second, our definition resembles the approach in a textbook regression model. That is, consider a cross-section model with one regressor, no constant, and the standard assumptions. There, one could write $\mathrm{y}_i=\mathrm{a}_i+\mathrm{c}\cdot\mathrm{x}_i+\mathrm{e}_i$, where $\mathrm{a}_i=\mathrm{d}\cdot\mathrm{x}_i$. However, one is typically interested in the ceteris paribus impact of $\mathrm{x}_i$, and to be able to vary $\mathrm{x}_i$ while keeping other determinants constant, one collects the terms involving $\mathrm{x}_i$ by writing $\mathrm{y}_i=\mathrm{b}\cdot\mathrm{x}_i+\mathrm{e}_i$, where $\mathrm{b}=\mathrm{c}+\mathrm{d}$ is the ceteris paribus impact. Similarly, we collect the terms involving $v_i$, so that $\nu$ is its ceteris paribus impact. This is just linking the impact of interest to one specific parameter, without loss of generality. This parameterization does not address the identification problem we will focus on, because $\alpha_i$ may still vary over $i$ by other, omitted variables, thereby making $\nu$ unidentified.

where $y$ and $\varepsilon$ stack all $y_{it}$ and $\varepsilon_{it}$, respectively. All deterministic variables constitute the matrix $D = [\iota_N \otimes \iota_T, I_N \otimes \iota_T, \iota_N \otimes [1, 2, \ldots, T]', I_N \otimes [1, 2, \ldots, T]', \iota_N \otimes I_T]$, where $\iota_n$ is the vector of ones of length $n$, and $I_n$ is the identity matrix of order $n$. The associated FE parameters form $\delta = [\alpha, \alpha_1, \ldots, \alpha_N, \tau, \tau_1, \ldots, \tau_N, \theta_1, \ldots, \theta_T]'$. The constant regressors are stacked into the matrix $Z$, where its $it$-th row $[v_i', w_t']$ consists of $K_z = K_v + K_w$ columns, and the corresponding parameter vector is $\gamma = [\nu', \omega']'$. All $it$-regressors are stacked in $X$.

We assume that the columns in $[D, Z, X]$ are linearly independent, except for the dependencies within $D$ and between $D$ and $Z$ set out in the next section ($Z$ itself has full column rank).

### 2.2 Multicollinearity and normalization

There is (perfect) multicollinearity in $[D, Z]$ for two reasons. The first is within $D$. For example, the vector of ones in $D$ is the sum of the $N$ vectors of country dummies in $D$. In general, $D$ has column rank $K_d - m_d$, where $m_d$ is the degree of multicollinearity, the number of dependent columns. In model (2) $m_d = 4$, that is, one due to the $\alpha$-family, one due to the $\tau$-family, and two because $\theta_t$ is combined with $\alpha$ and $\tau \cdot t$.

The second source of multicollinearity is that the $v_i$-columns in $Z$ are linear combinations of the vectors of country dummies in $D$, and similarly for the $w_t$-columns regarding the time dummies. This adds $m_z = K_z$ dependencies.

In total, the column rank of $[D, Z]$ is $K_d + K_z - m_d - m_z$. Hence, from the sum $D\delta + Z\gamma$ we cannot infer $\delta$ and $\gamma$.

We thus introduce $m_d + m_z$ normalizations. One can choose them freely, under two requirements, and we call the normalizations together the general normalization, indicated by $g$.

The first requirement is that the normalization is linear in the parameters. For later convenience, we formalize this by

$$N^g \begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = 0, \tag{3}$$

where the $(m_d + m_z) \times (K_d + K_z)$ normalization matrix $N^g$ has independent rows, each specifying one normalization. The $g$-superscript in the general-normalized parameter $\gamma^g$ makes explicit that it is a pseudo-true value, which generally differs from the true value $\gamma$ of the impact of the constant regressors, the value one is actually interested in.

A zero normalization is a special case that sets specific elements of $\delta^0$ and $\gamma^0$ to zero. Hence, each row of $N^0$ has a one at the place corresponding to the zero-normalized parameter, and (3) applies with $g$ substituted by 0. For example, if we normalize the constant to zero, so $\alpha^0 = 0$, then $N^0$ contains the row $[1, 0, .., 0]$. Instead, if we normalize the $i$-th country FE to zero, the row is $[0, .., 0, 1, 0, .., 0]$, where element $1 + i$ is one.

For the second requirement on the normalization, realize that the normalization should distribute the sum $D\delta + Z\gamma$ over the fixed effects and the constant regressors without changing the total and in a unique way, that is,

$$D\delta + Z\gamma = D\delta^g + Z\gamma^g, \tag{4}$$

for unique $\delta^g$ and $\gamma^g$. To achieve this, we append (3) to (4). Then, for any value of the sum, the requirement on $N^g$ becomes

$$\text{rank} \begin{bmatrix} D & Z \\ & N^g \end{bmatrix} = K_d + K_z. \tag{5}$$

In case of zero normalization, one typically removes columns of $[D, Z]$ to obtain a regressor matrix of full column rank, but we append (3) to (4) for $g = 0$ to get full column rank. That is equivalent. Our approach is more tractable here, as it allows us to account for many normalizations by changing just $N^g$, leaving $D$ and $Z$ untouched.[4]

The normalization ensures that $\delta^g$, $\gamma^g$, and $\beta$ can be estimated. Appendix A sets out two estimation methods. This is where typical fixed-effects modeling stops, motivated by a focus on $\beta$. We go further by analyzing $\delta^g$ and $\gamma^g$ in Sections 3 and 4.

# 3 Identifying the true value of the impact of constant regressors

Our overall approach is split into two parts, one about normalization and the other about a constraint. The normalization, as explained in Section 2.2, is irrelevant for the conditional distribution of the dependent variable $y$ and is only used to obtain an identified model. But the chosen normalization affects $\gamma^g$, making its estimate unusable for inference. Put differently, we are interested in the true value, $\gamma$, but that has not been identified so far. A simple example is when $\gamma^g$ is normalized to zero, which says nothing about $\gamma$, of course.

The constraint, discussed in Section 3.1, is on $\delta^g$ to identify $\gamma$. This is not limited to some specific normalization, as the $g$-normalization captures many of them. We focus on the $\nu$-part of $\gamma$, that is, the impact of time-constant regressors $v_i$; the approach for $w_t$ is similar.

[4] Appendix A.2 provides a $g$-specific column-removal approach for each $N^g$ normalization.

### 3.1 A testable constraint to identify the true value

The identification problem is that an observationally-equivalent model results by taking another value instead of $\nu$. That is, taking $\nu^g$ instead of $\nu$, defining $\alpha^g + \alpha_i^g$ such that

$$\alpha + \alpha_i + v_i'\nu = \alpha^g + \alpha_i^g + v_i'\nu^g, \tag{6}$$

and then substituting this into (1) gives the same $y_{it}$. One can thus estimate the pseudo-true values $\alpha^g$, $\alpha_i^g$, and $\nu^g$, but one cannot infer estimates of the true values $\alpha$, $\alpha_i$, and $\nu$ from them. This is a well-known and unsolved problem.

As a potential solution, consider the null hypothesis

$$H_0 : \alpha_i^g = 0 \text{ for all } i. \tag{7}$$

Note that this constraint is not on the unidentified $\alpha_i$, but on the normalized $\alpha_i^g$. The latter can be estimated, so this constraint is testable.

Under the constraint, the right-hand side of (6) shrinks to $\alpha^g + v_i'\nu^g$. The $\alpha + \alpha_i + v_i'\nu$ part in (1) now becomes $\alpha^g + v_i'\nu^g$, so we have a regression on essentially just $v_i$, as in a textbook regression without FE. Hence, $\nu^g$ is not influenced by the normalization, is unique, and is the ceteris paribus impact of $v_i$ on $y_{it}$. Now, recall from below (1) that we have denoted the ceteris paribus impact by $\nu$. Because there is only one such impact, $\nu = \nu^g$, so that $\nu$ is identified.

An alternative presentation gives some additional insights. Under constraint (7), equation (6) becomes

$$\alpha_i = \alpha^g - \alpha + v_i' \left(\nu^g - \nu\right), \tag{8}$$

where $\nu^g - \nu$ is unique, because there is no exact linear relationship among the constant regressors. Hence, $[\alpha_1, \ldots, \alpha_N]'$ lies in the column space of $[v_1, \ldots, v_N]'$. Now, recall from below (1) that the motivation for including $\alpha_i$ is to control for omitted $i$-variables, that is, for vectors outside the column space. From (8) we know that such variables are absent. Hence, there is no reason to add $\alpha_i$, that is, $\alpha_i = 0$ for all $i$. Substitution into (8) implies $\nu = \nu^g$, so the impact of the constant regressors is identified.

We thus have a testable constraint to tackle the identification problem. We use two ways to test this constraint. The first examines the constraint directly, so it is a diagnostic test. The second verifies whether the estimates of other parameters, $\beta$, are affected by the constraint, so we call that the sensitivity test. Appendix B sets out both approaches in detail and studies the Wald statistic for both in a Monte Carlo analysis. That suggests good size and power properties. Still, for $T < 20$ the diagnostic test is slightly oversized, but that is repairable and only makes our approach conservative. Section 5 illustrates how the tests work in practice.

### 3.2 Relevance of the normalization for the realism of the constraint

Although the chosen normalization $g$ is irrelevant for the conditional distribution of $y$, and for each $g$ hypothesis (7) identifies $\nu$, the specific choice can matter for the realism of the hypothesis. The reason is as follows, starting from the fact that $g$ consists of $1 + K_v$ normalizations on $\alpha^g$, $\alpha_i^g$, and $\nu^g$.

If the normalizations only concern $\alpha_i^g$, then the null hypothesis constrains its remaining $N - 1 - K_v$ elements. This correctly reflects that the null says the $N$ observations of the sum (6) are driven by a constant and $v_i$, that is, by $1 + K_v$ determinants.

However, if the normalization is $\alpha^g = 0$ and $\nu^g = 0$, then all explanatory power of the constant term and constant regressors for the sum on the left-hand side of (6) is concentrated into $\alpha_i^g$, and the null constrains all $N$ of them to zero. This is less realistic than the $N - 1 - K_v$ constraints resulting under the previous normalization, where the explanatory power of the constant term and constant regressors was exploited.

Researchers often normalize at least $\nu^g = 0$. We thus advocate a different approach. To increase the realism of the constraint, one should exploit the information in the constant regressors, so that only part of the sum $\alpha^g + \alpha_i^g + v_i'\nu^g$ ends up in $\alpha_i^g$. That is, one should normalize some of the $\alpha_i^g$ and leave $\alpha^g$ and $\nu^g$ free. In general, for a clean identification analysis, one should normalize parameters that will be constrained by the null hypothesis and leave the remaining involved parameters free. As long as $g$ fulfills this requirement, the specific choice of $g$ does not matter.[5]

### 3.3 Comparison to the literature

The leading existing approach to identify $\nu$ is due to Hausman and Taylor (1981). HT treat $\alpha_i$ as mean-zero random variable and use an IV procedure that depends on the following moment restrictions. First, they restrict that a subset of the constant regressors, $v_{1i}$, are uncorrelated with $\alpha_i$: $\mathbb{E}\{v_{1i}\alpha_i\} = 0$. Second, consider the other constant regressors, $v_{2i}$, which are allowed to correlate with $\alpha_i$ and are thus endogenous. HT restrict that the time-averages of a subset of $it$-regressors, $\bar{x}_{1i}$, are uncorrelated with $\alpha_i$, so $\mathbb{E}\{\bar{x}_{1i}\alpha_i\} = 0$, and that the $\bar{x}_{1i}$ are relevant instruments for $v_{2i}$. Let $k_1$ be the number of instruments and $g_2$ the number of endogenous constant regressors. Then HT require $k_1 \geq g_2$. All this identifies $\nu$.

If $k_1 > g_2$ (overidentification), HT use the Hausman principle to test their prior restrictions. That is, they compare the estimated $\beta$ when the two moment restrictions are imposed to the estimate without using them, which is the within-groups estimate.[6]

---

[5]Of course, in case of hypothesis testing, we exclude normalizations that make the null hypothesis impossible, for example, normalizations that fix $\alpha_i^g$ at a nonzero value for some $i$. Put differently, if the free elements in $\alpha_i^g$ are consistent with the constraint, then the other elements, which result from the free elements and the normalization, must also fulfill the constraint.

[6]HT encompasses many well-known special cases, including the random effects (RE), Mundlak

The difference between our approach and HT is fourfold. First, we treat $\alpha_i$ as fixed instead of random. Second, our constraint to identify $\nu$ is different. Third, we can test the constraint directly, using estimates of $\alpha_i^g$, instead of via $\beta$. Therefore, we can always test the constraint (without requiring overidentification), with substantial power to detect correlation between $v_i$ and $\alpha_i$, and our test does not depend on instrument strength. The final difference concerns a trade-off in estimation, involving the restrictiveness of the constraint and instrument weakness. We discuss these differences below. The overall conclusion is that our approach has pros and cons. It can be a useful and simple method on its own. It can also complement HT, as Section 5.5.3 illustrates.

