Essays in the Econometric Theory of Panel and Multidimensional Data

Hugo Stuart Harold Freeman

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Economics

University College London

June 27, 2023

I, Hugo Stuart Harold Freeman, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Date: 13 March, 2023

Abstract

This dissertation studies econometric models in the presence of unobserved heterogeneity when data is observed over multiple dimensions. Chapter 2 and 3 study this in the classic panel setting with two dimensions, which are usually individuals and time. Chapter 2 studies the setting where unobserved heterogeneity may enter non-linearly and nonseparably to the observed covariates. Established matrix completion methods and a group fixed-effect type estimator prove to approximate the model well. Chapter 3 studies the setting where unobserved heterogeneity enters linearly and separably, but is modelled as a generic functional transformation of unobserved characteristics. The factor model estimated with many factors approximates this form of unobserved heterogeneity well, and, like in Chapter 2, a group fixed-effects estimator also performs well in theory and in simulations. Chapter 4 studies this setting when three or more dimensions are observed in the data and restricts focus to the linear regression model. This chapter extends the notion of the group fixed-effects estimator to a nonparametric kernel style transformation that can be applied to any number of dimensions. The results in this chapter show that the current state-of-the-art factor model methods to approximate unobserved heterogeneity do not extend well to the setting with three or more dimensions. The results also show that the novel nonparametric kernel transformation proposed in this chapter control for unobserved heterogeneity sufficiently well to achieve the parametric rate of consistency under certain conditions.

Impact Statement

This dissertation makes several contributions to the literature on econometric theory and the results are useful for practitioners working with panel and multidimensional datasets. In particular, the methods proposed herein are useful for practitioners who want model estimates that are robust to many forms of unobserved heterogeneity.

Chapter 2 proposes two estimation procedures to estimate nonseparable models of unobserved heterogeneity that are useful to practitioners studying models with a binary covariate of interest, for example, to estimate average treatment effects. The results show that existing matrix completion methods can approximate counterfactuals in this model well, and that the novel group fixed-effects estimator can outperform these existing methods in certain settings.

Chapter 3 proposes two estimation procedures to estimate regression coefficients in linear models with an additive and separable unobserved heterogeneity term that is specified to be flexible to many different functional forms. This is useful to practitioners who are concerned with robustness of their estimates in linear models, where they may be concerned there is a high-dimensional and complicated form of unobserved heterogeneity. The state-of-the-art factor model approximates this model for unobserved heterogeneity well, and a novel group fixed-effects estimator also approximates this model well. The group fixed-effects estimator may also have useful inference properties, that are left for future research.

Chapter 4 extends the interactive fixed-effects model to the setting where data is observed over an arbitrary number of dimensions, and focuses on the more challenging case of three or more dimensions. This chapter is particularly useful for practitioners using data that varies over many dimensions - such as supermarket

data, or repeated network data like international trade - who are concerned that unobserved heterogeneity may interact over all dimensions of their data. The theoretical and simulation results highlight the benefit of the proposed novel estimators in controlling for this complicated form of unobserved heterogeneity, and the inadequacy in using existing state-of-the-art techniques. Prior to this work there were only limited methods available to practitioners working in the three or higher dimensional setting.

Acknowledgements

I am immensely thankful to my supervisor, Martin Weidner, for taking me on as a student and dedicating so much time to my development. I also thank Lars Nesheim for his support throughout my studies as my secondary advisor. I thank Iván Fernández-Val for giving me the opportunity to work with him on a joint project. I also would like to the UCL Economics Department more generally for their teaching and support throughout my studies, including Andrei, who gave me great guidance on my job market paper and job market talk.

I would also like to thank my fellow PhD students who have made my time at UCL truly enjoyable. I had a great time with Matt, Nathan, Elena, Alex, Cris, Mikkel, and Guanyi throughout my stay in London and we have made many lasting memories together.

I have to thank my wife, Nicola, who put up with me through my exams, put up with me through my upgrade, and put up with me through the job market. She has spent the last five or so years putting up with me and for that I am forever thankful.

I want to thank my parents and step-parents – Quintin, Kamala, Jeremy, and Donella – for their continued support through my studies into "extraterrestrial intelligence," as mum would put it. There may or may not be life on Mars, but there is certainly life in her household.

I would like to mention my son, Freddie, who has been a great adversary to my academic output these last few years, and made sure that I really earnt this degree.

Lastly, I would like to recognise the support of the European Research Council Grant, ERC-2018-CoG-819086-PANEDA, which sponsored my PhD studentship.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern econometric datasets often record observations of economic activity over multiple dimensions, for example, to record the economic behaviour of the same set of individuals over many time periods. This classic example is referred to as panel data, and the more general case of observing three or more dimensions is referred to as multidimensional data. The additional avenue of variation afforded by these datasets greatly improves our ability to control for latent behaviour that is not explicitly observed, but can be inferred by repeated observation. This latent behaviour is commonly known as unobserved heterogeneity, and controlling for this presents many opportunities and challenges for practitioners analysing such datasets. This dissertation considers the problem in a number of different settings.

Chapter 2 explores the panel data setting when discrete valued covariates are observed, and mainly focuses on the binary treatment effect case. This chapter allows for flexible interactions between the binary covariate of interest and latent characteristics that vary over either or both dimensions of the data. To achieve this, the chapter considers a nonseparable functional form that translates values of the latent characteristics in an arbitrary manner, subject to smoothness conditions across values of the latent characteristics. This smoothness condition is a technical requirement that may not be required in general, but is imposed as a sufficient condition for consistency. To deal with unobserved heterogeneity, this chapter proposes a matrix completion technique and a novel group fixed-effect method that treat counterfactual values of the dependent variable as missing data to be estimated. Theoretical

and simulation results show that both methods are consistent, and whilst the matrix completion method has large biases, the group fixed-effects estimator offers substantial bias reduction. The methods are implemented in an empirical application of the effect of election day registration on voter turnout in the U.S.

Chapter 3 studies linear regression models in the panel data setting where unobserved heterogeneity enters additively, and is a flexible transformation of latent characteristics. This is done by allowing the function that interacts latent characteristics to be nonseparable across dimensions of the data, and unspecified up to smoothness conditions over variation of the latent characteristics, much like in Chapter 2. The object of interest in this chapter are the slope coefficients on the observed covariates. The advantage of the linear and additive model is that it can deal with high dimensional variation in the covariate of interest, which the model and methods in Chapter 2 are not suitable for. This chapter shows that existing state-of-the-art factor models can consistently estimate the slope coefficients, but requires the estimated number of factors to increase asymptotically, which makes formal inference results difficult. The chapter also introduces a group fixed-effects estimator, which turns out to have a faster rate of consistency for the slope coefficients under certain smoothness conditions and restrictions on the number of latent characteristics. Simulation results confirm the theoretical findings and the methods are implemented in an empirical application on UK house prices.

Chapter 4 studies linear regression models with additive unobserved heterogeneity when data is observed over three or more dimensions. This setting poses significant technical challenges for existing estimation methods, and the chapter demonstrates the need for a nuanced approach. Focus is again on the estimation of slope coefficients related to observed covariates. The chapter shows that whilst existing factor model methods can consistently estimate the slope coefficients, the rate of consistency is slow when dimensions grow at roughly similar rates. Generalisations of the factor model to multidimensional settings suffer from many theoretical shortcomings and are largely avoided in the estimation approach developed in this chapter. Instead, the method proposed generalises a group fixed-effects style esti-

mator - similar to the method in Chapter 2 and Chapter 3 - to a nonparametric kernel type estimator that has superior asymptotic and finite sample bias properties to the current toolset available in the literature. Theoretical results suggest the parametric rate of consistency can be shown for this new estimator, and simulation results corroborate these findings. The methods are implemented to estimate the elasticity of demand for beer using the Dominick's supermarket dataset.

Chapter 2

Low-Rank Approximations of Nonseparable Panel Models

2.1 Introduction

Nonseparable models are useful to capture multidimensional unobserved heterogeneity, which is an important feature of economic data. The presence of this heterogeneity makes the effect of covariates on the outcome of interest different for each unit due to factors that are unobservable or unavailable to the researcher. In the absence of further restrictions, a different data generating process essentially operates for each unit, which creates identification and estimation challenges. One way to deal with these challenges is the use of panel data, where each unit is observed on multiple occasions. In this paper, we develop an approach to estimate nonseparable models from panel data based on homogeneity restrictions and low-rank factor approximations. Whilst homogeneity restrictions have been used previously in this context, the application of low-rank factor approximations is more novel.

The nonseparable model that we consider includes observed discrete covariates or treatments, multidimensional unobserved individual and time effects, and idiosyncratic errors. We construct the effects of interest as averages or quantiles of potential outcomes constructed from the model by exogenously manipulating the value of the treatments. These effects are generally not identified from the observed data because the treatment assignment is usually determined by the unobserved in-

dividual and time effects. Following the previous panel literature, we impose cross-section and time-series homogeneity restrictions to identify the effects of interest, see, e.g. Chamberlain (1982), Manski (1987), Honoré (1992), Evdokimov (2010), Graham and Powell (2012), Hoderlein and White (2012a) and Chernozhukov et al. (2013a).

The estimation of the nonseparable model is challenging due to the presence of the multidimensional unobserved individual and time effects. We cannot just exclude these effects because they are endogenous, i.e., related to the treatments. We deal with this problem by approximating their effect with a low-rank factor structure. This approach can be interpreted as a series or sieve approximation on the unobservables. We characterize the error of this approximation in terms of the functional singular value decomposition of the expectation of the outcome conditional on the treatment and unobserved effects. For smooth conditional expectation functions, the mean squared error of the approximation error vanishes with the rank of the factor structure at a polynomial rate.

We develop an estimator of the low-rank factor approximation in the case where the covariate of interest is binary. This is an empirically relevant case as it covers the treatment effect model for panel data. We also show how to extend the model to include additive controls and fixed effects. Here, we rely on the analogy between the estimation of treatment effects and the matrix completion problem previously noted by Athey et al. (2017) and Amjad et al. (2018). Thus, given that the principal components program is combinatorially hard in the presence of missing data, we consider the convex relaxation of this program that replaces a constraint in the rank of a matrix by a constraint in its nuclear norm, following Srebro and Jaakkola (2003) and Fazel (2003). The resulting estimator is the matrix-completion estimator.

The main theoretical result of the paper is to show that the matrix-completion estimator is consistent under asymptotic sequences where the two dimensions of the panel grow to infinity at the same rate. This result does not follow from the existing matrix completion literature that assumes that the matrix to complete has low-rank.

In our case, the underlying matrix of interest can have full rank, but we impose appropriate smoothness assumptions on the data generating process that guarantee that the singular values of the matrix form a rapidly decreasing sequence. This allows a low-rank approximation, and it also implies a bound on the nuclear norm of the matrix. Our consistency proof for the matrix completion estimator therefore crucially relies on the bound of the nuclear norm, but does not impose any low-rank conditions. Our proof strategy also avoids the high-level *restricted strong convexity* assumption (see e.g. Negahban and Wainwright (2012)). We instead provide interpretable conditions on the underlying process of the observable and unobservable variables directly.

The matrix-completion estimator is consistent, but can be biased in small samples. This bias comes from two different sources: approximation bias due to the low-rank factor structure approximation and shrinkage bias due to the nuclear norm regularization of the principal component analysis program Cai et al. (2010); Ma et al. (2011); Bai and Ng (2019b). We propose matching approaches to debias the estimator. For each treatment level, the simplest approach consists of finding the observation in the other treatment level that is the closest in terms of the estimated factor structure. We also propose a two-way matching procedure that combines matching with a differences-in-differences approach. The two-way procedure is related to several recent proposals such as the matching approach of Imai and Kim (2019) to estimate causal effects from panel data and the blind regression of Li et al. (2017b) for matrix completion. The difference with these proposals is in the information used to match the observations. Imai and Kim (2019) use the treatment variable and Li et al. (2017b) the outcome, whereas we use the estimated factor structure. In this sense, the estimation of the factor structure can be seen as a preliminary de-noising step of the data Chatterjee (2015). Amjad et al. (2018) proposed a similar debiasing procedure based on the estimated factor structure, but they rely on synthetic control methods instead of matching. In contemporaneous and independent work, Chernozhukov et al. (2020) have developed an alternative rotation-debiasing method that can be applied to make inference on heterogenous

treatment effects in low-rank models. This method consists of the application of iterative least squares to the left and right singular vectors of the matrix-completion estimator.

We illustrate our methods with an empirical application to the effect of election day registration (EDR) on voter turnout and numerical simulations. We estimate average and quantile effects using a state-level panel dataset on the 24 U.S. presidential elections between 1920 and 2012 collected by Xu (2017). We find that, after controlling for possible non-random adoption, EDR has a positive effect, especially at the bottom of the voter turnout distribution. Our methods uncover stronger effects than standard difference-in-differences methods that rely on restrictive parallel trend assumptions. The simulation results show that our theoretical results provide a good representation of the behavior of the estimators in small samples.

The rest of the paper is organized as follows. Section 2.2 describes the model and effects of interest. Section 2.3 introduces the low-rank factor approximation and derives the properties of its matrix-completion estimator. The matching methods to debias the matrix-completion estimator are discussed in Section 2.4. Section 2.5 reports the results of the numerical examples. All the proofs of the theoretical results are gathered in the Appendix.

2.2 Model and Effects of Interest

Throughout this paper we consider the following nonseparable and nonparametric panel data model:

Assumption 2.2.1 (Model)**.**

$$Y_{it} = g(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}), \qquad i \in \mathbb{N} = \{1, \ldots, N\},\ t \in \mathbb{T} = \{1, \ldots, T\}, \tag{2.1}$$

where i and t index individual units and time periods, respectively; Y_{it} is an observed outcome or response variable with support $\mathbb{Y} \subseteq \mathbb{R}$; g is an unknown function; $\boldsymbol{X}_{it}$ is a vector of observed covariates or treatments with finite support $\mathbb{X}$; $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are vectors of individual and time unobserved effects, possibly correlated with $\boldsymbol{X}_{it}$, with supports $\mathbb{A} \subseteq \mathbb{R}^{d_a}$ and $\mathbb{B} \subseteq \mathbb{R}^{d_b}$, respectively; and $\boldsymbol{U}_{it}$ is a vector of unobserved error

terms of unspecified dimension, for which we assume that

$$\boldsymbol{U}_{it} \overset{d}{=} \boldsymbol{U}_{js} \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T}, \qquad \text{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T}, \tag{2.2}$$

and

$$\boldsymbol{U}_{it} \perp\!\!\!\perp (\boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T}) \mid \boldsymbol{A}_i, \boldsymbol{B}_t, \qquad \text{for all } i \in \mathbb{N},\ t \in \mathbb{T}, \tag{2.3}$$

where $\boldsymbol{X}^{NT} = \{\boldsymbol{X}_{it} : i \in \mathbb{N}, t \in \mathbb{T}\}$, $\boldsymbol{A}^{N} = \{\boldsymbol{A}_i : i \in \mathbb{N}\}$, $\boldsymbol{B}^{T} = \{\boldsymbol{B}_t : t \in \mathbb{T}\}$, and $\perp\!\!\!\perp$ denotes stochastic independence. We also assume that, for all $i \in \mathbb{N}$, $t \in \mathbb{T}$, the support of $(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)$ is equal to the Cartesian product $\mathbb{X} \times \mathbb{A} \times \mathbb{B}$, and that $\mathrm{E} Y_{it}^2 < \infty$.

This model can be motivated from a purely statistical perspective as a latent variable model using the Aldous-Hoover representation for exchangeable random matrices, e.g. Xu et al. (2014), Chatterjee (2015), Orbanz and Roy (2015), and Li and Bell (2017).[1] We motivate it instead as a structural model where the unobserved effects $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are associated with individual heterogeneity and aggregate shocks, respectively. Additional exogenous covariates can be incorporated in the usual way by carrying out the analysis conditional on them. We focus on discrete covariates but, from a theoretical perspective, the extension to continuous covariates is straightforward by using appropriate smoothing methods — it is, however, not clear to us whether that extension would be practically useful with realistic sample sizes. We therefore think that it would complicate our presentation without much benefit.

The main restriction imposed by Assumption 2.2.1 is the unit and time homogeneity in (2.2). A sufficient condition for unit homogeneity is that the observations are identically distributed across i, which is a common sampling assumption for panel data. Time homogeneity has also been commonly used in panel data models (Chamberlain, 1982; Manski, 1987; Honoré, 1992; Evdokimov, 2010; Graham and Powell, 2012; Hoderlein and White, 2012a; Chernozhukov et al., 2013a). It implies that time is randomly assigned, conditional on covariates and unobserved effects.

[1] In the Aldous-Hoover representation, $\boldsymbol{A}_i$, $\boldsymbol{B}_t$ and $\boldsymbol{U}_{it}$ are independent uniform random variables.

The additional restrictions in (2.3) are exogeneity conditions on $(\boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T})$ with respect to $\boldsymbol{U}_{it}$, conditional on $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$. The most substantive is the exogeneity of $\boldsymbol{X}_{it}$. Given (2.2), this is a mild condition as time homogeneity already imposes that any relationship between $\boldsymbol{U}_{it}$ and $\boldsymbol{X}_{it}$ can only be unit and time-invariant. Taken together, (2.2) and (2.3) impose that

$$\boldsymbol{U}_{it} \mid \boldsymbol{A}_i, \boldsymbol{B}_t \overset{d}{=} \boldsymbol{U}_{js} \mid \boldsymbol{A}_j, \boldsymbol{B}_s, \qquad \text{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T}. \tag{2.4}$$

The product support condition guarantees overlap in the support of the unobserved effects for all values of the treatments. This condition is similar to the overlap condition used in cross section treatment effect models under unconfoundedness or selection on observables. Thus, together with (2.3), it implies that $P_{it}(x) := \Pr\left(X_{it} = x \mid \boldsymbol{A}^{N}, \boldsymbol{B}^{T}\right) > 0$, a.s., for all $i \in \mathbb{N}$, $t \in \mathbb{T}$ and $x \in \mathbb{X}$, where $P_{it}(x)$ is the analog of the propensity score in our setting. This condition is plausible in many applications. For example, in our empirical application in Section 2.5.1, $X_{it} = \mathbb{1}\{t \geq \tau_i\}$, where τ_i is the date of the law change in state i. In that case, if we consider τ_i to be a random variable with sufficiently large support conditional on the unobserved effects, then the condition $P_{it}(x) > 0$, a.s., is satisfied.

The model considered is similar to the static model in Chernozhukov et al. (2013a), but there are three important differences. First, the structural function g has time effects as arguments and therefore allows the relationship between Y_{it} and $\boldsymbol{X}_{it}$ to vary over time in an unrestricted fashion even under (2.2). For example, it can include location and scale time effects. Second, Chernozhukov et al. (2013a) impose that Y_{it} and $\boldsymbol{X}_{it}$ are identically distributed across i, which is stronger than the unit homogeneity in (2.4). Thus, unit homogeneity does not restrict the treatment assignment process. Third, they analyze short panels, whereas we rely on large T for identification. Our model also encompasses the nonseparable model with time effects in Freyberger (2017), where in our notation $Y_{it} = g_t(\boldsymbol{X}_{it}, \boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{B}_t + \boldsymbol{U}_{it})$.[2] We provide more examples of models covered by Assumption 2.2.1 below.

The structural function g is generally not identified, but can be used to construct

[2]Note that our model allows for g to depend on t because the dimension of $\boldsymbol{B}_t$ is unspecified.

interesting effects. Let $Y_{it}(\boldsymbol{x}) := g(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}(\boldsymbol{x}))$ be the potential outcome for individual i at time t obtained by setting exogenously $\boldsymbol{X}_{it} = \boldsymbol{x} \in \mathbb{X}$, where

$$\boldsymbol{U}_{it}(\boldsymbol{x}) \overset{d}{=} \boldsymbol{U}_{it} \mid \boldsymbol{A}^N, \boldsymbol{B}^T. \tag{2.5}$$

Here we impose rank similarity as the distribution of $\boldsymbol{U}_{it}(\boldsymbol{x})$ conditional on $\boldsymbol{A}^N$ and $\boldsymbol{B}^T$ does not change with $\boldsymbol{x}$. The main effects of interest are the average structural functions (ASFs)

$$\mu_t(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right], \;\; \mu(\boldsymbol{x}) := \frac{1}{T}\sum_{t=1}^{T} \mu_t(\boldsymbol{x}), \tag{2.6}$$

and the conditional average structural functions (CASFs)

$$\begin{aligned} \mu_t(\boldsymbol{x} \mid \mathbb{X}_0) &:= \frac{1}{N_t(\mathbb{X}_0)} \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{X}_{it} \in \mathbb{X}_0\} \mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right], \\ N_t(\mathbb{X}_0) &= \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{X}_{it} \in \mathbb{X}_0\}, \\ \mu(\boldsymbol{x} \mid \mathbb{X}_0) &:= \frac{1}{n(\mathbb{X}_0)} \sum_{t=1}^{T} N_t(\mathbb{X}_0) \mu_t(\boldsymbol{x} \mid \mathbb{X}_0), \;\; n(\mathbb{X}_0) = \sum_{t=1}^{T} N_t(\mathbb{X}_0), \end{aligned} \tag{2.7}$$

where $\mathbb{X}_0 \subseteq \mathbb{X}$, provided that $n(\mathbb{X}_0) > 0$. The ASFs and CASFs correspond to averages of the potential outcome $Y_{it}(\boldsymbol{x})$ at a given time period or aggregated over the observed time periods. In both cases the average is over the cross sectional units in the observed sample or finite population. Infinite-population versions of the effects can be obtained by taking probability limits as $N \to \infty$. If $\boldsymbol{X}_{it}$ includes only a binary treatment, the ASFs and CASFs can be used to form treatment effects. For example, $\mu(1) - \mu(0)$ is the time-aggregated average treatment effect and $\mu_t(1 \mid \{1\}) - \mu_t(0 \mid \{1\})$ is the average treatment effect on the treated at time t. Distribution structural functions (DSFs) can be constructed analogously replacing $Y_{it}(\boldsymbol{x})$ by $\mathbb{1}\{Y_{it}(\boldsymbol{x}) \leq y\}$ in (2.6) and (2.7) for $y \in \mathbb{Y}$. Quantile effects can then be formed by taking left-inverses of the DSFs and taking differences. For example, the

τ-quantile treatment effect at time t is $q_{t,\tau}(1) - q_{t,\tau}(0)$, where

$$q_{t,\tau}(\boldsymbol{x}) = \inf\left\{ y \in \mathbb{Y} : \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[\mathbb{1}\{Y_{it}(\boldsymbol{x}) \leq y\} \mid \boldsymbol{A}^N, \boldsymbol{B}^T \right] \geq \tau \right\}.$$

We provide some examples of data generating processes that satisfy Assumption 2.2.1. The purpose is to show that Assumption 2.2.1 covers a great variety of models commonly used in empirical analysis. Our estimation methods are generic in that we do not need to specify the data generating process, besides of satisfying Assumption 2.2.1. Of course, using more information about the data generating process would lead to more efficient estimators, but at the cost of robustness to model misspecification.

Example 2.2.1 (Linear factor model). Consider the linear panel model with factor structure in the error terms:

$$Y_{it}(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{f}_t + \sigma_i(\boldsymbol{x})\sigma_t(\boldsymbol{x})U_{it}(\boldsymbol{x}), \;\; U_{it}(\boldsymbol{x}) \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T \sim i.i.d.\ F_U,$$

where $U_{it}(\boldsymbol{x})$ is a zero mean random variable with marginal distribution F_U, which does not depend on $\boldsymbol{x}$. This is special case of Assumption 2.2.1 with $Y_{it} = Y_{it}(\boldsymbol{X}_{it})$, $\boldsymbol{A}_i = (\boldsymbol{\lambda}_i, \{\sigma_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\})$, $\boldsymbol{B}_t = (\boldsymbol{f}_t, \{\sigma_t(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\})$, and $\boldsymbol{U}_{it} = U_{it}(\boldsymbol{X}_{it})$. The average effect of changing the covariate from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$ at t is

$$\mu_t(\boldsymbol{x}_1) - \mu_t(\boldsymbol{x}_0) = \mu_t(\boldsymbol{x}_1 \mid \{\boldsymbol{x}_1\}) - \mu_t(\boldsymbol{x}_0 \mid \{\boldsymbol{x}_1\}) = (\boldsymbol{x}_1 - \boldsymbol{x}_0)^{\mathrm{T}}\boldsymbol{\beta}.$$

A version of this model was considered by Kim and Oka (2014) to analyze the effect of unilateral divorce laws on divorce rates in the U.S. This model encompasses the standard difference-in-differences model, $Y_{it}(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \lambda_i + f_t + \sigma_i(\boldsymbol{x})\sigma_t(\boldsymbol{x})U_{it}(\boldsymbol{x})$, by setting $\boldsymbol{\lambda}_i = (\lambda_i, 1)^{\mathrm{T}}$ and $\boldsymbol{f}_t = (1, f_t)^{\mathrm{T}}$.

Example 2.2.2 (Binary response model). Assume that the potential outcome $Y_{it}(\boldsymbol{x})$ is binary and generated by

$$Y_{it}(\boldsymbol{x}) = \mathbb{1}\{m(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t) \geq U_{it}(\boldsymbol{x})\}, \quad U_{it}(\boldsymbol{x}) \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T \sim i.i.d.\ \mathscr{U}(0,1),$$

for some unknown function m. Here, assuming that $U_{it}(\boldsymbol{x})$ is uniform is a normalization, since m can be arbitrary. This latent index model with unobserved effects is a special case of Assumption 2.2.1 with $Y_{it} = Y_{it}(\boldsymbol{X}_{it})$ and $\boldsymbol{U}_{it} = U_{it}(\boldsymbol{X}_{it})$. The ASFs at $\boldsymbol{x}$ and t is

$$\mu_t(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} m(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t).$$

Similar latent index models for count or censored responses are also covered by Assumption 2.2.1.

Example 2.2.3 (Treatment effect factor model). Assume that $\boldsymbol{X}_{it}$ contains only a binary treatment indicator, i.e., $\mathbb{X} = \{0,1\}$. The potential outcomes are generated by the linear factor model

$$Y_{it}(\boldsymbol{x}) = \boldsymbol{\lambda}_i(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{f}_t(\boldsymbol{x}) + \sigma_i(\boldsymbol{x})\sigma_t(\boldsymbol{x})U_{it}(\boldsymbol{x}),\ \ U_{it}(\boldsymbol{x}) \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T \sim i.i.d.\ F_U,\ \boldsymbol{x} \in \mathbb{X},$$

where $U_{it}(\boldsymbol{x})$ is a zero mean random variable with marginal distribution F_U, which does not depend on $\boldsymbol{x}$. This is special case of Assumption 2.2.1 with $Y_{it} = Y_{it}(\boldsymbol{X}_{it})$, $\boldsymbol{A}_i = (\{\boldsymbol{\lambda}_i(\boldsymbol{x}), \sigma_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\})$, $\boldsymbol{B}_t = (\{\boldsymbol{f}_t(\boldsymbol{x}), \sigma_t(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\})$, and $\boldsymbol{U}_{it} = U_{it}(\boldsymbol{X}_{it})$. The average treatment effect at t is

$$\mu_t(1) - \mu_t(0) = \frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{\lambda}_i(1)^{\mathrm{T}}\boldsymbol{f}_t(1) - \boldsymbol{\lambda}_i(0)^{\mathrm{T}}\boldsymbol{f}_t(0)],$$

and the average effect on the treated at t is

$$\mu_t(1 \mid \{1\}) - \mu_t(0 \mid \{1\}) = \frac{1}{N_t(1)}\sum_{i=1}^{N}\mathbb{1}\{\boldsymbol{X}_{it} = 1\}[\boldsymbol{\lambda}_i(1)^{\mathrm{T}}\boldsymbol{f}_t(1) - \boldsymbol{\lambda}_i(0)^{\mathrm{T}}\boldsymbol{f}_t(0)],$$

provided that $N_t(1) = \sum_{i=1}^{N}\mathbb{1}\{\boldsymbol{X}_{it} = 1\} > 0$. Versions of this model have been considered by Hsiao et al. (2012), Gobillon and Magnac (2016a), Athey et al. (2017), Li and Bell (2017), Xu (2017), Li (2018), Bai and Ng (2019a), Xiong and Pelger (2019), and Chan and Kwok (2020). Example 2.2.1 is a special case with $\boldsymbol{\lambda}_i(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{f}_t(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{f}_t$.

Throughout this paper we use standard panel data notation, with the two panel

dimensions being denoted by units i and time t. However, one could also consider pseudo-panel or network applications of our results, where the two panel dimensions are denoted by i and j, and Y_{ij} could, for example, be wage of worker i in firm j, consumption of member i in household j, a friendship indicator between individuals i and j, or the volume of trade from country i to country j. The existing literature on two-way heterogeneity in network models usually either makes stronger parametric assumptions than we impose here (e.g. Graham (2017), Dzemski (2019), Chen et al. (2020), Zeleneev (2020)) or uses stochastic blockmodels or graphon models, which typically ignore the effect of covariates (e.g. Holland et al. (1983), Wolfe and Olhede (2013), Gao et al. (2015), Auerbach (2019)). Our methods of estimating non-parametric models with two-way heterogeneity may therefore also be of interest in a network context.

2.3 Estimation via Factor Structure Approximation

A natural starting point to estimate the effects in (2.6) and (2.7) is to use empirical analogs. This amounts to replacing $\mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right]$ by an estimator. There are two complications with this approach. First, the potential outcome $Y_{it}(\boldsymbol{x})$ is not observable. We deal with this complication by noting that

$$\begin{aligned}\mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right] &= \mathrm{E}\left[g(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}(\boldsymbol{x})) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right] \\ &= \mathrm{E}\left[g(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right] = \mathrm{E}\left[g(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}) \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t\right] \\ &= \mathrm{E}\left[Y_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t\right],\end{aligned}$$

under the rank similarity in (2.5) and Assumption 2.2.1. Hence, we can write the expectation of the potential outcome as an expectation of the observed outcome. The second complication is that $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are not observable, so that we cannot directly estimate $\mathrm{E}\left[Y_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t\right]$. To deal with this complication, we start by

noticing that

$$\begin{aligned} \mathrm{E}[Y_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{B}_t = \boldsymbol{b}] &= \mathrm{E}[g(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{U}_{it}) \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{B}_t = \boldsymbol{b}] \\ &=: m(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}), \end{aligned} \tag{2.8}$$

where the function m does not vary with i and t, by implication (2.4) of Assumption 2.2.1. We show next how this function can be approximated and estimated using a low-rank factor structure.

2.3.1 Low-rank factor structure approximation

For ease of exposition, we assume in the rest of the paper that the covariate vector $\boldsymbol{X}_{it}$ includes only a binary treatment and $\mathbb{X} = \{0,1\}$. Accordingly, we denote the covariate and its values by X_{it} and x instead of $\boldsymbol{X}_{it}$ and $\boldsymbol{x}$. In what follows, x denotes a generic element of $\mathbb{X}$ and all the assumptions and results hold for all $x \in \mathbb{X}_1 \subseteq \mathbb{X}$, where $\mathbb{X}_1 = \mathbb{X}$ if we are interested in the entire population, $\mathbb{X}_1 = \{0\}$ if we are interested in the treated subpopulation, and $\mathbb{X}_1 = \{1\}$ if we are interested in the untreated subpopulation.

The approximation that we propose is based on the singular value decomposition of the function $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ for each $x \in \mathbb{X}$. We make two assumptions on this decomposition. The first assumption is a sampling condition on the unobserved effects that will be useful to define a norm for the eigenfunctions.

Assumption 2.3.1 (Sampling of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$)**.** (i) $\boldsymbol{A}_i$ is independent and identically distributed across $i \in \mathbb{N}$ with distribution $F_{\boldsymbol{A}}$, (ii) $\boldsymbol{B}_t$ is independent and identically distributed over $t \in \mathbb{T}$ with distribution $F_{\boldsymbol{B}}$, and (iii) $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are independent for all i, t.

For simplicity we consider the case where both $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are independently distributed across i and over t, but since we consider asymptotic sequences where both N and T become large one could also allow for appropriate weak dependence across both i and t. Formalizing this weak dependence would complicate both the assumption and the proof of the following results, which is why we decided to stick to independence in our presentation here.

The next assumption is a regularity condition on the function $m(x, \boldsymbol{a}, \boldsymbol{b})$.

Assumption 2.3.2 (Smoothness of $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$)**.** The function $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ admits a singular value decomposition

$$m(x, \boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{\infty} s_j(x)\, u_j(x, \boldsymbol{a})\, v_j(x, \boldsymbol{b}),$$

under the $L_2(F_{\boldsymbol{A}} \times F_{\boldsymbol{B}})$ norm, where the eigenfunctions $u_j(x, \boldsymbol{a})$ and $v_j(x, \boldsymbol{b})$ are orthonormal, i.e.,

$$\begin{aligned} &\mathbb{E}\, u_j(x, \boldsymbol{A}_i)^2 = 1, \qquad \mathbb{E}\, u_j(x, \boldsymbol{A}_i) u_k(x, \boldsymbol{A}_i) = 0, \\ &\mathbb{E}\, v_j(x, \boldsymbol{B}_t)^2 = 1, \qquad \mathbb{E}\, v_j(x, \boldsymbol{B}_t) v_k(x, \boldsymbol{B}_t) = 0, \qquad j \neq k \in \{1, 2, 3 \ldots\}, \end{aligned}$$

and the singular values $s_1(x) \geq s_2(x) \geq s_3(x) \geq \ldots \geq 0$ satisfy

$$\sum_{j=1}^{\infty} s_j(x) < \infty.$$

There is a large literature on singular value decompositions of functions, which shows that, under appropriate conditions, the singular values satisfy $s_j(x) \lesssim j^{-\alpha}$,[3] where the decay coefficient α depends on the dimensions of the arguments $\boldsymbol{a}$, $\boldsymbol{b}$, and on the smoothness of $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$. For sufficiently smooth functions, $\alpha > 1$ and therefore $\sum_{j=1}^{\infty} s_j(x) < \infty$. For example, if $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ is continuously differentiable up to order s and $\mathbb{A}$ and $\mathbb{B}$ are compact, then

$$s_j(x) \lesssim j^{-\frac{s}{d_a \wedge d_b}},$$

by Theorem 3.3 of Griebel and Harbrecht (2013), where $d_a \wedge d_b$ is the minimum of d_a and d_b. This implies that $\sum_{j=1}^{\infty} s_j(x) < \infty$ if $s > d_a \wedge d_b$. Assumption 2.3.2 is therefore a high-level smoothness assumption on $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$, very similar to the Assumption 2.2. in Menzel (2018), where an analogous condition on the singular values is imposed, with the same aim of controlling the behaviour of a

[3] Here, $s_j(x) \lesssim j^{-\alpha}$ means that there exists a constant $c > 0$ such that $s_j(x) \leq c\, j^{-\alpha}$, for all j.

function of unobserved two-dimensional heterogeneity.

The formulation of this smoothness assumption is convenient for our purposes, because it immediately leads to a low-rank approximation of $m(x,\boldsymbol{a},\boldsymbol{b})$. The low-rank approximation truncates the singular value decomposition to the first R elements,

$$m(x,\boldsymbol{a},\boldsymbol{b}) = \sum_{j=1}^{\infty} \underbrace{s_j(x)^{1/2}u_j(x,\boldsymbol{a})}_{=:\phi_j(x,\boldsymbol{a})} \underbrace{s_j(x)^{1/2}v_j(x,\boldsymbol{b})}_{=:\psi_j(x,\boldsymbol{b})} = \sum_{j=1}^{R} \phi_j(x,\boldsymbol{a})\psi_j(x,\boldsymbol{b}) + \zeta_R(x,\boldsymbol{a},\boldsymbol{b}). \tag{2.9}$$

The first term is the approximation and the second term is the approximation error. Under Assumption 2.3.2,

$$\mathrm{E}\,\zeta_R(x,\boldsymbol{A}_i,\boldsymbol{B}_t)^2 \to 0 \quad \text{as} \quad R\to\infty.$$

In other words, the approximation error can be made negligible by increasing the truncation point R. For example, if $s_j(x) \lesssim j^{-\alpha}$ with $\alpha > 1$, then

$$\begin{aligned}\mathrm{E}\,\zeta_R(x,\boldsymbol{A}_i,\boldsymbol{B}_t)^2 &= \mathrm{E}\left[\sum_{j=R+1}^{\infty} s_j(x)\,u_j(x,\boldsymbol{A}_i)\,v_j(x,\boldsymbol{B}_t)\right]^2 \\ &= \sum_{j,k=R+1}^{\infty} s_j(x)s_k(x)\mathrm{E}\left[u_j(x,\boldsymbol{A}_i)u_k(x,\boldsymbol{A}_i)\right]\mathrm{E}\left[v_j(x,\boldsymbol{B}_t)v_k(x,\boldsymbol{B}_t)\right] \\ &= \sum_{j=R+1}^{\infty} s_j(x)^2 \lesssim \sum_{j=R+1}^{\infty} j^{-2\alpha} \le \int_R^{\infty} j^{-2\alpha}\mathrm{d}j \lesssim R^{1-2\alpha},\end{aligned}$$

by Assumptions 2.3.1 and 2.3.2. Hence, $\zeta_R(x,\boldsymbol{A}_i,\boldsymbol{B}_t)$ converges in mean square to zero at a polynomial rate with R.

Combining (2.8) and (2.9), we obtain the approximate factor model

$$Y_{it} = \boldsymbol{\lambda}_i(X_{it})^{\mathrm{T}}\boldsymbol{f}_t(X_{it}) + \zeta_R(X_{it},\boldsymbol{A}_i,\boldsymbol{B}_t) + E_{it},\;\; E_{it} := Y_{it} - \mathrm{E}[Y_{it}\mid X_{it},\boldsymbol{A}_i,\boldsymbol{B}_t],\;\; (2.10)$$

where $\boldsymbol{\lambda}_i(x) = [\phi_1(x,\boldsymbol{A}_i),\ldots,\phi_R(x,\boldsymbol{A}_i)]^{\mathrm{T}}$, $\boldsymbol{f}_t(x) = [\psi_1(x,\boldsymbol{B}_t),\ldots,\psi_R(x,\boldsymbol{B}_t)]^{\mathrm{T}}$, and the composite error $v_{it} := \zeta_R(X_{it},\boldsymbol{A}_i,\boldsymbol{B}_t) + E_{it}$ contains the approximation error, $\zeta_R(X_{it},\boldsymbol{A}_i,\boldsymbol{B}_t)$, and the conditional expectation error, E_{it}. The factor structure can

be seen as a series or sieve approximation to the function $(\boldsymbol{a},\boldsymbol{b}) \mapsto m(x,\boldsymbol{a},\boldsymbol{b})$ with basis functions $\{\phi_j(x,\boldsymbol{a})\psi_j(x,\boldsymbol{b})\}_{j=1}^{\infty}$ if we let $R = R_{N,T}$ to grow with N and T such that $\zeta_R(x,\boldsymbol{a},\boldsymbol{b})$ vanishes as $N,T \to \infty$. The factor structure approximation is exact in some cases for fixed R. For instance, in Example 2.2.3

$$m(x,\boldsymbol{A}_i,\boldsymbol{B}_t) = \boldsymbol{\lambda}_i(x)^{\mathrm{T}}\boldsymbol{f}_t(x),$$

so that $\zeta_R(x,\boldsymbol{A}_i,\boldsymbol{B}_t) = 0$, a.s., if R is greater or equal to the number of factors.

In the model (2.10) the factor structure changes with the treatment level. In other words, we have a different pure factor model for each $x \in \mathbb{X}$, that is

$$Y_{it} = \boldsymbol{\lambda}_i(x)^{\mathrm{T}}\boldsymbol{f}_t(x) + v_{it} \text{ if } X_{it} = x.$$

This observation leads to our first estimation strategy where the data is partitioned by the treatment level and separate factors and factor loadings are estimated in each element of the partition by solving the least squares program

$$\min_{\{\boldsymbol{\lambda}_i\}_{i=1}^N,\{\boldsymbol{f}_t\}_{t=1}^T} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\left(Y_{it} - \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{f}_t\right)^2, \tag{2.11}$$

where $D_{it}(x) := \mathbb{1}\{X_{it} = x\}$. Unfortunately, we cannot solve this problem using standard principal component analysis due to the presence of missing data, that is, each observational unit (i,t) is not available at all treatment levels. In the next section, we apply matrix completion methods to deal with this problem.

2.3.2 Estimation by matrix completion methods

We start by expressing the program (2.11) in matrix form. Let $\boldsymbol{\Gamma}^R(x) = \boldsymbol{\lambda}^N(x)\boldsymbol{f}^T(x)^{\mathrm{T}}$, where $\boldsymbol{\lambda}^N(x) = [\boldsymbol{\lambda}_1(x),\ldots,\boldsymbol{\lambda}_N(x)]^{\mathrm{T}}$, a $N \times R$ matrix of factor loadings, and $\boldsymbol{f}^T(x) = [\boldsymbol{f}_1(x),\ldots,\boldsymbol{f}_T(x)]^{\mathrm{T}}$, a $T \times R$ matrix of factors. The least squares estimator of $\boldsymbol{\Gamma}^R(x)$ is the $N \times T$ matrix $\boldsymbol{\Gamma}$ with typical element Γ_{it} that solves

$$\min_{\{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}:\mathrm{rank}(\boldsymbol{\Gamma})\le R\}} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\left(Y_{it} - \Gamma_{it}\right)^2. \tag{2.12}$$

Let $\boldsymbol{Y}(x)$ be a $N \times T$ matrix whose (i,t) element is Y_{it} if $X_{it} = x$ and is missing otherwise. The previous program is closely related to the problem of completing the missing entries of $\boldsymbol{Y}(x)$ using a low rank approximation matrix $\boldsymbol{\Gamma}^R(x)$ Rennie and Srebro (2005); Candès and Recht (2009); Candes and Tao (2010). This connection was previously noticed by Athey et al. (2017) and Amjad et al. (2018) in the context of treatment effects models. The solution is the $N \times T$ matrix of rank R whose entries are the closest in the mean squared error sense to the corresponding entries of $\boldsymbol{Y}(x)$.

The previous program is combinatorially hard because of the constraint in the rank of the matrix Srebro and Jaakkola (2003). Following Fazel (2003) we consider the convex relaxation of this program. Let $\|\mathbf{M}\|_\infty$ be the spectral norm of a $\mathbb{R}^{N\times T}$-matrix $\mathbf{M}$, and define the nuclear norm (also called trace norm) of $\boldsymbol{\Gamma}$ as the corresponding dual norm $\|\boldsymbol{\Gamma}\|_1 := \max_{\{\mathbf{M}\in\mathbb{R}^{N\times T}:\|\mathbf{M}\|_\infty\leq 1\}} \mathrm{Tr}\left(\mathbf{M}'\boldsymbol{\Gamma}\right)$. This nuclear norm can equivalently be defined as the sum of the singular values of $\boldsymbol{\Gamma}$. Using this norm we can write the convex relaxation of the program (2.12) as follows,

$$\min_{\{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}:\|\boldsymbol{\Gamma}\|_1\leq R_1\}} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\left(Y_{it}-\Gamma_{it}\right)^2,$$

where R_1 is a positive constant such that $R = f(R_1)$, where f is an increasing function. Hence, $\zeta_R(x,\boldsymbol{A}_i,\boldsymbol{B}_t)$ vanishes in mean square as $R_1 \to \infty$. We replace the rank constraint, $\mathrm{rank}(\boldsymbol{\Gamma}) \leq R$, by a constraint on the nuclear norm of the matrix, $\|\boldsymbol{\Gamma}\|_1 \leq R_1$, i.e. we replace a constraint in the number of nonzero singular values by a constraint in the sum of singular values. This program is convex in $\boldsymbol{\Gamma}$ and can be reformulated in Lagrange form as

$$\min_{\{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}\}} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\left(Y_{it}-\Gamma_{it}\right)^2 + \rho(R_1)\|\boldsymbol{\Gamma}\|_1, \tag{2.13}$$

where $\rho(R_1) \geq 0$ is a regularization parameter, which is a one-to-one increasing function of R_1. There exist efficient algorithms to solve this program Mazumder et al. (2010).

Let $\widehat{\boldsymbol{\Gamma}}(x)$ be a solution to (2.13) with typical element $\widehat{\Gamma}_{it}(x)$. Then, we can form

estimators of the ASF and CASF as

$$\widehat{\mu}_t(x) = \frac{1}{N}\sum_{i=1}^{N}\left[D_{it}(x)Y_{it} + \{1 - D_{it}(x)\}\widehat{\Gamma}_{it}(x)\right],$$

and

$$\widehat{\mu}_t(x \mid \{x_0\}) = \frac{\sum_{i=1}^{N} D_{it}(x_0)\left[D_{it}(x)Y_{it} + \{1 - D_{it}(x)\}\widehat{\Gamma}_{it}(x)\right]}{\sum_{i=1}^{N} D_{it}(x_0)}.$$

In the next section, we provide conditions under which these estimators are consistent using asymptotic sequences where $N, T \to \infty$. These estimators, however, might display shrinkage biases in finite samples due to the nuclear norm regularization Cai et al. (2010); Ma et al. (2011); Bai and Ng (2019b). We propose two matching procedures to debias the estimator in Section 2.4.

2.3.3 Consistency of Matrix Completion Estimator

Let $\mathbf{\Gamma}^{\infty}(x)$ be the $N \times T$ matrix with typical element $\Gamma_{it}^{\infty}(x) = m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$ and $\boldsymbol{E}(x)$ be the $N \times T$ matrix with typical element

$$E_{it}(x) := \begin{cases} E_{it} = Y_{it} - \Gamma_{it}^{\infty}(x) & \text{if } X_{it} = x, \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Note that $\mathbf{\Gamma}^{\infty}(x) = \lim_{R\to\infty} \mathbf{\Gamma}^{R}(x)$ a.s. Furthermore, we introduce the notation $\mathbb{D}(x) = \{(i,t) \in \mathbb{N} \times \mathbb{T} : X_{it} = x\}$, and $n(x) = |\mathbb{D}(x)|$ for the number of observations with $X_{it} = x$.

Recall that

$$\widehat{\mathbf{\Gamma}}(x) \in \underset{\mathbf{\Gamma}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} Q_{NT}(\mathbf{\Gamma}, \rho, x), \quad Q_{NT}(\mathbf{\Gamma}, \rho, x) = \frac{1}{2}\sum_{(i,t)\in\mathbb{D}(x)} (Y_{it} - \Gamma_{it})^2 + \rho\|\mathbf{\Gamma}\|_1, \tag{2.15}$$

where $\rho := \rho(R_1)$. Here, if the argmin over $\mathbf{\Gamma} \in \mathbb{R}^{N\times T}$ is not unique, then we can choose $\widehat{\mathbf{\Gamma}}(x)$ arbitrarily from the set of minimizers — our results are not affected by that, we only require that $Q_{NT}(\widehat{\mathbf{\Gamma}}(x), \rho, x) \leq Q_{NT}(\mathbf{\Gamma}, \rho, x)$, for all $\mathbf{\Gamma} \in \mathbb{R}^{N\times T}$. We want to show that $\widehat{\mathbf{\Gamma}}(x)$ converges to $\mathbf{\Gamma}^{\infty}(x)$ as $N, T \to \infty$ in some sense such that

$\widehat{\mu}(x) - \mu(x) = o_P(1)$. For that we require additional assumptions.

Assumption 2.3.3 (Error Moments)**.** Conditional on $\boldsymbol{X}^{NT}$, $\boldsymbol{A}^N$ and $\boldsymbol{B}^T$, $E_{it}(x)$ is independent across $(i,t) \in \mathbb{D}(x)$, and there exists a constant $b < \infty$ that does not depend on i, t, N, T, such that

$$\mathrm{E}\left[E_{it}(x)^4 \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] \leq b.$$

Furthermore, we assume that $n(x)^{-1} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma_{it}^{\infty}(x)^2 = O_P(1)$.

For the purpose of showing Lemma 2.3.1 and Theorem 2.3.4 we could alternatively replace Assumption 2.3.3 by the two high-level conditions:

$$\frac{2}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma_{it}^{\infty}(x) E_{it} = o_P(1), \qquad \|\boldsymbol{E}(x)\|_{\infty} = O_P\left(\sqrt{N+T}\right),$$

where again $\|\cdot\|_{\infty}$ denotes the spectral norm. The first of those conditions is implied by Assumption 2.3.3 through application of the weak law of large numbers, while the second follows, for example, by the spectral norm inequality in Latała (2005). In principle, we could still derive those high-level conditions if we allowed for appropriate weak dependence of $E_{it}(x)$ across i and over t, but we again focus on the independent case for simplicity of presentation.

