Econometric Theory, **39**, 2023, 623–658.
doi:10.1017/S0266466622000081

POST-SELECTION INFERENCE IN THREE-DIMENSIONAL PANEL DATA

HAROLD D. CHIANG
University of Wisconsin-Madison

JOEL RODRIGUE
Vanderbilt University

YUYA SASAKI
Vanderbilt University

Three-dimensional panel models are widely used in empirical analysis. Researchers use various combinations of fixed effects for three-dimensional panels while the correct specification is unknown. When one imposes a parsimonious model and the true model is rich in complexity, the fitted model inevitably incurs the consequences of misspecification including potential bias. When a richly specified model is employed and the true model is parsimonious, then the consequences typically include a poor fit with larger standard errors than necessary. It is therefore useful for researchers to have good model selection techniques that assist in determining the "true" model or a satisfactory approximation. In this light, Lu, Miao, and Su (2021, *Econometric Reviews* 40, 867–898) propose methods of model selection. We advance this literature by proposing a method of post-selection inference for regression parameters. Despite our use of the lasso technique as the means of model selection, our assumptions allow for many and even all fixed effects to be nonzero. This property is important to avoid a degenerate distribution of fixed effects which often reflect economic sizes of countries in gravity analyses of trade. Using an international trade database, we document evidence that our key assumption of approximately sparse fixed effects is plausibly satisfied for gravity analyses of trade. We also establish the uniform size control over alternative data generating processes of fixed effects. Simulation studies demonstrate that the proposed method is less biased than under-fitting fixed effect estimators, is more efficient than over-fitting fixed effect estimators, and robustly allows for inference that is as accurate as the oracle estimator.

First arXiv version: March 30, 2019. We benefited from very useful comments by Peter C.B. Phillips (*editor*), Liangjun Su (*co-editor*), three anonymous referees, Antonio Galvao, Hiro Kasahara, Kengo Kato, Carlos Lamarche, Whitney Newey, participants in 2019 Cemmap/WISE Workshop on Advances in Econometrics and 2019 University of Tokyo Workshop on Advances in Econometrics. All remaining errors are ours. Code files are available upon request from the authors. Chiang is supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. Address correspondence to Yuya Sasaki, Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA; e-mail: yuya.sasaki@vanderbilt.edu.

1. INTRODUCTION

Consider the three-dimensional panel model

$$y_{ijt} = x_{ijt}'\beta + \underbrace{\alpha_i + \gamma_j + \lambda_t}_{\text{Fixed Effects}} + \varepsilon_{ijt} \tag{1.1}$$

for $(i,j,t) \in \{1,\ldots,N\} \times \{1,\ldots,M\} \times \{0,\ldots,T\}$, where y_{ijt} denotes an outcome variable of unit (i,j) at time t, x_{ijt} denotes k-dimensional explanatory variables of unit (i,j) at time t, and α_i, γ_j, and λ_t are fixed effects associated with indices i, j, and t, respectively. To fix ideas, consider the gravity model (Tinbergen, 1962) from the empirical trade literature where y_{ijt} denotes the logarithm of the volume of exports from country i to country j in year t, and the k-dimensional covariates x_{ijt} contain observed characteristics of the trade pair (i,j) in year t, including the log Gross Domestic Product (GDP) of country i in year t (GDP_{it}), the log GDP of country j in year t (GDP_{jt}), the log distance between countries i and j (DIST_{ij}), and a dummy variable capturing the presence of a bilateral trade agreement between countries i and j (TA_{ij}), among others. The fixed effects, α_i, γ_j, and λ_t, represent the unobserved exporting country effects, destination country effects, and year effects, respectively. Researchers are often interested in the coefficient of DIST_{ij} interpreted as the trade elasticity or the trade cost. Another important parameter of empirical interest is the coefficient of TA_{ij} interpreted as the effect of bilateral trade agreements on trade volumes. See Head and Mayer (2014) for a comprehensive review of gravity analysis.

To date, variants of the three-dimensional panel model (1.1) have been extensively used in empirical analysis of international trade (see Baltagi, Egger, and Erhardt, 2017 for a survey), housing (see Baltagi and Bresson, 2017 for a survey), migration (see Ramos, 2017 for a survey), and consumer prices. In these analyses, researchers employ various combinations of fixed effects, including (I) $\alpha_i + \gamma_j$, (II) $\alpha_i + \gamma_j + \lambda_t$, and (III) $\alpha_{it} + \gamma_{jt}$, among others.[1] See Balazsi, Matyas, and Wansbeek (2017), Tables 1.1–1.3) for a comprehensive list of empirical papers and their specifications of the combinations of fixed effects. Typically, researchers do not know which combination of fixed effects correctly specifies the model of their interest. If the true model is parsimonious and a researcher erroneously assumes a rich specification, then naïve fixed effect estimators generally entail exacerbated variances. On the other hand, if the true model is rich and a researcher erroneously assumes a parsimonious specification, then naïve fixed effect estimators generally entail misspecification biases. The lack of knowledge of the true model specification therefore leads to undesired econometric results in any event.

[1] Parameters β of certain types of controls are not identified under more general combinations of fixed effects. For example, the coefficients of GDP_{it} and GDP_{jt} are not identified under the fixed effect model (III) due to the collinearity. However, the coefficients of DIST_{ij} and TA_{ij} would be identifiable under any of the three models. In empirical analysis of bilateral trade flows, the latter two coefficients are of more common interest. In fact, substituting fixed effects (such as α_{it} and γ_{jt}) for observed proxies (such as GDP_{it} and GDP_{jt}) is "now common practice and recommended by major empirical trade economists" (Head and Mayer, 2014) because GDP is often inaccurately measured especially for poor countries.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

A recent paper by Lu, Miao, and Su (2021) develops a method of model selection—also see Lu and Su (2020). Their method provides useful guidance to empirical researchers aiming to choose the correct combination of fixed effects in three-dimensional panel models. When a researcher uses a selected model to compute estimates of β and their standard errors, it is also important that she takes into account the statistical effects of the model selection. To the best of our knowledge, the existing literature does not provide a method of post-selection inference for three-dimensional panel models. In this light, we extend the frontier of this existing econometric literature (Lu, Miao, and Su, 2021) by providing a method of inference for β accounting for the effect of the model selection. To this end, we make use of the lasso technique along with de-biasing. Our method, however, does not require exactly sparse fixed effects. In other words, our assumptions allow for many and even all of the fixed effects to be nonzero in a general combination of fixed effects. Furthermore, we argue in Section 7.1 that the approximately sparse fixed effects are plausible in gravity analysis based on world trade data.

All Models Are Wrong: We use such terminologies as “correct specifications” and “true models” throughout the paper following the convention in the literature. However, we want to acknowledge a common aphorism in statistics that “all models are wrong” (Box, 1976), and a reasonable position that there may be no true model in the probability space considered by econometricians (Phillips, 2005). Our use of the naïve terminologies is for the sake of succinctness, but we emphasize that these qualifications need to be borne in mind.

Related Literature: A three-dimensional panel model was suggested by Mátyás (1997) for gravity analysis. The literature on multidimensional panels is extensive, and is surveyed in the article collection edited by Mátyás (2017). Its chapter written by Baltagi, Egger, and Erhardt (2017) provides a comprehensive list of empirical research papers employing multidimensional panel models.

Methods of model selection in three-dimensional panels are developed by Lu, Miao, and Su (2021) and this contribution serves as the primary motivation for our paper. As stated earlier, we aim to extend this frontier of the literature by developing post-selection inference for the regression parameters. We emphasize that this is a nontrivial contribution to the literature, as consistent model selection does not guarantee that it has no effect on the subsequent inference and postmodel-selection estimators are often irregular (Leeb and Pötscher, 2005).

We use the lasso technique for model selection and post-selection inference, but our assumptions do allow for all fixed effects to be nonzero. This is because we rely on the approximate sparsity condition as opposed to the exact sparsity. Post-selection inference via lasso is studied by an extensive body of literature in various contexts. This literature includes, but is not limited to, Belloni et al. (2012) for IV models, and Belloni, Chernozhukov, and Hansen (2014), Javanmard and Montanari (2014)), van de Geer et al. (2014), Zhang and Zhang (2014), and Caner and Kock (2018a) for linear regression models, and by Belloni et al. (2018) and Caner and Kock (2018b) for generalized method of moments.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Lasso estimation for panel models is suggested by Koenker (2004), Galvao and Montes-Rojas (2010), Lamarche (2010), Kock (2013), Caner and Han (2014), Lu and Su (2016), Li, Qian, and Su (2016), Qian and Su (2016), and Caner, Han, and Lee (2018)), among others. Classification and estimation by lasso for panel models is proposed by Su, Shi, and Phillips (2016)—also see Lu and Su (2017), Su and Ju (2018), and Su, Wang, and Jin (2019). For post-selection inference with panel data using lasso, Belloni et al. (2016) work with de-meaned fixed effect models with high-dimensional controls using a postdouble-selection estimator. Kock (2016) and Kock and Tang (2019) work with correlated random effect panel models and dynamic panel models with sparse fixed effects via de-biased lasso, respectively. We extend the frontier of this literature to three-dimensional panels. In addition to studying three-dimensional panels instead of two-dimensional panels, this paper differs from Kock (2016) and Kock and Tang (2019) in the following four technical points. First, we extend the theory of nodewise lasso by allowing for different convergence rates and thereby incorporate a larger class of fixed effect models. Second, we use a different proof strategy with the sparsity requirement of $ss_l(\log(p \vee (NM)))^2/(N \wedge M) = o(1)$ inspired by Belloni et al. (2012, Lemma 8), whereas an adaptation of the proof strategies of Kock (2016)[2] and Kock and Tang (2019)[3] to our framework would require $ss_l^2(\log(p \vee (NM)))^2/(N \wedge M) = o(1)$. This feature further extends the class of models that can be handled under our framework. Third, the subgaussianity assumption of covariates, which is assumed by the majority of papers in the de-biased lasso literature, is not required. Fourth, we allow for nonsparse coefficients based on the notion of approximate sparsity following that of Belloni et al. (2012) instead of the L^v sparsity for $0 < v < 1$ as in Kock and Tang (2019). We take advantage of the existing techniques of Caner and Kock (2018a) and Kock and Tang (2019) in establishing the uniform size control property over alternative data generating processes of fixed effects.

With all these technical relations to the existing literature, we once again emphasize that our main contribution is a method for robust inference in the context of three-dimensional panels. Unlike two-dimensional panels, there are a number of alternative combinations of fixed effect specifications in three-dimensional panels, and hence model selection is more important in these models (Lu, Miao, and Su, 2021). We apply and extend state-of-the-art technology (e.g., Belloni et al., 2012; Kock, 2016; Kock and Tang, 2019) to this three-dimensional panel framework common to many applied settings.

Notation: We introduce the following notation. $\mathbf{1}_n$ denotes an n-dimensional vector of ones. $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$. $||\cdot||_0$ denotes the support cardinality (the L^0 norm), $||\cdot||_1$ denotes the L^1 norm, $||\cdot||$ denotes the Euclidean norm, and $||\cdot||_\infty$ denotes the (essential) supremum. $X_n \rightsquigarrow Y$ indicates weak convergence, and $X_n = o_p(Y_n)$ indicates that $X_n/||Y_n||$ converges in probability to

[2]See Assumption A3 (b) of Kock (2016).

[3]See Assumption 5 (c) of Kock and Tang (2019).

zero. For a given integer p and $l \in \{1,\ldots,p\}$, let e_l denote the column vector of the lth standard basis .

Organization: The rest of this paper is organized as follows. We introduce the model framework in Section 2. An overview of our proposed method is presented in Section 3. The main theoretical result is presented in Section 4, followed by sufficient conditions discussed in Section 5. We conduct simulation studies in Section 6. We discuss the key assumption in the context of gravity analysis of international trade and then apply our proposed method to empirical data in Section 7. Section 8 concludes the paper. Supplementary Material to this article is available online. It contains auxiliary lemmas for the theoretical results and additional simulation results.

2. THE MODEL FRAMEWORK

Consider the following representation of a general class of three-dimensional panel models with large N and large M.

$$\begin{aligned} y_{ijt} = x_{ijt}'\beta + \sum_{i'=1}^{N}\alpha_{i'}\mathbb{1}_{i=i'} + \sum_{j'=1}^{M}\gamma_{j'}\mathbb{1}_{j=j'} + \sum_{t'=1}^{T-1}\lambda_{t'}\mathbb{1}_{t=t'} \\ + \sum_{i'=1}^{N}\sum_{t'=1}^{T-1}\alpha_{i't'}\mathbb{1}_{i=i'}\mathbb{1}_{t=t'} + \sum_{j'=1}^{M}\sum_{t'=1}^{T-1}\gamma_{j't'}\mathbb{1}_{j=j'}\mathbb{1}_{t=t'} + \varepsilon_{ijt} \end{aligned} \quad \textbf{(2.1)}$$

This representation consists of a k-dimensional parameter vector β, N-dimensional parameter vector $\alpha_{[N]} = (\alpha_1,\ldots,\alpha_N)'$, M-dimensional parameter vector $\gamma_{[M]} = (\gamma_1,\ldots,\gamma_M)'$, $(T-1)$-dimensional parameter vector $\lambda_{[T-1]} = (\lambda_1,\ldots,\lambda_{T-1})'$, $N(T-1)$-dimensional parameter vector $\alpha_{[N(T-1)]} = (\alpha_{11},\ldots,\alpha_{N(T-1)})'$, and $M(T-1)$-dimensional parameter vector $\gamma_{[M(T-1)]} = (\gamma_{11},\ldots,\gamma_{M(T-1)})'$. In total, there are $k+N+M+(T-1)+N(T-1)+M(T-1)$ parameters involved in this representation (2.1). We remark that the summations over t' run from 1 to $T-1$ to avoid the perfect multicollinearity among the time fixed effects.