First, HT is a hybrid of fixed and random-effects approaches, as the HT subsets $\bar{x}_{2i}$ and $v_{2i}$ are allowed to correlate with $\alpha_i$, whereas $\bar{x}_{1i}$ and $v_{1i}$ are not. In contrast, we treat $\alpha_i$ as fixed, so that we work in a fully FE setting, allowing all regressors $\bar{x}_i$ and $v_i$ to correlate freely with $\alpha_i$. This can be valuable in practice, because motivating what regressors in $\bar{x}_i$ and $v_i$ are exogenous can be onerous (Breusch et al. (2011)). Another attractive feature of the FE approach is that it delivers estimates and standard errors of the normalized FE, which helps to see what the model misses.

Second, consider how the true value $\nu$ is identified. HT use the above moment restrictions for that. Instead, we use a constraint on $\alpha_i^g$, constraint (7). We can do this because we split the analysis. The first phase handles multicollinearity by some normalization, which is irrelevant for the conditional distribution of the dependent variable. That generates pseudo-true values one can estimate, including $\alpha_i^g$. In the second phase we study the constraint on $\alpha_i^g$ to identify $\nu$.

The third difference between our approach and HT concerns testing. HT impose the identifying restrictions during estimation. In case of exact identification ($k_1 = g_2$), those restrictions are fulfilled by construction, so there is no test on their validity. Note that this also applies to all special cases of HT in footnote 6, except for the RE method.

If there are additional exogenous regressors ($k_1 > g_2$), HT can test. How to interpret the outcome? The HT test focuses on $\beta$, which explains its success at finding correlation

(1978), hybrid or between-within (BW), and fixed effects vector decomposition (FEVD) approaches. All special cases restrict all constant regressors to be uncorrelated with $\alpha_i$, that is, $\mathbb{E}\{v_i\alpha_i\} = 0$. In HT notation, $g_2 = 0$. They differ regarding their restrictions on the correlation between $it$-regressors and $\alpha_i$, that is, $\mathbb{E}\{\bar{x}_i\alpha_i\}$, as follows.

The RE approach is the special case that restricts $\mathbb{E}\{\bar{x}_i\alpha_i\} = 0$. Hence $k_1 = K_x$, the number of $it$-regressors. $k_1 > 0$ enables the familiar Hausman test for random versus fixed effects.

Mundlak (1978) weakens the RE restriction by auxiliary model $\alpha_i = \bar{x}_i^{c\prime}\cdot\pi$ plus noise that has mean zero and is uncorrelated with $\bar{x}_i$, where the $c$ superscript denotes centering and $\pi$ is the Mundlak parameter. He thus restricts $\mathbb{E}\{\bar{x}_i\alpha_i\} = \mathbb{E}\{\bar{x}_i\bar{x}_i^{c\prime}\}\cdot\pi$. This is a correlated random effects approach. As HT set out, it is the special case where $k_1 = 0$, so there is no test of the moment restrictions. Although Mundlak (1978) has no constant regressors $v_i$, one can add them, as in the BW model of Allison (2009, p. 23).

Finally, FEVD introduced by Plümper and Troeger (2007) leaves $\mathbb{E}\{\bar{x}_i\alpha_i\}$ unrestricted. Hence $k_1 = 0$, so there is no test of the moment restrictions. The generality also underlies the estimators by Plümper and Troeger (2011), Honoré and Kesina (2017), and Pesaran and Zhou (2018).

between $\bar{x}_i$ and $\alpha_i$. The test is not based on $\nu$, because the within estimator does not estimate that. Hence, as Ahn and Moon (2014) write, the test is not designed to directly detect correlation between $v_i$ and $\alpha_i$, causing limited power for that.

For example, consider the random effects model with $\mathbb{E}\{\bar{x}_i\alpha_i\} = 0$ but $\mathbb{E}\{v_i\alpha_i\} \neq 0$, so that the RE assumptions are violated by the latter. The question is whether the HT test reveals this. Wooldridge (2010, p. 331) shows in his formula (10.87) that the test actually tests whether $\alpha_i$ is uncorrelated with $\bar{x}_i$, after taking out the correlation of $\bar{x}_i$ with $v_i$. If $\bar{x}_i$ and $v_i$ are uncorrelated, (10.87) holds, so the HT test has no power. For cases where they are correlated, Ahn and Moon (2014) show limited power as well. We corroborate these results, as applying the HT test to the data from the Monte Carlo exercise of Appendix B.4.3 yields powers that are all below 28%. Hence, one should be careful with interpreting non-rejection of the HT test as convincing support for $\mathbb{E}\{v_i\alpha_i\} = 0$.

In contrast, our split analysis delivers estimates of $\alpha_i^g$, which *always* yield a test of our identifying constraint. Moreover, we have substantial power to detect $\mathbb{E}\{v_i\alpha_i\} \neq 0$, as our Monte Carlo study reveals.

The fourth difference concerns the estimator. To start with, suppose that both HT and our test do not reject. First, assume the HT test result is for a model where (the full vector) $\bar{x}_i$ is exogenous. Given the power issues just described, it is wise to allow part of $v_i$ to be endogenous, but then the strengths of the instruments in $\bar{x}_i$ matter. In contrast, given the power of our test, our non-rejection suggests taking a simpler model, without the FE and without potential weak-instrument issues. Second, assume $\bar{x}_i$ is left endogenous. Then $v_i$ must be treated as exogenous in HT. Here our approach is simpler and, if our null hypothesis indeed holds, yields a more efficient estimator.[7]

Suppose now the HT test and our test both reject. Both approaches fail to identify $\nu$. But our FE framework delivers estimated $\alpha_i^g$ with confidence band, which help to find omitted regressors. This has paid off, as in Section 5.5.2 the estimated FE reveal the relevance of the real effective exchange rate for explaining exports, giving the idea to extend the theoretical gravity model of exports.

Finally, suppose the HT test does not reject, but ours does. This can be due to the power differences described above, or because of weak instruments. However, the cause can also be that our constraint is stricter than the HT moment restrictions. This is the price we pay for the above advantages regarding testing and estimation. If the reason for the test difference is that our constraint is invalid and the two HT moment restrictions hold and instruments are strong enough, we fail to identify $\nu$, while HT succeed. Still, our estimated $\alpha_i^g$ help to find omitted regressors.

[7]The efficiency gain can be small and is not the focus of the paper. Still, to understand it, realize that here HT uses the within transformation, whereas by leaving out $\alpha_i$ we avoid that. This makes our $\hat{\beta}$ more efficient and, if $v_i$ is correlated with $x_{it}$, this also makes our $\hat{\nu}$ more efficient.

# 4 Untangling normalization

## 4.1 Learning from the estimated normalized fixed effects

So far, we have tested constraint (7) that all normalized FE $\alpha_i^g$ are zero. Whether it holds is an empirical question. If it does, the FE have revealed that the true value $\nu$ is identified. The FE can be left out, and there is no need for normalizing them anymore. However, if the constraint does not hold, we want to learn from the estimated $\alpha_i^g$.

How much we can learn depends on the precision of the estimated $\alpha_i^g$. That increases with the number of observations over time, $T$. Still, a moderate $T$ can already provide valuable insights. For example, 17 observations per FE deliver clear information on omitted variables in Section 5.5.[8]

Learning from the FE also depends on the normalization. This section introduces a particularly convenient one, untangling normalization. It is denoted by $u$, so the special case $g = u$. Importantly, untangling is not used for identifying the true value $\nu$. After all, the identification analysis was for the general normalization $g$, so that identification is not driven by a particular normalization — it is due to (7) holding. Moreover, we introduce untangling *after* testing (7), knowing that identification is rejected.

Untangling pins down $\alpha_i^u$. The key is that it does so in a way that is attractive for interpretation and finding omitted regressors. Hence, the FE framework delivers estimates and covariance matrix of the normalized FE, and untangling enables us to make good use of that benefit.

## 4.2 The idea and advantages of untangling

The idea of untangling normalization is to handle multicollinearity by making the (normalized) FE orthogonal to each other and to constant regressors, if any. We can now interpret the FE as deviations from both the other FE and the constant regressors. They have been untangled, and each FE-type targets a specific feature of the data.

The main aspects of untangling are as follows. First, consider $\alpha^u$ and $\alpha_i^u$ as an example, so that we need one normalization. Untangling normalization sets the mean of the $\alpha_i^u$ to zero. Now the untangled constant $\alpha^u$ captures the overall level, and the untangled country FE $\alpha_i^u$ is the country deviation from the overall level. Hence, both

[8] The literature on (small-sample) bias gives more support. For notational convenience, set $\nu^g = 0$. If the regressors are strictly exogenous regarding the error, LSDV is an unbiased estimator of the $\alpha_i^g$, irrespective of $T$. If, in addition, the error is normally distributed, the estimated $\alpha_i^g$ are also normal.

For alternative error assumptions, consider Fernández-Val and Weidner (2018), who allow for predetermined regressors and study linear and the most commonly used nonlinear models. They review the literature on large-$N$ and large-$T$ approximations and conclude that the order of the bias in the asymptotic approximation corresponds with the inverse of the number of observations per parameter. For the estimator of $\alpha_i^g$ this means $1/T$. Buddelmeyer et al. (2008) use simulations to study the bias, and for their settings the biases are fairly small for $T = 20$. Both papers also show how bias correction can further improve small-sample properties.

effects do not interfere with each other and are assigned to separate parameters, in a unique way. Section 4.3 discusses the details for all FE.

Of course, the normalization of $\alpha_i^u$ is well known. It goes back to Suits (1984). The other normalizations in this section generalize his idea. In addition, we contribute by presenting an encompassing framework for estimation and testing in Appendices A and B, and our approach can be extended to, for example, three-dimensional panels.

For the second aspect of untangling, note that each country-specific regressor in $v_i$ requires one additional normalization. Untangling sets the country FE orthogonal to the variables in $v_i$. Now the $\alpha_i^u$ capture what is left over after the explanation by $v_i$, thereby exploiting the information in constant regressors. The details are in Section 4.4.

Untangling offers several advantages. It eases interpretation and is unique. In contrast, in typically-used normalizations, such as zero normalization, the overall level and the country deviations are "tangled" into the FE, and the normalization depends on ad-hoc choices. Moreover, by focusing on one specific feature of the data and exploiting constant regressors, estimates of $\alpha_i^u$ contribute to finding potentially important omitted regressors. In other normalizations, such as those where $\nu^0 = 0$, the information in constant regressors is ignored. Finally, in untangling, the orthogonalization minimizes the sum of all squared differences between the observed totals $\alpha^u + \alpha_i^u + v_i'\nu^u$ and a linear combination of the variables in $v_i$, as in linear regression, which stabilizes the solution $\nu^u$. Zero normalizations where $\nu^0$ is free can lead to erratic $\nu^0$ due to noise in the observations $i$ involved. Section 5.5 illustrates some advantages.

## 4.3 Untangling fixed effects

We first introduce the untangling normalizations concerning the FE themselves. Note that $\alpha$ is the homogeneous type of the intercept fixed effects, so we want $\alpha^u$ to capture the overall intercept, so that we do not normalize it. Likewise, $\tau^u$ should capture the overall trend in the model, so we do not normalize that either.

**Country-specific effects**

To untangle the country FE from the common constant, we normalize the mean of $\alpha_i^u$ to zero, so that they capture the country deviations from $\alpha^u$. In formula,

$$\sum_i \alpha_i^u = 0. \tag{9}$$

**Country-specific trends**

Similar to $\alpha_i^u$, we normalize the mean of the country-trend FE $\tau_i^u$ to zero, so that the untangled $\tau_i^u \cdot t$ capture the country deviations from the common trend $\tau^u \cdot t$:

$$\sum_i \tau_i^u = 0. \tag{10}$$

**Time-specific effects**
Similar to $\alpha_i^u$, we normalize the mean of $\theta_t^u$ to zero, so that they are the time deviations from the overall intercept $\alpha^u$. In addition, time FE pick up the common trend. Because we already have $\tau^u \cdot t$ in the model, we orthogonalize the time FE to it. This ensures that $\theta_t^u$ is trendless and is the time deviation from the common trend. In formula,

$$\sum_t \theta_t^u \quad = 0 \tag{11}$$

$$\sum_t \theta_t^u \cdot t = 0. \tag{12}$$

## 4.4 Untangling to exploit constant regressors

**Country-specific regressors**
To clean country FE from the information in the $v_i$, we make the $\alpha_i^u$ orthogonal to the $k$-th regressor $v_i^k$ for all $k$. This gives the following $K_v$ normalizations

$$\sum_i \alpha_i^u v_i^k = 0. \tag{13}$$

This looks like $\mathbb{E}\{v_i \alpha_i\} = 0$, a key relation in existing methods such as RE, FEVD, and partly in HT, which treat $\alpha_i$ as random. But the difference between (13) and $\mathbb{E}\{v_i \alpha_i\} = 0$ is fundamental. We use (13) just as a normalization, so not for identifying the true value $\nu$ — for the latter we take constraint (7) and that does not rely on untangling as it can use any normalization $g$. In contrast, in the existing literature $\mathbb{E}\{v_i \alpha_i\} = 0$ is a moment condition, used to identify the true value $\nu$. The advantages of our approach regarding the testability of identification are in Section 3.3.[9]

**Time-specific regressors**
Similarly, we clean the time FE from the $w_t$, resulting in $K_w$ normalizations

$$\sum_t \theta_t^u w_t^k = 0. \tag{14}$$

[9] There are additional differences with FEVD. Our approach can more easily handle a broad set of FE and constant regressor configurations, and it yields estimates of all FE with standard errors to facilitate the search for new regressors. But even in a panel with $\alpha_i$ effects only and for the choice $g = u$, our approach improves on FEVD, as follows. In this special case, our point estimates for $v_i$ and $x_{it}$ so far are the same as those of FEVD. However, we recognize that the estimate for $v_i$ concerns the pseudo-true value $\nu^u$. We then test $\alpha_i^u = 0$ for all $i$ and, if rejected, we admit that we have no estimate of the true value $\nu$. If our test does not reject, we remove the FE and realize an efficiency gain, as explained in footnote 7, realizing that FEVD equals HT for endogenous $\bar{x}_i$ and exogenous $v_i$. Moreover, we have a motivation for the claim that we have identified $\nu$, while FEVD views the estimate for $v_i$ as one of $\nu$, without a test, so by assumption.