We first provide a consistency result for the entries of $\widehat{\boldsymbol{\Gamma}}(x)$ that correspond to the observed values of $\boldsymbol{Y}(x)$.

Lemma 2.3.1. *Let the Assumptions 2.3.1, 2.3.2 and 2.3.3 hold, and assume that $\rho = \rho_{NT}$ is chosen such that $\rho_{NT}/\sqrt{N+T} \to \infty$ and $\rho_{NT}\sqrt{NT}/n(x) \to 0$ as $N, T \to \infty$. Then,*

$$\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)\right]^2 = o_P(1).$$

A necessary condition for the existence of the sequence $\rho = \rho_{NT}$ in Lemma 2.3.1 is $n(x)/\sqrt{(N+T)NT} \to \infty$, that is, the fraction $n(x)/(NT)$ of observations with $X_{it} = x$ can converge to zero, but not too fast. Apart from that,

Lemma 2.3.1 does not restrict the assignment process that determines $\boldsymbol{X}^{NT}$. Notice also that Lemma 2.3.1 does not require Assumption 2.2.1 because $\Gamma^{\infty}(x)$ is a reduced-form parameter.

Applying the Cauchy-Schwarz inequality

$$\left(\frac{1}{n(x)}\sum_{(i,t)\in\mathbb{D}(x)} a_{it}\right)^2 \leq \frac{1}{n(x)}\sum_{(i,t)\in\mathbb{D}(x)} a_{it}^2$$

with $a_{it} = \widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)$, Lemma 2.3.1 guarantees that

$$\frac{1}{n(x)}\sum_{(i,t)\in\mathbb{D}(x)} \left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)\right] = o_P(1).$$

Nevertheless, Lemma 2.3.1 is not directly useful to show the consistency of the estimators of the ASF, because it only guarantees L_2-consistency of $\widehat{\boldsymbol{\Gamma}}(x)$ over the set of entries (i,t) for which $X_{it} = x$. Those are exactly the observations for which an unbiased estimator of $\Gamma_{it}^{\infty}(x) = m(x,\boldsymbol{A}_i,\boldsymbol{B}_t)$ is already available, namely Y_{it}. The consistency result we would like to obtain is

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)\right]^2 = o_P(1), \tag{2.16}$$

but such a result will certainly require stronger assumptions on $\boldsymbol{X}^{NT}$ than we have imposed so far.

The existing literature on matrix completion relies on the concept of *restricted strong convexity* to derive (2.16). This approach shows that under certain conditions on a $\mathbb{R}^{N\times T}$-matrix $\boldsymbol{M}$ with entries M_{it}, and on $\boldsymbol{X}^{NT}$ (which determines the set $\mathbb{D}(x)$), there exists a constant $c > 0$ such that with high probability

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}^2 \leq \frac{c}{n(x)}\sum_{(i,t)\in\mathbb{D}(x)} M_{it}^2.$$

See Theorem 1 in Negahban and Wainwright (2012), Lemma 12 in Klopp et al. (2014), and Lemma 3 in Athey et al. (2017). Thus, if $M_{it} = \widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)$ and $\boldsymbol{X}^{NT}$

satisfy restricted strong convexity, then (2.16) would follow from Lemma 2.3.1.

We pursue a different strategy than the existing matrix completion literature to show that

$$\widehat{\mu}(x) := \frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_t(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,Y_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[1-D_{it}(x)]\,\widehat{\Gamma}_{it}(x)$$

is a consistent estimator of $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Gamma_{it}^{\infty}$, which under Assumption 2.2.1 is equal to $\mu(x)$ defined in (2.6). We believe that our approach is simpler in the setting of this paper where $\Gamma_{it}^{\infty}(x)$ is not necessarily of low-rank. In particular, we do not aim to show (2.16), but instead we derive consistency of $\widehat{\mu}(x)$ directly. However, the following theorem still requires additional assumptions on the assignment process that determines $\boldsymbol{X}^{NT}$, in the same way that additional conditions on $\boldsymbol{X}^{NT}$ are required to verify restricted strong convexity. For simplicity, we focus on consistency of $\widehat{\mu}(x)$ in the main text, but results for more general weighted averages of the form $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)\,\Gamma_{it}^{\infty}(x)$, with known weights $W_{it}(x)\in\mathbb{R}$, are presented in the appendix. For example, in the case of the treatment effects on the treated that we consider in the empirical application of Section 2.5.1, $W_{it}(x) = n(1)^{-1}X_{it}$.

Theorem 2.3.4. *Let the Assumptions 2.2.1, 2.3.1, 2.3.2 and 2.3.3 hold. Consider $N,T\to\infty$ at the same rate, and let $\rho=\rho_{NT}$ be chosen such that $\rho_{NT}/\sqrt{N+T}\to\infty$ and $\rho_{NT}/\sqrt{NT}\to 0$. Let $P_{it}(x) = \Pr\left(X_{it}=x \mid \boldsymbol{A}^N,\boldsymbol{B}^T\right)$, and assume that $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}P_{it}^{-1}(x) = O_P(1)$. Let $\boldsymbol{G}(x)$ be the $N\times T$ matrix with entries $G_{it}(x) = P_{it}^{-1}(x)(D_{it}(x)-P_{it}(x))$, and assume that $\|\boldsymbol{G}(x)\|_{\infty} = O_P(\sqrt{N+T})$, and*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}P_{it}^{-1}(x)\,G_{it}(x) = o_P(1), \qquad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Gamma_{it}^{\infty}(x)\,G_{it}(x) = o_P(1). \tag{2.17}$$

Then,

$$\widehat{\mu}(x) = \mu(x) + o_P(1).$$

To interpret the conditions in Theorem 2.3.4, notice that due to the definitions $D_{it}(x) = \mathbb{1}\{X_{it}=x\}$ and $P_{it}(x) = \Pr\left(X_{it}=x \mid \boldsymbol{A}^N,\boldsymbol{B}^T\right)$, $\mathrm{E}\left[G_{it}(x) \mid \boldsymbol{A}^N,\boldsymbol{B}^T\right] = 0$ by

construction, and $G_{it}(x)$ therefore plays a role very similar to the error term $E_{it}(x)$. In particular, the conditions in (2.17) can be verified by a weak law of large numbers, as long as $P_{it}^{-1}(x)$ is not too large, and $G_{it}(x)$ is not too strongly correlated across i and over t. Regarding the condition on the spectral norm $\|\boldsymbol{G}(x)\|_\infty = O_P(\sqrt{N+T})$, there are many results in the random-matrix theory literature that show this rate for mean-zero random matrices $\boldsymbol{G}(x)$, see, for example, Geman (1980), Silverstein (1989), Bai et al. (1988), Yin et al. (1988). In particular, if $G_{it}(x)$ is independent across both i and t, then this rate result follows from the very elegant spectral norm inequality in Latała (2005), see the proof of Lemma 2.3.1 in the appendix, where we apply that inequality to $E_{it}(x)$. However, that simple argument would require X_{it} to be independently distributed across i and t, conditional on $\boldsymbol{A}^N$, $\boldsymbol{B}^T$. More generally, we expect $\|\boldsymbol{G}(x)\|_\infty = O_P(\sqrt{N+T})$ to hold whenever the matrix entries $G_{it}(x)$ have zero mean, sufficiently bounded moments, and weak correlation across both i and t, see Section S.2 of the supplementary material of Moon and Weidner (2017) for details.

We have thus shown that consistent estimates for ASFs can be obtained via the matrix completion estimator even if the estimand $\Gamma_{it}^{\infty}(x) = m(x,\boldsymbol{A}_i,\boldsymbol{B}_t)$ itself is not of low rank. This is the main technical result of this paper. However, inference on $\mu(x)$ based on $\widehat{\mu}(x)$ can be problematic, because $\widehat{\mu}(x)$ is subject to both low-rank approximation and shrinkage biases. The low-rank approximation bias is due to the approximation error $\zeta_R(x,\boldsymbol{a},\boldsymbol{b})$ in the decomposition of $m(x,\boldsymbol{a},\boldsymbol{b})$ in equation (2.9). The shrinkage bias comes from bias in $\widehat{\boldsymbol{\Gamma}}(x)$ due to the presence of the nuclear norm penalization in the objective function of (2.15). To isolate this bias, consider a simple case where $Y_{it}(x)$ follows a deterministic pure factor model

$$Y_{it}(x) = \Gamma_{it}(x) = \sum_{j=1}^{R} s_j(x) u_j(x,\boldsymbol{A}_i) v_j(x,\boldsymbol{B}_i).$$

Then, the matrix completion estimator of $\Gamma_{it}(x)$ in (2.15) yields

$$\widehat{\Gamma}_{it}(x) = \sum_{j=1}^{R} [s_j(x) - \rho]_+ u_j(x,\boldsymbol{A}_i) v_j(x,\boldsymbol{B}_i)$$

where $[z]_+ = \max(z,0)$. Compared to $\mathbf{\Gamma}(x)$, $\widehat{\mathbf{\Gamma}}(x)$ has the same eigenvectors but the singular values are shrunk toward zero. This argument carries over to the case where $Y_{it}(x)$ follows an approximate factor structure Cai et al. (2010); Ma et al. (2011); Bai and Ng (2019b). Because of these biases, we explore alternative estimates for $\mu(x)$ in Section 2.4.

2.3.4 Covariates and fixed effects

As we mentioned in Section 2.2, exogenous covariates can be incorporated by conditioning on their values. This method can produce very noisy estimators in small samples unless the covariates take only on few values. Here we consider a semiparametric version of the model that imposes additivity in the effect of the exogenous covariates, which may be continuous, discrete or mixed. It also allows for additive unobserved individual and time effects that might vary across the covariate level x. These effects can be subsumed in the factor structure, but are usually considered separately in empirical analysis as the estimators perform better without regularizing them Athey et al. (2017).

Let $\boldsymbol{C}_{it}$ be a d_c-vector of covariates, $\boldsymbol{\alpha}(x) = (\alpha_1(x), \ldots, \alpha_N(x))$ be a N-vector of individual effects and $\boldsymbol{\delta}(x) = (\delta_1(x), \ldots, \delta_T(x))$ be a T-vector of time effects. Then, we can replace the program (2.13) by

$$\min_{\{\boldsymbol{\beta}\in\mathbb{R}^{d_c}, \boldsymbol{\alpha}\in\mathbb{R}^N, \boldsymbol{\delta}\in\mathbb{R}^T, \mathbf{\Gamma}\in\mathbb{R}^{N\times T}\}} \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{1}\{X_{it}=x\}\left(Y_{it} - \boldsymbol{C}_{it}^{\mathrm{T}}\boldsymbol{\beta} - \alpha_i - \delta_t - \Gamma_{it}\right)^2 + \rho(R_1)\|\mathbf{\Gamma}\|_1,$$

Chernozhukov et al. (2018), Moon and Weidner (2018) and Beyhum and Gautier (2019) provide algorithms to solve this program. Let $\widehat{\boldsymbol{\beta}}(x)$, $\widehat{\boldsymbol{\alpha}}(x) = (\widehat{\alpha}_1(x), \ldots, \widehat{\alpha}_N(x))$, $\widehat{\boldsymbol{\delta}}(x) = (\widehat{\delta}_1(x), \ldots, \widehat{\delta}_T(x))$, and $\widehat{\mathbf{\Gamma}}(x)$ be the solution of the previous program. We can form estimators of the ASF and CASF as

$$\widehat{\mu}_t(x) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{1}\{X_{it}=x\}Y_{it} + \mathbb{1}\{X_{it}\neq x\}\left\{\boldsymbol{C}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(x) + \widehat{\alpha}_i(x) + \widehat{\delta}_t(x) + \widehat{\Gamma}_{it}(x)\right\}\right],$$

and

$$\widehat{\mu}_t(x \mid \{x_0\}) = \frac{\sum_{i=1}^{N}\left[\mathbb{1}\{X_{it}=x_0=x\}Y_{it}+\mathbb{1}\{X_{it}=x_0\neq x\}\left\{\boldsymbol{C}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(x)+\widehat{\alpha}_i(x)+\widehat{\delta}_t(x)+\widehat{\Gamma}_{it}(x)\right\}\right]}{\sum_{i=1}^{N}\mathbb{1}\{X_{it}=x_0\}}.$$

2.4 Debiasing Using Matching Methods

The matrix completion estimator of the ASF is generally biased. As we explained in Section 2.3.3, the bias comes from two sources: low-rank approximation bias and shrinkage bias. One could attempt to correct the shrinkage bias by shifting the singular values of $\widehat{\boldsymbol{\Gamma}}(x)$ upwards. However, inference results on the ASFs based on matrix completion are generally very difficult to obtain even if $\boldsymbol{\Gamma}^{\infty}(x)$ is truly low rank. In our setting, the presence of the additional low-rank approximation bias makes this even more challenging. We instead discuss alternative estimators and show that they have significantly lower biases than the matrix completion estimators in the numerical simulations of Section 2.5.2.

To construct the estimators of $\boldsymbol{\Gamma}^{\infty}(x)$, we start by extracting the factor structure of $\widehat{\boldsymbol{\Gamma}}(x)$ in (2.15). Let $\widehat{\boldsymbol{\lambda}}_i(x)$ and $\widehat{\boldsymbol{f}}_t(x)$ be the $R\times 1$ vectors that satisfy

$$\widehat{\Gamma}_{it}(x) = \widehat{\boldsymbol{\lambda}}_i(x)^{\mathrm{T}}\widehat{\boldsymbol{f}}_t(x),$$

subject to the usual normalizations that $T^{-1}\sum_{t=1}^{T}\widehat{\boldsymbol{f}}_t(x)\widehat{\boldsymbol{f}}_t(x)^{\mathrm{T}}$ is the identity matrix of size R and $N^{-1}\sum_{i=1}^{N}\widehat{\boldsymbol{\lambda}}_i(x)\widehat{\boldsymbol{\lambda}}_i(x)^{\mathrm{T}}$ is a diagonal matrix. Next, we apply a matching procedure to this factor structure. In its simplest version, we estimate each entry $\boldsymbol{\Gamma}_{it}^{\infty}(x)$ such that $X_{it}\neq x$, by matching with the observation with $X_{js}=x$ that is the nearest neighbor in terms of the vectors $\widehat{\boldsymbol{\lambda}}_i(x)$ and $\widehat{\boldsymbol{f}}_t(x)$. In particular, $\breve{\Gamma}_{it}(x) = Y_{i^{**}(i,t,x),t^{**}(i,t,x)}$ where $i^{**}(i,t,x)\in\mathbb{N}$ and $t^{**}(i,t,x)\in\mathbb{T}$ are a solution to the program

$$\begin{aligned}&\min_{j\in\mathbb{N},s\in\mathbb{T}}\left\|\widehat{\boldsymbol{\lambda}}_i(x)-\widehat{\boldsymbol{\lambda}}_j(x)\right\|^2+\left\|\widehat{\boldsymbol{f}}_t(x)-\widehat{\boldsymbol{f}}_s(x)\right\|^2\\ &\text{s.t.}\quad X_{js}=x.\end{aligned}$$

We also consider a two-way matching procedure that combines matching with a difference-in-differences approach. It consists of two steps:

(i) For all $x \in \mathbb{X}$ and $(i,t) \in \mathbb{N} \times \mathbb{T}$ such that $X_{it} \neq x$, find the matches $i^*(i,t,x) \in \mathbb{N}$ and $t^*(i,t,x) \in \mathbb{T}$ that solve the program

$$\min_{j \in \mathbb{N}, s \in \mathbb{T}} \left\| \widehat{\boldsymbol{\lambda}}_i(x) - \widehat{\boldsymbol{\lambda}}_j(x) \right\|^2 + \left\| \widehat{\boldsymbol{f}}_t(x) - \widehat{\boldsymbol{f}}_s(x) \right\|^2$$
$$\text{s.t.} \quad X_{is} = X_{jt} = X_{js} = x.$$

(ii) Estimate $\Gamma_{it}(x)$ by

$$\widetilde{\Gamma}_{it}(x) = Y_{i,t^*(i,t,x)} + Y_{i^*(i,t,x),t} - Y_{i^*(i,t,x),t^*(i,t,x)}.$$

In other words, we find the match (j,s) with $X_{js} = x$ that not only is the closest to (i,t) in terms of the estimated factor structure, but also corresponds to a unit j with $X_{jt} = x$ and a time period s with $X_{is} = x$. Then, we estimate the counterfactual $\Gamma_{it}(x)$ as a linear combination of Y_{jt}, Y_{is} and Y_{js}.

The additional difference-in-differences step in the two-way procedure is useful to reduce bias. To see this, we can compare $\widetilde{\Gamma}_{it}(x)$ with the simple matching estimator $\breve{\Gamma}_{it}(x)$. Thus, abstracting from the estimation error in the factors and loadings,

$$\begin{aligned} \mathrm{E}[\breve{\Gamma}_{it}(x) - \Gamma_{it}(x) \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}] &= m(x, \boldsymbol{A}_{i^{**}(i,t,x)}, \boldsymbol{B}_{t^{**}(i,t,x)}) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t) \\ &= \mathscr{O}_P(\|\boldsymbol{A}_{i^{**}(i,t,x)} - \boldsymbol{A}_i\| + \|\boldsymbol{B}_{t^{**}(i,t,x)} - \boldsymbol{B}_t\|), \end{aligned}$$

by a first-order Taylor expansion of $(\boldsymbol{a}_i, \boldsymbol{b}_t) \mapsto m(x, \boldsymbol{a}_i, \boldsymbol{b}_t)$ around $(\boldsymbol{A}_i, \boldsymbol{B}_t)$; whereas

$$\begin{aligned} \mathrm{E}[\widetilde{\Gamma}_{it}(x) - \Gamma_{it}(x) \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}] &= m(x, \boldsymbol{A}_{i^*(i,t,x)}, \boldsymbol{B}_{t^*(i,t,x)}) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t) \\ &= \mathscr{O}_P(\|\boldsymbol{A}_{i^*(i,t,x)} - \boldsymbol{A}_i\|^2 + \|\boldsymbol{B}_{t^*(i,t,x)} - \boldsymbol{B}_t\|^2), \end{aligned}$$

by a second-order Taylor expansion of $(\boldsymbol{a}_i, \boldsymbol{b}_t) \mapsto m(x, \boldsymbol{a}_i, \boldsymbol{b}_t)$ around $(\boldsymbol{A}_i, \boldsymbol{B}_t)$. The

two-way matching removes the leading term of the Taylor expansion, reducing the bias of the matching by one order of magnitude because $i^{**}(i,t,x) \neq i$ or $t^{**}(i,t,x) \neq t$. On the other hand, $\|\boldsymbol{A}_{i^*(i,t,x)} - \boldsymbol{A}_i\| \geq \|\boldsymbol{A}_{i^{**}(i,t,x)} - \boldsymbol{A}_i\|$ and $\|\boldsymbol{B}_{t^*(i,t,x)} - \boldsymbol{B}_t\| \geq \|\boldsymbol{B}_{t^{**}(i,t,x)} - \boldsymbol{B}_t\|$ a.s. because the two-way procedure imposes the additional restrictions $X_{is} = X_{jt} = x$. Whether the first or second order bias dominates would generally be determined by the proportion of observations with $X_{js} = x$ and the distributions of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$. We provide a numerical comparison of the biases of the matching estimators in Section 2.5.2.

We develop the theory for a debiased estimator that allows for multiple matches and estimated factors and loadings. Multiple matches are expected to reduce dispersion at the cost of increasing bias. Let $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}(x,\boldsymbol{A}_i)$ and $\boldsymbol{f}_t = \boldsymbol{f}(x,\boldsymbol{B}_t)$ be the transformations of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ that are consistently estimated by $\widehat{\boldsymbol{\lambda}}_i$ and $\widehat{\boldsymbol{f}}_t$.[4] We define

$$\mathbb{N}_i = \left\{ j \in \mathbb{N} \setminus \{i\} : \left\| \widehat{\boldsymbol{\lambda}}_i - \widehat{\boldsymbol{\lambda}}_j \right\| \leq \tau_{NT} \right\}, \quad \mathbb{T}_t = \left\{ s \in \mathbb{T} \setminus \{t\} : \left\| \widehat{\boldsymbol{f}}_t - \widehat{\boldsymbol{f}}_s \right\| \leq \upsilon_{NT} \right\},$$

for some bandwidth parameters $\tau_{NT} > 0$ and $\upsilon_{NT} > 0$. The debiased estimator of $\mu(x)$ is then given by

$$\widetilde{\mu}(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{Y}_{it}(x),$$

with

$$\widetilde{Y}_{it}(x) = \begin{cases} Y_{it} & \text{if } X_{it} = x, \\ \dfrac{1}{n_{it}} \displaystyle\sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\}(Y_{is} + Y_{jt} - Y_{js}) & \text{if } X_{it} \neq x \text{ and } n_{it} > 0, \\ \frac{1}{n(x)} \sum_{(j,s) \in \mathbb{D}(x)} Y_{js} & \text{if } n_{it} = 0, \end{cases} \tag{2.18}$$

where $n_{it} := \sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\}$. Here, for $X_{it} \neq x$, we construct

[4] The matching method discussed here is also applicable to settings where the matching is based on variables other than the estimated factor structure. These include for example cross section and time series averages of the observable variables. See the appendix for a more general treatment.

the counterfactual $\widetilde{Y}_{it}(x)$ by averaging over all units $(j,s) \in \mathbb{N}_i \times \mathbb{T}_t$ that satisfy the constraint $X_{is} = X_{jt} = X_{js} = x$. Notice that if $X_{it} \neq x$ and $n_{it} = 0$, then we cannot construct a suitable counterfactual by that method. In that case we assign $\widetilde{Y}_{it}(x)$ the average of the observations with $X_{js} = x$ to make sure that $\widetilde{\mu}(x)$ is always well-defined, but our assumption below guarantees that this rarely happens.

This estimator has similar debiasing properties to the nearest neighbor described above, but it is more tractable theoretically because it varies more smoothly with respect to the factors and loadings.

Indeed, $\widetilde{\mu}(x)$ can be written as

$$\widetilde{\mu}(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_{it} Y_{it},$$

where the weights ω_{it} are functions of $\widehat{\boldsymbol{\lambda}}_j$ and $\widehat{\boldsymbol{f}}_s$ for all $j \in \mathbb{N}$ and $s \in \mathbb{T}$. To show that $\widetilde{\mu}(x)$ is a consistent estimator of $\mu(x)$, we use the following assumption:

Assumption 2.4.1 (Two-way Matching Estimator)**.** There exists a sequence $\xi_{NT} > 0$ such that $\xi_{NT} \to 0$ as $N, T \to \infty$, and

(i) $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{X_{it} \neq x \,\&\, n_{it} = 0\} = O_P(\xi_{NT})$.

(ii) Y_{it} is uniformly bounded over i, t, N, T.

(iii) Y_{it} is independent across both i and t, conditional on $\boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T}$.

(iv) The function $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ is at least twice continuously differentiable with uniformly bounded second derivatives.

(v) There exists $c > 0$ such that $\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \leq c \|\boldsymbol{\lambda}(\boldsymbol{a}_1) - \boldsymbol{\lambda}(\boldsymbol{a}_2)\|$ for all $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{A}$, and $\|\boldsymbol{b}_1 - \boldsymbol{b}_2\| \leq c \|\boldsymbol{f}(\boldsymbol{b}_1) - \boldsymbol{f}(\boldsymbol{b}_2)\|$ for all $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{B}$.

(vi) $\frac{1}{N} \sum_{i=1}^{N} \left(\left\| \widehat{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i \right\|^2 + \max_{j \in \mathbb{N}_i} \left\| \widehat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j \right\|^2 \right) = O_P(\xi_{NT})$.
$\frac{1}{T} \sum_{t=1}^{T} \left(\left\| \widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t \right\|^2 + \max_{s \in \mathbb{T}_t} \left\| \widehat{\boldsymbol{f}}_s - \boldsymbol{f}_s \right\|^2 \right) = O_P(\xi_{NT})$.

(vii) $\tau_{NT}^2 = O_P(\xi_{NT})$ and $\upsilon_{NT}^2 = O_P(\xi_{NT})$.

(viii) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left[\omega_{it}^2 \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] = O_P(NT\,\xi_{NT}^2)$.

(ix) Let $\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}$ be the outcome matrix $\boldsymbol{Y}^{NT}$, but with Y_{it} and Y_{js} replace by zero (or some other non-random number), and all other outcomes unchanged. We assume

$$\frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\{(i,t)\neq(j,s)\}\mathrm{E}\Bigg[\Bigg|\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) - \omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT})\Bigg|\,\Bigg|\,\boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\Bigg] = O_P\left(\xi_{NT}^2\right).$$

Remark 2.4.1 (Assumption 2.4.1)**.** Part (i) guarantees that $X_{it}\neq x$ and $n_{it}=0$ only happens for a small fraction of observations (i,t). We are therefore able to construct proper counterfactuals $\widetilde{Y}_{it}(x)$ for most observations. Part (ii) is a boundedness condition that is standard in the matrix completion literature. Part (iii) is an independence condition that is convenient to simplify the derivations but can be generalized to weak correlation across both i and t. We use part (iv) to bound the error terms of the Taylor expansions for the bias. Part (v) imposes an injectivity condition. The functions $\boldsymbol{a}\mapsto\boldsymbol{\lambda}(\boldsymbol{a})$ and $\boldsymbol{b}\mapsto\boldsymbol{f}(\boldsymbol{b})$ need to be such that $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ can be uniquely recovered from $\boldsymbol{\lambda}_i=\boldsymbol{\lambda}(\boldsymbol{A}_i)$ and $\boldsymbol{f}_t=\boldsymbol{f}(\boldsymbol{B}_t)$. A necessary condition is that the dimensions of $\boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$ are greater than or equal to the dimensions of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$, respectively. This holds in our factor structure approximation when let R grow with the sample size, provided that the dimensions of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are fixed. Part (vi) holds if $\widehat{\boldsymbol{\lambda}}_i-\boldsymbol{\lambda}_i$ and $\widehat{\boldsymbol{f}}_t-\boldsymbol{f}_t$ are of order $N^{-1/2}$ and $T^{-1/2}$. We expect this assumption to be satisfied for rates $\xi_{NT}\gg\max(N^{-1},T^{-1})$. The bandwidth parameters τ_{NT} and υ_{NT} should not be chosen too large according to part (vii). For example, if we want to achieve a rate $\xi_{NT}\ll\max(N^{-1/2},T^{-1/2})$, then we need $\tau_{NT}\ll\max(N^{-1/4},T^{-1/4})$ and $\upsilon_{NT}\ll\max(N^{-1/4},T^{-1/4})$. Part (viii) requires that any given outcome Y_{it} is not chosen too often with too high weight in the construction of the counterfactuals $\widetilde{Y}_{js}(x)$. Finally, part (ix) is a high-level assumption that could be justified by appropriate distributional assumptions on X_{it}, $\boldsymbol{A}_i$, $\boldsymbol{B}_t$, and on the estimators $\widehat{\boldsymbol{\lambda}}_i$ and $\widehat{\boldsymbol{f}}_t$. We prefer to present it as a high-level assumption, because formally working out

the distributional assumptions is quite cumbersome. Intuitively, if n_{it} is sufficiently large, then changing $\boldsymbol{Y}^{NT}$ to $\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}$ should not change the constructions of the counterfactual $\widehat{Y}_{it}(x)$ very much. If that is true for all (i,t), then the weights $\omega_{it}(\boldsymbol{Y}^{NT})$ should be very close to the weights $\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)$ and the assumption is satisfied.

Theorem 2.4.2. *Under Assumptions 2.2.1 and 2.4.1,*

$$\widetilde{\mu}(x) - \mu(x) = O_P(\xi_{NT}).$$

As discussed in the above remark, one can achieve rates

$$\xi_{NT} \ll \max(N^{-1/2}, T^{-1/2})$$

for sufficiently regular data generating processes, and if the bandwidth parameters τ_{NT} and υ_{NT} are chosen sufficiently small. By contrast, the low-rank approximation bias in $\widehat{\mu}(x)$ will usually prevent us from achieving such a convergence rate for $\widehat{\mu}(x)$. This finding is consistent with our Monte Carlo results in Section 2.5.2, where $\widetilde{\mu}(x)$ is found to typically have much smaller bias than $\widehat{\mu}(x)$.

2.5 Numerical Examples

2.5.1 Election day registration and voter turnout

We illustrate the methods of the paper with an empirical application to the effect of allowing voter registration during the election day on voter turnout in the U.S. Xu (2017). Voting in the U.S. used to require registration prior to the election day in most states. Registration increased the cost of voting and was considered as one possible reason for low turnout rates. In response, some states implemented Election Day Registration (EDR) laws that allowed eligible voters to register on election day when they arrive at the polling stations. These laws were not passed by all the states, and there was variation in the time of adoption across states. Thus, they were enacted by Maine, Minnesota and Wisconsin in 1976; Wyoming, Indiana and New Hampshire in 1994, and Connecticut in 2012.

We use a dataset on the 24 presidential elections for 47 states between 1920 and 2012 collected by Xu (2017). It includes state-level information about the turnout rate, Y_{it}, measured as the total ballots counted divided by voting-age population in state i at election t, and a treatment indicator for EDR, X_{it}, that equals one if the state i has an EDR law enacted at election t. Following Xu (2017), we exclude North Dakota where registration was never needed, and Alaska and Hawaii that were not states until 1959. Since there are only 9 states that are ever treated and the treatment started in the 1976 election, we focus on effects on the treated at the elections between 1976 and 2012. We estimate average treatment effects and quantile treatment effects at multiple quantile indices.

Figure 2.1 compares the average turnout of states that are ever treated with states that are never treated in elections prior to the first implementation of the EDR laws in 1976. It shows that ever treated states have higher turnout rates on average than never treated states without the EDR treatment. We consider several methods to deal with this likely nonrandom assignment of EDR to estimate the ATTs for each election after 1976. First, we do a naive comparison of means between treated and nontreated states in each election (Dmeans). Second, we consider a difference-in-differences method that uses the nontreated states as controls at each election (DiD). In particular, we estimate the effects from a linear regression with state effects and election effects interacted with a EDR indicator. This method yields the ATT for each election under a parallel trend assumption between treated and nontreated states.[5] Third, we compute our estimator based on matrix completion methods without debiasing (MC) with additive state and election effects and the parameter ρ such that the number of factors is $R = 6$. Fourth, we debias the MC estimates using the two-way matching method with 10 matches (TWM-10). Fifth, we consider the simple matching method with 5 matches (SM-5). We choose the number of matches roughly based on the numerical simulations of Section 2.5.2.

Figure 2.2 reports the estimates of the ATT of EDR at each election. The methods that account for possible nonrandom assignment of the EDR produce lower

[5]The DiD model is a special case our model with additive effects. In this case, it imposes that there are only additive state and election effects that are the same for both treatment levels.

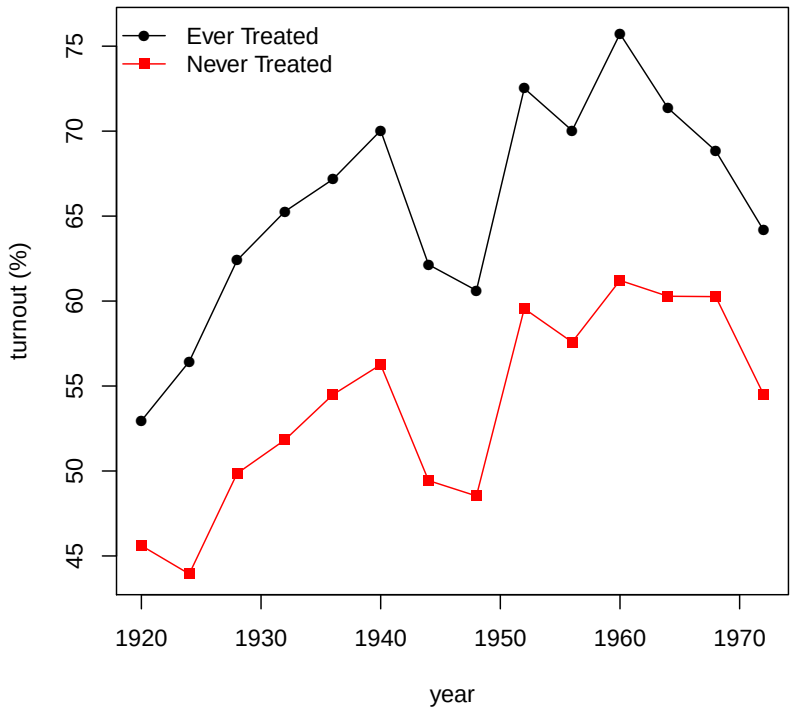

Figure 2.1: Pretrends in turnout rate

estimates of the effect than the naive comparison of means between treated and non-treated states. This finding agrees with the pre-EDR differences found in fig. 2.1. MC, TWM-10 and SM-5 estimates are generally larger and more stable across elections than DiD estimates. According to TWM-10, EDR laws increase voter turnout between 5 and 9% depending on the election. This effect is an economically significant relative to 55%, the average turnout rate for states without EDR. The estimates of the election-aggregated ATTs are 10.71%, 0.67%, 7.35%, 5.56%, and 4.87% for Dmeans, DiD, MC, TWM-10, and SM-3, respectively.

Figure 2.3 plots the estimates of the election-aggregated quantile treatment effect on the treated (QTT) of EDR as a function of the quantile index. We report estimates from four methods: a naive comparison of quantiles between treated and non-treated states (Dquantiles), our estimator based on matrix completion methods without debiasing (MC) with additive state and election effects and the parameter ρ such that the number of factors is $R = 3$, two-way matching with 10 matches

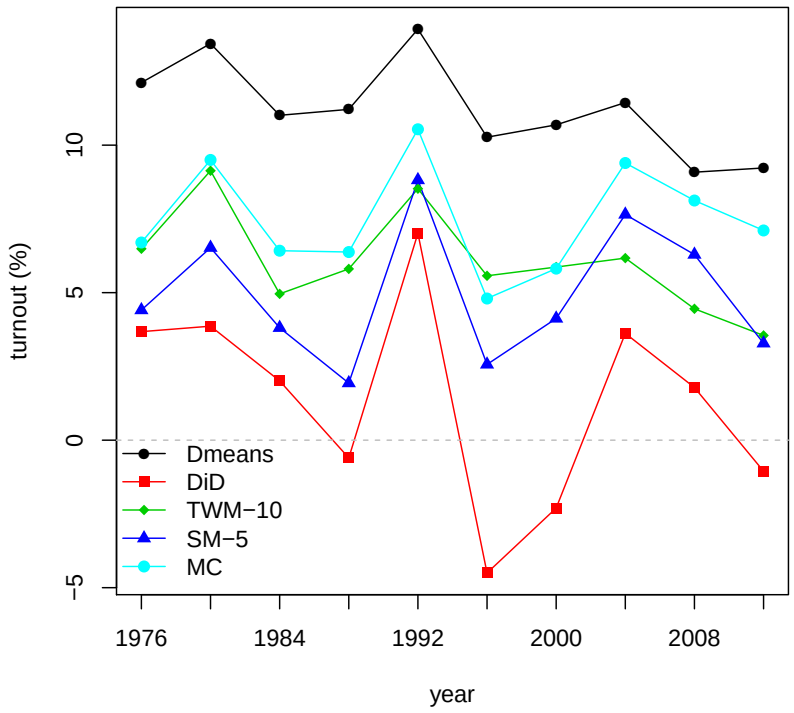

Figure 2.2: Average treatment effect on treated

(TWM-10), and simple matching with 5 matches (SM-5). The QTT is the difference of the quantiles between the observed turnout for the treated observations and the corresponding potential turnout have they not been treated. The quantiles of the observed turnout are estimated using sample quantiles. The estimates of the quantiles of the potential outcomes are obtained by inverting the corresponding estimates of the distribution, which are obtained by our methods replacing Y_{it} by the indicator $\mathbb{1}(Y_{it} \leq y)$ and repeating the procedure over a grid of values of y that includes the sample quantiles of observed turnout with indices $\{.10, .11, \ldots, .98\}$.[6] Here, we find that the effect of EDR is decreasing across the distribution of turnout and ranges between 10 and 0% according to TWM-10. EDR is therefore more effective at the bottom of the voter turnout distribution. Comparing with the Dquantiles estimates, we find that the sign of the selection bias switches from positive to negative around

[6] We rearrange the estimates of the distribution to guarantee that they are increasing with respect to y Chernozhukov et al. (2010).

the middle of the turnout distribution.

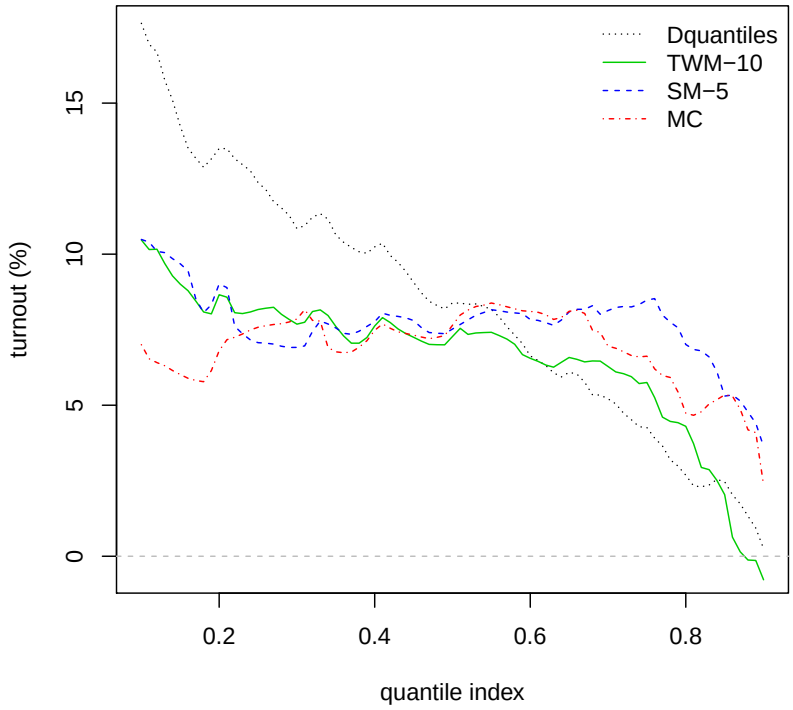

Figure 2.3: Time-averaged QTT

2.5.2 Monte Carlo simulations

To evaluate the performance of our methods in a controlled synthetic environment, we generate potential outcomes from an additive linear model where

$$Y_{it}(x) = x + g(A_i, B_t) + U_{it}(x), \quad x \in \{0,1\}, i \in \{1,\ldots,30\}, t \in \{1,\ldots,30\},$$

$U_{it}(x) \sim N(0, 1/4)$ independently over i, t and x, $A_i \sim U(0,1)$ independently over i, $B_t \sim U(0,1)$ independently over t, $U_{it}(x)$, A_j and B_s are independent for all i, t, j and s, and g is the Gaussian kernel, i.e.,

$$g(a,b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-b)^2}{\sigma^2}\right).$$

This design is similar to that used in Bordenave et al. (2020), with kernel function specification from the numerical simulations in Griebel and Harbrecht (2013).[7] The parameter σ controls the decay of the singular values of g and can be calibrated to make sure the singular values decay slowly. Smaller values for σ lead to greater dispersion in the kernel function $(a,b) \mapsto g(a,b)$ and a slower singular value decay, hence can be interpreted as a measure of smoothness.[8] The assignment of X_{it} that determines what potential outcomes are observed is similar to the election application. In particular, only observations for the first half of the units, $i \in \{1,\dots,15\}$, and the second half of the panel, $t \in \{15,\dots,30\}$, may be treated. For these observations, X_{it} is related to the unobserved effects (A_i, B_t) via $X_{it} = \mathbb{1}\{g(A_i,B_t) \geq c\}$, where c is a constant calibrated to $\Pr(g(A_i,B_t) \geq c) = .5$.

Table 2.1: Results for $\mu(0 \mid \{1\})$

	Bias	St. Dev.	RMSE
Dmeans	0.59	0.02	0.59
DiD	0.70	0.03	0.70
MC	0.74	0.02	0.74
TWM-1	0.03	0.14	0.14
TWM-5	0.03	0.11	0.12
TWM-10	0.04	0.10	0.11
TWM-30	0.07	0.09	0.12
SM-1	0.12	0.10	0.16
SM-5	0.15	0.07	0.17
SM-10	0.19	0.06	0.20
SM-30	0.31	0.05	0.31

Notes: based on $1,000$ simulations

We apply similar methods to Section 2.5.1 to estimate the CASFs $\mu_t(0 \mid \{1\})$, $t \in \{15,\dots,30\}$, and $\mu(0 \mid \{1\})$ using the observed variables X_{it} and $Y_{it} = Y_{it}(X_{it})$. Thus, we consider Dmeans, DiD, MC without additive effects and with the parameter ρ such that $R = 5$, and multiple versions of TWM and SM with the number of matches equal to 1, 5, 10, and 30. For each method, we compute the bias, standard deviation and rmse from $1,000$ simulations. Across the simulations, we redraw the

[7] We find similar results in a multiplicative model where $Y_{it}(x) = (1+x)g(A_i,B_t) + U_{it}(x)$. We omit these results for the sake of brevity.

[8] Smoothness here is specifically related to numerical smoothness, i.e. variability in the function within close neighbourhoods of its arguments.

values of $U_{it}(x)$ and hold A_i, B_t and X_{it} fixed. Table 2.1 reports the results for the time-aggregated CASF, $\mu(0 \mid \{1\})$, and Figure 2.4 plots the results for the CASF, $\mu_t(0 \mid \{1\})$, as a function of t. The results show that Dmeans, DiD and MC are severely biased relative to their standard deviations. All the matching estimators reduce bias and rmse, despite of increasing dispersion. As one would expect, increasing the number of matches reduces the variability of the matching estimators but increases their biases. The number of matches that minimizes the rmse is larger for the TWM than for the SM. Overall, these small-sample findings agree with the asymptotic results of Sections 2.3.3 and 2.4.

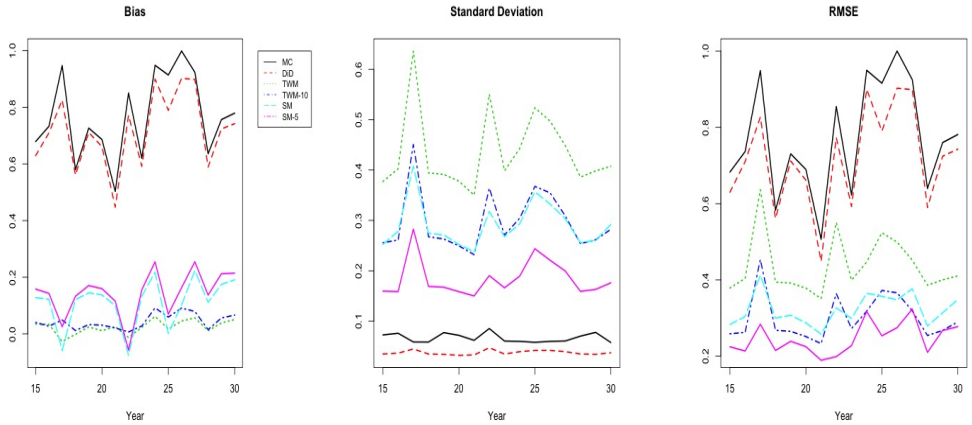

Figure 2.4: Results for $t \mapsto \mu_t(0 \mid \{1\})$.

Acknowledgements

This paper was prepared for the Econometrics Journal Special Session on "Econometrics of Panel Data" at the Royal Economic Society 2019 Annual Conference in Warwick University. We thank the editor Jaap Abbring, two anonymous referees, Shuowen Chen, and the participants of this conference and the 25th International Panel Data Conference for comments. This research was supported by the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and by the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

Chapter 3

Linear Panel Regressions with Two-Way Unobserved Heterogeneity

3.1 Introduction

We consider the following panel data model for $i = 1, \ldots, N$ cross-sectional units, and $t = 1, \ldots, T$ time periods,

$$Y_{it} = X_{it}'\beta + u_{it}, \qquad u_{it} = h(\alpha_i, \gamma_t) + \varepsilon_{it}, \tag{3.1}$$

where Y_{it} is an observed dependent variable, $X_{it} = (X_{it,1}, \ldots, X_{it,K})'$ is a K-vector of observed explanatory variables, and u_{it} is an unobserved error term. Within the unobserved error term, we have an unknown real-valued function $h(\cdot,\cdot)$ that depends on the (vector-valued) unobserved fixed effects $\alpha_i \in \mathbb{R}^{d_\alpha}$ and $\gamma_t \in \mathbb{R}^{d_\gamma}$, which are allowed to be arbitrarily correlated with the observed regressors X_{it}, while ε_{it} is a mean-zero error term that is uncorrelated with X_{it}. Our focus is on estimation of and inference on the parameter $\beta \in \mathbb{R}^K$ — the regression coefficient of X_{it} on Y_{it} when properly controlling for the unobserved α_i and γ_t.

The key model restrictions in (3.1) are the linearity in X_{it} as well as the additive separability between $X_{it}'\beta$ and u_{it}. If the unobserved error term u_{it} is of the more general form $u_{it} = g(\alpha_i, \gamma_t, \xi_{it})$, for some idiosyncratic errors ξ_{it} that are identically distributed across i and over t, and independent of the covariates X_{it}, then under appropriate regularity conditions we can define $h(\alpha_i, \gamma_t) = \mathbb{E}\left[u_{it} \,\middle|\, \alpha_i, \gamma_t\right]$ and

$\varepsilon_{it} = u_{it} - h(\alpha_i, \gamma_t)$ to again obtain model (3.1). The additive separability between $h(\alpha_i, \gamma_t)$ and ε_{it} is therefore not strictly required. However, throughout this paper we take the representation of the model in (3.1) as the starting point for our analysis.

Analogous to the singular value decomposition of a *matrix*, there exists, under weak regularity conditions, the singular value decomposition of a *function* $h : \mathbb{R}^{d_\alpha} \times \mathbb{R}^{d_\gamma} \to \mathbb{R}$, which reads

$$h(\alpha, \gamma) = \sum_{r=1}^{\infty} \sigma_r \, \varphi_r(\alpha) \, \psi_r(\gamma), \tag{3.2}$$

for some functional singular values $\sigma_r > 0$, and appropriate normalized functions $\varphi_r : \mathbb{R}^{d_\alpha} \to \mathbb{R}$ and $\psi_r : \mathbb{R}^{d_\gamma} \to \mathbb{R}$, $r \in \{1, 2, 3, \ldots\}$. Equation (3.2) allows us to rewrite model (3.1) as

$$Y_{it} = X_{it}' \beta + \sum_{r=1}^{\infty} \lambda_{ir} \, f_{tr} + \varepsilon_{it}, \tag{3.3}$$

with $\lambda_{ir} := \sigma_r \, \varphi_r(\alpha_i)$ and $f_{tr} := \psi_r(\gamma_t)$. Thus, our model can be viewed as a linear panel regression model with unobserved "factor structure" or "interactive fixed effects", but where the number of factors f_{tr} and corresponding factor loadings λ_{ir} is infinite. The same rewriting of a function $h(\alpha_i, \gamma_t)$ by an infinite sum $\sum_{r=1}^{\infty} \lambda_{ir} \, f_{tr}$ is used in Menzel (2021), but for a different model, and with the goal of analyzing the bootstrap for multidimensional data.

