Robust Estimation and Inference in Panels with Interactive Fixed Effects*

Timothy B. Armstrong‡ Martin Weidner§ Andrei Zeleneev¶

May 9, 2025

Abstract

We consider estimation and inference for a regression coefficient in panels with interactive fixed effects (i.e., with a factor structure). We demonstrate that existing estimators and confidence intervals (CIs) can be heavily biased and size-distorted when some of the factors are weak. We propose estimators with improved rates of convergence and bias-aware CIs that remain valid uniformly, regardless of factor strength. Our approach applies the theory of minimax linear estimation to form a debiased estimate, using a nuclear norm bound on the error of an initial estimate of the interactive fixed effects. Our resulting bias-aware CIs take into account the remaining bias caused by weak factors. Monte Carlo experiments show substantial improvements over conventional methods when factors are weak, with minimal costs to estimation accuracy when factors are strong.

*We thank the participants of the numerous seminars and conferences for helpful comments and suggestions. We also thank Riccardo D'Adamo and Chen-Wei Hsiang for their excellent research assistance. Any remaining errors are our own. Armstrong gratefully acknowledges support by the National Science Foundation Grant SES-2049765. Weidner gratefully acknowledges support through the European Research Council grant ERC-2018-CoG-819086-PANEDA. Zeleneev gratefully acknowledges the generous funding from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Ref: EP/X02931X/1).

‡University of Southern California. Email: timothy.armstrong@usc.edu

§University of Oxford. Email: martin.weidner@economics.ox.ac.uk

¶University College London. Email: a.zeleneev@ucl.ac.uk

1 Introduction

In this paper, we consider a linear panel regression model of the form

$$Y_{it} = X_{it}\beta + \sum_{k=1}^{K} Z_{k,it}\delta_k + \Gamma_{it} + U_{it}, \tag{1}$$

where $Y_{it}, X_{it}, Z_{1,it}, \ldots, Z_{K,it} \in \mathbb{R}$ are the observed outcome variable and covariates for units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. The error components $\Gamma_{it} \in \mathbb{R}$ and $U_{it} \in \mathbb{R}$ are unobserved, and the regression coefficients $\beta, \delta_1, \ldots, \delta_K \in \mathbb{R}$ are unknown. The parameter of interest is $\beta \in \mathbb{R}$, the coefficient on X_{it}. We are interested in "large panels", where both N and T are relatively large.

The error component U_{it} is modelled as a mean-zero random shock that is uncorrelated with the regressors X_{it} and $Z_{k,it}$ and that is at most weakly autocorrelated across i and over t. By contrast, the error component Γ_{it} can be correlated with X_{it} and $Z_{k,it}$ and can also be strongly autocorrelated across i and over t. Of course, further restrictions on Γ_{it} are required to allow estimation and inference on β. For example, the additive fixed effect model imposes that $\Gamma_{it} = \alpha_i + \gamma_t$, where α_i accounts for any omitted variable that is constant over time, and γ_t for any omitted variable that is constant across units. Instead of this additive fixed effect model we consider the so-called interactive fixed effect model, where

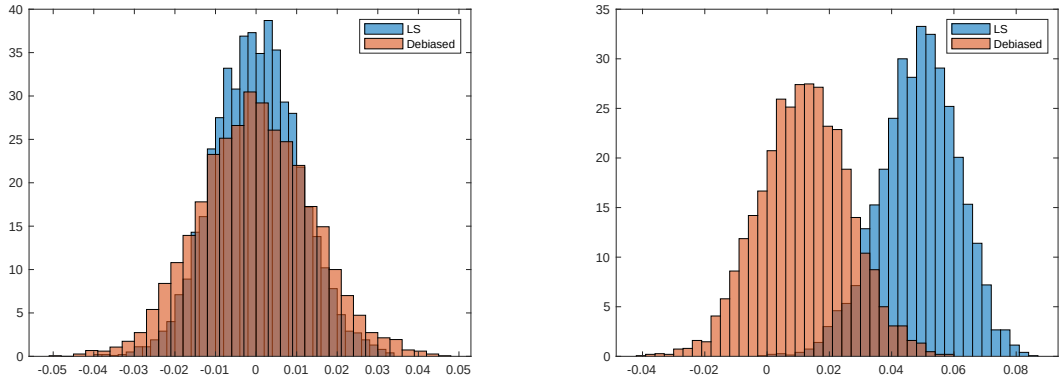
$$\Gamma_{it} = \sum_{r=1}^{R} \lambda_{ir}\, f_{tr}\ . \tag{2}$$

Here, the λ_{ir} and f_{tr} can either be interpreted as unknown parameters or as unobserved shocks. This model for Γ_{it} is also known as a factor model, with factor loadings λ_{ir} and factors f_{tr}. We will use the terms factor and interactive fixed effect interchangeably. The number of factors R is unknown, but is assumed to be small relative to N and T. The interactive fixed effect model is attractive because it introduces enough restrictions to allow estimation and inference on β while still incorporating or approximating a large class of data generating processes (DGPs) for Γ_{it}.

The existing econometrics literature on panel regressions with interactive fixed effects is quite large. Since the seminal work of Pesaran (2006) and Bai (2009), developing tools for estimation and inference on β in model (1)-(2) under large N and large T asymptotics has been a primary focus of this literature. Specifically, Pesaran (2006) introduces the common correlated effects (CCE) estimator, which uses cross-sectional averages of the observed variables as proxies for the unobserved factors. Bai (2009) derives the large N, T properties of the least-squares (LS) estimator that jointly minimizes the sum of squared residuals over the regression coefficients, factors, and factor loadings.[1]

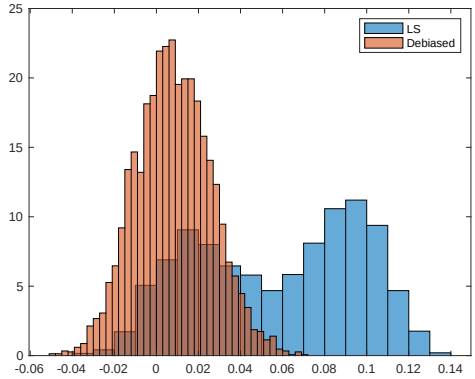
Bai (2009) shows that, under appropriate assumptions, the LS estimator for the regression

[1] This estimator was first introduced by Kiefer (1980).

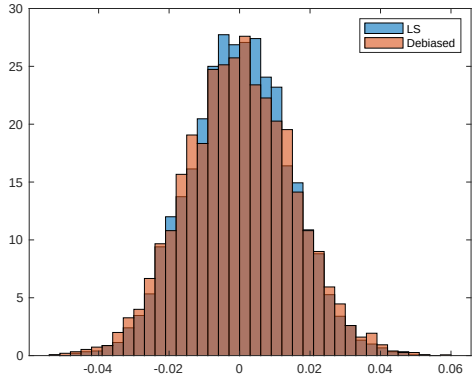


(a) No ID ($\kappa = 0.00$)

(b) Weak ID ($\kappa = 0.10$)



(c) Weak ID ($\kappa = 0.20$)



(d) Strong ID ($\kappa = 1.00$)

Figure 1: Finite sample distributions of the LS and the debiased estimators, $N = 100$, $T = 50$, $R = 1$

coefficients is $\sqrt{NT}$-consistent and asymptotically normally distributed as both N and T grow to infinity. One of the key assumptions imposed for this result is the so-called "strong factor assumption", which requires all the factor loadings λ_{ir} and factors f_{tr} to have sufficient variation across i and over t, respectively. If the strong factor assumption is violated, then the LS estimator for λ_{ir} and f_{tr} may be unable to pick up the true loadings and factors correctly, because the "weak factors"[2] in Γ_{it} cannot be distinguished from the noise in U_{it}. This can lead to substantial bias and misleading inference, due to omitted variables bias from Γ_{it} that is not picked up by the estimator.

To illustrate how this can lead to problems with conventional estimates and CIs for β, Figure 1 presents a subset of the results of our Monte Carlo study.[3] When the factors are nonexistent (panel a) or strongly identified (panel d), the distribution of the LS estimator (in blue) is centered at the true parameter value β (equal to 0 in this case). However, when the

[2] See, for example, Onatski (2010, 2012) for a discussion and formalization of the notion of weak factors.
[3] A detailed description of the numerical experiment is provided in Section 5.1.

factors are present but weak enough that they are difficult to estimate (panels b and c), the LS estimator is heavily biased and non-normally distributed. In our Monte Carlo study in Section 5, we show that this indeed leads to severe coverage distortion, with conventional CIs based on the LS estimator having almost zero coverage.

In this paper, we address this issue by developing new tools for estimation and inference on β in the model (1). We develop a debiased estimator along with a bound on the remaining bias, which we use to construct a bias-aware confidence interval. As illustrated in Figure 1, our debiased estimator (shown in red) substantially decreases the bias of the LS estimator when factors are weak, leading to a large improvement in overall estimation error. In addition, this improved performance under weak factors does not come at a substantial cost to performance when factors are strong or nonexistent: our debiased estimator performs similarly to the LS estimator in these cases. Importantly, our CI requires only an upper bound on the number of factors: we show that it is valid uniformly over a large class of DGPs that allows for weak, strong or nonexistent factors up to a specified upper bound on the number of factors. We derive rates of convergence that hold uniformly over this class of DGPs, and we show that our estimator achieves a faster uniform rate of convergence than existing approaches when weak factors are allowed. In the case where N and T grow at the same rate, our estimator achieves the parametric $\sqrt{NT}$ rate.

Our debiasing approach uses a preliminary estimate $\hat{\Gamma}_{\text{pre}}$ of the individual effect matrix Γ along with a bound $\hat{C}$ on the nuclear norm $\|\Gamma - \hat{\Gamma}_{\text{pre}}\|_*$ of its estimation error. Letting $\tilde{\Gamma} := \Gamma - \hat{\Gamma}_{\text{pre}}$, we then consider the augmented outcomes

$$\tilde{Y}_{it} := Y_{it} - \hat{\Gamma}_{\text{pre},it} = X_{it}\beta + \sum_{k=1}^{K} Z_{k,it}\delta_k + \tilde{\Gamma}_{it} + U_{it}.$$

Treating $\tilde{\Gamma}_{it}$ as nuisance parameters satisfying a convex constraint $\|\tilde{\Gamma}\|_* \leq \hat{C}$, we derive linear weights A_{it} such that the estimator $\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\tilde{Y}_{it}$ for β optimally uses this constraint, using the theory of minimax linear estimators (see Ibragimov and Khas'minskii, 1985; Donoho, 1994; Armstrong and Kolesár, 2018). In particular, the resulting weights A_{it} control the remaining omitted variables bias $\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\tilde{\Gamma}_{it}$ due to possible weak factors in $\tilde{\Gamma} = \Gamma - \hat{\Gamma}_{\text{pre}}$ not picked up by the initial estimate $\hat{\Gamma}_{\text{pre}}$.

A key step in deriving our CI is the construction of the preliminary estimator $\hat{\Gamma}_{\text{pre}}$ and bound $\hat{C}$ on the nuclear norm of its estimation error. Our CI is bias-aware: it uses the bound $\hat{C}$ to explicitly take into account any remaining bias in the debiased estimator. Our bound is feasible once an upper bound on the number of factors is specified. In our Monte Carlo study, we find that, while our CIs are often conservative, they are about as wide as an "oracle" CI that uses an infeasible critical value to correct the coverage of a CI based on the standard LS estimator.

While our results allow for arbitrary sequences of weak factors, our conditions on other aspects of the model are similar to Bai (2009) and Moon and Weidner (2015). An important

condition is that the covariate of interest X_{it} must not itself be entirely explained by a low dimensional factor model. For example, in a panel where X_{it} is the minimum hourly wage in state i and year t, we would require that states change their minimum wage laws in different years, and that this is done sufficiently often to generate variation in X_{it} that cannot be explained by a small number of factors f_t. This rules out settings where X_{it} is an indicator variable for a policy that affects a subset of the units and occurs only during a single time period: in this case, $X_{it} = \lambda_i \cdot f_t$ where λ_i is an indicator variable for unit i undergoing the policy change and f_t is an indicator variable for periods after the policy change. See Section 4 for formal conditions and further discussion.

A special case of the factor model is the grouped unobserved heterogeneity model considered by Bonhomme and Manresa (2015). In this model, $\Gamma_{it} = \alpha_{g(i),t}$, where $g(\cdot)$ is an unknown function mapping individuals i to a group index $g(i) \in \{1, \ldots, R\}$. This takes the form of the factor model (2) with $\lambda_{ir} = 1$ if $g(i) = r$ and 0 otherwise, and with $f_{tr} = \alpha_{r,t}$. The strong factor assumption corresponds to the strong group separation assumption imposed in this literature (e.g., Assumption 2(b) in Bonhomme and Manresa, 2015) which imposes that the group means $\alpha_{r,\cdot} = (\alpha_{r,1}, \ldots, \alpha_{r,T})'$ are sufficiently far away for different groups r. Our results apply in this setting and allow for this assumption to be relaxed. An interesting question for future research is whether it is possible to modify our approach to take advantage of the additional structure in the grouped unobserved heterogeneity model.

Related literature

The papers by Pesaran (2006) and Bai (2009) mentioned previously have motivated a large follow up literature on large N and T analysis of panel models with interactive effects. Bai and Wang (2016) provides a review with further references. Another literature has proposed alternative estimation methods along with asymptotic analysis in the regime with T fixed and N increasing. This includes the quasi-difference approach of Holtz-Eakin, Newey and Rosen (1988) and generalized method of moments approaches of Ahn, Lee and Schmidt (2001, 2013). More recent papers analyzing the fixed T large N regime include Robertson and Sarafidis (2015), Juodis and Sarafidis (2018), Westerlund, Petrova and Norkute (2019), Higgins (2021), Juodis and Sarafidis (2022). None of these papers provide inference methods that remain valid when factors are weak or rank-deficient (e.g. $f = 0$). Chamberlain and Moreira (2009) derive estimators that satisfy a Bayes-minimax property over a certain class of priors in a finite sample setting that includes a version of the model (2). This Bayes-minimax property does not, however, translate to a guarantee on coverage or estimation error under weak factors.

A special case of the violation of the strong factor assumption is when some factor are equal to zero, while all other factors are strong; the inference results of Bai (2009) are usually robust towards this specific violation of the strong factor assumption (Moon and Weidner, 2015). This robustness, however, does not carry over to more general weak factors in the DGP of Γ_{it}, as illustrated by Figure 1.

The problem of weak factors is related to the problem of omitted variable bias of LASSO

estimators in high dimensional regression that is the focus of debiased LASSO estimators (see Belloni, Chernozhukov and Hansen, 2014; Javanmard and Montanari, 2014; van de Geer, Bühlmann, Ritov and Dezeure, 2014; Zhang and Zhang, 2014). Just as LASSO estimators omit variables with coefficients that are large enough to cause omitted variables bias but too small to distinguish from zero, weak factors in Γ can be difficult to estimate, leading to omitted variables bias in conventional estimates of β. Our approach to using minimax linear estimation to debias an initial estimate mirrors the approach of Javanmard and Montanari (2014) to debiasing the LASSO. We discuss this connection further in Section 4.4. Hirshberg and Wager (2020) provide a general discussion and further references for minimax linear debiasing; we refer to this general approach as augmented linear estimation following their terminology. Minimax linear estimation itself goes back at least to Ibragimov and Khas'minskii (1985), with further results on this approach and its optimality properties in Donoho (1994), Armstrong and Kolesár (2018) and Yata (2021), among others. The particular form of the minimax estimator used for debiasing in our setup follows from a formula given in Armstrong, Kolesár and Kwon (2020).

Requiring Γ_{it} to have the factor structure (2) is equivalent to requiring the matrix of unobserved effects Γ to have rank at most R, i.e., having $\text{rank}(\Gamma) \leq R$. Bounding the nuclear norm of $\tilde{\Gamma}$ or Γ instead can also be seen as a convex relaxation of this requirement. Similar convexifications of the rank constraint have been widely used in the matrix completion literature (e.g., Recht, Fazel and Parrilo 2010 and Hastie, Tibshirani and Wainwright 2015 for recent surveys), and for reduced rank regression estimation (e.g., Rohde and Tsybakov 2011). In the econometrics literature, the numerous applications of this idea include, for example, estimation of pure factor models (Bai and Ng, 2017), estimation of panel regression models with homogeneous (Moon and Weidner, 2018; Beyhum and Gautier, 2019) and heterogeneous coefficients (Chernozhukov, Hansen, Liao and Zhu, 2019), estimation of treatment effects (Athey, Bayati, Doudchenko, Imbens and Khosravi, 2021; Fernández-Val, Freeman and Weidner, 2021), and many others.[4] However, none of these papers obtain asymptotically valid CIs or improved rates of convergence under weak factors.

In recent work, Chetverikov and Manresa (2022) propose an estimator that, like ours, achieves a faster rate of convergence than conventional approaches under weak factors.[5] While Chetverikov and Manresa (2022) allow for weak factors in some of their estimation results, they assume strong factors when constructing CIs. The estimation approach in Chetverikov and Manresa (2022) also differs from our approach by using modelling assumptions that place a factor structure on the covariate matrix X.

[4]For example, recent economic applications of nuclear norm and related penalization methods also include latent community detection (Alidaee, Auerbach and Leung, 2020; Ma, Su and Zhang, 2022), quantile regression (Belloni, Chen, Madrid Padilla and Wang, 2023; Wang, Su and Zhang, 2022; Feng, 2023), and estimation of panel threshold models and high-dimensional VARs (Miao, Li and Su, 2020 and Miao, Phillips and Su, 2023).

[5]The main focus of Chetverikov and Manresa (2022) is the grouped effects model of Bonhomme and Manresa (2015), which is a special case of the interactive fixed effects setting we consider here. However, the authors extend their results to the general interactive fixed effects setting.

Our focus is on allowing for weak factors without imposing additional assumptions on the error term U, such as homoskedasticity or full independence from the individual effects Γ and regressor X. Such additional structure allows for further identifying information by making it easier to distinguish between the error term U and the individual effects Γ, leading to a fundamentally different analysis. Zhu (2019) derives asymptotic upper and lower bounds for estimators and CIs in a setting with possible weak factors under homoskedastic and fully independent errors. The estimators and CIs constructed by Zhu (2019) take advantage of the additional structure of Zhu's setting, making them inapplicable in ours. However, the *lower* bounds derived by Zhu (2019) are immediately relevant: they show that no CI can be asymptotically valid under weak factors while mimicking the performance of the CI of Bai (2009) when factors are strong.

As discussed above, our assumptions rule out the case where X_{it} is an indicator variable for a policy that affects a subset of units starting in the same time period. Recent papers that analyze such settings include Ferman and Pinto (2021) and Arkhangelsky, Athey, Hirshberg, Imbens and Wager (2021). The fact that X_{it} is collinear with the confounding factor model in this setting presents a fundamental identification issue that requires placing additional conditions on the model. In contrast to this literature, our goal is to leverage variation in X_{it} that cannot be explained by a low dimensional factor model in settings where such variation exists.

Beyhum and Gautier (2022), Fan and Liao (2022), and Bai and Ng (2023) consider estimation and inference in various settings under a regime in which a lower bound on the strength of the factors can decrease with N and T, but is large enough that factors can be consistently estimated. This is analogous to the "semi-strong" regime in weak instrument and related settings; see Andrews and Cheng (2012). While the semi-strong regime requires careful theoretical analysis, the fact that factors can be consistently estimated leads to asymptotically unbiased and normal estimators for the main effect β. Our results apply to semi-strong and strong regimes as well, while also allowing for weak factor regimes in which factors cannot be consistently estimated.

Finally, Cox (2024) develops tools for inference in low-dimensional factor models with weak identification. In Cox (2024), the primary objects of interests are the covariance of the factors and the loadings. The baseline model in Cox (2024) does not include observed covariates, whereas we focus on estimation and inference on β, the coefficient on X_{it}, exclusively.[6]

The rest of this paper is organized as follows. Section 2 introduces the framework and describes construction of the debiased estimator and bias-aware CI. Section 3 provides implementation details. Section 4 provides formal statistical guarantees. Section 5 considers numerical and empirical illustrations. A supplementary appendix contains all proofs and additional results for the numerical and empirical illustrations.

[6] Cox (2024) mentions that observed covariates could, in principle, be incorporated in his framework as long as they are uncorrelated with the unobserved effects, which is a primary worry in the panel literature.

2 Construction of robust estimates and confidence intervals

2.1 Setup

We consider a panel setting in which we observe a scalar outcome Y_{it}, a scalar covariate X_{it} of interest and additional control covariates $\{Z_{k,it}\}_{k=1}^K$ for $i = 1, \ldots, N$, $t = 1, \ldots, T$, which follow the regression model (1). The error term U_{it} is assumed to be mean zero conditional on X, $\{Z_{k,it}\}_{k=1}^K$ and Γ,[7] but we allow for heteroskedasticity, which may depend on X_{it} and Γ_{it}, as well as some weak dependence. We write the model in matrix notation as

$$Y = X\beta + Z \cdot \delta + \Gamma + U\,, \qquad \mathbb{E}[U|X, Z, \Gamma] = 0\,, \tag{3}$$

where Z denotes the three dimensional array $\{Z_{k,it}\}$ and we define $Z \cdot \delta = \sum_{k=1}^K Z_k \delta_k$ where Z_k denotes the matrix with i,t-th element $Z_{k,it}$. We use λ to denote the $N \times R$ matrix of loadings λ_{ir} and f to denote the $T \times R$ matrix of factors f_{tr}, so that (2) can be written in matrix form as $\Gamma = \lambda f'$.

We are interested in the coefficient β of X_{it}, which can be interpreted as the effect of a treatment variable X_{it} in a constant treatment effects model (we discuss extensions to heterogeneous treatment effects in Remark 2.2). For concreteness, we use panel notation, and we refer to i and t as individuals and time periods respectively. However, we allow for other settings such as network data in which i and t both index individuals in a network. While we will assume a low rank structure on Γ, we allow for arbitrary dependence between the covariate X_{it} and the individual effect Γ_{it}.

A key ingredient in our approach is an initial estimate $\hat{\Gamma}$ of Γ and a bound on its estimator in the nuclear norm, which holds with probability approaching one:

$$\left\|\widetilde{\Gamma}\right\|_* \leq \hat{C}\,, \qquad \text{where} \qquad \widetilde{\Gamma} := \Gamma - \hat{\Gamma}\,. \tag{4}$$

Here, $\|\cdot\|_*$ denotes the nuclear norm of the argument matrix, and $\hat{C} \geq 0$ is a known or estimated constant. We describe our estimate $\hat{\Gamma}$ and bound $\hat{C}$ in Section 3, and we state a formal result giving conditions under which the bound holds with probability approaching one in Section 4. This bound depends on an upper bound for the number of factors R, which must be specified a priori. Importantly, our approach does not require specifying the exact number of factors: some of the factors may be zero, in addition to the possibility of being "weak" in the sense of being close to zero. We emphasize that obtaining an computable upper bound $\hat{C}$ that enables construction of our CI is itself one of the main technical contributions of this paper.[8]

[7] We note that this requires strict exogeneity and in particular rules out using lagged outcomes as covariates. We leave extensions to models with lagged outcomes as a topic for future research.

[8] As we discuss further in Section 4, a tighter CI can be constructed by bounding the difference between the estimate $\hat{\Gamma}$ and $\Gamma + P_\lambda U$, where $P_\lambda = \lambda(\lambda'\lambda)^+\lambda$, with M^+ denoting the Moore–Penrose inverse of a matrix M. The implementation in Section 3 is for the tighter CI that uses these arguments. For ease of exposition, however, we focus on using the bound (4) directly in the remainder of this section.

Remark 2.1. Although the main focus of this paper is on models with the linear factor structure (2), the methodology presented in this section applies to general interactive fixed effects models as long as it is possible to construct a preliminary estimator $\hat{\Gamma}$ and a bound $\hat{C}$ satisfying (4). For example, we conjecture that our method can also be extended to nonlinear factor models with $\Gamma_{it} = g(\lambda_i, f_t)$, where $g(\cdot, \cdot)$ is some unknown function (e.g., Zeleneev, 2019; Freeman and Weidner, 2023). As noted, for example, in Fernández-Val, Freeman and Weidner (2021), such Γ can be approximated by a low-rank matrix with a (slowly) growing rank R. Hence, we expect that our method can be applied in this setting with $\hat{\Gamma}$ constructed using a growing R and $\hat{C}$ adjusted for the low-rank approximation error (if needed), in the same spirit as sieve approximations are used in nonparametric estimation.

2.2 Augmented linear estimators and CIs

We first define a class of estimators and CIs, indexed by an $N \times T$ matrix A. We then provide a choice of the matrix A, based on finite sample optimality in an idealized setting. Our class of estimators is given in the following definition.

Definition 2.1. *Let $A = A(X, Z)$ be an $N \times T$ matrix of weights $A_{it} \in \mathbb{R}$ that can depend on the matrix X and array Z. Let $\hat{\Gamma}$ be an initial estimate of Γ, and let $\widetilde{Y} = Y - \hat{\Gamma}$. The* augmented linear estimator *with weight matrix A and initial estimate $\hat{\Gamma}$ is given by*

$$\hat{\beta}_A := \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \widetilde{Y}_{it} = \langle A, \widetilde{Y} \rangle_F. \tag{5}$$

Here, $\langle \cdot, \cdot \rangle_F$ denotes the entry-wise inner product between the argument matrices.

The estimator $\hat{\beta}_A = \langle A, \widetilde{Y} \rangle_F$ applies a linear estimator after an initial estimation step in which the initial estimate $\hat{\Gamma}$ is subtracted from the outcome Y. This mirrors applications of this idea in other settings going back to Javanmard and Montanari (2014); see Hirshberg and Wager (2020) for references (the term "augmented linear estimation" is used in the latter paper).