### 4.5 Untangling in matrix form

Normalizations (9)-(14) can be expressed in matrix form, giving a special case of $N^g$ in (3). That yields $N^u$ below. It fulfills rank requirement (5). As there are no normalizations on $\alpha$, $\tau$, $\nu$, and $\omega$, the corresponding columns in $N^u$ are zero.

$$\begin{array}{rcl} & \begin{array}{ccccccc} \alpha & \alpha_1 ... \alpha_N & \tau & \tau_1 ... \tau_N & \theta_1 ... \theta_T & \nu' & \omega' \end{array} & \text{Row implements:} \\ N^u = & \left[\begin{array}{ccccccc} 0 & 1...1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1...1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1...1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1...T & 0 & 0 \\ 0 & v_1 ... v_N & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_1 ... w_T & 0 & 0 \end{array}\right] & \begin{array}{l} \sum_i \alpha_i^u = 0 \\ \sum_i \tau_i^u = 0 \\ \sum_t \theta_t^u = 0 \\ \sum_t \theta_t^u \cdot t = 0 \\ \sum_i \alpha_i^u v_i^k = 0 \\ \sum_t \theta_t^u w_t^k = 0. \end{array} \end{array} \tag{15}$$

## 5 Application: the gravity model of trade

Many models contain parameters that one thinks are unidentifiable due to added FE. Consider the gravity model of trade.

### 5.1 The gravity model and the identification problem

The gravity model says that exports from one country to another depend positively on the exporting and importing countries' GDPs and negatively on distance between the countries. Distance can be both physical and economic distance, such as trade costs.

Anderson and Van Wincoop (2003) show it is important to include multilateral resistance terms for the importer and exporter to avoid estimation bias. It has been difficult, however, to find economic variables that capture these terms.

If gravity models use bilateral data over time, one could nevertheless control for the multilateral resistances by country-time FE. But then the impacts of country- and time-specific variables, such as exporter and importer GDP, are not identified. This is a well-known problem (Head and Mayer (2014)).

Our method can help for both problems. We can test whether FE matter after exploiting the information in country- and time-specific observables. If FE do not matter, they can be left out and we estimate the previously unidentified parameters. That would also suggest economic variables underlying the multilateral resistance terms.

### 5.2 Model specification

Consider exports from country $i$ to the US in year $t$. Taking one importer is for simplicity and to stay within the *it*-setting of previous sections; see Klaassen and Teulings

(2017) for a three-dimensional application. The model specifies

$$\begin{aligned} exp_{iUSt} = & \beta_1 gdp_{it} + \beta_2 reer_{it} + \omega_1 gdp_{USt} + \omega_2 gdp_{Wt} + \omega_3 reer_{USt} \\ & + \alpha + \alpha_i + \tau \cdot t + \tau_i \cdot t + \theta_t + \varepsilon_{it}, \end{aligned} \tag{16}$$

where $exp_{iUSt}$ represents real exports from country $i$ to the US, and $gdp_{it}$, $gdp_{USt}$ and $gdp_{Wt}$ are real GDP of country $i$, the US, and the world, respectively, all in constant dollars. Moreover, $reer_{it}$ and $reer_{USt}$ are the real effective exchange rates (REER) of country $i$ and the US, respectively, where exchange rates are defined as home currency units per unit of foreign currency, so an increase in REER means a depreciation. Using nominal variables does not affect the main results. All variables are in log. We thus have two $it$-regressors, $gdp_{it}$ and $reer_{it}$, and three $t$-regressors, $gdp_{USt}$, $gdp_{Wt}$, and $reer_{USt}$, so the latter are the constant regressors.[10]

The GDP variables are suggested by the common gravity model. The theory in Klaassen and Teulings (2017) proposes adding real exchange rates, both the bilateral rate $rer_{iUSt}$ and the partners' REERs, here $reer_{it}$ and $reer_{USt}$. Because triangular arbitrage implies $rer_{iUSt} = reer_{it} - reer_{USt}$, only the REERs are included as regressors.

We add the general set of FE from the gravity literature. Note that it is less standard to include $\tau_i \cdot t$. However, Bun and Klaassen (2007) and Baier et al. (2014) confirm the importance of adding this to account for trends in exports not explained by the regressors, as is often done in the time-series literature.

The model includes lags of all regressors as, for example, traders often entered into contracts in previous periods to export goods in period $t$, based on export determinants back then. Two lags turn out to be sufficient, and we add them in the form of first differences, for example, $\Delta gdp_{it}$ and $\Delta gdp_{i,t-1}$. We focus on the long-run effects, that is, the parameters of the level regressors. The results for the first differences do not alter our conclusions, and we ignore them in (16) for simplicity of exposition.

The error term $\varepsilon_{it}$ has mean zero conditional on the regressors in all times. We thus ignore feedback from bilateral exports to GDPs and REERs, which is in line with the gravity literature and seems reasonable given that bilateral exports are a limited fraction of total exports and thus GDP and that exchange rates are mainly driven by financial variables. The Wooldridge (2010, p. 325) test for strict exogeneity supports this, as leads of regressors have insignificant impacts. The error term is allowed to be

[10] The $gdp_{Wt}$ regressor resembles a Mundlak (1978) term, that is, an average of $gdp_{it}$ over $i$. Footnote 6 sets out the differences between Mundlak's and our approaches, which we can now illustrate. First, Mundlak uses averages as auxiliary regressors to try to control for the correlation between his random effect and his $it$-regressors; there are no constant regressors. Instead, we have constant regressors, which are not auxiliary and are motivated by economic theory. Second, $gdp_{USt}$ would be in Mundlak's time effect, creating large correlation with $gdp_{it}$ and its average, invalidating his random effects assumption. In contrast, we treat the time effect as fixed instead of random, thereby recognizing its correlation with the $it$- and constant regressors. And we can test instead of assume identification of the true value.

heteroscedastic and serially correlated.

We estimate the model using LSDV and the indirect approach of Appendix A.1. This suffices to illustrate our method. Our main results are robust to different specifications, such as omitting $\tau_i \cdot t$, accounting for non-stationarity and cointegration, and a multiplicative approach estimated by Gamma and Poisson pseudo maximum likelihood (GPML and PPML), following Silva and Tenreyro (2006), as Appendix C shows.

### 5.3 Data

The data concern $N = 17$ countries, namely the EU-15 countries except for Belgium and Luxembourg, expanded with Canada, Japan, Norway and Switzerland. The sample is from 1979-2011 ($T = 33$), resulting in 561 observations.

We use monthly nominal export data from the IMF Direction of trade statistics (DOTS) and convert them back into home currency using the monthly dollar exchange rate from the International Financial Statistics (IFS) of the IMF. We then sum to get yearly values and divide by the home export price index from the European Commission AMECO database (the base year for all data is 2010; rebasing does not affect our results). We divide by the home PPP of the dollar from the OECD Economic Outlook to obtain exports in constant dollars.

Nominal yearly GDP is from AMECO, and we use the AMECO exchange rate to express it in national currency. We then divide by the AMECO GDP deflator and the home PPP of the US dollar to get GDP in constant US dollars. West-German data is used as a proxy for Germany before 1991. Real world GDP in US dollars is from the OECD Economic Outlook.

Finally, we use consumer-price-based monthly REER data from the Bank for International Settlements (BIS), construct yearly averages, and invert.

### 5.4 Identifying the true value

This section illustrates the first contribution of the paper, our test to identify the true value of the impact of constant regressors, here $gdp_{USt}$, $gdp_{Wt}$, and $reer_{USt}$. The null hypothesis is $\theta_t^g = 0$ for all years $t$, which is the time equivalent of (7).

As indicated by the $g$-superscript, we can take any normalization. Following the advice in Section 3.2, and realizing that the model has a constant and trend and that the three $t$-regressors are included with two lags, we normalize 11 out of the 31 time FE. We also need one normalization on the country FE and one on the country-trend FE. What specific normalization we take is irrelevant for the identification tests. In particular, untangling is irrelevant here.

We now estimate the model and test for identification. We take Wald tests, based on the motivations in Appendix B.1, the supportive size and power results in Monte

Carlo Appendix B.4, and our panel sizes. The diagnostic test that all normalized time FE are zero is 22.77 with 20 degrees of freedom, implying a $p$-value of 0.30.[11] The sensitivity test is 0.67 with $p$-value 0.72, which shows that leaving out the time FE does not significantly alter the estimated impacts of $gdp_{it}$ and $reer_{it}$, so this signals no evidence of omitted variable bias. Hence, the tests do not reject identification of the impacts of the three constant regressors. In Section 5.5.2 we will argue that this is most likely not due to a lack of power, as already suggested by the Monte Carlo results in Appendix B.4.

Table 1: Estimation results for $exp_{iUSt}$ based on model (16)

| | Without $reer$s | | | With both $reer$s | | |
|---|---|---|---|---|---|---|
| Specification | 1 | 2 | 3 | 4 | 5 | 6 |
| Included | $\theta_t$ | $\theta_t$ | $\theta_t$ | $\theta_t$ | $\theta_t$,DL | DL |
| Normalization | 0 | 0,$u$0 | $u$ | $u$ | $u$ | – |
| Figure of FE | 1a | 1b,1c | 1d=2a | 2b | 2c | – |
| $gdp_{it}$ | 0.60 * | 0.60 * | 0.60 * | 0.80 * | 1.12 * | 1.13 * |
| | (0.23) | (0.23) | (0.23) | (0.24) | (0.25) | (0.24) |
| $reer_{it}$ | | | | 0.42 * | 0.66 * | 0.68 * |
| | | | | (0.12) | (0.15) | (0.14) |
| $gdp_{USt}$ | | | 3.43 * | 3.03 * | 2.35 * | 2.30 * |
| | | | (0.30) | (0.29) | (0.39) | (0.37) |
| $gdp_{Wt}$ | | | −3.59 * | −1.84 * | −1.18 | −1.23 |
| | | | (0.41) | (0.45) | (0.73) | (0.72) |
| $reer_{USt}$ | | | | −1.02 * | −1.02 * | −1.01 * |
| | | | | (0.10) | (0.13) | (0.13) |
| Wald tests | | | | | | |
| $\theta_t^g = 0$ | 770.04 * | 545.14 * | 222.57 * | 85.31 * | 22.77 | – |
| | [0.00] | [0.00] | [0.00] | [0.00] | [0.30] | |
| $\beta = \beta\|_{\theta_t^g=0}$ | 2.50 | 80.69 * | 1.34 | 9.93 * | 0.67 | – |
| | [0.11] | [0.00] | [0.25] | [0.01] | [0.72] | |
| $R^2_\theta$ | 0 | 0 | 0.65 | 0.91 | 0.98 | – |

Static models have 561 observations and distributed lag (DL) models 527. DL models have two lags of each regressor; we display the long-run effects. The normalizations used in specifications 1 and 2 are made explicit in the corresponding FE figures.
The first Wald is the diagnostic test motivated in Appendix B.2 and tests 33, 31, 29, 28, 20 independent constraints for specifications 1-5. The second Wald, labeled $\beta = \beta|_{\theta_t^g=0}$, is the sensitivity test of Appendix B.3, showing how sensitive the estimator for $\beta$ of the level variables is to setting $\theta_t^g = 0$, which concerns 1, 1, 1, 2, 2 constraints.
$R^2_\theta$ is the fraction of the variance of the untangled time FE from a model without $t$-regressors that is explained once (detrended) $t$-regressors are included.
Standard errors are between brackets and they are based on Newey and West (1987, 1994), which gives three lags. $p$-values are in square brackets. * indicates significance at the 5% level, the level we use throughout the paper.