Recall that conventional fixed effect models include

(I) $\alpha_i + \gamma_j$,
(II) $\alpha_i + \gamma_j + \lambda_t$, and
(III) $\alpha_{it} + \gamma_{jt}$,

among others. Model (I) entails $k+N+M$ possibly nonzero parameters $(\beta',\alpha_{[N]}',\gamma_{[M]}')'$, while the rest of the $(T-1)+N(T-1)+M(T-1)$ parameters $(\lambda_{[T-1]}',\alpha_{[N(T-1)]}',\gamma_{[M(T-1)]}')'$ are all zero. Similarly, Model (II) entails $k+N+M+(T-1)$ possibly nonzero parameters $(\beta',\alpha_{[N]}',\gamma_{[M]}',\lambda_{[T-1]})'$, while the rest of the $N(T-1)+M(T-1)$ parameters $(\alpha_{[N(T-1)]}',\gamma_{[M(T-1)]}')'$ are all zero. Likewise, Model (III) entails $k+N(T-1)+M(T-1)$ possibly nonzero parameters $(\beta',\alpha_{[N(T-1)]}',\gamma_{[M(T-1)]}')'$, while the rest of the $N+M+(T-1)$ parameters

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

$(\alpha'_{[N]}, \gamma'_{[M]}, \lambda'_{[T-1]})'$ are all zero. Furthermore, the representation (2.1) includes many combinations other than these three models.

When Model (I) is true, for example, then the representation (2.1) has $(T-1)+N(T-1)+M(T-1)$ redundant parameters and hence estimating the model (2.1) generally yields much larger standard errors for the parameters β of interest than necessary. This motivates the need for model selection. We propose to use the lasso to select such redundant fixed effect parameters out of the representation (2.1), and then conduct inference robustly accounting for the statistical effects of model selection.

Remark 1. The representative model (2.1) appears to be redundant, and all of Models (I)–(III) can be succinctly represented by Model (III), i.e.,

$$y_{ijt} = x'_{ijt}\beta + \sum_{i'=1}^{N}\sum_{t'=1}^{T}\alpha_{i't'}\mathbb{1}_{i=i'}\mathbb{1}_{t=t'} + \sum_{j'=1}^{M}\sum_{t'=1}^{T}\gamma_{j't'}\mathbb{1}_{j=j'}\mathbb{1}_{t=t'} + \varepsilon_{ijt}. \tag{2.2}$$

Therefore, the method that we present in this paper can be conducted based on (2.2) as well as (2.1). However, we emphasize important advantages of using the representation (2.1). When the true model is Model (I), for instance, the parsimonious representation (2.2) will select as many as $(N+M)(T-1)$ fixed effects, $\alpha_{[N(T-1)]}$ and $\gamma_{[M(T-1)]}$, whereas the representation (2.1) will select only $N+M$ fixed effects, $\alpha_{[N]}$ and $\gamma_{[M]}$. Therefore, the representation (2.1) allows for a much smaller number of regressors in the post-lasso selection, leading to preferred asymptotic properties. A similar remark applies to the case where the true model is Model (II). △

For ease of conducting econometric analysis, we further rewrite the representation (2.1) as

$$y_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{d}'_{1,it}\boldsymbol{\alpha} + \mathbf{d}'_{2,jt}\boldsymbol{\gamma} + \varepsilon_{ijt}, \tag{2.3}$$

where $\mathbf{x}_{ijt} = (x'_{ijt}, \mathbb{1}_{t=1}, \ldots, \mathbb{1}_{t=T-1})'$ and $\boldsymbol{\beta} = (\beta', \lambda_1, \ldots, \lambda_{T-1})'$ are of dimension $k_0 = k + (T-1)$, $\mathbf{d}_{1,it} = (\mathbb{1}_{i=1}, \ldots, \mathbb{1}_{i=N}, \mathbb{1}_{i=1}\mathbb{1}_{t=1}, \ldots, \mathbb{1}_{i=N}\mathbb{1}_{t=T-1})'$ and $\boldsymbol{\alpha} = (\alpha'_{[N]}, \alpha'_{[N(T-1)]})'$ are of dimension $N_0 = N + N(T-1)$, and $\mathbf{d}_{2,jt} = (\mathbb{1}_{j=1}, \ldots, \mathbb{1}_{j=N}, \mathbb{1}_{j=1}\mathbb{1}_{t=1}, \ldots, \mathbb{1}_{j=M}\mathbb{1}_{t=T-1})'$ and $\boldsymbol{\gamma} = (\gamma'_{[M]}, \gamma'_{[M(T-1)]})'$ are of dimension $M_0 = M + M(T-1)$.

Suppose that we can decompose the fixed effects $\boldsymbol{\alpha}$ into $\overline{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}$ and decompose the fixed effects $\boldsymbol{\gamma}$ into $\overline{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}$ so that

$$\|\overline{\boldsymbol{\alpha}}\| \text{ is bounded and } \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T-1}\left(\mathbf{d}'_{1,it}(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}})\right)^2 \lesssim \|\boldsymbol{\beta}\|_0 + \|\overline{\boldsymbol{\alpha}}\|_0 + \|\overline{\boldsymbol{\gamma}}\|_0 \tag{2.4}$$

and

$$\|\overline{\boldsymbol{\gamma}}\| \text{ is bounded and } \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T-1}\left(\mathbf{d}'_{2,it}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})\right)^2 \lesssim \|\boldsymbol{\beta}\|_0 + \|\overline{\boldsymbol{\alpha}}\|_0 + \|\overline{\boldsymbol{\gamma}}\|_0 \tag{2.5}$$

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

hold, where we recall that $\|\cdot\|_0$ denotes the support cardinality (the L^0 norm).[4] Such a decomposition is constructed by setting $\overline{\boldsymbol{\alpha}}_\ell$ equal to $\boldsymbol{\alpha}_\ell$ for those coordinates ℓ for which $|\boldsymbol{\alpha}_\ell|$ is large and setting $\overline{\boldsymbol{\alpha}}_\ell$ equal to zero for those coordinates ℓ for which $|\boldsymbol{\alpha}_\ell|$ is small, and similarly for $\boldsymbol{\gamma}$. Consequently, we can further rewrite the representation (2.3) as

$$y_{ijt} = \mathbf{x}'_{ijt}\boldsymbol{\beta} + \mathbf{d}'_{1,it}\overline{\boldsymbol{\alpha}} + \mathbf{d}'_{2,jt}\overline{\boldsymbol{\gamma}} + r_{ijt} + \varepsilon_{ijt},$$

where r_{ijt} is the approximation error defined by

$$r_{ijt} = \mathbf{d}'_{1,it}(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}) + \mathbf{d}'_{2,jt}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}),$$

which satisfies

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T-1} r_{ijt}^2 \lesssim \|\boldsymbol{\beta}\|_0 + \|\overline{\boldsymbol{\alpha}}\|_0 + \|\overline{\boldsymbol{\gamma}}\|_0.$$

Stacking the three-dimensional panel data across the NMT observations, we in turn construct the matrix representation

$$Y = X\boldsymbol{\beta} + D_1\overline{\boldsymbol{\alpha}} + D_2\overline{\boldsymbol{\gamma}} + R + \varepsilon = Z\overline{\boldsymbol{\eta}} + R + \varepsilon, \tag{2.6}$$

where $Y = (y_{111},\ldots,y_{NMT})'$, $R = (r_{111},\ldots,r_{NMT})'$, and $\varepsilon = (\varepsilon_{111},\ldots,\varepsilon_{NMT})'$, are vectors of dimension NMT, $X = (\mathbf{x}_{111},\ldots,\mathbf{x}_{NMT})'$ is a matrix of size $NMT \times k_0$, $D_1 = (\mathbf{d}_{1,11},\ldots,\mathbf{d}_{1,NT})' \otimes \mathbf{1}_M$ is a matrix of size $NMT \times N_0$, $D_2 = \mathbf{1}_N \otimes (\mathbf{d}_{2,11},\ldots,\mathbf{d}_{2,MT})'$ is a matrix of size $NMT \times M_0$, $Z = [X\; D_1\; D_2]$, and $\overline{\boldsymbol{\eta}} = [\boldsymbol{\beta}'\; \overline{\boldsymbol{\alpha}}'\; \overline{\boldsymbol{\gamma}}']'$ is a vector of dimension $p = k_0 + N_0 + M_0$. k_0, N_0 and M_0 can vary with sample size unless otherwise stated.

To understand the requirement (2.4), let us assume that α_{it} indicates the logarithm of country is GDP adjusted by its price index. If we have a large number of very small log GDPs, then we can set their $\overline{\alpha}_{it}$ to zero and treat their small differences $\alpha_{it} - \overline{\alpha}_{it}$ as approximation errors. Since α_{it} is not precisely zero and there is heterogeneity in both country-level GDPs and price indices, allowing for $R \neq 0$ accommodates a realistic scenario. The requirement that $||\overline{\boldsymbol{\alpha}}||$ is bounded in (2.4) can be interpreted as a finite bounded sum of squares of these price-adjusted GDPs worldwide.

Example 1. To better understand the decomposition of the fixed effects (FEs) in (2.4)–(2.6), consider Model (I) with $k_0 = 0$ for simplicity and the following two fixed effect designs.

---

[4] With this said, we emphasize that this decomposition is merely theoretical, and a researcher need not implement such a decomposition in practice. Precise requirements for the decomposition are stated in Assumptions 2 and 5 (4) ahead, followed by a discussion in Remark 4 in the context of our motivating application (1.1). In Section 7.1, we use world trade data to argue that these assumptions are plausible in the application (1.1).

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

1. (Exactly sparse FE) $\alpha_i = 1/\left(i\cdot(\log(i+1))^{3/2}\right)$ for all $i \in \{1,\ldots,\lfloor C_\alpha N^{1/3}\rfloor\}$, $\alpha_i = 0$ for all $i \in \{\lfloor C_\alpha N^{1/3}\rfloor+1,\ldots\}$, $\gamma_j = 1/\left(j\cdot(\log(j+1))^{3/2}\right)$ for all $j \in \{1,\ldots,\lfloor C_\gamma M^{1/3}\rfloor\}$, and $\gamma_j = 0$ for all $j \in \{\lfloor C_\gamma M^{1/3}\rfloor+1,\ldots\}$, where C_α and C_γ are some positive constants.
2. (Approximately sparse FE) $\alpha_i = 1/\left(i\cdot(\log(i+1))^{3/2}\right)$ for all $i \in \mathbb{N}$ and $\gamma_j = 1/\left(j\cdot(\log(j+1))^{3/2}\right)$ for all $j \in \mathbb{N}$.

Under the exactly sparse FE, let $\overline{\alpha}_i = \alpha_i$ for all $i \in \mathbb{N}$ and let $\overline{\gamma}_j = \gamma_j$ for all $j \in \mathbb{N}$. Since $||\overline{\boldsymbol{\alpha}}||$ and $||\overline{\boldsymbol{\gamma}}||$ are bounded and $\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}} = \boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}} = 0$, this decomposition satisfies (2.4) and (2.5). Under the approximately sparse FE, let $\overline{\alpha}_i = \alpha_i$ for all $i \in \{1,\ldots,\lfloor C_\alpha N^{1/3}\rfloor\}$, $\overline{\alpha}_i = 0$ for all $i \in \{\lfloor C_\alpha N^{1/3}\rfloor+1,\ldots\}$, $\overline{\gamma}_j = \gamma_j$ for all $j \in \{1,\ldots,\lfloor C_\gamma M^{1/3}\rfloor\}$, and $\overline{\gamma}_j = 0$ for all $j \in \{\lfloor C_\gamma M^{1/3}\rfloor+1,\ldots\}$, where C_α and C_γ are some positive constants. Since $||\overline{\boldsymbol{\alpha}}||$, $||\overline{\boldsymbol{\gamma}}||$, and $||R||$ are bounded while $||\overline{\boldsymbol{\alpha}}||_0$ and $||\overline{\boldsymbol{\gamma}}||_0$ diverge, this decomposition also satisfies (2.4) and (2.5). We will revisit this example more formally in Section 5. On the other hand, the following fixed effect design is ruled out by the above decomposition.

3. $\alpha_i = (-1)^i$ for all $i \in \mathbb{N}$ and $\gamma_j = (-1)^j$ for all $j \in \mathbb{N}$.

In order to achieve a bounded $||\overline{\boldsymbol{\alpha}}||$ as in (2.4), $\overline{\alpha}_i$ has to diminish at the rate of $1/(\sqrt{i}\log(i)^{1/2+\delta})$ after rearrangement for some $\delta > 0$. But this forces $||R||$ to grow at the rate of $\sqrt{N}$ under this fixed effect design, whereas we require $||R||$ to be of order $||\overline{\boldsymbol{\alpha}}||_0$ in (2.4) (which in turn should be of a strictly smaller order than $\sqrt{N}$ by a formal assumption to be stated ahead). More generally, fixed effects generated i.i.d. from a nondegenerate distribution are not accommodated by the decomposition. $\triangle$

If the true model is parsimonious, like Model (I), then a large number of the elements of the high-dimensional parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, will be zero. Thus, a large number of the elements of $\overline{\boldsymbol{\alpha}}$ and $\overline{\boldsymbol{\gamma}}$ will be zero. Furthermore, for those coordinates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ that are small in absolute value, the corresponding coordinates of $\overline{\boldsymbol{\alpha}}$ and $\overline{\boldsymbol{\gamma}}$ are set to zero in the decomposition in light of the relatively smaller approximation errors caused by setting them to zero. We propose to use the lasso technique to select such redundant parameters in $\overline{\boldsymbol{\alpha}}$ and $\overline{\boldsymbol{\gamma}}$ out of this high-dimensional model as the means of model selection for the purpose of obtaining smaller standard errors. Furthermore, accounting for the statistical effects of this model selection, we then conduct robust inference for the main parameters $\boldsymbol{\beta}$ in the panel model. Section 3 illustrates an overview of our proposed method. A formal theoretical analysis will then follow in Sections 4 and 5.