To analyze this class of estimators, note that subtracting the initial estimate from both sides of the equation (3) gives

$$\widetilde{Y} = X\beta + Z \cdot \delta + \widetilde{\Gamma} + U \tag{6}$$

(recall that $\widetilde{Y} = Y - \hat{\Gamma}$ and $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}$). Our choice of the matrix A will be motivated by a heuristic in which we consider the model (6) with $\widetilde{Y}$ as the observed outcome and $\widetilde{\Gamma}$ a nuisance parameter such that U is mean zero conditional on X, Z and $\tilde{\Gamma}$, with the bound (4) interpreted as a deterministic bound that holds with $\hat{C}$ nonrandom. This heuristic is not literally true, since $\widetilde{\Gamma}$ depends on U through the estimation error in the initial estimate $\hat{\Gamma}$. Nonetheless, the CIs and estimators we obtain will be asymptotically valid and consistent respectively, under conditions that we give in Section 4.

Following this heuristic, we consider the decomposition

$$\hat{\beta}_A - \beta = \text{bias}_{\beta,\delta,\widetilde{\Gamma}}(\hat{\beta}_A) + \langle A, U \rangle_F \tag{7}$$

where

$$\text{bias}_{\beta,\delta,\widetilde{\Gamma}}(\hat{\beta}_A) := (\langle A, X \rangle_F - 1)\,\beta + \langle A, Z \cdot \delta \rangle_F + \langle A, \widetilde{\Gamma} \rangle_F. \tag{8}$$

Under the heuristic where $\tilde{\Gamma}$ is nuisance parameter in the model (6), $\text{bias}_{\beta,\delta,\widetilde{\Gamma}}$ gives the bias of the estimator $\hat{\beta}_A$ conditional on X, Z and $\tilde{\Gamma}$. In reality, $\text{bias}_{\beta,\delta,\widetilde{\Gamma}}$ does not literally give the bias or conditional bias of $\hat{\beta}_A$, since conditioning on $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}$ means conditioning on an information set that depends on Y through the preliminary estimate $\hat{\Gamma}$. We nonetheless refer to $\text{bias}_{\beta,\delta,\Gamma}(\hat{\beta}_A)$ as a bias term, following our heuristic.

Let $\widehat{\text{se}}$ be an estimate of the standard deviation of $\langle A, U \rangle_F = \sum_{i=1}^N \sum_{t=1}^T A_{it} U_{it}$. For example, to allow for arbitrary heteroskedasticity in U_{it} while imposing independence across i and t, we can use $\widehat{\text{se}} = \sqrt{\sum_{i=1}^N \sum_{t=1}^T A_{it}^2 \hat{U}_{it}^2}$ where $\hat{U}_{it}$ denotes residuals from an initial regression. If $\text{bias}_{\beta,\delta,\Gamma}(\hat{\beta}_A)$ were zero, then we could form a CI by adding and subtracting a normal critical value times $\widehat{\text{se}}$. To take into account the possibility that $\text{bias}_{\beta,\delta,\Gamma}(\hat{\beta}_A)$ will in general be nonnegligible in our setting, we use the bound (4) to obtain an upper bound on the bias term. In particular, when (4) holds, we have $\left|\text{bias}_{\beta,\delta,\widetilde{\Gamma}}(\hat{\beta}_A)\right| \leq \overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$, where for general $C \geq 0$ we define

$$\begin{aligned}
\overline{\text{bias}}_C(\hat{\beta}_A) &:= \sup_{\beta,\delta,\widetilde{\Gamma}:\|\widetilde{\Gamma}\|_* \leq C} \text{bias}_{\beta,\delta,\widetilde{\Gamma}}(\hat{\beta}_A) \\
&= \begin{cases} \sup_{\widetilde{\Gamma}:\|\widetilde{\Gamma}\|_* \leq C} \langle A, \widetilde{\Gamma} \rangle_F & \text{if } \langle A, X \rangle_F = 1, \text{ and } \langle A, Z_k \rangle_F = 0, \text{ for } k = 1, \ldots K, \\ \infty & \text{otherwise} \end{cases} \\
&= \begin{cases} C s_1(A) & \text{if } \langle A, X \rangle_F = 1, \text{ and } \langle A, Z_k \rangle_F = 0, \text{ for } k = 1, \ldots K, \\ \infty & \text{otherwise.} \end{cases}
\end{aligned} \tag{9}$$

Here, for the second equality we used that the supremum over β and δ is unbounded unless $\langle A, X \rangle_F = 1$ and $\langle A, Z_k \rangle_F = 0$, and for the final step we used that the nuclear norm $\|\cdot\|_*$ is dual to the spectral norm, which we denote by $s_1(\cdot)$ since it is equal to the largest singular value of the argument matrix. We refer to $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$ as the worst-case bias of the estimator $\hat{\beta}_A$ (again, this terminology reflects the heuristic in which $\tilde{\Gamma}$ is treated as a nuisance parameter in (6) rather than estimation error from the initial estimate $\hat{\Gamma}$).

Note that, whereas $\text{bias}_{\beta,\delta,\widetilde{\Gamma}}(\hat{\beta}_A)$ depends on the unknown matrix of individual effects Γ through the matrix $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}$, $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$ is feasible to compute once a bound $\hat{C}$ is given.

Taking into account the possible bias leads to a *bias-aware* CI:

$$\left\{\hat{\beta}_A \pm \left[\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) + z_{1-\alpha/2}\widehat{\text{se}}\right]\right\}. \tag{10}$$

To motivate this CI, note that the probability that the lower endpoint is greater than β is

$$P\left(\hat{\beta}_A - \overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) - z_{1-\alpha/2}\widehat{\text{se}} > \beta\right) = P\left(\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}U_{it} + \text{bias}_{\beta,\delta,\tilde{\Gamma}}(\hat{\beta}_A) > \overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) + z_{1-\alpha/2}\widehat{\text{se}}\right)$$
$$\leq P\left(\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}U_{it} > z_{1-\alpha/2}\widehat{\text{se}}\right) \approx \alpha/2,$$

where the last step assumes that $\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}U_{it}$ is approximately normally distributed with zero mean and standard deviation close to $\widehat{\text{se}}$. We provide formal justifications for this later. By a similar argument, the probability that the upper endpoint is less than β can be bounded by $\alpha/2$, and these calculations together imply that the coverage of our CI is approximately at least $1-\alpha$.

Remark 2.2. In principle, our approach can be extended to a heterogeneous treatment effect model where the constant coefficient β is replaced by an individual specific coefficient β_{it} that is allowed to vary with i and t. In particular, if a bound on the nuclear norm of the matrix of coefficients β_{it} or on the error of preliminary estimates of these coefficients is available in addition to such a bound for Γ, we can use minimax linear debiasing to estimate a linear functional of the individual specific effects β_{it}. For example, the linear functional $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\beta_{it}$ gives the average treatment effect of a one-unit change in X_{it} over the NT units in a setting where β_{it} is interpreted as the causal effect of a change in the variable X_{it}. Deriving a computable bound on the nuclear norm error of an initial estimate of the coefficients β_{it} in this case is nontrivial, however, and we leave this question for future research.

2.3 Choice of weights $A = (A_{it})$

As described in the last subsection, one can construct valid confidence intervals for β of the form (10) for any choice of weight matrix A, subject to weak regularity conditions. To get a simple baseline procedure, we compute weights that are optimal in an idealized setting where $U_{it} \overset{iid}{\sim} N(0,\sigma^2)$ independently of X, Z and $\tilde{\Gamma}$ (again, this involves invoking the heuristic of treating $\tilde{\Gamma}$ as a nuisance parameter in (6) rather than estimation error from a preliminary estimate). In this idealized setting, $\hat{\beta}_A$ is then normally distributed with variance $\sigma^2\sum_{i=1}^{N}\sum_{t=1}^{T}A_{it}^2 = \sigma^2\|A\|_F^2$ (where $\|\cdot\|_F$ denotes the Frobenius norm), and with bias ranging from $-\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$ to $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$. Thus, if we choose worst-case MSE under i.i.d. normal errors as our criterion function for the weights, then the optimal weights are obtained by minimizing $\left(\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)\right)^2 + \sigma^2\|A\|_F^2$. By substituting the formula for $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$ from (9), we obtain the following baseline choice of weights, indexed by a tuning parameter b that corresponds to $\hat{C}/\sigma$.

Definition 2.2. *For $b > 0$, define the "optimal" $N \times T$ weight matrix by*

$$A_b^* := \underset{A \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \; b^2 s_1(A)^2 + \|A\|_F^2 \qquad s.t. \quad \langle A, X \rangle_F = 1 \text{ and } \langle A, Z_k \cdot \delta \rangle_F = 0,$$

Here, the constraint $\langle A, Z_k \cdot \delta \rangle_F = 0$ is imposed for all $k \in \{1, \dots, K\}$.

Heuristically, we expect that a good choice of b will correspond to $\hat{C}/\sigma$ such that the bound $\hat{C}$ on the nuclear norm holds with high probability. Conveniently, our nuclear norm bound in the exact factor model in Section 3 scales with the standard deviation σ in the homoskedastic case, which gives us a simple and feasible choice of the tuning parameter b.

We emphasize again that while the definition of A_b^* is motivated by the idealized setting $U_{it} \overset{iid}{\sim} N(0, \sigma^2)$, we do *not* assume that the error terms U_{it} satisfy this strong assumption. Choosing $A = A_b^*$ to construct the estimator $\hat{\beta}_A$ and the confidence intervals (10) under more general error distributions just means that the resulting estimates and confidence intervals will not be optimal (in finite samples), but we will nevertheless show them to be consistent and valid, respectively.

Remark 2.3. While we have used MSE to motivate our baseline choice of weights A_b^*, one could use other criteria corresponding to different weights on bias and variance. For example, optimizing CI length when $\hat{C}/\sigma = b$ would give the criterion $bs_1(A) + z_{1-\alpha}\|A\|_F$. If β gives the net welfare gain of an all-or-nothing policy change, then one can target minimax welfare regret as in Ishihara and Kitagawa (2021) and Yata (2021). In our Monte Carlo simulations however, we find that the exact choice of criterion has little effect on performance.

2.4 Practical implementation

The definition of A_b^* is a convex optimization problem that can easily be solved numerically for any given input X, Z, b. Using results from Armstrong, Kolesár and Kwon (2020), it follows that A_b^* can also be computed using the residuals of a nuclear norm regularized regression of X on $Z_1, \dots, Z_K$ and a matrix of individual effects. When there are no additional covariates Z, this nuclear norm regularized regression simplifies further: it can be solved by computing the singular value decomposition of X, and then performing soft thresholding on the singular values. The resulting weights A_b^* obtained from the residuals of this regression replace the largest singular values of X with a constant. We provide details in Appendix B.

In addition to giving alternative methods for computing the weights A_b^*, these results provide some intuition for these weights. The residuals from this nuclear norm regularized regression of X on $Z_1, \dots, Z_K$ and the individual effects "partial out" potential correlation of X with the estimation error $\widetilde{\Gamma}$, similar to the estimator of Robinson (1988) in the partially linear model. When there are no additional covariates Z, this amounts to removing the largest singular values of X and replacing them with a constant.

To summarize, we can compute an estimator $\hat{\beta}_A$ using Definition 2.1 using any matrix of weights A. We can also compute a CI $\left\{\hat{\beta}_A \pm \left[\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) + z_{1-\alpha/2}\widehat{\text{se}}\right]\right\}$ as in (10), once we

have a standard error $\widehat{\text{se}}$ and an upper bound $\hat{C}$ for the nuclear norm of the error in the initial estimate of Γ. Definition 2.2 gives us a heuristic for computing a reasonable choice of the matrix A, once we have an initial choice of b for the ratio $\hat{C}/\sigma$ of the nuclear norm bound to variance of U_{it}.

Thus, to apply our approach, we need an initial choice b to compute the weights A_b^* using Definition 2.2. We also need a robust upper bound $\hat{C}$ such that the bound (4) holds with high probability. Finally, we need a robust standard error $\widehat{\text{se}}$. Our CI then takes the form in (10) with $A = A_b^*$ and the given bound $\hat{C}$ and standard error $\widehat{\text{se}}$. In Section 3, we give details of these choices, as well as how to compute the initial estimate of Γ.

3 Implementation

In this section, we describe the implementation of our approach. Our approach relies on bounds for the nuclear norm of the initial estimate of Γ, derived formally in Section 4. As explained in Section 4, a tighter CI can be derived using a more nuanced argument that bounds the difference between $\hat{\Gamma}$ and $\Gamma + P_\lambda U$, where $P_\lambda = \lambda(\lambda'\lambda)^+\lambda$, and M^+ denotes the Moore–Penrose inverse of a matrix M. In particular, we show that the bound $\hat{C} \approx 2Rs_1(U)$ can be used. The bound R on the number of factors must be specified by the researcher, similar to other methods in this literature (e.g. Bai, 2009). Furthermore, the weights A_b^* are designed to be optimal when $U_{it} \overset{iid}{\sim} N(0, \sigma^2)$, which leads to the approximation $s_1(U)/\sigma \approx \sqrt{N} + \sqrt{T}$ (Geman, 1980). We therefore use $b = b^* := 2R(\sqrt{N} + \sqrt{T})$ as our default choice to calibrate $\hat{C}/\sigma$ when computing the weights in Definition 2.2. We then use an upper bound $\hat{C}$ that is valid under heteroskedasticity when computing $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_{A_{b^*}^*})$ in the construction of the CI.

Our initial estimate is formed in two steps. First, we form the least squares estimate $\hat{\Gamma}_{\text{LS}}$. We then apply our debiasing approach to get an estimate of the coefficients β and δ and form an $\hat{\Gamma}_{\text{pre}}$ by applying least squares to estimate Γ, with β and δ fixed at this initial debiased estimate. This estimate $\hat{\Gamma}_{\text{pre}}$ is then used as the initial estimate in our procedure. Thus, our procedure involves applying our debiasing approach twice. This appears to be necessary to get an initial estimate $\hat{\Gamma}_{\text{pre}}$ the best possible nuclear norm bounds on the estimation error.

Below we provide the details of our implementation algorithm.[9]

Algorithm 3.1 (Implementation for the factor model)**.**

Input *Data Y, X, Z and R pre-specified by the user, along with tuning parameter ε.*

Output *Estimator and CI for β.*

[9]Implementation of this algorithm in R is also available at `https://github.com/chenweihsiang/PanelIFE/tree/main`. We thank Chen-Wei Hsiang for his excellent assistance in preparing this R package.

1. *Compute the least squares (LS) estimator*

$$\left(\hat{\beta}_{\mathrm{LS}}, \hat{\delta}_{\mathrm{LS}}, \hat{\Gamma}_{\mathrm{LS}}\right) = \underset{\{\beta\in\mathbb{R}, \delta\in\mathbb{R}^K, G\in\mathbb{R}^{N\times T}\,:\,\mathrm{rank}(G)\leq R\}}{\mathrm{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - X_{it}\beta - Z_{it}'\delta - G_{it}\right)^2. \tag{11}$$

2. *Compute* $\widetilde{Y}_{\mathrm{pre}} = Y - \hat{\Gamma}_{\mathrm{LS}}$ *and let* $b^* = 2R(\sqrt{N}+\sqrt{T})$. *Let*

$$\hat{\beta}_{\mathrm{pre}} = \langle A^*_{b^*}, \widetilde{Y}_{\mathrm{pre}}\rangle_F.$$

Construct $\hat{\delta}_{\mathrm{pre}}$ *with the* j*-th element* $\hat{\delta}_{\mathrm{pre},j}$ *computed in the same way as* $\hat{\beta}_{\mathrm{pre}}$, *but with* X *and* Z_j *switched.*

3. *Compute* $\hat{\Gamma}_{\mathrm{pre}}$ *as*

$$\hat{\Gamma}_{\mathrm{pre}} = \underset{\{G\in\mathbb{R}^{N\times T}\,:\,\mathrm{rank}(G)\leq R\}}{\mathrm{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - X_{it}\hat{\beta}_{\mathrm{pre}} - Z_{it}'\hat{\delta}_{\mathrm{pre}} - G_{it}\right)^2.$$

The solution $\hat{\Gamma}_{\mathrm{pre}}$ *to this least squares problem is simply given by the leading* R *principal components of the residuals* $Y_{it} - X_{it}\hat{\beta}_{\mathrm{pre}} - Z_{it}'\hat{\delta}_{\mathrm{pre}}$. *Compute* $\widetilde{Y} = Y - \hat{\Gamma}_{\mathrm{pre}}$.

4. *Compute the final estimate*

$$\hat{\beta} = \hat{\beta}_{A^*_{b^*}} = \langle A^*_{b^*}, \widetilde{Y}\rangle_F.$$

To compute the CI, let $\hat{C} = (2+\varepsilon)Rs_1(\hat{U}_{\mathrm{pre}})$ *and* $\widehat{\mathrm{se}}^2 = \sum_{i=1}^{N}\sum_{t=1}^{T} A^{*2}_{b^*,it}\hat{U}^2_{\mathrm{pre},it}$, *where*

$$\hat{U}_{\mathrm{pre}} = Y - X\hat{\beta}_{\mathrm{pre}} - Z\cdot\hat{\delta}_{\mathrm{pre}} - \hat{\Gamma}_{\mathrm{pre}}.$$

Compute the CI

$$\hat{\beta}_{A^*_{b^*}} \pm \left[\overline{\mathrm{bias}}_{\hat{C}}(\hat{\beta}_{A^*_{b^*}}) + z_{1-\alpha/2}\widehat{\mathrm{se}}\right] \tag{12}$$

where $\overline{\mathrm{bias}}_{\hat{C}}(\hat{\beta}_{A^*_{b^*}}) = \hat{C}s_1(A^*_{b^*})$.

Remark 3.1 (Behavior of estimator under strong factors)**.** In Section 4.3, we show that, in the absence of weak factors, $\overline{\mathrm{bias}}_{\hat{C}}(\hat{\beta}_{A^*_{b^*}})$ becomes negligible relative to $\widehat{\mathrm{se}}$ in the construction of the CI in (12). Thus, the CI

$$\hat{\beta}_{A^*_{b^*}} \pm z_{1-\alpha/2}\widehat{\mathrm{se}}, \tag{13}$$

which uses our bias-corrected estimator but ignores bias when computing the critical value, will have correct asymptotic coverage in the strong factor case. In our Monte Carlos provided

in Section C.3, we find that this CI is (i) comparable to alternative non-robust CIs in terms of the length and (ii) despite its non-robustness, substantially less size-distorted if there is a weak factor(s) since it is based on the debiased estimator. In settings where the bias-aware CI (12) is too wide to yield precise inference, we recommend reporting the CI (13) alongside the bias-aware CI (12) as a compromise between ignoring weak factors and a fully robust approach.

Remark 3.2 (Choice of R and ε)**.** The quantity ε is used in the bound $\hat{C} = (2+\varepsilon)Rs_1(\hat{U}_{\text{pre}})$ on $\|\widetilde{\Gamma}\|_*$ needed to compute the CI in the final step. While $\varepsilon > 0$ is necessary for theoretical guarantees, in our Monte Carlos, we find that we get good coverage when choosing $\varepsilon = 0$.

In contrast, the choice of R has a substantive effect on the CI, both through the bound $\hat{C} = (2+\varepsilon)Rs_1(\hat{U}_{\text{pre}})$ and through the point estimate. Since the number of weak factors cannot be determined from the data, the researcher must specify an a priori bound on the total number of factors R. Nonetheless, the data can be informative about the number of strong or semi-strong factors, which provides a lower bound for R. We recommend forming an estimate $\hat{R}_s$ of the number of strong factors using one of the standard methods (e.g., Bai and Ng, 2002; Onatski, 2010; Ahn and Horenstein, 2013) and using this as a starting point for examining the sensitivity of the results to the choice of R. For example, by taking $R = \hat{R}_s + 1$, the researcher allows for the potential presence of an additional weak factor (or $R - \hat{R}_s$ weak factors in general for a bigger R).

Remark 3.3 (Lindeberg condition)**.** The asymptotic validity of the CI depends on asymptotic normality of the stochastic term $\langle A, U\rangle_F$ where $A = A^*_{b^*}$ is a non-random matrix of weights. This, in turn, depends on a Lindeberg condition on the weights A. To ensure that this holds, we can modify our optimization procedure for computing the weights $A = A^*_{b^*}$ by imposing a bound on the Lindeberg weights

$$\text{Lind}(A) = \frac{\max_{1\le i\le N, 1\le t\le T} A_{it}^2}{\sum_{i=1}^N \sum_{t=1}^T A_{it}^2}. \tag{14}$$

A similar approach to showing asymptotic validity is taken in Javanmard and Montanari (2014) in a different setting.

To make this approach practical, we need guidance on what makes $\text{Lind}(A)$ "small enough to use the central limit theorem" in a given sample size. A formal answer to this question is elusive, due to the difficulty of obtaining finite sample bounds on approximation error in the central limit theorem that are practically useful. As a heuristic, we can use comparisons to other settings where the central limit theorem is used. For example, the sample mean $\bar{W} = \frac{1}{n}\sum_{i=1}^n W_i$ with n observations corresponds to an estimator with Lindeberg constant $(1/n)^2/[n \cdot (1/n)^2] = 1/n$. If we are comfortable using the normal approximation in such a setting with, say, $n = 50$, then we can impose a bound $\text{Lind}(A) \le 1/50$. Noack and Rothe (2024) provide some discussion of these issues in a related setting involving inference in fuzzy regression discontinuity.

In our Monte Carlos, we find that $\text{Lind}(A)$ is very small for the weights used in Algorithm 3.1 once N and T are larger than, say, 20. Thus, imposing a bound on these weights does not appear to be necessary in practice in the data generating processes we have examined.

Remark 3.4 (Standard error)**.** The standard error $\widehat{\text{se}}^2 = \sum_{i=1}^N \sum_{t=1}^T A_{it}^2 \hat{U}_{\text{pre},it}^2$ assumes that U_{it} is uncorrelated across i and t, but allows for heteroskedasticity. Such an assumption will be reasonable if Γ_{it} captures all of the dependence in errors for the outcome. However, incorporating all dependence in Γ_{it} may lead to an unnecessarily conservative choice of the upper bound R on the number of factors, leading to a wider CI. To avoid such conservative bounds on Γ, one can incorporate any dependence that is not directly correlated with X_{it} into the error term U_{it}, and allow for such dependence when constructing the standard error. For example, to allow for (arbitrary) time dependence of U_{it} (while maintaining uncorrelatedness across i), one could simply use clustered standard errors (see, e.g., Arellano, 1987; Hansen, 2007)

$$\widehat{\text{se}}^2 = \sum_{i=1}^N \left(\sum_{t=1}^T A_{it} \hat{U}_{\text{pre},it} \right)^2.$$

4 Asymptotic results

This section gives formal asymptotic results for the estimators and CIs given in Sections 2 and 3. We consider the following decomposition of our regression model:

$$\widetilde{Y} := Y - \hat{\Gamma}_{\text{pre}} = X\beta + Z \cdot \delta + \Gamma - \hat{\Gamma}_{\text{pre}} + U = X\beta + Z \cdot \delta + \widetilde{\Gamma} + \widetilde{U}. \tag{15}$$

Here, $\widetilde{\Gamma}$ and $\widetilde{U}$ an be any $N \times T$ matrices chosen compatibly so that $\widetilde{\Gamma} + \widetilde{U} = \Gamma - \hat{\Gamma}_{\text{pre}} + U$. While our discussion so far has focused on the case where $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}_{\text{pre}}$ and $\widetilde{U} = U$, it turns out that allowing for other choices of $\widetilde{\Gamma}$ and $\widetilde{U}$ allows for an improvement in the width of our CI.

To formally state asymptotic results that allow for weak factors and an unknown error distribution, we introduce some additional notation. We consider uniform-in-the-underlying distribution asymptotics over a set $\mathcal{P}$ of distributions P for Γ and $X, Z_1, \ldots, Z_K, U$ and a set Θ of parameters $\theta = (\beta, \delta')'$. While we treat $\Gamma, X, Z_1, \ldots, Z_k$ as random variables determined by the unknown probability distribution P for notational purposes, we note that a fixed design setting in which $\Gamma, X, Z_1, \ldots, Z_k$ are non-random (sequences of) matrices can be incorporated by considering a set $\mathcal{P}$ that places a probability one mass on a given value of $\Gamma, X, Z_1, \ldots, Z_k$. We use $\mathbb{P}_{P,\theta}$ to denote probability under the given distribution P and parameters θ. Formally, we consider large N, large T asymptotics in which $N = N_n \to \infty$ and $T = T_n \to \infty$, and we consider sequences of distributions $\mathcal{P} = \mathcal{P}_n$ and parameter spaces $\Theta = \Theta_n$. Asymptotic statements are then taken in the sequence n. However, we suppress the dependence on an index sequence n in order to save on notation. For a sequence of vectors

or matrices $A_{N,T} = A_{N,T}(\theta, P)$ of fixed dimension (which may depend on θ, P), we use the notation $A_{N,T} = \mathcal{O}_{\Theta,\mathcal{P}}(r_{N,T})$ when, for every $\varepsilon > 0$, there exists C_ε such that

$$\limsup \sup_{P\in\mathcal{P},\theta\in\Theta} \mathbb{P}_{P,\theta}\left(r_{N,T}^{-1}\|A_{N,T}\| \geq C_\varepsilon\right) \leq \varepsilon,$$

and we use the notation $A_{N,T} = o_{\Theta,\mathcal{P}}(r_{N,T})$ when, for every $\varepsilon > 0$, we have

$$\limsup \sup_{P\in\mathcal{P},\theta\in\Theta} \mathbb{P}_{P,\theta}\left(r_{N,T}^{-1}\|A_{N,T}\| \geq \varepsilon\right) \to 0.$$

We use the notation $A_{N,T} \asymp_{\Theta,\mathcal{P}} r_{N,T}$ when $A_{N,T} = \mathcal{O}_{\Theta,\mathcal{P}}(r_{N,T})$ and $A_{N,T}^{-1} = \mathcal{O}_{\Theta,\mathcal{P}}(r_{N,T}^{-1})$. We use the notation $A_{N,T} \underset{\Theta,\mathcal{P}}{\overset{d}{\to}} \mathcal{L}$ to denote the statement

$$\limsup \sup_{\theta\in\Theta,P\in\mathcal{P}} \left|\mathbb{P}_{\theta,P}\left(A_{N,T} \leq t\right) - F_{\mathcal{L}}(t)\right| \to 1 \text{ for all } t$$

where $F_{\mathcal{L}}$ denotes the cdf of the probability law $\mathcal{L}$.