We conclude that three $t$-regressors and their lags have made time FE redundant, so we can safely leave out the FE. Estimates for the impacts of the $t$-regressors thus

[11] The standardized Wald test, discussed in footnote 19, yields a similar $p$-value of 0.33.

reflect their true values $\omega$ instead of only pseudo-true values. This is remarkable, because identifying such true values has been a notorious problem in the economics literature, not only the literature on gravity models. Moreover, the result supports gravity theory in the sense that a model with $reer_{USt}$ here no longer needs time FE to control for omitted $t$-regressors, and that $reer_{USt}$ helps to explain US multilateral resistance, a variable that has been considered unobservable in the gravity literature. The rightmost column in Table 1 shows the resulting estimates. All signs can be explained within the theoretical gravity model.

## 5.5 Untangling

The second contribution of the paper addresses the question what to do if the null hypothesis $\theta_t^g = 0$ is rejected. We create such a situation within the framework just used.[12] More specifically, we start from the empirical gravity model that is now common practice (Head and Mayer (2014)), which is (16) without the REERs ($\beta_2 = \omega_3 = 0$). Note that the latter is a constraint, no normalization. There are no lagged regressors.[13]

We thus deliberately take a step back by reducing the explanatory part of the model. This allows us to illustrate how an empirical researcher, starting from a well-known model, can use untangling normalization of Section 4 for better interpretation and thereby improve the specification. At the same time, we have untangling *after* the test outcome of not rejecting identification in Section 5.4, which highlights that the test in no way depends on untangling normalization, a key point to be realized.

### 5.5.1 Untangling helps interpretation

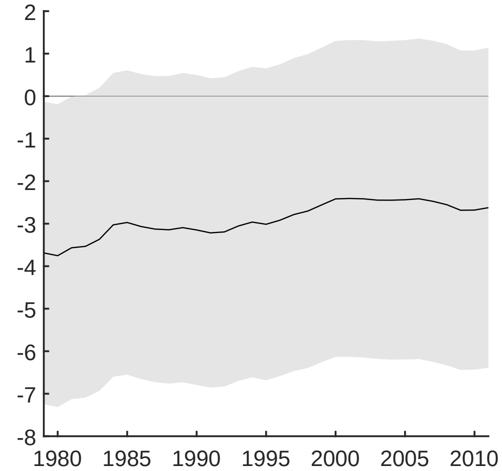
The first advantage of untangling is that it facilitates interpretation. This section illustrates that by varying the normalization, while keeping the rest of the model the same. Hence, there is one estimation, and then the estimates are simply transformed to fulfill other normalizations. All normalizations have one on the country FE and one on the country-trend FE, but we focus on the remaining four choices, two due to the time FE and two regarding the constant regressors $gdp_{USt}$ and $gdp_{Wt}$.

One typically chooses a zero normalization. We study two such choices, both having $\omega^0 = 0$. Figure 1a normalizes $\alpha^0 = \tau^0 = 0$ and shows the estimated normalized time FE $\theta_t^0$ and confidence band. Their mean is nonzero, reflecting that they are affected by the overall means of the dependent and explanatory variables. They exhibit some variation over time, but this seems a minor feature.

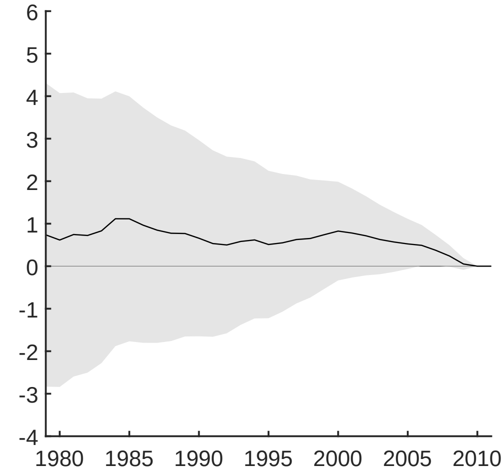
[12] Another possibility is to take other data. Appendix C exemplifies that by adding pre-1979 data.

[13] Throughout this section we consider the number of countries $N = 17$ to be sufficiently large to rely on asymptotic normality when discussing the estimated time FE, stimulated by our own simulation results underlying Appendix B.4 and the literature reported in footnote 8.

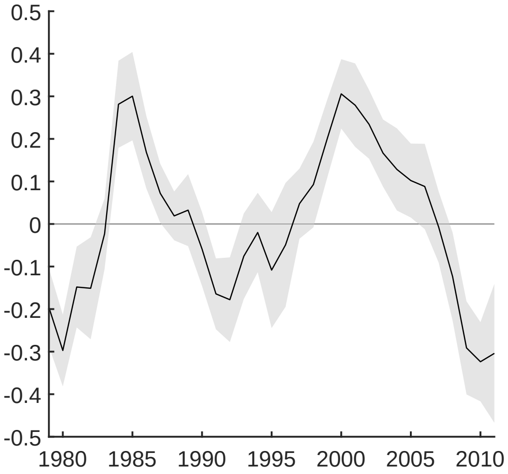
Figure 1: From zero-normalized to untangled time FE in the common gravity model



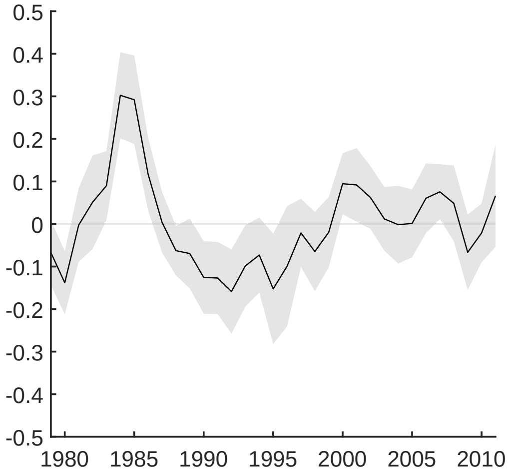

(a) $\theta_t^0$ if $\alpha^0 = \tau^0 = 0$ and $\omega^0 = 0$




(b) $\theta_t^0$ if $\theta_{T-1}^0 = \theta_T^0 = 0$ and $\omega^0 = 0$




(c) Untangled $\theta_t^{u0}$, but still $\omega^{u0} = 0$




(d) Untangled $\theta_t^u$

All bands around the time FE in the paper are 95% simultaneous confidence bands. They are sup-t bands, following the suggestion by Montiel Olea and Plagborg-Møller (2019). The sup-t band is the narrowest band in the class of confidence bands that scale up the pointwise band (which simply combines the confidence intervals of the underlying parameters) by one parameter to guarantee the required simultaneous coverage. That parameter is about 1.0 for the bands based on zero normalization and 1.6 based on untangling.

Figure 1b normalizes $\theta^0_{T-1} = \theta^0_T = 0$, giving a positive mean and a shrinking confidence band. The changes compared to Figure 1a exemplify the well-known effects of different zero normalizations on the FE estimates, which hamper their interpretation.

It is not yet clear how important the variation over time is, as that may be dominated by the constant or the trend. To have this curvature information visible right away, it would have been appealing to let the normalization split off the level and trend information from the time FE. That is what untangling normalization does, by (11) and (12), in a unique way.

Figure 1c shows the results, where for the moment we do not yet exploit the constant regressors. Hence, this is a combination of untangling and zero normalization, indicated by the superscript $u0$. These FE are more informative and easier to interpret than the zero-normalized ones. Note the economic downturn in the early 1990s, the dot-com bubble, and the recent financial crisis. Footnote 15 will give a more complete analysis of the correspondence between untangled FE and the business cycle.

In addition, untangling has resulted in a more informative confidence band. Untangling can thus better capitalize on an advantage of a fixed instead of random effects approach, which is that the former delivers insights into the accuracy of all estimated time effects. Overall, the estimated $\theta^{u0}_t$ indicate the model misses export determinants.

Figure 1d shows the FE from untangling normalization, so we now exploit the constant regressors $gdp_{USt}$ and $gdp_{Wt}$, using (14). Untangling shrinks and cleans the FE, so it better shows when exports deviate from what the common gravity model explains.

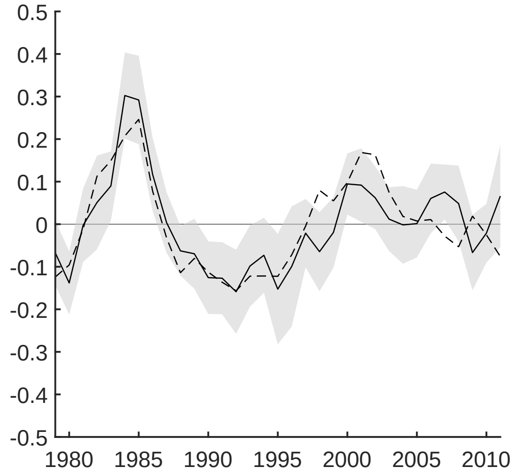
This section has only varied the normalization. Section 3.2 has shown that normalizations can matter for the restrictiveness of our identifying constraint $\theta^g_t = 0$ for all $t$. Hence, as a side issue, we now calculate our Wald tests for the normalizations in Figures 1a-1d. The bottom parts of columns 1-3 in Table 1 show the results.

The drop in the diagnostic Wald from Figure 1a to 1b (column 1 to 2) reflects that leaving the overall constant and trend free makes the constraint more realistic, confirming our advice in Section 3.2 to normalize parameters that also appear in the null hypothesis, here $\theta^g_t$.[14] How the $\theta^g_t$ are normalized does not matter, which is illustrated by the equal Walds for Figures 1b and 1c.

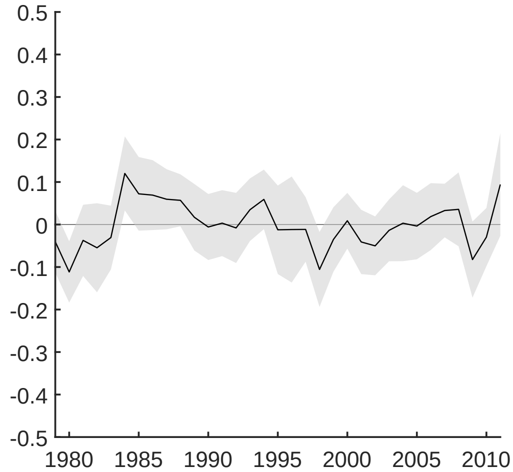
Moving to Figure 1d and thus exploiting the constant regressors reduces the Walds. This is again in line with our advice and reflects that the constraint concerns the FE left over after accounting for constant regressors. In sum, untangling normalization, or any other approach with the normalizations on $\theta^g_t$ only, provides the cleanest indication of how close one is to identification.

[14] Our advice also avoids chance dependence of the sensitivity Wald. That is, the small value for Figure 1a is because the additional constant and trend restrictions make that the estimated $\beta_1$ and thus the test depend on the coincidental level and scale of the variables. Our advice avoids this.

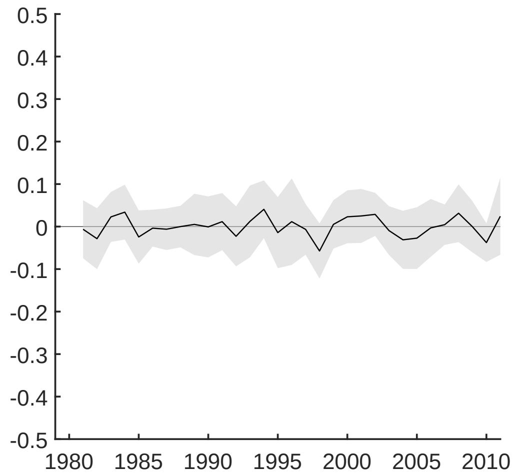
Figure 2: Untangled time FE $\theta_t^u$ when adding $reer$s to the common gravity model




(a) Common gravity model 3; dashed $-reer_{USt}$




(b) Gravity model with $reer$s: model 4




(c) Gravity model with $reer$s and lags: model 5

### 5.5.2 Untangling helps to find omitted variables

Figure 2a replicates 1d but adds minus US REER as a dashed line (after detrending, demeaning, and scaling). The resemblance with the FE is striking. For example, consider the eighties, where the dollar bubble stimulated and then hampered exports to the US. Hence, the US REER seems important in explaining time FE, illustrating the second advantage of untangling, that it can reveal omitted regressors. Of course, the relevance of exchange rates for exports is well known. But gravity theory typically abstracts from exchange rates. Klaassen and Teulings (2017) thus extend gravity theory and confirm that $reer_{it}$ and $reer_{USt}$ matter. Hence, here untangling not only reveals $t$-regressors, but also indirectly an $it$-regressor.

We thus add $reer_{it}$ and $reer_{USt}$ to the model, that is, we leave $\beta_2$ and $\omega_3$ free; see column 4 in Table 1.[15] Compared to column 3, we observe a notable change in the estimated impact of $gdp_{Wt}$. This indicates that adding the REERs substantially mitigates omitted variable bias.