Within a panel regression context, most of the existing literature assumes that the number of unobserved factors is finite, which, from our perspective, corresponds to a truncation of the infinite sequence of factors in (3.3), that gives

$$Y_{it} = X_{it}' \beta + \sum_{r=1}^{R} \lambda_{ir} \, f_{tr} + e_{it}, \tag{3.4}$$

where $e_{it} := \varepsilon_{it} + \sum_{r=R+1}^{\infty} \lambda_{ir} \, f_{tr}$. The interactive fixed effect model in (3.4) is one possible approximation of the model (3.1) that we explore in this paper, and we will show that this approximation can be used to estimate β consistently. However, we also explore another approximation of $h(\alpha_i, \gamma_t)$ using two-way grouped fixed

effects, see Section 3.2.2 below, and we also derive convergence rate results for the resulting grouped fixed effect estimator. Other approximation methods for $h(\alpha_i, \gamma_t)$ are also conceivable, but are not explored in this paper.[1]

For datasets with both N and T large, the two currently dominant estimation methods for the panel regression model in (3.4) are the common correlated effect (CCE) estimator of Pesaran (2006) and the least-squares (LS) estimator (also called quasi maximum likelihood estimator) in Bai (2009). Since those original papers by Pesaran and Bai, a large literature has emerged that has extended the CCE and LS estimation methods, and has analyzed the properties of those estimators in more general settings — see Chudik and Pesaran (2013), Bai and Wang (2016), and Karabiyik et al. (2019) for recent surveys. We follow that literature here by also considering panels with both N and T large, that is, for our asymptotic results we consider $N, T \to \infty$.[2]

The "conventional" interactive fixed effect model in (3.4) is a special case of our model (3.1), with $\alpha_i = \lambda_i = (\lambda_{i1}, \ldots, \lambda_{iR})'$, $\gamma_t = f_t = (f_{t1}, \ldots, f_{tR})'$, and $h(\alpha_i, \gamma_t) = \lambda_i' f_t$. The key question that we ask in this paper is what happens when the multiplicative factor structure $\lambda_i' f_t$ is replaced by a more general non-linear factor structure $h(\alpha_i, \gamma_t)$. However, we do maintain all other assumptions of model (3.4), in particular, the homogenous regression coefficient β, and the additive separability between $X_{it}'\beta$ and the unobserved error.

The main challenge that we need to tackle when considering this extension is that, if the data generating process is given by (3.1), then the error term e_{it} in (3.4) will generally be correlated with X_{it}, because e_{it} contains the truncated part $\sum_{r=R+1}^{\infty} \lambda_{ir} f_{tr}$ of the infinite factor structure,[3] and $\lambda_{ir} = \varphi_r(\alpha_i)$ and $f_{tr} = \psi_r(\gamma_t)$ are

[1] For example, to justify (3.2) we rely on the paper by Griebel and Harbrecht (2014), which also discusses the alternative "sparse grid" approximation. In our context, the sparse grid approximation would correspond to replacing $\sum_{r=1}^{R} \lambda_{ir} f_{tr}$ by $\sum_{r,q=1}^{R} \gamma_{rq} \lambda_{ir} f_{tq}$, with some sparsity condition on the matrix $\gamma = (\gamma_{rq})$.

[2] There is of course also work on model (3.4) in the context of short T panels, for example, Holtz-Eakin et al. (1988), Ahn et al. (2001, 2013), Sarafidis and Robertson (2009) Juodis and Sarafidis (2018, 2022), Westerlund et al. (2019),

[3] Notice that the majority of these truncated factors will be "weak", see Onatski Onatski (2010, 2012) and Chudik, Pesaran and Tosetti Chudik, Pesaran, and Tosetti (2011a) for the distinction between "strong" and "weak" factors.

functions of α_i and γ_t, which can be correlated with X_{it}. Once e_{it} is correlated with X_{it} in this way, then the existing results for the CCE and the LS estimator are not applicable anymore. The currently known results on the CCE and LS estimator in the presence of an infinite number of factors (e.g. Pesaran and Tosetti 2011, Chudik et al. 2011b, and Westerlund and Urbain 2013) require that the "unaccounted" factors $\sum_{r=R+1}^{\infty} \lambda_{ir} f_{tr}$ are uncorrelated with the regressors, so that they can be considered part of the error term e_{it} without generating an endogeneity problem.

For the case that X_{it} and e_{it} are correlated, there exist instrumental variable (IV) generalizations of both the CCE and LS method (e.g. Harding and Lamarche 2011, Lee et al. 2012, Robertson and Sarafidis 2015, Moon et al. 2018, and Norkutė et al. 2021), but those require observed instruments Z_{it} that are uncorrelated with e_{it}. We do not explore instrumental variable approaches in this paper.

The two main theoretical contributions of our paper are as follows: Firstly, we formally show that the LS estimator of Bai (2009) can still provide consistent estimates of β in model (3.1), as long as the number of factors $R = R_{NT}$ used in estimation grows to infinity jointly with N and T (a similar asymptotic with growing number of factors is considered in Beyhum and Gautier 2022). Secondly, we suggest an alternative estimator for β, which we denote the *two-way group fixed-effect estimator* (generalizing ideas in Bonhomme et al. 2021 on the discretization of one-way heterogeneity), and we provide conditions under which this new estimator is $\sqrt{NT}$-consistent as $N, T \to \infty$. In addition, we also suggest inference procedures using both of these estimators, but we do not formally derive inference results in this paper. Instead, we study the properties of our suggested confidence intervals in Monte Carlo simulations. We also apply the estimators to an empirical application on UK house price data.

When employing the LS estimator with factors from Bai (2009) to model (3.1), we are effectively estimating a misspecified model — the DGP is given by (3.1), but the estimating equation by (3.4). Galvao and Kato (2014) and Juodis (2020) have recently studied linear panel regression models with additive fixed effects under misspecification. We consider interactive fixed effects for estimation here, and the

type of misspecification we allow for is more restrictive. We therefore do not have to introduce any pseudo-true parameter, but we find that the LS estimator is still consistent for the true value of β under our assumptions.

It also natural to ask if our non-linear model $h(\alpha_i, \gamma_t)$ is truly necessary, and also if there is a way to test whether a more standard additive or multiplicative error component structure would be sufficient to capture unobserved heterogeneity. For example, Kapetanios et al. (2019) provide a test for whether the multiplicative error component structure is necessary or whether a simpler two-way fixed effect estimator would be sufficient. In many applications they find evidence that the standard two-way fixed effect should work well without the need for interactive fixed-effects. However, we do not pursue such a testing approach here, because if the main goal is inference on β, then size distortions due to pre-testing quickly become a concern (see e.g. Guggenberger 2010). Instead, our recommendation for applied researcher is to report two-way fixed effect estimates jointly with factor augmented estimates and grouped fixed effect estimates in one table that is then subjected to human interpretation.

In related work, allowing for the number of factors to grow with sample size has been considered in Li et al. (2017a), where they explicitly detail a factor model with the number of factors growing with sample size. The difference to this paper is our model admits an infinite number of factors even in small samples and considers finite factor estimation as an approximation to the true data generating process.

There also exist other work on non-linear generalizations of the interactive fixed effect and factor model specification. Zeleneev (2020) considers the same model (3.1) in the context of network data, but in his baseline discussion, the outcome Y_{ij} (instead of Y_{it} here) is symmetric in i and j. The main difference to our work, however, is that Zeleneev estimates the model based on a strategy that identifies agents with similar fixed effect values based on the distribution of their outcomes. His estimation method is accordingly also completely different to ours.

Bodelet and Shan (2020) also consider non-linear functions in place of the standard linear factor model. In our notation, their model assumes a series of

smooth univariate functions of the form $\sum_{q=1}^{Q} h_{iq}(\gamma_{tq})$ for unobserved heterogeneity. Their approach models individual specific responses to structural shocks but is different to our approach, which uses a homogeneous bivariate function. Therefore, their approach allows for discontinuities across how individual effects are modelled whereas our assumption is more restrictive since variation across individuals, via α_i, must be smooth.

Other papers on unobserved two-way heterogeneity in panel or network models either make more parametric assumptions (e.g. Graham 2017, Dzemski 2019, Chen et al. 2020), or employ stochastic block or graphon models (e.g. Holland et al. 1983, Wolfe and Olhede 2013, Gao et al. 2015, Auerbach 2019), and are therefore less closely related to our paper.

There are also recent papers that use matrix completion methods for the purpose of treatment effect estimation in panel models with two-way heterogeneity, e.g. Athey et al. (2017) and Amjad et al. (2018), Chernozhukov et al. (2020), and Fernández-Val et al. (2021). Those papers do not require the additive separability between the regressors and error term in (3.1), but as a result they also have to make stronger assumptions and employ more complicated estimation methods than we do here. The same is true for Freyberger (2017), who considers a non-separable model with interactive fixed effects. Alternative non-linear extensions of factor models are discussed, for example, in Cunha et al. (2010) and Gunsilius and Schennach (2019).

The rest of the paper is organized as follows. Section 3.2 introduces our suggested estimators and inference methods. Section 3.3 and Section 3.4 provide asymptotic results for the LS estimator of Bai (2009) and for our new two-way group fixed-effect estimator, respectively. Section 3.5 discusses the practical implementation. Monte Carlo simulations are presented in Section 3.6, and an empirical application is worked out in Section 3.7.

3.2 Estimation approaches

In this section, we introduce the two estimation approaches that are afterwards analyzed and used in the rest of the paper.

3.2.1 Least-squares interactive fixed effect estimator

Following Bai (2009) we consider

$$\left(\widehat{\beta}_{\mathrm{LS}}, \widehat{\lambda}, \widehat{f}\right) = \underset{(\beta,\lambda,f)\in\mathbb{R}^{K+N\times R+T\times R}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - X_{it}'\beta - \sum_{r=1}^{R}\lambda_{ir} f_{tr}\right)^2. \tag{3.5}$$

This estimator was introduced for the exact factor model in equation (3.4), and Bai (2009) shows that it is $\sqrt{NT}$-consistent and asymptotically normally distributed for $N,T\to\infty$ when the true number of factors is fixed and known. Moon and Weidner (2015) extend this result to the case where the true number of factors is chosen too large in the estimation.

To make the estimates $\widehat{\lambda}$ and $\widehat{f}$ in (3.5) unique, we choose the usual normalization $T^{-1}\widehat{f}'\widehat{f} = \mathbb{I}_R$, and $\widehat{\lambda}'\widehat{\lambda}$ to be a diagonal matrix. In addition, it is convenient to introduce the notation $X\cdot\beta$ for the $N\times T$ matrix with elements $X_{it}'\beta$.

As explained above, the model (3.1) that we consider in this paper can be rewritten as the factor model in (3.3) with an infinite number of factors in the true data generating process. This suggests that the least-squares estimators in (3.5) can still be consistent as long as the number of factors $R = R_{NT}$ used in the estimation is allowed to grow to infinity jointly with N and T. Estimation of $\left(\widehat{\beta}_{\mathrm{LS}}, \widehat{\lambda}, \widehat{f}\right)$ is done using an iterative scheme. That is, we start by initialising $\widehat{\beta}_{\mathrm{LS}}$, and then iterate between estimating the principal components of $Y - X\cdot\widehat{\beta}_{\mathrm{LS}}$ to obtain $\left(\widehat{\lambda}, \widehat{f}\right)$ and least squares of $Y = X\cdot\beta + \widehat{\lambda}\widehat{f}' + e$ to obtain $\widehat{\beta}_{\mathrm{LS}}$. The convergence metric we use is the sum of squares in (3.5). However, this iteration scheme can converge to a local minimum, and it is therefore important to repeat the procedure with multiple starting values of β. For more details on the numerical computation of the estimator in (3.5) we refer to Bai (2009) and Moon and Weidner (2015).

This least-squares estimator of Bai (2009) is very well-established in the panel regression literature. It is used regularly both in empirical and in methodological papers, e.g. Su and Chen (2013), Kim and Oka (2014), Lu and Su (2016), Gobillon and Magnac (2016b), Totty (2017), Su and Wang (2017), Moon and Weidner (2017), Giglio and Xiu (2021), to name just a few.

3.2.2 Group fixed effects estimator

Here, we introduce two-way grouped fixed effects estimator, which discretizes the unobserved heterogeneity that is parameterized by α_i and γ_t in the spirit of Bonhomme et al. (2021). We first describe the main idea of this estimator before explaining its practical implementation in more details.

3.2.2.1 Main idea

We partition the set $\{1,\dots,N\}$ of cross-sectional units into $G = G_{NT}$ groups such that individuals in the same group have similar values of α_i. Let $g_i \in \{1,\dots,G\}$ denote the group membership of individual i. Analogously, we partition the set $\{1,\dots,T\}$ of time periods into $C = C_{NT}$ groups such that time periods in the same group have similar values of γ_t. Let $c_t \in \{1,\dots,C\}$ denote the group membership of time period t. Details on how we construct those partitionings in practice are described below. Notice that within each group the values of the α_i and γ_t, respectively, need not be the same, but in the asymptotic theory in Section 3.4 the differences of those fixed effects within each group are asymptotically negligible.

Once we have obtained those groups, then we estimate β by applying pooled OLS to the linear fixed-effect model

$$Y_{it} = X_{it}'\beta + \delta_{i,c_t} + \nu_{t,g_i} + \varepsilon_{it}, \tag{3.6}$$

where $\delta_{i,c_t} \in \mathbb{R}$ and $\nu_{t,g_i} \in \mathbb{R}$ are nuisance parameters that are jointly estimated with β, that is, the basic two-way grouped fixed effect estimator for β can be written as

$$\widehat{\beta}_{\mathrm{G}} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \min_{\delta\in\mathbb{R}^{N\times C}} \min_{\nu\in\mathbb{R}^{T\times G}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(Y_{it} - X_{it}'\beta - \delta_{i,c_t} - \nu_{t,g_i}\right)^2. \tag{3.7}$$

Notice that within each pair of groups for i and t, that is, for fixed values of c_t and g_i, the model in (3.6) is simply a standard additive two-way fixed effect model $Y_{it} = X_{it}'\beta + \delta_i + \nu_t + \varepsilon_{it}$. However, as the group membership changes we allow the parameters δ_i and ν_t to change arbitrarily, as indicated by the additional subscripts c_t and g_i in (3.6). We could have written $\delta_{i,g_i,c_t} + \nu_{t,g_i,c_t}$ to indicate explicitly that both

the individual and time effect are allowed to change across groups, but the notation in (3.6) of course already allows for that generality. The parameters δ therefore form an $N \times C$ matrix, while the parameters ν form a $T \times G$ matrix.

In the introduction, we explained how the LS-estimator with interactive effects can be justified for model (3.1) by a truncation of the functional singular value expansion in (3.2). In other words, a particular approximation of the function $h(\alpha_i, \gamma_t)$ naturally leads to the estimator in (3.5).

The grouped fixed effect estimator in (3.7) can be justified analogously by a different approximation of the function $h(\alpha_i, \gamma_t)$. Under appropriate regularity conditions, by a joint Taylor expansion in α_i and γ_t around the corresponding group means $\overline{\alpha}_{g_i} = \frac{\sum_{j=1}^n \mathbb{1}\{g_i=g_j\}\,\alpha_j}{\sum_{j=1}^n \mathbb{1}\{g_i=g_j\}}$ and $\overline{\gamma}_{c_t} = \frac{\sum_{s=1}^T \mathbb{1}\{c_t=c_s\}\,\gamma_s}{\sum_{s=1}^T \mathbb{1}\{c_t=c_s\}}$, we find that

$$h(\alpha_i, \gamma_t) = \delta_{i,c_t} + \nu_{t,g_i} + O\left(\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2\right), \tag{3.8}$$

where for vectors $\|\cdot\|$ denotes the Euclidean norm, and

$$\delta_{i,c_t} := h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t}) + \frac{\partial h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t})}{\partial \alpha_i'}(\alpha_i - \overline{\alpha}_{g_i}), \quad \nu_{t,g_i} := \frac{\partial h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t})}{\partial \gamma_t'}(\gamma_t - \overline{\gamma}_{c_t}).$$

This shows that the leading order dependence of $h(\alpha_i, \gamma_t)$ on α_i and γ_t can be described by the additive specification $\delta_{i,c_t} + \nu_{t,g_i}$ used in (3.6). Since this two-way grouped fixed effect ignores the terms $O\left(\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2\right)$ entirely, it is of course crucial to construct the groups such that $\alpha_i - \overline{\alpha}_{g_i}$ and $\gamma_t - \overline{\gamma}_{c_t}$ are small. The clustering algorithm that we use to achieve that is described in Subsection 3.2.2.2 below.

Notice that a naive application of Bonhomme et al. (2021) to our two-way fixed effect model would *not* result in our estimating equation (3.6) but in $Y_{it} = X_{it}'\beta + \chi_{g_i,c_t} + \varepsilon_{it}$, where $\chi_{g,c}$ is a fixed effect specific to each pair of groups $(g,c) \in \{1,\ldots,G\} \times \{1,\ldots,C\}$. The analog of equation (3.8) for that alternative approach

reads

$$h(\alpha_i,\gamma_t) = \underbrace{h(\overline{\alpha}_{g_i},\overline{\gamma}_{c_t})}_{=\chi_{g_i,c_t}} + O\left(\|\alpha_i-\overline{\alpha}_{g_i}\|+\|\gamma_t-\overline{\gamma}_{c_t}\|\right),$$

that is, the approximation error would be of linear order in the discrepancies $\alpha_i - \overline{\alpha}_{g_i}$ and $\gamma_t - \overline{\gamma}_{c_t}$ within groups. By contrast, for our estimating equation (3.6) the resulting approximation error in (3.8) is of quadratic order, which explains why we prefer that approach.

Finally, notice that if our original model would only contain individual specific fixed effects α_i, that is, $Y_{it} = X_{it}'\beta + h(\alpha_i) + \varepsilon_{it}$, then the analog of (3.6) is the standard additive fixed effect model $Y_{it} = X_{it}'\beta + \delta_i + \varepsilon_{it}$, which requires no grouping at all, and also entails no approximation error since we can set $\delta_i = h(\alpha_i)$. The way in which we generalize the grouping ideas in Bonhomme et al. (2021) is therefore quite specific to the two-way fixed effect model in (3.1).

3.2.2.2 Hierarchical clustering algorithm

To make the two-way grouped fixed effect estimator in (3.7) operational we employ the following three-step algorithm:

A. Obtain the factor loading and factor estimates $\widehat{\lambda}$ and $\widehat{f}$ of the interactive fixed effect LS estimator in (3.5) for a relatively large number of factors R. Only keep the leading few R^* factor loading and factor estimates and denote those by $\widehat{\lambda}^* = (\widehat{\lambda}_{ir} : i = 1,\ldots,N,\ r = 1,\ldots,R^*)$ and $\widehat{f}^* = (\widehat{f}_{tr} : t = 1,\ldots,T,\ r = 1,\ldots,R^*)$.

B. Use the $\widehat{\lambda}_1^*,\ldots,\widehat{\lambda}_N^*$ as inputs into the clustering algorithm in Table 3.1 to partition the set of individuals $\{1,\ldots,N\}$. This algorithm returns the number G of chosen groups and the group membership $g_i \in \{1,\ldots,G\}$ of each individual. Analogously, we use the inputs $\widehat{f}_1^*,\ldots,\widehat{f}_T^*$ into the same algorithm to partition $\{1,\ldots,T\}$, resulting in the number of groups C and the group membership $c_t \in \{1,\ldots,C\}$ for each time period. Notice that the words partition, cluster, and group are used interchangeably in this paper.

C. Calculate the two-way grouped fixed effect estimator $\widehat{\beta}_{\mathrm{G}}$ via pooled OLS ac-

Algorithm

1: Input $\widehat{\lambda}_i^* \in \mathbb{R}^{R^*}$ for all $i = 1,\dots N$. Calculate all pairwise Euclidean distances $A_{ij} = \left\|\widehat{\lambda}_i^* - \widehat{\lambda}_j^*\right\|$, for $i \neq j$, and set $A_{ii} = \infty$. Initialize $\mathscr{P} = \{\{1\},\{2\},\dots,\{N\}\}$ as a partition of $\{1,\dots,N\}$.
2: **if** $\exists \mathscr{C}_* \in \mathscr{P}$ with $|\mathscr{C}_*| = 4$ **then** for that $\mathscr{C}_*$
3: Find the solution to

$$\min_{\{i,j,l,m:\mathscr{C}_*=\{i,j,l,m\}\}} A_{ij} + A_{lm},$$

and split $\mathscr{C}_*$ into $\{i,j\}$ and $\{l,m\}$, updating the partition $\mathscr{P}$.
4: **else if** $\exists \mathscr{C} \in \mathscr{P}$ with $|\mathscr{C}| = 1$ **then**
5: Find the solution to

$$\min_{\{i \in \cup_{\{\mathscr{C}\in\mathscr{P}:|\mathscr{C}|=1\}}\}} \min_{\{j \in \cup_{\{\mathscr{C}\in\mathscr{P}:|\mathscr{C}|\leq 3\}}\}} A_{ij},$$

and merge the clusters containing i and j into a single cluster, updating the partition $\mathscr{P}$.
6: **end if**
7: Repeat 2-6 until $\{|\mathscr{C}| : \mathscr{C} \in \mathscr{P}\} \subset \{2,3\}$.

Table 3.1: Hierarchical clustering with minimum single linkage.

cording to equation (3.7).

It is constructive to briefly describe our algorithm from Table 3.1 in words before we discuss features of this whole procedure. Step 1 defines the proxy variable to cluster on ($\widehat{\lambda}_i^*$ in this instance) and sets the distance metric we wish to use, Euclidean distance, which could easily be changed to another norm or metric. Then, we initialise each individual into their own cluster. Steps 2 and 3 then splits any groups of four into two groups of two, since we want groups of no larger than three in our final output.[4] The optimisation in Step 3 looks at all combinations of two by two splits within this group of four and takes the smallest sum of distances. This type of optimisation is only suitable for very small groups of individuals because it is a combinatorially hard problem.

Steps 4 and 5 then finds the solitary individual with the smallest distance to any other existing cluster and merges it to that cluster. Combined with Steps 2 and 3 we create an iteration that merges single clusters one at a time to groups of one, two or three, then splits any groups of four as and when they occur. This means Step 2 can only ever return one group of four. Doing this iteration one at a time is important so

[4]We avoid singleton groups, because for those groups the within transformation removes all information of the data. The restriction to groups of at most size three is somewhat arbitrary, but we want to maintain small group sizes to guarantee that the differences in the fixed effects within each group are small, and there is no incidental parameter problem for the linear fixed effect model in (3.6).

that we may split these groups of four immediately and have a larger choice set in Step 5 for each unmatched individual. Also, splitting groups of four into two by two groups rather than groups of one and three avoids infinite iterations. The repetition of Steps 2-5 is guaranteed to converge, and delivers a partition of $\{1,\dots,N\}$ into groups of size two or three.

Now to discuss the procedure as a whole. The choice of R in step A here is not too important since we only need this to generate proxy variables for clustering and otherwise dispose of β estimates from this initial LS step. The important hyperparameter is the number of proxies per observation, R^*, which we choose equal to two to five. We discuss the theoretical properties of the hyperparameter in Section 3.4.1 but here outline a heuristic approach to this choice. Choosing R^* to be more than one is important to capture cases when α_i and γ_t have higher dimension or when the function $h(.,.)$ admits eigenfunctions that are not individually injective maps from α_i or γ_t. The aim is that a linear combination of non-injective maps provides a better mapping to the closeness of the primitives α_i and γ_t. An archetypal example of this is discussed in Griebel and Harbrecht (2014) where they show that the first few eigenfunctions of the exponential kernel are individually clearly not injective maps.

It is also important to not use too many proxies so as to avoid clustering on noise. This can make for poor matches that result in large deviations between α_i and α_j, respectively γ_t and γ_s, that show up in the leading $O\left(\|\alpha_i-\alpha_j\|^2\right)$ and $O\left(\|\gamma_t-\gamma_s\|^2\right)$ remainder terms in (3.8). Maintaining closeness in these primitives when clustering is key to any argument using Taylor's theorem, however, optimising this proxy hyperparameter is still rough and does require further development. We defer discussion about the presence of noise in factors with relation to the LS estimator to Section 3.3.

There are, of course, other choices for proxies such as the cross-sectional moments employed in Bonhomme et al. (2021). However, as displayed in (3.2) and formulated in Griebel and Harbrecht (2014), using the eigenfunctions from the singular value decomposition are a more natural choice since these are direct functions

of the primitives α_i and γ_t and should in theory lead to closer proximity between these. Since we require cross-sectional and time-dependent clusters for our method, these eigenfunctions also provide a convenient means to find these. If one truly believes that other proxy variables have more precise injectivity with these primitives then they could always make those the the input to Step 1 in our clustering algorithm.

Another divergence from the existing literature is the use of clusters of size two or three, rather than letting these cluster sizes grow with sample size. Our motivation for using these small cluster sizes comes directly from the within-group β estimation, i.e. that we do not need consistent estimates of δ or ν since these are treated as nuisance parameters that are simply differenced out. Hence, for our purposes, it is more useful to have small groups that are very similar rather than to have large groups that have better central tendency estimates. This very conveniently removes one choice for the analyst, namely the setting of group sizes G or C.

This procedure is also a departure from the k-means approach taken in Bonhomme et al. (2021). For example, k-means with $k \approx N/2$ or $k \approx N/3$ only requires group sizes to be 2 or 3 on average. This allows for a large heterogeneity in group sizes, which we avoid with our hierarchical approach. Considering the distance metric in our algorithm is interchangeable, we expect our method to produce similar allocations to a k-means approach that manually limits cluster sizes to 2 or 3.

Other cluster methods also exist. For example, in the presence of heterogeneous coefficients β_i, Su et al. (2016) and Su et al. (2019) propose clustering on β_i. The procedure proposed there may suggest useful ways to incorporate heterogeneity in the slope coefficients in this setting, or indeed provide good cluster proxies for the unobserved heterogeneity term. It should be noted, however, that in those settings there exists a true group structure, which departs from our approach that considers groups as useful discretisations of the underlying parameter space.

3.2.2.3 Split-sample version of the estimator

As explained above, we estimate the group memberships g_i and c_t that enter into the estimator for β in (3.7) via a clustering method applied to $\widehat{\lambda}^*$ and $\widehat{f}^*$. However,

clustering in this way creates dependence across i and t through $\widehat{\lambda}^*$ and $\widehat{f}^*$. This dependence creates technical difficulties when establishing asymptotic convergence results. To mitigate this dependence we augment the clustering estimator by a simple sample splitting method. The resulting group fixed effect estimator with sample splitting is given by

$$\widehat{\beta}_{\text{GS}} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \min_{\delta} \min_{\nu} \sum_{i=1}^{N}\sum_{t=1}^{T}\left[Y_{it} - X_{it}'\beta - \sum_{s=1}^{S} \mathbb{1}\{(i,t)\in\mathscr{O}_s\}\left(\delta^{(s)}_{i,c_t^{(s)}} + \nu^{(s)}_{t,g_i^{(s)}}\right)\right]^2, \tag{3.9}$$

where S is the number of partitions, and the sets $\mathscr{O}_s$, $s=1,\ldots,S$, are the partitions of the sample space $\{1,\ldots,N\}\times\{1,\ldots,T\}$, that is, the observation (i,t) is a member of the s'th partition if and only if $(i,t)\in\mathscr{O}_s$. Compared to the original group fixed effect estimator in (3.6), the group membership indicators $g_i^{(s)}$ and $c_t^{(s)}$ and the group fixed effect $\delta^{(s)}_{i,c_t^{(s)}}$ and $\nu^{(s)}_{t,g_i^{(s)}}$ are all specific to the partition s. For the purpose of this paper, we choose the number of partitions to be $S=4$ and we split the sample space into four blocks as follows:

$$\begin{aligned}
\mathscr{O}_1 &= \{1,\ldots,\lfloor N/2\rfloor\}\times\{1,\ldots,\lfloor T/2\rfloor\},\\
\mathscr{O}_2 &= \{1,\ldots,\lfloor N/2\rfloor\}\times\{\lfloor T/2\rfloor+1,\ldots,T\},\\
\mathscr{O}_3 &= \{\lfloor N/2\rfloor+1,\ldots,N\}\times\{1,\ldots,\lfloor T/2\rfloor\},\\
\mathscr{O}_4 &= \{\lfloor N/2\rfloor+1,\ldots,N\}\times\{\lfloor T/2\rfloor+1,\ldots,T\},
\end{aligned} \tag{3.10}$$

where $\lfloor\cdot\rfloor$ is the floor function.

We still need to explain how the group memberships $g_i^{(s)}$ and $c_t^{(s)}$ are obtained here. The aim of the sample splitting is to avoid any stochastic dependence between $g_i^{(s)}$ and $c_t^{(s)}$ and the idiosyncratic noise ε_{it}. For each partition $s=1,\ldots,S$, we therefore construct the group memberships $g_i^{(s)}$ and $c_t^{(s)}$ without using outcomes Y_{it}

for observations (i,t) of that partition $\mathscr{O}_s$. For that purpose, we define the sets

$$\begin{aligned}
\mathscr{O}_1^* &= \{1,\ldots,N\} \times \{1,\ldots,\lfloor T/2 \rfloor\}, \\
\mathscr{O}_2^* &= \{1,\ldots,N\} \times \{\lfloor T/2 \rfloor + 1,\ldots,T\}, \\
\mathscr{O}_3^* &= \{1,\ldots,\lfloor N/2 \rfloor\} \times \{1,\ldots,T\}, \\
\mathscr{O}_4^* &= \{\lfloor N/2 \rfloor + 1,\ldots,N\} \times \{1,\ldots,T\},
\end{aligned} \tag{3.11}$$

and for $\tilde{s} = 1,\ldots,4$, we define the corresponding least-squares factor and loading estimates

$$\left(\widehat{\lambda}^{(\tilde{s})}, \widehat{f}^{(\tilde{s})}\right) = \underset{(\lambda,f)\in\mathbb{R}^{N_{\tilde{s}}^* \times R + T_{\tilde{s}}^* \times R}}{\operatorname{argmin}} \min_{\beta\in\mathbb{R}^K} \sum_{(i,t)\in\mathscr{O}_{\tilde{s}}^*} \left(Y_{it} - X_{it}'\beta - \sum_{r=1}^{R} \lambda_{ir} f_{tr}\right)^2, \tag{3.12}$$

which is simply the LS estimator in (3.5) applied only to the $N_{\tilde{s}}^* \times T_{\tilde{s}}^*$ subpanel of observations $(i,t) \in \mathscr{O}_{\tilde{s}}^*$, and we also impose the same normalization on the factors and loadings explained after (3.5).[5] Now, for the original partition $\mathscr{O}_s$, $s = 1,\ldots 4$, we construct the group membership $g_i^{(s)}$ of unit i by applying the clustering algorithm in Table 3.1 to the loading estimates $\widehat{\lambda}_i^{(\tilde{s})}$ obtained from the subpanel $\mathscr{O}_{\tilde{s}}$ with $\tilde{s} = \tilde{s}(s)$ given by

$$\tilde{s} = \begin{cases} 2 & \text{for } s = 1, \\ 1 & \text{for } s = 2, \\ 2 & \text{for } s = 3, \\ 1 & \text{for } s = 4. \end{cases}$$

Analogously, for the partition $\mathscr{O}_s$, $s = 1,\ldots 4$, we construct the group membership $c_t^{(s)}$ of time period t by applying the clustering algorithm in Table 3.1 to the factor

[5] Notice that factor model proxies can only be used to compare observations from the same factor estimation sample space. This is because factors are only identified up to rotations, where these rotations may differ across estimation samples.

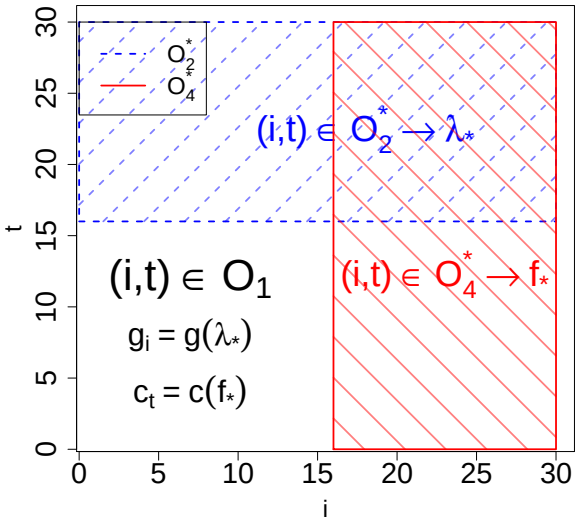

Figure 3.1: Sample split for partition 1

estimates $\widehat{f}_t^{(\tilde{s})}$ obtained from the subpanel $\mathscr{O}_{\tilde{s}}$ with $\tilde{s} = \tilde{s}(s)$ given by

$$\tilde{s} = \begin{cases} 4 & \text{for } s = 1, \\ 4 & \text{for } s = 2, \\ 3 & \text{for } s = 3, \\ 3 & \text{for } s = 4. \end{cases}$$

Figure 3.1 details an example of this sample splitting technique for clustering within partition $\mathscr{O}_1$. Here we see clearly how the partitions for proxy estimation $\mathscr{O}_2^*$ and $\mathscr{O}_4^*$ do not overlap with the partition we are grouping within, $\mathscr{O}_1$. This guarantees that we do not introduce any dependence between the group functions and the noise term by making sure grouping within each partition is not a function of the independent noise term, ε_{it}, from observations within that partition. This becomes important in our derivations in Section 3.4.2, where we require that the process $X_{it}\,\varepsilon_{it}$ remains zero mean and independently distributed after group means are projected out.

With these cluster assignments it then becomes straightforward to estimate (3.9) by first taking within-cluster mean-differences for each partition and then simply apply pooled OLS on the transformed variables.

Notice that by allowing the partitioning in (3.11) used to estimate proxy variables to extend over the whole sample of either N or T, we get better estimates than just using the original partition (3.10). As discussed earlier, it is crucial to avoid poor initial estimates of proxy variables to better approximate the residual terms in the Taylor expansion in expression (3.8).

3.3 Asymptotic results for the least squares estimator

Here, we derive convergence rate results for the least-squares estimator (3.5) for a data generating process given by (3.1). Thus, we generalize the consistency results in Bai (2009) and Moon and Weidner (2015) to the case where the underlying panel regression model does not satisfy the factor model in (3.4). However, as explained in the introduction, the factor model in (3.4) can be viewed as an approximation of (3.1), and this approximation idea can be formalized asymptotically, as long as we allow the number of factors $R = R_{NT}$ used in the least-squares estimator (3.5) to grow with N and T.

3.3.1 Consistency and convergence rate

From now on, we denote the true parameter β that generates the data by β^0. We rewrite model (3.1) as

$$Y_{it} = X_{it}' \beta^0 + \Gamma_{it} + \varepsilon_{it}, \tag{3.13}$$

where both Γ_{it} and ε_{it} are unobserved. Our main convergence rate results in Theorem 3.3.5 actually hold for any $N \times T$ matrix $\Gamma = (\Gamma_{it})$ that satisfies Assumption 3.3.4 below, but ultimately we are of course interested in the case $\Gamma_{it} = h(\alpha_i, \gamma_t)$. Arbitrary dependence between X_{it} and Γ_{it} is allowed for, so there is a potential endogeneity problem.

Remember that the components of the K-vector X_{it} are denoted by $X_{it,k}$, $k = 1, \ldots, K$. Let $X_k = (X_{it,k})$ and $\varepsilon = (\varepsilon_{it})$ be $N \times T$ matrices. For a matrix A we denote r'th largest singular value by $\sigma_r(A)$, that is, $\sigma_r^2(A)$ is equal to the r'th largest eigenvalue of AA'. Furthermore, for matrices we denote the spectral norm by $\|\cdot\|$,

and for vectors the norm $\|\cdot\|$ denotes the Euclidean norm. We write wpa1 for "with probability approaching one". We impose the following assumptions.

Assumption 3.3.1 (Bounded norms of X_k and ε).

(i) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it,k})^2 = O_P(1), \quad \text{for } k = 1,\dots,K.$

(ii) $\|\varepsilon\| = O_P\left(\sqrt{\max\{N,T\}}\right).$

Assumption 3.3.2 (Weak Exogeneity of X_k). $\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}\varepsilon_{it} = O_P(\sqrt{NT}), \quad \text{for } k = 1,\dots,K.$

Assumption 3.3.3 (Non-Collinearity of X_k). Consider linear combinations $\delta \cdot X := \sum_{k=1}^{K}\delta_k X_k$ of the regressors X_k with vectors $\delta \in \mathbb{R}^K$ such that $\|\delta\| = 1$. Assume that there exists a constant $b > 0$ such that

$$\min_{\{\delta\in\mathbb{R}^K,\,\|\delta\|=1\}} \sum_{r=2R_{NT}+1}^{\min(N,T)} \sigma_r^2\left[\frac{(\delta\cdot X)}{\sqrt{NT}}\right] \geq b\,, \qquad \text{wpa1.}$$

Assumption 3.3.4 (Singular value decay). There exists a constant $\rho > 3/2$ such that

$$\frac{1}{NT}\sum_{r=R_{NT}+1}^{\min(N,T)} \sigma_r^2(\Gamma) = O_P\left(R_{NT}^{1-2\rho}\right).$$

Here, $R = R_{NT}$ is the number of factors that is chosen in the computation of the least-squares estimator $\widehat{\beta}_{\mathrm{LS}}$ in (3.5). We require $R_{NT} \to \infty$ as $N,T \to \infty$ to obtain consistency of $\widehat{\beta}_{\mathrm{LS}}$.

Lemma 3.3.1 below justifies Assumption 3.3.4 for our main case of interest $\Gamma_{it} = h(\alpha_i,\gamma_t)$, and we therefore postpone the discussion of that assumption until we discuss that lemma. Assumptions 3.3.1-3.3.3 are very similar to the assumptions used in Bai (2009) and Moon and Weidner (2015) to show consistency of $\widehat{\beta}_{\mathrm{LS}}$,[6] and the following discussion of those assumptions will, accordingly, be brief.

[6]Compared to the assumptions imposed in the consistency Theorem 4.1 of Moon and Weidner (2015), the only two differences are that we allow for R_{NT} to grow asymptotically, and that Assumption 3.3.1(i) requires a bound on the Frobenius norm $\|X_k\|_F := \left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}^2\right)^2$ instead of a bound on the spectral norm $\|X_k\|$. Since $\|X_k\| \leq \|X_k\|_F$, our assumption here is technically stronger, but in practice, one likely will justify any bound on $\|X_k\|$ using the inequality $\|X_k\| \leq \|X_k\|_F$ anyway.

Assumption 3.3.1(i) follows from Markov's inequality as long as the second moment of $X_{it,k}$ is uniformly bounded. Assumption 3.3.1(ii) follows, for example, from the inequality in Latala (2005) if ε_{it} has mean zero, uniformly bounded fourth moment, and is independent across i and t. However, the assumption still holds if ε_{it} is weakly correlated across i and over t, see Moon and Weidner (2015). Assumption 3.3.2 is satisfied as long as $X_{it}\varepsilon_{it}$ has zero mean, uniformly bounded second moment, and is weakly correlated across i and over t.

To understand Assumption 3.3.3, notice first that for $R_{NT} = 0$ the expression $\sum_r \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}}\right]$ in that assumption becomes

$$\sum_{r=1}^{\min(N,T)} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}}\right] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\delta \cdot X)_{it}^2.$$

Thus, for $R_{NT} = 0$, the assumption is just a standard non-collinearity assumption on the regressors, which demands that every non-trivial linear combination $\delta \cdot X$ of the regressors has sufficient variation. Next, for $R_{NT} > 0$ we have

$$\sum_{r=2R_{NT}+1}^{\min(N,T)} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}}\right] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\delta \cdot X)_{it}^2 - \sum_{r=1}^{2R_{NT}} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}}\right],$$

that is, the assumption demands that the variation in the linear combination $\delta \cdot X$ does not only come from the leading $2R_{NT}$ singular values of this linear combination.

Of course, if $\text{rank}(\delta \cdot X) \leq 2R_{NT}$, then for $r > 2R_{NT}$ all the singular values $\sigma_r(\delta \cdot X)$ are equal to zero and the assumption is violated. Thus, a necessary condition for Assumption 3.3.3 is that $\text{rank}(\delta \cdot X) > 2R_{NT}$, that is, any linear combination of the regressors needs to be a "high-rank matrix". For example, a constant regressor $X_{it,1} = 1$ violates this assumption (it constitutes a rank one matrix, which could be easily absorbed into the unobserved Γ_{it}), but if the regressors are drawn from a DGP with random variation across both i and t, then they typically have full rank. Again, we refer to the existing papers on the least-squares estimator with interactive fixed effects for further discussion of this generalized non-collinearity condition on

the regressors.

Theorem 3.3.5 (Consistency of $\widehat{\beta}_{\mathrm{LS}}$). *Let Assumptions 3.3.1 – 3.3.4 hold, and furthermore assume that $R_{NT} = o(\min\{N,T\})$ as $N,T \to \infty$. Then we have*

$$\widehat{\beta}_{LS} - \beta^0 = O_P\Big(R_{NT}^{(3-2\rho)/2}\Big) + O_P\Big(R_{NT}\left(\min\{N,T\}\right)^{-1/2}\Big). \tag{3.14}$$

Therefore, by choosing $R_{NT} \propto (\min\{N,T\})^{\frac{1}{2\rho-1}}$ we obtain that

$$\widehat{\beta}_{LS} - \beta^0 = O_P\Big(\min\{N,T\}^{\frac{3-2\rho}{2(2\rho-1)}}\Big).$$

Assumption 3.3.4 demands $\rho > 3/2$, and the first term on the right-hand side of (3.14) is therefore decreasing in the number of factors R_{NT} used for estimation. By contrast, the second term on the right-hand side of (3.14) is increasing in R_{NT}. The final part of the theorem simply gives the rate for R_{NT} that optimally balances the trade-off between those two terms. This is analogous the bias-variance trade-off for bandwidth selection in non-parametric estimation. Indeed, the term $O_P\left(R_{NT}^{(3-2\rho)/2}\right)$ is due to the approximation error of the $N \times T$ matrix Γ (which can have full rank) by only a finite number of factors (of rank only R_{NT}). As expected, the approximation error is small when choosing a more flexible model (large R_{NT}).

The second term on the right-hand side of (3.14) also occurs when one considers a conventional interactive fixed effect model, where the true matrix Γ itself is assumed to have low rank and the approximation error is therefore not present (for $R_{NT} \geq \mathrm{rank}(\Gamma)$). For that case, the first paper to derive the large N, T asymptotic properties for $\widehat{\beta}_{\mathrm{LS}}$ was Bai (2009). He imposes assumptions (in particular, $R_{NT} = \mathrm{rank}(\Gamma) =$ constant, and all factors in Γ are "strong factors") that are strong enough to derive the result $\widehat{\beta}_{\mathrm{LS}} - \beta^0 = O_P(1/\sqrt{NT})$ when N and T grow at the same rate.[7] However, without such strong assumptions, the estimator $\widehat{\beta}_{\mathrm{LS}}$ may very well converge at a slower rate. For example, in Section 4.3 of Moon and Weidner (2015) a concrete data generating process is given for which $\widehat{\beta}_{\mathrm{LS}}$ only converges

[7] For general sequences of $N,T \to \infty$ one finds $\widehat{\beta}_{\mathrm{LS}} - \beta^0 = O_P(1/N + 1/T)$ under Bai's assumptions.

at the slower rate $(\min\{N,T\})^{-1/2}$.[8] The key difference between that example and Bai (2009) is that $R_{NT} > \text{rank}(\Gamma)$, that is, the number of factors in the estimation is larger than the true number of factors. More generally, as soon as the "strong factor" assumption or the known number of factors assumption ($R_{NT} = \text{rank}(\Gamma)$) are violated, there is no guarantee that $\widehat{\beta}_{\text{LS}}$ converges at the fast rate derived in Bai (2009). In the absence of those assumptions, Theorem 4.1 in Moon and Weidner (2015) shows that $\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\Big((\min\{N,T\})^{-1/2}\Big)$ when $R_{NT} \geq \text{rank}(\Gamma)$ is fixed. The second term on the right-hand side of (3.14) exactly generalizes that rate to the case where R_{NT} is allowed to grow asymptotically.

In our setting, we cannot impose the "strong factor" or known number of factor assumptions in Bai (2009), because, as explained in the introduction, the data generating process $\Gamma_{it} = h(\alpha_i, \gamma_t)$ typically generates an infinite sequence of factors of decreasing strength. Demanding all those factors in equation (3.3) to be strong factors makes no sense in our setting. Deriving a convergence rate for $\widehat{\beta}_{\text{LS}}$ faster than $(\min\{N,T\})^{-1/2}$ in our model therefore appears to very challenging, to say the least. This is of course, the key motivation for why we also consider the two-way grouped fixed effect estimator in this paper, see Section 3.4 below.

Remark 3.3.1. *If we change Assumption 3.3.4 to*

$$\sigma_r(\Gamma) \leq c\sqrt{NT}\, r^{-\rho}, \tag{3.15}$$

for all $r \in \{R_{NT}+1, \ldots \min\{N,T\}\}$, wpa1, and some constant $c > 0$, then the result in equation (3.14) *of Theorem 3.3.5 can be improved to*

$$\widehat{\beta}_{LS} - \beta^0 = O_P\left(R_{NT}^{1-\rho}\right) + O_P\left(R_{NT}\,(\min\{N,T\})^{-1/2}\right),$$

and we can then obtain consistency of $\widehat{\beta}$ under the weaker condition $\rho > 1$. Condition (3.15) *implies Assumption 3.3.4, but not vice versa, because Assumption 3.3.4 is a condition on the sum of the squared singular values, not on each of the singu-*

[8] In that example, the unnecessarily estimated loadings $\widehat{\lambda}$ and factors $\widehat{f}$ are correlated with the regressors, and by controlling for such endogenous $\widehat{\lambda}$ and $\widehat{f}$ one ends up reducing the convergence rate of $\widehat{\beta}_{\text{LS}}$ from $\sqrt{NT}$ to $(\min\{N,T\})^{1/2}$.

lar values separately. It turns out to be technically much easier to verify Assumption 3.3.4 than to verify (3.15) *for our main case of interest* $\Gamma_{it} = h(\alpha_i, \gamma_t)$,[9] *as we do in Lemma 3.3.1 below. This explains why we have chosen that formulation of the assumption and theorem in our baseline presentation.*

Despite the technical subtleties explained in the preceding remark, one should still interpret Assumption 3.3.4 as imposing a particular decay rate for the singular values Γ, as in display (3.15) of the remark. Thus, the leading few singular value can have a magnitude of $\sqrt{NT}$, as would be the case under the "strong factor assumption" in the usual interactive fixed effects model of Bai (2009). However, as N, T, r all converge to infinity we require the $\sigma_r(\Gamma)$ to converge at the polynomial rate $r^{-\rho}$ in order to satisfy the summability condition in Assumption 3.3.4.

The results in this section so far have not made any use of the structure $\Gamma_{it} = h(\alpha_i, \gamma_t)$. Theorem 3.3.5 is applicable to any other data generating process for Γ that satisfies Assumption 3.3.4. A full-rank matrix Γ satisfying that assumption could, for example, also be generated by a dynamic factor model (see e.g. Forni et al. 2000, 2005, Stock and Watson 2002).[10]

In the following we now focus exclusively on the case $\Gamma_{it} = h(\alpha_i, \gamma_t)$. The following lemma provides conditions on the function $h(\cdot,\cdot)$ that guarantee that Assumption 3.3.4 is satisfied.

Lemma 3.3.1. *Assume* $\alpha_i \in \Omega_\alpha$ *and* $\gamma_t \in \Omega_\gamma$*, and that* $h : \Omega_\alpha \times \Omega_\gamma \to \mathbb{R}$ *is* p *times continuously differentiable in both arguments, with uniformly bounded mixed-derivatives up to order p, and the domains* $\Omega_\alpha \subset \mathbb{R}^{n_\alpha}$ *and* $\Omega_\gamma \subset \mathbb{R}^{n_\gamma}$ *are smooth and bounded. Then for* $\Gamma_{it} = h(\alpha_i, \gamma_t)$ *Assumption 3.3.4 is satisfied for* $R_{NT} \to \infty$ *with* $\rho = \frac{p}{\min\{n_\alpha, n_\gamma\}}$.

Here, we measure the smoothness of the function $h(\cdot,\cdot)$ by p, which is the number of times it is continuously differentiable. The decay rate ρ of the singular values of Γ then depends on this measure of smoothness and the dimensions n_α and

[9] This is because not only the decay of $\sigma_r(\Gamma)$ as $r \to \infty$ needs to be controlled, but also the convergence rate of the expressions as $N, T \to \infty$.

[10] One can generate an infinite number of "static factors", as in (3.3), via a dynamic factor model with a finite number of dynamic factors.

n_γ of the arguments α_i and γ_t. The smoother the function $h(\cdot,\cdot)$, for fixed dimensions n_α and n_γ, the faster the eigenvalues of Γ converge to zero.

The proof of Lemma 3.3.1 crucially relies on the functional singular value decomposition in (3.2) and results on the decay rate of the corresponding singular values in Griebel and Harbrecht (2014). The only technical contribution of the proof is then to properly relate those known results on the functional singular value to the matrix singular values of Γ.

Notice that Lemma 3.3.1 requires no assumptions on the data generating process of α_i and γ_t, apart from boundedness of the domains Ω_α and Ω_γ, which can always be achieved by a reparameterization. Thus, those nuisance parameters can be arbitrarily correlated with each other (across i and over t) and with the regressors $X_{it,k}$. This result is analogous to the consistency Theorem 4.1 for $\widehat{\beta}_{\mathrm{LS}}$ in Moon and Weidner (2015), where also no assumptions on the interactive fixed effects are imposed at all, apart from $\mathrm{rank}(\lambda f') \leq R$.