4.1 General results

We first show asymptotic validity of the CI (10) under the following high level assumption imposed on the augmented model (15) and the weights A_{it}.

Assumption 1.

(i) $\inf_{\theta\in\Theta,P\in\mathcal{P}} \mathbb{P}_{\theta,P}\left(\|\widetilde{\Gamma}\|_* \leq \hat{C}\right) \to 1$;

(ii) $\frac{\langle A,\widetilde{U}\rangle_F}{\widehat{\text{se}}} \underset{\Theta,\mathcal{P}}{\overset{d}{\to}} N(0,1)$.

Theorem 1. *Suppose that Assumption 1 holds. Then*

$$\liminf \inf_{\theta\in\Theta,P\in\mathcal{P}} \mathbb{P}_{\theta,P}\left(\beta \in \left\{\hat{\beta}_A \pm \left[\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) + z_{1-\alpha/2}\widehat{\text{se}}\right]\right\}\right) \geq 1-\alpha.$$

4.2 Primitive conditions

We now apply these results to the initial estimate and bound given in Section 3, under the assumption of a linear factor model for Γ. We allow for a side condition on the Lindeberg weights $\text{Lind}(A)$ defined in (14), as described in Remark 3.3. Let $A_{b,c}^*$ be defined in the same way as A_b^*, with the modification that we impose the constraint $\text{Lind}(A) \leq c$:

$$\begin{aligned} &\min_A \|A\|_F^2 + b^2 s_1(A)^2, \\ &\text{s.t.} \quad \text{Lind}(A) \leq c, \quad \langle A, X\rangle_F = 1, \quad \langle A, Z_k\rangle_F = 0 \text{ for } k = 1,\ldots,K. \end{aligned} \tag{16}$$

In particular, the weights used in Algorithm 3.1 are given by $A^*_{b^*,\infty} = A^*_{b^*}$, and the weights $A^*_{b^*,c}$ with $c < \infty$ correspond to the modification described in Remark 3.3. In our asymptotic theory we will require $c = c_{NT}$ to converge to zero as $N, T \to \infty$.

We impose the following conditions.

Assumption 2 (Factor Model)**.** *Suppose that rank*$(\Gamma) \leq R$*, i.e.,* $\Gamma = \lambda f'$ *for some* $N \times R$ *matrix* λ *and some* $T \times R$ *matrix* f*, with probability one for all* $P \in \mathcal{P}$ *and the following conditions hold:*

(i) Write W *for* $X, Z_1, \ldots, Z_K$ *and* $W \cdot \gamma = X\beta + \sum_{k=1}^K Z_k \delta_k$ *where* $\gamma = (\beta, \delta')'$*. We assume that there exists* $\underline{s}^2 > 0$ *such that*

$$\min_{\gamma \in \mathbb{R}^{K+1} : \|\gamma\| = 1} \frac{1}{NT} \sum_{r=2R+1}^{\min\{N,T\}} s_r^2(W \cdot \gamma) \geq \underline{s}^2$$

with probability approaching 1 uniformly over $P \in \mathcal{P}$*;*

(ii) $s_1(X) = \mathcal{O}_{\Theta,\mathcal{P}}\left(\sqrt{NT}\right)$*,* $s_1(Z_k) = \mathcal{O}_{\Theta,\mathcal{P}}\left(\sqrt{NT}\right)$ *for* $k \in \{1, \ldots, K\}$*, and* $s_1(U) \asymp_{\Theta,\mathcal{P}} \max\{\sqrt{N}, \sqrt{T}\}$*;*

(iii) $\langle X, U \rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *and* $\langle Z_k, U \rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *for* $k \in \{1, \ldots, K\}$*;*

(iv) $(s_1(U) - s_r(U))/s_1(U) = o_{\Theta,\mathcal{P}}(1)$ *for any fixed positive integer* r*;*

(v) For any sequence of matrices $A = A_{N,T}(X, Z)$ *that is a function of* $X, Z_1, \ldots, Z_k$*, we have* $\langle A, U \rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(\|A\|_F)$*.*

Assumption 2(i) is a generalized non-collinearity condition, which requires that there is enough variation in the regressors after concentrating out $2R$ arbitrary factors. It is closely related to Assumption A of Bai (2009), but our version here avoids mentioning the unobserved factor loadings. The same generalized non-collinearity assumption is imposed in Moon and Weidner (2015). The assumption would be violated if some linear combination $W \cdot \gamma$ of the covariates were to have rank smaller or equal to $2R$. In particular, "low-rank regressors" are ruled out by this condition. Intuitively, Assumption 2(i) holds provided that X_{it} and Z_{it} have non-collinear idiosyncratic components. This intuition is formalized in Appendix A.7, where we also show that, when X is the only regressor in the model (i.e., $K = 0$), then Assumption 3(i) and (ii) below already guarantee that Assumption 2(i) holds.

Assumption 2(ii) places mild bounds on X and Z_k. For example, if the second moments of X_{it} are (uniformly) bounded, then $\mathbb{E}[s_1(X)^2] \leq \mathbb{E}[\|X\|_F^2] = \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}[X_{it}^2] = \mathcal{O}_{\Theta,\mathcal{P}}(NT)$, which, by Markov's inequality, implies $s_1(X) = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$. In addition Assumption 2(ii) also places a rate restriction on $s_1(U)$ that will hold as long as U_{it} does not exhibit too much dependence over i and t. This rate for $s_1(U)$ is closely related to Assumption 2(iv), which is discussed below. Assumption 2(iii) again holds as long as U_{it} does not exhibit too much

dependence over i and t, and is uncorrelated with X_{it} and Z_{it}. Finally, Assumption 2(v) holds as long as U is mean zero given X and Z and satisfies bounds on dependence and second moments.

Assumption 2(iv) is a high level assumption on the first few singular values of U (note that r is fixed as N and T converge to infinity). The singular values of U are the square roots of the eigenvalues of UU'. The random matrix theory literature shows that, if U is an appropriate noise matrix, the largest few eigenvalues of UU' converge to the Tracy-Widom law, after appropriate rescaling: if N and T grow at the same rate, then each of the largest eigenvalues of UU' grows at rate N, while the gaps between them grow at rate $N^{1/3}$. Johnstone (2001) establish the Tracy-Widom law for the largest eigenvalues of UU', for the case of i.i.d. normal error U_{it}. The subsequent literature has shown the universality of this result for more general error distributions, see e.g. Soshnikov (2002), Pillai and Yin (2012) and Yang (2019).

We also place conditions on the matrix X requiring that there is sufficient variation after controlling for individual effects and the additional covariates Z. The constant $c = c_{N,T}$ in the following assumption is the one that appears in our construction of $A^*_{b,c}$.

Assumption 3. *For all $P \in \mathcal{P}$, there exists uniformly bounded $\pi = \pi_P$ and random matrices H and V such that $X = Z \cdot \pi + H + V$ and the following conditions hold:*

(i) $\|V\|_F \asymp_{\Theta,\mathcal{P}} \sqrt{NT}$, $s_1(V) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$;

(ii) $\|H\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *and* $\langle H, V\rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$;

(iii) $\|Z_k\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *and* $\langle Z_k, V\rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *for* $k \in \{1, \dots, K\}$;

(iv) $(\mathbf{Z}'\mathbf{Z})^{-1} = \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{NT}\right)$ *where* $\mathbf{Z} = [\mathrm{vec}(Z_1), \dots, \mathrm{vec}(Z_K)]$;

(v) $\max_{i,t} V_{it}^2 = o_{\Theta,\mathcal{P}}(NTc_{N,T})$ *and* $\max_{i,t} Z_{k,it}^2 = o_{\Theta,\mathcal{P}}\left((NT)^2 c_{N,T}\right)$ *for* $k \in \{1, \dots, K\}$.

Assumption 3 uses a decomposition of X_{it} that depends on an individual effect H_{it} and a random variable V_{it} that is approximately independent and uncorrelated with $Z_{1,it} \dots, Z_{k,it}$ as well as being approximately uncorrelated with the individual effect H_{it}. Importantly, the individual effect H_{it} can be arbitrarily correlated with Γ_{it} and with the variables $Z_{k,it}$. Note also that we do not place any assumptions on the rank or nuclear norm of the matrix H_{it}.

Part (v) holds under a tail bound on V_{it} and $Z_{k,it}$. For example, if V_{it} are (uniformly) sub-Gaussian then $\max_{i,t} V_{it}^2 = \mathcal{O}_{\Theta,\mathcal{P}}(\log(N+T))$, and the condition $\max_{i,t} V_{it}^2 = o_{\Theta,\mathcal{P}}(NTc_{N,T})$ is satisfied provided that $NTc_{N,T}/\log(N+T) \to \infty$. The only other requirement on $c_{N,T}$ is the requirement that $c_{N,T} \max\{N, T\} \to 0$ in Theorem 4 below. Thus, our results allow for a range of choices of $c_{N,T}$.

Define $P_\lambda = \lambda(\lambda'\lambda)^+\lambda$ where M^+ denotes the Moore–Penrose inverse of a matrix M.

Theorem 2. *Let $\hat{\Gamma}_{\mathrm{pre}}$ be defined in Algorithm 3.1, with the modification described in Remark 3.3. Suppose that Assumption 2 holds, and that Assumption 3 holds as stated and with Z_k*

and X interchanged for each $k = 1, \ldots, K$, for the given sequence $c = c_{N,T}$. Then, for any $\varepsilon > 0$, Assumption 1(i) holds with

(i) $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}_{\text{pre}}$ *and* $\hat{C} = 3Rs_1(\hat{U}_{\text{pre}})(1+\varepsilon) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$;

(ii) $\widetilde{\Gamma} = \Gamma + P_\lambda U - \hat{\Gamma}_{\text{pre}}$ *and* $\hat{C} = 2Rs_1(\hat{U}_{\text{pre}})(1+\varepsilon) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$.

Theorem 2 is the main novel technical result that allows us to construct a feasible CI. It provides an explicit bound on the nuclear norm error of our initial estimate. As we show in the proof of Theorem 4 below, the term $\langle A, P_\lambda U\rangle_F$ is asymptotically negligible under our assumptions. Thus, redefining the target parameter to be $\Gamma + P_\lambda U$ instead of Γ and using the bound in part (ii) of the theorem does not affect the construction of the CI. This leads to a shorter CI using the bound in part (ii) compared to using the bound in part (i). For this reason, we use the bound in part (ii) in the implementation described in Section 3 and in our formal coverage results below.

We now turn to the rate of convergence of the debiased estimator and the coverage of the CI. The proofs of these theorems use the nuclear norm bounds in Theorem 2.

Theorem 3. *Let $\hat{\beta} = \hat{\beta}_{A^*_{b^*,c}}$ be defined in Algorithm 3.1, with the modification described in Remark 3.3. Suppose that Assumption 2 holds, and that Assumption 3 holds as stated and with Z_k and X interchanged for each $k = 1, \ldots, K$, for the given sequence $c = c_{N,T}$. Then*

$$\hat{\beta} - \beta = \mathcal{O}_{\Theta,\mathcal{P}}(1/\min\{N,T\}).$$

To obtain primitive conditions for a central limit theorem and asymptotic validity of the confidence interval, we impose that the errors are independent, but not necessarily identically distributed, conditional on X, Z and Γ.

Assumption 4. *There exist constants $\underline{\sigma} > 0$ and $\eta > 0$ such that, for all $P \in \mathcal{P}$, U_{it} is independent over i, t conditional on W, Γ and, for all i, t,*

$$\mathbb{E}_P[U_{it}|W,\Gamma] = 0, \quad \mathbb{E}_P[U_{it}^2|W,\Gamma] > \underline{\sigma}^2, \quad \mathbb{E}_P[U_{it}^4|W,\Gamma] < 1/\eta.$$

Theorem 4. *Let $\hat{\beta} = \hat{\beta}_{A^*_{b^*,c}}$ and $\hat{C} = 2Rs_1(\hat{U}_{\text{pre}})(1+\varepsilon)$ be defined in Algorithm 3.1, with the modification described in Remark 3.3 for $c = c_{N,T}$ with $c_{N,T}\max\{N,T\} \to 0$. Suppose that Assumptions 2(i)-(iv) hold, and that Assumption 3 holds as stated and with Z_k and X interchanged for each $k = 1, \ldots, K$, for the given sequence $c = c_{N,T}$, and that Assumption 4 holds. Let $\widehat{\text{se}}^2 = \sum_{i=1}^N \sum_{t=1}^T A_{it}^2 \hat{U}_{it}^2$ where $A = A^*_{b^*,c}$ and $\hat{U}_{it}$ is the residual from the least squares estimator. Then*

$$\hat{\beta} - \beta = \mathcal{O}_{\Theta,\mathcal{P}}(1/\min\{N,T\})$$

and

$$\liminf \inf_{\theta\in\Theta, P\in\mathcal{P}} \mathbb{P}_{\theta,P}\left(\beta \in \left\{\hat{\beta} \pm \left[\overline{\text{bias}}_{\hat{C}}(\hat{\beta}) + z_{1-\alpha/2}\widehat{\text{se}}\right]\right\}\right) \geq 1-\alpha.$$

4.3 Strong factor case

Numerous studies on the estimation of panel regressions with unobserved factors assume that these factors are "strong" or "semi-strong". This assumption implies that the unobserved error structure, $\Gamma_{it} + U_{it}$ in model (1), viewed as an $N \times T$ matrix, contains an $R = \text{rank}(\Gamma)$ factor component $\Gamma = \lambda f'$ with singular values that asymptotically diverge faster than the largest singular value, $s_1(U)$, of the idiosyncratic error part U. Specifically, as $N, T \to \infty$, the ratio $s_1(U)/s_R(\Gamma)$ approaches zero (at a certain rate) under the semi-strong (strong) factor assumptions. Both Pesaran (2006) and Bai (2009) impose conditions that imply strong factors in this sense, as do many subsequent papers.

A key motivation for the estimation approach in this paper is to avoid assuming strong factors, instead providing an inference method that remains uniformly valid regardless of factor strength. Nevertheless, it is natural to consider how our approach behaves when factors are, in fact, strong, if only to facilitate comparison with much of the existing literature. The following theorem, therefore, extends Theorem 2 to accommodate the case of strong factors.

Theorem 5. *Let $\hat{\Gamma}_{\text{pre}}$ be defined in Algorithm 3.1, with the modification described in Remark 3.3. Suppose that the hypotheses of Theorem 4 hold, and furthermore, assume that $s_1(U)/s_R(\Gamma) = o_{\Theta,\mathcal{P}}(1)$. Then Assumption 1(i) holds with $\widetilde{\Gamma} = \Gamma + M_\lambda U P_f + P_\lambda U M_f - \hat{\Gamma}_{\text{pre}}$ and $\hat{C} = o_{\Theta,\mathcal{P}}\left(s_1(U)\right) = o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$, where $P_f = f(f'f)^+f'$, $M_\lambda = \mathbb{I}_N - P_\lambda$, and $M_f = \mathbb{I}_T - P_f$.*

Theorem 5 additionally requires $s_1(U)/s_R(\Gamma) = o_P(1)$, i.e., that the factors are semi-strong or strong, and also that $R = \text{rank}(\Gamma)$. This implies that $\hat{\Gamma}_{\text{pre}}$ converges to $\Gamma + M_\lambda U P_f + P_\lambda U M_f$ in second leading order (see e.g. Lemma S.3 in the supplement to Moon and Weidner, 2015). Theorem 5 states that with this change of target matrix, the nuclear norm bound $\hat{C}$ is smaller than $s_1(U) \asymp_{\Theta,\mathcal{P}} (\max\{\sqrt{N}, \sqrt{T}\})$. This implies that, when N and T grow at the same rate, the worst-case bias $\overline{\text{bias}}_{\hat{C}}(\hat{\beta})$ used in the construction of the CI in Theorem 4 becomes negligible relative to the standard error $\widehat{\text{se}}$. At the same time, the extra term $\langle A, M_\lambda U P_f + P_\lambda U M_f\rangle_F$ is asymptotically negligible by exactly the same arguments given in the proof of Theorem 4 for the negligibility of the term $\langle A, P_\lambda U\rangle_F$. As a result, in the considered regime, the debiased estimator is asymptotically unbiased and the (non-bias) aware CI (13) is asymptotically valid.

Remark 4.1. Theorem 5 shows that the bias term is asymptotically negligible when N and T grow at the same rate and all R factors are strong. This justifies the CI (13) discussed in Remark 3.1 in the strong factor setting. More generally, in the case where R_w factors are weak and $R_s = R - R_w$ factors are strong, we conjecture that Theorem 5 could be extended

to show that the bias-aware CI (12) is valid with $\hat{C} = 2R_w s_1(\hat{U}_{\text{pre}})(1+\varepsilon)$. In other words, we conjecture that the worst-case bias of our estimator only depends on the number of weak factors.

4.4 Comparison to other results in the literature

Our debiasing approach leads to the faster rate $\min\{N,T\}$ compared to the rate $\min\{\sqrt{N},\sqrt{T}\}$ for $\hat{\beta}_{\text{LS}}$ (see, e.g., Moon and Weidner, 2015). While our results appear to be the first to demonstrate a $\min\{N,T\}$ rate of convergence under the conditions above, recent papers have proposed estimators that use additional structure to construct estimators that achieve the same or better rates. Chetverikov and Manresa (2022) impose a factor structure on X, which corresponds to imposing a low-rank assumption on the matrix H in our Assumption 3. They use this assumption to construct an estimator that, like ours, achieves a $\min\{N,T\}$ rate under weak factors. Zhu (2019) imposes homoskedastic and independent errors in addition to a factor structure on X, and shows that this allows for a faster $\sqrt{NT}$ rate of convergence, even under weak factors.

While robust to weak factors, our CI will be wider than a CI based on the strong factor asymptotics in Bai (2009). Ideally, one would like to form a CI that is *adaptive* to the strength of factors. Such a CI would be robust to weak factors, while being asymptotically equivalent to the CI in Bai (2009) when factors are strong. However, as shown by Zhu (2019), such an adaptive CI cannot be obtained, even if one imposes homoskedastic errors and additional structure on the covariate matrix X. Thus, while there may be some room for efficiency gains over our CI, one must allow for some increase in CI length relative to the CI in Bai (2009) in order to allow for weak factors.

As discussed in the introduction, our debiasing approach is analogous to the approach to debiasing the LASSO taken in Javanmard and Montanari (2014) and, more broadly, other papers in the debiased LASSO literature such as Belloni, Chernozhukov and Hansen (2014), van de Geer, Bühlmann, Ritov and Dezeure (2014) and Zhang and Zhang (2014). Interestingly, this analogy extends to the rates of convergence in our asymptotic results. The debiased lasso applies to a high dimensional regression model with s nonzero coefficients and n observations. The resulting estimator has bias of order s/n, up to log terms, and variance $1/n$. Note that s is the dimension of the constraint set for the unknown parameter, while n is the total number of observations. In our setting, the debiased estimator has bias of order $\max\{N,T\}/(NT)$ and variance $1/(NT)$. The set of matrices Γ with rank at most R has dimension of order $\max\{N,T\}$ so, just as with the debiased lasso, the bias term is of the same order of magnitude as the ratio of the dimension of the constraint set to the total number of observations. In the debiased lasso setting, one can justify a CI that ignores bias by assuming that s increases slowly enough relative to n for the order s/n bias term to be asymptotically negligible relative to the order $1/\sqrt{n}$ standard deviation term. Unfortunately, this cannot occur in our setting even if $R = 1$, since the bias term is of order $\max\{N,T\}/(NT)$ which is always of at least the

same order of magnitude as the standard deviation $1/\sqrt{NT}$. This necessitates our bias-aware approach.

5 Numerical Evidence

5.1 Simulation Study

We consider the following design:

$$Y_{it} = X_{it}\beta + \sum_{r=1}^{R} \kappa_r \lambda_{ir} f_{tr} + U_{it},$$
$$X_{it} = \sum_{r=1}^{R} \lambda_{ir} f_{tr} + V_{it},$$

where κ_r controls the strength of factor f_{tr}, and R stands for the number of factors. In addition, λ_i, f_t, U_{it} and V_{it} are all mutually independent across both i, t, and (i,t), and

$$\lambda_i \sim N(0, I_R) \perp f_t \sim N(0, I_R) \perp \begin{pmatrix} U_{it} \\ V_{it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_U^2 & 0 \\ 0 & \sigma_V^2 \end{pmatrix} \right).$$

In the designs considered below, we fix $(\beta, \sigma_U^2, \sigma_V^2) = (0, 1, 1)$ and vary N, T, the number of factors R, and their strengths controlled by κ_r. The number of simulations in all of the considered designs is 5000. As before, we are interested in estimation of and inference on β.

In Tables 1-3, we report the bias, standard deviation, and rmse for the benchmark LS estimator of Bai (2009) and for the proposed debiased estimator in various designs with 1 and 2 factors.[10] We also report the size of the corresponding tests (with 5% nominal size) and the average length of the CIs (with 95% nominal coverage). For simplicity and for brevity of the reported results, we assume that the number of factors is known. We consider a case when the number of factors is overspecified in Appendix C.2.

The LS estimator is heavily biased and the associated tests and CIs are heavily size distorted unless all the factors are strong. At the same time, the proposed estimator effectively reduces the "weak factors" bias without inflating the variance. As a result, the potential efficiency gains from using the debiased estimator can be very large when there is a weak factor, especially for larger sample sizes (see Appendix C.1 for additional simulation results). Importantly, even if all the factors are strong, the debiased estimator performs comparably to the LS estimator.

When weak factors are present, the LS CIs can have zero coverage because they are (i) centered around the biased LS estimator and (ii) too short. Hence, the average length of the LS CIs is not a proper benchmark to compare the average length of the bias-aware CIs.

[10] Note that the CCE estimator of Pesaran (2006) would not work in these designs, regardless of whether the factors are strong or not, because the cross-sectional averages of λ_{ir} equal zero.

To provide a relevant comparison, we also construct identification robust CIs by inverting the (absolute value of the) LS based t-statistic using appropriate identification robust critical values (instead of $z_{1-\alpha/2}$). Specifically, for a given design (here, for fixed N, T, and R), we (numerically) compute the least favorable (over κ) critical value for the absolute value of the t-statistic based on the LS estimator. We also construct analogous CIs by inverting the (absolute value of the) t-statistic based on the debiased estimator using the corresponding least favorable critical values. We refer to such CIs as the LS and debiased oracle CIs (because they are based on unknown design-specific least favorable critical values) and report their average length denoted by "length*" in the tables below.

Notice that the average length of the LS oracle CIs (the "length*" column under the LS heading) is at least comparable to but mostly significantly greater than the actual length of the bias-aware CIs (the "length" column under the debiased heading), especially for larger sample sizes (again, see Appendix C.1 for additional simulation results). Thus, the bias-aware CI outperforms the LS CI once one corrects the LS CI to compensate for its severe undercoverage.

Another important comparison is between the actual length and oracle length of the bias-aware CI. Throughout most of the designs, the oracle length of the bias-aware CI is slightly less than half the length of the actual bias-aware CI that we compute. This gives a bound on how conservative our CI is: our bias-aware critical value cannot be decreased by more than a factor of about two without sacrificing coverage in these Monte Carlos. There are two possible sources of this conservativeness: (1) the bound in Theorem 2 may be conservative or (2) there may be some additional structure in the initial error or its correlation with the data that our nuclear norm debiasing method does not exploit. While further improving the CI using the proof techniques in this paper appears difficult, we cannot rule out these possibilities. On the other hand, it is possible that these Monte Carlos overstate the conservativeness of our bias-aware CI: there may be other DGPs for which our bias-aware critical value cannot be decreased without sacrificing coverage.

Despite the simplicity of the design considered in this section, the presented findings seem to be characteristic of more complicated and settings. Specifically, in Appendix C.2, we consider a design with an additional covariate and non-Gaussian, heteroskedastic, and serially correlated errors and establish qualitatively similar results regardless of whether the correct number of factors is known or overspecified.

5.2 Empirical Illustration

In this section, we illustrate the finite sample properties of the proposed estimator and confidence intervals in a numerical experiment calibrated to imitate an actual empirical setting. Specifically, we calibrate our experiment based on the seminal studies of the effects of unilateral divorce law reforms on the US divorce rates by Friedberg (1998) and Wolfers (2006), subsequently revisited by Kim and Oka (2014) and Moon and Weidner (2015) in the context

of interactive fixed effects models.

For simplicity of the experiment, as a benchmark, we use the following static specification also considered in Friedberg (1998) and Wolfers (2006)

$$Y_{it} = X_{it}\beta + \alpha_i + \zeta_i t + \nu_i t^2 + \phi_t + U_{it},$$

where Y_{it} denotes the annual divorce rate (per 1,000 persons) in state i in year t, and X_{it} is a dummy variable indicating if state i had a unilateral divorce law in year t. Following Friedberg (1998) and Wolfers (2006), we also control for state-specific quadratic time trends and time effects.