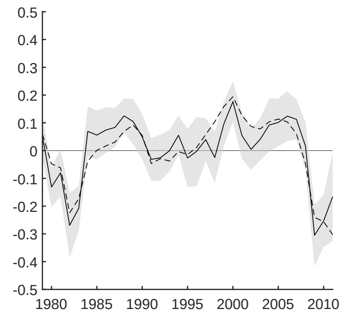
Figure 2b presents the estimated FE $\theta_t^u$. There is again a large reduction of the FE, and they are now close to zero. Taken together, adding three $t$-regressors has reduced the FE substantially, from those in Figure 1c to Figure 2b. We can explain almost all of the $T = 33$ time FE by just three variables. This can be quantified by $R_\theta^2 = 91\%$, which is defined in the note to Table 1.

We next add two lags of all regressors, in first-difference form, as in Section 5.2. The lags of $it$-regressors will take some noise out of $\theta_t^u$, and lagged $t$-regressors further explain $\theta_t^u$. We thus again take a distributed lag (DL) model and still allow for unrestricted serial correlation in the error. The estimates are in Table 1 specification 5.

Figure 2c shows the estimated $\theta_t^u$. They are close to zero and $R_\theta^2 = 98\%$. The confidence band includes zero for all $t$. This signals that the Wald insignificances in column 5 are not due to low power caused by aggregating the information on all FE into a single statistic.[16] In fact, the band in Figure 2b is close to zero and comes with clear rejections of both Wald tests, namely 85.31 ($p$-value 0.00) and 9.93 (0.01). This indicates that the tests have power, in line with the Monte Carlo results in Appendix B.4.

Finally, column 6 leaves out the FE to increase efficiency. That lowers standard

[15] An alternative way to show that untangling can help researchers to find omitted regressors is by taking this more general model 4 and then leaving out the t-regressors one by one. Leaving out US GDP gives the estimated untangled FE in the solid line in the figure on the right, and the dashed line is the omitted regressor itself. There is a strong resemblance. This again shows that untangling can reveal omitted regressors. If we redo this for world GDP, the resemblance is much weaker (figure not reported), but for US REER it is strong (the figure is not reported but is almost equal to Figure 2a).




[16] The Wald tests have the same values as in Section 5.4, confirming that untangling is irrelevant for the identification tests.

errors by about 1% of the point estimates.

### 5.5.3 Complementing Hausman-Taylor

Our approach can be used on its own, but it can also assist the Hausman and Taylor (1981) method set out in Section 3.3. The current section illustrates that. We again start from the common gravity model, so (16) without REERs, for which our approach gives the results in column 3 of Table 1. For HT we treat $\theta_t$ as random.

HT requires specifying which regressors are exogenous regarding $\theta_t$. If $gdp_{USt}$, say, is endogenous, business cycle co-movement across countries makes it likely that also $gdp_{it}$ and thus its cross-sectional average $\overline{gdp}_t$ and $gdp_{Wt}$ are endogenous. This gives two endogenous constant regressors and no instrument. To avoid violating the order condition, one must view all regressors as exogenous, giving a random effects model.

The HT test, here the Hausman test of random versus fixed effects, has $p$-value 0.70. We must be careful with interpreting this as evidence that the GDPs are truly exogenous, given the power characteristics discussed in Section 3.3. More specifically, Wooldridge (2010, p. 331) implies that the test projects out $gdp_{USt}$ and $gdp_{Wt}$ (and country dummies and trends) from $\overline{gdp}_t$ and then tests whether the remainder is uncorrelated with $\theta_t$. As US and world GDP capture a lot of the country-average GDP, this remainder may indeed be uncorrelated with $\theta_t$. But this holds even if both constant regressors are endogenous. Hence, the high $p$-value does not tell us whether the constant regressors are exogenous.

Using our approach, we plot the estimated untangled FE. As shown in Section 5.5.2, this reveals that $reer_{it}$ and $reer_{USt}$ have explantory power. Because REERs reflect competitiveness, which matters for GDP, omitting them in the previous specification may have caused endogeneity of GDP variables there. This exemplifies how our method can assist the HT test.

Our method also suggests adding $reer_{it}$ and $reer_{USt}$ as regressors, which offers two ways to continue with HT. First, suppose one allows $reer_{USt}$ to be endogenous. Because the US is an important trading partner of all countries $i$, there will be a negative correlation between $reer_{USt}$ and $reer_{it}$, so that $\overline{reer}_t$ is also endogenous. This means an exactly identified HT model, where $\overline{gdp}_t$ is an instrument for $reer_{USt}$. One cannot test this specification. The Stata HT estimates and standard errors for the constant regressors are: for $gdp_{USt}$ 2.30 (0.57), for $gdp_{Wt}$ 1.39 (4.12), and for $reer_{USt}$ -2.25 (1.62). Particularly the latter two estimates differ from ours and the standard errors are much larger, as our results from column 5 are 2.35 (0.39), -1.18 (0.73) and -1.02 (0.13), respectively. The reason is that $\overline{gdp}_t$ is a weak instrument for $reer_{USt}$.

Alternatively, one can impose that $reer_{USt}$ and $reer_{it}$ are exogenous. This gives a random effects model. The HT test does not reject, with $p$-value 0.93. As before, one must be careful with interpreting this as support for the specification. The estimation results are similar to ours.

We thus end up with two HT specifications. Which one to choose? Our non-rejection of $\theta_t^g = 0$ supports the second specification. This again illustrates that our approach can help out HT. In fact, our non-rejection motivates us to simply leave out all $\theta_t$, so that there is no need for HT here.

# 6 Conclusion

We have shown that in fixed-effect models the true value of the ceteris paribus impact of constant regressors is identified if normalized FE are zero, whatever the normalization. This constraint is testable. If it holds, our approach resolves a notorious problem in the literature. It can help researchers to motivate leaving out FE. Moreover, it is an alternative to Hausman and Taylor (1981), and it can help out an HT analysis.

We have applied our method to a panel gravity model for exports to the US. With only three $t$-regressors — US GDP, world GDP and US REER — the year FE become redundant, so that we have identified the true values of their impacts, even though that is typically considered beyond reach. This also supports gravity theory in the sense that a model with US REER here no longer needs time FE to control for omitted $t$-regressors, and that US REER helps to explain US multilateral resistance, a variable that has been considered unobservable in the gravity literature.

Our second contribution concerns the case where the constraint does not hold. For that, we have introduced untangling normalization. It disentangles FE-types from each other and from constant regressors, which eases the interpretation of the normalized FE. This also helps researchers to find omitted variables.

The gravity application has illustrated how untangling can visualize the information in estimated normalized FE, and how untangling has revealed the relevance of three $t$-regressors. The business cycle pattern in exports is well known. But untangling has also shown the importance of the US REER. We thus recommend giving exchange rates a more prominent role in gravity theory.

This paper has used parameter homogeneity to simplify the exposition, and that has turned out to be sufficient. In future work, one may want to allow for heterogeneity, to further shrink the FE and make our identifying constraint more realistic. Moreover, as the generalization to $ijt$-panels is straightforward, untangling can facilitate studies of financial or trade relations involving many sectors $j$. Finally, untangling illustrates the value of information in FE, which may stimulate further research on their estimation.

## References

AHN, S. C. AND H. R. MOON (2014): "Large-N and Large-T Properties of Panel Data Estimators and the Hausman Test," in *Festschrift in Honor of Peter Schmidt*, ed. by R. C. Sickles and W. C. Horrace, Springer New York, 219–258.

ALLISON, P. (2009): *Fixed Effects Regression Models*, SAGE Publications, Inc.

ANDERSON, J. E. AND E. VAN WINCOOP (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *The American Economic Review*, 93, 170–192.

BAIER, S. L., J. H. BERGSTRAND, AND M. FENG (2014): "Economic integration agreements and the margins of international trade," *Journal of International Economics*, 93, 339–350.

BASU, D. (2020): "Bias of OLS Estimators due to Exclusion of Relevant Variables and Inclusion of Irrelevant Variables," *Oxford Bulletin of Economics and Statistics*, 82, 209–234.

BEETSMA, R., F. KLAASSEN, W. ROMP, AND R. VAN MAURIK (2020): "What drives pension reforms in the OECD?" *Economic Policy*, 35, 357–402.

BREUSCH, T., M. B. WARD, H. T. M. NGUYEN, AND T. KOMPAS (2011): "On the Fixed-Effects Vector Decomposition," *Political Analysis*, 19, 123–134.

BUDDELMEYER, H., P. H. JENSEN, U. OGUZOGLU, AND E. WEBSTER (2008): "Fixed Effects Bias in Panel Data Estimators," IZA Discussion Papers 3487, Institute for the Study of Labor (IZA).

BUN, M. J. G. AND F. KLAASSEN (2007): "The Euro Effect on Trade is not as Large as Commonly Thought," *Oxford Bulletin of Economics and Statistics*, 69, 473–496.

DONALD, S., G. IMBENS, AND W. NEWEY (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117, 55–93.

EVANS, G. B. A. AND N. E. SAVIN (1982): "Conflict among the Criteria Revisited; The W, LR and LM Tests," *Econometrica*, 50, 737–748.

FERNÁNDEZ-VAL, I. AND M. WEIDNER (2018): "Fixed Effects Estimation of Large-T Panel Data Models," *Annual Review of Economics*, 10, 109–138.

GHYSELS, E., J. B. HILL, AND K. MOTEGI (2020): "Testing a large set of zero restrictions in regression models, with an application to mixed frequency Granger causality," *Journal of Econometrics*, 218, 633 – 654.

HAUSMAN, J. A. AND W. E. TAYLOR (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377–1398.

HEAD, K. AND T. MAYER (2014): *Gravity Equations: Workhorse,Toolkit, and Cookbook*, Elsevier, vol. 4 of *Handbook of International Economics*, chap. 3, 131–195.

HONORÉ, B. E. AND M. KESINA (2017): "Estimation of Some Nonlinear Panel Data Models With Both Time-Varying and Time-Invariant Explanatory Variables," *Journal of Business & Economic Statistics*, 35, 543–558.

Kiviet, J. F. (2012): "Monte Carlo Simulation for Econometricians," *Foundations and Trends(R) in Econometrics*, 5, 1–181.

Klaassen, F. and R. Teulings (2017): "Untangling Real Gravity," Tinbergen Institute Discussion Papers 17-121/VI, Tinbergen Institute.

Kripfganz, S. and C. Schwarz (2019): "Estimation of linear dynamic panel data models with time-invariant regressors," *Journal of Applied Econometrics*, 34, 526–546.

Lu, X. and L. Su (2020): "Determining individual or time effects in panel data models," *Journal of Econometrics*, 215, 60–83.

Magnus, J. R. and A. L. Vasnev (2007): "Local sensitivity and diagnostic tests," *Econometrics Journal*, 10, 166–192.

Mark, N. C. and D. Sul (2003): "Cointegration Vector Estimation by Panel DOLS and Long-run Money Demand," *Oxford Bulletin of Economics and Statistics*, 65, 655–680.

Montiel Olea, J. L. and M. Plagborg-Møller (2019): "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 34, 1–17.

Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

——— (1994): "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 61, 631–653.

Pesaran, M. H. and Q. Zhou (2018): "Estimation of time-invariant effects in static panel data models," *Econometric Reviews*, 37, 1137–1171.

Plümper, T. and V. E. Troeger (2007): "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects," *Political Analysis*, 15, 124–139.

——— (2011): "Fixed-Effects Vector Decomposition: Properties, Reliability, and Instruments," *Political Analysis*, 19, 147–164.

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

Suits, D. B. (1984): "Dummy Variables: Mechanics V. Interpretation," *The Review of Economics and Statistics*, 66, 177–180.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge MA: The MIT Press.

# Appendix

## A Estimation of normalized parameters

It is important to realize that this appendix (on estimation) and the next one (on testing) use the general setting of the $g$-normalization. Hence, both do *not* impose the untangling normalization of Section 4. The latter is just a special case that results by taking (15) for $N^g$, which happens to simplify the formulas.

Consider mean equation (2). Normalizations only affect the lower-dimensional part of the regressors, so the focus in this appendix is on $D\delta + Z\gamma$. We introduce two methods of estimation, an indirect and a direct approach.

### A.1 Indirect estimation: renormalizing zero-normalized estimates

The indirect estimation method consists of two steps. It first estimates (2) using a zero normalization and then renormalizes the estimates into the desired $g$ normalization.

#### A.1.1 Estimating zero-normalized parameters

Using zero normalization in the estimation step is convenient because the parameters can then be estimated in a standard way, as one just omits the dummies and regressors corresponding to the normalized parameters from the regressor matrix. After estimation, add zeros to the estimated parameter vector and rows and columns of zeros to the estimated covariance matrix corresponding to the zero-normalized parameters. We now have estimates for $\delta^0$ and $\gamma^0$ and the corresponding full covariance matrix.