From Theorem 3.3.5 and Lemma 3.3.1 we have the following corollary.

Corollary 3.3.1. *Let Assumptions 3.3.1 – 3.3.3 and the assumption on $h(.,.)$ in Lemma 3.3.1 be satisfied with $p > 3\min\{n_\alpha, n_\gamma\}/2$, and also let $R_{NT} \to \infty$ such that $R_{NT}/(\min\{N,T\})^{1/2} \to 0$. Then we have*

$$\widehat{\beta}_{LS} - \beta^0 = o_P(1).$$

This is our final consistency result for the least-squares estimator of Bai (2009) in a data generating process given by (3.1). The convergence rate of the estimator was already discussed after Theorem 3.3.5 above, in particular, the difficulty in showing a convergence rate faster than $(\min\{N,T\})^{1/2}$ in our setting.

3.3.2 Further discussion

Here, we want to present some further intuition on the formal results on $\widehat{\beta}_{\mathrm{LS}}$ presented above. The discussion in this subsection is purely heuristic and does not aim to provide any formal derivations.

Remember the functional singular value decomposition in equation (3.2) of

the introduction, which we now write as $h(\alpha_i, \gamma_t) = \sum_{r=1}^{\infty} \lambda_{ir}^0 f_{tr}^0$. For the sake of the following discussion, suppose that variation from $h(\alpha_i, \gamma_t)$ dominates the variation in $X_{it}'\beta$ and ε_{it} for the leading R_{NT} principal components of the residuals $Y_{it} - X_{it}'\beta - \sum_{r=1}^{R} \lambda_{ir} f_{tr} = \sum_{r=1}^{\infty} \lambda_{ir}^0 f_{tr}^0 - X_{it}'(\beta - \beta^0) + \varepsilon_{it}$. In this "best case scenario", the estimated factors $\sum_{r=1}^{R} \lambda_{ir} f_{tr}$ in the definition of $\widehat{\beta}_{\mathrm{LS}}$ in (3.5) will coincide with the leading R_{NT} components $\sum_{r=1}^{R} \lambda_{ir}^0 f_{tr}^0$ of $h(\alpha_i, \gamma_t)$, and we then have

$$\widehat{\beta}_{\mathrm{LS}} - \beta^0 = \zeta_{NT} + \xi_{NT},$$

where

$$\zeta_{NT} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' X_{it} \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' \varepsilon_{it}$$
$$\xi_{NT} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' X_{it} \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' \sum_{r=R+1}^{\infty} \lambda_{ir}^0 f_{tr}^0.$$

Under standard regularity conditions we have $\sqrt{NT}\,\zeta_{NT} \Rightarrow \mathcal{N}(0, \Sigma)$, and under the assumptions in the last subsection we have $\xi_{NT} = O_P\left(R_{NT}^{(3-2\rho)/2}\right)$. In this "best-case scenario" we can therefore have $R_{NT} \to \infty$ quick enough such that $\xi_{NT} = o_P(1/\sqrt{NT})$.

However, this is not a realistic scenario for $R_{NT} \to \infty$, because as R_{NT} grows, eventually the singular values of ε_{it} will dominate those of $\sum_{r=R+1}^{\infty} \lambda_{ir}^0 f_{tr}^0$, and the factor projection method will just project out idiosyncratic noise, or even contributions from $X_{it}'(\widehat{\beta}_{\mathrm{LS}} - \beta^0)$. This implies that the problematic variation associated with $\lambda_{ir}^0 f_{tr}^0$ for most singular values r remains. This explains why it is so difficult to show anything better than the convergence rate results in Theorem 3.3.1 for the estimator $\widehat{\beta}_{\mathrm{LS}}$ in our setting.

3.4 Asymptotic results for group fixed-effect estimator

The main goal of this section is to derive asymptotic results for the estimator $\widehat{\beta}_{\text{GS}}$ defined in (3.9), which is the sample-splitting version of the group fixed-effect estimator. But we are first going to discuss the initial group fixed-effect estimator $\widehat{\beta}_{\text{G}}$ defined in (3.7) without sample-splitting. We will not actually derive convergence rate results for $\widehat{\beta}_{\text{G}}$ itself, but the discussion of the approximation bias of $\widehat{\beta}_{\text{G}}$ will be a very useful precursor of the results for $\widehat{\beta}_{\text{GS}}$.

3.4.1 Results for $\widehat{\beta}_{\text{G}}$

We can rewrite our estimating equation for the group fixed-effect estimator in (3.6) as

$$Y = X \cdot \beta + \delta D_{\delta}' + D_{\nu} \nu' + \varepsilon, \tag{3.16}$$

where δ and ν are the $N \times C$ and $T \times G$ matrices of nuisance parameters, while D_{δ} and D_{ν} are $T \times C$ and $N \times G$ are binary matrices in which each row contains a single one, indicating the group membership of the corresponding unit or time period, respectively. By standard partitioned regression results we can then rewrite the group fixed-effect estimator in (3.7) as

$$\widehat{\beta}_{\text{G}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{Y}_{it}, \quad \widetilde{X}_k = M_N X_k M_T, \quad \widetilde{Y} = M_N Y M_T, \tag{3.17}$$

where $\widetilde{X}_{it} = \left(\widetilde{X}_{it,1}, \ldots, \widetilde{X}_{it,K} \right)$, $\widetilde{Y}_{it}$ and $\widetilde{X}_{it,k}$ are the entries of the $N \times T$ matrices $\widetilde{X}_k$ and $\widetilde{Y}$, respectively, and $M_N = \mathbb{I}_N - D_{\nu}(D_{\nu}'D_{\nu})^{-1}D_{\nu}'$ and $M_T = \mathbb{I}_T - D_{\delta}(D_{\delta}'D_{\delta})^{-1}D_{\delta}'$ are projection matrices of dimesion $N \times N$ and $T \times T$, respectively.

Using this representation of the group fixed-effect estimator and the model in (3.13) we obtain that

$$\widehat{\beta}_{\text{G}} - \beta^0 = \phi_{NT} + \kappa_{NT}, \tag{3.18}$$

where

$$\phi_{NT} := \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \varepsilon_{it}, \quad \kappa_{NT} := \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{\Gamma}_{it}, \tag{3.19}$$

with $\widetilde{\Gamma}$ defined analogously to $\widetilde{X}_k$ and $\widetilde{Y}$ in (3.17). In the definition of ϕ_{NT} we can equivalently write $\widetilde{\varepsilon}_{it}$ instead of ε_{it}, but since M_N and M_T are idempotent matrices, and $\widetilde{X}_{it}$ is already the projected regressor, this does not matter. The same is true, of course, for $\widetilde{\Gamma}_{it}$ vs Γ_{it} in the definition of κ_{NT}. However, the expressions in (3.19) turn out to be convenient as written.

Here, κ_{NT} is the approximation error of having replaced the nonlinear specification $\Gamma_{it} = h(\alpha_i, \gamma_t)$ in our model in (3.1) by the much simpler additive specification $\delta_{i,c_t} + \nu_{t,g_i}$ in the estimation equation (3.6). To see this, we can use standard matrix inequalities to bound the Euclidian norm of κ_{NT} by

$$\|\kappa_{NT}\| \leq \left\| \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \right\| \left(\max_k \left\| \widetilde{X}_k \right\|_F \right) \left\| \widetilde{\Gamma} \right\|_F, \tag{3.20}$$

where $\|\cdot\|_F$ refers to the Frobenius norm. Due to the definition of M_N and M_T we have

$$\left\| \widetilde{\Gamma} \right\|_F^2 = \min_{\delta \in \mathbb{R}^{N \times C}} \min_{\nu \in \mathbb{R}^{T \times G}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[h(\alpha_i, \gamma_t) - \delta_{i,c_t} - \nu_{t,g_i} \right]^2. \tag{3.21}$$

The last two displays show that κ_{NT} is small whenever $h(\alpha_i, \gamma_t)$ can be well approximated by $\delta_{i,c_t} + \nu_{t,g_i}$. In equation (3.8) we already informally discussed the magnitude of this approximation error, and found that it is of order $\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2$. We now want to provide a more formal discussion of this and show that κ_{NT} is asymptotically small under appropriate regularity conditions.

In Section 3.2.2.2 we described the clustering algorithms that delivers the group memberships g_i and c_t based on the initial estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$. The goal of the clustering is to group units i with approximately the same value of α_i, and

to group time periods t with approximately the same γ_t. It is therefore crucial that $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ are good proxies for α_i and γ_t. Specifically, we require that there exist functions $\lambda^* : \mathscr{A} \to \mathbb{R}^{R_*}$ and $f^* : \mathscr{C} \to \mathbb{R}^{R_*}$ such that $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ converge to the non-random limits $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$ as $N,T \to \infty$. The following assumption formalizes this and states all the regularity condition that we require on $h(\cdot,\cdot)$, $\lambda^*(\cdot)$, $f^*(\cdot)$, $\widehat{\lambda}_i^*$, $\widehat{f}_t^*$, and X_{it}.

Assumption 3.4.1. There exists a sequence $\xi_{NT} > 0$ such that $\xi_{NT} \to 0$ as $N,T \to \infty$, and

(i) The function $h(\cdot,\cdot)$ is at least twice continuously differentiable with uniformly bounded second derivatives.

(ii) Every unit i is a member of exactly one group $g_i \in \{1,\ldots,G\}$, and every time period t is a member of exactly one group $c_t \in \{1,\ldots,C\}$. The size of all G groups of units, and the size of all C groups of time periods is bounded uniformly by $Q_{\max}$.

(iii) There exists $B > 0$ such that $\|a-b\| \leq B\|\lambda^*(a)-\lambda^*(b)\|$ for all $a,b \in \mathscr{A}$, and $\|a-b\| \leq B\|f^*(a)-f^*(b)\|$ for all $a,b \in \mathscr{C}$, and the domains $\mathscr{A}$ and $\mathscr{C}$ are convex set.

(iv) $\frac{1}{N}\sum_{i=1}^N \left(\left\| \widehat{\lambda}_i^* - \lambda^*(\alpha_i) \right\|^2 \right) = O_P(\xi_{NT})$,
$\frac{1}{T}\sum_{t=1}^T \left(\left\| \widehat{f}_t^* - f^*(\gamma_t) \right\|^2 \right) = O_P(\xi_{NT})$.

(v) $\frac{1}{N}\sum_{i=1}^N \left\| \widehat{\lambda}_i^* - \widehat{\lambda}_{j(i)}^* \right\|^2 = O_P(\xi_{NT})$ for any matching function $j(i) \in \{1,\ldots,N\}$ such that $g_i = g_{j(i)}$, and $\frac{1}{T}\sum_{t=1}^T \left\| \widehat{f}_t^* - \widehat{f}_{s(t)}^* \right\|^2 = O_P(\xi_{NT})$ for any matching function $s(t) \in \{1,\ldots,T\}$ such that $c_t = c_{s(t)}$.

(vi) $\max_{k,i,t} \left| \widetilde{X}_{it,k} \right| = O_P(1)$, and $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{X}_{it}' \widetilde{X}_{it} = \Omega$, where Ω is a positive definite non-random matrix.

Lemma 3.4.1. *Under Assumption 3.4.1 we have*

$$\kappa_{NT} = O_P(\xi_{NT})$$

The lemma shows that the approximation error κ_{NT} vanishes at rate ξ_{NT} as $N,T \to \infty$. The assumption and lemma are formulated for arbitrary rates, but as will become clear from the following discussion, the best we can achieve in our setting is a rate of $\xi_{NT} = 1/\min(N,T)$, which coincides with $\xi_{NT} = 1/\sqrt{NT}$ in the special case that N and T grow at the same rate.

Part (i) of Assumption 3.4.1 requires the function $h(\cdot,\cdot)$ to be sufficiently smooth. This condition should not be surprising, because our informal discussion of the approximation error in equation (3.8) already relies on a second order Taylor expansion of $h(\cdot,\cdot)$, and the proof of Lemma 3.4.1 is based on exactly such an expansion.

Part (iii) and (iv) of the assumption are analogous to "Assumption 2 (injective moments)" in Bonhomme et al. (2021), except that they consider a one-way fixed effect setting while we consider a two-way fixed effect setting. Part (iii) requires the functions $\lambda^*(\cdot)$ and $f^*(\cdot)$ to be injective, that is, α_i and γ_t can be uniquely recovered from knowing $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$. A necessary condition for this is that

$$R^* \geq \max(d_\alpha, d_\gamma), \tag{3.22}$$

where d_α and d_γ are the dimensions of α_i and γ_t, respectively. Part (iv) requires the estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ to converge to $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$ at the average rate of $\xi_{NT}^{1/2}$. We expect that the estimated eigenfunctions of $h(\alpha_i,\gamma_t)$, which correspond to the estimated factor loadings and factors, proposed as cluster proxies in Section 3.2.2.2 satisfy this assumption by an application of Theorem 1 from Bai and Ng (2002). Since T observations are available for unit i we expect that $\widehat{\lambda}_i^*$ converges at a rate of $T^{1/2}$, and since N observations are available for time period t we expect that $\widehat{f}_t^*$ converges at a rate of $N^{1/2}$, see also, for example, Theorem 1 and 2 in Bai (2003). This explains why $\xi_{NT} = 1/\min(N,T)$ is the best rate we can achieve here.

Part (v) of Assumption 3.4.1 is a high-level assumption on the clustering mechanism used to obtain the group memberships g_i and c_t. For units i and j in the same group, and for time periods t and s in the same group, we demand the average differences $\widehat{\lambda}_i^* - \widehat{\lambda}_j^*$ and $\widehat{f}_t^* - \widehat{f}_s^*$ to be small as $N,T \to \infty$. In other words, we require that

the clustering mechanism does what it is intended to do, namely forming groups such that the estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ for units i and time periods t in the same group are close to each other. For a given clutstering algorithms (e.g. the one describe in Section 3.2.2.2) one could prove that this assumption holds under further regularity conditions on the distribution of α_i and γ_t, see, for example, Lemma 1 in Bonhomme et al. 2021. In particular, a necessary condition for part (v) of Assumption 3.4.1 to hold is the following:

Condition 3.4.1. $\frac{1}{N}\sum_{i=1}^{N}\left\|\alpha_i-\alpha_{j(i)}\right\|^2 = O_P(\xi_{NT})$ *for any matching function* $j(i)\in\{1,\ldots,N\}$ *such that* $g_i = g_{j(i)}$, *and* $\frac{1}{T}\sum_{t=1}^{T}\left\|\gamma_t-\gamma_{s(t)}\right\|^2 = O_P(\xi_{NT})$ *for any matching function* $s(t)\in\{1,\ldots,T\}$ *such that* $c_t = c_{s(t)}$.

This condition coincides with Assumption 3.4.1(v) in the unrealistic case that $\widehat{\lambda}_i^* = \alpha_i$ and $\widehat{f}_t^* = \gamma_t$. Starting from this unrealistic case and then applying the transformations $\lambda^*:\mathscr{A}\to\mathbb{R}^{R_*}$ and $f^*:\mathscr{C}\to\mathbb{R}^{R_*}$ and adding noise to the estimates then gives part (v) of Assumption 3.4.1. Crucially, for this regularity condition to hold, we need that $\xi_{NT}\gtrsim 1/\min(N^{2/d_\alpha},T^{2/d_\gamma})$, see Lemma 2 in Bonhomme et al. (2021) for the analogous results in a one-way fixed effect model (also Graf and Luschgy 2002). Since our actual clustering method is not based on the unobserved α_i and γ_t, but on $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ we require the stronger condition (in view of (3.22)) that

$$\xi_{NT}\gtrsim[\min(N,T)]^{-2/R^*}.$$

This is a necessary condition for Assumption 3.4.1(v) to be satisfied.[11] Therefore, if we want to achieve the best possible rate $\xi_{NT}=1/\min(N,T)$, then we need $R^*\leq 2$, which according to (3.22) implies that $d_\alpha\leq 2$ and $d_\gamma\leq 2$. This discussion shows that our group fixed-effect estimator $\widehat{\beta}_{\mathrm{G}}$ suffers from a curse of dimensionality with regards to the dimensions of α_i and γ_t. However, this should be unsurprising, given the semi-parametric nature of the estimation problem – with non-parametric com-

[11] Following the logic in Bonhomme et al. (2021) we believe that we actually only need $\xi_{NT}\gtrsim 1/\min(N^{2/d_\alpha},T^{2/d_\gamma})$, that is, our group fixed effect estimator $\widehat{\beta}_{\mathrm{G}}$ truly cannot achieve a convergence rate faster than $1/\min(N^{2/d_\alpha},T^{2/d_\gamma})$. Thus, if $R^*>\max(d_\alpha,d_\gamma)$, then $\xi_{NT}\gtrsim[\min(N,T)]^{-2/R^*}$ is probably not a necessary condition for the result of Lemma 3.4.1 itself, but only for our Assumption 3.4.1(v).

ponent $h(\alpha_i, \gamma_t)$. This also shows that there is a tradeoff between the LS estimator analyzed in Section 3.3 and the group fixed effects estimator discussed here – we will further compare those two estimators in our MC analysis below.

Finally, part (vi) of Assumption 3.4.1 requires some regularity conditions on the projected regressors $\widetilde{X}_k = M_N X_k M_T$ defined in (3.17).

This concludes our discussion of the approximation error κ_{NT}. We have argued that, under appropriate regularity conditions, including $\max(d_\alpha, d_\gamma) \leq 2$, we can use Lemma 3.4.1 to obtain $\kappa_{NT} = 1/\sqrt{NT}$, for N and T growing to infinity at the same rate. Since $\widehat{\beta}_{\mathrm{G}} - \beta^0 = \phi_{NT} + \kappa_{NT}$ we could then conclude that $\widehat{\beta}_{\mathrm{G}} - \beta^0 = O_P(1/\sqrt{NT})$, if we could also show that $\phi_{NT} = O_P(1/\sqrt{NT})$.

From the definition of ϕ_{NT} in (3.19) one might think that it is easy to derive this result on ϕ_{NT} by imposing an approximate exogeneity condition on the regressors. However, the problem is that $\widetilde{X}_k$ depends on the group assignments of units i and time periods t, which were constructed based on $\widehat{\lambda}^*$ and $\widehat{f}^*$, which depend on the errors ε. Thus, $\widetilde{X}_k$ depends on ε in complicated ways through the group assignment, making a proof of $\phi_{NT} = O_P(1/\sqrt{NT})$ technically challenging. In principle, we expect that

$$\sqrt{NT}\,\phi_{NT} \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{G}}) \tag{3.23}$$

holds for an appropriate covariance matrix Σ_{G}, and our simulations evidence suggest that this is indeed the case. However, we are not aiming to prove this result in this paper. As explained already in Section 3.2, this technical difficulty in analyzing $\widehat{\beta}_{\mathrm{G}}$ is exactly why we introduced the split-sample version of the group fixed-effect estimator, for which we are going to derive results in the following.

3.4.2 Results for $\widehat{\beta}_{\mathrm{GS}}$

The split-sample version of the group fixed effect estimator was introduced in Section 3.2.2.3 above. Using the Frisch-Waugh-Lovell theorem we can rewrite $\widehat{\beta}_{\mathrm{GS}}$ in

equation (3.9) as follows:

$$\widehat{\beta}_{\mathrm{GS}} = \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} Y_{it},$$

where the projected regressors $\widetilde{X}_{it}^{(s)} = \left(\widetilde{X}_{it,1}^{(s)}, \ldots, \widetilde{X}_{it,K}^{(s)} \right)'$ for each subpanel $s \in \{1,2,3,4\}$, each regressor $k = 1, \ldots, K$, and observations $(i,t) \in \mathscr{O}_s$ within that subpanel, are the residuals of the least-squares problem

$$\min_{\delta} \min_{\nu} \sum_{(i,t)\in\mathscr{O}_s} \left(X_{it,k} - \delta_{i,c_t^{(s)}} - \nu_{t,g_i^{(s)}} \right)^2. \tag{3.24}$$

Following the decomposition of $\widehat{\beta}_{\mathrm{G}}$ in (3.18), we can now introduce the analogous decomposition for $\widehat{\beta}_{\mathrm{GS}}$ by

$$\widehat{\beta}_{\mathrm{GS}} - \beta^0 = \phi_{NT}^{(\mathrm{GS})} + \kappa_{NT}^{(\mathrm{GS})}, \tag{3.25}$$

where

$$\phi_{NT}^{(\mathrm{GS})} := \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \varepsilon_{it},$$
$$\kappa_{NT}^{(\mathrm{GS})} := \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{\Gamma}_{it}^{(s)},$$

Here, $\phi_{NT}^{(\mathrm{GS})}$ is a variance term that we will show to be unbiased and asymptotically normal, and $\kappa_{NT}^{(\mathrm{GS})}$ is the approximation error from having replaced $h(\alpha_i, \gamma_t)$ by the linear grouped fixed effect in the estimation for $\widehat{\beta}_{\mathrm{GS}}$ in (3.9). The $\widetilde{\Gamma}_{it}^{(s)}$ are the residuals of the least-squares problem (3.24) when $X_{it,k}$ is replaced by $\Gamma_{it} = h(\alpha_i, \gamma_t)$.

For each of the four subpaneles $s \in \{1,2,3,4\}$, the discussion of the approximation error $\kappa_{NT}^{(\mathrm{GS})}$ is identical to the discussion of the approximation error κ_{NT} of $\widehat{\beta}_{\mathrm{G}}$, see, in particular, the bounds (3.20) and (3.21) above. It is therefore straightforward to obtain the analogue of Lemma 3.4.1 for the approximation error of the split-sample estimator.

Lemma 3.4.2. *Under Assumption B.2.1 (in appendix) we have*

$$\kappa_{NT}^{(GS)} = O_P(\xi_{NT})$$

Assumption B.2.1 is stated in the appendix, but it is simply a restatement of Assumption 3.4.1 for each subpanel $s \in \{1,2,3,4\}$. Those assumptions were discussed after Lemma 3.4.1 above. In particular, the best possible convergence rate we can hope for here is $\xi_{NT} = 1/\min(N,T)$, but that rate is only attainable for $d_\alpha \leq 2$ and $d_\gamma \leq 2$.

The key difference between $\widehat{\beta}_{\mathrm{G}}$ and $\widehat{\beta}_{\mathrm{GS}}$ is that for the split-sample estimator we can derive the asymptotic behavior of the variance term very easily $\phi_{NT}^{(\mathrm{GS})}$. For this purpose, we impose the following assumption.

Assumption 3.4.2.

(i) Conditional on X, α, γ, we assume that ε_{it} is independently distributed across i and over t, such that $\sigma_{it}^2 := \mathbb{E}\left(\varepsilon_{it}^2 \,\middle|\, X, \alpha, \gamma\right) \leq B < \infty$, for some constant B that is independent of i, t, N, T.

(ii) We have $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{s=1}^{4} \sum_{(i,t)\in\mathscr{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} = \Omega > 0$, and for each $s \in \{1,\ldots,S\}$ we have $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{(i,t)\in\mathscr{O}_s} \sigma_{it}^2 \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} = \Sigma^{(s)}$. Furthermore, we assume that, for $s \in \{1,2,3,4\}$, all the third-order sample moments of $\widetilde{X}_{it}^{(s)\prime} \varepsilon_{it}$ across $(i,t) \in \mathscr{O}_s$ are bounded as $N, T \to \infty$.

Assumption 3.4.2 together with the sample splitting method used to construct $\widehat{\beta}_{\mathrm{GS}}$ guarantees that, within each subpabel $s \in \{1,2,3,4\}$, the $\widetilde{X}_{it}^{(s)\prime} \varepsilon_{it}$ are zero mean and independently distributed across (i,t). Here, the split-panel construction is crucial, since it guarantees that $\widetilde{X}_{it}^{(s)}$ is independent of ε_{it}. The remaining conditions in Assumption 3.4.2 are regularity conditions to allow us to apply the Lyapunov central limit theorem for each subpanel and to guarantee that $\phi_{NT}^{(\mathrm{GS})}$ has a finite asymptotic variance. We therefore obtain the following lemma.

Lemma 3.4.3. *Under Assumption 3.4.2 we have, as $N,T \to \infty$,*

$$\sqrt{NT}\,\phi_{NT}^{(GS)} \Rightarrow \mathcal{N}(0,\Sigma_{GS}), \qquad \Sigma_{GS} = \Omega^{-1}\left(\sum_{s=1}^{4}\Sigma^{(s)}\right)\Omega^{-1}.$$

Combining equation (3.25) with Lemma 3.4.2 and Lemma 3.4.3 then gives the following theorem.

Theorem 3.4.3. *Under Assumption 3.4.2 and Assumption B.2.1 we have*

$$\widehat{\beta}_{GS} - \beta^0 = O_P\left(\frac{1}{\sqrt{NT}} + \xi_{NT}\right) = o_P(1)$$

Analogous to Corollary 3.3.1 for the least-squared estimator of Bai (2009), we have this obtained a consistency result for $\widehat{\beta}_{\mathrm{GS}}$ as well. We have not derived asymptotic inference results using either of these estimators, but in the following section we explain how we use those estimators to construct confidence intervals in our simulations and empirical application.

3.5 Implementation

The asymptotic results derived for $\widehat{\beta}_{\mathrm{LS}}$, $\widehat{\beta}_{\mathrm{G}}$, and $\widehat{\beta}_{\mathrm{GS}}$ in the last two sections are insightful for how those estimates should be used in practice. In particular, our discussions and derivations are helpful to appreciate the limitations and assumptions needed for the estimation approaches, and we will summarize those again in our conclusion section below.

In the following Monte Carlo simulations and empirical application we will employ the estimates $\widehat{\beta}_{\mathrm{LS}}$, $\widehat{\beta}_{\mathrm{G}}$, and $\widehat{\beta}_{\mathrm{GS}}$ in a way that goes beyond our formal asymptotic results. In particular, we will use all those estimators to construct confidence intervals and we will also apply Jackknife methods for bias correction. In this section, we want to briefly explain how those confidence intervals and bias corrected estimates are constructed.

To calculate standard errors for each estimator we ignore the approximation error discussed in our formal results and simply use formulas as if residuals were independently distributed. For example, in section 3.4.1 where we split the residual

term into ϕ and κ, we will ignore the κ term and estimate standard errors as if we are left with only ϕ. We use the jackknife corrections to address the residual terms related to approximation error in both the factor and grouped fixed-effects estimation models.

For factor model standard errors we construct the heteroscedasticity-consistent estimator from White (1980) as follows. Take $\Omega = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{X}_{it}$ and $\widehat{\Sigma} = \sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{u}_{it}^2\widetilde{X}_{it}'\widetilde{X}_{it}$ where $\widehat{u}_{it} = \widetilde{Y}_{it} - \sum_k \widehat{\beta}_{LS,k}\widetilde{X}_{it,k}$ and for a matrix A, in this context, $\widetilde{A}$ represents the matrix with factors projected. We must make a degrees of freedom correction for the factor projection by the ratio $dfc = \sqrt{\frac{NT}{(N-R)(T-R)}}$. Then the vector of standard errors are,

$$\text{se}(\widehat{\beta}_{\text{LS}}) = dfc \cdot \sqrt{\text{diag}\left(\Omega^{-1}\widehat{\Sigma}\Omega^{-1}\right)}.$$

As above, we use this same standard error estimator for jackknife corrected estimates.

For the grouped fixed-effects models we use clustered standard errors where clusters are taken as the combination of i and t clusters. That is, for the matrices of clusters D_α and D_γ for i and t respectively we take clusters as the Kronecker product between these two matrices, $D_\alpha \otimes D_\gamma$. Remember here that the columns of D_α, resp. D_γ, are the cluster assignments of i, resp. t with a 1 entry if that observation is in the cluster and a 0 otherwise. Take m as the index for cluster assignment with $M = GC$ the total number of clusters. Hence, $D_\alpha \otimes D_\gamma := \mathscr{D}$ is an NT by M matrix with $\mathscr{D}_m$ representing a column of this matrix and $\mathscr{D}_{n,m}$ representing an entry. A combination (i,t) can be identified by the row, n, of the matrix $\mathscr{D}$ as $t = \lceil n/N \rceil$ and $i = n - (\lceil n/N \rceil - 1)N$, which is similar to the usual matrix flattening procedure. Then, the column-vector $\mathscr{D}_m$ consists of a 1 if the (i,t) combination implied by that row, n, is in that column's cluster and 0 otherwise.

Define as above $\Omega = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{X}_{it}$ and $\widehat{u}_{it} = \widetilde{Y}_{it} - \sum_k \widehat{\beta}_{G,k}\widetilde{X}_{it,k}$ where in this context for matrix A, the matrix $\widetilde{A}$ represents the matrix with group fixed-effects projected out. Call the index function $n(i,t) = i + (t-1)N$, such that $\mathscr{D}_{n(i,t),m}$ returns the binary indicator of whether (i,t) is in the m^{th} combination cluster. Now define

$\widehat{\Sigma} = \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathscr{D}_{n(i,t),m}\widehat{u}_{it}^2\widetilde{X}_{it}'\widetilde{X}_{it}$. This collapses the familiar block-diagonal matrix where values within each block corresponds to a combination cluster and are unrestricted but zero outside each block. The clustered standard errors can thus be defined as

$$\text{se}(\widehat{\beta}_G) = dfc \cdot \sqrt{\text{diag}\left(\Omega^{-1}\widehat{\Sigma}\Omega^{-1}\right)}$$

where in this context $dfc = \sqrt{\frac{NT}{(N-G)(T-C)}}$. The standard error estimator is identical for the split sample version except there are many more combination clusters by the nature of this split sample estimators clustering method.

Finally, in our Monte Carlo simulations below we also explore whether Jackknife bias correction methods are able to reduce the approximation bias and the incidental parameter bias of the various estimates. We do not have any theoretical results on the leading order bias of the various estimates, but we nevertheless we follow Fernández-Val and Weidner (2016) to estimate the jackknife bias corrected analog to each estimator as follows. This procedure is closely related to Dhaene and Jochmans (2015). First, split the sample along the i dimension into two $N/2$ by T samples. For each of these samples run and call the related estimates from estimator E, $\widehat{\beta}_E^{1,1}$ and $\widehat{\beta}_E^{1,2}$, respectively. Repeat this process along the t dimension to return $\widehat{\beta}_E^{2,1}$ and $\widehat{\beta}_E^{2,2}$. Then the final jackknife bias corrected analog for estimator E is

$$\widehat{\beta}_{E,JK} = 3\widehat{\beta}_E - \frac{1}{2}\left(\widehat{\beta}_E^{1,1} + \widehat{\beta}_E^{1,2}\right) - \frac{1}{2}\left(\widehat{\beta}_E^{2,1} + \widehat{\beta}_E^{2,2}\right),$$

where $\widehat{\beta}_E$ is simply the estimate without any sample split. We maintain the assumption that standard errors are the same across split samples so we can simply take the standard error estimate from the whole sample.

3.6 Monte Carlo simulations

For our Monte Carlo simulations, we choose a data generating process with a single regressor ($K = 1$), and we generate outcomes and regressor as follows:

$$\begin{aligned} Y_{it} &= X_{it}\beta^0 + h(\alpha_i, \gamma_t) + \varepsilon_{it}, \\ X_{it} &= g(\alpha_i, \gamma_t) + \mu_{it}, \end{aligned} \tag{3.26}$$

with

$$\varepsilon_{it},\, \alpha_i,\, \gamma_t,\, \mu_{it} \sim \text{all mutually independent and i.i.d.} \mathcal{N}(0,1) \tag{3.27}$$

This setting assumes that the endogeneity in X_{it} depends on the specification of $g(.,.)$ vis–à–vis $h(.,.)$. The decay in singular values for either the unobserved term in Y_{it} or for X_{it} can be directly manipulated through the specification of $h(.,.)$ and $g(.,.)$, which will dictate the number of significant factors in each decomposition.

We set $\beta^0 = 1$ and,

$$h(a,b) = g(a,b) = \frac{1}{\sqrt{2\pi}\theta} \exp\left(-\frac{(a-b)^2}{\theta^2}\right), \quad \theta = (1/2)^3. \tag{3.28}$$

The θ value here dictates the speed of decay in singular values for $h(.,.)$ and $g(.,.)$, holding fixed the variation in their arguments, where a lower value implies a slower decay. This particular value for θ was chosen as it implies a slow decay in singular values such that the endogenous component of the unobserved term and X persists even as many factors are included. The value for θ carries no fundamental economic meaning. Note, the nature of bias in this simulation is by design monotonic and positive for illustrative purposes.

Table 3.2 below shows the results from 10,000 Monte Carlo simulations. These results display our theoretical result on bias reduction succinctly. We see that as we increase the number of factors the average bias reduces and the standard deviation of estimates increases. We also see a significant improvement in bias using the grouped fixed-effects estimator, without a large increase in standard deviation. The GFE split

sample estimator performs much worse in terms of bias, which is expected given the significantly smaller candidate pool for clustering in this estimator. The jackknife analog to each estimator reduces bias in all cases except the factor model with 5 factors, but significantly increases standard deviation in all cases. Note that we only report factor model estimated after first applying a within transformation, but we actually do not find any substantial difference compared to not applying the within transformation first.

Table 3.2: Monte Carlo simulations

	Bias	St. Dev.	Mean $\widehat{se}$	CDF(β^0)	Cover	MC cover
OLS	0.5201	0.0089	0.0085	0.00	0%	0%
Fixed-effects	0.5166	0.0095	0.0086	0.00	0%	0%
LS ($R = 5$)	0.0420	0.0115	0.0105	0.00	3%	0%
LS ($R = 20$)	0.0160	0.0148	0.0110	0.14	64%	28%
LS ($R = 50$)	0.0138	0.0317	0.0135	0.33	56%	66%
LS JK ($R = 5$)	-0.4210	0.0319	0.0105	1.00	0%	0%
LS JK ($R = 20$)	-0.0079	0.0283	0.0110	0.61	53%	78%
LS JK ($R = 50$)	-0.0019	0.0798	0.0135	0.51	26%	96%
GFE	0.0025	0.0177	0.0179	0.44	95%	89%
GFE jackknife	0.0002	0.0322	0.0179	0.50	73%	99%
GFE splits	0.0210	0.0182	0.0126	0.13	57%	25%

N = T = 100 with 10,000 repetitions.
All results refer to estimation of β. Bias is simply the mean of the bias across simulations. Standard deviation is the standard deviation of the estimates, again across simulations. LS JK is the Jackknife version of the least squares estimator. Mean $\widehat{se}$ is the mean across simulations of the standard error estimate. CDF(β^0) is value of the empirical CDF across simulations evaluated at the true value of $\beta^0 = 1$. Cover is defined here as the percentage of the 95% confidence intervals containing the true β^0. MC cover reports coverage if estimates are normally distributed with mean bias from column 2 and standard deviation from column 3.

If we compare the mean standard error estimates to standard deviation across simulations we see evidence that the standard error calculation may underestimate the true standard error of the estimator. In light of discussion in Section 3.5, we explicitly ignore fixed-effects approximation error and assumed only a noise term remains when estimating standard errors, which may explain this discrepancy. The divergence between estimated standard errors and standard deviation across simulations is particularly noticeable for the factor model with a large number of factors and for jackknife bias corrected estimators. For large factor models it is likely our

inference approach misses out dependence structures introduced by the factor projection. This divergence is less pronounced for the group fixed-effects estimator without bias correction. It is also worth noting the assumption of equal standard errors across the components of the jackknife estimator appears to be violated from the difference in standard deviations between jackknife and non-jackknife estimators. These results suggest an alternative method, for example using bootstrap, is necessary to do feasible inference in this setting with these estimators. As mentioned, we leave matters of inference for future research. Since we do not expressly advocate a particular inference approach for any estimator used in this paper we do not discuss this issue any further and leave it for future research.

In Table 3.2 we also compare where the true value of $\beta^0 = 1$ lies in the empirical CDF of each estimator to the coverage based on a normal distribution with the simulated mean bias and standard deviation as the distribution parameters. We see that in instances where bias is low and $\beta^0 = 1$ is close to the median of the empirical CDF then 2|CDF(β^0) - 0.5| $\times$100% is approximately equal to 1 - MC cover. This is some evidence that the estimators may approach the normal distribution, where the simulations correctly estimate the standard deviations. However, given that the estimated standard errors are usually far from the simulation standard deviations, this still does not present a feasible inference procedure.

To compare the rates of convergence across estimators we repeat the above simulation exercise across different sample sizes, namely $N = T \in \{20, 40, 80, 160\}$. The results are displayed in Table 3.3. The table shows that for this range of data the convergence rates for the GFE estimators are all better or equal to the parametric rate. Note, in this setting the parametric rate suggests the bias should halve for each increment in sample size. The factor model looks to be decaying at about the parametric rate, however, for the specification with a small number of factors (factors equal $N^{1/4}$ and $N^{3/8}$) the bias is substantially above standard deviation. This suggest there is a statistically significant persistence in bias for this estimator. For the factor model with $N^{9/20}$ factors, which is near the upper bound of number of factors, $\min\{N,T\}^{1/2}$, as per Theorem 3.3.5, the bias does converge to within two

standard deviations of zero. The standard deviations for each estimator do look to settle on the parametric rate by at the latest the last sample size increment, which is seen by comparison of the second last and last columns across estimators. In Appendix B.1 we also include a simulation exercise with lagged dependent variables, which highlights the importants of having a correctly specified model.

Table 3.3: Convergence rate simulation

$N = T =$	20	40	80	160
Mean Bias				
(Standard Deviation)				
LS ($N^{1/4}$ factors)	0.4461	0.2782	0.2696	0.1609
	(0.0983)	(0.0405)	(0.0194)	(0.0097)
LS ($N^{3/8}$ factors)	0.3546	0.1748	0.0860	0.0177
	(0.1245)	(0.0388)	(0.0159)	(0.0067)
LS ($N^{9/20}$ factors)	0.2763	0.1077	0.0320	0.0122
	(0.1411)	(0.0386)	(0.0158)	(0.0071)
Group fixed-effects	0.2690	0.0064	0.0045	0.0008
	(0.1525)	(0.0545)	(0.0224)	(0.0110)
GFE jackknife	0.2458	0.0334	0.0023	0.0004
	(0.2225)	(0.0942)	(0.0406)	(0.0201)
GFE splits	0.3829	0.1524	0.0249	0.0036
	(0.1200)	(0.0657)	(0.0237)	(0.0111)

10,000 Monte Carlo rounds.
All results refer to estimation of β. Mean bias is simply the mean of the bias across simulations. Standard deviation is the standard deviation of the estimates, again across simulations.

3.7 Empirical application

We apply our estimation procedure to an analysis of the UK housing market, following Giglio et al. (2016) (GMS16). Specifically, we study the effects of extremely long lease agreements on the price of housing, when compared to freehold agreements. In the UK housing market it is common for real estate property to be sold under each agreement. GMS16 posit that any change in price due to exogenous variation in whether the property was sold under extremely long lease or freehold must be attributed to so–called "housing bubbles associated with a failure of the transversatility condition". The empirical challenge in making this comparison, and much discussed in GMS16, is to sufficiently control for observable and unobservable co-

variates such that variation in the variable of interest can be reasonably described as exogenous.

In the following, we compare estimates using our method with the more flexible approach taken in their paper. We note first that given differences in data, these results should not be directly compared with GMS16. Rather, this should be seen as an internal validity check across estimation models, i.e., to check if the aggregated setting produce similar estimates to the granular setting from GMS16 within the same set of data.

Consider the granular model from GMS16

$$Y_{iprt} = ExtremelyLongLease_i\beta + controls'_{it}\delta + \phi_{prt} + \varepsilon_{iprt} \tag{3.29}$$

where i are individual *transactions* (i.e. not necessarily properties), p is property type, r are regions and t is the month of transaction. Controls include hedonic variables, e.g. number of bedrooms, bathrooms and floorspace. ϕ_{prt} is a scalar fixed effect particular to the region, property type and month, and is identified via variation across transactions i. Compare this to an aggregated setting,

$$Y_{rt} = ExtremelyLongLease_{rt}\beta + controls'_{rt}\delta + h(\alpha_r, \gamma_t) + \varepsilon_{rt} \tag{3.30}$$

where Y_{rt}, $ExtremelyLongLease_{rt}$ and $controls_{rt}$ are the sample means aggregated to the region and transaction month. The multidimensional array with entries ϕ_{prt} varies with higher rank than the matrix with entries $h(\alpha_r, \gamma_t)$ because the latter is constant across p if extended to the equivalent multidimensional array with dimensions across (p, r, t). This is why we believe the model in (3.29) will better capture fixed-effects.

For purposes of this exercise we take the granular model with fixed-effects below as being, in theory, the better model to approximate unobserved heterogeneity. Hence we refer to this as the benchmark model. We use this benchmark approach to understand how well each estimator performs in practical instances where granular levels of aggregation are not always available, for example when data is aggregated

for privacy reasons or for other feasibility reason. Hence, estimates close to the granular model estimates should be seen as performing "well" in this setting.

Table 3.4 shows that when we control for fixed effects in the granular model there is a 0.3% reduction in price when a long leasehold transaction is made compared to a freehold. Whilst this is statistically significant, it translates to a decrease in the median house price of less than £1,000 so is arguably a small reduction economically. The OLS estimates do not change much across the different aggregation schemes and perhaps unsurprisingly the panel aggregated OLS has a much higher standard deviation due to the lower effective sample size. In the panel setting the factor model shows a convergence to the granular model with fixed effects as factors are increased and, interestingly, also to the grouped fixed-effects estimate, which is the closest to the benchmark estimates.[12] These results show a similar pattern to the simulation exercise where, according to the benchmark model, we see a bias reduction as the number of factors increases and when using the group fixed-effects estimator.

Table 3.4: Empirical Results

	Model	Estimate	Standard Errors
Granular Model	Ordinary Least Squares	0.203	0.0054
(3.29)	with Fixed Effects	-0.003	0.0006
Panel Model	Ordinary Least Squares	0.229	0.106
(3.30)	LS factor model (5 factors)	0.024	0.012
	LS factor model (15 factors)	0.007	0.007
	LS factor model (30 factors)	0.007	0.008
	Group fixed-effects	0.006	0.020

UK housing market results for N = 2088 and T = 48.

3.8 Conclusions

Panel regressions are very popular estimation tools, because they allow to control for omitted variables that are unobserved and potentially correlated with the observed covariates. Both Pesaran (2006) and Bai (2009), and most of the literature

[12] In Table 3.4, our usual computation for the clustered standard errors of the group fixed-effect estimator was infeasible here due to the sample size. These standard error estimates are generated by resampling region clusters with replacement over 10,000 resamples.

following those seminal papers, assume that those unobserved omitted variables take the form of a low-rank matrix, which can be interpreted as a static factor model or interactive fixed effects. In this paper, we deviate from this interactive fixed effect model by assuming that the unobserved omitted variables enter the model in the more general form $h(\alpha_i, \gamma_t)$, where $h(\cdot,\cdot)$ is an unknown smooth function, and α_i and γ_t are (multidimensional) fixed effects that can be arbitrarily correlated across i, over t, and with the observed covariates.

We first explore the behavior of Bai's least-squares esimator in this new setting. We show that this LS esimator estimator is still consistent, as long as the number of factors used in the estimation is allowed to grow asymptotically. However, as explained in detail in Section 3.3, it seems impossible to derive convergence rates faster than $(\min\{N,T\})^{1/2}$ for this estimator in our setting.

We therefore develop a new estimation approach called the two-way grouped fixed effects approach, which generalize ideas in Bonhomme et al. (2021) to our two-way setting. We derive convergence rate results for the resulting new estimators and show that, depending on the dimension of α_i and γ_t, and the relative size of N and T, convergence rates up to $\sqrt{NT}$ can be achieved with our new estimation approach.

We also explore the performance of those various estimators in simulations and in an empirical application. We find that both Bai's least-squares esimator and our grouped fixed effect estimators tend to perform well in practice. Interestingly, the theoretical convergence rate of $(\min\{N,T\})^{1/2}$ for the LS esimator may often understate the performance of this esimator in practice.

We also find that Jackknife bias correction helps to further reduce the bias of the various estimators, but at the cost of increasing the variance. Overall, the (Jackknife corrected) group fixed-effects estimator tends to have the smallest bias, but not necessarily the smallest variance. The empirical application shows that, according to our benchmark estimation, the LS estimation approach improves with more factors and that the group fixed-effects estimator does indeed provide a bias reduction compared to the LS estimator.

In the simulation exercise and empirical application we implemented standard error calculations for each estimator, but we leave formal inference results in the setting of our paper as an open question for future research.

Chapter 4

Multidimensional Interactive Fixed-Effects

4.1 Introduction

Models of multidimensional data – panel data with more than two dimensions – are fast becoming popular in econometric analysis as large data sets with a multidimensional structure become available. For example, in gravity models of trade that are repeated over time one may be interested in studying trade patterns between an importer, i, an exporter, j, that is repeated every quarter or year, t. One may also be interested in studying demand elasticities through consumption data that may vary by product, i, store, j, with repeated observation over week or month, t.[1] In these examples it is clear that there may exist unobserved characteristics in each dimension that can determine variation across both the dependent and independent variables that needs to be controlled for to avoid issues with endogeneity. For example, this could be shifts in taste preferences, that are unobserved by the econometrician, that may effect sales of particular products in certain stores differently over time. Thus far, most analysis has addressed unobserved heterogeneity in the higher-dimensional setting by using a combination of additive scalar fixed-effects. These additive scalar fixed-effects approaches, however, can only accommodate variation in unobserved heterogeneity over a subset of dimensions with any

[1] A non-exhaustive list of related examples can be found in the introduction of Matyas (2017) in trade, housing and prices, migration, country productivity and consumer price setting.

one of the scalar fixed-effects terms. For example, in the three-dimensional model this type of fixed-effect approach can only control for variation over ij, it and jt, but not over all ijt. In the face of more complicated relationships that admit multiplicative variation across dimensions, these additive effects are unsatisfactory to control for unobserved heterogeneity. This paper develops tools to control for unobserved heterogeneity in the form of interactive fixed-effects in models of multidimensional panel data. The main body of the paper focuses on the linear and additively separable model. More generic applications of these tools are discussed in the introduction but are not formally studied.

To fix ideas consider linear parameter estimation in the following interactive fixed-effects model with three dimensions,

$$Y_{ijt} = X_{ijt}'\beta + \sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)} + \varepsilon_{ijt}, \tag{4.1}$$

where all terms in $\sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$ are unobserved and L is unknown but small relative to sample size. Reducing the problem to three dimensions is without loss of generality for the methods considered herein. Additive fixed-effects are omitted for brevity but are subsumed by the interactive fixed-effect term or can be removed with a simple within transformation. Let X_{ijt} be arbitrarily correlated with the unobserved interactive fixed-effects term, $\sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$, but uncorrelated with the noise term, ε_{ijt}. The challenge to estimating β is isolating variation in X_{ijt} that is not correlated with the interactive fixed-effects term. This paper develops the multidimensional group fixed-effects and kernel weighted transformations to project out this unobserved heterogeneity and also shows settings were standard factor methods work well. The kernel weighted within transformation is a novel contribution to the best of the author's knowledge. The group fixed-effects method uses similar clustering techniques from Bonhomme et al. (2021) and the within-cluster transformation in Freeman and Weidner (2022).

This paper makes two main contributions to the literature. The first is to show that the three or higher dimensional model can be couched in a standard

two-dimensional panel data model and to derive sufficient conditions for consistency using these methods. The second contribution is to introduce kernel weighted fixed-effects methods and extend group fixed-effects methods to the multidimensional setting. The asymptotic results show that under certain conditions the group and kernel weighted fixed-effects can retrieve the parametric rate of convergence; and shows the merits of the two-dimensional panel methods. With existing proof techniques, it has not yet be shown that the panel methods can in general achieve the parametric rate in the three dimensional setting. The simulation results corroborate these theoretical findings and an empirical application that estimates the demand elasticity of beer demonstrates how these methods work in practice.