We follow Kim and Oka (2014) and use their data to construct a balanced panel with $N = 48$ states and $T = 33$ years. As in Moon and Weidner (2015), first we profile out the individual trends and time effects from Y_{it} and X_{it} to form the projected model

$$Y_{it}^{\perp} = X_{it}^{\perp}\beta + U_{it}^{\perp}$$

and obtain the estimates $\hat{\beta}$ and $\hat{\sigma}^2_{U^\perp}$. We also extract the first principal component of the matrix of regressors $X^\perp$ denoted by $\Gamma^{X^\perp} = \lambda_i^{X^\perp} f_t^{X^\perp}$.

In our numerical experiment, we fix $X^\perp$, $\{\lambda_i^{X^\perp}\}_{i=1}^N$, and $\{f_t^{X^\perp}\}_{t=1}^T$, and consider the following DGP

$$Y_{it}^{\perp} = X_{it}^{\perp}\hat{\beta} + \kappa\lambda_i^{X^\perp} f_t^{X^\perp} + U_{it}^{\perp},$$

where we introduce an additional factor $f_t^{X^\perp}$ and a parameter κ controlling the strength of $f_t^{X^\perp}$. For every repetition, we draw $U_{it}^{\perp}$ as iid $N(0, \hat{\sigma}^2_U)$ and treat all the other parts of the DGP as fixed.

As before, we compare the LS estimator and inference performance with the proposed approach for various values of κ. Both approaches use the correctly specified number of factors $R = 1$. The results are based on 5,000 simulations and provided in Table 4. We report the same statistics as in Section 5.1.

The results are qualitatively similar to the results in Section 5.1. The LS estimator is heavily biased when the factor is weak, and the standard tests and confidence intervals are severely size distorted. Compared to the LS estimator, the debiased estimator has a substantially smaller bias, standard deviation, and rmse when the factor is weak. It also performs competitively if the factor is strong. The LS CIs are much shorter than the bias-aware CIs but have very poor coverage. The oracle CIs based on the LS estimator have the correct coverage and are also considerably wider than the naive CIs and comparable with the bias-aware CIs. Again, the oracle CIs based on the debiased estimator are considerably shorter than the bias-aware CIs and LS oracle CIs, indicating that there is a potential scope for improvement.

Overall, our empirically calibrated simulation study shows that the presence of a weak

factor can lead to poor performance of conventional estimators and inference procedures in an actual empirical setting. It also demonstrates that in such settings, the gains from using the debiased estimator could be substantial.

Finally, we also report estimation and inference results for the actual data set. For consistency with the numerical experiment above, we focus on the same single covariate X_{it}. In Appendix D, we also consider a specification with dynamic treatment effects as in Wolfers (2006). Similarly to Kim and Oka (2014) and Moon and Weidner (2015), we estimate

$$Y_{it} = X_{it}\beta + \alpha_i + \zeta_i t + \nu_i t^2 + \phi_t + \sum_{r=1}^{R} \lambda_{ir} f_{tr} + U_{it}$$

for various values of R using the LS and the debiased approaches and construct 95% CIs for β. As before, we first profile out the individual trends and time effects, and then use the residual outcomes and regressors as inputs for the LS and debiased estimators.

The results are provided in Table 5. We construct three types of CIs based on the debiased estimator using $\hat{C}$ as in Remark 4.1 with different values of R_w. The first one is constructed assuming that there are no weak factors among R factors ($R_w = 0$), i.e., under the same assumption under which the standard LS CI is valid. This is the same CI as introduced in Remark 3.1. In this application, these CIs are as short as or even shorter than the LS CIs. Thus, if the researcher wants to obtain shorter CIs at the cost of non-robustness to the potential presence of weak factors, they can still do that using our debiased estimator. As pointed out in Remark 3.1 and documented in Section C.3, such CIs are still likely to have much better coverage than the LS ones when there is a weak factors since they are based on the debiased estimator.

The second type of CIs is constructed assuming that among R factors there is up to one weak factor ($R_w = 1$). The corresponding bias-aware CIs are substantially wider than the non-robust ones. However, as the numerical experiment considered earlier in this sections suggests, this is how wide identification robust CIs appear to have to be in this setting. In the considered application, we find that the potential presence of one weak factor is likely to be sufficient to nullify the significance of the previously obtained non-robust estimates.

Finally, we also report our bias-aware CIs as in (12) corresponding to $R_w = R$. These CIs are uniformly valid regardless of the strength of identification of the factors.

The results for a specification with dynamic treatment effects are qualitatively similar and provided in Appendix D.

Table 1: Simulation results for the experiment in Section 5.1, $N = 100$, $R = 1$

	LS						Debiased					
κ	bias	std	rmse	size	length	length*	bias	std	rmse	size	length	length*
						$T = 20$						
0.00	-0.0000	0.0171	0.0171	7.1	0.061	0.299	0.0002	0.0206	0.0206	0.0	0.304	0.137
0.05	0.0242	0.0178	0.0300	37.3	0.062	0.300	0.0095	0.0207	0.0228	0.0	0.304	0.137
0.10	0.0478	0.0200	0.0518	79.3	0.062	0.302	0.0181	0.0215	0.0281	0.0	0.305	0.137
0.15	0.0690	0.0249	0.0734	91.6	0.063	0.308	0.0244	0.0235	0.0339	0.0	0.306	0.138
0.20	0.0792	0.0382	0.0879	85.7	0.067	0.324	0.0250	0.0276	0.0372	0.0	0.309	0.138
0.25	0.0670	0.0531	0.0855	64.8	0.074	0.358	0.0189	0.0306	0.0360	0.0	0.311	0.139
0.50	0.0049	0.0244	0.0248	8.2	0.087	0.425	0.0013	0.0239	0.0240	0.0	0.314	0.139
1.00	0.0004	0.0232	0.0232	5.9	0.088	0.427	0.0001	0.0237	0.0237	0.0	0.315	0.140
						$T = 50$						
0.00	-0.0002	0.0103	0.0103	5.9	0.039	0.228	-0.0001	0.0136	0.0136	0.0	0.173	0.079
0.05	0.0244	0.0108	0.0267	67.5	0.039	0.228	0.0064	0.0137	0.0151	0.0	0.173	0.080
0.10	0.0484	0.0124	0.0500	98.2	0.039	0.230	0.0121	0.0143	0.0187	0.0	0.174	0.080
0.15	0.0683	0.0189	0.0709	96.8	0.040	0.237	0.0135	0.0164	0.0213	0.0	0.175	0.080
0.20	0.0580	0.0390	0.0699	72.4	0.046	0.269	0.0084	0.0180	0.0198	0.0	0.177	0.080
0.25	0.0229	0.0306	0.0382	33.5	0.053	0.308	0.0032	0.0164	0.0167	0.0	0.177	0.080
0.50	0.0016	0.0144	0.0145	5.7	0.055	0.324	0.0002	0.0151	0.0151	0.0	0.177	0.080
1.00	0.0001	0.0142	0.0142	5.1	0.055	0.324	-0.0001	0.0151	0.0151	0.0	0.178	0.080
						$T = 100$						
0.00	-0.0001	0.0073	0.0073	6.1	0.028	0.183	-0.0001	0.0108	0.0108	0.0	0.122	0.057
0.05	0.0246	0.0077	0.0258	91.0	0.028	0.183	0.0051	0.0108	0.0120	0.0	0.122	0.057
0.10	0.0486	0.0093	0.0495	99.9	0.028	0.185	0.0089	0.0117	0.0147	0.0	0.123	0.057
0.15	0.0619	0.0224	0.0658	92.9	0.030	0.197	0.0068	0.0132	0.0149	0.0	0.124	0.058
0.20	0.0239	0.0267	0.0358	47.4	0.037	0.243	0.0023	0.0124	0.0127	0.0	0.124	0.058
0.25	0.0077	0.0120	0.0143	17.9	0.039	0.256	0.0009	0.0119	0.0120	0.0	0.124	0.058
0.50	0.0009	0.0103	0.0103	5.9	0.039	0.260	0.0001	0.0118	0.0118	0.0	0.124	0.058
1.00	0.0001	0.0102	0.0102	5.4	0.039	0.260	-0.0000	0.0118	0.0118	0.0	0.124	0.058
						$T = 300$						
0.00	-0.0000	0.0042	0.0042	5.3	0.016	0.121	0.0000	0.0056	0.0056	0.0	0.080	0.033
0.05	0.0247	0.0046	0.0252	100.0	0.016	0.122	0.0047	0.0056	0.0073	0.0	0.080	0.033
0.10	0.0482	0.0070	0.0487	99.8	0.016	0.123	0.0057	0.0067	0.0088	0.0	0.080	0.033
0.15	0.0178	0.0173	0.0248	60.5	0.021	0.161	0.0015	0.0063	0.0065	0.0	0.081	0.033
0.20	0.0047	0.0064	0.0080	16.4	0.022	0.170	0.0005	0.0061	0.0061	0.0	0.081	0.033
0.25	0.0023	0.0060	0.0064	7.6	0.023	0.171	0.0003	0.0061	0.0061	0.0	0.081	0.033
0.50	0.0003	0.0057	0.0058	4.9	0.023	0.172	0.0001	0.0060	0.0060	0.0	0.081	0.033
1.00	0.0001	0.0057	0.0057	4.9	0.023	0.172	0.0000	0.0060	0.0060	0.0	0.081	0.033

$\text{Lind}(A) \in \{0.0063, 0.0028, 0.0015, 0.0006\}$ for $T \in \{20, 50, 100, 300\}$.

Table 2: Simulation results for the experiment in Section 5.1, $N = 100$, $T = 50$, $R = 2$

κ_1 \ κ_2	LS										Debiased									
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	1.00	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	1.00
bias																				
0.00	-0.000	0.016	0.031	0.035	0.018	0.008	0.004	0.002	0.001	0.000	-0.000	0.006	0.010	0.010	0.005	0.002	0.001	0.000	0.000	-0.000
0.05	0.016	0.033	0.049	0.060	0.052	0.036	0.030	0.027	0.026	0.025	0.006	0.012	0.017	0.018	0.014	0.010	0.009	0.008	0.007	0.007
0.10	0.031	0.049	0.066	0.080	0.083	0.067	0.057	0.051	0.050	0.049	0.010	0.017	0.022	0.025	0.022	0.018	0.015	0.014	0.014	0.013
0.15	0.035	0.060	0.080	0.097	0.108	0.099	0.082	0.072	0.070	0.068	0.010	0.018	0.025	0.029	0.028	0.024	0.020	0.018	0.017	0.017
0.20	0.018	0.052	0.083	0.108	0.123	0.114	0.089	0.068	0.064	0.060	0.005	0.014	0.022	0.028	0.029	0.024	0.018	0.014	0.013	0.012
0.25	0.008	0.036	0.067	0.099	0.114	0.088	0.054	0.032	0.028	0.025	0.002	0.010	0.018	0.024	0.024	0.017	0.010	0.006	0.005	0.005
0.30	0.004	0.030	0.057	0.082	0.089	0.054	0.027	0.015	0.012	0.010	0.001	0.009	0.015	0.020	0.018	0.010	0.005	0.003	0.002	0.002
0.40	0.002	0.027	0.051	0.072	0.068	0.032	0.015	0.007	0.005	0.004	0.000	0.008	0.014	0.018	0.014	0.006	0.003	0.001	0.001	0.001
0.50	0.001	0.026	0.050	0.070	0.064	0.028	0.012	0.005	0.004	0.002	0.000	0.007	0.014	0.017	0.013	0.005	0.002	0.001	0.001	0.000
1.00	0.000	0.025	0.049	0.068	0.060	0.025	0.010	0.004	0.002	0.000	-0.000	0.007	0.013	0.017	0.012	0.005	0.002	0.001	0.000	-0.000
std																				
0.00	0.009	0.009	0.012	0.019	0.019	0.013	0.011	0.011	0.010	0.010	0.013	0.013	0.014	0.015	0.015	0.014	0.014	0.014	0.014	0.014
0.05	0.009	0.009	0.010	0.014	0.021	0.015	0.012	0.011	0.011	0.011	0.013	0.013	0.013	0.015	0.015	0.014	0.014	0.014	0.014	0.014
0.10	0.012	0.010	0.010	0.012	0.019	0.020	0.014	0.013	0.013	0.013	0.014	0.013	0.014	0.015	0.016	0.015	0.015	0.014	0.014	0.014
0.15	0.019	0.014	0.012	0.012	0.017	0.025	0.021	0.019	0.019	0.020	0.015	0.015	0.015	0.016	0.016	0.017	0.016	0.016	0.016	0.016
0.20	0.019	0.021	0.019	0.017	0.025	0.043	0.045	0.040	0.039	0.039	0.015	0.015	0.016	0.016	0.018	0.020	0.020	0.019	0.019	0.019
0.25	0.013	0.015	0.020	0.025	0.043	0.064	0.055	0.037	0.034	0.032	0.014	0.014	0.015	0.017	0.020	0.023	0.021	0.018	0.018	0.017
0.30	0.011	0.012	0.014	0.021	0.045	0.055	0.036	0.021	0.019	0.019	0.014	0.014	0.015	0.016	0.020	0.021	0.018	0.016	0.016	0.016
0.40	0.011	0.011	0.013	0.019	0.040	0.037	0.021	0.016	0.015	0.015	0.014	0.014	0.014	0.016	0.019	0.018	0.016	0.016	0.016	0.016
0.50	0.010	0.011	0.013	0.019	0.039	0.034	0.019	0.015	0.015	0.015	0.014	0.014	0.014	0.016	0.019	0.018	0.016	0.016	0.015	0.015
1.00	0.010	0.011	0.013	0.020	0.039	0.032	0.019	0.015	0.015	0.015	0.014	0.014	0.014	0.016	0.019	0.017	0.016	0.016	0.015	0.015
rmse																				
0.00	0.009	0.019	0.033	0.039	0.026	0.015	0.012	0.011	0.010	0.010	0.013	0.014	0.017	0.018	0.016	0.014	0.014	0.014	0.014	0.014
0.05	0.019	0.034	0.050	0.061	0.056	0.039	0.032	0.029	0.028	0.027	0.014	0.017	0.021	0.023	0.021	0.017	0.016	0.016	0.016	0.016
0.10	0.033	0.050	0.066	0.081	0.085	0.070	0.058	0.053	0.051	0.050	0.017	0.021	0.026	0.029	0.027	0.023	0.021	0.020	0.020	0.020
0.15	0.039	0.061	0.081	0.098	0.109	0.102	0.085	0.075	0.073	0.071	0.018	0.023	0.029	0.033	0.033	0.029	0.026	0.024	0.024	0.023
0.20	0.026	0.056	0.085	0.109	0.125	0.122	0.099	0.079	0.075	0.071	0.016	0.021	0.027	0.033	0.034	0.032	0.027	0.024	0.023	0.022
0.25	0.015	0.039	0.070	0.102	0.122	0.109	0.077	0.049	0.044	0.041	0.014	0.017	0.023	0.029	0.032	0.028	0.023	0.019	0.018	0.018
0.30	0.012	0.032	0.058	0.085	0.099	0.077	0.045	0.026	0.023	0.021	0.014	0.016	0.021	0.026	0.027	0.023	0.018	0.016	0.016	0.016
0.40	0.011	0.029	0.053	0.075	0.079	0.049	0.026	0.017	0.016	0.016	0.014	0.016	0.020	0.024	0.024	0.019	0.016	0.016	0.016	0.016
0.50	0.010	0.028	0.051	0.073	0.075	0.044	0.023	0.016	0.015	0.015	0.014	0.016	0.020	0.024	0.023	0.018	0.016	0.016	0.016	0.015
1.00	0.010	0.027	0.050	0.071	0.071	0.041	0.021	0.016	0.015	0.015	0.014	0.016	0.020	0.023	0.022	0.018	0.016	0.016	0.015	0.015

Table 3: Simulation results for the experiment in Section 5.1, $N = 100$, $T = 50$, $R = 2$

κ_1 \ κ_2	LS										Debiased									
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	1.00	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	1.00
size																				
0.00	7.6	52.3	88.3	80.0	42.8	17.9	10.6	6.7	6.3	5.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.05	52.3	96.2	99.8	99.6	95.9	88.3	81.6	73.7	70.8	68.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.10	88.3	99.8	100.0	100.0	100.0	99.7	99.2	98.6	98.4	98.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.15	80.0	99.6	100.0	100.0	99.8	99.5	98.7	97.9	97.4	96.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.20	42.8	95.9	100.0	99.8	98.3	93.1	87.3	80.0	76.9	73.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.25	17.9	88.3	99.7	99.5	93.1	76.1	60.5	45.2	39.9	36.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.30	10.6	81.6	99.2	98.7	87.3	60.5	38.9	23.3	18.9	16.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.40	6.7	73.7	98.6	97.9	80.0	45.2	23.3	11.3	9.2	7.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	6.3	70.8	98.4	97.4	76.9	39.9	18.9	9.2	7.4	6.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.00	5.9	68.0	98.0	96.6	73.8	36.2	16.2	7.9	6.4	5.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
length																				
0.00	0.031	0.031	0.032	0.034	0.037	0.038	0.039	0.039	0.039	0.039	0.257	0.257	0.258	0.260	0.261	0.262	0.262	0.262	0.262	0.262
0.05	0.031	0.031	0.032	0.033	0.036	0.038	0.038	0.039	0.039	0.039	0.257	0.257	0.258	0.259	0.261	0.262	0.262	0.262	0.262	0.262
0.10	0.032	0.032	0.032	0.032	0.034	0.037	0.039	0.039	0.039	0.039	0.258	0.258	0.258	0.260	0.261	0.262	0.263	0.263	0.263	0.263
0.15	0.034	0.033	0.032	0.032	0.033	0.036	0.039	0.040	0.040	0.040	0.260	0.259	0.260	0.261	0.263	0.264	0.265	0.265	0.265	0.265
0.20	0.037	0.036	0.034	0.033	0.034	0.037	0.042	0.045	0.045	0.046	0.261	0.261	0.261	0.263	0.265	0.266	0.267	0.267	0.268	0.268
0.25	0.038	0.038	0.037	0.036	0.037	0.043	0.049	0.052	0.052	0.052	0.262	0.262	0.262	0.264	0.266	0.268	0.268	0.269	0.269	0.269
0.30	0.039	0.038	0.039	0.039	0.042	0.049	0.053	0.054	0.054	0.054	0.262	0.262	0.263	0.265	0.267	0.268	0.269	0.269	0.269	0.269
0.40	0.039	0.039	0.039	0.040	0.045	0.052	0.054	0.055	0.055	0.055	0.262	0.262	0.263	0.265	0.267	0.269	0.269	0.269	0.269	0.270
0.50	0.039	0.039	0.039	0.040	0.045	0.052	0.054	0.055	0.055	0.055	0.262	0.262	0.263	0.265	0.268	0.269	0.269	0.269	0.270	0.270
1.00	0.039	0.039	0.039	0.040	0.046	0.052	0.054	0.055	0.055	0.055	0.262	0.262	0.263	0.265	0.268	0.269	0.269	0.270	0.270	0.270
length*																				
0.00	0.345	0.346	0.353	0.373	0.406	0.420	0.424	0.427	0.427	0.428	0.114	0.114	0.114	0.115	0.115	0.115	0.115	0.115	0.115	0.115
0.05	0.346	0.346	0.349	0.360	0.391	0.416	0.424	0.427	0.428	0.429	0.114	0.114	0.114	0.115	0.115	0.115	0.115	0.115	0.115	0.115
0.10	0.353	0.349	0.348	0.354	0.375	0.409	0.424	0.430	0.432	0.432	0.114	0.114	0.114	0.115	0.115	0.115	0.115	0.115	0.115	0.116
0.15	0.373	0.360	0.354	0.354	0.365	0.399	0.428	0.441	0.443	0.445	0.115	0.115	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.116
0.20	0.406	0.391	0.375	0.365	0.371	0.410	0.459	0.491	0.497	0.503	0.115	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.116	0.116
0.25	0.420	0.416	0.409	0.399	0.410	0.475	0.536	0.568	0.573	0.576	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.116	0.116	0.116
0.30	0.424	0.424	0.424	0.428	0.459	0.536	0.579	0.595	0.598	0.599	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.116	0.116	0.117
0.40	0.427	0.427	0.430	0.441	0.491	0.568	0.595	0.604	0.606	0.607	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.117	0.117	0.117
0.50	0.427	0.428	0.432	0.443	0.497	0.573	0.598	0.606	0.608	0.609	0.115	0.115	0.115	0.116	0.116	0.116	0.116	0.117	0.117	0.117
1.00	0.428	0.429	0.432	0.445	0.503	0.576	0.599	0.607	0.609	0.610	0.115	0.115	0.116	0.116	0.116	0.116	0.117	0.117	0.117	0.117

Table 4: Simulation results for the empirically calibrated experiment, $N = 48$, $T = 33$, $R = 1$

	LS						Debiased					
κ	bias	std	rmse	size	length	length*	bias	std	rmse	size	length	length*
0.00	-0.0007	0.0647	0.0647	6.9	0.236	1.062	-0.0010	0.0797	0.0797	0.0	1.374	0.683
0.20	0.0920	0.0656	0.1130	35.0	0.237	1.067	0.0517	0.0805	0.0957	0.0	1.376	0.685
0.40	0.1822	0.0703	0.1953	81.9	0.240	1.083	0.1013	0.0842	0.1317	0.0	1.381	0.690
0.60	0.2620	0.0890	0.2767	93.4	0.248	1.116	0.1392	0.0966	0.1694	0.0	1.392	0.698
0.80	0.2999	0.1467	0.3339	84.6	0.262	1.180	0.1428	0.1267	0.1910	0.0	1.406	0.703
1.00	0.2356	0.2168	0.3202	57.5	0.286	1.288	0.0972	0.1470	0.1762	0.0	1.418	0.700
1.20	0.1134	0.1955	0.2260	26.8	0.307	1.381	0.0420	0.1258	0.1326	0.0	1.424	0.693
1.40	0.0427	0.1168	0.1243	12.6	0.315	1.419	0.0177	0.1042	0.1057	0.0	1.426	0.690
1.60	0.0235	0.0921	0.0950	8.8	0.317	1.428	0.0104	0.0992	0.0997	0.0	1.427	0.690
1.80	0.0157	0.0879	0.0893	7.3	0.318	1.431	0.0070	0.0981	0.0984	0.0	1.428	0.689
2.00	0.0112	0.0867	0.0875	6.8	0.318	1.432	0.0050	0.0977	0.0978	0.0	1.428	0.689

Table 5: LS and debiased estimates and 95% CIs for β

	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	$R = 6$
			LS			
	0.047	0.160	0.101	0.043	0.028	0.091
	$[-0.06, 0.15]$	$[0.04, 0.28]$	$[-0.02, 0.22]$	$[-0.07, 0.16]$	$[-0.10, 0.16]$	$[-0.04, 0.22]$
			Debiased			
	0.089	0.162	0.130	0.084	0.071	0.106
$R_w = 0$	$[-0.01, 0.19]$	$[0.07, 0.26]$	$[0.05, 0.21]$	$[0.01, 0.16]$	$[-0.01, 0.15]$	$[0.04, 0.18]$
$R_w = 1$	$[-0.77, 0.95]$	$[-0.56, 0.88]$	$[-0.45, 0.71]$	$[-0.40, 0.57]$	$[-0.34, 0.48]$	$[-0.24, 0.45]$
$R_w = R$	$[-0.77, 0.95]$	$[-1.18, 1.50]$	$[-1.43, 1.69]$	$[-1.62, 1.79]$	$[-1.67, 1.81]$	$[-1.61, 1.82]$

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics 101*(2), 219–255.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics 174*(1), 1–14.

Alidaee, H., E. Auerbach, and M. P. Leung (2020). Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*.

Andrews, D. W. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics & Statistics 49*(4).

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021, December). Synthetic Difference-in-Differences. *American Economic Review 111*(12), 4088–4118.

Armstrong, T. B. and M. Kolesár (2018). Optimal inference in a class of regression models. *Econometrica 86*(2), 655–683.

Armstrong, T. B., M. Kolesár, and S. Kwon (2020). Bias-Aware Inference in Regularized Regression Models. *arXiv:2012.14823 [econ, stat]*.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association 0*(0), 1–15.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002, January). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2017). Principal components and regularized estimation of factor models. *arXiv preprint arXiv:1708.08137*.

Bai, J. and S. Ng (2023). Approximate factor models with weaker loadings. *Journal of Econometrics 235*(2), 1893–1916.

Bai, J. and P. Wang (2016). Econometric analysis of large factor models. *Annual Review of Economics 8*, 53–80.

Belloni, A., M. Chen, O. H. Madrid Padilla, and Z. Wang (2023). High-dimensional latent panel quantile regression with an application to asset pricing. *The Annals of Statistics 51*(1), 96–121.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies 81*(2), 608–650.

Beyhum, J. and E. Gautier (2019). Square-root nuclear norm penalized estimator for panel data models with approximately low-rank unobserved heterogeneity. *arXiv preprint arXiv:1904.09192*.

Beyhum, J. and E. Gautier (2022). Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. *Journal of Business & Economic Statistics*, 1–12.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). A Wiley-Interscience publication. John Wiley and Sons.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Chamberlain, G. and M. J. Moreira (2009). Decision theory applied to a linear panel data model. *Econometrica 77*(1), 107–133.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2019). Inference for heterogeneous effects using low-rank estimation of factor slopes.

Chetverikov, D. and E. Manresa (2022). Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*.

Cox, G. F. (2024). Weak identification in low-dimensional factor models with one or two factors. *The Review of Economics and Statistics 03*, 1–45.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 238–270.

Fan, J. and Y. Liao (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association 117*(538), 909–924.

Fan, K. (1951). Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences 37*(11), 760–766.

Feng, J. (2023). Nuclear norm regularized quantile regression with interactive fixed effects. *Econometric Theory*, 1–31.

Ferman, B. and C. Pinto (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics 12*(4), 1197–1221.

Fernández-Val, I., H. Freeman, and M. Weidner (2021). Low-rank approximations of nonseparable panel models. *The Econometrics Journal 24*(2), C40–C77.