3. OVERVIEW OF THE METHOD: PRACTICAL GUIDE OF IMPLEMENTATION

Our proposed method consists of four steps. The first step is a lasso estimation of the parameter vector $\boldsymbol{\eta}$ entailing model selection. The second step is an auxiliary step to calculate an approximate inverse of the Gram matrix to be used in the subsequent two steps. The third step de-biases the regularized lasso estimate from the first step. The fourth step is a calculation of the asymptotic variance of the de-biased lasso estimator of $\boldsymbol{\beta}$.

Step 1: For the representative equation (2.6), define the lasso estimator

$$\widehat{\boldsymbol{\eta}} \in \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^p} \|Y - Z\boldsymbol{\eta}\|^2 + 2\mu P(\boldsymbol{\eta}), \tag{3.1}$$

where $\mu \in [0,\infty)$ is a regularization tuning parameter and the penalty function P is defined by

$$P(\boldsymbol{\eta}) = \left\|\widehat{\Upsilon}_1\boldsymbol{\beta}\right\|_1 + \frac{1}{\sqrt{N}}\left\|\widehat{\Upsilon}_2\boldsymbol{\alpha}\right\|_1 + \frac{1}{\sqrt{M}}\left\|\widehat{\Upsilon}_3\boldsymbol{\gamma}\right\|_1$$

for some diagonal normalization matrix $\widehat{\Upsilon}_m$ and for each $m \in \{1,2,3\}$.[5] In practice, the regularization tuning parameter μ can be chosen using cross validation methods via software packages.

Step 2: The next step is an auxiliary process to obtain a $p \times p$ matrix $\widehat{\Theta}$ of the approximate inverse of the Gram matrix to be used in Step 3. We define the nodewise lasso estimator[6]

$$\widehat{\phi}^{\ell} \in \arg\min_{\phi\in\mathbb{R}^{p-1}} \left\{\left\|Z^{\ell} - Z^{-\ell}\phi\right\|^2 + 2\mu^{\ell}_{\text{node}}\left\|\frac{1}{\sqrt{NM}}S^{-\ell}\widehat{\Upsilon}^{\ell}_{\text{node}}\phi\right\|_1\right\} \tag{3.2}$$

of the ℓth column Z^{ℓ} of Z on all the other $(p-1)$ columns $Z^{-\ell}$ of Z for each $\ell \in \{1,\ldots,p\}$, where $\mu^{\ell}_{\text{node}} \in [0,\infty)$ is a regularization tuning parameter, $\widehat{\Upsilon}^{\ell}_{\text{node}}$ is some diagonal normalization matrix for each $\ell \in \{1,\ldots,p\}$, and $S^{-\ell}$ is the $(p-1)\times(p-1)$ matrix obtained by removing the ℓth row and the ℓth column of

$$S = \begin{bmatrix} \sqrt{NM}I_{k_0} & 0 & 0 \\ 0 & \sqrt{M}I_{N_0} & 0 \\ 0 & 0 & \sqrt{N}I_{M_0} \end{bmatrix}.$$

As above, the regularization tuning parameter μ^{ℓ}_{node} may be chosen using cross validation methods via software packages.

Once the nodewise lasso estimates $\widehat{\phi}^{\ell}$ are obtained, a $p \times p$ matrix $\widehat{\Theta}$ approximating the inverse Gram matrix can be constructed by

[5] See Remark B.1 in Appendix B.1 in the Supplementary Material.

[6] Similarly to the rate normalization factors $N^{-1/2}$ and $M^{-1/2}$ in (3.1), the role of matrix $S^{-\ell}$ is to adjust for differences in effective sample sizes for the main covariates and different fixed effects. For the i-specific fixed effects, the effective sample size is of the order $O(M)$. In contrast, for each of the j-specific fixed effects, only $O(N)$ of them are observed. Without this rate adjustment matrix S, we would over-penalize the fixed effects. Note that, here we regress each column on columns. These are transposes of rows in nodewise regression. The $\hat{\Theta}$ in equation (3.3) is not symmetric, but it converges to a symmetric limit in probability asymptotically.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

$$\widehat{\Theta}=\begin{bmatrix}\widehat{\tau}_1^{-2} & 0 & \cdots & 0 & 0\\ 0 & \widehat{\tau}_2^{-2} & & 0 & 0\\ \vdots & & \ddots & & \vdots\\ 0 & 0 & & \widehat{\tau}_{p-1}^{-2} & 0\\ 0 & 0 & \cdots & 0 & \widehat{\tau}_p^{-2}\end{bmatrix}\begin{bmatrix}1 & -\widehat{\phi}_1^1 & \cdots & -\widehat{\phi}_{p-2}^1 & -\widehat{\phi}_{p-1}^1\\ -\widehat{\phi}_1^2 & 1 & \cdots & -\widehat{\phi}_{p-2}^2 & -\widehat{\phi}_{p-1}^2\\ \vdots & & \ddots & & \vdots\\ -\widehat{\phi}_1^{p-1} & -\widehat{\phi}_2^{p-1} & \cdots & 1 & -\widehat{\phi}_{p-1}^{p-1}\\ -\widehat{\phi}_1^p & -\widehat{\phi}_2^p & \cdots & -\widehat{\phi}_{p-1}^p & 1\end{bmatrix}, \tag{3.3}$$

with $\widehat{\tau}_\ell$ given by

$$\widehat{\tau}_\ell^2=\frac{1}{NM}\left\|Z^\ell-Z^{-\ell}\widehat{\phi}^\ell\right\|^2+\frac{\mu_{\text{node}}^\ell}{NM}\left\|\frac{1}{\sqrt{NM}}S^{-\ell}\widehat{\Upsilon}_{\text{node}}^\ell\widehat{\phi}^\ell\right\|_1$$

for each $\ell\in\{1,\dots,p\}$ where $\widehat{\phi}_l^\ell$ denotes the lth coordinate of the nodewise lasso estimate $\widehat{\phi}^\ell$ for each $\ell\in\{1,\dots,p\}$ and $l\in\{1,\dots,p-1\}$.

Step 3: Shrinkage by the regularization $\mu P(\boldsymbol{\eta})$ forces a subvector of the lasso estimates $\widehat{\boldsymbol{\eta}}$ to be zero, and this mechanism serves as the means of model selection. Since this regularization biases the second-stage lasso estimator $\widehat{\boldsymbol{\eta}}$, we further "de-bias" it according to

$$\widetilde{\boldsymbol{\eta}}_\ell=\widehat{\boldsymbol{\eta}}_\ell+\frac{1}{NM}\widehat{\Theta}_\ell'Z'(Y-Z\widehat{\boldsymbol{\eta}}), \tag{3.4}$$

for each $\ell\in[p]$, where $\widehat{\Theta}_\ell$ is the ℓth column of $\widehat{\Theta}$ and $\widehat{\Theta}$ is the $p\times p$ approximate inverse Gram matrix constructed in Step 2. The subvectors of $\widetilde{\boldsymbol{\eta}}$ are denoted by $\widetilde{\boldsymbol{\eta}}=\left(\widetilde{\boldsymbol{\beta}}',\widetilde{\boldsymbol{\alpha}}',\widetilde{\boldsymbol{\gamma}}\right)'$.

Step 4: The asymptotic variance of $\sqrt{NM}\left(\widetilde{\boldsymbol{\beta}}_\ell-\boldsymbol{\beta}_\ell\right)$ for $\ell\in\{1,\dots,k_0\}$ is approximated by

$$\widehat{V}_{\ell\ell}=\widehat{\Theta}_\ell'\widehat{\Omega}\widehat{\Theta}_\ell,$$

where $\widehat{\Theta}_\ell$ is defined in Step 3, $\widehat{\Omega}$ is given by

$$\widehat{\Omega}=\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M\left(\sum_{t=1}^T Z_{ijt}\widehat{\varepsilon}_{ijt}\right)\left(\sum_{t=1}^T Z_{ijt}\widehat{\varepsilon}_{ijt}\right)',$$

and $\widehat{\varepsilon}_{ijt}$ is the residual from the lasso in Step 1.

Remark 2 (The De-biased Lasso). A derivation of the de-biased lasso (3.4) follows naturally from the K.K.T. condition of equation (3.1). See Proof of Theorem 1 for details.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

4. THE THEORY

In this section, we present the main theoretical results for the asymptotic normality of the de-biased lasso estimator. Define the de-biased lasso estimator by

$$\widetilde{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}} + \frac{\mu}{NM}\widehat{\Theta}' P'(\widehat{\boldsymbol{\eta}}), \tag{4.1}$$

where P' denotes the subgradient of P.

Remark 3. The de-biased lasso estimator $\widetilde{\boldsymbol{\eta}}_l = \widehat{\boldsymbol{\eta}}_l + \frac{\mu}{NM}\widehat{\Theta}_l' P'(\widehat{\boldsymbol{\eta}})$ can also be rewritten by replacing $\mu P'(\widehat{\boldsymbol{\eta}})$ by $Z'(Y - Z\widehat{\boldsymbol{\eta}})$ following the K.K.T. condition, i.e., $\widetilde{\boldsymbol{\eta}}_l = \widehat{\boldsymbol{\eta}}_l + \frac{1}{NM}\widehat{\Theta}_l' Z'(Y - Z\widehat{\boldsymbol{\eta}})$. This representation yields the concrete de-biased lasso formula proposed in (3.4). △

Recall that the subvectors of $\widetilde{\boldsymbol{\eta}}$ are denoted by $\widetilde{\boldsymbol{\eta}} = [\widetilde{\boldsymbol{\beta}}', \widetilde{\boldsymbol{\alpha}}', \widetilde{\boldsymbol{\gamma}}]'$, corresponding to $\overline{\boldsymbol{\eta}} = [\boldsymbol{\beta}' \ \overline{\boldsymbol{\alpha}}' \ \overline{\boldsymbol{\gamma}}']'$. This section presents a general limit distribution result for each coordinate of the de-biased lasso estimator $\widetilde{\boldsymbol{\beta}}$ for the coefficients of x_{ijt}. We focus on short panels with fixed T and large (N, M), although an extension to large T cases may be feasible with alternative assumptions. While we maintain high-level assumptions in the current section to allow for general applicability, we will also provide lower-level sufficient conditions in Section 5.1 based on common sampling assumptions made in the gravity analysis literature. Consider the following assumption.

Assumption 1 (Asymptotic Normality). For all (N, M), there exists a $p \times p$ matrix $\widehat{\Theta}$ such that the following conditions hold for an (N, M)-dependent choice of μ as $N, M \to \infty$.

(i) $\max_{l \in [k_0]} \left| \sqrt{NM}\left(\widehat{\Theta}_l' Z'Z \big/ (NM) - e_l'\right)(\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}})\right| = o_p(1)$.

(ii) $\max_{l \in [k_0]} \left| \widehat{\Theta}_l' Z'R \big/ \sqrt{NM} \right| = o_p(1)$.

(iii) For each $l \in [k_0]$, there exists $V_{ll} \in (0, \infty)$ that can depend on (N, M) such that

$$V_{ll}^{-1/2}\widehat{\Theta}_l' Z'\varepsilon \Big/ \sqrt{NM} \rightsquigarrow N(0, 1).$$

In the current general theoretical discussion, Assumption 1 merely requires the existence of some $\widehat{\Theta}_l$ satisfying the three conditions, and does not say how it should be constructed. Recall that the overview of the method in Section 3 suggests a concrete way to construct such $\widehat{\Theta}_l$. Section 5 ahead will discuss lower-level sufficient conditions to guarantee that such a concrete construction of $\widehat{\Theta}_l$ satisfies the three high-level conditions in Assumption 1. The following theorem provides an asymptotic linear representation for the de-biased lasso estimator $\widetilde{\boldsymbol{\eta}}$ and the asymptotic normality for $\widetilde{\boldsymbol{\beta}}_l$ for each $l \in [k_0]$.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

THEOREM 1 (Asymptotic Normality). *Suppose that Assumptions 1 (i) and (ii) are satisfied. Then,*

$$\widetilde{\boldsymbol{\eta}}_l - \overline{\boldsymbol{\eta}}_l = \frac{1}{NM}\widehat{\Theta}_l' Z'\varepsilon + o_p\left(1/\sqrt{NM}\right)$$

for each $l \in [p]$. Furthermore, if Assumption 1 (iii) is also satisfied, then we have

$$\sqrt{NM}(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l) \rightsquigarrow N(0, V_{ll})$$

for each $l \in [k_0]$.

A proof is found in Appendix B.1.

5. SUFFICIENT CONDITIONS, VARIANCE ESTIMATION, JOINT INFERENCE, AND UNIFORMITY

In this section, we first propose lower-level sufficient conditions for the high-level general statements in Assumption 1. These conditions provide a theoretical guarantee for the practical procedure outlined in Section 3 to work. While the general limit distribution result in Theorem 1 did not specify a concrete form of the asymptotic variance V_{ll}, the current section also provides a formula for it under these sufficient conditions. Furthermore, we propose an analog variance estimator $\widehat{V}_{ll}$, and show its consistency under these sufficient conditions. We also present as a corollary a theoretical guarantee for joint hypothesis testing involving multiple parameters. Finally, given the asymptotic normality result and the consistent variance estimation result, we establish the uniform size control property.

For parsimony, we will assume $\widehat{\Upsilon}_1 = I_{k_0}$, $\widehat{\Upsilon}_2 = I_{N_0}$, $\widehat{\Upsilon}_3 = I_{M_0}$ and $\widehat{\Upsilon}^{\ell}_{\text{node}} = I_{p-1}$ for all $\ell \in [k_0]$ throughout this section, although these restrictions are not at all essential—see Appendix B.1 in the Supplementary Material for essential requirements. We also use the following notation to denote the supports of the parameters: $J_1 = \text{supp}(\boldsymbol{\beta})$, $J_2 = \text{supp}(\overline{\boldsymbol{\alpha}})$, $J_3 = \text{supp}(\overline{\boldsymbol{\gamma}})$, and $J = \text{supp}(\overline{\boldsymbol{\eta}})$. Their cardinalities are denoted by $s_1 = |J_1|$, $s_2 = |J_2|$, $s_3 = |J_3|$, and $s = |J|$. We note that s is nondecreasing in N and/or M. Similarly to the decomposition (2.6) for the main regression model, we also consider the decomposition

$$Z^{\ell} = Z^{-\ell}\phi^{\ell} + R^{\ell} + \zeta^{\ell}, \qquad \textbf{(5.1)}$$
$$E[Z^{-\ell}\zeta^{\ell}] = 0$$

for each coordinate $\ell \in [k_0]$ of the regressors.