The within-cluster transformation can be motivated by considering a very simple extension to the usual within transformation. First, consider methods to project additive fixed-effects of the form $a_{ij} + b_{it} + c_{jt}$, which is usually projected using

$$\dot{Y}_{ijt} = Y_{ijt} - \bar{Y}_{\cdot jt} - \bar{Y}_{i\cdot t} - \bar{Y}_{ij\cdot} + \bar{Y}_{\cdot\cdot t} + \bar{Y}_{\cdot j\cdot} + \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}, \tag{4.2}$$

applied equivalently to X_{ijt}, where the bar variables denote the average taken over the "dotted" index for the entire sample. That is, $\bar{Y}_{\cdot jt} := \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{ijt}$, $\bar{Y}_{\cdot\cdot t} := \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} Y_{ijt}$, etc. The within-cluster transformation simply constrains the sample these averages are taken over to just within each unit's cluster. With a slight abuse of notation, this is done using,

$$\widetilde{Y}_{ijt} = Y_{ijt} - \bar{Y}_{i^* jt} - \bar{Y}_{ij^* t} - \bar{Y}_{ijt^*} + \bar{Y}_{i^* j^* t} + \bar{Y}_{i^* jt^*} + \bar{Y}_{ij^* t^*} - \bar{Y}_{i^* j^* t^*} \tag{4.3}$$

where the bar variables combined with the star indices denote means taken within that indice's cluster. For example, $\bar{Y}_{i^* jt}$ is the mean value of all i^*'s assigned to i's cluster, $\bar{Y}_{i^* jt^*}$ is the mean across both i^* in i's cluster and t^* in t's cluster, and so on. This is equivalent to including fixed-effects of the form $a_{ijg_3(t)} + b_{ig_2(j)t} + c_{g_1(i)jt}$, where $g_n(\cdot)$ maps to the cluster assignment for units in dimension n. Furthermore, the kernel weighted method simply uses weights rather than cluster assignments to take these averages, for example $\bar{Y}_{i^* jt} = \sum_{i'} w_{i,i'} Y_{i'jt}$ for some weights on each i'

for i, that need not be symmetric. This projects out the more generic fixed-effect, $\sum_{t'} w_{t,t'} a_{ijt'} + \sum_{j'} w_{j,j'} b_{ij't} + \sum_{i'} w_{i,i'} c_{i'jt}$. It is therefore apparent that with a relatively small change to how the within transformation is performed, much more general fixed-effects can be controlled for, including, as is shown later, the interactive fixed-effects considered in this paper.

The model for interactive fixed-effects has precedent in the standard two-dimensional panel data setting. For instance take the model considered in Bai (2009) and similar to Pesaran (2006),

$$Y_{it} = X_{it}'\beta + \sum_{\ell=1}^{L} \lambda_{i\ell} f_{t\ell} + e_{it}. \tag{4.4}$$

In that setting, Bai (2009) show that the interactive term $\sum_{\ell=1}^{L} \lambda_{i\ell} f_{t\ell}$ also sufficiently captures variation in additive individual and time effects without the need to specify these separately, so these are again naturally omitted. For multidimensional applications it may be preferable to simply transform the problem in (4.1) to a two dimensional problem and estimate (4.4) directly using the transformed data. However, and as will be explained in further detail in Section 4.3.1, problems persist when L is large and only a subset of the unobserved heterogeneity parameters are low-dimensional. For consistent estimation of β, transforming the multidimensional array to a matrix then estimating (4.4) requires either: (a) all fixed-effects are low-dimensional, or; (b) that a subset of the fixed-effects are low-dimensional and the analyst knows which ones are. The requirement that the analyst has this knowledge can be highly restrictive. Furthermore, only a very slow rate of convergence can be shown for this approach even when the analyst does know which fixed-effects live in a low-dimensional space. Alternatively, the within-cluster and kernel weighted transformations analysed in this paper requires only that a subset of the fixed-effect parameters are low-dimensional, though the analyst does not need to know which of the fixed-effect parameters make up this subset. Further, when fixed-effects in all dimensions can be estimated well, the usual parametric rate of consistency is possible with these within transformations.

The demand elasticity for beer application uses Dominick's supermarket data from the Chicago area from 1991-1995, where price and quantity vary over product, store, and month. The log-log and logit models are estimated. The log-log model is implemented with and without cross-elasticities to demonstrate the limited ability of the fixed-effects estimators to project relevant covariates that vary across all dimensions. The elasticities for the fixed-effects estimators, including the simple additive fixed-effects, are all similar within each model, indicating that whilst fixed-effects probably exist, they are most likely of a simple form. The estimates with and without cross-elasticities are also substantially different, which indicates that if cross-elasticities are important then even the more sophisticated fixed-effects estimators cannot project them out. Hence, relevant covariates with high variation across all dimensions still need to be included in the regression line. The estimates from the log-log with cross-elasticities closely reflect the own-price elasticities in Table 1 from Hausman et al. (1994).

The technical component of this paper is highly related to the numerical analysis literature on low-rank approximations of multidimensional arrays. As pointed out in De Silva and Lim (2008), the optimisation problem of finding low-rank approximations in the tensor setting is not well-posed, hence most results in this literature rely on numerical evidence. See Kolda and Bader (2009) for a summary of the multidimensional array decomposition problem and Vannieuwenhoven et al. (2012); Rabanser et al. (2017) for examples of numerical results. As such, it is necessary to innovate on this tensor low-rank problem to find appropriate analytical results. To this end, this paper utilises well-posed components of the numerical analysis literature for use in nuisance parameter applications. These applications have the advantage that they do not require the multidimensional array of fixed-effects to be reconstructed, hence do not attempt to directly solve the low-rank tensor problem. It is worth a note that Elden and Savas (2011), along with related papers, suggest a reformulation of the low multilinear rank problem that may have promising applications in econometrics, but this is left for future research.

Some extensions of this modelling approach are now informally discussed but

not considered further in this paper. Under sufficient regularity conditions, the methods considered in this paper may also control for variation from arbitrary functions of the fixed-effects. Similar to that considered in Zeleneev (2020) and Freeman and Weidner (2022), the functional representation of model (4.1) could be,

$$Y_{ijt} = X'_{ijt}\beta + h(\varphi_i^{(1)}, \varphi_j^{(2)}, \varphi_t^{(3)}) + \varepsilon_{ijt},$$

for vector-valued $\varphi_i^{(1)}$, $\varphi_j^{(2)}$ and $\varphi_t^{(3)}$. The set of fixed-effects to be transformed by the function $h(\cdot,\cdot,\cdot)$ could also extend to fixed-effects over multiple indices, e.g. α_{ij} from above. It should be noted that the setting considered in Zeleneev (2020) requires that the transformation is non-smooth, and it is not trivial to see that a "within-type" transformation will sufficiently project this type of heterogeneity. With sufficient smoothness conditions on the function transforming the fixed-effects, existing literature could be generalised to show consistency using the proposed within-cluster transformation in the multidimensional case.

Models with discrete explanatory variables (Chernozhukov et al., 2013b; Hoderlein and White, 2012b; Evdokimov, 2010; Fernández-Val et al., 2021), provide another interesting application of these group fixed-effects estimators. Take the following regression line for discrete valued X_{ijt},

$$Y_{ijt} = h\left(X_{ijt}, \varphi_i^{(1)}, \varphi_j^{(2)}, \varphi_t^{(3)}, \varepsilon_{ijt}\right).$$

Then, under sufficient smoothness conditions on the function h, the unobserved heterogeneity may also be projected out with a group fixed-effect estimator. Tensor completion techniques also have useful generalisations in this setting, for example, Tomioka et al. (2010); Li et al. (2019); Xu (2020), for some examples of methods that consider sparse multidimensional arrays. The sparse multidimensional array problem has similar complexities to the low-rank tensor approximation problem in that they do not extend from the matrix problems in a straightforward way, hence require non-trivial extensions.

It is also important to consider unobserved heterogeneity in applications that

admit discrete dependent variables. For example, for binary response variable with known $F(\cdot)$,

$$P\left(Y_{ijt}=1\middle|X_{ijt},\varphi_i^{(1)},\varphi_j^{(2)},\varphi_t^{(3)}\right)=F\left(X_{ijt}'\beta+\sum_{\ell=1}^{L}\varphi_{i\ell}^{(1)}\varphi_{j\ell}^{(2)}\varphi_{t\ell}^{(3)}\right). \tag{4.5}$$

Estimation of the unobserved heterogeneity term may then be performed with a similar iterative scheme as that proposed in Chen et al. (2021) or the sufficient statistic approach in Chapter 6 of Matyas (2017). The incidental parameter problem in this setting can be alleviated using methods in this paper by allowing cluster sizes to grow with data size coupled with taking grouped fixed-effects along fewer dimensions, for example in Bonhomme et al. (2021) and also Appendix C.2.

Menzel (2021) consider a special case of multidimensional data for bootstrapping methods where the data is D-adic. That is, each dimension of the data refers to the same set of observations, like a network graph where each index refers to an individual in the network. An example of the multidimensional version of this could be a binary indicator of a three step path, $Y_{ijk}=G_{ij}G_{jk}$, detailing if there exists a path from i to k. In any case, the type of multidimensional data considered in that work is a distinct special case of the type of data structures considered in this paper.

The paper is organised as follows. Section 4.2 introduces the model, and notation and preliminaries; Section 4.3 details the estimators and associated assumptions with convergence results; Section 4.4 discusses the convergence results along with some alternative assumptions, and further motivates the estimation approach; Section 4.5 displays the simulation results; Section 4.6 shows the beer demand estimation empirical application; and Section 4.7 concludes.

4.2 Model

Let β^0 denote the true parameter value for the slope coefficients. The model in full dimensional generality is, [2]

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{X}_k \beta_k^0 + \mathscr{A} + \boldsymbol{\varepsilon}, \tag{4.6}$$

where $\boldsymbol{Y}, \boldsymbol{X}_k, \boldsymbol{\varepsilon} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_d}$. $\mathscr{A} = \sum_{\ell=1}^{L} \varphi_\ell^{(1)} \circ \cdots \circ \varphi_\ell^{(d)}$ where $\varphi_\ell^{(n)} \in \mathbb{R}^{N_n}$ for each $n = 1, \ldots, d$ and "$\circ$" is the outer product. L is naturally restricted to have upper bound $\min_n \{\prod_{n' \neq n} N_{n'}\}$, see Kruskal (1989). $\boldsymbol{\varepsilon}$ is a noise term uncorrelated with all $\boldsymbol{X}_k$ and all unobserved fixed-effects terms. Take $i_n \in \{1, \ldots, N_n\}$ for all $n \in \{1, \ldots, d\}$ as the dimension specific index, where N_n is the sample size of dimension n. The regressors $\boldsymbol{X}_k$ may be arbitrarily correlated with $\mathscr{A}$. Throughout this paper all dimensions are considered to grow asymptotically, that is $N_n \to \infty$ for all n.

Model (4.6) can be seen as a natural extension of the Bai (2009) model to three (or more) dimensions with the $\mathscr{A}$ term interpreted as a "higher-dimensional" factor stucture. Similar to this strain of the literature, all terms in $\mathscr{A}$ are considered fixed nuisance parameters. There are potentially many extensions to the factor model setting in Bai (2009) to the higher dimension case. This paper starts with what seems the most natural extension.

The term $\mathscr{A}$ may also incorporate additive fixed effects that vary in any strict subset of the dimensions. For example, in the three dimensional setting one may want to control for the additive terms, $a_{ij} + b_{it} + c_{jt}$. These can be controlled for using $L = \min\{N_1, N_2\} + \min\{N_1, N_3\} + \min\{N_2, N_3\}$, with the first $\min\{N_1, N_2\}$ terms $\sum_{\ell=1}^{\min\{N_1,N_2\}} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} = a_{ij}$ by setting $\varphi_{t\ell}^{(3)} = 1$ for $\ell = 1, \ldots, \min\{N_1, N_2\}$, and so on for the b_{it} and c_{jt}. These could also be controlled for directly using the standard within-transformation before considering the model in (4.6).

---

[2] For example, in index notation this model can be written as,

$$Y_{i_1,i_2,\ldots,i_d} = \sum_{k=1}^{K} X_{i_1,i_2,\ldots,i_d;k} \beta_k^0 + \mathscr{A}_{i_1,i_2,\ldots,i_d} + \varepsilon_{i_1,i_2,\ldots,i_d}$$

with $\mathscr{A}_{i_1,i_2,\ldots,i_d} = \sum_{\ell=1}^{L} \varphi_{i_1\ell}^{(1)} \ldots \varphi_{i_d\ell}^{(d)}$.

This paper comprises of two main modelling approaches. The first is to embed the multidimensional model into a standard panel data model by simply flattening all arrays into matrices. The second approach uses weighted differences across each dimension to reduce each $\varphi^{(n)}$ component of $\mathscr{A}$ separately for each n. For this reason the model assumptions are split out and stated in Section 4.3 alongside each estimation approach.

4.2.1 Notation and preliminaries

For a d-order tensor, $\mathbf{A}$, a factor-n flattening, denoted as $\mathbf{A}_{(n)}$, is the rearrangement of the tensor into a matrix with dimension n varying along the rows and the remaining dimensions simultaneously varying over the columns. That is, $\mathbf{A}_{(n)} \in \mathbb{R}^{N_n \times N_{n+1}N_{n+2}\ldots N_1 \ldots N_{n-1}}$. The Frobenius norm, $\|\cdot\|_F$, of a matrix or tensor is the entry-wise norm, $\|\mathbf{A}\|_F^2 = \sum_{i_1=1}^{N_1} \ldots \sum_{i_d=1}^{N_d} A_{i_1 \ldots i_d}^2$. The spectral norm, denoted $\|\cdot\|$, is the largest singular value of a matrix. For a d-order tensor, $\mathbf{A}$, the multilinear rank, denoted $\mathbf{r}$, is a vector of matrix ranks after factor-n flattening in each dimension, with each component of this vector $r_n = rank\big(\mathbf{A}_{(n)}\big)$. Tensor rank, different to multilinear rank, is defined as the least number of outer products of vectors to replicate the tensor. That is, for tensor $\mathbf{A}$ and vectors $u_\ell^{(n)} \in \mathbb{R}^{N_n}$, tensor rank is the smallest L such that $\mathbf{A} = \sum_{\ell=1}^{L} u_\ell^{(1)} \circ \cdots \circ u_\ell^{(d)}$, where $\circ$ is the outer product of a vector. The notation $a \lesssim b$ means the asymptotic order of a is bounded by the asymptotic order of b.

The n-mode product between a tensor $\mathbf{A}$ and matrix B is denoted $\mathbf{A} \times_n B$ and has elements

$$(\mathbf{A} \times_n B)_{i_1,\ldots,j,\ldots,i_d} = \sum_{i_n=1}^{N_n} A_{i_1,\ldots,i_n,\ldots,i_d} B_{j,i_n},$$

which is equivalent to saying the flattening $(\mathbf{A} \times_n B)_{(n)} = B\mathbf{A}_{(n)}$. This can be referred to as "hitting" the tensor $\mathbf{A}$ with matrix B in the n^{th} dimension, though this terminology is only stated to help with understanding of the definition.

The singular value decomposition is used in both estimation approaches in this paper so it is important to understand some of its properties. The singular value

decomposition of a matrix, $A \in \mathbb{R}^{N_1 \times N_2}$ is

$$A = U\Sigma V' = \sum_{r=1}^{\min\{N_1,N_2\}} \sigma_r u_r v_r', \tag{4.7}$$

where U is the matrix of left singular vectors, u_r, V is the matrix of right singular vectors, v_r, and Σ is a diagonal matrix of singular values, σ_r, with values running in descending order down the diagonal. For a rank-r matrix, the first r entries on the diagonal of Σ are strictly positive and the remaining entries are zero.

Take the approximation problem,

$$\min_{A'} \|A - A'\|_F \text{ such that } rank(A') = k. \tag{4.8}$$

It is well known from the Eckart-Young-Mirsky theorem that the solution to this approximation problem is the first k terms of the singular value decomposition, i.e. $\sum_{r=1}^{k} \sigma_r u_r v_r'$. The Eckart-Young-Mirsky theorem effectively picks out the row and column subspaces that best explain variation in the matrix A as the leading columns of the matrix U, respectively of V. The sum of squared error at the minimiser is thus $\sum_{r=k+1}^{\min\{N_1,N_2\}} \sigma_r^2$. This is commonly called a low-rank approximation and forms the cornerstone for estimation of unobserved heterogeneity in the factor model and interactive fixed-effects models in Bai and Ng (2002); Bai (2009); Moon and Weidner (2015) amongst others.

The Eckart-Young-Mirsky theorem, however, does not extend to the three or higher dimensional setting, see De Silva and Lim (2008) for details. This is why the multidimensional problem either needs to be translated to the two-dimensional setting to utilise the Eckart-Young-Mirsky theorem, or the fixed-effects parameters need to be shrunk separately, as is done with the group fixed-effects and kernel methods.

4.3 Estimation

This section details the three estimation approaches used. The first subsection details how to apply standard two dimensional estimators to the problem and the as-

sumptions required for consistent estimation. The second and third subsections detail the group fixed-effect and kernel weighted approaches and the required assumptions for consistency in those settings.

4.3.1 Matrix low-rank approximation estimator

This section provides a description of some matrix methods that can be applied directly to the multidimensional model and stipulates the assumptions required for consistency. Note that Kapetanios et al. (2021) employ a similar approach for three-dimensional arrays in conjunction with the Pesaran (2006) common correlated effects estimator.

Consider recasting the multidimensional array problem into a two dimensional panel problem by flattening Y and X in the n-th dimension,

$$Y_{(n)} = X'_{(n)}\beta^0 + \varphi^{(n)}\Gamma'_n + \varepsilon_{(n)}$$

where $Y_{(n)}, X_{(n)}, \varepsilon_{(n)} \in \mathbb{R}^{N_n \times \prod_{n' \neq n} N_{n'}}$, $\varphi^{(n)}$ is an $N_n \times r_n$ matrix and Γ_n is an $\prod_{n' \neq n} N_{n'} \times r_n$ matrix that accounts for variation in the remaining $\varphi^{(n')}$ for all $n' \neq n$. The term r_n is indexed by the dimension n because it may vary non-trivially according to the flattened dimension. It should then be apparent that this is exactly the model described in (4.4), that is, the standard linear model with factor structure unobserved heterogeneity as studied in Bai (2009).

The two-dimensional estimator for a given flattening, n, optimises the following objective function,

$$R(\beta, \widehat{r}_n, n) = \min_{\substack{\varphi^{(n)} \in \mathbb{R}^{N_n \times \widehat{r}_n}, \\ \Gamma_n \in \mathbb{R}^{\prod_{n' \neq n} N_{n'} \times \widehat{r}_n}}} \left\| Y_{(n)} - X'_{(n)}\beta - \varphi^{(n)}\Gamma'_n \right\|_F^2. \tag{4.9}$$

Then $\widehat{\beta}^{2D}_{(n)} = \operatorname{argmin}_\beta R(\beta, \widehat{r}_n, n)$ is the slope estimate for the two-dimensional setup. The analyst must choose both the dimension to flatten in, n, and the rank of the estimated interactive fixed-effects term, $\widehat{r}_n$. It is well known that the minimum in (4.9) is achieved using the leading $\widehat{r}_n$ terms from the singular value decomposition of the error term, $Y_{(n)} - X'_{(n)}\beta$. This gives $\widehat{\varphi}^{(n)}$ as the first $\widehat{r}_n$ columns of $\widehat{U}\widehat{\Sigma}$ and

$\widehat{\Gamma}_n$ as the first $\widehat{r}_n$ columns of $\widehat{V}$ where $\widehat{U}$, $\widehat{\Sigma}$ and $\widehat{V}$ are the terms from (4.7) of the singular value decomposition of $Y_{(n)} - X'_{(n)}\beta$. Because this error term is a function of β, an iteration is naturally required between estimating β and finding the singular value decomposition of the error term. This is a well studied iteration procedure, for convergence details see Bai (2009); Moon and Weidner (2015).

In the following assumptions let $\widehat{r}_n$ be the estimated number of factors for the (n)-flattening of the regression line when applying the least square methods in (4.9). Also, let $\mathscr{L} \subset \{1,\dots,d\}$ be a non-empty subset of the dimensions. The tensor rank parameter, L, may without loss be restricted to the upper bounded by $L \leq \min_n\{\prod_{n'\neq n} N_{n'}\}$. This is a result of elementary bounds on the tensor rank of an arbitrary tensor. In the following, the multilinear rank of $\mathscr{A}$ is restricted such that it is low-rank along at least one of the flattenings.

Assumption 4.3.1 (Bounded norms of covariates and exogenous error)**.**

(i). $\left\|X_k\right\|_F = O_p\left(\prod_{n=1}^d \sqrt{N_n}\right)$ for each k

(ii). $\left\|\varepsilon_{(n^*)}\right\| = O_p\left(\max\{\sqrt{N_{n^*}}, \prod_{m\neq n^*}\sqrt{N_m}\}\right)$ for each $n^* \in \mathscr{L}$

Assumption 4.3.2 (Weak exogeneity)**.** $\operatorname{vec}(X_k)'\operatorname{vec}(\varepsilon) = O_p\left(\prod_{n=1}^d \sqrt{N_n}\right)$ for each k

Assumption 4.3.3 (Low multilinear rank)**.** For some positive integer, c, $r_{n^*} < c$ for all $n^* \in \mathscr{L}$, where r_n is the n^{th} component of the multilinear rank of $\mathscr{A}$.

Assumption 4.3.4 (Non-singularity)**.** Let $\sigma_s(A)$ be the s^{th} singular value for a matrix A. Consider linear combinations $\delta_{n^*} \cdot X_{(n^*)} = \sum_k \delta_{n^*,k} X_{(n^*),k}$. For each dimension $n^* \in \mathscr{L}$ that satisfies Assumption 4.3.3, then for $K \times 1$ unit vector δ_{n^*},

$$\min_{\{\delta_{n^*}\in\mathbb{R}^K, \|\delta_{n^*}\|=1\}} \sum_{s=r_{n^*}+\widehat{r}_{n^*}+1}^{\min\{N_{n^*}, \prod_{m\neq n^*} N_m\}} \sigma_s^2\left(\frac{(\delta_{n^*} \cdot X_{(n^*)})}{\prod_n \sqrt{N_n}}\right) > b > 0 \qquad wpa1.$$

Assumptions 4.3.1, 4.3.2 and 4.3.4 are standard regularity assumptions already well established in the literature, e.g. see Moon and Weidner (2015). Assumption 4.3.1.(i) ensures that the covariates have bounded norms, for example having

bounded second moments. Assumption 4.3.1.(ii) allows for some weak correlation across dimensions, see Moon and Weidner (2015), or is otherwise implied if the noise terms are independently distributed with bounded fourth moments, see Latała (2005). Assumption 4.3.2 is implied if $X_{i_1,i_2,\ldots,i_d;k}\varepsilon_{i_1,i_2,\ldots,i_d}$ are zero mean, bounded second moment and only admits weak correlation across dimensions for each $k = 1,\ldots,K$. Assumption 4.3.4 simply states that, after factor projection, the set of covariates still collectively admit full-rank variation.

Assumption 4.3.3 is new and asserts that there exists at least one flattening of the interactive term, $\mathscr{A}$, that is low-dimensional or simply low-rank. Given that the true value for L is left mostly unrestricted at this stage, this requires that at least one of the unobserved terms $\varphi^{(n)}$ is low dimensional. Note that not all dimensions must satisfy Assumption 4.3.3 for the below result. If the correct dimension is chosen then variation from the interactive term can be sufficiently projected out using the factor model approach. This makes up the statement of the following Proposition.

Proposition 4.3.1. *Let $\widehat{\beta}^{2D}_{(n^*)}$ be the estimator from Bai (2009) after first flattening along dimension $n^* \in \mathscr{L}$. If Assumptions 4.3.1-4.3.4 hold, the subset $\mathscr{L}$ is non-empty, and the estimated number of factors $\widehat{r}_{n^*} \geq r_{n^*}$, then, for each $n^* \in \mathscr{L}$ satisfying Assumption 4.3.3,*

$$\left\|\widehat{\beta}^{2D}_{(n^*)} - \beta^0\right\| = O_p\left(\frac{1}{\sqrt{\min\{N_{n^*}, \prod_{n\neq n^*} N_n\}}}\right). \tag{4.10}$$

Proposition 4.3.1 follows directly from Moon and Weidner (2015) since the flattening procedure reduces the problem to the standard linear factor model. Notice that this result only applies to estimates in the dimension(s) that satisfy the low-rank assumption in Assumption 4.3.3. That is, implicit in Proposition 4.3.1 is that the analyst has chosen the correct dimension to flatten over when reformulating the problem as a two-dimensional panel. Assumption 4.3.3 can be relaxed to $r_{n^*} = o\left(\min\{N_{n^*}, \prod_{n\neq n^*} N_n\}\right)$ as long as the estimated number of factors is allowed to increase with data size at a faster rate than this. The constraint $\widehat{r}_{n^*} \geq r_{n^*}$ can also be changed to $\widehat{r}_{n^*} \geq c$, however, this is more conservative than required for the

statement of the result.

The estimation procedure from Proposition 4.3.1 can also be augmented to flatten over multiple indices. For instance, the analyst may flatten such that both the rows and columns in the matrix contain multiple indices from the original array. Of course, this augmentation makes Assumption 4.3.3 harder to satisfy as it requires multiple parameters to vary in low-dimensional space. To see this take the tensor $\mathscr{A}$ flattened over the first two indices as $\mathscr{A}_{(1,2)} \in \mathbb{R}^{N_1 N_2 \times \prod_{n \notin \{1,2\}} N_n}$. If the parameters $\varphi^{(n)}$ for $n = 3, \ldots, d$ are high-dimensional, Assumption 4.3.3 is only satisfied when both $\varphi^{(1)}$ and $\varphi^{(2)}$ and their product space is low-dimensional. Clearly this is more restrictive than requiring only one of the parameter spaces to be low-dimensional. However, flattening along multiple dimensions can improve the convergence rate in Proposition 4.3.1 to $O_p\left(\frac{1}{\sqrt{\min\{N_1 N_2, \prod_{n \notin \{1,2\}} N_n\}}}\right)$, so there are benefits if this more restrictive assumption can be made. Further discussion of the matrix method results are relegated to Section 4.4.1, in particular some avenues to choosing the dimension to flatten over.

4.3.2 Group fixed-effects

This section describes the group fixed-effects estimator. Take again the model in array notation,

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{X}_k \beta_k^0 + \mathscr{A} + \boldsymbol{\varepsilon}.$$

A cluster assignment, $\mathscr{C}$, is a length d list of partition matrices, $C_n \in \{0,1\}^{N_n \times G_n}$, where G_n is the number of clusters in dimension n and each entry of C_n is a binary indicator of a unit's membership to a given cluster. Clusters are assigned separately along each dimension. Let $\Theta_{\mathscr{C}}$ be the space of group fixed-effects parameters associated to cluster assignment $\mathscr{C}$. Each $\boldsymbol{\theta} \in \Theta_{\mathscr{C}}$ is an ordered set of size d of $\bigtimes_{n=1}^{d} N_n$ tensors. For each n in $\{1, \ldots, d\}$, the tensor θ_n varies freely over dimensions $\{1, \ldots, n-1\}$ and $\{n+1, \ldots, d\}$ but is constant within each cluster along dimension n.[3] The objective function for the group fixed-effect estimation of β

[3] This parameter space is exemplified in Remark 4.3.1 for the three dimensional setting for clarity.

under cluster assignment $\mathscr{C}$ is

$$Q(\beta,\mathscr{C}) = \min_{\theta\in\Theta_{\mathscr{C}}} \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k\beta_k - \sum_{n=1}^{d} \theta_n \right\|_F^2 \tag{4.11}$$

and $\widehat{\beta}_{GFE,\mathscr{C}} := \text{argmin}_{\beta\in\mathbb{R}^K} Q(\beta,\mathscr{C})$.

The minimum within (4.11) is obtained from the within-cluster transformation in (4.3) from the Introduction. Remark 4.3.1 below details this objective function for the three-dimensional setting for clarity. It should be clear that the parameter space $\Theta_{\mathscr{C}}$ is indexed by cluster assignment $\mathscr{C}$ because this assignment defines how the parameters may vary. That is, this is the estimated parameter space under a specific group fixed-effects estimator, which may only be an approximation of the true parameter space.

The within-cluster transformation can be stated in more general fashion as follows. Take M_n to be an $N_n \times N_n$ matrix defined as $\mathbb{I}_{N_n} - C_n(C_n'C_n)^{\dagger}C_n'$, where † is the Moore-Penrose generalised inverse. Then, the within-cluster transformation can be formed by the following series of n-mode products, $\mathbf{Y} \times_1 M_1 \times_2 M_2 \times_3 \cdots \times_d M_d$. This sequentially differences out the group specific means from each dimension separately. Then the Frisch-Waugh-Lovell theorem straightforwardly applies.

Cluster assignments may be known or estimated, with suggestions of how to estimate these discussed in Section 4.4.2. In the case these are estimated from the error term, $\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k\beta_k$, there is an iteration between estimating cluster assignments and estimating slope coefficients, like in Section 4.3.1, but this can also be optimised with a grid space over β. Alternatively, estimated proxy variables used to form groups may come from the matrix method procedure, then no iteration is required. Both implementations are discussed in Section 4.4.5.

Assumption 4.3.5 (Clustering)**.**

Let $j_n(i_n)$ be any unit in the same cluster as i_n from using cluster assignment $\mathscr{C}$. Then,

(i). For all n as $N_n \to \infty$, $\frac{1}{N_n}\sum_{i_n=1}^{N_n} \left\| \varphi_{i_n}^{(n)} \right\|^2 \lesssim O_p(1)$, and,

(ii). For a non-empty subset $\mathscr{M} \subset \{1,\ldots,d\}$ take for any $n^* \in \mathscr{M}$ a sequence $\xi_{N_{n^*}} \to 0$ as $N_{n^*} \to \infty$. Then,

$$\frac{1}{N_{n^*}} \sum_{i_n=1}^{N_{n^*}} \left\| \varphi_{i_{n^*}}^{(n^*)} - \varphi_{j_{n^*}(i_{n^*})}^{(n^*)} \right\|^2 = O_p(\xi_{N_{n^*}})$$

Assumption 4.3.5.(i) restricts fixed-effects parameter space to have finite second moments. This implies that as $\{N_1,\ldots,N_d\} \to \infty$, cluster allocations cannot become increasingly disparate in the underlying parameter space. Assumption 4.3.5.(ii) states that for at least one dimension the clustering procedure finds matches with asymptotically negligible difference in the underlying parameter space. Since cluster assignments are not always estimated, and actually sometimes group assignments may be given extraneously, it is useful to state Assumption 4.3.5 in generic terms that ignore these clustering mechanics. These assumptions restrict the cluster assignment to uncover closeness in the true parameter space, which implies a restriction on the underlying parameter space and on how clusters are assigned.

Below is a refinement to the regularity conditions contained within the Assumptions listed in Section 4.3.1 that account for the within-cluster transformation.

Assumption 4.3.6 (Regularity conditions)**.** Let $\widetilde{T}_{i_1,\ldots,i_d}$ be the entries of tensor $\mathbf{T}$ after the group fixed-effects from the minimiser of (4.11) are differenced out. Then,

(i). $\left(\frac{1}{\prod_n N_n} \sum_{i_1} \cdots \sum_{i_d} \widetilde{X}_{i_1,\ldots,i_d} \widetilde{X}'_{i_1,\ldots,i_d}\right) = O_p(1)$ converges to a nonrandom positive definite matrix as $N_1,\ldots,N_d \to \infty$.

(ii). $\frac{1}{\prod_n N_n} \sum_{i_1} \cdots \sum_{i_d} \widetilde{X}_{i_1,\ldots,i_d} \varepsilon_{i_1,\ldots,i_d} = O_p\left(\frac{1}{\sqrt{\prod_n N_n}}\right)$.

Assumption 4.3.6.(i) is very similar to Assumption 4.3.4 except that here full rank is required after the within-cluster projection rather than the factor projection. Assumption 4.3.6.(ii) is an exogeneity condition that requires weak exogeneity in the covariates after the within-cluster transformation, which can be viewed as similar to Assumption 4.3.2. This is stricter than Assumption 4.3.2 because the noise

term ε can foreseeably impact cluster allocation if clusters are estimated as functionals of a residual term. This limitation is alleviated by, for instance, making sure cluster assignments are based on variables extraneous to the regression line, hence independent of ε, or perhaps through some sample splitting methods such as that proposed in Freeman and Weidner (2022).

Proposition 4.3.2 (Upper bound on group fixed-effects estimator)**.** *Let Assumptions 4.3.5 and 4.3.6 hold for cluster allocation $\mathscr{C}$. Let $\mathscr{M}$ be the set defined in Assumption 4.3.5.(ii). Then, for tensor rank L_N that may depend on sample size,*

$$\|\widehat{\beta}_{GFE,\mathscr{C}} - \beta^0\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathscr{M}} \sqrt{\xi_{N_{n^*}}}\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

Discussed in Section 4.4.2 are methods and restrictions that restrict $\xi_{N_{n^*}}$ from Assumption 4.3.5 and Proposition 4.3.2 to $1/N_{n^*}$. This suggests that as long as $\mathscr{M} = \{1, \ldots, d\}$ and L_N is bounded the parametric rate of convergence is achievable. Related to $\mathscr{M}$, an implicit requirement on the latent parameters from this subset of dimensions is some form of low-dimensionality in the vectors $\varphi_{i_{n^*}}^{(n^*)}$ for $n^* \in \mathscr{M}$. For a discussion on the curse of dimensionality using clustering methods, see Bonhomme et al. (2021) that suggests the dimension of these parameters should be ≤ 2 to be well-clustered. A sufficient condition for parameters in these dimension to be low-dimensional is low multilinear rank for each $n^* \in \mathscr{M}$. Hence, it is expected that the set $\mathscr{M}$ should be a subset of $\mathscr{L}$, from Assumption 4.3.3. The advantage with the group fixed-effects methods is that the analyst does not need to choose which n does admit low multilinear rank, hence it is more flexible. However, since the matrix factor methods do not suffer such a large curse of dimensionality, if the low multilinear rank dimension is known then there is still some advantage in using this method, for example if the smallest multilinear rank parameter is bounded but of order 5-10.

Remark 4.3.1. *In the three dimensional setting, the group fixed-effect objective*

function,

$$Q(\beta, \mathbf{g}) = \min_{\alpha,\gamma,\delta} \sum_{i,j,t} \left(Y_{ijt} - X'_{ijt}\beta - \theta_{1;g_1(i)jt} - \theta_{2;ig_2(j)t} - \theta_{3;ijg_3(t)}\right)^2 \tag{4.12}$$

where $\theta_1 \in \mathbb{R}^{ncol(g_1)\times N_2\times N_3}$, $\theta_2 \in \mathbb{R}^{N_1\times ncol(g_2)\times N_3}$, *and* $\theta_3 \in \mathbb{R}^{N_1\times N_2\times ncol(g_3)}$; *and* $ncol(\cdot)$ *returns the number of columns of a matrix. Then, taking the within-cluster transformation on* $\mathbf{Y}$ *and* $\mathbf{X}$ *from* (4.3) *is equivalent to differencing out the minimisers from* (4.12). $\mathbf{g}$ *is a list of group assignment for each dimension. For example,* $g_1(i)$ *maps to the group identity of individual i. This is why* θ_1 *is restricted to vary across only* $ncol(g_1)$ *different values in the first dimension, which is less than* N_1.

Notice that the optimisers for θ_1, θ_2 *and* θ_3 *from* (4.12) *can be described as combinations of the within-cluster projection as follows,*

$$\begin{aligned}
\widehat{\theta}_{1;g_1(i)jt} &\approx \bar{\mathscr{A}}_{i^*jt} - \bar{\mathscr{A}}_{i^*j^*t} + \bar{\mathscr{A}}_{i^*j^*t^*} \\
\widehat{\theta}_{2;ig_2(j)t} &\approx \bar{\mathscr{A}}_{ij^*t} - \bar{\mathscr{A}}_{i^*j^*t} \\
\widehat{\theta}_{2;ijg_3(t)} &\approx \bar{\mathscr{A}}_{ijt^*} - \bar{\mathscr{A}}_{i^*jt^*},
\end{aligned}$$

though this representation is not unique.

Additional to controlling for any additive terms, this projection leaves the following interactive fixed-effects residual,

$$\widetilde{\mathscr{A}_{ijt}} = \sum_{\ell=1}^{L} (\varphi_{i\ell}^{(1)} - \bar{\varphi}_{i^*\ell}^{(1)})(\varphi_{j\ell}^{(2)} - \bar{\varphi}_{j^*\ell}^{(2)})(\varphi_{t\ell}^{(3)} - \bar{\varphi}_{t^*\ell}^{(3)}). \tag{4.13}$$

where $\bar{\varphi}_{i^*\ell}^{(1)}$ *is the group mean of* $\varphi_{i^*\ell}^{(1)}$ *for the* i^**'s in i's group, and so on for the other terms. Hence, sufficient projection of the interactive fixed-effects terms relies on the weaker condition that parameters converge to their group means, namely,* $\varphi_{i\ell}^{(1)} \to \bar{\varphi}_{i^*\ell}^{(1)}$, $\varphi_{j\ell}^{(2)} \to \bar{\varphi}_{j^*\ell}^{(2)}$ *or* $\varphi_{t\ell}^{(3)} \to \bar{\varphi}_{t^*\ell}^{(3)}$ *for each* ℓ. *Indeed, the group mean differencing could be seen as a weighted mean difference across the population, with equal weight given to observations within the cluster and zero weight to observations outside of each cluster. This fact is utilised for the more generic kernel weighted*

difference estimator in Section 4.3.3, which is synonymous to a Nadaraya-Watson type estimator for each fixed-effects term.

So far it has been shown that with the relatively innocuous shift from the within transformation to the within-cluster transformation, any additive terms are automatically controlled for and there are conditions to also control for the interactive term. Choice of clusters for this transformation is key to suffice this less restrictive condition. Given a set of proxies to cluster on, clustering or matching methods can be used to find these groups, for example Bonhomme et al. (2021). Developing a set of proxies to cluster on is important and is discussed in Section 4.4.2.

4.3.3 Kernel weighted fixed-effects

Let $\widehat{\varphi}_{i_n}^{(n)}$ generically denote a proxy measure for unit i_n in dimension n that may be known or estimated. The use of this notation will become clear in the statement of Proposition 4.3.3 and in discussion of how to estimate these proxy measures in Section 4.4.2. Let $\mathscr{W}$ be an ordered set of weight matrices, where the n^{th} item $W_{(n)} \in \mathbb{R}^{N_n \times N_n}$ has elements,

$$w_{i_n,j_n}^{(n)} := \frac{k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)}{\sum_{i_n'=1}^{N_n} k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{i_n'}^{(n)}\right\|\right)}, \tag{4.14}$$

where k is a kernel function, and h_n is a bandwidth parameter. Let Δ be a d-list of $\bigtimes_{n=1}^{d} N_n$ tensors $\delta_n \in \mathbb{R}^{N_1 \times \cdots \times N_d}$. For a given set of proxy measures and kernel function, the kernel weighted fixed-effects estimator optimises

$$S(\beta, \mathscr{W}) = \min_{\delta \in \Delta} \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k \beta_k - \sum_{n=1}^{d} \delta_n \times_n W_{(n)} \right\|_F^2. \tag{4.15}$$

Then, $\widehat{\beta}_{KER,\mathscr{W}} := \operatorname{argmin}_{\beta \in \mathbb{R}^K} S(\beta, \mathscr{W})$. The notation $\times_n$ is the n-mode product defined at the beginning of this section. Like in the discrete group case, these weighted fixed-effects can be projected out using the weighted-within version of (4.3), where discrete groups are replaced with weights. Again, if proxy measures are estimated from the error term $\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k \beta_k$, then there is an iteration between slope estima-

tion and estimation of kernel weights, much like in the other estimators presented already. This can also be optimised over a grid space for β if β is low-dimensional. Alternatively, estimates used to calculate weights may come from the matrix method procedure, and again no iteration is required. Both implementations are discussed in Section 4.4.5.

Like the within-cluster transformation, the weighted-within transformation can also be stated in more general fashion as follows. Take M_n to be an $N_n \times N_n$ matrix defined as $\mathbb{I}_{N_n} - W_{(n)}(W'_{(n)}W_{(n)})^{\dagger}W'_{(n)}$, where † is the Moore-Penrose generalised inverse. Then, the weighted-within transformation can be formed by the following series of n-mode products, $\mathbf{Y} \times_1 M_1 \times_2 M_2 \times_3 \cdots \times_d M_d$. This sequentially differences out the weighted means from each dimension separately. Then, again, the Frisch-Waugh-Lovell theorem straightforwardly applies.

Assumption 4.3.7 (Kernels)**.** Denote the kernel function used as $k(\cdot)$ and let this function be bounded. Then for $a \geq 0$ and $h > 0$ there exists an $\alpha > 0$ such that $k(a/h)a \lesssim O(h^{\alpha})$.

Assumption 4.3.7 refers to a bandwidth parameter, h, and restricts the kernels to penalise distance at a rate equal to or faster than $O(h^{\alpha}/a)$. For consistency using the kernel methods, the sequence $h \to 0$ is considered, such that an upper bound on α is the critical object of interest.

As an example of a class of kernel functions that satisfies Assumption 4.3.7, the exponential class of the form considered in Remark 4.3.2 may be utilised.

Remark 4.3.2. *For $c_1, c_2 > 0$, let $k'(a) \propto c_1 exp(-c_2 a^2)$ for all $a \geq 0$ and $k' \in \mathscr{K}'$. Then* $\operatorname{argmax}_a k'(a/h)a = h/\sqrt{2c_2}$*, and,*

$$\max_a k'(a/h)a \propto \frac{c_1}{\sqrt{2c_2}} e^{-1/2} h = O(h)$$

Thus, Assumption 4.3.7 is satisfied for the exponential class of kernel functions $\mathscr{K}'$ with $\alpha = 1$. Further, for $h \to 0$, it suffices that $\alpha \in (0, 1]$.

Assumption 4.3.7 is stated more generically than Remark 4.3.2 as there is a larger class of bounded kernel functions that satisfy the sufficient restriction for the

result below. The point here is to show that Assumption 4.3.7 is satisfied for some very standard kernel functions, hence is not too restrictive.

Assumption 4.3.8 (Regularity of proxy measures)**.** Let $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$ be the proxy space for the fixed-effects and let $k(\cdot)$ be a bounded kernel function. Let $K_{i_n}(h_n) := \max_{j_n} k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)$. For $0 < e_{i_n} < K_{i_n}(h_n)$ define

$$M_n\left(\widehat{\varphi}_{i_n}^{(n)}, e_{i_n}\right) := \sum_{j=1}^{N_n} \mathbb{1}\left(k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right) > e_{i_n}\right)$$

Then for any $e_{i_n} \in (0, K_{i_n}(h_n))$,

$$\operatorname*{plim}_{N_n \to \infty} \frac{M_n\left(\widehat{\varphi}_{i_n}^{(n)}, e_{i_n}\right)}{N_n} \geq c_{i_n}^{(n)} \in (0,1]. \tag{4.16}$$

for all $i_n \in 1, \ldots, N_n$.

The upper bound on e_{i_n}, K_{i_n}, is expected to be $k(0)$ for most classes of kernels. That is, the kernel function evaluated at $\widehat{\varphi}_{i_n}^{(n)} = \widehat{\varphi}_{j_n}^{(n)}$ should maximise the value of the kernel function. For example, the Gaussian kernel function is maximised at $k(0) = (1/\sqrt{2\pi})$. An example of a low-level condition for Assumption 4.3.8 is presented in Remark 4.3.3.

Assumption 4.3.8 is a restriction on the data generating process of the fixed-effect proxy parameter space. This is similar to Assumption 5.5 in Altonji and Matzkin (2005), except in this case related to the fixed-effect parameter space. Note that for these to be satisfied, the probability over the support of the fixed-effect space must be strictly positive. Whilst this paper focuses on fixed-effects, that is, effects that are taken as given and not modelled as random variables, it is still useful to understand that these parameters are sampled from some space. This is the space that the restriction in Assumption 4.3.8 pertains to.

Assumption 4.3.8 places a restriction on the bounds of the kernel function and on the space of proxy measures used. The restriction on the proxy measures may be satisfied if they are generated such that the neighbourhood around each realisation grows proportionally with the sample size, such as in Remark 4.3.3,

Remark 4.3.3 (Regularity of proxy measures)**.** *Let* $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$ *and redefine* $M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right)$ *as*

$$M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right) := \sum_{j_n=1}^{N_n} \mathbb{1}\left(\widehat{\varphi}_{j_n}^{(n)} \in B_{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right)\right),$$

where $B_{\varepsilon}(x)$ *is the* ε*-neighbourhood around x. Then Assumption 4.3.8 is satisfied for the dimension n for any* $\varepsilon > 0$ *if,*

$$\operatorname*{plim}_{N_n \to \infty} \frac{M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n^*)}\right)}{N_n} \geq c_{i_n}^{(n)} \in (0,1].$$

for all $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$.

Defined in Remark 4.3.3 is essentially the building blocks of the probability space for the fixed-effects proxy parameter space. This is a lower level restriction than Assumption 4.3.8 in that it relates to the space of the proxy measures without reference to the kernel functions used.

Assumption 4.3.9 (Regularity conditions)**.** Let $\check{T}_{i_1,\ldots,i_d}$ be the entries of tensor $\mathbf{T}$ after the kernel weighted fixed-effects from the minimiser of (4.15) are differenced out. Then,

(i). $\left(\frac{1}{\prod_n N_n}\sum_{i_1}\cdots\sum_{i_d}\check{X}_{i_1,\ldots,i_d}\check{X}'_{i_1,\ldots,i_d}\right) = O_p(1)$ converges to a nonrandom positive definite matrix as $N_1,\ldots,N_d \to \infty$.

(ii). $\frac{1}{\prod_n N_n}\sum_{i_1}\cdots\sum_{i_d}\check{X}_{i_1,\ldots,i_d}\varepsilon_{i_1,\ldots,i_d} = O_p\left(\frac{1}{\sqrt{\prod_n N_n}}\right)$.

Assumption 4.3.9 is exactly Assumption 4.3.6 but with the kernel weighted fixed-effects in place of the group fixed-effects.

Proposition 4.3.3 (Upper bound on kernel estimator)**.** *Let the class of kernel functions used to formulate the weights and the proxy measure used in these kernel functions for the kernel weighted fixed-effects estimator satisfy Assumption 4.3.7 and 4.3.8. Also, let Assumption 4.3.9 hold for the set of regressors. Let* $\frac{1}{N_{n^*}}\sum_{i_{n^*}}\left\|\varphi_{i_{n^*}}^{(n^*)} - \widehat{\varphi}_{i_{n^*}}^{(n^*)}\right\|^2 = O_p(C_{n^*}^{-2})$ *for* $n^* \in \mathscr{M}'$ *and* $\frac{1}{N_{n'}}\sum_{i_{n'}}\left\|\varphi_{i_{n'}}^{(n')} - \widehat{\varphi}_{i_{n'}}^{(n')}\right\|^2 = O_p(1)$ *for* $n' \notin \mathscr{M}'$*, where* $\mathscr{M}'$ *is a non-empty subset of dimensions. Let* h_n *be the*

bandwidth parameter from Assumption 4.3.8. Then, for L_N that may depend on sample size,

$$\left\|\widehat{\beta}_{KER,\mathscr{W}} - \beta^0\right\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathscr{M}'} \sqrt{O_p\left(C_{n^*}^{-2}\right) + O_p\left(h_{n^*}^{2\alpha}\right)}\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

For $h_n^\alpha \lesssim O(C_n^{-1})$ this reduces to

$$\left\|\widehat{\beta}_{KER,\mathscr{W}} - \beta^0\right\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathscr{M}'} O_p\left(C_{n^*}^{-1}\right)\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

Proposition 4.3.3 shows that the convergence rate for the kernel estimator is bounded by the convergence rate of the proxy estimates. That is, as long as the bandwidth parameter approaches zero sufficiently fast, the kernel estimator converges at a rate no worse than the convergence of the proxies when proxies are estimations of the true parameter values at or slower than $\sqrt{N_n}$-convergence. This is expected and also a good result that the kernel method does not hinder the convergence rate from these proxies. These kernel methods do, however, suffer a curse of dimensionality since Assumption 4.3.8 becomes increasingly difficult to justify as the dimension of the fixed-effects increases. Discussions in Section 4.4.2 suggest $O_p(C_{n^*}^{-1})$ can be $O_p\left(1/\sqrt{N_{n^*}}\right)$. This shows the parametric rate is attainable if $\mathscr{M}' = \{1, \ldots, d\}$ and L_N is fixed.