Freeman, H. and M. Weidner (2023). Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics 237*(1), 105498.

Friedberg, L. (1998). Did unilateral divorce raise divorce rates? Evidence from panel data. *American Economic Review*, 608–627.

Geman, S. (1980). A limit theorem for the norm of random matrices. *The Annals of Probability 8*(2), 252–261.

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

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Higgins, A. (2021). Fixed T estimation of linear panel data models with interactive fixed effects. *arXiv preprint arXiv:2110.05579*.

Hirshberg, D. A. and S. Wager (2020). Augmented minimax linear estimation. arXiv: 1712.00038.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 56*(6), 1371–95.

Horn, R. A. and C. R. Johnson (2013). *Matrix Analysis* (2 ed.). Cambridge University Press.

Ibragimov, I. and R. Khas'minskii (1985). On Nonparametric Estimation of the Value of a Linear Functional in Gaussian White Noise. *Theory of Probability & Its Applications 29*(1), 18–32.

Ishihara, T. and T. Kitagawa (2021). Evidence Aggregation for Treatment Choice. *arXiv preprint arXiv:2108.06473*.

Javanmard, A. and A. Montanari (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research 15*(82), 2869–2909.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics 29*(2), 295–327.

Juodis, A. and V. Sarafidis (2018). Fixed T dynamic panel data estimators with multifactor errors. *Econometric Reviews 37*(8), 893–929.

Juodis, A. and V. Sarafidis (2022). A linear estimator for factor-augmented fixed-T panels with endogenous regressors. *Journal of Business & Economic Statistics 40*(1), 1–15.

Kiefer, N. (1980). A time series-cross section model with fixed effects with an intertemporal factor structure. *Unpublished manuscript, Department of Economics, Cornell University*.

Kim, D. and T. Oka (2014). Divorce law reforms and divorce rates in the usa: An interactive fixed-effects approach. *Journal of Applied Econometrics*.

Ma, S., L. Su, and Y. Zhang (2022). Detecting latent communities in network formation models. *The Journal of Machine Learning Research 23*(1), 13971–14031.

Miao, K., K. Li, and L. Su (2020). Panel threshold models with interactive fixed effects. *Journal of Econometrics 219*(1), 137–170.

Miao, K., P. C. Phillips, and L. Su (2023). High-dimensional vars with common factors. *Journal of Econometrics 233*(1), 155–183.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Moon, H. R. and M. Weidner (2018). Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987*.

Noack, C. and C. Rothe (2024). Bias-Aware Inference in Fuzzy Regression Discontinuity Designs. *Econometrica 92*(3), 687–711.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multi-

factor error structure. *Econometrica 74*(4), 967–1012.

Pillai, N. S. and J. Yin (2012). Edge universality of correlation matrices. *The Annals of Statistics*, 1737–1763.

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review 52*(3), 471–501.

Robertson, D. and V. Sarafidis (2015). IV estimation of panels with factor residuals. *Journal of Econometrics 185*(2), 526–541.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica 56*(4), 931–954.

Rohde, A. and A. B. Tsybakov (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics 39*(2), 887–930.

Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics 108*(5), 1033–1056.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Wang, Y., L. Su, and Y. Zhang (2022). Low-rank panel quantile regression: Estimation and inference. *arXiv preprint arXiv:2210.11062*.

Westerlund, J., Y. Petrova, and M. Norkute (2019). CCE in fixed-T panels. *Journal of Applied Econometrics 34*(5), 746–761.

Wolfers, J. (2006). Did unilateral divorce laws raise divorce rates? a reconciliation and new results. *American Economic Review 96*, 1802–1820.

Yang, F. (2019). Edge universality of separable covariance matrices. *Electron. J. Probab 24*(123), 1–57.

Yata, K. (2021). Optimal decision rules under partial identification. *arXiv preprint arXiv:2111.04926*.

Zeleneev, A. (2019). Identification and estimation of network models with nonparametric unobserved heterogeneity. *Department of Economics, Princeton University*.

Zhang, C.-H. and 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*(1), 217–242.

Zhu, Y. (2019). How well can we learn large factor models without assuming strong factors? *arXiv preprint arXiv:1910.10382*.

Supplementary Appendix to "Robust Estimation and Inference in Panels with Interactive Fixed Effects"

Timothy B. Armstrong, Martin Weidner and Andrei Zeleneev

May 9, 2025

A Proofs

This section contains proofs of the results in the main text. Section A.1 states and proves a general result on rates of convergence using high level conditions on the covariates X and Z and the bound $\hat{C}$ on $\|\Gamma - \hat{\Gamma}\|_*$. The proofs of Theorems 1-4 are provided in Sections A.2-A.5, respectively.

Notation In this section, mirroring the notation used in Assumption 2(i), we write W for $X, Z_1, \ldots, Z_K$ and $W \cdot \gamma = X\beta + \sum_{k=1}^K Z_k \delta_k$ for a generic $\gamma = (\beta, \delta')'$. In particular, we also use $\hat{\gamma} = (\hat{\beta}, \hat{\delta}')'$, $\hat{\gamma}_{\text{LS}} = (\hat{\beta}_{\text{LS}}, \hat{\delta}_{\text{LS}}')'$, and $\hat{\gamma}_{\text{pre}} = (\hat{\beta}_{\text{pre}}, \hat{\delta}_{\text{pre}}')'$.

A.1 General result for rates of convergence

We first prove a result giving rates of convergence for estimators $\hat{\beta} = \langle A^*_{b,c}, \widetilde{Y} \rangle_F$ given in Definition 2.1 with weights $A^*_{b,c}$ given in (16) under a high level condition on the bound $\hat{C}$ on the initial estimation error in (4). Lemma 9(ii) and Theorem 2(i) verify this condition for $\hat{\Gamma} = \hat{\Gamma}_{\text{LS}}$ and $\hat{\Gamma} = \hat{\Gamma}_{\text{pre}}$ (and provide appropriate bounds $\hat{C}$), respectively.

We make the following assumption on the class of distributions of X, $Z_1, \ldots, Z_k$ and U and the sequence $c = c_{N,T}$ used in the Lindeberg constraint. After we prove the main result of this section, we will also verify this assumption under a set of primitive conditions.

Assumption 5. *There exists a sequence of $N \times T$ random matrices Ξ such that*

$$\|\Xi\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT}), \quad |\langle \Xi, X \rangle_F|^{-1} = \mathcal{O}_{\Theta,\mathcal{P}}((NT)^{-1}),$$
$$s_1(\Xi) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),$$

and, with probability approaching one,

$$\text{Lind}(\Xi) \leq c_{N,T} \quad \textit{and} \quad \langle \Xi, Z_k \rangle_F = 0 \textit{ for } k = 1, \ldots, K.$$

Theorem 6. *Let $\hat{\beta} = \langle A^*_{b,c}, \widetilde{Y} \rangle_F$ for $\widetilde{Y} = \Gamma - \hat{\Gamma}$ and some sequences $c = c_{N,T}$ and $b = b_{N,T}$. Suppose Assumption 5 and Assumption 2(v) hold and that Assumption 1(i) holds with $\widetilde{\Gamma} = \Gamma - \hat{\Gamma}$ and $\hat{C} = \mathcal{O}_{\Theta,\mathcal{P}}(\overline{C}_{N,T})$ for some sequence $\overline{C}_{N,T}$. Then*

$$|\hat{\beta} - \beta| = \mathcal{O}_{\Theta,\mathcal{P}}\left(\max\left\{\overline{C}_{N,T}/b_{N,T}, 1\right\} \cdot \max\left\{(NT)^{-1/2}, b_{N,T} \cdot \max\{\sqrt{N}, \sqrt{T}\}/(NT)\right\}\right).$$

Proof. We have

$$|\hat{\beta} - \beta| \leq \left|\langle A^*_{b,c}, U\rangle_F\right| + \overline{\text{bias}}_{\tilde{C}}(A^*_{b,c}) = \left|\langle A^*_{b,c}, U\rangle_F\right| + \tilde{C} s_1(A^*_{b,c})$$

where $\tilde{C} = \|\Gamma - \hat{\Gamma}\|_*$. Thus,

$$\begin{aligned}
|\hat{\beta} - \beta|^2 &\leq 2\left|\langle A^*_{b,c}, U\rangle_F\right|^2 + 2\tilde{C}^2 s_1(A^*_{b,c})^2 \\
&\leq 2\max\left\{\frac{\left|\langle A^*_{b,c}, U\rangle_F\right|^2}{\|A^*_{b,c}\|_F^2}, \frac{\tilde{C}^2}{b^2}\right\} \cdot \left[\|A^*_{b,c}\|_F^2 + b^2 s_1(A^*_{b,c})^2\right]. \qquad (17)
\end{aligned}$$

Consider the oracle weights $\tilde{A} = \Xi/\langle \Xi, X\rangle_F$. With probability approaching one uniformly over θ, P, the weights $\tilde{A}$ are feasible for (16), so that

$$\begin{aligned}
&\|A^*_{b,c}\|_F^2 + b^2 s_1(A^*_{b,c})^2 \leq \|\tilde{A}\|_F^2 + b^2 s_1(\tilde{A})^2 = \frac{\|\Xi\|_F^2 + b^2 s_1(\Xi)^2}{\left|\langle \Xi, X\rangle_F\right|^2} \\
&= \mathcal{O}_{\Theta,\mathcal{P}}((NT)^{-1}) + b^2 \cdot \mathcal{O}_{\Theta,\mathcal{P}}(\max\{N,T\}/(NT)^2). \qquad (18)
\end{aligned}$$

Plugging this into (17) gives the result. □

Next, we verify that Assumption 5 holds under low level conditions provided in Assumption 3.

Lemma 7. *Suppose that Assumption 3 holds. Then Assumption 5 holds with Ξ_{it} given by the residual in the regression of V_{it} on Z_{it}, i.e., $\text{vec}(\Xi) = M_{\mathbf{Z}}\text{vec}(V)$ where $M_{\mathbf{Z}} = I_{NT} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.*

Proof. First, notice that $\Xi = V - Z \cdot \hat{\varphi} = V - \sum_{k=1}^K Z_k \hat{\varphi}_k$ where $\hat{\varphi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\text{vec}(V)$. Also, it follows from Assumption 3(iii) and (iv) that $\|\hat{\varphi}\| = \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\sqrt{NT}}\right)$.

Next we verify all the conditions required by Assumption 5.

Verification of $\langle \Xi, Z_k\rangle_F = 0$ for $k = 1, \ldots, K$. By construction.

Verification of $\|\Xi\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$. $\|\Xi\|_F = \|\text{vec}(\Xi)\| \leq \|\text{vec}(V)\| = \|V\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$.

Verification of $s_1(\Xi) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$. Notice that

$$s_1(\Xi) = s_1\left(V - \sum_{k=1}^K Z_k \hat{\varphi}_k\right) \leq s_1(V) + \sum_{k=1}^K |\hat{\varphi}_k|\, s_1(Z_k) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),$$

using the fact that $|\hat{\varphi}_k|\, s_1(Z_k) = \mathcal{O}_{\Theta,\mathcal{P}}(1)$ since $\hat{\varphi}_k = \mathcal{O}_{\Theta,\mathcal{P}}(1/\sqrt{NT})$ and $s_1(Z_k) \leq \|Z_k\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$.

Verification of $|\langle \Xi, X\rangle_F|^{-1} = \mathcal{O}_{\Theta,\mathcal{P}}((NT)^{-1})$. Using the fact that $\langle \Xi, Z_k\rangle_F = 0$ for each k, we

have

$$\langle \Xi, X \rangle_F = \langle \Xi, H \rangle_F + \langle \Xi, V \rangle_F, = \langle V, H \rangle_F - \sum_{k=1}^{K} \hat{\varphi}_k \langle Z_k, H \rangle_F + \|V\|_F^2 - \sum_{k=1}^{K} \hat{\varphi}_k \langle Z_k, V \rangle_F.$$

The first term is $\mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT}) = o_{\Theta,\mathcal{P}}(NT)$ by Assumption 3(ii). The second term is $\mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT}) = o_{\Theta,\mathcal{P}}(NT)$ since $\hat{\varphi}_k = \mathcal{O}_{\Theta,\mathcal{P}}(1/\sqrt{NT})$ and $\langle Z_k, H \rangle_F \leq \|H\|_F \cdot \|Z_k\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(NT)$ under these assumptions. Similarly, the fourth term is $\mathcal{O}_{\Theta,\mathcal{P}}(1) = o_{\Theta,\mathcal{P}}(NT)$. Thus, $\langle \Xi, X \rangle_F = \|V\|_F^2 + o_{\Theta,\mathcal{P}}(NT)$ and the result follows since $\|V\|_F^2 \asymp_{\Theta,\mathcal{P}} NT$ by Assumption 3(i).

Verification of $\text{Lind}(\Xi) \leq c_{N,T}$ *with probability approaching one.*

$$\text{Lind}(\Xi) = \frac{\max_{i,t} \Xi_{it}^2}{\|\Xi\|_F^2},$$

where

$$\|\Xi\|_F^2 = \|V\|_F^2 - 2 \sum_{k=1}^{K} \hat{\varphi}_k \langle V, Z_k \rangle_F + \left\| \sum_{k=1}^{K} Z_k \hat{\varphi}_k \right\|_F^2,$$

where $\sum_{k=1}^{K} \hat{\varphi}_k \langle V, Z_k \rangle_F = \mathcal{O}_{\Theta,\mathcal{P}}(1)$ and $\left\| \sum_{k=1}^{K} Z_k \hat{\varphi}_k \right\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(1)$, so $\|\Xi\|_F^2 \asymp_{\Theta,\mathcal{P}} NT$. Next,

$$\begin{aligned}
\max_{i,t} \Xi_{it}^2 &= \max_{i,t} \left(V_{it} - \sum_{k=1}^{K} \hat{\varphi}_k Z_{k,it} \right)^2 \\
&\leq (K+1)^2 \left(\max_{i,t} V_{it}^2 + \sum_{k=1}^{K} \hat{\varphi}_k^2 \max_{i,t} Z_{k,it}^2 \right) = o_{\Theta,\mathcal{P}}(NT c_{N,T}).
\end{aligned}$$

Hence, $\text{Lind}(\Xi) = o_{\Theta,\mathcal{P}}(c_{N,T})$, which completes the proof. □

A.2 Proof of Theorem 1

The probability that the upper endpoint of the CI is less than β is

$$\begin{aligned}
&\mathbb{P}_{\theta,P} \left(\hat{\beta} + \overline{\text{bias}}_{\hat{C}}(\hat{\beta}) + z_{1-\alpha/2} \widehat{\text{se}} < \beta \right) \\
&= \mathbb{P}_{\theta,P} \left(\langle A, X\beta + Z \cdot \delta + \widetilde{\Gamma} \rangle_F - \beta + \overline{\text{bias}}_{\hat{C}}(\hat{\beta}) + \langle A, \widetilde{U} \rangle_F < -z_{1-\alpha/2} \widehat{\text{se}} \right) \\
&\leq \mathbb{P}_{\theta,P} \left(\langle A, X\beta + Z \cdot \delta + \widetilde{\Gamma} \rangle_F - \beta < -\overline{\text{bias}}_{\hat{C}}(\hat{\beta}) \right) + \mathbb{P}_{\theta,P} \left(\langle A, \widetilde{U} \rangle_F < -z_{1-\alpha/2} \widehat{\text{se}} \right).
\end{aligned}$$

The first term is, by definition, bounded by $\mathbb{P}_{\theta,P}(\|\widetilde{\Gamma}\|_* > \hat{C})$, which converges to zero uniformly over $\theta \in \Theta, P \in \mathcal{P}$ by Assumption 1(i). The second term converges to $\alpha/2$ uniformly over $\theta \in \Theta, P \in \mathcal{P}$ by Assumption 1(ii). Applying a symmetric argument to the probability that the lower endpoint of the CI is greater than β gives the result.

A.3 Proof of Theorem 2

To prove Theorem 2, we first state and prove a series of auxiliary lemmas in Section A.3.1 below. In Section A.3.2, we then prove the final results.

A.3.1 Auxiliary Lemmas

Lemma 8. *Consider*

$$\widehat{\Gamma} = \underset{\{G\,:\,\mathrm{rank}(G)\leq R\}}{\operatorname{argmin}} \|Y - W\cdot\hat{\gamma} - G\|_F^2$$

for some $\hat{\gamma}$*. Suppose that* $\Gamma = \lambda f'$ *for some* $N\times R$ *matrix* λ *and* $T\times R$ *matrix* f*. Then, we have*

(i) $\|\widehat{\Gamma} - \Gamma\|_* \leq 3Rs_1(\hat{U})$*;*

(ii) $\|\widehat{\Gamma} - \Gamma - P_\lambda\hat{U}\|_* \leq 2Rs_1(\hat{U})$*;*

where $\hat{U} \coloneqq U - W\cdot(\hat{\gamma}-\gamma)$*.*

Proof. Let

$$Q(G) \coloneqq \|Y - W\cdot\hat{\gamma} - G\|_F^2,$$

$\hat{Y} \coloneqq Y - W\cdot\hat{\gamma} = \Gamma + \hat{U}$, and $\Gamma^\dagger \coloneqq \Gamma + P_\lambda\hat{U}$, where $M_\lambda = \mathbb{I}_N - P_\lambda$ and $\mathbb{I}_N$ stands for a $N\times N$ identity matrix. Notice that we have $M_\lambda\Gamma^\dagger = 0$, and therefore $\mathrm{rank}\left(\Gamma^\dagger\right)\leq R$.

Next, note that

$$Q(G) = \left\|\hat{Y} - G\right\|_F^2 = \left\|\Gamma + \hat{U} - G\right\|_F^2 = \left\|P_\lambda(\Gamma^\dagger - G) + M_\lambda(\hat{U} - G)\right\|_F^2,$$

so we also have

$$\begin{aligned}
Q(\widehat{\Gamma}) &= \left\|P_\lambda\left(\widehat{\Gamma} - \Gamma^\dagger\right)\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma} - \hat{U}\right)\right\|_F^2 \\
&= \left\|\widehat{\Gamma} - \Gamma^\dagger\right\|_F^2 - \left\|M_\lambda\widehat{\Gamma}\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma} - \hat{U}\right)\right\|_F^2 \\
&= \left\|\widehat{\Gamma} - \Gamma^\dagger\right\|_F^2 + \left\|M_\lambda\hat{U}\right\|_F^2 - 2\mathrm{Tr}\left(\hat{U}'M_\lambda\widehat{\Gamma}\right) \\
&\geq \left\|\widehat{\Gamma} - \Gamma^\dagger\right\|_F^2 + \left\|M_\lambda\hat{U}\right\|_F^2 - 2\,s_1(\hat{U})\,\left\|M_\lambda\widehat{\Gamma}\right\|_*.
\end{aligned}$$

Combining this with

$$Q(\widehat{\Gamma}) \leq Q(\Gamma^\dagger) = \left\|M_\lambda\hat{U}\right\|_F^2,$$

we obtain that

$$\begin{aligned}\left\|\widehat{\Gamma}-\Gamma^{\dagger}\right\|_{*}^{2} &\leq 2R\left\|\widehat{\Gamma}-\Gamma^{\dagger}\right\|_{F}^{2}, \\ &\leq 4R\, s_1(\hat{U})\, \left\|M_{\lambda}\widehat{\Gamma}\right\|_{*} \\ &\leq 4R^2\, s_1(\hat{U})\, s_1(M_{\lambda}\widehat{\Gamma}),\end{aligned}$$

and therefore

$$\left\|\widehat{\Gamma}-\Gamma^{\dagger}\right\|_{*} \leq 2\,R\sqrt{s_1(\hat{U})\, s_1(M_{\lambda}\widehat{\Gamma})}. \tag{19}$$

Next, since $\widehat{\Gamma}$ is given by the R leading principal components of $\hat{Y}$, we know that $\hat{Y}\hat{Y}' \geq (\widehat{\Gamma})(\widehat{\Gamma})'$, that is the difference $\hat{Y}\hat{Y}' - \widehat{\Gamma}\widehat{\Gamma}'$ is positive-definitive, which implies that

$$\begin{aligned}[s_1(M_{\lambda}\widehat{\Gamma})]^2 = \mu_1\left(M_{\lambda}\widehat{\Gamma}\widehat{\Gamma}'M_{\lambda}\right) &= \max_{\{v\in\mathbb{R}^N\,:\,\|v\|=1\}} v'M_{\lambda}\widehat{\Gamma}\widehat{\Gamma}'M_{\lambda}v = \hat{v}'M_{\lambda}\widehat{\Gamma}\widehat{\Gamma}'M_{\lambda}\hat{v} \\ &\leq \hat{v}'M_{\lambda}\hat{Y}\hat{Y}'M_{\lambda}\hat{v} \leq \max_{\{v\in\mathbb{R}^N\,:\,\|v\|=1\}} v'M_{\lambda}\hat{Y}\hat{Y}'M_{\lambda}v = \mu_1(M_{\lambda}\hat{Y}\hat{Y}'M_{\lambda}) = [s_1(M_{\lambda}\hat{Y})]^2 \\ &= [s_1(M_{\lambda}\hat{U})]^2 \leq [s_1(\hat{U})]^2.\end{aligned}$$

We have thus shown that $s_1(M_{\lambda}\widehat{\Gamma}) \leq s_1(U)$, and, combining this with (19), we obtain

$$\left\|\widehat{\Gamma}-\Gamma^{\dagger}\right\|_{*} \leq 2\,R\, s_1(\hat{U}),$$

which proofs the second statement of the lemma. To prove the first, notice that

$$\left\|\widehat{\Gamma}-\Gamma\right\|_{*} \leq \left\|\widehat{\Gamma}-\Gamma^{\dagger}\right\|_{*} + \left\|P_{\lambda}\hat{U}\right\|_{*} \leq 3Rs_1(\hat{U}),$$

where the last inequality uses $\left\|P_{\lambda}\hat{U}\right\|_{*} \leq Rs_1(\hat{U})$. □

Lemma 9. *Under Assumptions 2(i)-(iii),*

(i) $\hat{\gamma}_{\text{LS}} - \gamma = \mathcal{O}_{\Theta,\mathcal{P}}(1/\min\{\sqrt{N},\sqrt{T}\})$*;*

(ii) $\left\|\hat{\Gamma}_{\text{LS}}-\Gamma\right\|_{*} \leq \hat{C}$ *for some* $\hat{C} = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N},\sqrt{T}\})$.

Proof. The first statement follows from the proof of Theorem 4.1 in Moon and Weidner (2015). Assumptions 2 (i)-(iii) are uniform analogues of Assumptions NC, SN, and EX in Moon and Weidner (2015). The derived rate of convergence is immediately uniform over $\theta \in \Theta, P \in \mathcal{P}$ because the proof of Theorem 4.1 in Moon and Weidner (2015) explicitly bounds $\|\hat{\gamma}_{\text{LS}} - \gamma_0\|$.

Next, we combine this result with Lemma 8(i) to obtain

$$\begin{aligned}\left\|\hat{\Gamma}_{\text{LS}} - \Gamma\right\|_* &\leq 3Rs_1(U - W \cdot (\hat{\gamma}_{\text{LS}} - \gamma)) \\ &\leq 3R\left(s_1(U) + s_1(W \cdot (\hat{\gamma}_{\text{LS}} - \gamma))\right) \\ &= \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}), \end{aligned} \tag{20}$$

where the last equality follows from Assumption 2(ii) and $\hat{\gamma}_{\text{LS}} - \gamma = \mathcal{O}_{\Theta,\mathcal{P}}(1/\min\{\sqrt{N}, \sqrt{T}\})$. □

Lemma 10. *Suppose that Assumption 2 holds, and that Assumption 5 holds as stated and with Z_k and X interchanged for each $k = 1, \ldots, K$. Then*

$$\hat{\gamma}_{\text{pre}} - \gamma = \mathcal{O}_{\Theta,\mathcal{P}}\left(1/\min\{N, T\}\right).$$

Proof. The result is immediate from Lemma 9(ii) and Theorem 6, using the fact that b^* is bounded from above and below by a constant times $\max\{\sqrt{N}, \sqrt{T}\}$. □

Lemma 11. *Suppose that Assumption 2 holds, and that Assumption 5 holds as stated and with Z_k and X interchanged for each $k = 1, \ldots, K$. Then, $s_1(\hat{U}_{\text{pre}}) \asymp_{\Theta,\mathcal{P}} \max\{\sqrt{N}, \sqrt{T}\}$ and*

$$s_1(U) \leq s_1(\hat{U}_{\text{pre}})(1 + o_{\Theta,\mathcal{P}}(1)).$$

Proof. First note that, letting $\Delta_\Gamma = \hat{\Gamma}_{\text{pre}} - \Gamma$, we have

$$\left|s_1(\hat{U}_{\text{pre}}) - s_1(U - \Delta_\Gamma)\right| \leq s_1(W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)) = o_{\Theta,P}(\max\{\sqrt{N}, \sqrt{T}\}) \tag{21}$$

the equality follows from Assumption 2(ii) and Lemma 10. Also, notice that

$$s_1(U - \Delta_\Gamma) \leq s_1(U) + s_1(\Delta_\Gamma) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),$$

where the equality follows from Assumption 2(ii) and $s_1(\Delta_\Gamma) \leq \|\Delta_\Gamma\|_* = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$, which can be verified analogously to (20) plugging $\hat{\gamma}_{\text{pre}}$ instead of $\hat{\gamma}_{\text{LS}}$ and using the result of Lemma 10. Combining this with (21), we conclude

$$s_1(\hat{U}_{\text{pre}}) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}). \tag{22}$$

Next, using the fact that $\text{rank}\,(\Delta_\Gamma) \leq 2R$ and the general singular value inequality $s_{i+j-1}(A + B) \leq s_i(A) + s_j(B)$ (see equation (7.3.13) in Problem 7.3.P16 of Horn and Johnson 2013, or alternatively Fan 1951) with $A = U - \Delta_\Gamma$, $B = \Delta_\Gamma$, $i = 1$, $j = 2R + 1$ gives $s_{2R+1}(U) \leq s_1(U - \Delta_\Gamma)$. Thus,

$$s_1(U) \leq s_{2R+1}(U) + o_{\Theta,P}(\max\{\sqrt{N}, \sqrt{T}\}) \leq s_1(\hat{U}_{\text{pre}}) + o_{\Theta,P}(\max\{\sqrt{N}, \sqrt{T}\}), \tag{23}$$

where we apply Assumption 2(iv) for the first inequality and (21) for the second inequality. Notice that this, together with Assumption 2(ii) and (22), implies $s_1(\hat{U}_{\text{pre}}) \asymp_{\Theta,\mathcal{P}} \max\{\sqrt{N}, \sqrt{T}\}$. Together with (23), this completes the proof. □

A.3.2 Proof of Theorem 2

First, notice that Assumption 5 holds according to Lemma 7, so we can apply Lemmas 10 and 11.