5.1. Sufficient Conditions for Assumption 1

We present the sufficient conditions as five modules: Assumptions 2–6, listed below.

Assumption 2 (Approximate Sparsity). (1) $\|\overline{\boldsymbol{\eta}}\| \leq K$. (2) $\|R\| \leq c_s \lesssim \sqrt{s}$ with probability $1-o(1)$. (3) $\|Z'R\| = o_p\left(\sqrt{NM}\right)$.

Recall that the fixed effects $\boldsymbol{\alpha}$ are decomposed into $\overline{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}}$ such that (2.4) is satisfied, and the fixed effects $\boldsymbol{\gamma}$ are decomposed into $\overline{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}$ such that (2.5) is satisfied. These conditions, (2.4) and (2.5), are imposed to satisfy Assumption 2 (1) and (2). Assumption 2 (3) is a high-level condition that limits the growth rate of the product of the data matrix and the approximation errors. It can be relaxed to a weaker condition,[7] but we present the current condition for ease of interpretation. If there is no approximation error (i.e., $R=0$), then Assumption 2 (2) and (3) trivially holds, and a sufficient condition for Assumption 2 (1) is that Y_{ijt} has a finite second moment.[8] Note that we do not impose the orthogonality between Z^{-l} and r^l in (4.1), and this is replaced by the weaker condition stated as Assumption 2 (3). That said, it is also plausible to assume that Z^{-l} is orthogonal to r^l in certain applications—for example, because GDP or price does not directly feed back into a distance measure in the gravity analysis of international trade—and Assumption 2 (3) can be trivially satisfied in such cases. We present concrete examples of fixed effect designs that satisfy and fail to satisfy Assumption 2 in Example 2.

Remark 4 (Discussion of the Approximate Sparsity Condition). We emphasize that the approximate sparsity condition in Assumption 2 (together with Assumption 5 (4) to be stated below) allows for many and even all the fixed effects (i.e., $\boldsymbol{\eta}$ as opposed to $\overline{\boldsymbol{\eta}}$) to be nonzero. The assumption should be interpreted as a requirement for how the fixed effects can be decomposed into the sparse components ($\overline{\boldsymbol{\alpha}}$ and $\overline{\boldsymbol{\gamma}}$) and the remaining components ($\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}}$ and $\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}$) generating $R = D_1(\boldsymbol{\alpha}-\overline{\boldsymbol{\alpha}}) + D_2(\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}})$. Indeed, the assumption implicitly imposes a non-trivial restriction on sampling procedures. For example, an i.i.d. sampling of fixed effects is not accommodated.[9] This feature does not, however, contradict our sampling assumption, which is stated below as Assumption 3. With this said, the same limitations apply to all the preceding papers (cf. Section 1) that employ (approximate) sparsity conditions on fixed effects in panel data. In fact, approximate sparsity is a rather plausible assumption for the sampling process in the context of our motivating application (1.1). In gravity analysis of trade, researchers initially used only the G7 countries, later added the Organisation for Economic Co-operation and Development (OECD) countries, and smaller economies have been added more recently. Nearly half of all import and export flows are determined by the top 10 largest economies. Newly added countries to the sample tend to have very small trade volumes. This sampling process entails fixed effects taking smaller values as sample size increases, and it therefore goes along with the approximate sparsity requirement. △

[7]For example, $\sup_{\substack{\|\xi\|=1 \\ \|\xi\|_0=Cs}} \|\xi' Z' R\| = o_p(\sqrt{NM})$ for some finite positive C.

[8]This follows from $E[YY'/NM] = \overline{\eta}' E[ZZ'/NM]\overline{\eta} + E[\varepsilon\varepsilon'/NM] \geq \Lambda_{\min}(\Psi)\|\overline{\eta}\|^2$.

[9]An i.i.d. sampling of fixed effects would entail a fixed proportion of (approximately) nonzero fixed effects relative to the sample size as the sample size increases, and this would contradict the (approximate) sparsity condition.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Assumption 3 (Moments). For each (N,M), the random vectors $(Y'_{ij1}, Z'_{ij1}, \dots, Y'_{ijT}, Z'_{ijT})'$, $(i,j) \in [N] \times [M]$, are independently distributed. Furthermore, there exists $q \in (4,\infty)$ and $K \in (0,\infty)$ not depending on (N,M) such that the following conditions hold for all $l \in [k_0]$.

(1) $\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} E\left[\max_{t\le T}\|X_{ijt}\|_\infty^{2q}\right]\right)^{1/2q} \le B_{NM}$ and $\left(E|X_{ijt,l}|^{2q}\right)^{1/2q} \le K$ hold for all i,j,t,l, where B_{MN} satisfies $B_{NM}\sqrt{\log(p \vee (NM))} \lesssim (NM)^{1/2-1/q}$;

(2) $\|(D_1,D_2)\|_\infty = 1$; and

(3) $\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} E[\varepsilon_{ijt}^{2q}] \vee \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T} E[(\zeta_{ijt}^{l})^{2q}] \le K^{2q} < \infty$.

Discussions of this assumption are found in Remark 5 ahead. For any squared matrix A, define the sparse eigenvalues by

$$\varphi_{\min}(A,m) = \inf_{\substack{\|\xi\|=1\\ \|\xi\|_0\le m}} \xi' A \xi \quad \text{and} \quad \varphi_{\max}(A,m) = \sup_{\substack{\|\xi\|=1\\ \|\xi\|_0\le m}} \xi' A \xi.$$

Define the $p \times p$ rate-adjusted Gram matrix

$$\bar{\Psi} = \begin{bmatrix} \frac{1}{NM}X'X & \frac{1}{M\sqrt{N}}X'D_1 & \frac{1}{N\sqrt{M}}X'D_2 \\ \frac{1}{M\sqrt{N}}D_1'X & \frac{1}{M}D_1'D_1 & \frac{1}{\sqrt{NM}}D_1'D_2 \\ \frac{1}{N\sqrt{M}}D_2'X & \frac{1}{\sqrt{NM}}D_2'D_1 & \frac{1}{N}D_2'D_2 \end{bmatrix}. \tag{5.2}$$

Let $[n] = \{1,\dots,n\}$ for any $n \in \mathbb{N}$. With this notation, we state the following assumption of sparse eigenvalues for the rate-adjusted Gram matrix $\bar{\Psi}$.

Assumption 4 (Sparse Eigenvalues). For any $C > 0$, there exist constants $0 < \underline{k} < \overline{k} < \infty$, which do not depend on (N,M), such that

$$\underline{k} \le \varphi_{\min}(\bar{\Psi}, Cs) \le \varphi_{\max}(\bar{\Psi}, Cs) \le \overline{k}$$

with probability approaching one.

A discussion of this assumption, jointly with Assumption 5, is found in Remark 5 ahead. Since the rate-adjusted Gram matrix $\bar{\Psi}$ has a specific structure consisting of D_1 and D_2 that contain only zeros and ones as their elements, simpler statements of sufficient conditions for Assumption 4 could be written only in terms of the spectra of X. Variants of its sufficient conditions exist in the literature. For example, suppose that $\|X_{ijt}\|_\infty$ is uniformly bounded almost surely and our Assumption 5 (3) holds, then sufficient conditions can be established by adapting Theorem 3.6 and its subsequent Lemmas in Rudelson and Vershynin (2008). That said, we state the weaker statement as in Assumption 4 for the sake of generality. For each (N,M), we write $\Psi = E[\bar{\Psi}]$ depending on (N,M). Using this notation, the auxiliary decomposition (5.1) is made according to the following conditions.

Assumption 5 (Nuisance Parameters). The following conditions are satisfied.

(1) $\max_{l\in[k_0]}\|\phi^l\|_0 \le s_l$ and $\max_{l\in[k_0]}\|\phi^l\| + (s_l)^{-1/2}\|\phi^l\|_1 \le K$;

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

(2) For all $l \in [k_0]$, $\|R^l\| \leq c_{s_l} \lesssim \sqrt{s_l}$;
(3) For all (N,M), $0 < L < \Lambda_{\min}(\Psi) < \Lambda_{\max}(\Psi) < U < \infty$ for L, U independent of (N,M); and
(4) $\max_{l\in[k_0]}(s_l \vee s)\sqrt{\frac{(\log(p\vee(NM)))^2}{N\wedge M}} = o(1)$.

Accounting for possible dependence, we define the cluster-robust variance matrix

$$\Omega = E\left[\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\sum_{t=1}^{T} Z_{ijt}\varepsilon_{ijt}\right)\left(\sum_{t=1}^{T} Z_{ijt}\varepsilon_{ijt}\right)'\right].$$

For each (N,M), we write $\Theta = (E[\frac{Z'Z}{NM}])^{-1}$ depending on (N,M). Let Θ_l denote the lth column of Θ. We state the following assumption of finite and nonzero variance.

Assumption 6 (Variance). For any (N,M) and for all $l \in [k_0]$, $\|\Omega\| < \infty$ and $\Theta_l'\Omega\Theta_l \geq \underline{k} > 0$ for a constant $\underline{k}$ which is independent of the sample size.

Remark 5. Notice that the conditions above are imposed on the Gram matrices, $\bar{\Psi}$ and Ψ, re-weighted by effective sample size, rather than the original Gram matrices, $Z'Z/NM$ and $E[Z'Z/NM]$. Assumption 3 is weaker than the common assumptions required in the literature, such as the subgaussianity or uniform boundedness. With this said, we admit that Assumption 3 (1) would also imply that the dimension of X_{ijt} cannot increase too fast, which can be interpreted as the price that we pay for not imposing the subgaussianity. Assumption 4 is also assumed by Belloni et al. (2012) and Belloni et al. (2016). It requires some small submatrices of the big $p\times p$ re-weighted Gram matrix to be well-behaved. Lower level sufficient conditions are also possible by using Lemma P1 in Belloni et al. (2018), but are not pursued here. Assumption 5 (1) and (2) impose sparsity on the nodewise regression parameters and the approximation errors. Assumption 5 (1) is equivalent to requiring the sparsity of rows of Ψ^{-1}. Assumption 5 (3) requires Ψ, the expectation of the re-weighted Gram matrix, to be positive definite uniform over (N,M). These are rather standard in the literature. Assumption 5 limits the models that can be handled in terms of their dimensionality and sparsity. Note that we need only $ss_l(\log(p\vee(NM)))^2/(N\wedge M) = o(1)$, whereas an adaptation of the proof strategies of Kock (2016) and Kock and Tang (2019) to our framework would entail $ss_l^2(\log(p\vee(NM)))^2/(N\wedge M) = o(1)$. Finally, Assumption 6 requires Ω in the sandwich form to be well-behaved. $\triangle$

Example 2. Recall the exactly sparse FE and the approximately sparse FE from Example 1. These two examples are admissible under our assumptions. In both of these cases, let $\overline{\alpha}_i = \alpha_i$ for all $i \in \{1,\ldots,\lfloor C_\alpha N^{1/3}\rfloor\}$, $\overline{\alpha}_i = 0$ for all $i \in \{\lfloor C_\alpha N^{1/3}\rfloor + 1,\ldots\}$, $\overline{\gamma}_j = \gamma_j$ for all $j \in \{1,\ldots,\lfloor C_\gamma M^{1/3}\rfloor\}$, and $\overline{\gamma}_j = 0$ for all $j \in \{\lfloor C_\gamma M^{1/3}\rfloor + 1,\ldots\}$. In case (b), one can pick a pair of any arbitrary positive constants, C_α and C_γ. Assumption 2 (1) follows from the fact that $\sum_{i=1}^{\infty} 1/\left(i^2\cdot\log(i+1)^3\right)$ is finite. Assumption 2 (2) holds as $||R|| = 0$ in case (a) and $||R||$ is bounded in case (b), while s diverges. Assumption 2 (3) holds under Assumption 3, since $r_{ijt} = 0$ in

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

case (a) and $\sup_{i,j,t} |r_{ijt}| \to 0$ in case (b). On the other hand, the third fixed effect design introduced in Example 1 is not accommodated by our assumptions. Recall:

3. $\alpha_i = (-1)^i$ for all $i \in \mathbb{N}$ and $\gamma_j = (-1)^j$ for all $j \in \mathbb{N}$.

In order to achieve a bounded $||\overline{\boldsymbol{\alpha}}||$ to satisfy Assumption 2 (1), $\overline{\alpha}_i$ has to diminish at the rate of $1/\left(\sqrt{i}\log(i)^{1/2+\delta}\right)$ after rearrangement for some $\delta > 0$. But this forces $||R||$ to grow at the rate of $\sqrt{N}$ under this fixed effect design, whereas we require $||R||$ to be at most of order $\sqrt{N}$ by Assumptions 2 (2) and 5 (4). Since these requirements are not compatible, we rule out this fixed effect design. More generally, fixed effects generated i.i.d. from a nondegenerate distribution are not accommodated by our assumptions. △

The following proposition states that Assumptions 2–6 are sufficient for the high-level conditions in Assumption 1, with a concrete variance formula motivating the practical guideline of Section 3.

PROPOSITION 1. *Assumptions 2–6 imply Assumption 1 with* $V_{ll} = \Theta_l' \Omega \Theta_l$.

A proof is found in Appendix B.2. Combining Theorem 1 and Proposition 1 together, we state the following corollary.