Remark 4.3.4. *The motivation for the kernel weighted fixed-effect estimator is very similar to the group fixed-effect estimator. Take* (4.3)*, stated again here in the three-dimensional setting for the kernel weighted transformation,*

$$\check{Y}_{ijt} = Y_{ijt} - \bar{Y}_{i^*jt} - \bar{Y}_{ij^*t} - \bar{Y}_{ijt^*} + \bar{Y}_{i^*j^*t} + \bar{Y}_{i^*jt^*} + \bar{Y}_{ij^*t^*} - \bar{Y}_{i^*j^*t^*}.$$

In the Introduction, the within-cluster transformation took the average within each starred indices' cluster. For the kernel weighted difference this average is instead taken as a weighted average over the whole sample. For example, the term $Y_{i^*j^*t}$

from this is,

$$Y_{i^*j^*t} = \sum_{i'=1}^{N_1} \sum_{j'=1}^{N_2} w^{(1)}_{i,i'} w^{(2)}_{j,j'} Y_{i'j't},$$

where $w^{(1)}_{i,i'}$ *and* $w^{(2)}_{j,j'}$ *are defined in* (4.14)*. The arguments for the group fixed-effects estimator then translate directly to the kernel weighted fixed-effect estimator, where smooth weights are applied instead of the binary weights implied by the within-cluster differencing.*

4.4 Discussion of estimators

This section serves to discuss the results in Section 4.3, motivate further some of the chosen methods, and provide some methods to estimate cluster assignments or proxies for kernel weights. A few iteration procedures are also discussed at the end of this section.

4.4.1 Matrix method results

As stated already, Proposition 4.3.1 takes for granted the dimension to flatten across admits a low rank interactive fixed-effect term for the least square method in Bai (2009). Under Assumption 4.3.1.(ii) the singular values of the flattened normalised noise term dissipates as follows;

$$\frac{1}{\sqrt{\prod_n N_n}} \left\| \varepsilon_{(n)} \right\| = O_p \left(\frac{1}{\sqrt{\min\{N_n, \prod_{m \neq n} N_m\}}} \right).$$

Since $\mathscr{A}$ is a collection of fixed-effects, the normalised singular values of its flattenings are $O_p(1)$, that is, the singular values are not asymptotically negligible like those of the noise term.[4] This ensures that, after flattening $\mathscr{A}$, each of the singular

[4]To see this consider the standard two dimension model and take the Frobenius norm any arbitrary component of the interactive fixed-effects term, $\lambda_r f_r'$, normalised by $1/\sqrt{NT}$

$$\frac{1}{\sqrt{NT}} \left\| \lambda_r f_r' \right\|_F = \sqrt{\frac{1}{NT} \sum_i \sum_t (\lambda_{ir} f_{tr})^2} = \sqrt{\frac{1}{N} \sum_i \lambda_{ir}^2} \sqrt{\frac{1}{T} \sum_t f_{tr}^2} = O(1).$$

The last equality comes from λ_{ir} and f_{tr} being bounded fixed-effects.

values eventually dominate those of the noise term. These conditions make up similar restrictions imposed in Ahn and Horenstein (2013) that allow for the use of the eigenvalue ratio test (ER) to diagnose the number of factors. Hence, in large samples, the analyst maybe able to use this test or similar to not only decide how many factors to use but also decide which dimension is likely to be low-dimensional.

Consider the factor model applied to a flattening that may not be low-rank. For a concrete example of when this can occur see the data generating process in the simulations in Section 4.5, where $\varphi^{(n)}$ are designed to be low-dimensional for some n, and high-dimensional otherwise. Along the dimensions of $\mathscr{A}$ that do not conform to the low-rank assumption in Assumption 4.3.3, the tail singular values may become difficult to discern from the singular values of the noise term in small samples. This means variation from those tail factors are less likely to be projected out from the factor model unless many factors are used in this projection. If r_n for $n \notin \mathscr{L}$ is allowed to increase adversely, for example at exactly the upper bound, then factor projection may never sufficiently project all relevant factors. Also, as the number of estimated factors increases, Assumption 4.3.4 becomes harder to satisfy since variation in the set of covariates is also projected out. This demonstrates the importance of choosing the correct dimension to flatten over, which is supported by the simulation results in Section 4.5.

Hence, a standard factor model that estimates at least r_n factors should result in consistent estimation of the slope coefficients, see Moon and Weidner (2015). However, this relies on an important structural feature of the unobserved heterogeneity term. When flattened in the chosen dimension – the first dimension in the above example – the rank of the matrix after flattening must be low relative to data size. This implies that to successfully project out the variation in the fixed-effect term either the matrix of fixed-effects from any flattening is low-rank, or, at least one flattening leads to a low-rank matrix of fixed-effects and the analyst knows which flattening this is. To use the above example again, this means the analyst knows that $\varphi^{(1)}\Gamma_1'$ is low-rank, hence flattening in the first dimension is the correct way to recast the model to a panel data model, and so forth for the other flattenings. Whilst

requiring low-rankness in at least one dimension may be an acceptable restriction, having knowledge of which dimension this low-rankness resides in is potentially more restrictive.

To understand the problem, consider the following two examples, one where the flattening is not low-rank and one where it is. First, assume $\varphi^{(1)}$ varies in a high-dimensional parameter space, e.g. with $N_1 < N_2N_3$, $\varphi^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ and $\Gamma \in \mathbb{R}^{N_2N_3 \times N_1}$ with each column mutually orthogonal for both these matrices. Then the product of these matrices is full-rank and any factor projection approach will not fully control for this term. On the contrary, consider $\varphi^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ where all columns are linearly dependent. Then the matrix $\varphi^{(1)}\Gamma'$ is rank-1 regardless of L and of how $\varphi^{(2)}$ and $\varphi^{(3)}$ vary, thus can be projected with a factor model estimated with 1 factor. Hence it is important which dimension the analyst chooses to flatten over.

Well established diagnostics in Bai and Ng (2002), Ahn and Horenstein (2013) and Hallin and Liška (2007) can be used to determine the number of factors. These diagnostics can be repeated across different flattenings, which may be informative of the dimension to use for flattening. Note these procedures require an initial guess of β and relies on this guess not eradicating the factor structure in the residual; see the beginning of Section 4.4.5 for a concrete example of this. It should also be noted that these diagnostics are not without restrictions and can lead to spurious conclusions on the optimal number of factors. For example, the eigenvalue ratio test in Ahn and Horenstein (2013) can undershoot the number of factors when singular values decay quickly for the leading few factors. This does not interfere with the asymptotic result in that paper but can have implications in small sample estimation. Indeed, however, these diagnostics can be helpful in both the matrix recasting of the problem and the group fixed-effects estimation in the sequel.

4.4.2 Estimating cluster and kernel proxies

Discussed here are some important functionals of multidimensional arrays that are useful for estimating proxies to cluster on or use for kernel weights. This includes a discussion on how to uncover proxies from multidimensional array data, and a dis-

cussion on why standard matrix methods do not extend well to the multidimensional setting.

First, consider how to cluster in the within-cluster transformation. In most clustering algorithms, for example K-means or K-nearest neighbour, there is some notion of a distance metric between units considered for each cluster. To arrive at a distance there must be some space to measure that distance over. For example, using some vector u_i and the Euclidean norm of differences, $\|u_i - u_j\|$, to measure the distance between units. Algorithms to arrive at these groupings are well established when the distance metric and variable to take distance over are given. However, in this setting there is no clear variable through which to take distance over. Motivated here are methods to extract proxies that serve to measure distance across units in a way that isolates variation in each dimension of the unobserved heterogeneity term.

It is important to find proxies that isolate variation in each dimension since clustering is to be performed one index at a time. Discussed here are decompositions of multidimensional arrays that can perform this, Kolda and Bader (2009) contains a nice summary of some candidate decompositions. The method discussed here uses the higher order singular value decomposition (HOSVD), and focuses on components of this decomposition that have well formulated theoretical properties. The HOSVD is traditionally used in pursuit of a low-rank tensor decomposition by either direct truncation of left singular vectors or by some iteration approach similar to this, see for example the higher order orthogonal iteration scheme. The problem of direct truncation, however, is not well-posed because the solution to the low-rank tensor problem may not be unique and reformulating the original tensor after the aforementioned truncation is not guaranteed to be lower tensor rank. See De Silva and Lim (2008) for an extensive explanation of the ill-posedness issues. Hence, this method cannot be used in the pursuit of analytic consistency results. Problems also arise in this setting where the reformulated tensors can be arbitrarily well approximated by a tensor of lower tensor rank, which is a result of the border rank issue of the tensor rank decomposition.

Reconsider the three dimensional model with heterogeneity of the following

form

$$\mathscr{A}_{ijt} = \sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}. \tag{4.17}$$

As shown in De Silva and Lim (2008), the Eckart–Young–Mirsky theorem cannot be relied upon to guarantee an optimal low-rank approximation for the multidimensional array $\mathscr{A}$. This motivates the use of the group fixed-effects and kernel weighted fixed-effects as alternative solutions.

Also reconsider the singular value decomposition for matrices, applied to each of the n-flattenings of $\mathscr{A} \in \mathbb{R}^{N_1 \times,\ldots,\times N_d}$ as

$$\mathscr{A}_{(n)} = U^{(n)} \Sigma_n V^{(n)\prime}. \tag{4.18}$$

By the same logic as in the matrix case and formalised with the Eckart-Young-Mirsky theorem, variation over the rows of each $U^{(n)}$ explains variation over the n^{th} dimension of the multidimensional array $\mathscr{A}$. Thus, if $\mathscr{A}_{(n)}$ is low rank, the leading few columns of $U^{(n)}$ provide good proxies for closeness in n^{th} dimension of $\mathscr{A}$. If $\mathscr{A}_{(n)}$ is not low rank then these leading columns still provide the best proxies for bias reduction using the group or kernel fixed-effects. Hence, by reconsidering the tensor problem as a sequence of matrix problems, the usual singular value decomposition properties can be utilised for this reduction. This shows that a simple rearrangement of the data provides readily available techniques to measure closeness in each dimension separately.

Consider for any of the dimensions n the corresponding matrix of left singular vectors from above, $U^{(n)}$, estimated with noise ε_{ijt}. That is, each U_n are calculated from the object $\mathscr{V} = \mathscr{A} + \boldsymbol{\varepsilon}$. Under reasonable regularity conditions on the noise term $\boldsymbol{\varepsilon}$, the left singular vectors from this decomposition comprise of a signal of the underlying fixed-effect parameter and noise from $\boldsymbol{\varepsilon}$. For example, in the three dimensional case, define the L_1-vector $\widehat{U}_i^{(1)}$ as the i-th row of the left singular matrix

of $\mathscr{A} + \boldsymbol{\varepsilon}$ flattened in the first dimension. Then the vector $\widehat{U}_i^{(1)}$ may comprise of,

$$\widehat{U}_i^{(1)} = \varphi_i^{(1)} + O_p\left(\frac{1}{\sqrt{\min\{N_1, N_2N_3\}}}\right).$$

Likewise, for any dimension n define $\widehat{U}^{(n)}$ as the matrix of singular vectors from $\mathscr{A} + \boldsymbol{\varepsilon}$ flattened in the n-th dimension, where $\mathscr{A}$ is the unobserved fixed-effects component of interest and $\boldsymbol{\varepsilon}$ is the usual idiosyncratic noise term. Then, Bai and Ng (2002) detail conditions required for the following "up-to-rotation" consistency result, which has been amended to this paper's setting;

Lemma 4.4.1 (Theorem 1 from Bai and Ng (2002))**.** *For any fixed integer $k \geq 1$, there exists an $(r_n \times k)$ matrix H_n^k with $rank(H_n^k) = \min\{k, r_n\}$ and $C_n = \min\left\{\sqrt{N_n}, \prod_{n' \neq n}\sqrt{N_{n'}}\right\}$ such that for each n under some regularity conditions*

$$C_n^2 \left\| \widehat{U}_{i_n}^{(n)} - H_n^{k\prime} \varphi_{i_n}^{(n)} \right\|^2 = O_p(1).$$

This establishes a consistency result for estimating cluster proxies and suggests these left singular vectors are viable options to cluster in each dimension if the true error, $\mathscr{V} = \mathscr{A} + \boldsymbol{\varepsilon}$, is observed. It also makes concrete the limitation implied by the value of C_n for each index - that short indices have poorly estimated proxies. Given that the error term displayed in (4.13) is multiplicative across dimension, the error from this poor approximation should become negligible as long as enough other dimension proxies are well estimated. Also, the presence of the rotation matrices, H_n^k, in Lemma 4.4.1 can be ignored since these do not change relative distances of each unit under standard distance metrics used in either the cluster or kernel methods.

However, it is not necessarily justified to assume the true error, $\mathscr{V} = \mathscr{A} + \boldsymbol{\varepsilon}$, is observed. In fact, an estimate of the error $\widehat{\mathscr{V}}_{ijt} = Y_{ijt} - X_{ijt}\widehat{\beta} = X_{ijt}(\beta^0 - \widehat{\beta}) + \mathscr{A}_{ijt} + \varepsilon_{ijt}$, clearly depends on the estimate $\widehat{\beta}$, hence should be bound by the rate of convergence for this estimate. For this reason, the uniform convergence result from Lemma 4.4.1 is unlikely to be useful. Proposition A.1 in Bai (2009) does provide

bounds for the mean squared deviation,

$$\frac{1}{N_n}\sum_{i_n=1}^{N_n}\left\|\widehat{U}_{i_n}^{(n)}-H_n^{k\prime}\varphi_{i_n}^{(n)}\right\|^2=O_p(\|\beta^0-\widehat{\beta}\|^2)+O_p\left(\frac{1}{\min\{N_n,\prod_{n'\neq n}N_{n'}\}}\right),$$

if $\widehat{U}_{i_n}^{(n)}$ come from the Bai (2009) least squares problem in each dimension. This establishes, ignoring H_n^k, $\frac{1}{N_n}\sum_{i_n=1}^{N_n}\left\|\widehat{U}_{i_n}^{(n)}-\varphi_{i_n}^{(n)}\right\|^2=O_p\left(\frac{1}{N_n}\right)$ if each dimension sample size grows at the same rate and the convergence rate for $\|\beta^0-\widehat{\beta}\|$ in Proposition 4.3.1 is used. Then, C_n^{-2} from Proposition 4.3.3 is $1/N_n$ and the parametric rate for the kernel weighted estimator can be established.

4.4.3 Group fixed-effect convergence result

Before discussing the result from Proposition 4.3.2, an alternative restriction on the cluster assignments is proposed. If clustering is performed on a proxy measure of the fixed-effect then Assumption 4.3.5 can be stated in terms of the proxies, which forms the statement of Remark 4.4.1. This requires that the proxies form an injective mapping to the true fixed-effect parameters. An example of this are the conditions imposed in Freeman and Weidner (2022), stated in similar terms here:

Remark 4.4.1 (Clustering)**.** *The statement of Assumption 4.3.5 can be reformulated in terms of the cluster proxies as follows. Let* $\widehat{\varphi}_{i_n}^{(n)}:=\widehat{\varphi}^{(n)}\left(\varphi_{i_n}^{(n)}\right)\in\mathbb{R}^{\widehat{r}_n}$ *be the proxy for individual* i_n *used to cluster along dimension n. Then,*

(i). For all n as $N_n\to\infty$*,*

$$\frac{1}{N_n}\sum_{i_n}^{N_n}\left\|\widehat{\varphi}_{i_n}^{(n)}-\widehat{\varphi}_{j_n(i_n)}^{(n)}\right\|^2\lesssim O_p(1)$$

(ii). For a non-empty subset $\mathscr{M}\subset\{1,\ldots,d\}$ *take for any* $n^*\in\mathscr{M}$ *a sequence* $\xi_{N_{n^*}}\to 0$ *as* $N_{n^*}\to\infty$*. Then,*

$$\frac{1}{N_{n^*}}\sum_{i_{n^*}}^{N_{n^*}}\left\|\widehat{\varphi}_{i_{n^*}}^{(n^*)}-\widehat{\varphi}_{j_{n^*}(i_{n^*})}^{(n^*)}\right\|^2=O_p(\xi_{N_{n^*}})$$

(iii). Let $\varphi_{i_n}^{(n)}\in\Phi_n$ *be the* r_n*-column vector of fixed effects, where* Φ_n *are convex*

sets for each n. For each $a, b \in \Phi_n$ *there exists a scalar* $c_n > 0$ *such that* $\|a-b\| \leq c_n \cdot \left\|\widehat{\varphi}^{(n)}(a) - \widehat{\varphi}^{(n)}(b)\right\|$

If these alternate restrictions hold along with Assumption 4.3.6, then the bound in Proposition 4.3.2 holds for the GFE estimator.

Restrictions (i) and (ii) in Remark 4.4.1 are exactly Assumption 4.3.5.(i) and (ii) but with cluster proxies in place of the true parameter values. These are high level restrictions on the clustering mechanism that requires the mechanism to find closeness in the proxy space. Restriction (iii) in Remark 4.4.1 is an injectivity assumption on the proxy functions that demands closeness in the underlying parameter space given closeness in the proxy space. This requires that the proxies do actually provide a mapping to the true parameter space, that is, that they are reasonable proxies. An example of proxies that do this are the singular vectors from Section 4.4.2 that fit the requirements of Lemma 4.4.1. To see this expand the term $\|a-b\|$ and use the triangle inequality to see, $\|a-b\| \leq \|a-\widehat{\varphi}^{(n)}(a)\| + \|\widehat{\varphi}^{(n)}(b)-b\| + \|\widehat{\varphi}^{(n)}(a)-\widehat{\varphi}^{(n)}(b)\|$, where the first two terms are bound at the rate $O_p(C_n^{-1})$. Note the rotation matrices are ignored for brevity and C_n is the convergence rate from Lemma 4.4.1. Hence, asymptotically, Remark 4.4.1.(iii) can be achieved with $c_n = 1$. Again, this argument requires knowledge of the true error term $\mathscr{V} = \mathscr{A} + \boldsymbol{\varepsilon}$, which may be unreasonable. By similar discussion from the kernel weighted estimator, this condition might relax to the mean squared deviation from Section 4.4.2, but is not formally discussed further here.

This display also makes clear the bottle-neck when clustering in high-dimensional objects. The distance of the proxies, $\|\widehat{\varphi}^{(n)}(a) - \widehat{\varphi}^{(n)}(b)\|$, is difficult to bound using clustering methods when the dimension of the proxies are larger than two, see Graf and Luschgy (2002) and further discussion in Bonhomme et al. (2021). This implies that a low-dimensional set of proxies must bound the true parameter values for clustering methods to work well in this setting. Hence, whilst the relationship in restriction (iii) of Remark 4.4.1 may be satisfied for an arbitrarily high-dimensional set of proxies, for a reasonable family of cluster mechanisms to bound these proxies as per restrictions (i) and (ii) of this remark, restriction (iii)

must also hold for a low-dimensional set of proxies. This can be highly restrictive with, for example, fixed-effects $\varphi_{i_n}^{(n)}$ that are high-dimensional.

How the sequences $\xi_{N_{n^*}}$ converges to zero and how L_N is bounded are important for the convergence result in Proposition 4.3.2. First note that for fixed L_N the first term in the result simplifies to $O_p\left(\prod_{n^*\in\mathscr{M}}\sqrt{\xi_{N_{n^*}}}\right)$. Also note, if the conditions for Lemma 4.4.1 hold and clustering is based on singular vector estimates that adhere to Remark 4.4.1, then it is possible to achieve $\xi_{N_{n^*}} = O_p\left(\frac{1}{\min\{N_{n^*},\prod_{n\neq n^*}N_n\}}\right)$. If each N_n grow at the same rate then the consistency result is,

$$\left\|\widehat{\beta}_{GFE,\mathscr{C}} - \beta^0\right\| = \sqrt{L_N}O_p\left(N_n^{-|\mathscr{M}|/2}\right).$$

In the worst case scenario $\sqrt{L_N}$ is upper bound by $\sqrt{L_N} \lesssim N_n^{(d-1)/2}$, which is taken from $L_N \leq \min_n \prod_{n'\neq n} N_{n'}$. The convergence result is then $O_p\left(N_n^{(d-1-|\mathscr{M}|)/2}\right)$, which is of course conservative but shows that if $|\mathscr{M}| = d$, then consistency is guaranteed albeit at the slow rate of $N_n^{1/2}$. This means that all dimensions must have good cluster assignments, which is obviously not an ideal worst case but shows the limitations of this method when L_N is unrestricted.

For the special case of $d = 3$ it can be shown that $L_N \leq \min_n \prod_{n'\neq n} r_{n'}$. From the discussion above, it is expected that $n \in \mathscr{M}$ is sufficient for $n \in \mathscr{L}$, that is, r_n is small for the set of dimensions $n \in \mathscr{M}$. This tightens the bound in Proposition 4.3.2 to $O_p\left(N_n^{\max\{-|\mathscr{M}|/2,1-|\mathscr{M}|\}}\right)$, such that only $|\mathscr{M}| \geq 1$ is required for consistency. The analogous tensor rank bound is so far not known for the case with $d \geq 4$.

4.4.4 Curse of Dimensionality

The curse of dimensionality appears in both the group fixed-effects estimator and the kernel weighted fixed-effects estimator. To see this, revisit the objective when finding groups or weights, which is to find closeness in the vector space of $\varphi_{i_n}^{(n)}$ using some proxies or estimates, $\widehat{\varphi}_{i_n}^{(n)}$. For $\varphi_{i_n}^{(n)} \in \mathbb{R}^L$ it is expected that closeness around each $\varphi_{i_n}^{(n)}$ is increasingly difficult as L increases, which is a standard result in nonparametric analysis.

Take for example the condition in Remark 4.3.3 that demands the neighbour-

hood of each $\varphi_{i_n}^{(n)}$ is well populated. As a concrete example, if $\varphi_{i_n}^{(n)} \sim U(0,1)^L$, then $\mathbb{P}\{\varphi_{j_n}^{(n)} \in B_\varepsilon(\varphi_{i_n}^{(n)})\} = O(\varepsilon^L)$. Then for a fixed sized neighbourhood around $\varphi_{i_n}^{(n)}$, call it $\widetilde{N}_{i_n}^{(n)}$, the dimension size N_n must grow exponentially in L for this to stay fixed in expectations. To see this note $\mathbb{E}(\widetilde{N}_{i_n}^{(n)}) = N_n O(\varepsilon^L)$, such that $N_n = O(\varepsilon^{-L})$ to keep $\mathbb{E}(\widetilde{N}_{i_n}^{(n)})$ fixed. This shows that for generically distributed fixed-effects parameters the regularity condition for the proxy measures in Assumption 4.3.8 suffers exponentially from the curse of dimensionality.

The group fixed-effect estimator also suffers as it uses quantization methods to approximate the true fixed-effect parameter values. This is thoroughly discussed in Graf and Luschgy (2002) and Bonhomme et al. (2021), which conclude each fixed-effect should have dimension less than or equal two.

It may be possible to break the curse of dimensionality by noting, without loss of generality, each entry in the vector for $\varphi_{i_n}^{(n)}$ can be written as mutually orthogonal entries, if just one dimension is considered. Then, the weighted-within transformations may be run sequentially to purge variation in each member of the vector $\varphi_{i_n}^{(n)}$, for example using a backfitting style algorithm from Breiman and Friedman (1985). What is not yet apparent is if the steps in this algorithm can be straightforwardly applied to the weighted within transformations proposed here, and if the vectors, $\varphi_{i_n}^{(n)}$, can truly be written as the sum of orthogonal parts when all dimensions are considered simultaneously. Hence, this line of exploration is left for future research.

4.4.5 Implementation

Estimation can be performed either as a simple least squares problem with groups or weights pre-estimated or by an iterative procedure. Discussed below is a suggestion of how to pre-estimate weights along with two possible iteration procedures. Both approaches are applicable to the kernel weighted fixed-effect estimator and the group fixed-effect estimator.

First, the kernel weighted fixed-effect estimator with weights estimated from the matrix methods procedure is detailed. Optimise the two-dimensional least squares objective function $R(\beta, \widehat{r}_n, n)$ from (4.9) for each dimension n. Obtain the estimates of $\varphi_{i_n}^{(n)}$ from each of these estimation procedures. Use the estimates $\widehat{\varphi}_{i_n}^{(n)}$

to form the kernel weights in (4.14) then perform the weighted-within transformation on $\mathbf{Y}$ and $\mathbf{X}$, then perform pooled OLS on the transformed data. This procedure effectively utilises the computational simplicity of the matrix methods to find closeness in each dimension fixed-effect before projecting these out directly. This is convenient because it also avoids any unnecessary iteration from cluster methods or from trying to directly estimate the error term.

Second, an iterative procedure is discussed that could be used for either estimator, but is detailed just for the group fixed-effect estimator. The rest of the subsection is dedicated to a few options on how to perform this. Numerical results suggest the iterative procedure performs well, which suggests errors are well estimated.

Consider taking cluster proxies from the estimated error term $\mathbf{W} = \mathbf{Y} - \mathbf{X}'\widetilde{\beta}$. Define $\widetilde{\beta}$ as the interim estimator used to obtain $\mathbf{W}$, and notice that this forms the basis of an iterative procedure, between forming clusters and estimating slope coefficients. This is illustrated in the following two-step procedure. For the below let $\widehat{r}_n$ be a hyperparameter that defines the number of singular vectors to use in the clustering stage.

1. For given $\widetilde{\beta}$, take the left singular matrices from each n-flattening of $\mathbf{W} = \mathbf{Y} - \mathbf{X}'\widetilde{\beta}$ to obtain $\left\{\widehat{U}_n\right\}_{n=1}^{d}$.

2. Cluster on the leading $\widehat{r}_n$ columns of $\widehat{U}_n$ to generate cluster assignments in the n^{th} dimension. Use these cluster assignments in the within-cluster transformation on $\mathbf{Y}$ and $\mathbf{X}$ then perform pooled OLS to obtain $\widehat{\beta}$.

3. Iterate steps 1 and 2 until convergence in the slope coefficients

This procedure may also be used as a debias estimator for a given initial estimate of $\widetilde{\beta}$ by ignoring step 3. Iteration here may not be stable given that step 1 and 2 do not optimise the same objective function, hence for theoretical purposes it may be convenient to only consider this as a debias procedure. In practice, iterating between step 1 and 2 after some initial grid search to initialise β may be optimal.

Of course, other clustering or transformations may be used in place of the residual clustering and within-cluster transformation. In the below, two alternatives

are provided. The first maintains the within-cluster transformation but considers a different set of proxies. The second approach considers a kernel weighted transformation procedure that uses a generic set of proxies. At this stage and in the below estimator refinements the analyst may be concerned with the number of parameters required to conduct these transformation. Appendix C.2 discusses a number of ways to reduce the size of the parameter space, including only projecting fixed-effects over a subset of dimensions and letting group sizes increase to reduce the number of groups.

Whether used as an iterative scheme or an update, the above method has some identification issues. As an illustration take the data generating process for model (4.1) with just one covariate,

$$X_{ijt} = -\mathscr{A}_{ijt} + \mu_{ijt},$$

where μ_{ijt} is a white noise term. Consider an initial guess of $\widetilde{\beta} = 0$ when the true value is $\beta^0 = 1$. This leaves the residual term from Step 1 to base cluster assignment on as, $\mathbf{W} = \mathbf{Y} - \mathbf{X}\widetilde{\beta} = \mathbf{Y}$, which reduces to $\mathbf{W} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$. Thus, clustering is based solely on noise and can be reasonably described as random. The associated within-cluster transformation will not project variation in the $\mathscr{A}$ terms that appear in both $\mathbf{Y}$ and $\mathbf{X}$ such that the OLS step in stage 2 produces

$$\widehat{\beta} \approx \frac{Var(\mu_{ijt})}{Var(\mu_{ijt}) + Var(\mathscr{A}_{ijt})} + o_p(1).$$

For $\frac{Var(\mu_{ijt})}{Var(\mathscr{A}_{ijt})} \to 0$, $\widehat{\beta} \to 0$ and the algorithm does not update the initial guess of $\widetilde{\beta} = 0$. This problem also arises in the matrix methods in Section 4.3.1 and is a more fundamental issue with this algorithmic approach. For this reason in practice it is important that the variation in X_{ijt} is not completely dominated by the fixed-effect term $\mathscr{A}_{ijt}$, i.e. there is a non-negligible source of variation coming from the term μ_{ijt}.

This example clearly displays some identification issues with the above method. Worth noting is that this may be alleviated with a grid search approach,

though this can be computationally infeasible even for a moderate number of covariates since the grid grows exponentially in the number of covariates. To avoid this, proposed below is a method to extract cluster allocations from only variation in the set of covariates. As discussed below, this clustering may also be conducted on control variables extraneous to the regression line. The two-step procedure works as follows.

1. Take the left singular matrices from each n-flattening of X to obtain $\left\{\widehat{U}_n\right\}_{n=1}^{d}$.

2. Cluster on the leading $\widehat{r}_n$ columns of $\widehat{U}_n$ to generate cluster assignments in the n^{th} dimension. Use these cluster assignments in the within-cluster transformation on $\mathbf{Y}$ and $\mathbf{X}$ then perform pooled OLS to obtain $\widehat{\beta}$.

An advantage of using covariate clustering is that it can make use of control variables that are a good signal of cluster but are not included in the regression line. For example, a control variable Z_i that is constant across j and t may be a good candidate to cluster along the i dimensions but will be projected out with the within-cluster transformation, so cannot be used directly in the pooled OLS estimation of β stage. This refinement also makes optimisation over β a convex problem, and no iteration is required because clustering is not a function of β estimates like in the first iteration procedure.

4.5 Simulation

Table 4.1 shows simulation results for the following DGP,

$$Y_{ijt} = X_{ijt}\beta + \mathscr{A}_{ijt} + \mathscr{B}_{ijt} + \varepsilon_{ijt}$$
$$X_{ijt} = \mathscr{A}_{ijt} + \mathscr{B}_{ijt} + \nu_{ijt}$$

with $\mathscr{A}_{ijt} = \sum_{\ell=1}^{N_1} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$, $\mathscr{B}_{ijt} = \alpha_{ij} + \gamma_{it} + \delta_{jt}$. Also,

$$\varepsilon_{ijt}, \nu_{ijt}, \alpha_{ij}, \gamma_{it} \text{ and } \delta_{jt} \overset{i.i.d.}{\sim} N(0,1) \text{ and for each } \ell,\ \varphi_{i\ell}^{(1)}, \varphi_{t\ell}^{(3)} \overset{i.i.d.}{\sim} N(0,1).$$
$$\varphi_{j1}^{(2)} \overset{i.i.d.}{\sim} N(0,1) \text{ with } \varphi_{j1}^{(2)} = \varphi_{j2}^{(2)} = \cdots = \varphi_{jN_1}^{(2)}$$

$\mathscr{A}_{ijt}$ and $\mathscr{B}_{ijt}$ are normalised to have unit variance. $\mathscr{A}$ is specified such that it is rank 1 when flattened in the second dimension and rank N_1 when flattened in either dimension one or three. That is, the multilinear rank is $\mathbf{r} = (N_1, 1, N_1)$. This comes directly from the data generating process for each $\varphi^{(n)}$, where the matrix $\varphi^{(2)}$ is designed to be rank-1 and the matrices $\varphi^{(1)}$ and $\varphi^{(3)}$ are designed to be rank-N_1.

In Table 4.1, the estimators OLS and Fixed-effects are simply the pooled OLS estimator and the pooled OLS estimator after additive fixed-effects are projected out, respectively. As expected both of these two have poor bias. The four GFE estimators perform well with reasonably low bias and standard deviation. GFE (K-means) is GFE estimator with clustering based on the K-means algorithm, with proxies taken from the residual. GFE (K-means on X) is the same estimator with proxies taken from the scalar covariate of interest. GFE (1-NN) and GFE (1-NN on X) are likewise the same estimators but using the one nearest neighbours clustering. The factor model is used after first flattening along each dimension as Factor(dim = n), where n is the dimension used for flattening. In each case, 2 factors are projected. The results show the theoretical result succinctly, where the bias is close to zero when the correct dimension is flattened over (the second dimension in this case) and very poor bias when the incorrect dimension is used (the first and third dimensions). Lastly, the kernel differencing estimator is estimated with Gaussian kernel function with bandwidths 0.5, 1 and 1.5; which are standardised to be equivalent to standard deviations of the proxy measures. All kernel estimators have comparable bias but substantially better standard deviation for bandwidth equal 1 and 1.5.

This analysis is repeated for the four dimensional case in Table 4.2, where the second and third dimensions admit low-dimensional unobserved interactive fixed-effects parameters. For computational reasons, the GFE nearest neighbour estimators are omitted. The simulations suggest similar results as the three dimensional case, where the factor models perform well when flattened in the low-dimensional dimensions (second and third) and poorly in the high-dimensional dimensions (first and fourth).

3-D	Mean bias	St. dev.	MSE
OLS	0.6668	0.0033	4.45e-01
Fixed-effects	0.4997	0.0114	2.50e-01
GFE (K-means)	0.0118	0.0096	2.32e-04
GFE (K-means on X)	0.0129	0.0112	2.91e-04
GFE (1-NN)	0.0112	0.0153	3.61e-04
GFE (1-NN on X)	0.0111	0.0154	3.61e-04
Kernel (h = 0.5)	0.0030	0.0090	8.94e-05
Kernel (h = 1.0)	0.0031	0.0068	5.58e-05
Kernel (h = 1.5)	0.0037	0.0062	5.24e-05
Factor (dim = 1)	0.4319	0.0135	1.87e-01
Factor (dim = 2)	0.0030	0.0050	3.40e-05
Factor (dim = 3)	0.4319	0.0135	1.87e-01

Table 4.1: 3D model ($N_1 = N_2 = N_3 = 36$), with 10,000 Monte Carlo rounds. All results are in relation to β estimation.

4-D	Mean bias	St. dev.	MSE
OLS	0.6670	0.0018	4.45e-01
Fixed-effects	0.4981	0.0282	2.49e-01
GFE (K-means)	0.0012	0.0049	2.58e-05
GFE (K-means on X)	0.0013	0.0051	2.82e-05
Kernel (h = 0.5)	5.23e-05	0.0114	1.00e-04
Kernel (h = 1.0)	4.42e-05	0.0057	3.26e-05
Kernel (h = 1.5)	-1.32e-05	0.0045	2.03e-05
Factor (dim = 1)	0.3733	0.0311	1.40e-01
Factor (dim = 2)	0.0030	0.0030	1.82e-05
Factor (dim = 3)	0.0030	0.0030	1.82e-05
Factor (dim = 4)	0.3734	0.0311	1.40e-01

Table 4.2: 4D model ($N_1 = N_2 = N_3 = N_4 = 20$), with 10,000 Monte Carlo rounds.

A two-dimensional simulation exercise is also performed to compare the grouped fixed-effects approach to the factor model approach in a setting where theoretical results for the factor model are well known. Table 4.3 shows the results of this two-way setting where the data generating process is a factor model with two factors. The GFE estimators have less bias than the factor model even when the factor model overestimates the number of factors. To see this, compare the factor estimates with 2, 4 and 6 factors projected out with the GFE estimator. For increase in variance of order ≈ 4, the GFE estimator reduces bias by an order ≈ 10. This is a surprising improvement in estimates for a setting that is purpose designed for the

factor model. Where this comparison falls down is for models with a larger number of factors because generally clustering does not perform well when the latent parameter space has dimension greater than 2.

	Mean bias	St. dev.	MSE
OLS	0.6672	0.0033	4.45e-01
Fixed-effects	0.5000	0.0043	2.50-e01
GFE	0.0002	0.0090	8.05e-05
Kernel (h = 0.5)	0.0003	0.0055	3.06e-05
Kernel (h = 1.0)	0.0003	0.0054	2.87e-05
Kernel (h = 1.5)	0.0004	0.0053	2.81e-05
Factor (R = 2)	0.0024	0.0047	2.76e-05
Factor (R = 4)	0.0031	0.0048	3.25e-05
Factor (R = 6)	0.0024	0.0049	3.03e-05

Table 4.3: 2D model ($N_1 = N_2 = 216$), with 10,000 Monte Carlo rounds.

4.6 Empirical application - demand estimation for beer

The methods proposed in this paper are applied to estimated the demand elasticity for beer. Price and quantity for beer sales is taken from the Dominick's supermarket dataset for the years 1991-1995 and is related to supermarkets across the Chicago area. Price and quantity vary over three dimensions in this example – product (i), store (j) and month (t). Fixed-effects that interact across all three dimensions can control for taste shocks to beer consumption that differ over both product and store. Take for instance a large sporting event (temporary t shock) that changes preferences differently across locations (j) and across certain subsets of sponsored beer (i). For example, in the stadiums for the many NBA finals playoffs the Chicago Bulls played in the early 1990's, Miller Lite beer advertisements could be seen alongside advertisements for the substitute product Canadian Club whisky. This suggests these events attracted large marketing campaign spends for these and other beer substitute brands that most likely also included price offers at local supermarkets. Whilst the impact of these advertisements and price offers on the demand for or price of beer is not clear and, further, that it is reasonably safe to assume the

econometrician does not observe the plethora of marketing campaigns around these events, the analyst would most likely still want to control for aggregate shocks like these. For this reason it is important to use methods that robustly control for unobserved fixed-effects, such as unobserved marketing campaigns, that may impact both quantity demanded and prices in unforeseen ways.

Models for demand estimation ideally account for endogenous variation in prices and quantity. The classic instrumental variable approach is to find a variable that varies exogenously to the production process but can reasonably describe price fluctuations. A popular instrument in the estimation of beer demand is the commodity price for barley, one of the product's main ingredients, see e.g. Saleh (2014); Tremblay and Tremblay (1995); Richards and Rickard (2021). Since the price of barley is arguably not driven by the demand for it by any one supplier of beer, it can be a useful variable to instrument for price shifts. In the following, it is taken as given that the price of barley is exogenous with respect to the noise term, ε.

For validity the instrument is also required to be strong, in the sense that it is strongly correlated with price. In this dataset correlation between the price of barley, which varies over only t, and price of beer depends on how beer price is aggregated. If beer price is first integrated over i and j, such that it only varies over t, then it is highly correlated with the price of barley, at 0.61. However, if beer price is not aggregated at all it is only correlated at 0.05. This suggests there are important product and store level price drivers for beer that are not accounted for by fluctuations in the price of barley. This implies that price fluctuations in barley alone may not be viable to fully capture beer prices when considering variation over all three dimensions. For exogeneity, the price of barley must be independent of common unobserved shocks to both price and demand, which translates to being independent of $\varphi^{(3)}_{t,\ell}$ and any scalar fixed-effects that vary over t in the interactive fixed-effects model.

The second column from Table 4.4 refers to the estimates for demand elastici-

ties for the following regression model,

$$\log(quantity_{ijt}) = \log(price_{ijt})\beta + \mathscr{A}_{ijt} + \varepsilon_{ijt} \tag{4.19}$$

where $\mathscr{A}_{ijt}$ is the usual interactive fixed-effects term from the prequel. This amounts to estimating the standard log-log model for demand with fixed-effects. That is,

$$quantity_{ijt} = price_{ijt}^{\beta} \exp(\mathscr{A}_{ijt} + \varepsilon_{ijt}).$$

Again, no controls are included here since they are low-dimensional and subsumed by the fixed-effects term. This model specification estimates reasonably similar elasticities as the logit case across each of the different fixed-effects estimators but relatively large differences in estimates for pooled OLS and IV. The similar elasticities for the different fixed-effects estimators within Table 4.4 again suggests that whilst some form of fixed-effects should be included, they may not need be as complex as implied by the GFE and kernel methods.

The third column from Table 4.4 reports estimates of the same log-log model controlling for the average log price of other products,

$$\log(quantity_{ijt}) = \log(price_{ijt})\beta + \delta \sum_{i' \neq i} \log(price_{i'jt}) + \mathscr{A}_{ijt} + \varepsilon_{ijt}. \tag{4.20}$$

This model assumes homogeneous cross-elasticity over all other beer products. That is, it refers to the demand model,

$$quantity_{ijt} = price_{ijt}^{\beta} \prod_{i' \neq i} price_{i'jt}^{\delta_{ii'}} \exp(\mathscr{A}_{ijt} + \varepsilon_{ijt}),$$

where $\delta_{ii'} = \delta$ for all i and i'. Whilst this may oversimplify the system of cross-elasticities in the market for beer, it does significantly change the estimates for β in the log-log model. This suggests that cross-elasticities should probably be controlled for since β estimates do seem sensitive to their inclusion. It also shows that for a covariate with full rank variation over all dimensions, not even the more

complex fixed-effects estimators can control for these. Note that most estimators returned a negative value for δ, which opposes the theory that other brands of beer, on aggregate, are substitutes. However, since prices are aggregated in such a crude way, the cross-elasticity estimates should not be taken too seriously. If interested in the cross-elasticities, then some care should be taken to segment or group products in such a way that actual substitution is being identified here, not just aggregate market forces. For this model, all fixed-effects estimates are within statistical noise of each other, this time with the control variable approach being closely aligned. These are also similar to the own-price elasticity estimates from Table 1 in Hausman et al. (1994). IV is estimated with very high variation in both log-log models, which may be due to barley being a weak instrument, or due to losing the richness in variation over products and stores after first-stage fitting of prices.

Estimator	$\widehat{\beta}$ (St. dev.) no cross elas.	$\widehat{\beta}$ (St. dev.) with cross elas.
Pooled OLS	1.18 (0.31)	1.22 (0.32)
Pooled IV	-4.87 (1.69)	-4.04 (1.34)
Additive Fixed-effects	-1.86 (0.31)	-3.10 (0.30)
Factor (dim = 1)	-1.60 (0.26)	-2.90 (0.27)
Factor (dim = 2)	-1.83 (0.30)	-3.07 (0.30)
Factor (dim = 3)	-2.17 (0.30)	-3.18 (0.29)
GFE	-1.85 (0.33)	-2.86 (0.30)
Kernel (Gaussian)	-1.83 (0.28)	-2.92 (0.29)

Table 4.4: Log-log demand elasticities (73 products, 41 stores, 57 months).

Standard deviations were bootstrapped by resampling along each dimension separately. In the first dimension, product 1 is fixed across bootstrap samples. Column 2 displays estimates for the model (4.19) with no cross elasticities. Column 3 displays estimates for the model (4.20), which controls for cross elasticities.

Table 4.5 refers to estimates from the standard logit demand model,

$$\log(quantity_{ijt}) - \log(quantity_{1jt}) = price_{ijt}\beta + \mathscr{A}_{ijt} + \varepsilon_{ijt}$$

where $\mathscr{A}_{ijt}$ is the usual interactive fixed-effects and no covariates are included since the set of available covariates are rank-deficient and automatically projected out with standard scalar fixed-effects and from differencing out the outside option. The outside option is encoded as product number 1 and is the aggregate

consumption of products with small quantities consumed. This serves the purpose of creating an outside option to do the necessary logit demand transformation as well as to avoid issues related to an unbalanced panel for the many niche products with sparse consumption amounts. Own price elasticity is calculated as $\eta_{ijt} = price_{ijt}\beta(1 - quantity_{ijt}/\sum_{ijt} quantity_{ijt})$ and the mean elasticity is taken as the mean of this measure for each estimator. The pooled instrumental variable approach estimates relatively large elasticities, but with much higher standard errors. All of the fixed-effects approaches estimate statistically similar slope coefficients and elasticities at the mean. This implies that whilst some fixed-effects may exist in the true model for demand, they are unlikely complex enough to require the high-dimensional projections from the GFE or kernel methods. To robustly test for the existence of fixed-effects in an IV model there must be an instrument with variation over all dimensions such that fixed-effects can be projected out alongside the IV model. This of course also takes for granted that the IIA logit model is the true model for demand.

Estimator	Coefficient (bootstrap st. dev.)	Elasticity at mean
Pooled OLS	-0.60 (0.04)	-3.26 (0.22)
Pooled IV	-0.72 (0.27)	-3.91 (1.49)
Additive FE	-0.32 (0.05)	-1.74 (0.27)
Factor (dim = 1)	-0.29 (0.04)	-1.58 (0.22)
Factor (dim = 2)	-0.32 (0.05)	-1.74 (0.27)
Factor (dim = 3)	-0.37 (0.05)	-2.01 (0.27)
GFE	-0.32 (0.05)	-1.74 (0.27)
Kernel (Gaussian)	-0.30 (0.05)	-1.63 (0.27)

Table 4.5: Logit demand estimates (73 products, 41 stores, 57 months).

Standard deviations were bootstrapped by resampling along each dimension separately. In the first dimension, product 1 is fixed across bootstrap samples as the outside option and the remaining products are resampled with replacement.

4.7 Conclusion

This paper develops methods to generalise the interactive fixed-effect to multi-dimensional datasets with more than two dimensions. Theoretical results show that standard matrix methods can be applied to this setting but require additional

knowledge of the data generating process. The multiplicative interactive error from the group fixed-effects and kernel methods show a potential improvement on the asymptotic rate of convergence and suggest a more robust approach to projecting fixed-effects. Simulations corroborate these theoretical results and show the relative advantage of using a standard factor model when the structure of the interactive term is known. They also show the robustness of the group fixed-effects and kernel weighted fixed-effects estimators to not having this same knowledge. Inference in these models is still an open question for further research.

The model is applied to a simple demand model for beer consumption. The application demonstrates that if fixed-effects do exist in this setting, they are unlikely to be complex enough to require the GFE or kernel methods to control for them. This is a useful analysis, as it provides a robustness check for the specification of fixed-effects in model specifications. It also showed that in datasets with many dimensions, the instrumental variable approach can be limited if the instrument used only varies over a subset of dimensions, hence is weak.

Appendix A

Appendix – Chapter 2

A.1 Proofs

We start with a preliminary result that relates the nuclear norm of $\mathbf{\Gamma}^{\infty}(x)$ with the sum of the singular values of the function $(\boldsymbol{a},\boldsymbol{b}) \mapsto m(x,\boldsymbol{a},\boldsymbol{b})$. This link will be useful to bound the approximation error of $\widehat{\mathbf{\Gamma}}(x)$. We define

$$\|m(x,\cdot,\cdot)\|_* := \sum_{j=1}^{\infty} s_j(x).$$

Lemma A.1.1. *Let Assumptions 2.3.1 and 2.3.2 hold. Then, as $N,T \to \infty$,*

$$\|\mathbf{\Gamma}^{\infty}(x)\|_1 \leq \sqrt{NT}\,\|m(x,\cdot,\cdot)\|_* + o_P(\sqrt{NT}) = O_P(\sqrt{NT}).$$

Lemma A.1.1 implies that $\|\mathbf{\Gamma}^{\infty}(x)\|_1$ grows with N and T at the same rate as any low-rank matrix $\boldsymbol{M}$ with elements that are of order one with bounded second moments such that $\|\boldsymbol{M}\|_1 \leq \sqrt{\text{rank}(\boldsymbol{M})}\,\|\boldsymbol{M}\|_2 = \sqrt{\text{rank}(\boldsymbol{M})\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}^2} = O_P(\sqrt{NT})$. This result will be useful for the proofs of Lemma 2.3.1 and of Theorem 2.3.4. The proof of Lemma A.1.1 is provided at the end of the appendix.

The following technical lemma provides the key step in the proof of Lemma 2.3.1 in the main text.

Lemma A.1.2. *Under Assumptions 2.3.1 and 2.3.2,*

$$\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \left(\widehat{\Gamma}_{it}(x) - \Gamma^{\infty}_{it}(x)\right)^2 \leq \frac{2\rho \, \|\boldsymbol{\Gamma}^{\infty}(x)\|_1}{n(x)} - \frac{2}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma^{\infty}_{it}(x) E_{it},$$

for all $\rho \geq \|\boldsymbol{E}(x)\|_{\infty}$.

Notice that Lemma A.1.2 is a non-stochastic finite sample result, which only requires that $E_{it}(x)$ and $\widehat{\boldsymbol{\Gamma}}(x)$ are as defined in (2.14) and (2.15). The proof of Lemma A.1.2 is provided at the end of the appendix. We are now ready to provide the proof of the lemma in the main text.