We prove the second statement of the theorem. The proof of the first statement is analogous.

Applying the second result of Lemma 8 with $\hat{\gamma} = \gamma_{\text{pre}}$ and $\hat{U} = U - W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)$, we obtain

$$2Rs_1(\hat{U}) \geq \left\|\hat{\Gamma}_{\text{pre}} - \Gamma - P_\lambda \hat{U}\right\|_* \geq \left\|\hat{\Gamma}_{\text{pre}} - \Gamma - P_\lambda U\right\|_* - \left\|P_\lambda W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)\right\|_*.$$

Note that

$$\left\|P_\lambda W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)\right\|_* \leq Rs_1\left(W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)\right) = o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),$$

where the equality follows from Assumption 2(ii) and Lemma 10. Similarly,

$$s_1(\hat{U}) \leq s_1(U) + s_1(W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)) = s_1(U) + o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}).$$

Hence,

$$\left\|\hat{\Gamma}_{\text{pre}} - \Gamma - P_\lambda U\right\|_* \leq 2Rs_1(U) + o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}).$$

Combining this with Lemma 11 proves the second statement of the theorem.

A.4 Proof of Theorem 3

The result follows from the first result of Theorem 2 and Theorem 6, along with Lemma 7 verifying Assumption 5.

A.5 Proof of Theorem 4

The first statement of the theorem follows from Theorem 3 once we verify Assumption 2(v). Notice that Assumption 2(v) is immediate from Assumption 4 and Chebyshev's inequality. Similarly, later in the proof, we will also invoke some of the previously derived results which rely on Assumption 2(v).

To prove the second statement of the theorem, we verify Assumption 1 with $\tilde{\Gamma} = \Gamma + P_\lambda U - \hat{\Gamma}_{\text{pre}}$ and $\tilde{U} = U - P_\lambda U$. Part (i) of Assumption 1 holds by construction, so we just

need to verify part (ii) with this choice of $\tilde{U}$. This will follow if we can show

$$\langle A, U\rangle_F / \widehat{\text{se}} \xrightarrow[\Theta,\mathcal{P}]{d} N(0,1) \tag{24}$$

and

$$\langle A, P_\lambda U\rangle_F / \widehat{\text{se}} = o_{\Theta,\mathcal{P}}(1) \tag{25}$$

where $A = A^*_{b,c}$. Section A.5.1 verifies (24) and Section A.5.2 verifies (25).

A.5.1 Verification of (24)

In this section, we verify that (24) holds under the hypotheses of Theorem 4. Specifically, we show that $\langle A, U\rangle_F / \widehat{\text{se}} \xrightarrow[\Theta,\mathcal{P}]{d} N(0,1)$ for $\widehat{\text{se}}^2 = \sum_{i=1}^N \sum_{t=1}^T A_{it}^2 \hat{U}_{it}^2$ with any sequence of matrices A satisfying $\text{Lind}(A) \leq c_{N,T}$ with $c_{N,T}$ satisfying the condition $c_{N,T} \max\{N,T\} \to 0$ given in the statement of the theorem.

To this end, we first prove a bound on $\|\hat{U} - U\|_F$ (Lemma 12), and then use this to show consistency of the standard error (Lemma 13, using a condition verified in Lemma 14). Lemma 15 completes the proof. We note that the conditions of Lemma 12 hold under the conditions of Theorem 4 by Lemma 9.

Lemma 12. *Let* $\hat{U} = Y - W \cdot \hat{\gamma} - \hat{\Gamma}$, *where*

$$\hat{\Gamma} = \underset{\{G \in \mathbb{R}^{N\times T} \,:\, \text{rank}(G) \leq R\}}{\text{argmin}} \sum_{i=1}^N \sum_{t=1}^T \left(Y_{it} - W_{it}'\hat{\gamma} - G_{it}\right)^2.$$

Suppose that

(i) $\hat{\gamma} - \gamma = \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\min\{\sqrt{N},\sqrt{T}\}}\right)$;

(ii) $\|X\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *and* $\|Z_k\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(\sqrt{NT})$ *for* $k \in \{1,\ldots,K\}$;

(iii) $s_1(X) = \mathcal{O}_{\Theta,\mathcal{P}}\left(\sqrt{NT}\right)$, $s_1(Z_k) = \mathcal{O}_{\Theta,\mathcal{P}}\left(\sqrt{NT}\right)$ *for* $k \in \{1,\ldots,K\}$, *and* $s_1(U) = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N},\sqrt{T}\})$.

Then,

$$\left\|\hat{U} - U\right\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{N,T\}).$$

Proof. Using $\hat{U} = W \cdot (\gamma - \hat{\gamma}) + \Gamma - \hat{\Gamma} + U$,

$$\left\|\hat{U} - U\right\|_F^2 = \|W \cdot (\hat{\gamma} - \gamma)\|_F^2 + \left\|\hat{\Gamma} - \Gamma\right\|_F^2 + 2\langle W \cdot (\hat{\gamma} - \gamma), \hat{\Gamma} - \Gamma\rangle_F.$$

To prove the result, we show that all the terms on the right hand side of the equation above are $\mathcal{O}_{\Theta,\mathcal{P}}(\max\{N,T\})$.

First,

$$\|W \cdot (\hat{\gamma} - \gamma)\|_F \leq \|X\|_F \left|\hat{\beta} - \beta\right| + \sum_{k=1}^{K} \|Z_k\|_F \left|\hat{\delta}_k - \delta_k\right| = \mathcal{O}_{\Theta,\mathcal{P}}\left(\max\{\sqrt{N}, \sqrt{T}\}\right).$$

where we used conditions (i) and (ii).

Second,

$$\left\|\hat{\Gamma} - \Gamma\right\|_F \leq \left\|\hat{\Gamma} - \Gamma\right\|_* = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\})$$

where the equality holds analogously to the previously derived result (20) with $\hat{\Gamma}$ and $\hat{\gamma}$ replacing $\hat{\Gamma}_{\text{LS}}$ and $\hat{\gamma}_{\text{LS}}$ correspondingly.

Third,

$$\left|\langle W \cdot (\hat{\gamma} - \gamma), \hat{\Gamma} - \Gamma\rangle_F\right| \leq \left|\langle X(\hat{\beta} - \beta), \hat{\Gamma} - \Gamma\rangle_F\right| + \sum_{k=1}^{K} \left|\langle Z_k(\hat{\delta}_k - \delta_k), \hat{\Gamma} - \Gamma\rangle_F\right|,$$

where

$$\left|\langle X(\hat{\beta} - \beta), \hat{\Gamma} - \Gamma\rangle_F\right| \leq \|X\|_F \left\|\hat{\Gamma} - \Gamma\right\|_F \left|\hat{\beta} - \beta\right| = \mathcal{O}_{\Theta,\mathcal{P}}\left(\max\{N, T\}\right).$$

Similarly,

$$\sum_{k=1}^{K} \left|\langle Z_k(\hat{\delta}_k - \delta_k), \hat{\Gamma} - \Gamma\rangle_F\right| = \mathcal{O}_{\Theta,\mathcal{P}}\left(\max\{N, T\}\right),$$

which implies

$$\left|\langle W \cdot (\hat{\gamma} - \gamma), \hat{\Gamma} - \Gamma\rangle_F\right| = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{N, T\})$$

and completes the proof. □

Lemma 13. *Suppose that the hypotheses of Lemma 12 are satisfied. Suppose, in addition, that the following conditions hold:*

(i) for any collections of weights $\{\omega_{it}\}_{1\leq i\leq N, 1\leq t\leq T}$, which are non-random conditional on W and Γ, such that $|\omega_{it}| \leq \overline{\omega}$ a.s. for all W and Γ and for all i, t, N, and T, we have

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \omega_{it} U_{it}^2 - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \omega_{it}\mathbb{E}\left[U_{it}^2 | W, \Gamma\right] = \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\sqrt{NT}}\right);$$

(ii) for some $\underline{\sigma}^2 > 0$, $\mathbb{E}\left[U_{it}^2 | W, \Gamma\right] \geq \underline{\sigma}^2$ a.s. for all i, t, N, and T;

(iii) $\text{Lind}(A) \leq c_{N,T}$ *and* $\max\{N, T\}\, c_{N,T} \to 0$.

Then,

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}A_{it}^2\hat{U}_{it}^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}A_{it}^2U_{it}^2}-1=o_{\Theta,\mathcal{P}}(1),$$

where $\hat{U}$ is defined in Lemma 12.

Proof. For simplicity of notation, we use $\sum_{i,t}\equiv\sum_{i=1}^{N}\sum_{t=1}^{T}$ and $\max_{i,t}\equiv\max_{1\leq i\leq N,1\leq t\leq T}$ throughout the proof.

Notice that

$$\begin{aligned}\frac{\sum_{i,t}A_{it}^2\hat{U}_{it}^2}{\sum_{i,t}A_{it}^2U_{it}^2}-1&=\frac{\sum_{i,t}A_{it}^2\left(\hat{U}_{it}^2-U_{it}^2\right)}{\sum_{i,t}A_{it}^2U_{it}^2}\\&=\frac{\sum_{i,t}A_{it}^2\left(\hat{U}_{it}-U_{it}\right)\left(\hat{U}_{it}-U_{it}+2U_{it}\right)}{\sum_{i,t}A_{it}^2U_{it}^2}\\&=\frac{\sum_{i,t}A_{it}^2\left(\hat{U}_{it}-U_{it}\right)^2}{\sum_{i,t}A_{it}^2U_{it}^2}+\frac{2\sum_{i,t}A_{it}^2U_{it}\left(\hat{U}_{it}-U_{it}\right)}{\sum_{i,t}A_{it}^2U_{it}^2}.\end{aligned}\tag{26}$$

The first term in (26) can be bounded as

$$\frac{\sum_{i,t}A_{it}^2\left(\hat{U}_{it}-U_{it}\right)^2}{\sum_{i,t}A_{it}^2U_{it}^2}\leq\frac{\max_{i,t}A_{it}^2\left\|\hat{U}-U\right\|_F^2}{\sum_{i,t}A_{it}^2U_{it}^2},$$

and and the second term in (26) can be bounded as

$$\begin{aligned}\frac{\sum_{i,t}A_{it}^2U_{it}\left(\hat{U}_{it}-U_{it}\right)}{\sum_{i,t}A_{it}^2U_{it}^2}&\leq\frac{\left(\sum_{i,t}A_{it}^4U_{it}^2\right)^{1/2}\left(\sum_{i,t}\left(\hat{U}_{it}-U_{it}\right)^2\right)^{1/2}}{\sum_{i,t}A_{it}^2U_{it}^2}\\&\leq\sqrt{\frac{\max_{i,t}A_{it}^2\left\|\hat{U}-U\right\|_F^2}{\sum_{i,t}A_{it}^2U_{it}^2}},\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality.

Hence, to complete the proof, it is sufficient to demonstrate

$$\frac{\max_{i,t}A_{it}^2\left\|\hat{U}-U\right\|_F^2}{\sum_{i,t}A_{it}^2U_{it}^2}=o_{\Theta,\mathcal{P}}(1).$$

Next, notice that

$$\frac{1}{NT}\sum_{i,t}\frac{A_{it}^2}{\max_{i,t}A_{it}^2}U_{it}^2=\frac{1}{NT}\sum_{i,t}\frac{A_{it}^2}{\max_{i,t}A_{it}^2}\mathbb{E}\left[U_{it}^2|W,\Gamma\right]+\mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\sqrt{NT}}\right)$$

$$\geq \frac{\underline{\sigma}^2}{NT\ \mathrm{Lind}(A)} + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\sqrt{NT}}\right)$$
$$\geq \frac{\underline{\sigma}^2}{NT\ c_{N,T}} + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{1}{\sqrt{NT}}\right) >_{\Theta,\mathcal{P}} 0,$$

where we used condition (i), (ii), and (iii) consequently, and the last inequality (which holds holds wpa1 uniformly) is ensured by condition (iii).

Then

$$\begin{aligned}
\frac{\max_{i,t} A_{it}^2 \left\|\hat{U} - U\right\|_F^2}{\sum_{i,t} A_{it}^2 U_{it}^2} &= \frac{\frac{1}{NT}\left\|\hat{U} - U\right\|_F^2}{\frac{1}{NT}\sum_{i,t} \frac{A_{it}^2}{\max_{i,t} A_{it}^2} U_{it}^2} \\
&\leq \frac{c_{N,T}\left\|\hat{U} - U\right\|_F^2}{\underline{\sigma}^2 + \mathcal{O}_{\Theta,\mathcal{P}}\left(\sqrt{NT}\ c_{N,T}\right)} \\
&\leq \frac{c_{N,T}\left\|\hat{U} - U\right\|_F^2}{\underline{\sigma}^2 + o_{\Theta,\mathcal{P}}(1)} \\
&= o_{\Theta,\mathcal{P}}(1),
\end{aligned}$$

where the last inequality uses condition (iii), and the last equality follows from $\left\|\hat{U} - U\right\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(\max\{N,T\})$ (the result of Lemma 12) and condition (iii). This completes the proof. □

Lemma 14. *Condition (i) of Lemma 13 holds under Assumption 4.*

Proof. The quantity in condition (i) of Lemma 13 has mean zero and variance conditional on W, Γ bounded by

$$\frac{\overline{\omega}^2}{(NT)^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_P[U_i^4 | W, \Gamma] \leq \frac{\overline{\omega}^2/\eta}{NT}.$$

This gives the $\mathcal{O}_{\Theta,\mathcal{P}}(1/\sqrt{NT})$ rate as claimed.

□

Lemma 15. *Suppose that the hypotheses of Lemma 13 are satisfied, and that Assumption 4 holds. Then* $\langle A, U\rangle_F / \widehat{\mathrm{se}} \xrightarrow[\Theta,\mathcal{P}]{d} N(0,1)$.

Proof of Lemma 15. First, we verify

$$\frac{\sum_{i,t} A_{it}^2 U_{it}^2}{\sum_{i,t} A_{it}^2 \sigma_{it}^2} - 1 = o_{\Theta,\mathcal{P}}(1).$$

Here $\sigma_{it}^2 \equiv \sigma_{it}^2(W,\Gamma) = \mathbb{E}[U_{it}^2 | W, \Gamma]$, where we drop the dependence of $\sigma_{it}^2(W,\Gamma)$ on W and Γ

for brevity of notation. Notice that

$$\frac{\sum_{i,t} A_{it}^2 U_{it}^2}{\sum_{i,t} A_{it}^2 \sigma_{it}^2} - 1 = \underbrace{\frac{\sqrt{NT}\max_{i,t} A_{it}^2}{\sum_{i,t} A_{it}^2 \sigma_{it}^2}}_{\underset{\Theta,\mathcal{P}}{\to} 0} \underbrace{\frac{1}{\sqrt{NT}} \sum_{i,t} \frac{A_{it}^2}{\max_{i,t} A_{it}^2}(U_{it}^2 - \sigma_{it}^2)}_{\mathcal{O}_{\Theta,\mathcal{P}}(1)} = o_{\Theta,\mathcal{P}}(1),$$

where the first factor (uniformly) converges to zero due to conditions (ii) and (iii) of Lemma 13, and the second factor is (uniformly) bounded in probability due to condition (i) of Lemma 13. Combining this result with the result of Lemma 13, we obtain

$$\sqrt{\frac{\sum_{it} A_{it}^2 \hat{U}_{it}^2}{\sum_{it} A_{it}^2 \sigma_{it}^2}} - 1 = o_{\Theta,\mathcal{P}}(1). \tag{27}$$

Second, we demonstrate

$$\frac{\sum_{i,t} A_{it} U_{it}}{\sqrt{\sum_{i,t} A_{it}^2 \sigma_{it}^2}} \underset{\Theta,\mathcal{P}}{\overset{d}{\to}} N(0,1). \tag{28}$$

Let $Q_{it} = A_{it}U_{it}/\sqrt{\sum_{i,t} A_{it}^2 \sigma_{it}^2}$ and $S_{N,T} = \sum_{i,t} Q_{it}$. Following the lines of the proof of Lemma F.1 in Armstrong and Kolesár (2018) (and using Assumption 4 and conditions (ii) and (iii) of Lemma 13), we conclude that for all sequences of $W = W_{N,T}$ and $\Gamma = \Gamma_{N,T}$ we have for any fixed $\varepsilon > 0$

$$\sum_{i,t} \mathbb{E}\left[Q_{it}^2 \mathbb{1}\{|Q_{it}| > \varepsilon\} | W, \Gamma\right] \underset{\Theta,\mathcal{P}}{\to} 0. \tag{29}$$

Note that (29) is a uniform version of the Lindeberg condition (applied conditional on W and Γ). Hence, following the lines of the proof of the Lindeberg CLT (see, for example, Theorem 27.2 and its proof in Billingsley, 1995), we establish that, for any fixed $t \in \mathbb{R}$,

$$\left|\mathbb{E}\left[e^{iS_{N,T}t} | W, \Gamma\right] - e^{-t^2/2}\right| \underset{\Theta,\mathcal{P}}{\leqslant} r_{N,T} \quad \text{a.s.}$$

for some $r_{N,T} \downarrow 0$. Hence, we also have

$$\left|\mathbb{E}\left[e^{iS_{N,T}t}\right] - e^{-t^2/2}\right| \underset{\Theta,\mathcal{P}}{\to} 0,$$

which implies $S_{N,T} \underset{\Theta,\mathcal{P}}{\overset{d}{\to}} N(0,1)$ and verifies (28). (27) and (28) together deliver the result. □

A.5.2 Verification of (25)

Using (27), we have

$$\frac{\langle A, P_\lambda U\rangle_F}{\widehat{\text{se}}} = \frac{\langle A, P_\lambda U\rangle_F}{\sqrt{\sum_{i,t} A_{it}^2 \sigma_{it}^2}}(1 + o_{\Theta,\mathcal{P}}(1)),$$

so it is sufficient to show that

$$\frac{\langle A, P_\lambda U\rangle_F}{\sqrt{\sum_{i,t} A_{it}^2 \sigma_{it}^2}} = o_{\Theta,\mathcal{P}}(1). \tag{30}$$

Before we proceed notice that since the low rank representation $\Gamma = \lambda' f$ is arbitrary (even though it is not unique), we can proceed with any compatible mapping $\lambda = \lambda(\Gamma)$ and make λ non-random conditional on Γ.

First, we bound $\text{var}_P(\langle A, P_\lambda U\rangle_F | W, \Gamma)$. Note that

$$\begin{aligned}
\text{var}_P(\langle A, P_\lambda U\rangle_F | W, \Gamma) &= \text{var}_P(\langle P_\lambda A, U\rangle_F | W, \Gamma) \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}_P[(P_\lambda A)_{it}^2 U_{it}^2 | W, \Gamma] \\
&\leq \overline{\sigma}^2 \left\| P_\lambda A \right\|_F^2,
\end{aligned}$$

where $\overline{\sigma}^2$ is a uniform upper bound on $\mathbb{E}_P[U_{it}^2 | W, \Gamma]$. Thus,

$$\text{var}_P\left(\frac{\langle A, P_\lambda U\rangle_F}{\sqrt{\sum_{i,t} A_{it}^2 \sigma_{it}^2}} \middle| W, \Gamma \right) \leq \frac{\overline{\sigma}^2}{\underline{\sigma}^2} \frac{\|P_\lambda A\|_F^2}{\|A\|_F^2}.$$

Note that to complete the proof it is sufficient to show that

$$\frac{\|P_\lambda A\|_F^2}{\|A\|_F^2} \underset{\Theta,\mathcal{P}}{\to} 0$$

because then (30) will follow Chebyshev's inequality.

Since $\text{rank}\,(P_\lambda) \leq R$, we have

$$\frac{\|P_\lambda A\|_F^2}{\|A\|_F^2} \leq \frac{R s_1(P_\lambda A)^2}{\|A\|_F^2} \leq \frac{R s_1(A)^2}{\|A\|_F^2}. \tag{31}$$

Since $A = A^*_{b^*,c}$, according to (18), we also have

$$s_1(A)^2 \leq \mathcal{O}_{\Theta,\mathcal{P}}(\max\{N,T\}/(NT)^2), \tag{32}$$

where we also used $b^{*2} \propto \max\{N,T\}$.

Next, we also want to bound $\|A\|_F^2$ from below. Consider the following minimization problem

$$\min_A \|A\|_F^2 \quad \text{s.t.} \quad \langle A, X\rangle_F = 1.$$

Notice that by the Gauss-Markov theorem the solution is given by $A_{\text{GM}} = X/\|X\|_F^2$. Since the constructed $A = A^*_{b^*,c}$ also needs to satisfy the constraint $\langle A, X\rangle_F = 1$, we also have $\|A\|_F^2 \geq \|A_{\text{GM}}\|_F^2 = 1/\|X\|_F^2$. Combining this with (31) and (32), we obtain

$$\frac{\|P_\lambda A\|_F^2}{\|A\|_F^2} \leq Rs_1(A)\,\|X\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(1/\min\{N,T\}),$$

where we used $\|X\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(NT)$, which follows from Assumptions 3(i)-(iii). This completes the proof.

A.6 Proof of Theorem 5

To prove Theorem 5, we first state and prove an auxiliary lemma in Section A.6.1 below. In Section A.6.2, we then prove the final results.

A.6.1 Auxiliary Lemma

Lemma 16. *Let A, λ, f be $N \times T$, $N \times R$, and $T \times R$ matrices, respectively. Assume that the matrices $P_\lambda A P_f$, λ, and f all have rank equal to R. Then we have*

$$\text{rank}(A) = R \qquad \Longleftrightarrow \qquad M_\lambda A M_f = (M_\lambda A P_f)\,(P_\lambda A P_f)^+\,(P_\lambda A M_f).$$

Notice that $(M_\lambda A P_f)\,(P_\lambda A P_f)^+\,(P_\lambda A M_f)$ in the statement of the lemma can equivalently be written as $M_\lambda A\,(P_\lambda A P_f)^+\,A M_f$, because $(P_\lambda A P_f)^+ = P_f\,(P_\lambda A P_f)^+\,P_\lambda$.

Proof. First, consider the special case

$$\lambda = \begin{pmatrix} \mathbb{I}_R \\ 0_{(N-R)\times R} \end{pmatrix}, \qquad f = \begin{pmatrix} \mathbb{I}_R \\ 0_{(T-R)\times R} \end{pmatrix}. \tag{33}$$

In that case, let $[P_\lambda A P_f]_\# = \lambda' A f$ be the non-zero $R\times R$ block of the $N\times T$ matrix $P_\lambda A P_f$, and analogously, let $[M_\lambda A P_f]_\#$, $[P_\lambda A M_f]_\#$, $[M_\lambda A M_f]_\#$ be the non-zero $(N-R)\times R$, $R\times(T-R)$, $(N-R)\times(T-R)$ blocks of the $N\times T$ matrices $M_\lambda A P_f$, $P_\lambda A M_f$, $M_\lambda A M_f$, respectively. With those definitions we have

$$A = \begin{pmatrix} [P_\lambda A P_f]_\# & [P_\lambda A M_f]_\# \\ [M_\lambda A P_f]_\# & [M_\lambda A M_f]_\# \end{pmatrix}$$

For any $i = 1, \ldots, N - R$ and $t = 1, \ldots, T - R$, we now construct an $(R + 1) \times (R + 1)$ submatrix of this matrix as follows

$$\begin{pmatrix} [P_\lambda A P_f]_\# & [P_\lambda A M_f]_\# \, e_t \\ e_i' \, [M_\lambda A P_f]_\# & e_i' \, [M_\lambda A M_f]_\# \, e_t \end{pmatrix}$$

where e_k refers to the k'th standard basis vector of appropriate dimension. The determinant of this submatrix is given by

$$\det([P_\lambda A P_f]_\#) \left[e_i' \, [M_\lambda A M_f]_\# \, e_t - e_i' \, [M_\lambda A P_f]_\# \, ([P_\lambda A P_f]_\#)^{-1} \, [P_\lambda A M_f]_\# \, e_t \right] = 0.$$

If this determinant is zero for any such $(R + 1) \times (R + 1)$ submatrix, then we can conclude that $\text{rank}(A) \leq R$, which together with our assumption $\text{rank}(P_\lambda A P_f) = R$ implies that $\text{rank}(A) = R$. Conversely, if $\text{rank}(A) = R$, then the determinant of any such $(R+1) \times (R+1)$ submatrix needs to be zero.