#### A.1.2 Renormalizing the estimates

The second step renormalizes $\delta^0$ and $\gamma^0$ into the general-normalized parameters $\delta^g$ and $\gamma^g$, that is, it redistributes the sum $D\delta^0 + Z\gamma^0$ over $D\delta^g$ and $Z\gamma^g$. For both normalizations, system (3)-(4) holds, so we obtain

$$\begin{bmatrix} D & Z \\ & N^g \end{bmatrix} \begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = \begin{bmatrix} D & Z \\ & N^0 \end{bmatrix} \begin{bmatrix} \delta^0 \\ \gamma^0 \end{bmatrix}. \tag{A.1}$$

To solve for $\delta^g$ and $\gamma^g$, define the matrix on the left by $R^g = [D, Z; N^g]$ and premultiply (A.1) with $R^{g\prime}$. Because of (5), $R^g$ has full column rank, so we obtain

$$\begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = R^{0g} \begin{bmatrix} \delta^0 \\ \gamma^0 \end{bmatrix}, \tag{A.2}$$

where $R^{0g} = \left(R^{g\prime} R^g\right)^{-1} R^{g\prime} R^0$ is the renormalization matrix that converts the zero into the general normalization, and the matrix only consists of observables.[17]

Hence, to obtain the estimates and estimated variance for the $g$-normalized parameters, we take both for the zero-normalized parameters and apply (A.2). Therefore, no additional estimation or standard error correction is needed.

## A.2 Direct estimation: incorporating normalization into regressors

The second estimation method transforms the regressors in the $(D\delta + Z\gamma)$-part of (2) such that they incorporate the normalization and that the regressor matrix becomes full column rank. Now we can directly estimate the transformed model and obtain estimates of the normalized parameters. This direct approach is also useful for estimating models under constraints, which we will need in Appendix B. Note that leaving out regressors when applying zero normalization is a special case of this approach.

For a set of general-normalized parameters $\delta^g$ and $\gamma^g$, we first split off some resultant parameters by writing them as a function of the free parameters based on the normalization. This can be done as follows.

Because $N^g$ has full row rank $m_d + m_z$, we can take $m_d + m_z$ independent columns of $N^g$ and collect them in $N_r^g$, which is thus invertible. Let $P$ be the column permutation matrix that forms $N_r^g$ and puts the remaining columns in $N_f^g$, while keeping the initial column ordering in both $N_r^g$ and $N_f^g$. We split $\delta^g$ and $\gamma^g$ accordingly. That is,

$$N^g P = \begin{bmatrix} N_f^g & N_r^g \end{bmatrix} \text{ and } P' \begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = \begin{bmatrix} \delta_f^g \\ \gamma_f^g \\ \delta_r^g \\ \gamma_r^g \end{bmatrix}. \tag{A.3}$$

Hence, the choice of $P$ determines what are the free and what are the resultant parameters, but $P$ does not affect the normalization itself.

Using normalization description (3), writing $N^g$ as $N^g P P'$, and using (A.3) gives

$$\begin{bmatrix} \delta_r^g \\ \gamma_r^g \end{bmatrix} = -N_r^{g-1} N_f^g \begin{bmatrix} \delta_f^g \\ \gamma_f^g \end{bmatrix}. \tag{A.4}$$

Thus the full parameter vector is a function of the free parameters:

$$\begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = F^g \begin{bmatrix} \delta_f^g \\ \gamma_f^g \end{bmatrix}, \tag{A.5}$$

[17] Computing $R^{0g}$ depends on multiplications involving $R^g$ and $R^0$, which have many rows. This can be simplified. First, select all $K_d + K_z$ independent rows in $R^g$ by Gaussian elimination, making the resulting $\widetilde{R}^g$ a square matrix of full rank. To maintain the equalities in (A.1), we then select the same rows in $R^0$ and obtain the square matrix $\widetilde{R}^0$. Finally, use $\widetilde{R}^{0g} = \widetilde{R}^{g-1}\widetilde{R}^0$ instead of $R^{0g}$ in (A.2). This yields the same $g$-normalized parameters.

where

$$F^g = P \begin{bmatrix} I_{K_d+K_z-m_d-m_z} \\ -N_r^{g-1} N_f^g \end{bmatrix}. \tag{A.6}$$

We partition $F^g$ into four blocks $[F_{11}^g\ F_{12}^g;\ F_{21}^g\ F_{22}^g]$ such that (A.5) yields

$$D\delta^g + Z\gamma^g = D^g \delta_f^g + Z^g \gamma_f^g, \tag{A.7}$$

where $D^g = DF_{11}^g + ZF_{21}^g$ and $Z^g = DF_{12}^g + ZF_{22}^g$. We have thus incorporated the $g$-normalization into the regressor matrix. Hence, we are in a standard setting, where $\delta_f^g$ and $\gamma_f^g$ can be estimated and (A.5) then gives the estimate of the full vector.[18]

Results (A.6) and (A.7) simplify if the normalization does not involve $\gamma^g$, that is, if the rightmost $K_z$ columns in $N^g$ are zero. This holds, for example, for the normalizations advocated in Sections 3.2 and 4. In this paragraph we thus consider $\gamma^g = \gamma_f^g$. Then (A.5) implies $[F_{21}^g, F_{22}^g] = [0, I_{K_z}]$. Moreover, the rightmost $K_z$ columns in $N_f^g$, which refer to $\gamma^g$ by construction of $P$, contain only zeros. Then the same holds for $N_r^{g-1} N_f^g$. Hence, considering the complete $F^g$ matrix, (A.6) implies that its rightmost $K_z$ columns consist of zeros except for a block $I_{K_z}$. The $P$ matrix in (A.6) permutes the rows such that $I_{K_z}$ ends up at the rows corresponding to the elements of $\gamma^g$, that is, the bottom rows. Hence, above those rows, the rightmost $K_z$ columns in $F^g$ contain only zeros. Hence, $F^g$ is block diagonal:

$$F^g = \begin{bmatrix} F_{11}^g & 0 \\ 0 & I_{K_z} \end{bmatrix}. \tag{A.8}$$

As a result, $D^g = DF_{11}^g$ and $Z^g = Z$. Hence, $Z$ is no longer transformed, reflecting there is no normalization on $\gamma^g$.

# B Testing constraints that identify $\gamma$

We are interested in $\gamma$, the true value of the impact of the constant regressors. The presence of the fixed effects $\delta$ makes that we can only estimate $\gamma^g$, leaving $\gamma$ unidentified. But $\alpha_i^g = 0$ for all $i$ implies $\nu^g = \nu$, and $\theta_t^g = 0$ for all $t$ implies $\omega^g = \omega$, as explained in Section 3.1, so that we have constraints that identify (parts of) $\gamma = [\nu', \omega']'$. This appendix sets out the diagnostic and sensitivity tests of such constraints. Both tests are special cases of the following more general testing procedure.

[18]To show that $[D^g, Z^g]$ has independent columns, start from (5), so that $[D, Z; N^g]$ has full column rank. Because also $F^g$ has full column rank, the same holds for $[D, Z; N^g]\, F^g$. Substituting (A.6) and the first equality in (A.3) yields $[D, Z; N^g]\, F^g = [[D, Z]\, F^g; 0]$. Because all columns on the left are linearly independent and on the right there is a block of only zeros under $[D, Z]\, F^g$, the latter matrix has independent columns. That matrix is $[D^g, Z^g]$.

## B.1 Testing under normalizations

As long as parameters are not pinned down in normalizations, such as $\nu^g$ if normalizations are on $\alpha_i^g$, we can use a standard way to test constraints, such as a Wald test. But parameters that are pinned down due to normalizations are linked to each other and their estimator can have a singular covariance matrix, invalidating standard testing. Things change if we incorporate the normalization into the constraint.

Consider the null hypothesis

$$H_0 : C \begin{bmatrix} \delta^g \\ \gamma^g \end{bmatrix} = c, \tag{B.1}$$

where $C$ is the constraint matrix with independent rows, and $c$ is a vector of constraint values. Testing this involves two problems. First, the estimator of $[\delta^{g\prime}, \gamma^{g\prime}]'$ has a singular covariance matrix, due to the normalization. This is resolved by substituting (A.5) into (B.1), giving $CF^g \left[\delta_f^{g\prime}, \gamma_f^{g\prime}\right]' = c$, where the vector of free parameters can be estimated in the standard way with a non-singular covariance matrix (see Appendix A.2).

The second problem is that rows in $C$ may be redundant due to the normalization. For example, if the normalization makes that $\alpha_N^g$ follows from the other $\alpha_i^g$, then constraining all $\alpha_i^g$ makes at least one row in $C$ redundant. More formally, $CF^g$ may have dependent rows. We remove those from $CF^g$ and denote the result by $C^g$. Taking out the corresponding rows from $c$ yields $c^g$. We thus rewrite

$$H_0 : C^g \begin{bmatrix} \delta_f^g \\ \gamma_f^g \end{bmatrix} = c^g. \tag{B.2}$$

To test this hypothesis, the researcher can choose a statistic, depending on the application at hand. That choice is beyond the focus of our paper. One example is the Wald test using the $\chi_Q^2$-distribution, where $Q$ is the number of independent constraints.[19] It may also be informative to study the constraints individually, for example, by t-tests.

## B.2 Diagnostic test

The $g$-normalized FE $\alpha_i^g$ and $\theta_t^g$ represent $i$-variables and $t$-variables that are relevant for $y_{it}$ but omitted from the model. Our first test is about existence of such omitted variables, so the null hypothesis constrains $\alpha_i^g$ and/or $\theta_t^g$ to zero. This means that

[19] Here we have in mind the null hypothesis that all $\alpha_i^g = 0$, in an application with large $T$. If $N$ grows, so do the degrees of freedom of $\chi_Q^2$. Donald et al. (2003) show that even for $N \to \infty$ the $\chi_Q^2$-distribution applies. Because such an approach is correct for fixed $Q$ as well, they prefer it over a standardized test, such as the standardized Wald statistic $(\text{Wald} - Q)/\sqrt{2Q}$ with a standard normal asymptotic distribution. Lu and Su (2020) test for the presence of FE when $N$ and $T$ pass to infinity simultaneously. Ghysels et al. (2020) introduce a test for many zero restrictions, but a Wald test is preferable here because of higher power, as discussed in Appendix B.4.

part of $\delta^g$ is constrained, and we denote that part by a subscript 0. Hence, the null is $\delta_0^g = 0$. This is a special case of (B.1). Hence, defining $C$ accordingly, setting $c = 0$, and following the approach of the previous section gives a test, such as the Wald test. This is our diagnostic test.

The constraint is sufficient for identification of $\nu$ and/or $\omega$. It is not necessary, because $\alpha_i^g$ and/or $\theta_t^g$ may be uncorrelated with the included regressors, ensuring identification even if they are nonzero. Hence, the diagnostic test may reject even if the true values are identified.

### B.3 Sensitivity test

The second test avoids the stringency of the diagnostic test by accounting for the fact that, even if omitted variables exist, they need not matter for estimating some parameters. This is similar to the idea underlying an omitted variables bias test. We compare the unconstrained estimator $\widehat{\beta}$ to the estimator $\widetilde{\beta}$ under the constraint $\delta_0^g = 0$, which is equivalent to setting its free part $\delta_{f0}^g = 0$. If $\widehat{\beta} - \widetilde{\beta} = 0$, the estimator of $\beta$ is insensitive to the FE, which supports leaving them out.

We can compute $\widehat{\beta} - \widetilde{\beta}$, but we do not know its variance, as $\widehat{\beta}$ and $\widetilde{\beta}$ are correlated. However, if we focus on LSDV, $\widehat{\beta} - \widetilde{\beta}$ can be written as a transformation of $\widehat{\delta}_{f0}^g$, based on Magnus and Vasnev (2007):

$$\begin{bmatrix} \widehat{\delta}_{f\emptyset}^g \\ \widehat{\gamma}_f^g \\ \widehat{\beta} \end{bmatrix} - \begin{bmatrix} \widetilde{\delta}_{f\emptyset}^g \\ \widetilde{\gamma}_f^g \\ \widetilde{\beta} \end{bmatrix} = -\left( \left[D_\emptyset^g, Z^g, X\right]' \left[D_\emptyset^g, Z^g, X\right] \right)^{-1} \left[D_\emptyset^g, Z^g, X\right]' D_0^g \cdot \widehat{\delta}_{f0}^g, \tag{B.3}$$

where $\delta_{f\emptyset}^g$ collects the elements of $\delta_f^g$ that are not in $\delta_{f0}^g$, and $D_\emptyset^g$ and $D_0^g$ are the corresponding submatrices of $D^g$ defined below (A.7).[20] We know the distribution of $\widehat{\delta}_{f0}^g$ and thereby of $\widehat{\beta} - \widetilde{\beta}$. We can thus test whether its realization differs significantly from zero, for example by a Wald test. This is our sensitivity test. It essentially takes (B.2) and uses a specific linear combination of $\delta_{f0}^g$, illustrating that here not the mere absence of omitted variables ($\delta_{f0}^g = 0$) is crucial, but rather how much a combination of them matters for estimating parameters of interest.