Proof of Lemma 2.3.1:

The definition of $E_{it}(x)$ in (2.14) guarantees that $\mathrm{E}\left[E_{it}(x) \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] = 0$, and Assumption 2.3.3 furthermore guarantees that $E_{it}(x)$ is independent across i and t and has a finite fourth moment, conditional on $\boldsymbol{X}^{NT}$, $\boldsymbol{A}^N$ and $\boldsymbol{B}^T$. Furthermore, $\Gamma^{\infty}_{it}(x) = m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$ only depends on $\boldsymbol{A}^N$ and $\boldsymbol{B}^T$. We therefore find

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma^{\infty}_{it}(x) E_{it}\right)^2 \middle| \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] \\
&\quad = \frac{1}{n^2(x)} \sum_{(i,t)\in\mathbb{D}(x)} [\Gamma^{\infty}_{it}(x)]^2 \, \mathrm{E}\left[E^2_{it} \,\middle|\, \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] \\
&\quad \leq \frac{b^{1/2}}{n^2(x)} \sum_{(i,t)\in\mathbb{D}(x)} [\Gamma^{\infty}_{it}(x)]^2 = O_P(1/n(x)),
\end{aligned}$$

where b is the constant from Assumption 2.3.3. From this we conclude that

$$\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma^{\infty}_{it}(x) E_{it} = O_P\left(\frac{1}{n^{1/2}(x)}\right) = o_P(1). \tag{A.1}$$

Next, applying Assumption 2.3.3 and Theorem 2 in Latała (2005) we find

$$\begin{aligned}
\mathrm{E}\left[\|\boldsymbol{E}(x)\|_\infty \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] \leq C\Bigg\{ & \max_t \sqrt{\sum_i \mathrm{E}\left[E_{it}(x)^2 \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right]} \\
& + \max_i \sqrt{\sum_t \mathrm{E}\left[E_{it}(x)^2 \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right]} \\
& + \left(\sum_{i,t} \mathrm{E}\left[E_{it}(x)^4 \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right]\right)^{1/4} \Bigg\} \\
\leq C b^{1/4} & \left\{\sqrt{N} + \sqrt{T} + n(x)^{1/4}\right\} = O_P\left(\sqrt{N+T}\right),
\end{aligned}$$

where C is a universal constant. We therefore have $\|\boldsymbol{E}(x)\|_\infty = O_P\left(\sqrt{N+T}\right)$, and since we assume that $\rho = \rho_{NT}$ satisfies $\rho_{NT}/\sqrt{N+T} \to \infty$ we conclude that

$$\rho_{NT} \geq \|\boldsymbol{E}(x)\|_\infty$$

with probability approaching one. We can therefore apply Lemma A.1.2 to find that, with probability approaching one, we have

$$\begin{aligned}
\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \left(\widehat{\Gamma}_{it}(x) - \Gamma^\infty_{it}(x)\right)^2 &\leq \frac{2\rho_{NT}\,\|\boldsymbol{\Gamma}^\infty(x)\|_1}{n(x)} - \frac{2}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \Gamma^\infty_{it}(x) E_{it} \\
&= \frac{2\rho_{NT}\, O_P(\sqrt{NT})}{n(x)} + o_P(1) \\
&= o_P(1),
\end{aligned}$$

where we applied (A.1) and Lemma A.1.1, as well as the condition $\rho_{NT}\sqrt{NT}/n(x) \to 0$.

□

In the following consider a generic reduced form parameter

$$\nu_0(x) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T W_{it}(x)\, \Gamma^\infty_{it}(x), \tag{A.2}$$

with corresponding estimator

$$\widehat{\nu}(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)\widehat{\Gamma}_{it}(x), \tag{A.3}$$

where $W_{it}(x)$ are given weights.

The following proposition provides a finite-sample non-stochastic bound for the error of this reduced form estimator.

Proposition A.1.1. *Let the Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. Let $P_{it}(x)$ be non-zero real numbers for all $(i,t) \in \mathbb{N}\times\mathbb{T}$. Define*

$$\begin{aligned}
V_{it}(x) &:= \frac{W_{it}(x)P_{it}^{-1}(x)(D_{it}(x)-P_{it}(x))}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)^2 P_{it}^{-1}(x)},\\
c_1 &:= \frac{1-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)P_{it}^{-1}(x)V_{it}(x)}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)^2 P_{it}^{-1}(x)},\\
c_2 &:= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}(x)\Gamma_{it}^{\infty}(x),\\
c_3 &:= \frac{2\rho}{c_1 NT}\|\mathbf{\Gamma}^{\infty}(x)\|_1 - \frac{2}{c_1 NT}\sum_{(i,t)\in\mathbb{D}(x)} E_{it}(x)\Gamma_{it}^{\infty}(x) + \left(\frac{c_2}{c_1}\right)^2,\\
c_4 &:= \sqrt{c_3} + \frac{|c_2|}{c_1},
\end{aligned}$$

and let $\boldsymbol{V}(x)$ be the $N\times T$ matrix with elements $V_{it}(x)$. If $c_1 > 0$ and $\rho > \|\boldsymbol{E}(x)\|_\infty + c_4\|\boldsymbol{V}(x)\|_\infty$, then

$$|\widehat{\nu}(x) - \nu_0(x)| \leq c_4.$$

The proof of Proposition A.1.1 is provided at the end of the appendix. Proposition A.1.1 is the key step required for the proof of Theorem 2.3.4. However, before proving this main text result we want to provide an informal remark on the usefulness of Proposition A.1.1 more generally.

Remark A.1.1 (Consistency of $\widehat{\nu}(x)$)**.** Proposition A.1.1 holds for all $P_{it}(x) \in \mathbb{R}\setminus\{0\}$, but for the proposition to be useful in showing consistency of $\widehat{\nu}(x)$ we need

to choose $P_{it}(x)$ such that c_2 and $\|\boldsymbol{V}(x)\|_\infty$ are not too large. The easiest way to guarantee this is to consider X_{it} to be random and weakly correlated across both i and t, and to define $P_{it}(x)$ as the propensity score, that is,

$$P_{it}(x) = \Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right),$$

which is assumed to be positive and not too small — e.g. we need that

$$q := \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)^2 P_{it}^{-1}(x)\right]^{-1}$$

converges to some positive constant. Then $V_{it}(x)$ has mean zero, analogous to $E_{it}(x)$, and

$$\begin{aligned}
c_1 &= q + O_P(1/\sqrt{NT}),\\
c_2 &= O_P(1/\sqrt{NT})\\
c_3 &= \frac{2\rho}{qNT}\|\boldsymbol{\Gamma}^\infty(x)\|_1 + O_P(1/\sqrt{NT}),\\
c_4 &= \sqrt{\frac{2\rho}{qNT}\|\boldsymbol{\Gamma}^\infty(x)\|_1} + \text{smaller order terms.}
\end{aligned}$$

Thus, if, like in Lemma 2.3.1, $\rho = \rho_{NT}$ such that $\rho_{NT}/\sqrt{N+T} \to \infty$ and $\rho_{NT}/\sqrt{NT} \to 0$ as $N, T \to \infty$, then

$$\widehat{\nu}(x) = \nu_0(x) + o_P(1).$$

The following proof formalizes this heuristic argument for the case that $W_{it}(x) = 1$.

Proof of Theorem 2.3.4: Let $W_{it}(x) = 1$, and let $\nu_0(x)$ and $\widehat{\nu}(x)$ be as defined in (A.2) and (A.3) above. We then have

$$\begin{aligned}
&\mu(x) = \nu_0(x),\\
&\widehat{\mu}(x) = \widehat{\nu}(x) + \frac{1}{NT}\sum_{(i,t)\in\mathbb{D}(x)} E_{it}(x) - \frac{1}{NT}\sum_{(i,t)\in\mathbb{D}(x)}\left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^\infty(x)\right]. \qquad \text{(A.4)}
\end{aligned}$$

We drop all the arguments x in the rest of this proof. We want to apply Proposition A.1.1 with $P_{it} = \Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right) > 0$. Let $G_{it} = P_{it}^{-1}(D_{it} - P_{it})$ be as defined in Theorem 2.3.4, and also define $q := \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{it}^{-1}\right]^{-1}$. Since $P_{it} \in [0,1]$ we also have $q \in [0,1]$, and the theorem assumes that $q^{-1} = O_P(1)$. Using Lemma A.1.1 we know that $\|\boldsymbol{\Gamma}^{\infty}\|_1 = O_P(\sqrt{NT})$, and we have already found that $\sum_{(i,t)\in\mathbb{D}} \Gamma_{it}^{\infty} E_{it} = O_P\left(n^{1/2}\right)$ in (A.1) above. Using this together the other assumptions in the theorem we find that

$$
\begin{aligned}
V_{it} &= q\,G_{it} \\
c_1 &= q\left(1 - \frac{q}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{it}^{-1} G_{it}\right) = q\,[1 - o_P(1)], \\
c_2 &= \frac{q}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} G_{it}\,\Gamma_{it}^{\infty} = o_P(1), \\
c_3 &= \frac{2\rho\, O_P(\sqrt{NT})}{c_1 NT} - \frac{O_P\left(n^{1/2}\right)}{c_1 NT} + \left(\frac{c_2}{c_1}\right)^2 = o_P(1), \\
c_4 &= \sqrt{c_3} + \frac{|c_2|}{c_1} = o_P(1).
\end{aligned}
$$

We furthermore have

$$
\|\boldsymbol{V}\|_{\infty} = q\,\|\boldsymbol{G}\|_{\infty} = O_P(1)\,O_P(\sqrt{N+T}) = O_P(\sqrt{N+T}).
$$

In the proof of Lemma 2.3.1 we already argued that $\|\boldsymbol{E}\|_{\infty} = O_P\left(\sqrt{N+T}\right)$. Since we assume that $\rho = \rho_{NT}$ satisfies $\rho_{NT}/\sqrt{N+T} \to \infty$ we conclude that

$$
\rho > \|\boldsymbol{E}\|_{\infty} + c_4\|\boldsymbol{V}\|_{\infty}
$$

with probability approach one. We can therefore apply Proposition A.1.1 to find that with probability approach one we have

$$
|\widehat{\nu} - \nu_0| \leq c_4 = o_P(1).
$$

We have thus shown that $\widehat{\nu} = \nu_0 + o_P(1)$.

Furthermore, analogous to the result in (A.1) we can show that $\sum_{(i,t)\in\mathbb{D}} E_{it} = O_P\left(n^{1/2}\right)$, and we therefore have $\frac{1}{NT}\sum_{(i,t)\in\mathbb{D}} E_{it} = o_P(1)$. Finally, applying Lemma 2.3.1 we have Next, from we know that

$$\left[\frac{1}{n}\sum_{(i,t)\in\mathbb{D}}\left(\widehat{\Gamma}_{it}-\Gamma_{it}^{\infty}\right)\right]^2 \leq \frac{1}{n}\sum_{(i,t)\in\mathbb{D}}\left(\widehat{\Gamma}_{it}-\Gamma_{it}^{\infty}\right)^2 = o_P(1),$$

and therefore $\frac{1}{NT}\sum_{(i,t)\in\mathbb{D}(x)}\left[\widehat{\Gamma}_{it}(x)-\Gamma_{it}^{\infty}(x)\right] = o_P(1)$. Plugging those result into (A.4) we find $\widehat{\mu}(x) = \mu(x) + o_P(1)$.

□

In this section we present and prove a more general version of Theorem 2.4.2. Let $\boldsymbol{\phi}_i = \boldsymbol{\phi}(x,\boldsymbol{A}_i)$ and $\boldsymbol{\psi}_t = \boldsymbol{\psi}(x,\boldsymbol{B}_t)$ be transformations of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$. Let $\widehat{\boldsymbol{\phi}}_i$ and $\widehat{\boldsymbol{\psi}}_t$ be corresponding estimators. In the main text we presented the special case where $\widehat{\boldsymbol{\phi}}_i$ and $\widehat{\boldsymbol{\psi}}_t$ were equal to the factor loadings and factors obtained from $\widehat{\boldsymbol{\Gamma}}(x)$, but many other choices of $\widehat{\boldsymbol{\phi}}_i$ and $\widehat{\boldsymbol{\psi}}_t$ are conceivable. We again define

$$\mathbb{N}_i = \left\{j\in\mathbb{N}\setminus\{i\} : \left\|\widehat{\boldsymbol{\phi}}_i-\widehat{\boldsymbol{\phi}}_j\right\| \leq \tau_{NT}\right\}, \quad \mathbb{T}_t = \left\{s\in\mathbb{T}\setminus\{t\} : \left\|\widehat{\boldsymbol{\psi}}_t-\widehat{\boldsymbol{\psi}}_s\right\| \leq \upsilon_{NT}\right\},$$

for some bandwidth parameters $\tau_{NT} > 0$ and $\upsilon_{NT} > 0$. A debiased estimator of the reduced form parameter in (A.2) is given by

$$\widetilde{\nu}(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)\widetilde{Y}_{it}(x),$$

where $\widetilde{Y}_{it}(x)$ is defined as in (2.18). In the main text we only discussed the special case $W_{it}(x) = 1$. We can write $\widetilde{\nu}(x)$ as

$$\widetilde{\nu}(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \omega_{it} Y_{it},$$

where the weights ω_{it} are functions of $\widehat{\boldsymbol{\phi}}_j$ and $\widehat{\boldsymbol{\psi}}_s$ for all $j\in\mathbb{N}$ and $s\in\mathbb{T}$. Assumption 2.4.1 in the main text is generalized as follows.

Assumption A.1.1. There exists a sequence $\xi_{NT} > 0$ such that $\xi_{NT} \to 0$ as $N, T \to \infty$, and

(a) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)\,\mathbb{1}\{X_{it} \neq x \,\&\, n_{it} = 0\} = O_P(\xi_{NT})$.

(b) Y_{it} and $W_{it}(x)$ are uniformly bounded over i, t, N, T.

(c) Y_{it} is independent across both i and t, conditional on $\boldsymbol{X}^{NT}$, $\boldsymbol{A}^{N}$, $\boldsymbol{B}^{T}$.

(d) The function $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ is twice continuously differentiable with uniformly bounded second derivatives.

(e) There exists $c > 0$ such that $\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \leq c\,\|\boldsymbol{\phi}(\boldsymbol{a}_1) - \boldsymbol{\phi}(\boldsymbol{a}_2)\|$ for all $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{A}$, and $\|\boldsymbol{b}_1 - \boldsymbol{b}_2\| \leq c\,\|\boldsymbol{\psi}(\boldsymbol{b}_1) - \boldsymbol{\psi}(\boldsymbol{b}_2)\|$ for all $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{B}$.

(f) $\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i\right\|^2 + \max_{j\in\mathbb{N}_i}\left\|\widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\right\|^2\right) = O_P(\xi_{NT})$.
$\frac{1}{T}\sum_{t=1}^{T}\left(\left\|\widehat{\boldsymbol{\psi}}_t - \boldsymbol{\psi}_t\right\|^2 + \max_{s\in\mathbb{T}_t}\left\|\widehat{\boldsymbol{\psi}}_s - \boldsymbol{\psi}_s\right\|^2\right) = O_P(\xi_{NT})$.

(g) $\tau_{NT}^2 = O_P(\xi_{NT})$ and $\upsilon_{NT}^2 = O_P(\xi_{NT})$.

(h) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left[\omega_{it}^2 \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T}\right] = O_P(NT\,\xi_{NT}^2)$.

(i) Let $\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}$ be the outcome matrix $\boldsymbol{Y}^{NT}$, but with Y_{it} and Y_{js} replace by zero (or some other non-random number), and all other outcomes unchanged. We assume

$$\frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\{(i,t)\neq(j,s)\}\,\mathrm{E}\left[\left|\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) - \omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT})\right|\,\middle|\,\boldsymbol{X}^{NT}, \boldsymbol{A}^{N}, \boldsymbol{B}^{T}\right] = O_P\left(\xi_{NT}^2\right).$$

The generalized version of Theorem 2.4.2 is given in the following.

Theorem A.1.2. *Under Assumptions 2.2.1 and A.1.1,*

$$\widetilde{v}(x) - v_0(x) = O_P(\xi_{NT}).$$

Proof of Theorem A.1.2 (containing Theorem 2.4.2 as a special case)

Define $m_{it}(x) := m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$. We decompose

$$\widetilde{\nu}(x) - \nu_0(x) = e_0(x) + e_1(x) + e_2(x), \tag{A.5}$$

where

$$e_0(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(x) \, \mathbb{1}\{X_{it} \neq x \,\&\, n_{it} = 0\} \left[m_{it}(X_{it}) - m_{it}(x) \right],$$

and

$$\begin{aligned} e_1(x) &:= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{X_{it} \neq x \,\&\, n_{it} > 0\} \, W_{it}(x) \, e_{1,it}(x), \\ e_{1,it}(x) &:= \frac{\sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\} \left[m_{is}(x) + m_{jt}(x) - m_{js}(x) - m_{it}(x) \right]}{\sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\}}, \end{aligned}$$

and

$$e_2(x) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_{it} \, E_{it},$$

In the following we consider $e_0(x)$, $e_1(x)$, $e_2(x)$ separately.

Bound on $e_0(x)$: Assumption A.1.1(i) and (ii) guarantee that

$$\begin{aligned} |e_0(x)| &\leq \left(\max_{it} |m_{it}(X_{it}) - m_{it}(x)| \right) \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(x) \, \mathbb{1}\{X_{it} \neq x \,\&\, n_{it} = 0\} \\ &= O_P(\xi_{NT}). \end{aligned} \tag{A.6}$$

Bound on $e_1(x)$: Assumption A.1.1(iv) guarantees that there exists a constant $b > 0$ such that

$$\begin{aligned} \left| m(x, \boldsymbol{a}, \boldsymbol{b}) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t) - (\boldsymbol{a} - \boldsymbol{A}_i)' \frac{\partial m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)}{\partial \boldsymbol{A}_i} - (\boldsymbol{b} - \boldsymbol{B}_t)' \frac{\partial m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)}{\partial \boldsymbol{B}_t} \right| \\ \leq b \left(\|\boldsymbol{a} - \boldsymbol{A}_i\|^2 + \|\boldsymbol{b} - \boldsymbol{B}_t\|^2 \right). \end{aligned}$$

Using this we find that

$$m_{is}(x)+m_{jt}(x)-m_{js}(x)-m_{it}(x)\leq 2b\left(\left\|\boldsymbol{A}_i-\boldsymbol{A}_j\right\|^2+\|\boldsymbol{B}_t-\boldsymbol{B}_s\|^2\right),$$

and therefore

$$\begin{aligned}\left|e_{1,it}(x)\right| &\leq \frac{2b\sum_{j\in\mathbb{N}_i}\sum_{s\in\mathbb{T}_t}\mathbb{1}\{X_{is}=X_{jt}=X_{js}=x\}\left(\left\|\boldsymbol{A}_i-\boldsymbol{A}_j\right\|^2+\|\boldsymbol{B}_t-\boldsymbol{B}_s\|^2\right)}{\sum_{j\in\mathbb{N}_i}\sum_{s\in\mathbb{T}_t}\mathbb{1}\{X_{is}=X_{jt}=X_{js}=x\}}\\ &\leq 2b\left(\max_{j\in\mathbb{N}_i}\left\|\boldsymbol{A}_i-\boldsymbol{A}_j\right\|^2+\max_{s\in\mathbb{T}_t}\|\boldsymbol{B}_t-\boldsymbol{B}_s\|^2\right).\end{aligned}$$

We thus find

$$\begin{aligned}|e_1(x)| &\leq 2b\Big(\max_{ij}|W_{it}(x)|\Big)\left(\frac{1}{N}\sum_{i=1}^{N}\max_{j\in\mathbb{N}_i}\left\|\boldsymbol{A}_i-\boldsymbol{A}_j\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\max_{s\in\mathbb{T}_t}\|\boldsymbol{B}_t-\boldsymbol{B}_s\|^2\right)\\ &\leq 2bc\left(\max_{ij}|W_{it}(x)|\right)\left(\frac{1}{N}\sum_{i=1}^{N}\max_{j\in\mathbb{N}_i}\left\|\boldsymbol{\phi}(\boldsymbol{A}_i)-\boldsymbol{\phi}(\boldsymbol{A}_j)\right\|^2\right.\\ &\left.+\frac{1}{T}\sum_{t=1}^{T}\max_{s\in\mathbb{T}_t}\|\boldsymbol{\psi}(\boldsymbol{B}_t)-\boldsymbol{\psi}(\boldsymbol{B}_s)\|^2\right)\\ &= 2bc\left(\max_{ij}|W_{it}(x)|\right)\left(\frac{1}{N}\sum_{i=1}^{N}\max_{j\in\mathbb{N}_i}\left\|\boldsymbol{\phi}_i-\boldsymbol{\phi}_j\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\max_{s\in\mathbb{T}_t}\|\boldsymbol{\psi}_t-\boldsymbol{\psi}_s\|^2\right).\end{aligned}$$

Using the triangle inequality, the definition of $\mathbb{N}_i$, and the general inequality $(x_1+x_2+x_3)^2\leq 3(x_1^2+x_2^2+x_3^2)$, for $x_1,x_2,x_3\in\mathbb{R}$, we have

$$\begin{aligned}\max_{j\in\mathbb{N}_i}\left\|\boldsymbol{\phi}_i-\boldsymbol{\phi}_j\right\|^2 &\leq \max_{j\in\mathbb{N}_i}\left(\left\|\widehat{\boldsymbol{\phi}}_i-\widehat{\boldsymbol{\phi}}_j\right\|+\left\|\widehat{\boldsymbol{\phi}}_i-\boldsymbol{\phi}_i\right\|+\left\|\widehat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right\|\right)^2\\ &\leq \max_{j\in\mathbb{N}_i}\left(\tau_{NT}+\left\|\widehat{\boldsymbol{\phi}}_i-\boldsymbol{\phi}_i\right\|+\left\|\widehat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right\|\right)^2\\ &\leq 3\tau_{NT}^2+3\left\|\widehat{\boldsymbol{\phi}}_i-\boldsymbol{\phi}_i\right\|^2+3\max_{j\in\mathbb{N}_i}\left\|\widehat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right\|^2.\end{aligned}$$

Analogously we find

$$\max_{s\in\mathbb{T}_t}\|\boldsymbol{\psi}_t-\boldsymbol{\psi}_s\|^2\leq 3\upsilon_{NT}^2+3\left\|\widehat{\boldsymbol{\psi}}_t-\boldsymbol{\psi}_t\right\|^2+3\max_{s\in\mathbb{T}_t}\left\|\widehat{\boldsymbol{\psi}}_s-\boldsymbol{\psi}_s\right\|^2.$$

We thus obtain

$$|e_1(x)| \leq 6\,b\,c\left(\max_{ij}|W_{it}(x)|\right)\Bigg\{\tau_{NT}^2+\upsilon_{NT}^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\widehat{\boldsymbol{\phi}}_i-\boldsymbol{\phi}_i\right\|^2+\max_{j\in\mathbb{N}_i}\left\|\widehat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right\|^2\right) + \frac{1}{T}\sum_{t=1}^{T}\left(\left\|\widehat{\boldsymbol{\psi}}_t-\boldsymbol{\psi}_t\right\|^2+\max_{s\in\mathbb{T}_t}\left\|\widehat{\boldsymbol{\psi}}_s-\boldsymbol{\psi}_s\right\|^2\right)\Bigg\}$$
$$= O_P(\xi_{NT}). \tag{A.7}$$

Bound on $e_2(x)$: We have

$$[e_2(x)]^2 = \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT})\,E_{it}\,E_{js} = T_0+T_1+T_2,$$

where

$$T_0 := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{it}^2(\boldsymbol{Y}^{NT})\,E_{it}^2,$$
$$T_1 := \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\{(i,t)\neq(j,s)\} \times\left[\omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT})-\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\right]E_{it}\,E_{js},$$
$$T_2 := \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\{(i,t)\neq(j,s)\}\,\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)E_{it}\,E_{js}.$$

We have

$$\mathrm{E}\left[T_0\,\middle|\,\boldsymbol{X}^{NT},\boldsymbol{A}^N,\boldsymbol{B}^T\right] \leq \left(\max_{i,t}|E_{it}|\right)^2\frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left[\omega_{it}^2\left(\boldsymbol{Y}^{NT}\right)\,\middle|\,\boldsymbol{X}^{NT},\boldsymbol{A}^N,\boldsymbol{B}^T\right]$$
$$= O_P(\xi_{NT}^2),$$

and

$$\begin{aligned}
&\left|\mathrm{E}\left[T_1 \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right]\right| \\
&\leq \left(\max_{i,t} |E_{it}|\right)^2 \frac{1}{(NT)^2} \sum_{i,j=1}^{N} \sum_{t,s=1}^{T} \mathbb{1}\{(i,t) \neq (j,s)\} \\
&\times \mathrm{E}\left[\left|\omega_{it}(\boldsymbol{Y}^{NT})\, \omega_{js}(\boldsymbol{Y}^{NT}) - \omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\right| \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] \\
&= O_P(\xi_{NT}^2).
\end{aligned}$$

where we used that Y_{it} (and thus E_{it}) is uniformly bounded, together with Assumption A.1.1(viii) and (ix). Next, for $(i,t) \neq (j,s)$ we

$$\begin{aligned}
&\mathrm{E}\left[\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) E_{it} E_{js} \,\middle|\, \boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] \\
&= \omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \mathrm{E}\left[E_{it} E_{js} \,\middle|\, \boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] \\
&= \omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \\
&\qquad \mathrm{E}\left[E_{it} \,\middle|\, \boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] \mathrm{E}\left[E_{js} \,\middle|\, \boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] \\
&= 0,
\end{aligned}$$

where we used $\mathrm{E}\left[E_{it} \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] = 0$ together with the assumption that Y_{it} (and thus E_{it}) is independent across both i and t, conditional on $\boldsymbol{X}^{NT}$, $\boldsymbol{A}^N$, $\boldsymbol{B}^T$. By the law of iterated expectations the last display result also implies that for $(i,t) \neq (j,s)$ we have

$$\mathrm{E}\left[\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) \omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right) E_{it} E_{js} \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] = 0.$$

Using this we obtain that

$$\mathrm{E}\left[T_2 \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T\right] = 0.$$

Combining those results on T_0, T_1, T_2 we obtain

$$\mathrm{E}\left\{ [e_2(x)]^2 \,\middle|\, \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T \right\} = O_P(\xi_{NT}^2),$$

which implies $e_2 = O_P(\xi_{NT})$. Together with (A.5), (A.6), and (A.7) this gives the statement of the theorem.

□

Proof of Lemma A.1.1: Let $\boldsymbol{u}_j(x)$ be the N-vector with elements $u_j(x,\boldsymbol{A}_i)$, and let $\boldsymbol{v}_j(x)$ be the T-vector with elements $v_j(x,\boldsymbol{B}_t)$. Then we have $\boldsymbol{\Gamma}^\infty(x) = \sum_{j=1}^\infty s_j(x)\,\boldsymbol{u}_j(x)\,\boldsymbol{v}_j^{\mathrm{T}}(x)$, and therefore

$$\begin{aligned}
\|\boldsymbol{\Gamma}^\infty(x)\|_1 &\leq \sum_{j=1}^\infty s_j(x)\,\left\|\boldsymbol{u}_j(x)\right\|\,\left\|\boldsymbol{v}_j(x)\right\| \\
&= \sqrt{NT}\sum_{j=1}^\infty s_j(x)\sqrt{\frac{1}{N}\sum_{i=1}^N [u_j(x,\boldsymbol{A}_i)]^2}\sqrt{\frac{1}{T}\sum_{t=1}^T [v_j(x,\boldsymbol{B}_t)]^2} \\
&\leq \sqrt{NT}\sum_{j=1}^\infty s_j(x)\left(1+\frac{\frac{1}{N}\sum_{i=1}^N [u_j(x,\boldsymbol{A}_i)]^2-1}{2}\right)\left(1+\frac{\frac{1}{T}\sum_{t=1}^T [v_j(x,\boldsymbol{B}_t)]^2-1}{2}\right) \\
&= \sqrt{NT}\sum_{j=1}^\infty s_j(x) + \sqrt{NT}\,R_{NT} \\
&= \sqrt{NT}\,\|m(x,\cdot,\cdot)\|_* + \sqrt{NT}\,R_{NT},
\end{aligned}$$

where for the second inequality we used that $\sqrt{z} \leq 1+\frac{z-1}{2}$, for all $z \geq 0$, and we defined $R_{NT} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T r_{it}$, with

$$r_{it} = \sum_{j=1}^\infty s_j(x)\left\{\frac{[u_j(x,\boldsymbol{A}_i)]^2+[v_j(x,\boldsymbol{B}_t)]^2}{4} + \frac{[u_j(x,\boldsymbol{A}_i)]^2[v_j(x,\boldsymbol{B}_t)]^2}{4} - \frac{3}{4}\right\}.$$

Assumption 2.3.2 guarantees that $[u_j(x,\boldsymbol{A}_i)]^2$ and $[v_j(x,\boldsymbol{B}_t)]^2$ have mean equal to one, which implies that r_{it} has mean zero. Assumption 2.3.1 and the WLLN therefore guarantees that $R_{NT} = o_P(1)$. We have thus shown that $\|\boldsymbol{\Gamma}^\infty(x)\|_1 \leq \sqrt{NT}\,\|m(x,\cdot,\cdot)\|_* + o_P(\sqrt{NT})$, and since $\|m(x,\cdot,\cdot)\|_*$ is finite and non-random we also have $\|\boldsymbol{\Gamma}^\infty(x)\|_1 = O_P(\sqrt{NT})$.

□

Proof of Lemma A.1.2 The nuclear norm (or trace norm) can be defined by

$$\|\mathbf{\Gamma}\|_1 = \max_{\{\boldsymbol{M}\in\mathbb{R}^{N\times T}:\|\boldsymbol{M}\|_\infty\leq 1\}} \underbrace{\mathrm{Tr}\left(\boldsymbol{M}'\mathbf{\Gamma}\right)}_{=\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}\Gamma_{it}}. \tag{A.8}$$

Our assumption $\rho \geq \|\boldsymbol{E}(x)\|_\infty$ guarantees that a possible choice in this maximization is $\boldsymbol{M} = \rho^{-1}\boldsymbol{E}(x)$, and we therefore have

$$\rho\,\|\mathbf{\Gamma}\|_1 \geq \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,E_{it}(x)\,\Gamma_{it}.$$

Using this and the model $Y_{it} = \Gamma_{it}^{\infty}(x) + E_{it}(x)$, for $X_{it} = x$, we find that

$$\begin{aligned}
&Q_{NT}(\mathbf{\Gamma},\rho,x)\\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,(Y_{it}-\Gamma_{it})^2 + \rho\|\mathbf{\Gamma}\|_1\\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,(\Gamma_{it}^{\infty}(x)+E_{it}(x)-\Gamma_{it})^2 + \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,E_{it}(x)\,\Gamma_{it}\\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,(\Gamma_{it}^{\infty}(x)-\Gamma_{it})^2 + \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,\Gamma_{it}^{\infty}(x)E_{it}(x)\\
&+ \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,E_{it}^2(x).
\end{aligned}$$

By definition we have

$$Q_{NT}(\widehat{\mathbf{\Gamma}}(x),\rho,x) \leq Q_{NT}(\mathbf{\Gamma}^{\infty}(x),\rho,x) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,E_{it}^2(x) + \rho\|\mathbf{\Gamma}^{\infty}(x)\|_1$$

Combining the results in the last two displays gives the statement of the lemma.

□

Proof of Proposition A.1.1 In this proof we drop the argument x everywhere, and we define $\theta = NT\nu$ and $\theta_0 = NT\nu_0$. Define the NT-vectors $\boldsymbol{\gamma} = \mathrm{vec}(\mathbf{\Gamma})$, $\boldsymbol{\gamma}^{\infty} =$

$\text{vec}(\mathbf{\Gamma}^{\infty})$, $\boldsymbol{w} = \text{vec}(W_{it} : i \in \mathbb{N}, t \in \mathbb{T})$, $\boldsymbol{d} = \text{vec}(D_{it} : i \in \mathbb{N}, t \in \mathbb{T})$, and $\boldsymbol{p} = \text{vec}(P_{it} : i \in \mathbb{N}, t \in \mathbb{T})$. Then, $\text{diag}(\boldsymbol{p})$ is an $NT \times NT$ diagonal matrix. For $\rho > 0$ and $\theta \in \mathbb{R}$ we define

$$L_{NT}(\theta, \rho) = \min_{\{\mathbf{\Gamma} \in \mathbb{R}^{N \times T} : \theta = \boldsymbol{w}'\boldsymbol{\gamma}\}} Q_{NT}(\mathbf{\Gamma}, \rho),$$

which is the profile objective function that minimizes $Q_{NT}(\mathbf{\Gamma}, \rho)$ over almost all parameters $\mathbf{\Gamma}$, only keeping our parameter of interest fixed at $\theta = \boldsymbol{w}'\boldsymbol{\gamma} = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\Gamma_{it}$. Our goal is to show that the minimizing value

$$\widehat{\theta} := \underset{\theta \in \mathbb{R}}{\text{argmin}}\, L_{NT}(\theta, \rho) = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\widehat{\Gamma}_{it}$$

is close to $\theta := \boldsymbol{w}'\boldsymbol{\gamma}^{\infty} = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\Gamma_{it}^{\infty}$. Using the definition of $Q_{NT}(\mathbf{\Gamma}, \rho)$ and $Y_{it} = \Gamma_{it}^{\infty} + E_{it}$, for $D_{it} = 1$, we find that

$$L_{NT}(\theta, \rho) \leq Q_{NT}(\mathbf{\Gamma}^{\infty}, \rho) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\, E_{it}^2 + \rho\|\mathbf{\Gamma}^{\infty}\|_1. \tag{A.9}$$

If for a given value of $\theta = \boldsymbol{w}'\boldsymbol{\gamma}$ we have that the matrix $\boldsymbol{M}(\theta)$ with elements $M_{it}(\theta) := D_{it}\, E_{it} - \frac{\boldsymbol{w}'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} \frac{(D_{it} - P_{it})W_{it}}{P_{it}}$ satisfies $\|\boldsymbol{M}(\theta)\|_{\infty} \leq \rho$, then by the definition of $\|\cdot\|_1$ in (A.8) we have $\rho\|\mathbf{\Gamma}\|_1 \leq \text{Tr}(\mathbf{\Gamma}'\boldsymbol{M}(\theta)) = \sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}(\theta)\Gamma_{it}$. Using this and $Y_{it} = \Gamma_{it}^{\infty} + E_{it}$, for $D_{it} = 1$, we find that

$$\begin{aligned}
Q_{NT}(\mathbf{\Gamma}, \rho) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\,(Y_{it} - \Gamma_{it})^2 + \rho\|\mathbf{\Gamma}\|_1 \\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\,[(\Gamma_{it}^{\infty} - \Gamma_{it}) + E_{it}]^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{ D_{it}\, E_{it} - \frac{[(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})'\boldsymbol{w}]}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} \frac{(D_{it} - P_{it})W_{it}}{P_{it}} \right\} \Gamma_{it} \\
&= \underbrace{\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\,(\Gamma_{it} - \Gamma_{it}^{\infty})^2 - \frac{[(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})'\boldsymbol{w}]}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{(D_{it} - P_{it})W_{it}}{P_{it}}\,(\Gamma_{it} - \Gamma_{it}^{\infty})}_{=:Q_{NT}^{(\text{low},1)}(\mathbf{\Gamma})} \\
&\qquad + \underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}(\theta)\Gamma_{it}^{\infty} + \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\, E_{it}^2}_{=:Q_{NT}^{(\text{low},2)}},
\end{aligned}$$

where in the last step we added and subtracted $\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}(\theta)\Gamma_{it}^{\infty}$, and we mul-

tiplied out $[(\Gamma_{it}^{\infty} - \Gamma_{it}) + E_{it}]^2$, which leads to some simplifications. Notice that $D_{it} E_{it} = E_{it}$ by construction of E_{it}, so that some occurrences of D_{it} above could be dropped, but we find it clearer to keep track of D_{it} explicitly here.

Next, we define the $NT \times NT$ idempotent matrices $\mathbf{P} = \frac{\operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w} \boldsymbol{w}'}{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}}$ and $\mathbf{R} = \mathbf{I}_{NT} - \mathbf{P}$. We then have

$$
\begin{aligned}
& Q_{NT}^{(\text{low},1)}(\boldsymbol{\Gamma}) \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \operatorname{diag}(\boldsymbol{d}) (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) - \frac{[(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \boldsymbol{w}]}{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}} \left[\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \operatorname{diag}(\boldsymbol{d} - \boldsymbol{p}) (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) \right] \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \left(\mathbf{P}' + \mathbf{R}' \right) \operatorname{diag}(\boldsymbol{d}) (\mathbf{P} + \mathbf{R}) (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) \\
&- (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{P}' \operatorname{diag}(\boldsymbol{d} - \boldsymbol{p}) (\mathbf{P} + \mathbf{R}) (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{R}' \operatorname{diag}(\boldsymbol{d}) \mathbf{R} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{P}' \operatorname{diag}(2\boldsymbol{p} - \boldsymbol{d}) \mathbf{P} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}), \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{R}' \operatorname{diag}(\boldsymbol{d}) \mathbf{R} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) \\
&+ \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{P}' \operatorname{diag}(\boldsymbol{p} - \boldsymbol{d}) \mathbf{P} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) + \frac{1}{2} \frac{[(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \boldsymbol{w}]^2}{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}}
\end{aligned}
$$

where all the "mixed terms" (that involve both $\mathbf{P}$ and $\mathbf{R}$) cancel because we have $\mathbf{P}' \operatorname{diag}(\boldsymbol{p}) \mathbf{R} = 0$, and in the last step we used that $\mathbf{P}' \operatorname{diag}(\boldsymbol{p}) \mathbf{P} = \frac{\boldsymbol{w} \boldsymbol{w}'}{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}}$. We have

$$
\min_{\{\boldsymbol{\Gamma} \in \mathbb{R}^{N \times T} : \theta = \boldsymbol{w}' \boldsymbol{\gamma}\}} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty})' \mathbf{R}' \operatorname{diag}(\boldsymbol{d}) \mathbf{R} (\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\infty}) = 0,
$$

because $\boldsymbol{\gamma}^{*} = \mathbf{R} \boldsymbol{\gamma}^{\infty} + \theta \frac{\operatorname{diag}(\boldsymbol{p})^{-1} w}{w' \operatorname{diag}(\boldsymbol{p})^{-1} w}$ is a possible choice in the minimization problem, which satisfies $w' \boldsymbol{\gamma}^{*} = \theta$ and $\mathbf{R} (\boldsymbol{\gamma}^{*} - \boldsymbol{\gamma}^{\infty}) = 0$. We therefore have

$$
\begin{aligned}
& \min_{\{\boldsymbol{\Gamma} \in \mathbb{R}^{N \times T} : \theta = \boldsymbol{w}' \boldsymbol{\gamma}\}} Q_{NT}^{(\text{low},1)}(\boldsymbol{\Gamma}) \\
&= \frac{1}{2} (\theta - \theta_0)^2 \left(\frac{1}{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}} + \frac{\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \operatorname{diag}(\boldsymbol{p} - \boldsymbol{d}) \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w}}{(\boldsymbol{w}' \operatorname{diag}(\boldsymbol{p})^{-1} \boldsymbol{w})^2} \right) \\
&= \frac{1}{2} (\theta - \theta_0)^2 \left(\frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}^2 P_{it}^{-1}} + \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}^2 P_{it}^{-2} (P_{it} - D_{it})}{(\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}^2 P_{it}^{-1})^2} \right) \\
&= \frac{NT}{2} c_1 (\nu - \nu_0)^2,
\end{aligned}
$$

with c_1 as defined in the statement of the proposition, and $\nu - \nu_0 = (NT)^{-1}(\theta - \theta_0)$.

Thus, if $M_{it}(\theta) = D_{it}E_{it} - (\nu - \nu_0)V_{it}$ satisfies $\|\boldsymbol{M}(\theta)\|_\infty \leq \rho$, then we have

$$\begin{aligned} L_{NT}(\theta,\rho) &\geq \min_{\{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}:\,\theta=\boldsymbol{w}'\boldsymbol{\gamma}\}} Q_{NT}^{(\text{low},1)}(\boldsymbol{\Gamma}) + Q_{NT}^{(\text{low},2)} \\ &= \frac{NT}{2}c_1(\nu-\nu_0)^2 + \sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}(\theta)\Gamma_{it}^{\infty} + \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}E_{it}^2, \end{aligned}$$

and combing this with (A.9) gives

$$\begin{aligned} \frac{L_{NT}(\theta,\rho) - L_{NT}(\theta_0,\rho)}{NT} &\geq \frac{c_1}{2}(\nu-\nu_0)^2 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}(\theta)\Gamma_{it}^{\infty} - \frac{\rho}{NT}\|\boldsymbol{\Gamma}^{\infty}\|_1 \\ &= \frac{c_1}{2}(\nu-\nu_0)^2 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}E_{it}\Gamma_{it}^{\infty} \\ &- (\nu-\nu_0)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}\Gamma_{it}^{\infty} - \frac{\rho}{NT}\|\boldsymbol{\Gamma}^{\infty}\|_1. \end{aligned}$$

Using the assumption $c_1 > 0$ and definitions of c_2 and c_3 in the proposition this inequality can equivalently be written as

$$\begin{aligned} \frac{2\left[L_{NT}(NT\nu,\rho) - L_{NT}(NT\nu_0,\rho)\right]}{c_1 NT} &\geq (\nu-\nu_0)^2 - \frac{2c_2}{c_1}(\nu-\nu_0) + \left(\frac{c_2}{c_1}\right)^2 - c_3 \\ &= \left(\nu - \nu_0 - \frac{c_2}{c_1}\right)^2 - c_3. \end{aligned} \tag{A.10}$$

Notice that $c_3 > 0$ because our assumptions guarantee that $\|\boldsymbol{E}\|_\infty < \rho$ and therefore $\rho\|\boldsymbol{\Gamma}^{\infty}\|_1 \geq \sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}\Gamma_{it}^{\infty}$, according to (A.8).

The inequality in (A.10) was derived under the assumption that $\|\boldsymbol{M}(NT\nu)\|_\infty \leq \rho$. Define $\nu_+^*(\varepsilon) \in \mathbb{R}$ and $\nu_-^*(\varepsilon) \in \mathbb{R}$ by

$$\nu_\pm^*(\varepsilon) := \nu_0 \pm (c_4 + \varepsilon), \qquad \text{for} \quad 0 < \varepsilon \leq \frac{\rho - \|\boldsymbol{E}\|_\infty - c_4\|\boldsymbol{V}\|_\infty}{\|\boldsymbol{V}\|_\infty}.$$

Our assumption $\|\boldsymbol{E}\|_\infty + c_4\|\boldsymbol{V}\|_\infty < \rho$ guarantees that such an $\varepsilon > 0$ exists. Using the triangle inequality we find that

$$\|\boldsymbol{M}(NT\nu_\pm^*(\varepsilon))\|_\infty = \|\boldsymbol{E} - (\nu_\pm^*(\varepsilon) - \nu_0)\boldsymbol{V}\|_\infty \leq \|\boldsymbol{E}\|_\infty + |\nu_\pm^*(\varepsilon) - \nu_0|\|\boldsymbol{V}\|_\infty \leq \rho,$$

where the final inequality follows from the definition of $\nu_{\pm}^{*}(\varepsilon)$. The conditions for (A.10) is therefore satisfies by $\nu = \nu_{\pm}^{*}(\varepsilon)$, that is, we have

$$\begin{aligned}
\frac{2\left[L_{NT}(NT\nu_{\pm}^{*}(\varepsilon),\rho) - L_{NT}(NT\nu_0,\rho)\right]}{c_1 NT} &\geq \left(\nu_{\pm}^{*}(\varepsilon) - \nu_0 - \frac{c_2}{c_1}\right)^2 - c_3 \\
&= \left(c_4 + \varepsilon \mp \frac{c_2}{c_1}\right)^2 - c_3 \\
&= \left(\sqrt{c_3} + \varepsilon + \frac{|c_2| \mp c_2}{c_1}\right)^2 - c_3 \\
&\geq \left(\sqrt{c_3} + \varepsilon\right)^2 - c_3 \\
&> 0.
\end{aligned}$$

where we used the definition $c_4 = \sqrt{c_3} + \frac{|c_2|}{c_1}$.

$L_{NT}(NT\nu,\rho)$ is a convex function of $\nu = \theta/NT$, because it was obtained via profiling of the convex function $Q_{NT}(\mathbf{\Gamma},\rho)$. The value ν_0 lies in the interval $[\nu_{+}^{*}(\varepsilon), \nu_{-}^{*}(\varepsilon)]$, and we have shown that $L_{NT}(NT\nu_0,\rho) < L_{NT}(NT\nu_{\pm}^{*}(\varepsilon),\rho)$. It must therefore be the case that the optimal $\widehat{\nu} = NT\widehat{\theta}$ that minimizes $L_{NT}(NT\nu,\rho)$ also lies in the interval $[\nu_{+}^{*}(\varepsilon), \nu_{-}^{*}(\varepsilon)]$ — otherwise we obtain a contradiction to the convexity of $L_{NT}(NT\nu,\rho)$. Thus, we have shown that

$$|\widehat{\nu} - \nu_0| \leq c_4 + \varepsilon,$$

and because we can choose $\varepsilon > 0$ arbitrarily small it must be the case that

$$|\widehat{\nu} - \nu_0| \leq c_4,$$

which is what we wanted to show.

□

Appendix B

Appendix – Chapter 3

B.1 Simulations with lagged dependent variable

In Table B.1 we display the simulation results for the following DGP,

$$\begin{aligned} Y_{it} &= Y_{i,t-1}\rho + X_{it}\beta + h(\alpha_i, \gamma_t) + \varepsilon_{it}, \\ X_{it} &= g(\alpha_i, \gamma_t) + \mu_{it}, \end{aligned} \tag{A.1}$$

where all parameters are set to the same values as Section 3.6, along with $\rho = 0.5$. Note that even though γ_t is simulated to be independent across t, there is no direct omitted variable bias from simply ignoring $Y_{i,t-1}$ in the regression. However, and as we see in Table B.1, omitting $Y_{i,t-1}$ makes factor estimation more difficult because of the additional $Y_{i,t-1}\rho$ term in the fitted residual. To see this take the fitted residual with and without lagged Y projected out,

$$\begin{aligned} \widehat{W}_1 &= (Y - X\widehat{\beta}) = X(\beta - \widehat{\beta}) + Y_{-1}\rho + h(\alpha, \gamma) + \varepsilon \\ \widehat{W}_2 &= (Y - X\widehat{\beta} - Y_{-1}\widehat{\rho}) = X(\beta - \widehat{\beta}) + Y_{-1}(\rho - \widehat{\rho}) + h(\alpha, \gamma) + \varepsilon, \end{aligned}$$

where Y_{-1} is simply the matrix of lagged Y. We see then when lagged Y, or any control variable for that matter, is not projected out, then it makes identifying factors related to $h(\alpha, \gamma)$ more difficult due to $Y_{-1}\rho$ in the residual. Hence, whilst the presence of the lagged dependent term in the residual may not be directly problematic, it obfuscates estimation of the factors. This is especially highlighted by the fact that increased number of factors do not necessarily improve bias.

Table B.1: Lagged dependent variable simulation

	Without lagged Y	With lagged Y
Mean bias (Standard deviation)		
OLS	0.5232 (0.0347)	0.5626 (0.0101)
Fixed-effects	0.4974 (0.0319)	0.5110 (0.0095)
LS (10 factors)	-0.0601 (0.0133)	0.0191 (0.0123)
LS (20 factors)	-0.1445 (0.0134)	0.0156 (0.0150)
LS (40 factors)	-0.2281 (0.0174)	-0.0379 (0.0838)
LS jackknife (10 factors)	-0.1970 (0.0249)	-0.0378 (0.0228)
LS jackknife (20 factors)	-0.2697 (0.0258)	-0.0088 (0.0290)
LS jackknife (40 factors)	-0.3302 (0.0406)	-0.0614 (0.2443)
GFE	-0.0358 (0.0249)	0.0179 (0.0193)
GFE jackknife	-0.0144 (0.0445)	0.0153 (0.0341)

10,000 Monte Carlo rounds.
All results refer to estimation of β. Mean bias is simply the mean of the bias across simulations. Standard deviation is the standard deviation of the estimates, again across simulations.

B.2 Proofs

B.2.1 Proofs for Section 3.3

We first establish a technical lemma, which is afterwards used to prove the main text theorem. Remember that we write $\|\cdot\|$ for the spectral norm of a matrix. Define the projection matrix $P_A = A(A'A)^{\dagger}A'$ for any matrix A and remember we write the annihilation matrix $M_A = \mathbb{I} - P_A$. Here, $\dagger$ refers to the Moore-Penrose inverse.

Lemma A.1. *Let Assumption 3.3.3 hold and consider $N,T \to \infty$. Furthermore, assume that*

$$Y = \sum_{k=1}^{K} X_k \beta_k^0 + e^* + e, \tag{A.2}$$

with $rank(e^*) = R_{NT} \leq \min(N,T)/2$, $\|e\| = \mathcal{O}_P(\eta_{NT})$, $\|X_k\| = \mathcal{O}_P(\sqrt{NT})$, and $\frac{1}{\sqrt{NT}} Tr(X_k e') = \mathcal{O}_P(\xi_{NT})$, for $k = 1,\ldots,K$. Then, the LS estimator in (3.5) calculated with $R = R_{NT}$ factors in the estimation procedure, satisfies $\widehat{\beta}_{LS} - \beta^0 = \mathcal{O}_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$.