The assumption $\text{rank}(P_\lambda A P_f) = R$ guarantees that $\det([P_\lambda A P_f]_\#) \neq 0$. Thus, we have $\text{rank}(A) = R$ if and only if

$$e_i' \left[[M_\lambda A M_f]_\# - [M_\lambda A P_f]_\# \, ([P_\lambda A P_f]_\#)^{-1} \, [P_\lambda A M_f]_\# \right] \, e_t = 0.$$

for all $i = 1, \ldots, N - R$ and $t = 1, \ldots, T - R$. This can equivalently be written as

$$[M_\lambda A M_f]_\# = [M_\lambda A P_f]_\# \, ([P_\lambda A P_f]_\#)^{-1} \, [P_\lambda A M_f]_\#,$$

or equivalently

$$M_\lambda A M_f = M_\lambda A P_f \, (P_\lambda A P_f)^+ \, P_\lambda A M_f.$$

We have thus shown the lemma for the special case that λ and f are of the form (33).

For any other λ and f that have full rank R (as assumed in the lemma) we can choose an orthogonal $N \times N$ matrix O_1 and an orthogonal $T \times T$ matrix O_2 such that $O_1 \lambda = \binom{\mathbb{I}_R}{0_{(N-R)\times R}}$ and $O_2 f = \binom{\mathbb{I}_R}{0_{(T-R)\times R}}$. By applying the result already proven to the transformed data $O_1 A O_2'$, $O_1 \lambda$, $O_2 f$ we then obtain the result of the lemma more generally, because the statement of the lemma is invariant under such orthogonal transformations. □

A.6.2 Proof of Theorem 5

First, let $\hat{U} \coloneqq U - W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)$ and notice that we have

$$s_1(\hat{U}) = s_1(U) + o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}) = s_1(U)(1 + o_{\Theta,\mathcal{P}}(1)) \asymp_{\Theta,\mathcal{P}} \max\{\sqrt{N}, \sqrt{T}\}, \quad (34)$$

where we used Assumption 2(ii) and Lemma 10. Together with the hypothesis of the theorem it implies that

$$\frac{s_1(\hat{U})}{s_R(\Gamma)} = o_{\Theta,\mathcal{P}}(1). \tag{35}$$

This condition implies that $s_R(\Gamma) \neq 0$, which together with our assumption $\text{rank}(\Gamma) \leq R$ implies that $\text{rank}(\Gamma) = R$. Since $\Gamma = \lambda f'$, this also implies that the rank of both λ and f is equal to R as well.

Next, define

$$\Gamma^\dagger := \Gamma + M_\lambda \hat{U} P_f + P_\lambda \hat{U} M_f + M_\lambda \hat{U} \Gamma^+ \hat{U} M_f.$$

Using Lemma 16 and that $\Gamma = P_\lambda \Gamma P_f$, we conclude that

$$\text{rank}\left(\Gamma^\dagger\right) = R.$$

Our first goal in the following is to derive a bound on $\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger\|_*$. Denote

$$Q(G) \coloneqq \|Y - W \cdot \hat{\gamma}_{\text{pre}} - G\|_F^2 = \left\|\hat{U} + \Gamma - G\right\|_F^2.$$

Then

$$\begin{aligned}
Q(\Gamma^\dagger) &= \left\|\hat{U} + \Gamma - \Gamma^\dagger\right\|_F^2 \\
&= \left\|\hat{U} - M_\lambda \hat{U} P_f - P_\lambda \hat{U} M_f - M_\lambda \hat{U} \Gamma^+ U M_f\right\|_F^2 \\
&= \left\|P_\lambda \hat{U} P_f + M_\lambda \hat{U} M_f - M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_F^2 \\
&= \left\|M_\lambda \hat{U} M_f - M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_F^2 + \left\|P_\lambda \hat{U} P_f\right\|_F^2 \\
&= \left\|M_\lambda \hat{U} M_f\right\|_F^2 - 2\text{Tr}\left(M_f \hat{U}' M_\lambda \hat{U} \Gamma^+ \hat{U} M_\lambda\right) + \left\|M_\lambda \hat{U} \Gamma^+ U M_f\right\|_F^2 + \left\|P_\lambda \hat{U} P_f\right\|_F^2,
\end{aligned}$$

where in the last step we used that $\|A + B\|_F^2 = \text{Tr}((A+B)'(A+B)) = \|A\|_F^2 + 2\text{Tr}(A'B) + \|B\|_F^2$. Furthermore, using

$$\left\|M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_F \leq R\, s_1(M_\lambda \hat{U} \Gamma^+ \hat{U} M_f) \leq R\,[s_1(\hat{U})]^2 s_1(\Gamma^+) = \frac{R\,[s_1(\hat{U})]^2}{s_R(\Gamma)}, \tag{36}$$

and[11]

$$\left|\text{Tr}\left(M_f \hat{U}' M_\lambda \hat{U} \Gamma^+ \hat{U}\right)\right| \leq R\, s_1(M_f \hat{U}' M_\lambda \hat{U} \Gamma^+ \hat{U})$$

[11] Here, we applied the general inequality $|\text{Tr}(A)| \leq \text{rank}(A)\, s_1(A)$, which is an immediate consequence of von Neumann's trace inequality, with $A = M_f \hat{U}' M_\lambda \hat{U} \Gamma^+ \hat{U}$, where $\text{rank}(A) \leq \text{rank}(\Gamma^+) = R$.

$$\leq R\,[s_1(\hat{U})]^3\,s_1(\Gamma^+) = \frac{R\,[s_1(\hat{U})]^3}{s_R(\Gamma)}, \tag{37}$$

we thus obtain

$$\begin{aligned} Q(\Gamma^\dagger) &\leq \left\|M_\lambda \hat{U} M_f\right\|_F^2 + \left\|P_\lambda \hat{U} P_f\right\|_F^2 + \mathcal{O}_{\Theta,\mathcal{P}}\left\{\frac{[s_1(\hat{U})]^4}{[s_R(\Gamma)]^2} + \frac{[s_1(\hat{U})]^3}{s_R(\Gamma)}\right\} \\ &= \left\|M_\lambda \hat{U} M_f\right\|_F^2 + \left\|P_\lambda \hat{U} P_f\right\|_F^2 + o_{\Theta,\mathcal{P}}\left[(s_1(\hat{U}))^2\right]. \end{aligned} \tag{38}$$

where in the last step we used (35).

Next, for any $N \times T$ matrix G we have

$$\begin{aligned} Q(G) &= \left\|G - \Gamma - \hat{U}\right\|_F^2 \\ &= \left\|P_\lambda\left(G - \Gamma - \hat{U}\right)P_f\right\|_F^2 + \left\|P_\lambda\left(G - \hat{U}\right)M_f\right\|_F^2 \\ &\quad + \left\|M_\lambda\left(G - \hat{U}\right)P_f\right\|_F^2 + \left\|M_\lambda\left(G - \hat{U}\right)M_f\right\|_F^2, \end{aligned}$$

where we used that $M_\lambda \Gamma = 0$ and $\Gamma M_f = 0$. In particular, we have

$$\begin{aligned} Q(\widehat{\Gamma}_{\text{pre}}) &= \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \Gamma - \hat{U}\right)P_f\right\|_F^2 + \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 \\ &\quad + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)P_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2. \end{aligned}$$

We have

$$\begin{aligned} \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 &= \left\|M_\lambda \hat{U} M_f - M_\lambda \widehat{\Gamma}_{\text{pre}} M_f\right\|_F^2 \\ &\geq \underset{\{G\,:\,\text{rank}(G)\leq R\}}{\text{argmin}} \left\|M_\lambda \hat{U} M_f - G\right\|_F^2 \\ &= \left\|M_\lambda \hat{U} M_f\right\|_F^2 - \sum_{r=1}^{R}[s_r(M_\lambda \hat{U} M_f)]^2 \\ &\geq \left\|M_\lambda \hat{U} M_f\right\|_F^2 - R[s_1(\hat{U})]^2, \end{aligned}$$

where for the first inequality we used that $M_\lambda \widehat{\Gamma}_{\text{pre}} M_f$ is a matrix of rank at most R, that is, minimizing over all such matrices can only make the expression smaller, and in the next step we used that the solution to this least squares minimization problem over $\{G : \text{rank}(G) \leq R\}$ is given by the principal components of $M_\lambda \hat{U} M_f$. We thus obtain that

$$\begin{aligned} Q(\widehat{\Gamma}_{\text{pre}}) &\geq \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \Gamma\right)P_f - P_\lambda \hat{U} P_f\right\|_F^2 + \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 \\ &\quad + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)P_f\right\|_F^2 + \left\|M_\lambda \hat{U} M_f\right\|_F^2 - R[s_r(\hat{U})]^2. \end{aligned} \tag{39}$$

Since $\widehat{\Gamma}_{\text{pre}}$ minimizes $Q(G)$ over all rank $\leq R$ matrices, and rank $\left(\Gamma^{\dagger}\right) = R$, we know that $Q(\widehat{\Gamma}_{\text{pre}}) \leq Q(\Gamma^{\dagger})$. Combining (38) and (39) we thus obtain that

$$\begin{aligned}\left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\Gamma\right)P_f - P_\lambda\hat{U}P_f\right\|_F^2 + \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\hat{U}\right)M_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\hat{U}\right)P_f\right\|_F^2 \\ \leq R[s_1(\hat{U})]^2 + o_{\Theta,\mathcal{P}}\left[(s_1(\hat{U}))^2\right] + \left\|P_\lambda\hat{U}P_f\right\|_F^2 \\ = \mathcal{O}_{\Theta,\mathcal{P}}([s_1(\hat{U})]^2).\end{aligned}$$

Since all three terms on the lhs are positive, this implies separately for each of them

$$\begin{aligned}\left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\Gamma\right)P_f\right\|_F &\leq \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\Gamma\right)P_f - P_\lambda\hat{U}P_f\right\|_F + \left\|P_\lambda\hat{U}P_f\right\|_F \\ &= \mathcal{O}_{\Theta,\mathcal{P}}(s_1(\hat{U})), \\ \left\|P_\lambda\widehat{\Gamma}_{\text{pre}}M_f\right\|_F &= \mathcal{O}_{\Theta,\mathcal{P}}(s_1(\hat{U})), \\ \left\|M_\lambda\widehat{\Gamma}_{\text{pre}}P_f\right\|_F &= \mathcal{O}_{\Theta,\mathcal{P}}(s_1(\hat{U})).\end{aligned} \tag{40}$$

The first result in the last display implies

$$s_1\left(P_\lambda\widehat{\Gamma}_{\text{pre}}P_f - \Gamma\right) \leq \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}}-\Gamma\right)P_f\right\|_F = \mathcal{O}_{\Theta,\mathcal{P}}(s_1(\hat{U})),$$

and by Weyl inequality for singular values we thus find that

$$\begin{aligned}s_R\left(P_\lambda\widehat{\Gamma}_{\text{pre}}P_f\right) &\geq s_R\left(\Gamma\right) - s_1\left(P_\lambda\widehat{\Gamma}_{\text{pre}}P_f - \Gamma\right) \\ &\geq s_R\left(\Gamma\right) - \mathcal{O}_{\Theta,\mathcal{P}}(s_1(\hat{U})) \\ &\geq 0, \qquad \text{wpa1},\end{aligned} \tag{41}$$

where in the last step we used (35) again. We thus have, wpa1, that $P_\lambda\widehat{\Gamma}_{\text{pre}}P_f$ has rank equal to R, and by applying Lemma 16 we thus find

$$M_\lambda\widehat{\Gamma}_{\text{pre}}M_f = (M_\lambda\widehat{\Gamma}_{\text{pre}}P_f)\left(P_\lambda\widehat{\Gamma}_{\text{pre}}P_f\right)^{+}(P_\lambda\widehat{\Gamma}_{\text{pre}}M_f).$$

Together with the bounds in (40) and (41) and (35) this implies that

$$\begin{aligned}\left\|M_\lambda\widehat{\Gamma}_{\text{pre}}M_f\right\|_F &\leq R\, s_1(M_\lambda\widehat{\Gamma}_{\text{pre}}M_f) \\ &\leq \frac{R\, s_1(M_\lambda\widehat{\Gamma}_{\text{pre}}P_f)\, s_1(P_\lambda\widehat{\Gamma}_{\text{pre}}M_f)}{s_R(P_\lambda\widehat{\Gamma}_{\text{pre}}P_f)} \\ &= \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^2}{s_R(\Gamma)}\right).\end{aligned} \tag{42}$$

We can now improve on the lower bound in $Q(\widehat{\Gamma}_{\text{pre}})$ in (39). As before, we have

$$Q(\widehat{\Gamma}_{\text{pre}}) = \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \Gamma - \hat{U}\right)P_f\right\|_F^2 + \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)P_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2.$$

Similarly, using the definition of $\Gamma^\dagger$ we obtain

$$\left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger - P_\lambda\hat{U}P_f\right\|_F^2 = \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \Gamma - \hat{U}\right)P_f\right\|_F^2 + \left\|P_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)P_f\right\|_F^2 + \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - M_\lambda\hat{U}\Gamma^+\hat{U}M_f\right)M_f\right\|_F^2.$$

Taking the difference of the equalities in the last two displays we obtain

$$\begin{aligned}
Q(\widehat{\Gamma}_{\text{pre}}) - &\left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger - P_\lambda\hat{U}P_f\right\|_F^2 \\
&= \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 - \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - M_\lambda\hat{U}\Gamma^\dagger\hat{U}M_f\right)M_f\right\|_F^2 \\
&= \left\|M_\lambda\left(\widehat{\Gamma}_{\text{pre}} - \hat{U}\right)M_f\right\|_F^2 + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^4}{[s_R(\Gamma)]^2}\right) \\
&= \left\|M_\lambda\hat{U}M_f\right\|_F^2 - 2\text{Tr}(\hat{U}'M_\lambda\widehat{\Gamma}_{\text{pre}}M_f) + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^4}{[s_R(\Gamma)]^2}\right) \\
&\geq \left\|M_\lambda\hat{U}M_f\right\|_F^2 - 2Rs_1(\hat{U})s_1(M_\lambda\widehat{\Gamma}_{\text{pre}}M_f) + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^4}{[s_R(\Gamma)]^2}\right) \\
&= \left\|M_\lambda\hat{U}M_f\right\|_F^2 + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^3}{s_R(\Gamma)}\right) + \mathcal{O}_{\Theta,\mathcal{P}}\left(\frac{[s_1(\hat{U})]^4}{[s_R(\Gamma)]^2}\right) \\
&= \left\|M_\lambda\hat{U}M_f\right\|_F^2 + o_{\Theta,\mathcal{P}}\left(\left[s_1(\hat{U})\right]^2\right),
\end{aligned}$$

where we used (42) and (36), the trace term was bounded analogously to (37), and in the final step we also used (35). Again using that $Q(\widehat{\Gamma}_{\text{pre}}) \leq Q(\Gamma^\dagger)$ and (38) we now obtain

$$\left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger - P_\lambda\hat{U}P_f\right\|_F^2 \leq o_{\Theta,\mathcal{P}}\left(\left[s_1(\hat{U})\right]^2\right) + \left\|P_\lambda\hat{U}P_f\right\|_F^2. \tag{43}$$

Our next step is to bound $\|P_\lambda\hat{U}P_f\|_F$. Note that, using Assumption 4,

$$\begin{aligned}
\mathbb{E}\left[\|P_\lambda U P_f\|_F^2 |\Gamma\right] &= \mathbb{E}\left[\text{vec}(P_\lambda U P_f)'\text{vec}(P_\lambda U P_f)|\Gamma\right] \\
&= \mathbb{E}\left[\text{Tr}\left[\text{vec}(P_\lambda U P_f)\text{vec}(P_\lambda U P_f)'\right]|\Gamma\right] \\
&= \text{Tr}\Big[(P_f \otimes P_\lambda)\underbrace{\mathbb{E}\left[\text{vec}(U)\text{vec}(U)'|\Gamma\right]}_{\leq \overline{\sigma}^2\mathbb{I}_{NT}}(P_f \otimes P_\lambda)\Big] \\
&\leq \overline{\sigma}^2\text{Tr}\left[P_f \otimes P_\lambda\right]
\end{aligned}$$

$$= \overline{\sigma}^2 \mathrm{Tr}(P_f)\mathrm{Tr}(P_\lambda)$$
$$= \overline{\sigma}^2 R^2$$

for some $\overline{\sigma}^2 > 0$ for all Γ. Hence, we also have

$$\mathbb{E}\left[\|P_\lambda U P_f\|_F^2\right] \leq \overline{\sigma}^2 R^2,$$

and, using Markov's inequality we conclude $\|P_\lambda U P_f\|_F^2 = \mathcal{O}_{\Theta,\mathcal{P}}(1)$. Finally, using Assumption 3 and Lemma 10, we obtain

$$\|P_\lambda \hat{U} P_f\|_F \leq \|P_\lambda U P_f\|_F + \|W \cdot (\hat{\gamma}_{\text{pre}} - \gamma)\|_F = o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}).$$

Combining this result with (43), we obtain

$$\begin{aligned}\left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger\right\|_F &\leq \left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger - P_\lambda \hat{U} P_f\right\|_F + \left\|P_\lambda \hat{U} P_f\right\| \\ &= o_{\Theta,\mathcal{P}}\left(s_1(U)\right),\end{aligned}$$

where the last equality also uses (34).

Since $\|A\|_* \leq \sqrt{\mathrm{rank}(A)}\,\|A\|_F$, this also implies that

$$\left\|\widehat{\Gamma}_{\text{pre}} - \Gamma^\dagger\right\|_* = o_{\Theta,\mathcal{P}}\left(s_1(U)\right) = o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}).$$

We have thus shown that

$$\left\|\widetilde{\Gamma} - M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_* = o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),$$

which implies

$$\begin{aligned}\left\|\widetilde{\Gamma}\right\|_* &= \left\|M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_* + o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}) \\ &= \sqrt{R}\left\|M_\lambda \hat{U} \Gamma^+ \hat{U} M_f\right\|_F + o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}) \\ &= o_{\Theta,\mathcal{P}}(\max\{\sqrt{N}, \sqrt{T}\}),\end{aligned}$$

where in the last step we used (36), (35), and (34) again. This completes the proof.

A.7 Verification of Assumption 2(i)

In this section, we verify that, in the absence of Z, Assumption 3(i) and (ii) imply Assumption 2(i). Thus, our goal in this section is to prove that

$$\frac{1}{NT} \sum_{r=2R+1}^{\min\{N,T\}} s_r^2(X) \geq \underline{s}^2 > 0, \tag{44}$$

with probability approaching one.

By the variational principle for eigenvalues, we have

$$\sum_{r=1}^{2R} s_r^2(X) = \sum_{r=1}^{2R} \lambda_r(X'X) = \max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(X'X\, P_Q),$$

where P_Q is the projection matrix onto the subspace spanned by the columns of Q. Using this and $\mathrm{Tr}(X'X) = \sum_{r=1}^{\min\{N,T\}} s_r^2(X)$, we obtain

$$\frac{1}{NT}\sum_{r=2R+1}^{\min\{N,T\}} s_r^2(X) = \frac{1}{NT}\mathrm{Tr}(X'X) - \frac{1}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(X'X\, P_Q).$$

Using $X = H + V$, where H and V are as defined in Assumption 3, and substituting into the above, we obtain:

$$\begin{aligned}\frac{1}{NT}\sum_{r=2R+1}^{\min\{N,T\}} s_r^2(X) \geq\ & \frac{1}{NT}\mathrm{Tr}(H'H) - \frac{1}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(H\, P_Q\, H') \\ & + \frac{1}{NT}\mathrm{Tr}(V'V) - \frac{1}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(V\, P_Q\, V') \\ & + \frac{2\,\mathrm{Tr}(V'H)}{NT} - \frac{2}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(V\, P_Q\, H').\end{aligned}$$

For the three lines on the right-hand side we have

$$\begin{aligned}&\frac{1}{NT}\mathrm{Tr}(H'H) - \frac{1}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(H\, P_Q\, H') = \frac{1}{NT}\sum_{r=2R+1}^{\min\{N,T\}} s_r^2(H) \geq 0, \\ &\frac{1}{NT}\mathrm{Tr}(V'V) - \frac{1}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(V\, P_Q\, V') \geq \frac{1}{NT}\|V\|_F^2 - \frac{2R\,[s_1(V)]^2}{NT}, \\ &\frac{2\,\mathrm{Tr}(V'H)}{NT} - \frac{2}{NT}\max_{Q \in \mathbb{R}^{T\times 2R}} \mathrm{Tr}(V\, P_Q\, H') \geq \frac{2\,\mathrm{Tr}(V'H)}{NT} - \frac{4R\, s_1(H)s_1(V)}{NT},\end{aligned}$$

and therefore

$$\begin{aligned}\frac{1}{NT}\sum_{r=2R+1}^{\min\{N,T\}} s_r^2(X) &\geq \frac{1}{NT}\|V\|_F^2 + \frac{2\,\mathrm{Tr}(V'H)}{NT} - \frac{4R\, s_1(H)s_1(V)}{NT} - \frac{2R\,[s_1(V)]^2}{NT} \\ &= \frac{1}{NT}\|V\|_F^2 + o_{\Theta,\mathcal{P}}(1) \\ &\asymp_{\Theta,\mathcal{P}} 1,\end{aligned}$$

we in the last two steps we used Assumption 3(i) and (ii), as well as $s_1(H) \leq \|H\|_F$. This confirms that (44) holds.

More generally, if in addition, each of the other regressors Z_k, $k = 1, \ldots, K$, can be

decomposed as $V_k + H_k$ such that Assumption 3(i) and (ii) holds equally for V_k and H_k, and the $K+1$ vector $V_{\text{vec},it}$ that combines V_{it} and $V_{k,it}$ satisfies the standard non-collinearity condition

$$\operatorname*{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{\text{vec},it}V'_{\text{vec},it} > 0,$$

then by the same arguments as for $K = 0$ above, the Assumption 2(i) holds for general K.

B Computational details

The optimal weights A_b^* given in Definition 2.2 can be computed directly using convex programming. Alternatively, we can obtain these weights from a nuclear norm regularized "partialling out" regression of X on Z and a matrix of individual effects. This follows by applying a result from Armstrong, Kolesár and Kwon (2020) to our setting, as we now describe. We first consider the general case with covariates (Section B.1), and then obtain a further simplification by specializing to the case with no additional covariates Z.

B.1 General case

The weights A_b^* minimize $\left(\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)\right)^2 + \sigma^2\|A\|_F^2$ when $\hat{C}/\sigma = b$. Equivalently, we can minimize $\sigma^2\|A\|_F^2$ subject to a bound on $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$:

$$\min_A \sigma^2\|A\|_F^2 \quad \text{s.t.} \quad \overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A) \le B. \tag{45}$$

We can then vary the bound B to optimize any increasing function of the variance $\sigma^2\|A\|_F^2$ and worst-case bias $\overline{\text{bias}}_{\hat{C}}(\hat{\beta}_A)$.

Let Π_μ^*, ψ_μ^* solve the nuclear norm regularized regression

$$\min_{\Pi,\psi} \|X - Z\cdot\psi - \Pi\|_F^2/2 + \mu\|\Pi\|_* \tag{46}$$

where μ indexes the penalty on the nuclear norm. Let

$$\Omega_\mu^* = X - Z\cdot\psi_\mu^* - \Pi_\mu^* \tag{47}$$

denote the matrix of residuals from this regression. Let

$$\hat{\beta}_{\tilde{A}_\mu^*} = \langle \tilde{A}_\mu^*, \widetilde{Y}\rangle_F = \frac{\langle \Omega_\mu^*, \widetilde{Y}\rangle_F}{\langle \Omega_\mu^*, X\rangle_F} \quad \text{where} \quad \tilde{A}_\mu^* = \frac{\Omega_\mu^*}{\langle \Omega_\mu^*, X\rangle_F} \tag{48}$$

and let

$$\overline{B}_\mu = \frac{1}{\|\Pi_\mu^*\|_*}\frac{\langle \Omega_\mu^*, \Pi_\mu^*\rangle_F}{\langle \Omega_\mu^*, X\rangle_F} \quad \text{and} \quad V_\mu = \sigma^2\frac{\|\Omega_\mu^*\|_F^2}{\langle \Omega_\mu^*, X\rangle_F^2}. \tag{49}$$

The following theorem follows immediately from applying Theorem 2.1 in Armstrong, Kolesár and Kwon (2020) to our setup (in applying the formulas from this paper, we use the fact that $\langle \Omega_\mu^*, Z \cdot \psi_\mu^* \rangle_F = 0$ by the first order conditions for ψ, since ψ is unconstrained).

Theorem 17. *Let Π_μ^*, ψ_μ^* be a solution to (46) and let Ω_μ^* be the matrix of residuals in (47), and suppose $\|\Omega_\mu^*\| > 0$. Then $\tilde{A}_\mu^*$ and the corresponding estimator $\hat{\beta}_{\tilde{A}_\mu^*}$ given in (48) solve (45) for $B = \hat{C}\overline{B}_\mu$, with minimized value V_μ, where $\overline{B}_\mu$ and V_μ are given in (49).*

Thus, to compute the MSE optimizing weights A_b^*, it suffices to compute the weights $\tilde{A}_\mu^*$ for each $\mu > 0$, and then minimize $\hat{C}^2\overline{B}_\mu^2 + V_\mu$ over the one-dimensional parameter μ. We can also minimize other criteria, as in Remark 2.3 by choosing μ to minimize other functions of worst-case bias $\hat{C}\overline{B}_\mu$ and variance V_μ.

B.2 No additional covariates

In the case where there are no additional covariates, the nuclear norm regularized "partialling out" regression (46) reduces to

$$\min_{\Pi} \|X - \Pi\|_F^2/2 + \mu\|\Pi\|_*. \tag{50}$$

The solution Π_μ^* can then be computed using soft thresholding on the singular values of X. We describe the solution here, and refer to Moon and Weidner (2018, Lemma S.1) for a detailed derivation.

Let the singular value decomposition of X be given by $X = V_X S_X W_X'$ where V_X is an $N \times N$ orthogonal matrix (i.e. $V_X'V_X = I_N$), W_X is a $T \times T$ orthogonal matrix (i.e. $W_X'W_X = I_T$) and S_X is a $N \times T$ rectangular diagonal matrix, with j-th diagonal element given by the j-th singular value $s_j(X)$ of X. Let $\tilde{S}_X(\mu)$ be the $N \times T$ diagonal matrix with j-th diagonal element given by $\max\{s_j(X) - \mu, 0\}$ (i.e. we perform soft thresholding on the j-th singular value).