COROLLARY 1 (Asymptotic Normality). *If Assumptions 2–6 are satisfied, then*

$$\sqrt{NM}(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l) \rightsquigarrow N(0, V_{ll})$$

for each $l \in [k_0]$, *where* $V_{ll} = \Theta_l' \Omega \Theta_l$.

5.2. Asymptotic Variance Estimation

Based on the asymptotic variance formula presented in Proposition 1, the cluster-robust asymptotic variance of $\sqrt{NM}\left(\widetilde{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_\ell\right)$ can be estimated by

$$\widehat{V}_{ll} = \widehat{\Theta}_l' \widehat{\Omega} \widehat{\Theta}_l,$$

as suggested in Section 3. This estimator is consistent under the current assumptions as formally stated in the following theorem.

THEOREM 2 (Variance Estimation). *If Assumptions 2–6 are satisfied, then*

$$\max_{l \in [k_0]} \left| \widehat{V}_{ll} - V_{ll} \right| = o_p(1).$$

A proof is found in Appendix B.3.

5.3. Joint Hypotheses Testing

In Corollary 1 and Theorem 2, one can replace $(\tilde{\beta}_l - \beta_l) = e_l'(\tilde{\beta} - \beta)$ by $\rho'(\tilde{\beta} - \beta)$ for any $\rho = (\rho_1, \ldots, \rho_{k_0})' \in \mathbb{R}^{k_0}$ with $\|\rho\|_1 = 1$ and we will still have the asymptotic normality and consistency of the corresponding variance estimator. These results facilitate joint hypothesis testing involving multiple parameters. Suppose that the researcher is interested in testing the hypothesis $H_0 : H'\beta = \theta_0$ for a vector $\theta_0 \in \mathbb{R}^q$ with fixed q where H is a $k_0 \times q$ matrix with rank q with each column H_j normalized to $\|H_j\| = 1$. To this end, define the Wald test statistic by

$$W = NM(H'\tilde{\beta} - \theta_0)' \left\{(H', O')\hat{\Theta}\hat{\Omega}\hat{\Theta}'(H', O')'\right\}^{-1} (H'\tilde{\beta} - \theta_0),$$

where O is an $(N_0 + M_0) \times q$ matrix of zeros. The following Corollary provides the asymptotic distribution of this Wald test statistic.

COROLLARY 2 (Asymptotics of the Wald Test Statistic). *Suppose that Assumptions 2–6 are satisfied with k_0 fixed and that Θ_X, the upper left $k_0 \times k_0$ submatrix of Θ, has its eigenvalues bounded and bounded away from zero. Then, it holds that $\sqrt{NM}\{(\rho', o')\hat{\Theta}\hat{\Omega}\hat{\Theta}'(\rho', o')'\}^{-1/2}\rho'(\tilde{\beta} - \beta) \rightsquigarrow N(0,1)$, where $o = (0, \ldots, 0) \in \mathbb{R}^{N_0+M_0}$, and $\left|(\rho', o')\hat{\Theta}\hat{\Omega}\hat{\Theta}'(\rho', o')' - (\rho', o')\Theta\Omega\Theta'(\rho', o')'\right| = o_p(1)$. Hence,*
$W \rightsquigarrow \chi_q^2$ holds under the null hypothesis $H_0 : H'\beta = \theta_0$.

The proof can be found in Appendix B.4.

5.4. Uniformity

We can further enhance the results of Corollary 1 and Theorem 2, so that they hold uniformly over the set of η with sparse approximation $\overline{\eta}$ that lies in ℓ_0-ball $\mathcal{B}_{\ell_0}(s) := \{\overline{\eta} \in \mathbb{R}^p : \|\overline{\eta}\|_0 \leq s\}$. The asymptotic normality result in Theorem 1 paves the way for statistical inference given a fixed data generating process. Furthermore, this result can be extended to a size control that is uniformly valid over alternative sizes and over alternative data generating processes up to a given sequence of the sparsity parameter. We state this uniformity result as a corollary below.

COROLLARY 3. *If Assumptions 2–6 are satisfied, then*

$$\sup_{a \in \mathbb{R}} \sup_{\eta \in \mathbb{R}^p : \overline{\eta} \in \mathcal{B}_{\ell_0}(s)} \left| P\left(\frac{\sqrt{NM}(\hat{\beta}_l - \beta_l)}{\widehat{V}_{ll}^{1/2}} \leq a \right) - \Phi(a) \right| \to 0$$

for each fixed $l \in [k_0]$

Our proof follows similar strategies to those of Theorem 3 in Caner and Kock (2018a) or Theorem 3 in Kock and Tang (2019), and can be found in Appendix B.5.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

6. SIMULATION STUDIES

6.1. Simulation Setting

Consider the following three fixed effect models of three-dimensional panel data.

$$
\begin{aligned}
\text{Model (I):}\quad & y_{ijt} = x_{ijt}\beta + \alpha_i + \gamma_j + \varepsilon_{ijt,} \\
\text{Model (II):}\quad & y_{ijt} = x_{ijt}\beta + \alpha_i + \gamma_j + \lambda_t + \varepsilon_{ijt,} \\
\text{Model (III):}\quad & y_{ijt} = x_{ijt}\beta + \alpha_{it} + \gamma_{jt} + \varepsilon_{ijt.}
\end{aligned}
$$

Model (I) is nested by Model (II), and Model (II) is in turn nested by Model (III). Therefore, Model (I) is the most parsimonious and subject to under-fitting, whereas Model (III) is the richest and subject to over-fitting. If a researcher runs a fixed effect estimator under Model (I) when Model (II) or (III) is true, then the estimates generally suffer from misspecification biases. If a researcher runs a fixed effect estimator under Model (III) when Model (I) or (II) is true, then the estimates generally suffer from larger standard errors than necessary.

We run simulations for varying sizes of N and $M = N-1$, while the length of time is set to $T = 5$ throughout. This setting follows from our asymptotic theory where N and M increase but T does not. The i and j fixed effects are generated by $\alpha_i = s_\alpha \big/ \left(i \cdot (\log(i+1))^{3/2}\right)$ and $\gamma_j = s_\gamma \big/ \left(j \cdot (\log(j+1))^{3/2}\right)$, where $s_\alpha = s_\gamma = 1$. Notice that these fixed effects designs are introduced in Example 2 as concrete examples that satisfy our approximate sparsity requirements. The t fixed effects are generated by $\lambda_t = 0$ for all t but for one year t when a universal shock of $\lambda_t = 2$ is applied. The it and jt fixed effects are generated by $\alpha_{it} = s_\alpha \big/ \left(i \cdot (\log(i+1))^{3/2}\right)$ and $\gamma_{jt} = s_\gamma \big/ \left(j \cdot (\log(j+1))^{3/2}\right)$, where $s_\alpha = s_\gamma = 1$. We generate X dependently on the fixed effects according to the mixture

$$x_{ijt} = m_x + s_x \cdot \left[(1-\rho) \cdot \tilde{x}_{ijt} + \rho F_{ijt}\right],$$

where $m_x = 0$, $s_x = 2$, $\rho = 0.5$, $\tilde{x}_{ijt} \sim N(0,1)$, and F_{ijt} is the standardized sum of fixed effects for the unit (i,j,t), i.e.,

$$
\begin{aligned}
\text{Under Model (I):}\quad & F_{ijt} = \left(\alpha_i + \gamma_j\right) \Bigg/ \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\alpha_i+\gamma_j\right)^2}, \\
\text{Under Model (II):}\quad & F_{ijt} = \left(\alpha_i + \gamma_j + \lambda_t\right) \Bigg/ \sqrt{\frac{1}{NMT}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\left(\alpha_i+\gamma_j+\lambda_t\right)^2}, \\
\text{Under Model (II):}\quad & F_{ijt} = \left(\alpha_{it} + \gamma_{jt}\right) \Bigg/ \sqrt{\frac{1}{NMT}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}\left(\alpha_{it}+\gamma_{jt}\right)^2}
\end{aligned}
$$

for each $(i,j,t) \in \{1,\ldots,N\} \times \{1,\ldots,M\} \times \{1,\ldots,T\}$. The error term is generated by $\varepsilon_{ijt} \sim N(m_\varepsilon, s_\varepsilon^2)$ independently where $m_\varepsilon = 0$ and $s_\varepsilon = 10$. The main coefficient of interest is set to $\beta = 1$. Each set of simulations consists of 10,000 Monte Carlo iterations of data generation, estimation, and inference.

We compare five methods of estimation and inference. These are the Ordinary Least Squares (OLS) estimator without any individual fixed effects, the fixed effect estimator based on Model (I), the fixed effect estimator based on Model (II), the fixed effect estimator based on Model (III), and our proposed de-biased lasso estimator and post-selection inference.[10] Note that OLS is always under-fitting the true data generating model, and hence is expected to produce misspecification biases. The fixed effect estimator based on Model (I) is correctly specified when the true data generating model is Model (I), but is under-fitting Models (II) and (III). The fixed effect estimator based on Model (II) is over-fitting Model (I), correctly specified when the true data generating model is Model (II), and under-fitting Model (III). The fixed effect estimator based on Model (III) is over-fitting Models (I) and (II), but is correctly specified when the true data generating model is Model (III).

6.2. Simulation Results

Table 1 displays Monte Carlo simulation results under Model (I) (top panel), Model (II) (middle panel), and Model (III) (bottom panel) with the sample size $N = 200$ ($NMT = 1,900$). Similarly, Tables D.1 and D.2 in the Supplementary Material display Monte Carlo simulation results with the smaller sample sizes $N = 10$ ($NMT = 450$) and $N = 15$ ($NMT = 1,050$), respectively. The displayed statistics are the averages, biases, standard deviations, and root mean squared errors of the estimates. Also displayed are the coverage frequencies of the true value of β by the 95% confidence intervals. The first column of each table shows the OLS results without any individual fixed effects. The next three columns of each table show results of fixed effect estimators based on estimating equations of Models (I)–(III). We shall call them FE-I, FE-II, and FE-III for succinctness. The last two columns of each table show results of our proposed de-biased lasso estimator with valid post-selection inference based on (2.1) and (2.2). We shall call them POST (2.1) and POST (2.2) for succinctness.

In the top panel of Table 1, where the true data generating model is Model (I), OLS is biased while FE-I, FE-II, and FE-III yield little bias. These results are consistent with the current simulation setting as OLS misspecifies the true model while FE-I, FE-II, and FE-III correctly specify the true model. The bias of POST (2.1) is between that of OLS and those of FE-I, FE-II, and FE-III but much closer to the latter group. In other words, POST (2.1) and POST (2.2) are de-biased to a large extent but not to the full extent so that desired balance between the bias

[10]Throughout, we use cross validation to choose the tuning parameter. This is implemented by the `cv.glmnet` function in R. For penalty weights, we set 1 for X, $1/\sqrt{N}$ for D_1, and $1/\sqrt{M}$ for D_2.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

TABLE 1. Monte Carlo Simulation Results Under Model (I) (Top Panel), Model (II) (Middle Panel), and Model (III) (Bottom Panel) with Size $N = 20$ ($NMT = 1,900$).

True model = (I)		Fixed effect estimators			**POST**	POST
$N = 20$ ($NMT = 1,900$)	OLS	FE-I	FE-II	FE-III	**(2.1)**	(2.2)
Under-fitting or over-fitting	Under	Correct	Over	Over	**Robust**	Robust
Average	1.256	1.000	0.999	0.998	**0.981**	1.027
Bias	0.256	0.000	−0.001	−0.002	**−0.019**	0.027
Standard deviation	0.165	0.233	0.233	0.243	**0.205**	0.206
Root mean square error	0.305	0.233	0.233	0.243	**0.206**	0.207
95% Coverage	0.656	0.946	0.945	0.934	**0.956**	0.957
True model = (II)		Fixed effect estimators			**POST**	POST
$N = 20$ ($NMT = 1,900$)	OLS	FE-I	FE-II	FE-III	**(2.1)**	(2.2)
Under-fitting or over-fitting	Under	Under	Correct	Over	**Robust**	Robust
Average	1.255	1.257	0.999	1.003	**1.037**	1.051
Bias	0.255	0.257	−0.001	0.003	**0.037**	0.051
Standard deviation	0.170	0.219	0.233	0.243	**0.209**	0.210
Root mean square error	0.307	0.338	0.233	0.243	**0.212**	0.216
95% Coverage	0.676	0.777	0.945	0.933	**0.953**	0.951
True model = (III)		Fixed effect estimators			**POST**	POST
$N = 20$ ($NMT = 1,900$)	OLS	FE-I	FE-II	FE-III	**(2.1)**	(2.2)
Under-fitting or over-fitting	Under	Under	Under	Correct	**Robust**	Robust
Average	1.425	1.395	1.302	0.999	**1.049**	1.051
Bias	0.425	0.395	0.302	−0.001	**0.049**	0.051
Standard deviation	0.173	0.180	0.194	0.243	**0.218**	0.214
Root mean square error	0.459	0.434	0.359	0.243	**0.223**	0.220
95% Coverage	0.306	0.399	0.647	0.934	**0.944**	0.943

and variance is maintained. OLS yields a smaller standard deviation than FE-I or FE-II, and FE-III yields by far the largest standard deviation. These results are also consistent with the fact that OLS is the most parsimonious while FE-III is the most redundant in specification. The standard deviation of POST (2.1) is between that of OLS and those of FE-I, FE-II, and FE-III. Furthermore, POST (2.1) yields an even smaller root mean square error than the oracle estimator, FE-I. The coverage frequency by FE-I, as the oracle estimator, is closer to the nominal level 95% than those of OLS, FE-II, or FE-III. Furthermore, POST (2.1) yields a coverage frequency as close to the nominal level as the oracle estimator, FE-I.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

TABLE 2. Monte Carlo Fractions of Nonzero Estimates of the Four Types of Fixed Effects, α_i, γ_j, α_{it}, and γ_{jt}, Under Each of the Models (I), (II), and (III).