To interpret (B.3), distinguish two parts on the right. At the end, we have the diagnostic part, $\widehat{\delta}_{f0}^g$, which measures the magnitude of the misspecification due to constraining FE to zero. The remainder indicates how much one unit of misspecification

[20]To derive (B.3), combine model equation (2) with (4) and (A.7). Then split off $\delta_{f0}^g$ by partitioning the regressors into $L = \left[D_\emptyset^g, Z^g, X\right]$ and $D_0^g$, and collecting the parameters for $L$ by $\lambda = [\delta_{f\emptyset}^{g\prime}, \gamma_f^{g\prime}, \beta']'$. Now the model becomes $y = L\lambda + D_0^g \delta_{f0}^g + \varepsilon$. Then $[\widehat{\lambda}', \widehat{\delta}_{f0}^{g\prime}]'$ depends on $([L, D_0^g]'[L, D_0^g])^{-1}$. Applying the partitioned inverse formula to the latter, and substituting the resulting expression for $\widehat{\delta}_{f0}^g$ into that for $\widehat{\lambda}$ yields (B.3).

matters for the estimate of $\beta$ (and the other parameters), so it is a derivative. Even a large and/or significant $\widehat{\delta}^g_{f0}$ can barely matter for estimating $\beta$, if the derivative in that direction is low. Magnus and Vasnev (2007) emphasize the importance of analyzing the derivative in addition to the diagnostic. Our sensitivity test accounts for both aspects.

### B.4 Monte Carlo study

This appendix presents a concise Monte Carlo analysis of the above diagnostic and sensitivity tests. Given their popularity, we focus on Wald tests. We only consider FE $\alpha_i$, so the null hypothesis is that all normalized country fixed effects are zero; as in (7). The results, however, are one-to-one applicable to a setting with only time FE $\theta_t$.

For some designs, exact finite-sample results of the Wald tests exist. Still, practitioners typically use the $\chi^2$-approximation. That yields oversize for small $T$, so one relevant question is how quickly that disappears when $T$ grows in a panel. Moreover, what happens if we let $N$ and thus the number of fixed effects grow large? How powerful are the Wald tests for our hypothesis? How do correlations between fixed effects and observed regressors matter? Together with the empirical application in Section 5, the Monte Carlo answers to these questions will illustrate the potential of our approach.

#### B.4.1 Design

The model contains country FE, one constant regressor $v_i$, and one $it$-regressor $x_{it}$,

$$y_{it} = \alpha^0 + \alpha_i^0 + \nu^0 v_i + \beta x_{it} + \varepsilon_{it}, \tag{B.4}$$

where we have normalized the FE. The exact normalization is irrelevant in the Monte Carlo analysis, but a concrete example is $\alpha^0_{N-1} = \alpha^0_N = 0$, so that the null hypothesis is $\alpha^0_1 = \ldots = \alpha^0_{N-2} = 0$. We estimate by LSDV and study degrees-of-freedom-corrected Wald tests, using 5% $\chi^2$-critical values.

The data generating process (DGP) allows for correlation between $\alpha_i$, $v_i$, and $x_{it}$, in line with our fixed effects framework. More specifically, we generate

$$y_{it} = \alpha_i + \varepsilon^y_{it} \tag{B.5}$$

where

$$\alpha_i = \alpha_v \varepsilon^v_i + \alpha_x \varepsilon^x_i + \varepsilon^\alpha_i \tag{B.6}$$

$$v_i = \nu_x \varepsilon^x_i + \varepsilon^v_i \tag{B.7}$$

$$x_{it} = \varepsilon^x_i + \varepsilon^\chi_{it}. \tag{B.8}$$

We leave out a constant, $v_i$, and $x_{it}$ in (B.5), because they would not affect our Wald tests given the above model. The parameters $\alpha_v, \alpha_x, \nu_x$ govern the key correlations. The innovations $\varepsilon_{it}^y, \varepsilon_i^v, \varepsilon_i^x, \varepsilon_{it}^\chi, \varepsilon_i^\alpha$ are iid with zero means. Without loss of generality in our exercise, we fix the variances of $\varepsilon_{it}^y, \varepsilon_i^v, \varepsilon_i^x$ at unity.[21] The variance of $\varepsilon_{it}^\chi$ is $\sigma_\chi^2 > 0$, and that of $\varepsilon_i^\alpha$ is $\sigma_\alpha^2 \geq 0$. All innovations are normally distributed, though we also address the uniform distribution.

Instead of choosing the five DGP parameter values directly, we first move to a more orthogonal parameter space, as in Kiviet (2012). The five base parameters underlying that space follow in a straightforward manner. First, $\sigma_\chi^2 > 0$ is a base parameter by itself and quantifies the variation of $x_{it}$ within a country over time relative to the pure cross-sectional variation due to $\varepsilon_i^x$. Second, the variance $\mathrm{Var}(\alpha_i) = \alpha_v^2 + \alpha_x^2 + \sigma_\alpha^2 \geq 0$ is the total variation of the country effect. Finally, consider the correlations between $\alpha_i$, $v_i$, and $\varepsilon_i^x$, that is, $\rho_{\alpha v} \in [-1, 1]$, $\rho_{\alpha x} \in (-1, 1)$, and $\rho_{vx} \in (-1, 1)$, where the first two are only defined if $\mathrm{Var}(\alpha_i) > 0$.[22] The correlation between $\alpha_i$ and $v_i$ consists of a direct and an indirect part via $\varepsilon_i^x$, and we simply define the direct part $\rho_{\alpha v}^d$ by splitting off the indirect part, so $\rho_{\alpha v}^d = \rho_{\alpha v} - \rho_{\alpha x}\rho_{vx}$ (one could also take the partial correlation). We take $\rho_{\alpha v}^d$, $\rho_{\alpha x}$, and $\rho_{vx}$ as base parameters.

Below, we choose base parameters within the above ranges. We then derive the DGP parameters and generate $N \times T$ observations.[23] We replicate this 100,000 times, redrawing all five innovations at each replication.

#### B.4.2 Size

Under the null hypothesis that all normalized country fixed effects $\alpha_i^0$ in model (B.4) are zero, the (non-normalized) $\alpha_i$ do not depend on $i$, as shown in Section 3.1. The reverse is also true. Hence, to calculate the actual sizes of the Wald tests, we generate data under $\mathrm{Var}(\alpha_i) = 0$.

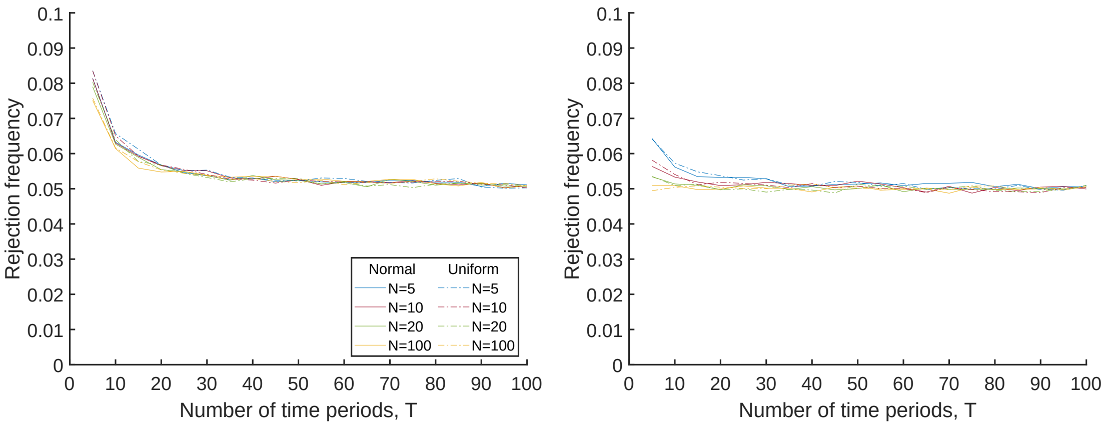
Figure B.1a presents the sizes of the Wald diagnostic test for various $N$ and $T$. It is reassuring that the number of fixed effects, $N$, has virtually no effect on size. This could have been expected from Donald et al. (2003).[19] Size also hardly depends on the distribution of the innovations, as the solid line for the normal and the dashed line for the uniform distribution are so close. The test, however, is oversized for small $T$, say $T < 20$. This is in line with the cross-section results in Evans and Savin (1982). They

[21] The Wald tests are invariant to these fixations. First, multiplying all five innovation variances by some positive factor does not change the Wald tests, so we can set that of $\varepsilon_{it}^y$ to unity. Moreover, multiplying $\varepsilon_i^v$ by a nonzero factor is absorbed by dividing $\nu^0$ by that factor, and similarly for $\varepsilon_i^x$, $\varepsilon_{it}^\chi$, and $\beta$, without changing the Wald tests, so that we can fix the variances of $\varepsilon_i^v$ and $\varepsilon_i^x$ at unity.

[22] If DGP parameter $\sigma_\alpha^2 > 0$, we further know that $\mathrm{Var}(\alpha_i) > 0$, $\rho_{\alpha v} \in (-1, 1)$, and that the correlation matrix of $[\alpha_i, v_i, \varepsilon_i^x]'$ becomes positive definite with determinant $1 + \rho_{\alpha x}^2\rho_{vx}^2 - \rho_{\alpha v}^{d2} - \rho_{\alpha x}^2 - \rho_{vx}^2 > 0$.

[23] If $\mathrm{Var}(\alpha_i) > 0$, we get $\alpha_v = \rho_{\alpha v}^d\sqrt{\mathrm{Var}(\alpha_i)/(1-\rho_{vx}^2)}$, $\alpha_x = \rho_{\alpha x}\sqrt{\mathrm{Var}(\alpha_i)}$, $\nu_x = \rho_{vx}/\sqrt{1-\rho_{vx}^2}$, and $\sigma_\alpha^2 = \mathrm{Var}(\alpha_i)\left[1 - \rho_{\alpha v}^{d2}/(1-\rho_{vx}^2) - \rho_{\alpha x}^2\right]$, where the term in square brackets is positive because the determinant in footnote 22 is so. If $\mathrm{Var}(\alpha_i) = 0$, the formula for $\nu_x$ remains, and $\alpha_v = \alpha_x = \sigma_\alpha^2 = 0$.

Figure B.1: Actual sizes of Wald tests of the absence of fixed effects $\alpha_i^0$




(a) Diagnostic test

(b) Sensitivity test

Each gridpoint is based on 100,000 normal or uniform draws from DGP (B.5) using $\mathrm{Var}(\alpha_i) = 0$, while the other base parameters turn out to be irrelevant. We estimate model (B.4) and compute the Wald tests of Appendices B.2 and B.3. Further details are in Section B.4.1.

suggest using the likelihood ratio test to avoid oversize, and our results for that test (not reported) corroborate that. Still, we keep our focus on Wald, because it is easier to use, oversize only makes our identification strategy conservative, and applications may exhibit a substantial $T$.

Figure B.1b shows the sizes for the sensitivity test. The distribution of the innovations has virtually no effect, and the test is better sized for small $T$ than the diagnostic test.

### B.4.3 Power

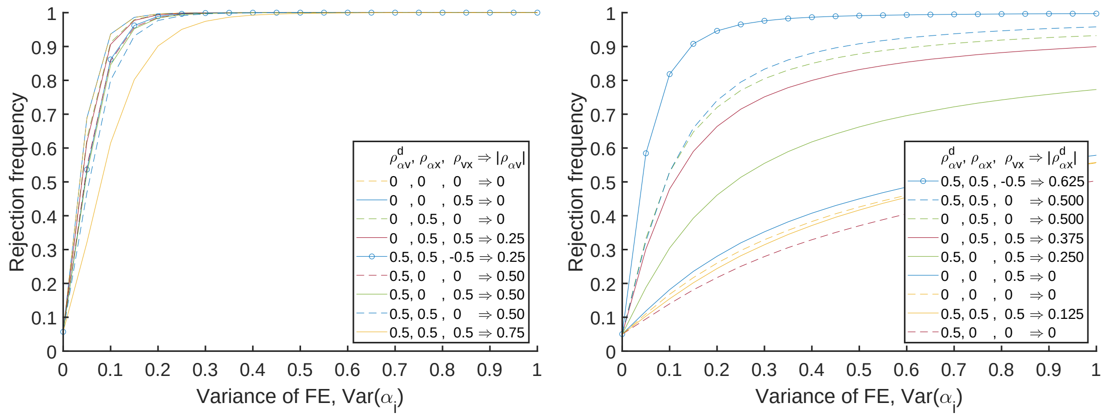
Based on the results of the previous section, we study the powers of both Wald tests for $N = 20$ and $T = 20$ with normally-distributed innovations.[24] Figure B.2 displays the powers as a function of the magnitude of the fixed effects, $\mathrm{Var}(\alpha_i)$. Powers have not been size corrected. There are power curves for nine representative combinations $\left(\rho^d_{\alpha v}, \rho_{\alpha x}, \rho_{vx}\right)$, eight containing all combinations of $\rho$s from $\{0, 0.5\}$, and one for $(0.5, 0.5, -0.5)$. The figure note motivates the choices of the base parameters.