Then the solution Π_μ^* to (50) and residuals $\Omega_\mu^* = X - \Pi_\mu^*$ are given by

$$\Pi_\mu^* = V_X \tilde{S}_X(\mu) W_X', \quad \Omega_\mu^* = V_X(S_X - \tilde{S}_X(\mu))W_X',$$

Note that $S_X - \tilde{S}_X(\mu)$ is a $N \times T$ diagonal matrix with j-th diagonal element given by $\min\{s_j(X), \mu\}$. Thus, the weights $\tilde{A}_\mu^* = \Omega_\mu^*/\langle \Omega_\mu^*, X\rangle_F$ used in the estimator $\hat{\beta} = \langle \tilde{A}_\mu^*, \tilde{Y}\rangle_F$ given in (48) can be obtained by replacing the singular values $s_j(X)$ that are larger than μ with the constant μ, and then dividing by the constant $\langle \Omega_\mu^*, X\rangle_F = \langle S_X - \tilde{S}_X(\mu), S_X\rangle_F = \sum_{j=1}^{\min\{N,T\}} \min\{s_j(X), \mu\} s_j(X)$.

C Additional Numerical Results

C.1 Additional simulation results for Section 5.1

In this section, we provide additional simulation results for the numerical experiment described in Section 5.1. For brevity, here we only report results for $R = 1$. Tables 6 and 7 report the same statistics as Table 1 in the main text but for smaller and bigger samples sizes $N = 50$ and $N = 300$, respectively. The results are similar to the ones presented in Section 5.1. We find that, when there is a weak factor, our method effectively reduces the bias and improves estimation precision in smaller sample sizes too, i.e., also for $N = 50$. The gains from using our method are even more considerable for $N = 300$, which is consistent with our estimator having a faster rate of convergence than the LS estimator.

Table 6: Simulation results for the experiment in Section 5.1, $N = 50$, $R = 1$

	LS						Debiased					
κ	bias	std	rmse	size	length	length*	bias	std	rmse	size	length	length*
						$T = 20$						
0.00	-0.0006	0.0242	0.0242	7.4	0.086	0.348	-0.0007	0.0300	0.0300	0.0	0.364	0.184
0.05	0.0233	0.0249	0.0340	22.5	0.087	0.349	0.0088	0.0302	0.0314	0.0	0.364	0.184
0.10	0.0466	0.0268	0.0538	56.5	0.087	0.351	0.0177	0.0310	0.0357	0.0	0.365	0.185
0.15	0.0683	0.0309	0.0750	78.5	0.089	0.357	0.0252	0.0327	0.0413	0.0	0.367	0.186
0.20	0.0847	0.0401	0.0937	83.8	0.091	0.368	0.0293	0.0361	0.0465	0.0	0.370	0.186
0.25	0.0879	0.0555	0.1040	76.0	0.097	0.390	0.0281	0.0406	0.0494	0.0	0.372	0.187
0.50	0.0115	0.0398	0.0414	12.1	0.122	0.493	0.0025	0.0359	0.0360	0.0	0.380	0.189
1.00	0.0004	0.0330	0.0330	6.1	0.124	0.499	-0.0006	0.0346	0.0346	0.0	0.381	0.189
						$T = 50$						
0.00	0.0003	0.0147	0.0147	6.0	0.055	0.279	-0.0001	0.0211	0.0211	0.0	0.230	0.118
0.05	0.0247	0.0152	0.0290	44.0	0.055	0.280	0.0070	0.0212	0.0223	0.0	0.230	0.118
0.10	0.0487	0.0169	0.0515	87.9	0.055	0.282	0.0134	0.0218	0.0256	0.0	0.231	0.118
0.15	0.0703	0.0214	0.0734	95.9	0.056	0.287	0.0173	0.0238	0.0294	0.0	0.232	0.119
0.20	0.0784	0.0368	0.0866	86.3	0.060	0.306	0.0158	0.0267	0.0310	0.0	0.234	0.119
0.25	0.0583	0.0510	0.0774	59.2	0.068	0.344	0.0097	0.0273	0.0290	0.0	0.236	0.119
0.50	0.0035	0.0207	0.0210	6.5	0.078	0.398	0.0003	0.0236	0.0236	0.0	0.237	0.120
1.00	0.0003	0.0201	0.0201	5.0	0.078	0.399	-0.0003	0.0234	0.0234	0.0	0.237	0.120
						$T = 100$						
0.00	0.0002	0.0105	0.0105	6.0	0.039	0.229	-0.0001	0.0137	0.0137	0.0	0.173	0.080
0.05	0.0247	0.0109	0.0270	68.4	0.039	0.229	0.0064	0.0138	0.0152	0.0	0.173	0.080
0.10	0.0487	0.0124	0.0502	98.1	0.039	0.231	0.0120	0.0143	0.0187	0.0	0.174	0.080
0.15	0.0684	0.0188	0.0709	97.3	0.040	0.238	0.0134	0.0165	0.0212	0.0	0.175	0.080
0.20	0.0577	0.0390	0.0697	72.5	0.046	0.271	0.0082	0.0179	0.0197	0.0	0.177	0.081
0.25	0.0225	0.0300	0.0375	33.9	0.053	0.310	0.0031	0.0164	0.0167	0.0	0.177	0.081
0.50	0.0016	0.0145	0.0146	5.6	0.055	0.325	0.0001	0.0153	0.0153	0.0	0.177	0.081
1.00	0.0001	0.0143	0.0143	5.1	0.055	0.326	-0.0001	0.0152	0.0152	0.0	0.177	0.081
						$T = 300$						
0.00	-0.0001	0.0059	0.0059	5.8	0.023	0.170	-0.0002	0.0077	0.0077	0.0	0.127	0.049
0.05	0.0245	0.0066	0.0254	96.7	0.023	0.171	0.0060	0.0078	0.0098	0.0	0.127	0.049
0.10	0.0481	0.0088	0.0489	99.8	0.023	0.173	0.0100	0.0089	0.0134	0.0	0.128	0.049
0.15	0.0482	0.0271	0.0553	83.6	0.026	0.196	0.0059	0.0103	0.0119	0.0	0.129	0.049
0.20	0.0117	0.0143	0.0185	33.3	0.031	0.234	0.0015	0.0090	0.0091	0.0	0.129	0.049
0.25	0.0047	0.0093	0.0104	12.6	0.032	0.239	0.0006	0.0087	0.0087	0.0	0.129	0.049
0.50	0.0004	0.0084	0.0085	5.6	0.032	0.241	-0.0001	0.0086	0.0086	0.0	0.129	0.049
1.00	-0.0001	0.0084	0.0084	5.5	0.032	0.241	-0.0002	0.0086	0.0086	0.0	0.129	0.049

$\text{Lind}(A) \in \{0.0109, 0.0049, 0.0028, 0.0011\}$ for $T \in \{20, 50, 100, 300\}$. The results are based on 5,000 simulations.

Table 7: Simulation results for the experiment in Section 5.1, $N = 300$, $R = 1$

	LS						Debiased					
κ	bias	std	rmse	size	length	length*	bias	std	rmse	size	length	length*
						$T = 20$						
0.00	0.0001	0.0096	0.0096	6.4	0.036	0.219	0.0001	0.0115	0.0115	0.0	0.246	0.088
0.05	0.0242	0.0106	0.0264	72.7	0.036	0.220	0.0091	0.0117	0.0148	0.0	0.246	0.088
0.10	0.0474	0.0136	0.0493	96.6	0.036	0.223	0.0169	0.0127	0.0211	0.0	0.247	0.088
0.15	0.0633	0.0235	0.0675	93.0	0.038	0.233	0.0192	0.0159	0.0249	0.0	0.249	0.088
0.20	0.0475	0.0382	0.0610	65.4	0.044	0.267	0.0120	0.0182	0.0218	0.0	0.250	0.088
0.25	0.0192	0.0276	0.0336	31.4	0.049	0.297	0.0047	0.0155	0.0162	0.0	0.251	0.088
0.50	0.0015	0.0134	0.0135	6.3	0.051	0.310	0.0004	0.0134	0.0134	0.0	0.252	0.089
1.00	0.0002	0.0132	0.0132	5.6	0.051	0.310	0.0001	0.0133	0.0133	0.0	0.252	0.089
						$T = 50$						
0.00	-0.0001	0.0060	0.0060	5.8	0.023	0.173	-0.0002	0.0078	0.0078	0.0	0.127	0.048
0.05	0.0246	0.0067	0.0254	96.8	0.023	0.173	0.0060	0.0079	0.0099	0.0	0.127	0.048
0.10	0.0482	0.0090	0.0490	99.8	0.023	0.175	0.0100	0.0090	0.0134	0.0	0.128	0.049
0.15	0.0478	0.0272	0.0550	83.6	0.026	0.199	0.0058	0.0103	0.0118	0.0	0.129	0.049
0.20	0.0117	0.0144	0.0186	32.5	0.031	0.237	0.0014	0.0090	0.0091	0.0	0.129	0.049
0.25	0.0047	0.0093	0.0105	12.8	0.032	0.243	0.0005	0.0088	0.0088	0.0	0.129	0.049
0.50	0.0004	0.0085	0.0085	5.7	0.032	0.245	-0.0001	0.0087	0.0087	0.0	0.129	0.049
1.00	-0.0001	0.0084	0.0084	5.6	0.032	0.245	-0.0002	0.0086	0.0086	0.0	0.129	0.049
						$T = 100$						
0.00	0.0001	0.0041	0.0041	5.0	0.016	0.123	0.0001	0.0055	0.0055	0.0	0.080	0.033
0.05	0.0248	0.0046	0.0253	100.0	0.016	0.123	0.0047	0.0056	0.0073	0.0	0.080	0.033
0.10	0.0482	0.0071	0.0488	99.9	0.016	0.125	0.0056	0.0067	0.0088	0.0	0.080	0.033
0.15	0.0178	0.0172	0.0248	61.1	0.021	0.163	0.0015	0.0063	0.0065	0.0	0.081	0.033
0.20	0.0048	0.0064	0.0080	16.3	0.022	0.172	0.0006	0.0060	0.0061	0.0	0.081	0.033
0.25	0.0024	0.0059	0.0064	7.6	0.023	0.173	0.0003	0.0060	0.0060	0.0	0.081	0.033
0.50	0.0004	0.0057	0.0057	4.8	0.023	0.174	0.0001	0.0060	0.0060	0.0	0.081	0.033
1.00	0.0001	0.0057	0.0057	4.9	0.023	0.174	0.0001	0.0060	0.0060	0.0	0.081	0.033
						$T = 300$						
0.00	-0.0000	0.0024	0.0024	5.5	0.009	0.105	-0.0001	0.0036	0.0036	0.0	0.043	0.018
0.05	0.0249	0.0028	0.0250	100.0	0.009	0.106	0.0030	0.0037	0.0048	0.0	0.043	0.018
0.10	0.0310	0.0169	0.0352	95.9	0.011	0.123	0.0011	0.0040	0.0042	0.0	0.043	0.018
0.15	0.0036	0.0037	0.0051	21.8	0.013	0.148	0.0002	0.0039	0.0039	0.0	0.044	0.018
0.20	0.0014	0.0034	0.0037	7.5	0.013	0.149	0.0001	0.0039	0.0039	0.0	0.044	0.018
0.25	0.0007	0.0033	0.0034	5.2	0.013	0.149	-0.0000	0.0039	0.0039	0.0	0.044	0.018
0.50	0.0000	0.0033	0.0033	4.5	0.013	0.149	-0.0000	0.0039	0.0039	0.0	0.044	0.018
1.00	-0.0000	0.0033	0.0033	4.5	0.013	0.149	-0.0001	0.0039	0.0039	0.0	0.044	0.018

$\text{Lind}(A) \in \{0.0025, 0.0011, 0.0006, 0.0002\}$ for $T \in \{20, 50, 100, 300\}$. The results are based on 5,000 simulations.

C.2 A design with an additional covariate and heteroskedastic serially correlated errors

Similarly to Section 5.1, we consider

$$\begin{aligned} Y_{it} &= X_{it}\beta + Z_{it}\delta + \kappa\lambda_i f_t + U_{it}, \\ X_{it} &= \lambda_i f_t + V_{it}^X, \\ Z_{it} &= \lambda_i f_t + V_{it}^Z, \end{aligned}$$

where λ_i, f_t, and (V_{it}^X, V_{it}^Z) are all mutually independent across i, t, and (i,t), and

$$\lambda_i \sim N(0,1) \perp f_t \sim N(0,1) \perp \begin{pmatrix} V_{it}^X \\ V_{it}^Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_V^2 & \rho\sigma_V^2 \\ \rho\sigma_V^2 & \sigma_V^2 \end{pmatrix} \right).$$

Conditional on λ_i, f_t, and (V_{it}^X, V_{it}^Z) for $i \in \{1,\dots,N\}$ and $t \in \{0\dots,T\}$, we construct serially correlated errors U_{it} as

$$U_{it} = \varepsilon_{it} + \theta_\varepsilon \varepsilon_{it-1},$$

where $\varepsilon_{it} = \sigma_\varepsilon(X_i, Z_i, \lambda_i, f_t)e_{it}$ and e_{it} are independently drawn from a scaled Student's t-distribution with 5 degrees of freedom and variance normalized to 1. The conditional variance of ε_{it} is given by

$$\sigma_\varepsilon^2(X_i, Z_i, \lambda_i, f_t) = \frac{\sigma_U^2}{1+\theta_\varepsilon^2}\left(\frac{1}{2} + \Lambda\left(\frac{X_{it} + Z_{it} + \lambda_i f_t}{3}\right)\right),$$

where $\Lambda(\cdot)$ stands for the logistic CDF.

In this experiment, we fix $(\beta, \delta, \sigma_U^2, \sigma_V^2, \rho_V, \theta_\varepsilon) = (0, 1, 1, 1, 1/\sqrt{2}, 1/\sqrt{2})$. As in Section 5, we compare the performance of the standard LS based approach and our approach for various values of κ. Both approaches are implemented with $R = 1$ and $R = 2$, i.e., when the number of factors is correctly specified and when it is overspecified. To adjust for serial correlation of U_{it}, we compute HAC standard errors for the LS estimator accounting for up to two lags of serial correlation, and we compute clustered standard errors for our approach allowing for an arbitrary form of serial correlation.

For brevity, we only report results for estimation and inference on β and for $(N,T) = (100, 50)$ in Table 8 below. The results are qualitatively similar to the ones presented in Section 5. Our findings suggest that the standard LS based approach might perform poorly when there is a weak factor, and our approach reduces the weak factors bias and improves the quality of point estimation in more complicated settings with additional covariates and non-Gaussian, heteroskedastic, and serially correlated errors. Importantly, when the number of factors is overspecified ($R = 2$), the bias of the LS estimator does not vanish, so one cannot protect themselves from the weak factors bias by simply conservatively overspecifying

R. At the same time, overspecifying the number of factors does not result in loss of efficiency for our estimator. Nonetheless, overspecifying R leads to wider confidence intervals. While this might seem as a disadvantage of our approach, the necessity of having wider confidence intervals when the true number of factors is unknown has been established in Zhu (2019). Specifically, Zhu (2019) shows that the uncertainty in the number of factors necessarily results in a dramatic loss of inference efficiency when robustness to weak factors is required.

Table 8: Simulation results for the experiment in Section C.2

	LS						Debiased					
κ	bias	std	rmse	size	length	length*	bias	std	rmse	size	length	length*
						$R=1$						
0.00	-0.0001	0.0195	0.0195	6.2	0.073	0.189	-0.0003	0.0208	0.0209	0.0	0.295	0.145
0.05	0.0132	0.0195	0.0236	12.3	0.073	0.189	0.0087	0.0209	0.0226	0.0	0.295	0.145
0.10	0.0264	0.0197	0.0329	29.9	0.073	0.190	0.0176	0.0210	0.0274	0.0	0.295	0.146
0.15	0.0393	0.0201	0.0442	55.1	0.073	0.191	0.0260	0.0212	0.0336	0.0	0.296	0.147
0.20	0.0512	0.0213	0.0555	76.0	0.074	0.192	0.0332	0.0220	0.0398	0.0	0.298	0.147
0.25	0.0543	0.0286	0.0614	75.9	0.075	0.194	0.0337	0.0255	0.0423	0.0	0.301	0.148
0.50	0.0018	0.0218	0.0219	6.2	0.078	0.203	0.0015	0.0224	0.0224	0.0	0.309	0.149
1.00	-0.0000	0.0204	0.0204	5.6	0.078	0.203	-0.0002	0.0218	0.0218	0.0	0.309	0.149
						$R=2$						
0.00	-0.0001	0.0196	0.0196	6.9	0.071	0.184	-0.0003	0.0210	0.0210	0.0	0.490	0.140
0.05	0.0131	0.0197	0.0237	13.7	0.071	0.185	0.0087	0.0210	0.0227	0.0	0.491	0.140
0.10	0.0263	0.0199	0.0330	32.0	0.071	0.185	0.0174	0.0212	0.0274	0.0	0.491	0.140
0.15	0.0389	0.0204	0.0439	55.8	0.072	0.186	0.0255	0.0215	0.0333	0.0	0.493	0.141
0.20	0.0494	0.0221	0.0542	73.9	0.072	0.187	0.0314	0.0224	0.0386	0.0	0.495	0.142
0.25	0.0476	0.0304	0.0565	67.4	0.073	0.190	0.0288	0.0262	0.0389	0.0	0.499	0.142
0.50	0.0015	0.0207	0.0208	6.9	0.076	0.197	0.0011	0.0219	0.0220	0.0	0.506	0.143
1.00	0.0000	0.0206	0.0206	6.4	0.076	0.197	-0.0002	0.0219	0.0219	0.0	0.507	0.143

The results are based on 5,000 simulations.

C.3 Coverage of non-robust CIs based on the debiased estimator

In this section, we present additional simulation results for the numerical experiment considered in Section 5.1.

Specifically, in Table 9 below, we report the size of the t-test (with nominal size 5%) based on the non-robust CI provided in (13) (with 95% nominal coverage) and its average length. As before, we also report the same statistics for the LS CI.

We find that the non-bias aware CIs based on the debiased estimator are comparable in terms of the length to the LS CIs and that their coverage is close to 95% in the absence of weak factors. This is in line with our asymptotic analysis provided in Section 4.3. At the same time, even if there is a weak factor, our non-robust CIs enjoy much better coverage

than the LS ones. For example, for $(N,T) = (300, 300)$ and $\kappa = 0.10$, our non-robust CI has coverage 91.4% whereas the coverage of the LS CI is only 4.1%.

Table 9: Simulation results for the non-robust CI (13)

	$(N,T) = (100,100)$				$(N,T) = (300,100)$				$(N,T) = (300,300)$			
	LS		Debiased		LS		Debiased		LS		Debiased	
κ	size	length	size	length	size	length	size	length	size	length	size	length
0.00	6.1	0.028	5.9	0.041	5.0	0.016	5.8	0.021	5.5	0.009	5.2	0.014
0.05	91.0	0.028	8.9	0.041	100.0	0.016	13.7	0.021	100.0	0.009	13.1	0.014
0.10	99.9	0.028	17.0	0.041	99.9	0.016	23.3	0.021	95.9	0.011	8.6	0.014
0.15	92.9	0.030	17.5	0.041	61.1	0.021	10.0	0.021	21.8	0.013	6.8	0.014
0.20	47.4	0.037	10.1	0.041	16.3	0.022	7.8	0.021	7.5	0.013	6.8	0.014
0.25	17.9	0.039	8.5	0.041	7.6	0.023	7.7	0.021	5.2	0.013	6.8	0.014
0.50	5.9	0.039	8.0	0.041	4.8	0.023	7.5	0.021	4.5	0.013	6.8	0.014
1.00	5.4	0.039	8.0	0.041	4.9	0.023	7.6	0.022	4.5	0.013	6.8	0.014

The results are based on 5,000 simulations.

D Additional Results for Empirical Illustration

In his section, we revisit the empirical application considered in Section 5.2 by considering an alternative specification with dynamic treatment effects. Specifically, in the spirit of Wolfers (2006), Kim and Oka (2014) and Moon and Weidner (2015), we consider

$$Y_{it} = \sum_{k=1}^{4} X_{k,it}\beta_k + \alpha_i + \zeta_i t + \nu_i t^2 + \phi_t + \sum_{r=1}^{R} \lambda_{ir} f_{tr} + U_{it},$$

where $X_{k,it}$ are the treatment dummies defined as

$$X_{k,it} = \mathbf{1}\{D_i + 4(k-1) \le t \le D_i + 4k - 1\} \quad \text{for} \quad k \in \{1, \dots, 3\},$$
$$X_{4,it} = \mathbf{1}\{D_i + 12 \le t\},$$

where D_i denotes the year in which state i adopted a unilateral divorce law. Here, instead of introducing bi-annual dummies as in Wolfers (2006), we consider a coarser dynamics of treatment effects to ensure that our regressors $X_{k,it}$ have sufficient variation necessary for debiasing.

As before, we estimate and construct 95% CIs for β_k using the LS and our approaches. The results are provided in Table 10 below. They are qualitatively similar to the results reported in Section 5.2. While it is also possible to obtain shorter confidence intervals and establish significance of certain dynamic effects using our approach in the absence of weak

factors, we again find that the potential presence of one weak factors is sufficient to render the estimated effects insignificant.

Table 10: LS and debiased estimates and 95% CIs for dynamic effects of divorce law reform

	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	$R=6$
			LS			
years 1-4	0.033	0.084	0.093	0.040	0.012	0.085
	$[-0.07, 0.14]$	$[-0.04, 0.21]$	$[-0.03, 0.21]$	$[-0.08, 0.16]$	$[-0.11, 0.14]$	$[-0.04, 0.21]$
years 5-8	-0.081	-0.026	0.025	-0.028	-0.078	0.088
	$[-0.23, 0.07]$	$[-0.20, 0.15]$	$[-0.15, 0.20]$	$[-0.19, 0.14]$	$[-0.24, 0.09]$	$[-0.06, 0.24]$
years 9-12	-0.253	-0.247	-0.186	-0.244	-0.297	-0.065
	$[-0.45, -0.05]$	$[-0.49, -0.00]$	$[-0.41, 0.04]$	$[-0.46, -0.03]$	$[-0.50, -0.09]$	$[-0.26, 0.13]$
years 13+	-0.198	-0.255	-0.234	-0.313	-0.365	-0.110
	$[-0.46, 0.06]$	$[-0.54, 0.03]$	$[-0.50, 0.03]$	$[-0.57, -0.06]$	$[-0.61, -0.12]$	$[-0.35, 0.13]$
			Debiased			
years 1-4	0.081	0.147	0.137	0.098	0.079	0.118
$R_w = 0$	$[-0.03, 0.19]$	$[0.05, 0.25]$	$[0.05, 0.22]$	$[0.02, 0.18]$	$[0.00, 0.16]$	$[0.05, 0.19]$
$R_w = 1$	$[-0.80, 0.96]$	$[-0.58, 0.88]$	$[-0.45, 0.72]$	$[-0.39, 0.59]$	$[-0.33, 0.49]$	$[-0.23, 0.47]$
$R_w = R$	$[-0.80, 0.96]$	$[-1.21, 1.51]$	$[-1.45, 1.72]$	$[-1.63, 1.83]$	$[-1.64, 1.80]$	$[-1.63, 1.86]$
years 5-8	-0.008	0.054	0.099	0.056	0.031	0.125
$R_w = 0$	$[-0.17, 0.15]$	$[-0.08, 0.19]$	$[-0.01, 0.21]$	$[-0.05, 0.16]$	$[-0.07, 0.13]$	$[0.04, 0.22]$
$R_w = 1$	$[-1.34, 1.33]$	$[-1.04, 1.14]$	$[-0.77, 0.97]$	$[-0.67, 0.79]$	$[-0.57, 0.63]$	$[-0.39, 0.64]$
$R_w = R$	$[-1.34, 1.33]$	$[-2.00, 2.10]$	$[-2.30, 2.50]$	$[-2.56, 2.67]$	$[-2.57, 2.63]$	$[-2.51, 2.76]$
years 9-12	-0.147	-0.139	-0.098	-0.126	-0.157	0.003
$R_w = 0$	$[-0.36, 0.06]$	$[-0.33, 0.05]$	$[-0.25, 0.05]$	$[-0.27, 0.02]$	$[-0.29, -0.02]$	$[-0.12, 0.13]$
$R_w = 1$	$[-2.04, 1.75]$	$[-1.70, 1.43]$	$[-1.34, 1.15]$	$[-1.17, 0.92]$	$[-1.01, 0.70]$	$[-0.73, 0.74]$
$R_w = R$	$[-2.04, 1.75]$	$[-3.08, 2.81]$	$[-3.54, 3.34]$	$[-3.88, 3.63]$	$[-3.89, 3.58]$	$[-3.78, 3.79]$
years 13+	-0.178	-0.228	-0.196	-0.225	-0.262	-0.071
$R_w = 0$	$[-0.45, 0.09]$	$[-0.46, 0.01]$	$[-0.38, -0.01]$	$[-0.40, -0.05]$	$[-0.44, -0.09]$	$[-0.23, 0.09]$
$R_w = 1$	$[-2.71, 2.35]$	$[-2.31, 1.86]$	$[-1.85, 1.46]$	$[-1.62, 1.17]$	$[-1.40, 0.88]$	$[-1.05, 0.90]$
$R_w = R$	$[-2.71, 2.35]$	$[-4.16, 3.71]$	$[-4.80, 4.40]$	$[-5.26, 4.80]$	$[-5.26, 4.74]$	$[-5.14, 4.99]$