True model	α_i	γ_j	α_{it}	γ_{jt}
(I)	0.107	0.104	0.065	0.060
(II)	0.136	0.132	0.149	0.140
(III)	0.171	0.184	0.286	0.252

We can make similar observations both in the middle panel and in the bottom panel of Table 1, where the true data generating models are Models (II) and (III), respectively. To avoid repetitive writing, we relegate detailed discussion of these simulation results to Appendix C in the Supplementary Material. Summarizing these omitted details, we once again confirm superior simulation performance of POST (2.1) over the respective oracle estimators in these two panels as well.

The simulation results reported above demonstrate that the proposed method (POST (2.1)) can be used as a robustly applicable method of inference when a researcher does not know the correct fixed effect specification in practice. POST (2.1) is more precise than biased parsimonious estimators, is more efficient than redundant estimators, and allows for at least as accurate inference as the oracle estimator.

Finally, we report how many of each type of the fixed effects, α_i, γ_j, α_{it}, and γ_{jt}, are selected to be zero/nonzero by POST (2.1) under each of the Models (I), (II), and (III). Table 2 summarizes the fractions of nonzero estimates for each of the four types of the fixed effects. Under Model (I), most (93–94%) of the α_{it} and γ_{jt} fixed effects are estimated to be exact zeros, as expected. Under Model (II), less but still most (85–86%) of the α_{it} and γ_{jt} fixed effects are estimated to be exact zeros. Under Model (III), on the other hand, a much smaller fraction (71–75%) of the α_{it} and γ_{jt} fixed effects are estimated to be exact zeros, as expected.

6.3. Alternative Data Generating Processes

Following Examples 1 and 2, we employed a fixed effect design that satisfies our decomposition and approximate sparsity assumption. Therefore, as expected from the theory, we presented simulation results for the case in which our proposed method works and exhibits robust and superior performance.

We also ran additional simulations based on alternative fixed effect designs. The first alternative design is similar to our baseline design in the sense that the scale diminishes at the rate of $1/\left(i\cdot(\log(i+1))^{3/2}\right)$, except that we allow for stochastically generated fixed effects—see Appendix D.2.1 in the Supplementary material. Simulation results are qualitatively very similar to those that we presented above, and hence we draw the same conclusion that our proposed method exhibits robust and superior performance even if we allow for stochastically generated fixed effects.

The second alternative design is based on the counter-example introduced in Examples 1 and 2. Specifically, $\alpha_i = (-1)^i$ for all $i \in \mathbb{N}$ and $\gamma_j = (-1)^j$ for all $j \in \mathbb{N}$ – see Appendix D.2.2 in the Supplementary Material. As emphasized in Examples 1 and 2, this design violates our assumptions. Interestingly, however, POST (2.1) still performs better than the other estimators in terms of the root mean square error. On the other hand, the coverage accuracy by POST (2.1) is not as good as that by the oracle estimator, as expected. We observe essentially the same pattern of simulation results under the third alternative design, where the fixed effects are generated i.i.d. from a nondegenerate distribution—see Appendix D.2.3 in the Supplementary Material.

7. GRAVITY ANALYSIS OF INTERNATIONAL TRADE

In this section, we present a gravity analysis of international trade. We retrieved data from the International Monetary Fund (IMF's), Direction of Trade Statistics (DOTS) Database, a common source of trade flows and trade costs used in gravity analysis. The DOTS data report exports and imports of merchandise goods by import source and export destination. This is a standard source of trade flow data used in gravity analysis—see Head and Mayer (2014). After removing observations with zero trade volumes, our data set is an unbalanced panel of 103,502 bilateral trade pairs of 188 countries for the years 1995, 2000, 2005, 2010, and 2015.

7.1. Sparsity of Fixed Effects

Since the number of bilateral observations is much larger than the number of fixed effects, we may consistently estimate the fixed effects even under the richest specification. In this light, we first use estimated fixed effects to examine the plausibility of our key assumption of the approximately sparse fixed effects. Following standard gravity equations, we include the logarithm of distance between the origin i and destination j, a dummy variable taking the value of 1 if the origin i and destination i are contiguous, and a dummy variable taking the value of 1 if the origin i and destination j share a common language.

Figure 1 displays histograms of estimated fixed effects α_{it} (left panel) and γ_{it} (right panel) in absolute value. Both histograms graphically indicate that most of the estimated fixed effects concentrate around zero. This pattern implies the plausibility of our assumption of the approximate sparsity. Specifically, let $\overline{\alpha}_{it} = \alpha_{it}$ if $|\alpha_{it}| \geq 1$ and $\overline{\alpha}_{it} = 0$ otherwise. Likewise, let $\overline{\gamma}_{it} = \gamma_{it}$ if $|\gamma_{it}| \geq 1$ and $\overline{\gamma}_{it} = 0$ otherwise. Then, we have $\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{T-1} \left(\mathbf{d}'_{1,it}(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}})\right)^2 \approx 135.102$, $\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{T-1} \left(\mathbf{d}'_{2,it}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})\right)^2 \approx 166.684$, and $\|\boldsymbol{\beta}\|_0 + \|\overline{\boldsymbol{\alpha}}\|_0 + \|\overline{\boldsymbol{\gamma}}\|_0 \approx 481$ (out of the total of 1,884 dimensions). Therefore, the decompositions (2.4) and (2.5)

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

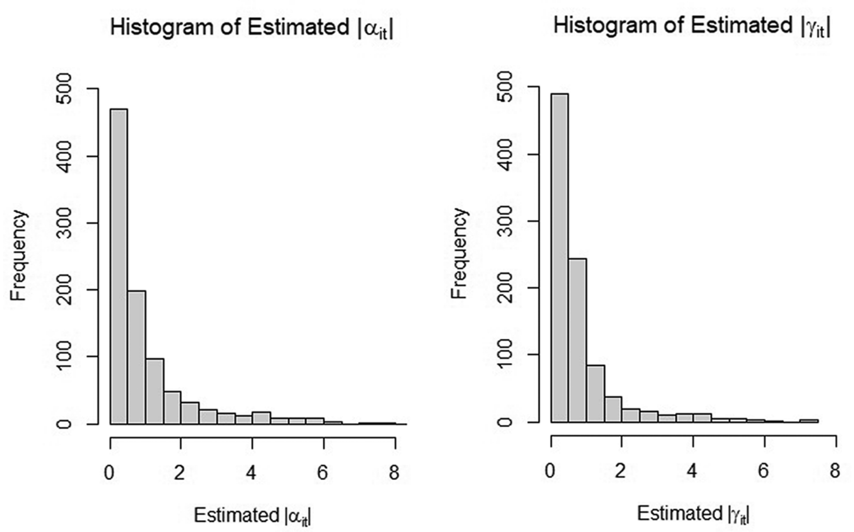

FIGURE 1. Histograms of estimated fixed effects α_{it} (left panel) and γ_{it} (right panel) in absolute value.

subject to Assumptions 2–5 are reasonable for this common international trade data set used for gravity analysis.

7.2. Estimation and Inference

We next apply our proposed method of estimation and inference to this data set. As in Section 7.1, we consider the standard gravity equation, which includes the logarithm of distance between the origin i and destination j, a dummy variable taking the value of 1 if the origin i and destination j are contiguous, and a dummy variable taking the value of 1 if the origin i and destination j share a common language. The estimate and its standard error are computed for each of these three key regressors. We focus on POST (2.1) for its robust and superior performance as demonstrated through our simulation studies in Section 6.[11] While we present results based on the entire data set consisting of the years 1995, 2000, 2005, 2010, and 2015, we also present results based on two subsamples of the data set as well—one is the subsample that consists of the years 1995 and 2000, and the other is the subsample that consists of the years 2010 and 2015. The motivation to split the data in this manner is to avoid pooling data across two possibly different structural regimes divided by a number of historical events that happened during 2001–2009,

[11] As in simulations, we use cross validation to choose the tuning parameter. This is implemented by the `cv.glmnet` function in R. For penalty weights, we set 1 for X, $1/\sqrt{N}$ for D_1, and $1/\sqrt{M}$ for D_2.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Table 3. Estimates and their Standard Errors (in Parentheses) for Log Distance, Contiguity, and Common Language in Gravity Analysis of International Trade. The full sample consists of the years 1995, 2000, 2005, 2010, and 2015.

Dependent variable: log trade volume	Full sample	1995 & 2000	2000 & 2015
Log distance	−1.677	−1.544	−1.713
	(0.016)	(0.023)	(0.023)
Contiguity	0.901	0.791	0.852
	(0.001)	(0.002)	(0.002)
Common language	0.746	0.765	0.769
	(0.008)	(0.009)	(0.011)
Proportion of zero α_i	0.005	0.000	0.005
Proportion of zero α_{it}	0.081	0.170	0.218
Proportion of zero γ_j	0.016	0.011	0.021
Proportion of zero γ_{jt}	0.096	0.176	0.218

such as the accession of China to the World Trade Organization (WTO) in 2001 and the great recession of 2008 among others.

Table 3 summarizes the results. The first column shows the results based on the full sample consisting of the years 1995, 2000, 2005, 2010, and 2015. The second column shows the results based on the subsample consisting of the years 1995 and 2000. The third column shows the results based on the subsample consisting of the years 2010 and 2015. Across all three sets of results, we make the following three points of observations. First, the coefficient of log distance is significantly negative. Specifically, a one percent increase in geographical distance results in about a 1.5–1.7% reduction in trade volumes. Second, coefficients of contiguity and common language are significantly positive. In other words, trade volumes tend to be greater between pairs of countries that are contiguous and share a common language.

Third, most of the α_i and γ_j fixed effects are selected as nonzero effects, while larger proportions of α_{it} and γ_{jt} fixed effects are zero, implying that time heterogeneity matters much less than country heterogeneity. Indeed, in each subsample, roughly one fifth of all of the time and country-specific fixed effects are set to zero using our proposed approach. This fraction of zero estimates is at odds with the exact sparsity, but most of the nonzero estimates in fact concentrate around near zero as shown in the kernel density plots in Figure 2 implying that the approximate sparsity assumption may well be satisfied.

Comparing the estimates in Table 3 across different time periods, observe that the trade elasticity has increased from the period 1995–2000 to the period 2010–2015. Specifically, a 1% increase in geographical distance results in about a 1.5% reduction in trade volume during the period 1995–2000, and about a 1.7%

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

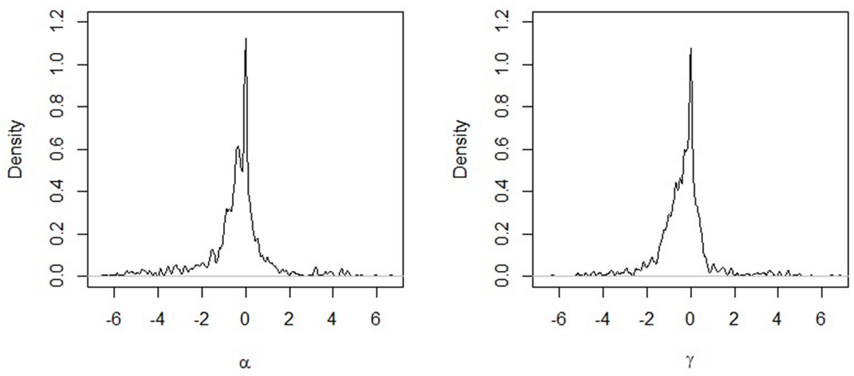

FIGURE 2. Kernel density plots of estimated fixed effects α (left) and γ (right).

reduction in trade volume during the period 2010–2015. In this sense, our findings suggest that distance became a greater deterrent to international trade after 2008, consistent with greater geographic concentration of global supply chains.

8. SUMMARY AND DISCUSSIONS

Three-dimensional panel models are used widely in empirical analysis of international trade, housing, migration, and consumer prices, among other applications. Empirical researchers use various combinations of fixed effects for three-dimensional panels. When a researcher imposes a parsimonious model and the true model is rich, then estimation based on the assumed parsimonious model generally incurs misspecification biases. When a researcher employs a rich model and the true model is parsimonious, then estimation based on the redundantly rich model generally incurs larger standard errors than necessary. It is therefore useful for researchers to have a mechanism to determine an accurate specification in applications. With this motivation, Lu, Miao, and Su (2021) propose methods of model selection in three-dimensional panel data. In this paper, we advance this literature by proposing a method of post-selection inference for regression parameters. We use the lasso technique as the means of model selection and to de-bias the lasso estimate, but our assumptions allow for many and even all fixed effects to be nonzero. Furthermore, we discuss the plausibility of the assumption of approximately sparse fixed effects in the context of gravity analysis of international trade. Simulation studies demonstrate that the proposed method is less biased than fixed-effect estimators based on parsimonious models, is more efficient than fixed-effect estimators based on redundant models, and allows for as accurate inference as the oracle estimator.

We emphasize that the objective of our method and theory lies in post-selection inference on β. In other words, model selection per se is out of the scope of this paper for the following two reasons. First, using the lasso-based approach as in our method for the purpose of consistent model selection would require imposing an additional strong assumption, such as the beta-min assumption, which we do not assume in this paper. Second, alternative existing approaches, such as the one taken by Lu, Miao, and Su (2021) accomplish model selection without requiring such an assumption. We therefore remark that our method should be used primarily for the purpose of making inferences about β, and not for the purpose of routine model selection.

We conclude the paper by suggesting a few directions for future research. First, our model framework does not allow for *ij* fixed effects, while i, j, t, *it*, and *jt* fixed effects are allowed. Although allowing for *ij* fixed effects is not of interest in our motivating example,[12] it may be possible to allow for such fixed effects provided that the asymptotic setting allows for large T as well as large N and/or large M. Formal theoretical development for this case is left for future research. Second, although we focus on three-dimensional panels for their relevance to important applications as discussed in Section 1, the proposed methodology can be extended to higher-dimensional panel models with more than three-dimensions of fixed effects. Finally, our framework focuses on shrinkage over fixed effects, as is the case with a few other preceding papers cited in Section 1. Since (approximately) non-zero fixed effects can be considered to absorb outliers, the proposed method can be also interpreted as an outlier-robust estimation method. In this sense, it complements the existing robust M-estimation methods such as the quantile regression and regression with Huber loss. We desire to see further research on comparison between absolute shrinkage of fixed effects and existing robust methods.