Figure B.2a concerns the diagnostic test. For all $\rho$-combinations power quickly increases when the FE become stronger. Empirical Section 5 presents more evidence that the diagnostic test has serious power.

One can understand this power from the intuition in Ghysels et al. (2020). They

[24]This choice gives a representative view. Still, there are some moderate dependencies of power on panel sizes. The powers of both Wald tests depend positively on $N$, because for higher $N$ the null hypothesis contains more constraints. The powers also depend positively on $T$, because higher $T$ gives more observations to estimate each $\alpha_i^0$.

Figure B.2: Powers of Wald tests of the absence of fixed effects $\alpha_i^0$ for $N = T = 20$




(a) Diagnostic test (b) Sensitivity test

Each gridpoint is based on 100,000 normal draws from DGP (B.5) and base parameters below. We estimate model (B.4) and compute the Wald tests of Appendices B.2 and B.3. Further details are in Section B.4.1.

The five base parameters are as follows. First, under the alternative hypothesis, $\text{Var}(\alpha_i) > 0$, and we cover that by a grid from 0 to 1, using that beyond 1 power increases monotonically to unity. This grid is realistic, as in Section 5 we can compute the variance of the estimated FE to obtain some idea about realistic values of $\text{Var}(\alpha_i)$, and in a model without constant regressors we get 1.6, while a model with them gives 0.03.

Next, each of $\rho^d_{\alpha v}$, $\rho_{\alpha x}$, and $\rho_{vx}$ has a grid $\{0, 0.5, -0.5\}$, giving 27 combinations. For nine of them the power curves are visible, because each of the other 18 curves coincides with an included curve.

Finally, $\sigma_\chi^2 = 1$. The diagnostic test power is robust to changes in $\sigma_\chi^2$. However, the power of the sensitivity test depends on it, not so much for $\sigma_\chi^2 > 0.1$, but if $\sigma_\chi^2$ goes to zero, power drops a lot. In the latter case, $x_{it}$ gets nearly time invariant, so in this sense comparable to $v_i$. Fortunately, in this case we still have the diagnostic test, which has high power. Note that for the two *it*-regressors in our empirical application in Section 5 we can estimate $\sigma_\chi^2$-values, which gives 1.7 and 6.9, so our choice $\sigma_\chi^2 = 1$ is sensible.

argue that the power of Wald tests increases if regressors become less correlated. Our regressors are almost all dummies, which are orthogonal, and for orthogonal regressors the authors show that Wald works well. They derive this for hypotheses consisting of many constraints. In fact, in our setting, a larger $N$ and thus more constraints makes Wald somewhat more powerful. Finally, their judgment is based on a Wald test that has lost power by adjusting for severe size distortion. We largely avoid such size distortion by using the degrees-of-freedom-corrected Wald test, as advised by Evans and Savin (1982), and that further explains the substantial power we find.

The ordering of the power curves yields insights into the power determinants, as follows. The top two items in the legend concern $\rho^d_{\alpha v} = \rho_{\alpha x} = 0$. The only source of the FE $\alpha_i$ is pure randomness $\varepsilon_i^\alpha$, uncorrelated with observables. This is where the test has maximum power. Intuitively, the $\alpha_i$ that are hidden in $y_{it}$ are neither explained by $v_i$, nor by $x_{it}$, so they are fully picked up by large (absolute) estimated $\alpha_i^0$, causing high power. As this holds for any $\rho_{vx}$, the two top power curves coincide.

Going down the legend reduces power (slightly). The most important power indicator is the total correlation $\rho_{\alpha v} = \rho^d_{\alpha v} + \rho_{\alpha x}\rho_{vx}$ in the right column of the legend, which is monotonically linked to power. Another driver is $\rho_{\alpha x}$. Hence, correlations between the FE and both regressors matter, where stronger correlation lowers power.

Figure B.2b gives the power of the sensitivity test for the same nine $\rho$-combinations as before. This test reflects the influence of leaving out the FE on the estimated $\beta$, similar to the idea of an omitted variable bias test. It thus provides an indirect signal compared to the direct signal in the diagnostic test, so that it is not surprising that the sensitivity test has lower power. As before, power is increasing in $\mathrm{Var}(\alpha_i)$, reflecting that the diagnostic part is relevant in the sensitivity test formula (B.3).

Because of the link to omitted variable bias, let us consider $\rho_{\alpha x}$. As usual, a high value by itself causes an upward bias in the estimated $\beta$, giving power to the test. From Basu (2020) we conclude that the indirect correlation term $\rho_{\alpha v}\rho_{vx}$ mitigates the bias. Hence, as before, we define the direct part of the correlation as $\rho^d_{\alpha x} = \rho_{\alpha x} - \rho_{\alpha v}\rho_{vx}$, and use this as a simple bias indicator. Its values are on the right side in the legend.

The legend is again ordered from high to low power, and the ordering confirms that the bias indicator is positively related to power, though not perfectly. We see that the lowest power occurs when there is low omitted variable bias, and that the test has substantial power when omitted variable bias is high. Both are reassuring.

# C Robustness analysis

This appendix confirms that our results are robust to various deviations from the baseline specification. We focus on the identification tests, for which the precise normalization is irrelevant, but we also report the estimates under untangling normalization.

## C.1 Leaving out country-specific trends $\tau_i \cdot t$

The number of papers that include trend FE $\tau_i \cdot t$ into a gravity type of model is growing. They are also relevant here. More specifically, a Wald test of $\tau_i = 0$ for all $i$ (leaving $\tau$ unrestricted) is 267, much higher than the critical value of 26 based on the $\chi^2_{16}$-distribution. Moreover, leaving out country trends affects the estimates, as follows from comparing specification 7 in Table C.1 to the baseline with FE, replicated as 5. For example, the estimate for $gdp_{it}$ changes, which can be explained by the fact that $gdp_{it}$ is the only $i$-dependent regressor with a clear trend, so that it will try to fit the omitted country trends as well.

Excluding $\tau_i \cdot t$, however, does not change our main results. The $t$-regressors still explain most of the time effects ($R^2_\theta = 97\%$), and the two Wald tests confirm that the $\theta^u_t$ are jointly insignificant and removing them does not notably affect the $\beta$ estimates.

Table C.1: Sensitivity of results for $exp_{iUSt}$ model (16)

| Specification | 5 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|
| Estimation | LSDV | No $\tau_i \cdot t$ | DOLS | GPML | PPML | Pre-1979 |
| $gdp_{it}$ | 1.12 * | 2.26 * | 0.46 | 1.12 * | 0.25 | 1.39 * |
| | (0.25) | (0.19) | (0.33) | (0.25) | (0.36) | (0.20) |
| $reer_{it}$ | 0.66 * | 0.83 * | 0.50 * | 0.67 * | 0.36 * | 0.28 * |
| | (0.15) | (0.21) | (0.16) | (0.15) | (0.10) | (0.12) |
| $gdp_{USt}$ | 2.35 * | 1.55 * | 2.81 * | 2.26 * | 3.05 * | 2.27 * |
| | (0.39) | (0.50) | (0.33) | (0.39) | (0.38) | (0.40) |
| $gdp_{Wt}$ | −1.18 | −1.06 | −1.21 | −1.19 | −2.46 * | −1.67 * |
| | (0.73) | (1.13) | (0.75) | (0.72) | (0.85) | (0.47) |
| $reer_{USt}$ | −1.02 * | −0.97 * | −1.07 * | −1.01 * | −0.43 * | −1.26 * |
| | (0.13) | (0.17) | (0.13) | (0.13) | (0.18) | (0.10) |
| Wald tests | | | | | | |
| $\theta_t^u = 0$ | 22.77 | 20.31 | 22.91 | 22.33 | 55.78 * | 82.73 * |
| | [0.30] | [0.44] | [0.29] | [0.33] | [0.00] | [0.00] |
| $\beta = \beta\|_{\theta_t^u=0}$ | 0.67 | 3.24 | 0.94 | 0.55 | 2.95 | 0.79 |
| | [0.72] | [0.20] | [0.62] | [0.76] | [0.23] | [0.68] |
| $R_\theta^2$ | 0.98 | 0.97 | 0.98 | 0.98 | 0.97 | 0.95 |

All models include $\theta_t$ and have two lags for every regressor. Model 7 leaves out the country-specific trends, that is, $\tau_i = 0$. Model 8 explicitly accounts for cointegration between $exp_{iUSt}$ and $gdp_{it}$ and it uses DOLS estimation with two leads and lags for $gdp_{it}$. For models 9 and 10, the Wald sensitivity tests no longer use the transformation in (B.3), but the corresponding one for the maximum likelihood (ML) estimator, as derived by Magnus and Vasnev (2007). This can be applied to GPML and PPML, because the transformation relies on the first-order condition of ML, which is identical for ML and PML for both the Gamma and Poisson approaches. Model 11 is the same as 5, but the results are based on the enlarged 1965-2011 sample (765 observations), leading to 34 instead of 20 degrees of freedom for the Wald diagnostic test, giving critical value 48.60 instead of 31.41. The note to Table 1 provides further details.

## C.2 Non-stationarity and cointegration

We first test whether $exp_{iUSt}$, $gdp_{it}$, and $reer_{it}$ have a unit root for all countries, using the four Fisher type tests in Stata. The test equation accounts for a drift term, lagged differences, and for $exp_{iUSt}$ and $gdp_{it}$ it also has a trend. All parameters are country specific, and we account for time effects. The results indicate that $exp_{iUSt}$ and $gdp_{it}$ have a unit root, but $reer_{it}$ is stationary.

Next, we apply the Pedroni panel cointegration tests. The test equation contains country effects and trends, and the cointegrating parameter is country specific. We add time effects. There is strong evidence for cointegration between $exp_{iUSt}$ and $gdp_{it}$.

Although the LSDV estimates used earlier remain consistent in the presence of cointegration, the standard errors require adjustment. We thus perform dynamic OLS (DOLS), as proposed by Mark and Sul (2003). That is, we estimate the cointegrating regression of $exp_{iUSt}$ on $gdp_{it}$, adding two leads and lags of the first differences of the regressor (combinations of 0-3 leads and lags yield similar results), allowing their coefficients to be country specific, and including the full set of fixed effects. This gives the DOLS estimate of $\beta_1$ and its standard error. Then we fix $\beta_1$ at this value and estimate the remaining parameters using LSDV.

Model 8 in Table C.1 displays the results. They do not differ much from the baseline results in model 5. Both Wald statistics do not reject, and the estimated $\theta_t^u$ (not shown) are comparable to those in Figure 2c. The time FE are for 98% explained by the $t$-regressors, in line with the baseline.

### C.3 Multiplicative model, estimated by Gamma and Poisson PML

Instead of our constant conditional mean restriction on $\varepsilon_{it}$, motivating LSDV estimation, one may prefer a multiplicative approach by assuming that restriction for $\exp(\varepsilon_{it})$ and then use GPML or PPML. Models 9 and 10 display the results.

The GPML results are close to those of LSDV, so the difference in moment restrictions does not matter for our data. For PPML we reject $\theta_t^u = 0$. However, the plot of estimated $\theta_t^u$ is similar to that of GPML (both not shown) and LSDV in Figure 2c, and only two are outside the band. We are thus still close to the LSDV conclusion, which is confirmed by the again high $R_\theta^2 = 97\%$. The Wald sensitivity test does not reject, in line with LSDV.

### C.4 Enlarged sample: 1965-2011

So far, we have considered a sample from 1979-2011. That has been sufficient to illustrate our contributions. Now, and only in this section, we add pre-1979 data, thereby including some economically unstable years.

Table C.1 specification 11 displays the results. The main difference with the baseline sample is that the Wald diagnostic test now rejects that all untangled time FE $\theta_t^u = 0$. However, the plot of estimated $\theta_t^u$ is similar to that of Figure 2c and only two are outside the band, reflecting that the time FE are almost completely explained by the $t$-regressors ($R_\theta^2 = 95\%$). Even the big economic swings before 1979 are captured quite well by the $t$-regressors. Moreover, nonzero time effects might be uncorrelated with the included regressors such that $\omega^u$ can still equal the true value. This illustrates the stringency of our Wald diagnostic test; it is sufficient but not necessary for $\omega^u = \omega$.

In contrast to the diagnostic test, the Wald sensitivity test does not reject. That is, there is no evidence that the $t$-variables driving $\theta_t^u$ are correlated with the two $it$-regressors. The latter are similar to the included $t$-regressors, as both concern GDP and REER. This suggests that leaving out $\theta_t^u$ does not cause omitted variable bias in the estimated $\omega^u$ as well. This corroborates our qualifications regarding the diagnostic test rejection above. Furthermore, all estimates are quite similar to those in the baseline sample, where we do not reject $\theta_t^u = 0$. Hence, despite the rejection of the diagnostic test, we tentatively conclude that also in the enlarged sample the estimated $\omega^u$ reflect the true value acceptably well.