A. MATHEMATICAL APPENDIX

We provide proofs of the main results in Appendix B. Auxiliary lemmas and their proofs are found in Appendix B in the Supplementary Material. Throughout, we use the following short-hand notation: $Q = S/\sqrt{NM}$ and $a = p \vee (NM)$. Also, for a matrix A, denote $\|A\|_\infty = \max_{i,j} |A_{i,j}|$.

B. PROOFS OF THE MAIN RESULTS

B.1. Proof of Theorem 1

Proof. The K.K.T. condition for the lasso program (3.1) gives

$$-Z'(Y - Z\widehat{\boldsymbol{\eta}}) + \mu P'(\widehat{\boldsymbol{\eta}}) = 0.$$

[12] In gravity models for international trade, the main parameters of interest are the coefficient of DIST_{ij}, interpreted as the trade elasticity or trade cost, and the coefficient of TA_{ij}, interpreted as the effects of bilateral trade agreements on trade volume. These parameters will not be identified once *ij* fixed effects enter the model.

Substituting the representation (2.6) yields

$$Z'Z(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})+\mu P'(\widehat{\boldsymbol{\eta}})=Z'\varepsilon+Z'R.$$

Multiplying both sides by $\widehat{\Theta}_l'/\sqrt{NM}$, we have

$$\frac{1}{\sqrt{NM}}\widehat{\Theta}_l'Z'Z(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})+\frac{\mu\widehat{\Theta}_l'P'(\widehat{\boldsymbol{\eta}})}{\sqrt{NM}}=\frac{\widehat{\Theta}_l'Z'\varepsilon}{\sqrt{NM}}+\frac{\widehat{\Theta}_l'Z'R}{\sqrt{NM}}.$$

Therefore, we have

$$\sqrt{NM}e_l'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})+\sqrt{NM}\left(\widehat{\Theta}_l'Z'Z/(NM)-e_l'\right)(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})+\frac{\mu\widehat{\Theta}_l'P'(\widehat{\boldsymbol{\eta}})}{\sqrt{NM}}=\frac{\widehat{\Theta}_l'Z'\varepsilon}{\sqrt{NM}}+\frac{\widehat{\Theta}_l'Z'R}{\sqrt{NM}}$$

or

$$\sqrt{NM}e_l'\left(\underbrace{\widehat{\boldsymbol{\eta}}+\frac{\mu}{NM}\widehat{\Theta}'P'(\widehat{\boldsymbol{\eta}})}_{=\widetilde{\eta}}-\overline{\boldsymbol{\eta}}\right)+\sqrt{NM}\left(\widehat{\Theta}_l'Z'Z/(NM)-e_l'\right)(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})=\frac{\widehat{\Theta}_l'Z'\varepsilon}{\sqrt{NM}}+\frac{\widehat{\Theta}_l'Z'R}{\sqrt{NM}}.$$

By Assumption 1 (i) and (ii) and the definition of the de-biased lasso in (4.1), we obtain

$$\sqrt{NM}e_l'(\widetilde{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})=\frac{1}{\sqrt{NM}}\widehat{\Theta}_l'Z'\varepsilon+o_p(1).$$

Applying Assumption 1 (iii) for each $l\in[k_0]$ yields the weak convergence result. □

B.2. Proof of Proposition 1

Proof. The sufficiency of Assumptions 2–5 for Assumption 1 (i) is provided in Lemma B.5 in the Supplementary Appendix. The sufficiency of Assumptions 2–5 for Assumption 1 (ii) is provided in Lemma B.6 in the Supplementary Appendix. The sufficiency of Assumptions 3–6 for Assumption 1 (iii) is provided in Lemma B.7 in the Supplementary Appendix. □

B.3. Proof of Theorem 2

Proof. We introduce an intermediate object defined by

$$\tilde{\Omega}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)'.$$

Lemma B.8 under Assumptions 2–5 yields $\max_{l\in[p]}\|\hat{\Theta}_l\|_0\le Cs_l$ with probability $1-o(1)$ for some C large enough for all $l\in[k_0]$. Therefore, we obtain the decomposition

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

$$
\begin{aligned}
&|\hat{\Theta}_l'\hat{\Omega}\hat{\Theta}_l - \Theta_l'\Omega\Theta_l| \\
\leq &|\hat{\Theta}_l'\hat{\Omega}\hat{\Theta}_l - \hat{\Theta}_l'\Omega\hat{\Theta}_l| + |\hat{\Theta}_l'\Omega\hat{\Theta}_l - \Theta_l'\Omega\Theta_l| \\
\leq &\|\hat{\Theta}_l\|^2 \max_{\substack{\|\xi\|=1 \\ \|\xi\|_0 \leq Cs_l}} \xi'(\hat{\Omega}-\tilde{\Omega})\xi + \|\hat{\Theta}_l\|_1^2\|\tilde{\Omega}-\Omega\|_\infty + \|\hat{\Theta}_l - \Theta_l\|^2 \\
&\max_{\substack{\|\xi\|=1 \\ \|\xi\|_0 \leq Cs_l}} \xi'\Omega\xi + 2\|\Omega\Theta_l\|\|\hat{\Theta}_l - \Theta_l\| \qquad \textbf{(B.1)}
\end{aligned}
$$

for all $l \in [k_0]$. By Lemma B.4 under Assumptions 3–5, it suffices to bound $\max_{\substack{\|\xi\|=1 \\ \|\xi\|_0 \leq Cs_l}} \xi'(\hat{\Omega}-\tilde{\Omega})\xi$and $\|\hat{\Theta}_l\|_1^2\|\tilde{\Omega}-\Omega\|_\infty$ on the right hand side.

We first bound $\|\hat{\Theta}_l\|_1^2\|\tilde{\Omega}-\Omega\|_\infty$ on the right hand side of (B.1). Observe that $\max_{l\in[p]}\|\hat{\Theta}_l\|_0 = O(s_l)$ with probability approaching one due to Lemma B.8. In addition, we have $\max_{1\leq l\leq p}\|\hat{\Theta}_l\| = O_p(1)$. To see this, notice that $\max_{1\leq l\leq p}\|\Theta_l\| = O(1)$ following Assumption 5 (1) and the fact that $\max_{1\leq l\leq p}|1/\tau_l^2| = O(1)$ from equation (B.11) in Lemma B.4, which holds under Assumptions 3–5. We also have $\max_{1\leq l\leq p}\|\hat{\Theta}_l - \Theta_l\| = o_p(1)$ from Lemma 4 under Assumptions 3–5. These facts together yield that $\|\hat{\Theta}_l\|_1 = O_p(\sqrt{s_l})$ uniformly over $l \in [k_0]$.

By an application of Lemma B.2, we have

$$
\begin{aligned}
\|\tilde{\Omega}-\Omega\|_\infty &\leq T^2 \max_{t\in[T]}\max_{l\in[p]}\left|\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M (Z_{ijt,l}^2\varepsilon_{ijt}^2 - E[Z_{ijt,l}^2\varepsilon_{ijt}^2])\right| \\
&\lesssim \sqrt{\frac{\sigma^2\log a}{NM}} + \frac{B\log a}{NM}
\end{aligned}
$$

with probability at least $1-o(1)$, where

$$
\begin{aligned}
\sigma^2 &= \max_{t\in[T],l\in[p]} \frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M E[Z_{ijt,l}^4\varepsilon_{ijt}^4] \\
&\leq \max_{t\in[T],l\in[p]} \sqrt{\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M E[Z_{ijt,l}^8]}\sqrt{\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M E[\varepsilon_{ijt}^8]} = O(1)
\end{aligned}
$$

under Assumption 3, and

$$
\begin{aligned}
B^2 &= E[\max_{i,j,t}\|Z_{ijt}\varepsilon_{ijt}\|_\infty^2] \\
&\leq (E[\max_{i,j,t}\|Z_{ijt}\varepsilon_{ijt}\|_\infty^q])^{2/q} \\
&\leq (NM)^{2/q}\left(\frac{1}{NM}\sum_{i=1}^N\sum_{j=1}^M\sum_{t=1}^T E[\|Z_{ijt}\varepsilon_{ijt}\|_\infty^q]\right)^{2/q}
\end{aligned}
$$

$$
\leq (NM)^{2/q}\left\{\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}E[\|Z_{ijt}\|_{\infty}^{2q}]\right)^{1/2}\left(\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{t=1}^{T}E[\varepsilon_{ijt}^{2q}]\right)^{1/2}\right\}^{2/q}
$$

$$
\lesssim (NM)^{2/q}B_{NM}^{2}O(1)
$$

under Assumption 3. Therefore, we obtain

$$
\sqrt{\frac{\sigma^2\log a}{NM}}+\frac{B\log a}{NM}\lesssim\sqrt{\frac{\log a}{NM}}+\frac{B_{NM}\log a}{(NM)^{1-1/q}}=O\left(\sqrt{\frac{\log a}{NM}}\right),
$$

where the last rate follows from Assumption 3 (i). Combining these results, we obtain

$$
\|\hat{\Theta}_l\|_1^2\|\tilde{\Omega}-\Omega\|_{\infty}=O_p\left(\sqrt{\frac{s_l^2\log a}{NM}}\right).
$$

We next bound $\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\leq Cs_l}}\xi'(\hat{\Omega}-\tilde{\Omega})\xi$ on the right hand side of (B.1). Note that $\hat{\varepsilon}=\varepsilon+R-Z(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})$. Thus,

$$
\begin{aligned}
&\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\leq Cs_l}}\xi'(\hat{\Omega}-\tilde{\Omega})\xi\\
&=\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\leq Cs_l}}\xi'\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{NM}\left\{\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)'\right.\\
&\quad-\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)'\\
&\quad+\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)'+\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)'\\
&\quad-\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)\left(\sum_{t=1}^{T}Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)'-\left(\sum_{t=1}^{T}Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)\left(\sum_{t=1}^{T}Z_{ijt}\varepsilon_{ijt}\right)'\\
&\quad-\left(\sum_{t=1}^{T}Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)\left(\sum_{t=1}^{T}Z_{ijt}r_{ijt}\right)'\\
&\quad\left.+\left(\sum_{t=1}^{T}Z_{ijt}X_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)\left(\sum_{t=1}^{T}Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)'\right\}\xi\\
&=:(1)+(2)+(3)+(4)+(5)+(6)+(7)+(8).
\end{aligned}
$$

We bound each of the last eight terms separately. First, Cauchy–Schwartz's inequality yields

$$
\begin{aligned}
(8) &= \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \xi' \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{1}{NM}\left\{\left(\sum_{t=1}^{T} Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)\left(\sum_{t=1}^{T} Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)'\right\}\xi \\
&\le T^2 \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \xi' Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})' Z_{ijt}Z_{ijt}'\xi \\
&\lesssim \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\xi' Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)^2}\sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left((\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})' Z_{ijt}Z_{ijt}'\xi\right)^2}.
\end{aligned}
$$

Due to the sparsity of all the feasible ξ, we have $\|\xi\|_1 \le \sqrt{s_l}\|\xi\|$. Thus, by Assumption 3, Lemma B.1, and Lemma B.3 with $\mu = C\sqrt{NM\log a}$ under Assumptions 2, 3 (1), and 4, we have

$$
\begin{aligned}
&\max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\xi' Z_{ijt}Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)^2 \\
&\le \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} (\max_{i,j} |\xi' Z_{ijt}|^2)\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Z_{ijt}'(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\right)^2 \\
&\le \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} (\max_{i,j} \|\xi'\|_1^2\cdot(1\vee\|X_{ijt}\|_\infty^2))\frac{1}{NM}\|Z(\widehat{\boldsymbol{\eta}}-\overline{\boldsymbol{\eta}})\|^2 \\
&\lesssim s_l\cdot O_p(1\vee E[\max_{i,j,t}\|X_{ijt}\|_\infty^2])O_p\left(\frac{s\log a}{NM}\right) = O_p\left(\frac{s\cdot s_l B_{NM}^2\log a}{(NM)^{1-1/q}}\right). \qquad \textbf{(B.2)}
\end{aligned}
$$

Therefore, $(8) = O_p\left(\frac{s\cdot s_l B_{NM}^2\log a}{(NM)^{1-1/q}}\right)$. Similarly, for (1) and (3), we have

$$
\begin{aligned}
&\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \xi' \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{1}{NM}\left\{\left(\sum_{t=1}^{T} Z_{ijt}r_{ijt}\right)\left(\sum_{t=1}^{T} Z_{ijt}\varepsilon_{ijt}\right)'\right\}\xi \\
&\le T^2 \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \xi' Z_{ijt}r_{ijt}\varepsilon_{ijt}Z_{ijt}'\xi \\
&\le T^2 \max_{t\in[T]} \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\xi' Z_{ijt}r_{ijt})^2}\sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\varepsilon_{ijt}Z_{ijt}'\xi)^2}.
\end{aligned}
$$

Thus, by Assumptions 2 and 3 (1),

$$
\begin{aligned}
\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\xi' Z_{ijt} r_{ijt})^2 &\le \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM} \max_{i,j,t} |\xi' Z_{ijt}| \sum_{i=1}^{N}\sum_{j=1}^{M} r_{ijt}^2 \\
&\le \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM} \|\xi\|_1^2 \cdot (1 \vee \max_{i,j,t} \|Z_{ijt}\|_\infty^2) \cdot s \\
&= O_p\left(\frac{s \cdot s_l B_{NM}^2}{(NM)^{1-1/q}}\right) \qquad \textbf{(B.3)}
\end{aligned}
$$

for all feasible ξ. Since $Z'Z/NM = Q\bar{\Psi}Q$ and$\|Q\xi\| \le \|\xi\|$,

$$
\begin{aligned}
\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\varepsilon_{ijt} Z_{ijt}'\xi)^2 &\le \max_{i,j,t} |\varepsilon_{ijt}|^2 \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(Z_{ijt}'\xi)^2 \\
&\le \max_{i,j,t} |\varepsilon_{ijt}|^2 \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \xi' Q\bar{\Psi}Q\xi \\
&\le \max_{i,j,t} |\varepsilon_{ijt}|^2 \max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}} \xi' \bar{\Psi}\xi \\
&\le O_p\left(E\max_{i,j,t} |\varepsilon_{ijt}|^2\right) \varphi_{\max}^2(\bar{\Psi}, Cs_l) = O_p\left((NM)^{1/q}\right), \qquad \textbf{(B.4)}
\end{aligned}
$$

where the second inequality is due to Assumption 4 and the last uses Assumption 3 (3).

Since all the remaining terms consist of the products of the above three components, using (B.2)–(B.4), we obtain

$$
\begin{aligned}
\max_{\|\xi\|=1 \|\xi\|_0\le Cs_l} \xi'(\hat{\Omega}-\tilde{\Omega})\xi \le{}& O_p\left(\sqrt{\frac{s \cdot s_l B_{NM}^2}{(NM)^{1-2/q}}}\right) + O_p\left(\sqrt{\frac{s \cdot s_l B_{NM}^2 \log a}{(NM)^{1-2/q}}}\right) \\
&+ O_p\left(\sqrt{\frac{s \cdot s_l B_{NM}^2}{(NM)^{1-2/q}}}\right) + O_p\left(\frac{s \cdot s_l B_{NM}^2}{(NM)^{1-1/q}}\right) \\
&+ O_p\left(\frac{s \cdot s_l B_{NM}^2 \sqrt{\log a}}{(NM)^{1-1/q}}\right) + O_p\left(\sqrt{\frac{s \cdot s_l B_{NM}^2 \log a}{(NM)^{1-2/q}}}\right) \\
&+ O_p\left(\frac{s \cdot s_l B_{NM}^2 \sqrt{\log a}}{(NM)^{1-1/q}}\right) + O_p\left(\frac{s \cdot s_l B_{NM}^2 \log a}{(NM)^{1-1/q}}\right) \\
={}& O_p\left(\sqrt{\frac{s \cdot s_l B_{NM}^2 \log a}{(NM)^{1-2/q}}}\right).
\end{aligned}
$$

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Using the rate for $\max_{l\in[k_0]}\|\hat{\Theta}_l-\Theta_l\|$ from Lemma B.4 under Assumptions 2, 3, and 6, we have

$$(\text{B.1}) = O_p\left(\sqrt{\frac{s\cdot s_l B_{NM}^2 \log a}{(NM)^{1-2/q}}}\right) + \frac{s_l \log a}{NM}O(1) + O(1)O_p(1)O_p\left(\sqrt{\frac{s_l\log a}{NM}}\right) = o_P(1)$$

as desired. □

Remark 6. As emphasized in the main text, recall that Assumption 5 (4) requires $ss_l(\log(p\vee(NM)))^2/(N\wedge M)=o(1)$ instead of $ss_l^2(\log(p\vee(NM)))^2/(N\wedge M)=o(1)$. This is due to the fact that we made use of the bound $|\hat{\Theta}_l'\hat{\Omega}\hat{\Theta}_l-\hat{\Theta}_l'\tilde{\Omega}\hat{\Theta}_l|\le\|\hat{\Theta}_l\|^2\max_{\substack{\|\xi\|=1\\ \|\xi\|_0\le Cs_l}}\xi'(\hat{\Omega}-\tilde{\Omega})\xi$ with probability approaching unity following Lemma B.8. On the other hand, in Kock (2016) and Kock and Tang (2019), the bound based on the dual norm inequality $|\hat{\Theta}_l'\hat{\Omega}\hat{\Theta}_l-\hat{\Theta}_l'\tilde{\Omega}\hat{\Theta}_l|\le\|\hat{\Theta}_l\|_1^2\|\hat{\Omega}-\tilde{\Omega}\|_\infty$ is used in place. △

B.4. Proof of Corollary 2

Proof. The first statement follows from a minor modification of the proofs of Corollary 1 and Theorem 2 with the facts that $\|\rho'\hat{\Theta}\|_1=\left\|\sum_{l\in[k_0]}\rho_l\hat{\Theta}_l\right\|_1\le\sum_{l\in[k_0]}|\rho_l|\max_{l\in[k_0]}\left\|\hat{\Theta}_l\right\|_1\le\max_{l\in[k_0]}\left\|\hat{\Theta}_l\right\|_1$, $\|\Theta(\rho',o')'\|\le\Lambda_{\max}(\Theta_X)\|\rho\|\le 1/\Lambda_{\max}(\Theta_X^{-1})=O(1)$, and $\|(\hat{\Theta}-\Theta)'(\rho',o')'\|=\left\|\sum_{k\in[k_0]}\left(\hat{\Theta}_l-\Theta_l\right)\rho_l\right\|\le\sum_{k\in[k_0]}\left\|\hat{\Theta}_l-\Theta_l\right\||\rho_l|\le\max_{l\in[k_0]}\left\|\hat{\Theta}_l-\Theta_l\right\|$. The second statement follows from the first one and the fact that $\sqrt{NM}\{(H',O')\hat{\Theta}\hat{\Omega}\hat{\Theta}'(H',O')'\}^{-1/2}(H'\tilde{\beta}-\theta_0)\rightsquigarrow N(0,I_q)$, where I_q is the $q\times q$ identity matrix. □

B.5. Proof of Corollary 3

Proof. Our proof follows the same strategy as those of Theorem 3 in Caner and Kock (2018a) or Theorem 3 in Kock and Tang (2019). Fix an $l\in[k_0]$. For an $\epsilon>0$, define

$$A_{1,(N,M)} := \left\{\sup_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)}\left|\sqrt{NM}\left(\widehat{\Theta}_l'Z'Z/(NM)-e_l'\right)(\widehat{\eta}-\overline{\eta})\right|<\epsilon\right\},$$

$$A_{2,(N,M)} := \left\{\sup_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)}\left|\frac{\hat{V}_{ll}^{1/2}}{V_{ll}^{1/2}}-1\right|<\epsilon\right\}, \quad \text{and}$$

$$A_{3,(N,M)} := \left\{\sup_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)}\left|\frac{1}{\sqrt{NM}}(\widehat{\Theta}_l'-\Theta_l')Z'\varepsilon\right|<\epsilon\right\}.$$

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Note that $V_{ll}^{1/2}$ is bounded away from zero and the probabilities of these three sets all tend to one asymptotically. Thus, for every $a \in \mathbb{R}$,

$$
\begin{aligned}
&\left| P\left(\frac{\sqrt{NM}(\beta_l - \beta_l)}{\widehat{V}_{ll}^{1/2}} \le a \right) - \Phi(a) \right| \\
&= \left| P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}} \left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l' \right) (\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \le a \right) - \Phi(a) \right| \\
&\le \left| P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}} \left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l' \right) (\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right.\right. \\
&\qquad \left.\left. \le \; a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)} \right) - \Phi(a) \right| \\
&\qquad + P\left(\cup_{i=1}^{3} A_{i,(N,M)}^{c} \right).
\end{aligned}
$$

Since $\hat{V}_{ll}^{1/2}$ does not depend on $\overline{\eta}$ and is bounded away from zero, there exists a positive constant D such that

$$
\begin{aligned}
&P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}} \left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l' \right) (\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \le a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)} \right) \\
&= P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{V_{ll}^{1/2}} - \frac{\sqrt{NM}}{V_{ll}^{1/2}} \left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l' \right) (\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right. \\
&\qquad \left. \le a\sqrt{\frac{\widehat{V}_{ll}}{V_{ll}}}, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)} \right) \\
&\le P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{V_{ll}^{1/2}} \le a(1+\epsilon) + 2D\epsilon \right).
\end{aligned}
$$

Note that the right hand side is independent of η. Therefore,

$$
\begin{aligned}
\sup_{\eta \in \mathbb{R}^p : \overline{\eta} \in \mathcal{B}_{\ell_0}(s)} &P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}} \left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l' \right) (\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right. \\
&\qquad \left. \le a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)} \right) \\
&\le P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{V_{ll}^{1/2}} \le a(1+\epsilon) + 2D\epsilon \right).
\end{aligned}
$$

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Corollary 1 then implies that for large N and M, one has

$$\sup_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)} P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}}\left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l'\right)(\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right.$$
$$\left. \leq a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)}\right)$$
$$\leq \Phi(a(1+\epsilon)+2D\epsilon)+\epsilon.$$

The continuity of Φ then implies that, for any $\delta > 0$, one can pick ϵ sufficiently small so it holds that

$$\sup_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)} P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}}\left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l'\right)(\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right.$$
$$\left. \leq a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)}\right)$$
$$\leq \Phi(a)+\delta+\epsilon.$$

Using a symmetric argument, one can also show that

$$\inf_{\eta\in\mathbb{R}^p:\overline{\eta}\in\mathcal{B}_{\ell_0}(s)} P\left(\frac{\widehat{\Theta}_l Z'\varepsilon/(NM)^{1/2}}{\widehat{V}_{ll}^{1/2}} - \frac{\sqrt{NM}}{\widehat{V}_{ll}^{1/2}}\left(\frac{\widehat{\Theta}_l' Z'Z}{NM} - e_l'\right)(\widehat{\boldsymbol{\eta}} - \overline{\boldsymbol{\eta}}) \right.$$
$$\left. \leq a, A_{1,(N,M)}, A_{2,(N,M)}, A_{3,(N,M)}\right)$$
$$\geq \Phi(a)+\delta+\epsilon.$$

Combining these intermediate results yields the claimed statement. □

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit: https://doi.org/10.1017/S0266466622000081

REFERENCES

Balazsi, L., L. Matyas, & T. Wansbeek (2017) Fixed effects models. In L. Matyas (ed.), *The Econometrics of Multi-Dimensional Panels*, chapter 1, pp. 1–34. Springer.

Baltagi, B.H., & G. Bresson (2017) Modelling housing using multi-dimensional panel data. In L. Matyas (ed.), *The Econometrics of Multi-Dimensional Panels*, chapter 12, pp. 349–376. Springer.

Baltagi, B.H., P.H. Egger, & K. Erhardt (2017) The estimation of gravity models in international trade. In L. Matyas (ed.), *The Econometrics of Multi-Dimensional Panels*, chapter 11, pp. 323–348. Springer.

Belloni, A., D. Chen, V. Chernozhukov, & C. Hansen (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80, 2369–2429.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, & K. Kato (2018) High-dimensional econometrics and regularized GMM, preprint, arXiv:1806.01888.

https://doi.org/10.1017/S0266466622000081 Published online by Cambridge University Press

Belloni, A., V. Chernozhukov, D. Chetverikov, & Y. Wei (2018) Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of Statistics* 46, 3643–3675.

Belloni, A., V. Chernozhukov, & C. Hansen (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, & D. Kozbur (2016) Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics* 34, 590–605.

Box, G.E.P. (1976) Science and statistics. *Journal of the American Statistical Association* 71, 791–799.

Caner, M., & X. Han (2014) Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business & Economic Statistics* 32, 359–374.

Caner, M., X. Han, & Y. Lee (2018) Adaptive elastic net GMM estimation with many invalid moment conditions: simultaneous model and moment selection. *Journal of Business & Economic Statistics* 36, 24–46.

Caner, M., & A.B. Kock (2018a) Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso. *Journal of Econometrics* 203, 143–168.

Caner, M., & A.B. Kock (2018b) High dimensional linear GMM, preprint, arXiv:1811.08779.

Galvao, A.F., & G.V. Montes-Rojas (2010) Penalized quantile regression for dynamic panel data. *Journal of Statistical Planning and Inference* 140, 3476–3497.

Head, K., & T. Mayer (2014) Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*, vol. 4, pp. 131–195. Elsevier.

Javanmard, A., & A. Montanari (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* 15, 2869–2909.

Kock, A.B. (2013) Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory* 29, 115–152.

Kock, A.B. (2016) Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models. *Journal of Econometrics* 195, 71–85.

Kock, A.B., & H. Tang (2019) Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory* 35, 295–359.

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

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

Leeb, H., & B.M. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.

Li, D., J. Qian, & L. Su (2016) Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association* 111, 1804–1819.

Lu, X., K. Miao, & L. Su (2021) Determination of different types of fixed effects in three-dimensional panels. *Econometric Reviews* 40, 867–898.

Lu, X., & L. Su (2016) Shrinkage estimation of dynamic panel data models with interactive fixed effects. *Journal of Econometrics* 190, 148–175.

Lu, X., & L. Su (2017) Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics* 8, 729–760.

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

Mátyás, L. (1997) Proper econometric specification of the gravity model. *The World Economy* 20, 363–368.

Mátyás, L. (2017) *The Econometrics of Multi-Dimensional Panels*. Springer.

Phillips, P.C.B. (2005) Automated discovery in econometrics. *Econometric Theory* 21, 3–20.

Qian, J., & L. Su (2016) Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *Journal of Econometrics* 191, 86–109.

Ramos, R. (2017) Modelling migration. In L. Matyas (ed.), *The Econometrics of Multi-Dimensional Panels*, chapter 13, pp. 377–395. Springer.
Rudelson, M., & R. Vershynin (2008) On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61, 1025–1045.
Su, L., & G. Ju (2018) Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* 206, 554–573.
Su, L., Z. Shi, & P.C.B. Phillips (2016) Identifying latent structures in panel data. *Econometrica* 84, 2215–2264.
Su, L., X. Wang, & S. Jin (2019) Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics* 37, 334–349.
Tinbergen, J.J. (1962) *Shaping the World Economy: Suggestions for an International Economic Policy*. Twth Century Fund.
van de Geer, S., P. Bühlmann, Y. Ritov, & R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* 42, 1166–1202.
Zhang, C.-H., & S.S. Zhang (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 217–242.