Proof of Lemma A.1. This proof is relatively minor modification of the consistency proof for the LS estimator in Moon and Weidner (2015), and more technical

details can be found there. For simplicity we just write R, η, ξ instead of R_{NT}, η_{NT}, ξ_{NT} in this proof. We rewrite the definition of $\widehat{\beta}_{\text{LS}}$ as

$$\begin{aligned}\widehat{\beta}_{\text{LS}} &= \underset{\beta}{\operatorname{argmin}}\, \mathscr{L}_{NT}(\beta), \\ \mathscr{L}_{NT}(\beta) &:= \min_{\{\lambda\in\mathbb{R}^{N\times R},\, f\in\mathbb{R}^{T\times R}\}} \frac{1}{NT}\text{Tr}\left[\left(Y - X\cdot\beta - \lambda f'\right)\left(Y - X\cdot\beta - \lambda f'\right)'\right].\end{aligned} \tag{A.3}$$

Since $\text{rank}(e^*) = R$ we can write $e^* = \lambda^* f^{*\prime}$ for some $N\times R$ matrix λ^* and $T\times R$ matrix $f^{*\prime}$.

We now first establish a lower bound on $\mathscr{L}_{NT}(\beta)$. Let $\Delta\beta = \beta - \beta^0$. Consider the definition of $\mathscr{L}_{NT}(\beta)$ in equation (A.3) and plug in the model $Y = \beta\cdot X + \lambda^* f^{*\prime} + e$. We then have

$$\begin{aligned}&\mathscr{L}_{NT}(\beta) = \\ &\min_{\{\lambda\in\mathbb{R}^{N\times R},\, f\in\mathbb{R}^{T\times R}\}} \frac{1}{NT}\text{Tr}\left[\left(\Delta\beta\cdot X + e + \lambda^* f^{*\prime} - \lambda f'\right)\left(\Delta\beta\cdot X + e + \lambda^* f^{*\prime} - \lambda f'\right)'\right] \\ &\geq \min_{\left\{\tilde{\lambda}\in\mathbb{R}^{N\times(2R)},\, \tilde{f}\in\mathbb{R}^{T\times(2R)}\right\}} \frac{1}{NT}\text{Tr}\left[\left(\Delta\beta\cdot X + e - \tilde{\lambda}\tilde{f}'\right)\left(\Delta\beta\cdot X + e - \tilde{\lambda}\tilde{f}'\right)'\right] \\ &= \frac{1}{NT}\min_{\tilde{f}\in\mathbb{R}^{T\times(2R)}} \text{Tr}\left[\left(\Delta\beta\cdot X + e\right) M_{\tilde{f}}\left(\Delta\beta\cdot X + e\right)'\right] \\ &= \frac{1}{NT}\min_{\tilde{f}\in\mathbb{R}^{T\times(2R)}} \Big\{\text{Tr}\left[\left(\Delta\beta\cdot X\right) M_{\tilde{f}}\left(\Delta\beta\cdot X\right)'\right] + \text{Tr}\left(ee'\right) - \text{Tr}\left(eP_{\tilde{f}}e'\right) \\ &\qquad\qquad + 2\text{Tr}\left[\left(\Delta\beta\cdot X\right)e'\right] - 2\text{Tr}\left[\left(\Delta\beta\cdot X\right)P_{\tilde{f}}e'\right]\Big\} \\ &\geq \frac{1}{NT}\Bigg\{\sum_{r=2R+1}^{T}\mu_r\left[(\Delta\beta\cdot X)'(\Delta\beta\cdot X)\right] + \text{Tr}\left(ee'\right) - 2R\|e\|^2 \\ &\qquad\qquad + 2\text{Tr}\left[\left(\Delta\beta\cdot X\right)e'\right] - 4R\|e\|\|\Delta\beta\cdot X\|\Bigg\} \\ &\geq b\,\|\Delta\beta\|^2 + \frac{1}{NT}\text{Tr}\left(ee'\right) + \mathscr{O}_P\left(\frac{R\,\eta^2}{NT}\right) + \mathscr{O}_P\left(\frac{(\xi + R\,\eta)\,\|\Delta\beta\|}{\sqrt{NT}}\right).\end{aligned} \tag{A.4}$$

Here, we applied the inequality $|\text{Tr}(A)| \leq \text{rank}(A)\|A\|$ with $A = (\Delta\beta\cdot X)P_{\tilde{f}}e'$ and also with $A = eP_{\tilde{f}}e'$. We also used that $\min_{\tilde{f}}\text{Tr}\left[\left(\Delta\beta\cdot X\right)M_{\tilde{f}}\left(\Delta\beta\cdot X\right)'\right] =$

$\sum_{r=2R+1}^{T} \mu_r \left[(\Delta\beta \cdot X)'(\Delta\beta \cdot X)\right]$. In the last step of (A.4) we applied the various assumptions in the lemma.

Next, we establish an upper bound on $\mathscr{L}_{NT}(\beta^0)$. We can choose $\lambda = \lambda^*$ and $f = f^*$ in the minimization problem in (A.3), and therefore

$$\mathscr{L}_{NT}(\beta^0) \leq \frac{1}{NT} \operatorname{Tr}\left(ee'\right). \tag{A.5}$$

Since we could choose $\beta = \beta^0$ in the minimization of β, the optimal $\widehat{\beta}_{\mathrm{LS}}$ needs to satisfy $\mathscr{L}_{NT}(\widehat{\beta}_{\mathrm{LS}}) \leq \mathscr{L}_{NT}(\beta^0)$. Together with (A.4) and (A.5) this gives

$$b\,\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\|^2 \leq \mathscr{O}_P\left(\frac{(\xi + R\eta)\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\|}{\sqrt{NT}}\right) + \mathscr{O}_P\left(\frac{R\eta^2}{NT}\right) \tag{A.6}$$

Since $R \to \infty$ as $N, T \to \infty$, we have

$$\mathscr{O}_P\left(\frac{R\eta^2}{NT}\right) \leq \mathscr{O}_P\left(\left(\frac{R\eta}{\sqrt{NT}}\right)^2\right) \leq \mathscr{O}_P\left(\left(\frac{\xi + R\eta}{\sqrt{NT}}\right)^2\right),$$

and (A.6) thus implies

$$\begin{aligned}\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\|^2 &\leq \mathscr{O}_P\left(\frac{(\xi + R\eta)\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\|}{b\sqrt{NT}}\right) + \mathscr{O}_P\left(\frac{1}{b}\left(\frac{\xi + R\eta}{\sqrt{NT}}\right)^2\right) \\ &=: 2B_1\,\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\| + (B_2)^2,\end{aligned}$$

with random variables $B_1 = \mathscr{O}_P\left(\frac{\xi + R\eta}{\sqrt{NT}}\right)$ and $B_2 = \mathscr{O}_P\left(\frac{\xi + R\eta}{\sqrt{NT}}\right)$, and where we used that b is a positive constant. Completing the square gives

$$\left(\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\| - B_1\right)^2 \leq (B_2)^2 + (B_1)^2,$$

by taking the square root we thus obtain

$$\|\widehat{\beta}_{\mathrm{LS}} - \beta^0\| \leq B_1 + \sqrt{(B_2)^2 + (B_1)^2}.$$

Since B_1 and B_2 are both of order $\mathcal{O}_P\left(\frac{\xi+R\eta}{\sqrt{NT}}\right)$ it thus follows that

$$\|\widehat{\beta}_{\mathrm{LS}}-\beta^0\| = \mathcal{O}_P\left((\xi+R\eta)/\sqrt{NT}\right),$$

which is what we wanted to show. ■

Using Lemma A.1 we are now ready to prove Theorem 3.3.5.

Proof of Theorem 3.3.5. To apply Lemma A.1 we first need to define e and e^* such that (A.2) is an implication of our model (3.13). Decompose $\Gamma = \sum_{r=1}^{\min\{N,T\}} \lambda_r^* f_r^{*\prime}$, which is a reformulation of the singular value decomposition of a matrix. Define $e^* = \sum_{r=1}^{R_{NT}} \lambda_r^* f_r^{*\prime}$ such that $\mathrm{rank}(e^*) = R_{NT}$. Also define $e = S + \varepsilon$ where $S = \Gamma - \sum_{r=1}^{R_{NT}} \lambda_r^* f_r^{*\prime}$. With these definitions model (3.13) can be rewritten as (A.2) and it remains to show Assumptions 3.3.1-3.3.3 are sufficient for Lemma A.1 and to characterise the sequences η_{NT} and ξ_{NT}.

First, use the norm inequality $\|S+\varepsilon\| \leq \|S\| + \|\varepsilon\|$ with $\|\varepsilon\| = O_P(\sqrt{\max\{N,T\}})$ from Assumption 3.3.1 (ii) to show $\|e\| \leq \|S\| + O_P(\sqrt{\max\{N,T\}})$. To bound $\|S\|$ use the fact that the spectral norm is bounded by the Frobenius norm and Assumption 3.3.4 to show

$$\begin{aligned}\|S\|^2 \leq \|S\|_F^2 &= \sum_{r=R_{NT}+1}^{\infty} \sigma_r^2(\Gamma) \\ &\leq O_P\big(NTR_{NT}^{1-2\rho}\big).\end{aligned}$$

This shows that $\|e\|$ is asymptotically bounded in probability by the sequence η_{NT} with

$$\eta_{NT} = \sqrt{\max\{N,T\}} + \sqrt{NT}R_{NT}^{(1-2\rho)/2}.$$

That is, $\|e\| = O_P(\eta_{NT})$.

Secondly, the bound on $\|X_k\|$ is direct from Assumption 3.3.1.(i) again because the spectral norm is bounded by the Frobenius norm. That is, $\|X_k\|^2 \leq \|X_k\|_F^2 = \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}^2 = O_P(NT)$.

Lastly, we need to show that $\frac{1}{\sqrt{NT}}\text{Tr}(X_k e') = O_P(\xi_{NT})$ and to find ξ_{NT}. To do this we decompose e and use the Cauchy-Schwarz inequality, the triangle inequality and linearity of the trace operator in the following,

$$\begin{aligned}\left|\frac{1}{\sqrt{NT}}\text{Tr}(X_k e')\right| &= \left|\frac{1}{\sqrt{NT}}\text{Tr}(X_k(S+\varepsilon)')\right| \\ &\leq \frac{1}{\sqrt{NT}}\|X_k\|_F \|S\|_F + \frac{1}{\sqrt{NT}}|\text{Tr}(X_k\varepsilon')| \\ &= O_P(1)\|S\|_F + O_P(1).\end{aligned} \tag{A.7}$$

The third line follows from Assumption 3.3.1.(i) and Assumption 3.3.2. From above we know $\|S\|_F = O_P\big(\sqrt{NT}R_{NT}^{(1-2\rho)/2}\big)$, hence we have found $\xi_{NT} = \sqrt{NT}R_{NT}^{(1-2\rho)/2} + 1$.

Thus, we have shown that all conditions for Lemma A.1 are satisfied and found the rates η_{NT} and ξ_{NT}. This shows that LS estimation in (3.5) on the model (3.13) with $R = R_{NT}$ factors satisfies $\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$, with

$$\begin{aligned}O_P\left(\frac{(\xi_{NT}+R_{NT}\eta_{NT})}{\sqrt{NT}}\right) &= O_P\Big(R_{NT}^{(1-2\rho)/2}\Big) + O_P\Big(\frac{1}{\sqrt{NT}}\Big) + O_P\Big(R_{NT}^{(3-2\rho)/2}\Big) \\ &\quad + O_P\left(R_{NT}\sqrt{\frac{\max\{N,T\}}{NT}}\right) \\ &= O_P\Big(R_{NT}^{(3-2\rho)/2}\Big) + O_P\Big(R_{NT}\min\{N,T\}^{-1/2}\Big).\end{aligned}$$

■

Proof of Remark 1. Note that if we weaken the singular value decay to that supposed in Remark 1, i.e. $\sigma_r(\Gamma) = c\sqrt{NT}r^{-\rho}$, and otherwise maintain Assumptions 3.3.1-3.3.3 we can further bound the bias in LS estimation found in Theo-

rem 3.3.5 as follows. For $\|S\|_F$, note,

$$\begin{aligned}
\|S\|_F^2 &= \sum_{r=R_{NT}+1}^{\infty} \sigma_r^2(\Gamma) \\
&\leq \sum_{r=R_{NT}+1}^{\infty} cNTr^{-2\rho} \qquad wpa.1 \qquad \text{(Assumption 3.3.4)} \\
&\leq cNT \int_{R_{NT}}^{\infty} r^{-2\rho} dr \qquad wpa.1 \qquad \text{(integral bound)} \\
&= \frac{c}{2\rho-1} NTR_{NT}^{1-2\rho} \qquad wpa.1
\end{aligned}$$

In the third line we use an integral bound and the fourth line simply evaluates this integral. From line two all arguments are $wpa.1$, hence $\|S\|_F = \mathcal{O}_P(\sqrt{NT}R_{NT}^{(1-2\rho)/2})$, where $(c/2\rho-1)$ is the bounding constant. We can then directly bound

$$\begin{aligned}
\|S\| &= \max_{r\in\{R_{NT}+1,\ldots,\min\{N,T\}\}} \sigma_r(\Gamma) \\
&= \mathcal{O}_P\big(\sqrt{NT}(R_{NT}+1)^{-\rho}\big),
\end{aligned}$$

where we use the convention that singular values are indexed in descending order. We then simplify the last bound to $\|S\| = \mathcal{O}_P\big(\sqrt{NT}R_{NT}^{-\rho}\big)$, replacing $R_{NT}+1$ with R_{NT} as $R_{NT} \to \infty$. We can then rely on the same working in the proof of Theorem 3.3.5 to show that the conditions in Lemma A.1 are satisfied with $\xi_{NT} = \sqrt{NT}R_{NT}^{(1-2\rho)/2} + 1$ and $\eta_{NT} = \sqrt{\max\{N,T\}} + \sqrt{NT}R_{NT}^{-\rho}$, where the second term in η_{NT} is slightly different to Theorem 3.3.5. Hence, $\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$, with

$$\begin{aligned}
O_P\left(\frac{(\xi_{NT}+R_{NT}\eta_{NT})}{\sqrt{NT}}\right) &= O_P\Big(R_{NT}^{(1-2\rho)/2}\Big) + O_P\Big(\frac{1}{\sqrt{NT}}\Big) + O_P\Big(R_{NT}^{1-\rho}\Big) \\
&\quad + O_P\left(R_{NT}\sqrt{\frac{\max\{N,T\}}{NT}}\right) \\
&= O_P\Big(R_{NT}^{1-\rho}\Big) + O_P\Big(R_{NT}\min\{N,T\}^{-1/2}\Big).
\end{aligned}$$

■

To prove Lemma 3.3.1 we rely on the following result from (Griebel and Harbrecht, 2014), which we state without proof.

Let $H^p(\Omega_\alpha \times \Omega_\gamma)$ denote the Sobolev space $W^{p,k}$ on the product domain $(\Omega_\alpha \times \Omega_\gamma)$ for $k = 2$, which is in turn a Hilbert space. In the one dimensional case, this space admits functions in $L^2(\mathbb{R})$-space whose derivatives up to order p are also in $L^2(\mathbb{R})$-space. In multiple dimensions this definition extends as follows. Let $\nabla := \{\nabla^\alpha, \nabla^\gamma\}$ be a multi-index that captures all the dimensions of α and γ respectively. Define the mixed partial derivative as,

$$f^{(\nabla)} = \frac{\partial^{|\nabla|} f}{\partial a_1^{\nabla_1^\alpha} \dots \partial a_{d_\alpha}^{\nabla_{d_\alpha}^\alpha} \partial c_1^{\nabla_1^\gamma} \dots \partial c_{d_\gamma}^{\nabla_{d_\gamma}^\gamma}},$$

where $a \in \Omega_\alpha$ and $c \in \Omega_\gamma$ with $(\Omega_\alpha \times \Omega_\gamma)$ the domain of f. Then $|\nabla| = |\nabla_\alpha| + |\nabla_\gamma|$ and the bivariate function h is said to be in Hilbert space of order p if the mixed partial derivative exists (weakly) and

$$\left\| h^{(\nabla)} \right\|_{L_2} \leq \infty \text{ for all } |\nabla| \leq p.$$

This space of functions places a bound on the function itself as well as its derivative, which is why we refer to it as a smoothness condition.

Lemma A.2 (Theorem 3.5 in Griebel and Harbrecht 2014)**.** *Let $h \in H^p(\Omega_\alpha \times \Omega_\gamma)$ and $p > \min\{n_\alpha, n_\gamma\}/2$, then*

$$\left\| h - \sum_{l=1}^{R} \sigma_l (\varphi_l \otimes \psi_l) \right\|_{L^2(\Omega_1 \times \Omega_2)} = O\left(R^{\frac{1}{2} - \frac{p}{\min\{n_\alpha, n_\gamma\}}} \right). \tag{A.8}$$

In the following proof we use the Frobenius norm, which as a reminder is defined as $\|A\|_F^2 = \sum_{i=1}^N \sum_{t=1}^T |A_{it}|^2$ for any $N \times T$ matrix A.

Proof of Lemma 3.3.1. From Lemma A.2 we have,

$$\begin{aligned}
&\mathbb{E}\Big[\Big(h(\alpha_i,\gamma_t)-\sum_{s=1}^{R}\sigma_r\varphi_r(\alpha_i)\psi_r(\gamma_t)\Big)^2\Big] \\
&\qquad=\int_{\Omega_\alpha}\int_{\Omega_\gamma}\big(h(a,c)-\sum_{s=1}^{R}\sigma_r\varphi_r(a)\psi_r(c)\big)^2 f_{\alpha_i,\gamma_t}(a,c)dadc \\
&\qquad\leq\int_{\Omega_\alpha}\int_{\Omega_\gamma}\big(h(a,c)-\sum_{s=1}^{R}\sigma_r\varphi_r(a)\psi_r(c)\big)^2 dadc\sup_{a,c} f_{\alpha_i,\gamma_t}(a,c) \qquad \text{(A.9)} \\
&\qquad=\left\|h-\sum_{l=1}^{R}\sigma_l\varphi_l\otimes\psi_l\right\|^2_{L^2(\Omega_\alpha\times\Omega_\gamma)} O(1) \\
&\qquad=O\left(R^{1-\frac{2p}{\min\{n_\alpha,n_\gamma\}}}\right),
\end{aligned}$$

where in the second line we use a supremum bound on the probabilities, in the third line we use the definition of the $L^2(\Omega_\alpha\times\Omega_\gamma)$-norm and in the final line we use Lemma A.2. This shows that, in expectations, the entry-wise functional representation decays at polynomial rate $r^{1-2\rho}$, with $\rho = p/\min\{n_\alpha,n_\gamma\}$.

Using the Markov inequality gives

$$\Big(\Gamma_{it}-\sum_{\ell=1}^{r}\sigma_\ell\varphi_\ell(\alpha_i)\psi_\ell(\gamma_t)'\Big)^2=\mathscr{O}_P\left(r^{1-\frac{2p}{\min\{n_\alpha,n_\gamma\}}}\right),$$

which we use to bound singular values of the matrix Γ as follows.

We know

$$\Gamma_{it}=h(\alpha_i,\gamma_t)=\sum_{r=1}^{\infty}\sigma_r\varphi_r(\alpha_i)\psi_r(\gamma_t)=\sum_{r=1}^{\infty}\sigma_r w_{ir}v_{tr}$$

and in matrix form,

$$\Gamma=h(\alpha,\gamma)=\sum_{r=1}^{\infty}\sigma_r\varphi_r(\alpha)\psi_r(\gamma)'=\sum_{r=1}^{\infty}\sigma_r w_r v_r'.$$

Hence, we have

$$\begin{aligned}
\sum_{\ell=r+1}^{\min\{N,T\}} \sigma_\ell^2(\Gamma) &= \min_{\lambda\in\mathbb{R}^{N\times r}} \min_{f\in\mathbb{R}^{T\times r}} \left\|\Gamma - \lambda\, f'\right\|_F^2 \\
&\leq \left\|\Gamma - \sum_{\ell=1}^{r} \sigma_\ell \varphi_\ell(\alpha)\psi_\ell(\gamma)'\right\|_F^2 \\
&= \sum_i \sum_t \left(\sum_{\ell=r+1}^{\infty} \sigma_\ell \varphi_\ell(\alpha_i)\psi_\ell(\gamma_t)\right)^2 \\
&= \sum_i \sum_t \mathscr{O}_P\left(r^{1-\frac{2p}{\min\{n\alpha,n\gamma\}}}\right) \\
&= NT\,\mathscr{O}_P\left(r^{1-\frac{2p}{\min\{n\alpha,n\gamma\}}}\right).
\end{aligned}$$

Hence, we have $\frac{1}{NT}\sum_{\ell=r+1}^{\min\{N,T\}} \sigma_\ell^2(\Gamma) = \mathscr{O}_P\left(r^{1-2\rho}\right)$ with $\rho = p/\min\left\{n_\alpha, n_\gamma\right\}$, and Assumption 3.3.4 is satisfied. ■

B.2.2 Proofs for Section 3.4

Proof of Lemma 3.4.1. From Section 3.4 we have

$$\kappa_{NT} := \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{\Gamma}_{it},$$

with $\widetilde{\Gamma}$ defined analogously to $\widetilde{X}_k$ and $\widetilde{Y}$.

Take

$$\|\kappa_{NT}\| := \left\|\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{\Gamma}_{it}\right\|.$$

Using the inequality $\|Az\| \leq \|A\|\,\|z\|$ for general matrices A and vectors z we find

$$\|\kappa_{NT}\| \leq \left\|\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1}\right\| \left\|\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{\Gamma}_{it}\right\|.$$

Use $\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it,k}\widetilde{\Gamma}_{it}\right| \leq \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\widetilde{X}_{it,k}\widetilde{\Gamma}_{it}\right|$ and Hölder's inequality such that

$$\begin{bmatrix}\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it,1}\widetilde{\Gamma}_{it}\right| \\ \vdots \\ \left|\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it,K}\widetilde{\Gamma}_{it}\right|\end{bmatrix} \leq \begin{bmatrix}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\widetilde{X}_{it,1}\widetilde{\Gamma}_{it}\right| \\ \vdots \\ \sum_{i=1}^{N}\sum_{t=1}^{T}\left|\widetilde{X}_{it,K}\widetilde{\Gamma}_{it}\right|\end{bmatrix} \leq \begin{bmatrix}\|\mathrm{vec}(X_1)\|_\infty \\ \vdots \\ \|\mathrm{vec}(X_K)\|_\infty\end{bmatrix} \left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1,$$

where $\mathrm{vec}(A)$ vectorises a matrix A such that $\|\mathrm{vec}(A)\|_\infty = \max_{i,t}|A_{it}|$ yields the maximum norm and $\|\mathrm{vec}(A)\|_1 = \sum_{i=1}^{N}\sum_{t=1}^{T}|A_{it}|$ yields the entry-wise 1-norm of such a matrix.

Take the $\|\cdot\|$ to show

$$\begin{aligned}\left\|\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{\Gamma}_{it}\right\| &= \left(\sum_k\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it,k}\widetilde{\Gamma}_{it}\right|^2\right)^{1/2} \\ &\leq \left(\sum_k\left(\left\|\mathrm{vec}(\widetilde{X}_k)\right\|_\infty\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1\right)^2\right)^{1/2} \\ &= \left(\sum_k\left(\left\|\mathrm{vec}(\widetilde{X}_k)\right\|_\infty\right)^2\right)^{1/2}\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1 \\ &\leq \left(\sum_k\left\|\mathrm{vec}(\widetilde{X}_k)\right\|_\infty\right)\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1,\end{aligned}$$

where in the last line we use that $\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1$ is a scalar and that $\left\|\mathrm{vec}(X_k)\right\|_\infty > 0\ \forall k$. Thus we can bound the norm of κ_{NT} by

$$\|\kappa_{NT}\| \leq \left\|\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1}\right\|\left(\sum_{k=1}^{K}\left\|\mathrm{vec}(\widetilde{X}_k)\right\|_\infty\right)\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1.$$

Concentrate on $\left\|\mathrm{vec}(\widetilde{\Gamma})\right\|_1$. Let n_i^N be the size of each i's cluster and n_t^T be the size of each t's cluster, then

$$\widetilde{\Gamma}_{it} = h(\alpha_i,\gamma_t) - \frac{1}{n_i^N}\sum_{j\in g_i}h(\alpha_j,\gamma_t) - \frac{1}{n_t^T}\sum_{s\in c_t}h(\alpha_i,\gamma_s) + \frac{1}{n_i^N}\frac{1}{n_t^T}\sum_{j\in g_i}\sum_{s\in c_t}h(\alpha_j,\gamma_s).$$

Take the following Taylor expansions,

$$h(\alpha_j,\gamma_s) = h(\alpha_i,\gamma_t) + \frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + \frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r(i,j,t,s)$$
$$h(\alpha_j,\gamma_t) = h(\alpha_i,\gamma_t) + \frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha_i'}(\alpha_j-\alpha_i) + r'(i,j,t)$$
$$h(\alpha_i,\gamma_s) = h(\alpha_i,\gamma_t) + \frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r''(t,s,i),$$

where r, r' and r'' are remainder terms from the Taylor expansion.

From these expansions we have

$$\frac{1}{n_i^N}\sum_{j\in g_i} h(\alpha_j,\gamma_t) = h(\alpha_i,\gamma_t) + \frac{1}{n_i^N}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + r'(i,j,t)\right),$$

$$\frac{1}{n_t^T}\sum_{s\in c_t} h(\alpha_i,\gamma_s) = h(\alpha_i,\gamma_t) + \frac{1}{n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r''(t,s,i)\right),$$

and

$$\begin{aligned}
\frac{1}{n_i^N}\frac{1}{n_t^T}\sum_{j\in g_i}\sum_{s\in c_t} h(\alpha_j,\gamma_s) &= \frac{1}{n_i^N n_t^T}h(\alpha_i,\gamma_t) \\
&+ \frac{1}{n_i^N n_t^T}\left(\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} h(\alpha_j,\gamma_s)+\sum_{\substack{j\in g_i,\\ j\neq i}} h(\alpha_j,\gamma_t)+\sum_{\substack{s\in c_t,\\ s\neq t}} h(\alpha_i,\gamma_s)\right) \\
&= h(\alpha_i,\gamma_t) \\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i)+\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t)+r(i,j,t,s)\right) \\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i)+r'(i,j,t)\right) \\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t)+r''(t,s,i)\right) \\
&= h(\alpha_i,\gamma_t)+\frac{1}{n_i^N}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i)+r'(i,j,t)\right) \\
&+ \frac{1}{n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t)+r''(t,s,i)\right) \\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s).
\end{aligned}$$

We explicitly split the sum in the second line to make clearer the fact that almost all terms cancel out once we difference these identities. From the last line it should be clear that,

$$\widetilde{\Gamma}_{it} = \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s).$$

From $h(.,.)$ being twice continuously differentiable and a uniformly bounded second derivative, we have from Cauchy-Schwarz

$$r(i,j,t,s) = O\left(\left\|\alpha_i-\alpha_j\right\|^2+\left\|\gamma_t-\gamma_s\right\|^2\right).$$

For the entry-wise 1-norm, we have,

$$\begin{aligned}\left\|\text{vec}(\widetilde{\Gamma})\right\|_1 &= \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s) \right| \\ &\leq \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\left\|\alpha_i-\alpha_j\right\|^2\right) \right| \\ &+ \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\left\|\gamma_t-\gamma_s\right\|^2\right) \right|.\end{aligned}$$

Now, concentrate on the first term,

$$\begin{aligned}&\sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\left\|\alpha_i-\alpha_j\right\|^2\right) \right| \\ &\leq \sum_i \sum_t \left| \frac{(n_i^N-1)(n_t^T-1)}{n_i^N n_t^T} \max_{\substack{j\in g_i,\\ j\neq i}} O\left(\left\|\alpha_i-\alpha_j\right\|^2\right) \right| \\ &= O(T) \sum_i \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\alpha_i-\alpha_j\right\|^2\end{aligned}$$

Use Assumption 3.4.1(iii) to show for $j \in g_i$,

$$\begin{aligned}\left\|\alpha_i-\alpha_j\right\|^2 &\leq B^2 \left\|\lambda(\alpha_i)-\lambda(\alpha_j)\right\|^2 \\ &= B^2 \left\|\lambda(\alpha_i)-\widehat{\lambda}_i-(\lambda(\alpha_j)-\widehat{\lambda}_j)+\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2 \\ &\leq B^2 \left(\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|+\left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|+\left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|\right)^2.\end{aligned}$$

An application of Cauchy-Schwarz and Assumption 3.4.1(iv) gives

$$\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\alpha_i-\alpha_j\right\|^2 \leq B^2 \sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left(\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|^2+\left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|^2+\left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2 \right),$$

hence we have

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\left\|\alpha_i-\alpha_j\right\|^2\right)\right| = NTO_P(\xi_{NT}).$$

The t-dimension analogy is direct such that

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\left\|\gamma_t-\gamma_s\right\|^2\right)\right| = NTO_P(\xi_{NT}).$$

Lastly, use Assumption 3.4.1.(vi), which implies $\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} = O_p(1/NT)$, to show

$$\begin{aligned}\|\kappa_{NT}\| &= O_p(\xi_{NT})\\ \Rightarrow \kappa_{NT} &= O_p(\xi_{NT})\end{aligned}$$

■

For each partition $\mathscr{O}_q$, with $q\in\{1,2,3,4\}$, the $N_q\times G^{(q)}$ matrix $D_\nu^{(q)}$, respectively $T_q\times C^{(q)}$ matrix $D_\delta^{(q)}$, represent the i, respectively t, cluster assignment matrices for $(i,t)\in\mathscr{O}_q$ where the columns of each matrix are binary indicators of cluster assignment. That is, any given column of $D_\nu^{(q)}$ represents a cluster equal to 1 if that row is a member of the cluster and 0 otherwise, and likewise for $D_\delta^{(q)}$. Here $G^{(q)}$ are the number of i clusters and $C^{(q)}$ are the number of t clusters in $\mathscr{O}_q$. For each partition define the annihilation matrix $M_\nu^{(q)} = \mathbb{I}_{N_q} - D_\nu^{(q)}\left([D_\nu^{(q)}]'D_\nu^{(q)}\right)^{-1}[D_\nu^{(q)}]'$ and $M_\delta^{(q)} = \mathbb{I}_{T_q} - D_\delta^{(q)}\left([D_\delta^{(q)}]'D_\delta^{(q)}\right)^{-1}[D_\delta^{(q)}]'$. To perform within-cluster mean-differences we can then take, for matrix $A^{(q)}$ being the partition $\mathscr{O}_q$ of matrix A, $\check{A}^{(q)} = M_\nu^{(q)}A^{(q)}M_\delta^{(q)}$.[1] Take $\check{A}$ as the block matrix with blocks $\check{A}^{(q)}$. Further, for each regressor, k, let $\check{X}_k$ be defined similarly for each k separately such that $\check{X}_{it}$ a K dimensional column vector.

[1] Note these are very similar to the $\tilde{A}$ variables in the main text, but here we make the distinction that projection is done at the partition level.

Assumption B.2.1. Let $\mathscr{O}_q$ denote partitions for cluster formation and $\mathscr{O}_q^*$ denote partitions for proxy sampling. Across each partition, $\alpha_i^{(q)}$ has common support $\mathscr{A}$ for each q, $\gamma^{(q)}$ has common support $\mathscr{C}$ for each q, and both of these are bounded and convex sets. Also, assume each partition is of equal size, up to rounding error, such that they all grow proportionally with N,T. There exists a sequence $\xi_{NT} > 0$ common to all partitions such that $\xi_{NT} \to 0$ as $N,T \to \infty$, and

(i) The function $h(\cdot,\cdot)$ is at least twice continuously differentiable with uniformly bounded second derivatives.

(ii) For each q, every unit $i \in \mathscr{O}_q$ is a member of exactly one group $g_i^{(q)} \in \{1,\ldots,G^{(q)}\}$, and every time period t is a member of exactly one group $c_t^{(q)} \in \{1,\ldots,C^{(q)}\}$. The size of all $G^{(q)}$ groups of units, and the size of all $C^{(q)}$ groups of time periods is bounded uniformly by $Q_{\max}$ for all q.

(iii) There exists $B > 0$ such that for all q there is, $\|a-b\| \leq B\left\|\lambda^{(q)}(a) - \lambda^{(q)}(b)\right\|$ for all $a,b \in \mathscr{A}$, and $\|a-b\| \leq B\left\|f^{(q)}(a) - f^{(q)}(b)\right\|$ for all $a,b \in \mathscr{C}$.

(iv) For each q there is,
$\frac{1}{N_q^* T_q^*}\sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathscr{O}_q^*\}\left(\left\|\widehat{\lambda}_i^{(q)} - \lambda^{(q)}(\alpha_i)\right\|^2\right) = O_P(\xi_{NT})$,
$\frac{1}{N_q^* T_q^*}\sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathscr{O}_q^*\}\left(\left\|\widehat{f}_t^{(q)} - f^{(q)}(\gamma_t)\right\|^2\right) = O_P(\xi_{NT})$.

(v) For each q there is,
$\frac{1}{N_q^* T_q^*}\sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathscr{O}_q^*\}\left\|\widehat{\lambda}_i^{(q)} - \widehat{\lambda}_{j(i)}^{(q)}\right\|^2 = O_P(\xi_{NT})$ for any matching function $(j(i),t) \in \mathscr{O}_q$ such that $g_i^{(q)} = g_{j(i)}^{(q)}$, and
$\frac{1}{N_q^* T_q^*}\sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathscr{O}_q^*\}\left\|\widehat{f}_t^{(o)} - \widehat{f}_{s(t)}^{(o)}\right\|^2 = O_P(\xi_{NT})$ for any matching function $(i,s(t)) \in \mathscr{O}_q$ such that $c_t^{(q)} = c_{s(t)}^{(q)}$.

(vi) $\max_{k,i,t}\left|\check{X}_{it,k}\right| = O_P(1)$, and $\operatorname{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \check{X}_{it}'\check{X}_{it} = \Omega$, where Ω is a positive definite non-random matrix.

Proof of Lemma 3.4.2. Recall from the proof of Lemma 3.4.1 the definition of

κ_{NT}. Take the split sample version as follows,

$$\begin{aligned}\kappa_{NT}^{(GS)} &:= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\check{X}_{it}'\check{X}_{it}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\check{X}_{it}'\check{\Gamma}_{it} \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\check{X}_{it}'\check{X}_{it}\right)^{-1}\sum_{o=1}^{4}\sum_{(i,t)\in\mathscr{O}_q}[\check{X}_{it}^{(q)}]'\check{\Gamma}_{it}^{(q)}.\end{aligned}$$

By Assumption B.2.1 and the proof steps of Lemma 3.4.1 we have that for each partition $\sum_{(i,t)\in\mathscr{O}_q}[\check{X}_{it}^{(q)}]'\check{\Gamma}_{it}^{(q)} = O_P(N_qT_q\xi_{NT})$, where N_q and T_q are the number of i and t, respectively, in partition q. Thus we have $\sum_{o=1}^{4}\sum_{(i,t)\in\mathscr{O}_q}[\check{X}_{it}^{(q)}]'\check{\Gamma}_{it}^{(q)} = \sum_{o=1}^{4}O_P(N_qT_q\xi_{NT}) \leq O_P(NT\xi_{NT})$. The statement of the lemma then follows from $\sum_{i=1}^{N}\sum_{t=1}^{T}\check{X}_{it}'\check{X}_{it} = O_P(NT)$. ■

Proof of Lemma 3.4.3. Using the definition of ϕ_{NT}^{GS} in the main text we have

$$\sqrt{NT}\,\phi_{NT}^{\mathrm{GS}} := \widehat{\Omega}^{-1}\sum_{s=1}^{4}\phi_{NT}^{(s)}$$

where

$$\widehat{\Omega} := \frac{1}{NT}\sum_{s=1}^{4}\sum_{(i,t)\in\mathscr{O}_s}\widetilde{X}_{it}^{(s)\prime}\widetilde{X}_{it}^{(s)}, \qquad \phi_{NT}^{(s)} := \frac{1}{\sqrt{NT}}\sum_{(i,t)\in\mathscr{O}_s}\widetilde{X}_{it}^{(s)\prime}\varepsilon_{it}.$$

By construction, the projected regressors $\widetilde{X}_{it}^{(s)}$ for subpanel $s\in\{1,2,3,4\}$ only depend on $X=(X_{it})$, and on outcomes Y_{it} (and thus error terms ε_{it}) that are not in that subpanel, i.e. $(i,t)\notin\mathscr{O}_s$. Therefore, under Assumption 3.4.2(i), we have that for $s\in\{1,2,3,4\}$, conditional on $\{\widetilde{X}_{it}^{(s)} : (i,t)\in\mathscr{O}_s\}$, the $\widetilde{X}_{it}^{(s)\prime}\varepsilon_{it}$ are mean zero and independently distributed across all the observations $(i,t)\in\mathscr{O}_s$ in that subpanel. Using the regularity conditions in Assumption 3.4.2(ii), for each $s\in\{1,2,3,4\}$, we can therefore apply Lyapunov's CLT to find

$$\left(\widehat{\Sigma}^{(s)}\right)^{-1}\phi_{NT}^{(s)} \Rightarrow \mathscr{N}(0,\mathbb{1}_K), \qquad \widehat{\Sigma}^{(s)} := \sum_{(i,t)\in\mathscr{O}_s}\sigma_{it}^2\widetilde{X}_{it}^{(s)\prime}\widetilde{X}_{it}^{(s)},$$

and the limiting distributions of $\left(\widehat{\Sigma}^{(s)}\right)^{-1}\phi_{NT}^{(s)}$ are independent across s. Using that

$\widehat{\Sigma}^{(s)}$ converges to the constant $\Sigma^{(s)}$ we thus find that

$$\sum_{s=1}^{4} \phi_{NT}^{(s)} \Rightarrow \mathcal{N}\left(0, \sum_{s=1}^{4} \Sigma^{(s)}\right).$$

Since $\widehat{\Omega}$ converges to $\Omega > 0$, the continuous mapping theorem then gives the statement of the lemma. ■

Appendix C

Appendix – Chapter 4

C.1 Proofs

Proof of Proposition 4.3.2. In the following, let $\text{vec}(\widetilde{\mathbf{X}})$ be the $\prod_n N_n \times K$ matrix of vectorised covariates after the within-cluster transformations where each column is a vectorised transformed covariate. The vec operator on other variables is the standard vectorisation operator. Also let $N = \prod_n N_n$ and the subscript $i = 1, \ldots, N$ be the index for the vectorised data when i has no subscript. Then,

$$\begin{aligned}
\beta_{GFE,\mathscr{C}} &= \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{Y}}) \\
&= \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\mathbf{X}})\beta^0 + \text{vec}(\widetilde{\mathscr{A}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right) \\
&= \beta^0 + \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\mathscr{A}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right),
\end{aligned}$$

such that,

$$\begin{aligned}
\left\|\beta_{GFE,\mathscr{C}} - \beta^0\right\| &= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\mathscr{A}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right) \right\| \\
&\leq \|\kappa_N\| + \|\omega_N\|
\end{aligned}$$

where

$$\begin{aligned}
\|\kappa_N\| &:= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathscr{A}}) \right\|; \\
\|\omega_N\| &:= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\varepsilon}}) \right\|.
\end{aligned}$$

The terms κ_N and ω_N are dealt with separately.

First to bound κ_N. Notice,

$$\|\kappa_N\| \leq \left\| \left(\text{vec}\left(\widetilde{\mathbf{X}}\right)' \text{vec}\left(\widetilde{\mathbf{X}}\right) \right)^{(-1)} \right\| \left\| \text{vec}\left(\widetilde{\mathbf{X}}\right)' \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\|$$

Focus on the right hand part, and let $\langle .,. \rangle_F$ be the Frobenius inner product,

$$\left\| \text{vec}\left(\widetilde{\mathbf{X}}\right)' \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\| = \left\| \begin{bmatrix} \langle \widetilde{X}_1, \widetilde{\mathscr{A}} \rangle_F \\ \vdots \\ \langle \widetilde{X}_K, \widetilde{\mathscr{A}} \rangle_F \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \sum_{i=1}^N \left| \widetilde{X}_{i,1} \widetilde{\mathscr{A}_i} \right| \\ \vdots \\ \sum_{i=1}^N \left| \widetilde{X}_{i,K} \widetilde{\mathscr{A}_i} \right| \end{bmatrix} \right\| \tag{A.1}$$

where the triangle inequality is used entry-wise. By Hölder's inequality

$$\sum_{i=1}^N \left| \widetilde{X}_{i,k} \widetilde{\mathscr{A}_i} \right| \leq \left\| \text{vec}\left(\widetilde{\mathbf{X}}_k\right) \right\| \left\| \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\| \quad \text{for each } k = 1, \ldots, K$$

This bounds the norm in (A.1) as,

$$\left\| \text{vec}\left(\widetilde{\mathbf{X}}\right)' \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\| \leq \sqrt{\sum_{k=1}^K \left\| \text{vec}\left(\widetilde{\mathbf{X}}_k\right) \right\|^2} \left\| \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\|.$$

From Assumption 4.3.1.(i) there is $\sqrt{\sum_{k=1}^K \left\| \text{vec}\left(\widetilde{\mathbf{X}}_k\right) \right\|^2} = O_p\left(\sqrt{\prod_n N_n}\right)$. This leaves $\left\| \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\|$. Take $g_n(i_n)$ as the indices in i_n's cluster such that $|g_n(i_n)|$ is the cluster size. Also, let $\bar{\varphi}_{i_n^*}^{(n)}$ be the cluster average for i_n's cluster. Then,

$$\begin{aligned} \left\| \text{vec}\left(\widetilde{\mathscr{A}}\right) \right\|^2 = \left\| \widetilde{\mathscr{A}} \right\|_F^2 &= \sum_{i_1,\ldots,i_d} \left(\sum_{\ell=1}^L \prod_{n=1}^d \left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)} \right) \right)^2 \\ \text{(Jensen's inequality)} \qquad &\leq L^2 \sum_{i_1,\ldots,i_d} \sum_{\ell=1}^L \frac{1}{L} \prod_{n=1}^d \left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)} \right)^2 \\ &\leq L \left(\prod_{n=1}^d N_n \right) \prod_{n=1}^d \frac{1}{N_n} \sum_{i_n} \left\| \varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)} \right\|^2. \end{aligned}$$

Expand the term, $\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2$,

$$\begin{aligned}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 &= \left\|\varphi_{i_n}^{(n)} - \frac{1}{|g_n(i_n)|}\sum_{j_n\in g_n(i_n)}\varphi_{j_n}^{(n)}\right\|^2 \\ &\leq \frac{1}{|g_n(i_n)|^2}\left(\sum_{j_n\in g_n(i_n)}\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|\right)^2 \\ &\leq \max_{j_n\in g_n(i_n)}\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2 \qquad \text{(A.2)}\end{aligned}$$

Then by Assumption 4.3.5,

$$\left\|\text{vec}\left(\widetilde{\mathbf{X}}\right)'\text{vec}\left(\widetilde{\mathscr{A}}\right)\right\| \leq \sqrt{L}\left(\prod_{n=1}^{d}N_n\right)O_p\left(\prod_{n\in\mathscr{M}}\sqrt{\xi_{N_n}}\right)$$

Lastly, Assumption 4.3.6.(i) implies the left hand term of $\|\kappa_N\|$ is $O_p(1/\prod_n N_n)$. This leaves

$$\|\kappa_N\| = \sqrt{L}O_p\left(\prod_{n^*\in\mathscr{M}}\sqrt{\xi_{N_{n^*}}}\right)$$

Finally, to bound $\|\omega_N\|$. Note that

$$\|\omega_N\| \leq \left\|\left(\text{vec}\left(\widetilde{\mathbf{X}}\right)'\text{vec}\left(\widetilde{\mathbf{X}}\right)\right)^{(-1)}\right\|_F\left\|\text{vec}\left(\widetilde{\mathbf{X}}\right)'\text{vec}\left(\widetilde{\boldsymbol{\varepsilon}}\right)\right\|.$$

Use Assumption 4.3.2 to bound the right hand term, $\left\|\text{vec}\left(\widetilde{\mathbf{X}}\right)'\text{vec}\left(\widetilde{\boldsymbol{\varepsilon}}\right)\right\| = O_p\left(\sqrt{\prod_n N_n}\right)$. Then, as above, the left hand term is $O_p(1/\prod_n N_n)$ such that

$$\|\omega_N\| = O_p\left(\frac{1}{\sqrt{\prod_n N_n}}\right).$$

■

Proof of Remark 4.4.1. Begin from A.2 in the proof of Proposition 4.3.2. The

right hand terms, $\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2$, are bound as,

$$\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2 \leq c_n^2 \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|^2,$$

by Remark 4.4.1.(iii). The result then follows immediately by applying the conditions Remark 4.4.1.(i) and (ii) after this inequality. ■

Proof of Proposition 4.3.3. As in the proof of Proposition 4.3.2, after the weighted within transformation, the estimation error can be written as,

$$\left\|\beta_{KFE,\mathscr{C}} - \beta^0\right\| \leq \|\kappa_N\| + \|\omega_N\|.$$

As above, $\|\omega_N\|$ is bounded at the parametric rate, so $\|\kappa_N\|$ is again the focus here. Again from above this can be bound as,

$$\|\kappa_N\| \leq \left\| \left(\operatorname{vec}\left(\widetilde{\mathbf{X}}\right)'\operatorname{vec}\left(\widetilde{\mathbf{X}}\right)\right)^{(-1)} \right\| \sqrt{\sum_{k=1}^{K} \left\|\operatorname{vec}\left(\widetilde{\mathbf{X}}_k\right)\right\|^2} \left\|\operatorname{vec}\left(\widetilde{\mathscr{A}}\right)\right\|.$$

As above, this is bounded as

$$\|\kappa_N\| \leq O_p\left(\prod_{n=1}^{d} N_n^{-1/2}\right) \left\|\operatorname{vec}\left(\widetilde{\mathscr{A}}\right)\right\|$$

The interactive fixed-effect approximation error can be summarised as,

$$\left\|\operatorname{vec}\left(\widetilde{\mathscr{A}}\right)\right\|^2 = \sum_{i_1,\ldots,i_d} \left(\sum_{\ell=1}^{L} \prod_{n=1}^{d} \left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)}\right)\right)^2.$$

By similar steps as the proof of Proposition 4.3.2 this can be bound by

$$\left\|\operatorname{vec}\left(\widetilde{\mathscr{A}}\right)\right\|^2 \leq L\left(\prod_{n=1}^{d} N_n\right) \prod_{n=1}^{d} \frac{1}{N_n} \sum_{i_n} \left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2.$$

Hence,

$$\|\kappa_N\|^2 \leq LO_p\left(\prod_{n=1}^{d}\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)}-\bar{\varphi}_{i_n^*}^{(n)}\right\|^2\right)$$

Concentrate on the term, $\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)}-\bar{\varphi}_{i_n^*}^{(n)}\right\|^2$, in each dimension separately,

$$\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)}-\bar{\varphi}_{i_n^*}^{(n)}\right\|^2 \leq \frac{1}{N_n}\sum_{i_n}\frac{\left(\sum_{j_n}k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)}-\widehat{\varphi}_{j_n}^{(n)}\right\|\right)\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|\right)^2}{\left(\sum_{j_n}k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)}-\widehat{\varphi}_{j_n}^{(n)}\right\|\right)\right)^2}, \quad \text{(A.3)}$$

where elementary norm bounds are used to bound the left hand side. Use as shorthand $\widehat{a}_{ij}^{(n)} := \left\|\widehat{\varphi}_{i_n}^{(n)}-\widehat{\varphi}_{j_n}^{(n)}\right\|$. This can be bound as, with $P_{N_n} := \sum_{j_n}k\left(\widehat{a}_{ij}^{(n)}/h_n\right)$,

$$\frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n}k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|^2 +$$
$$+\frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n}\sum_{j_n'\neq j_n}k\left(\widehat{a}_{ij}^{(n)}/h_n\right)k\left(\widehat{a}_{ij'}^{(n)}/h_n\right)\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n'}^{(n)}\right\|$$

Call

$$A := \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n}k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|^2$$

and

$$B := \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n}\sum_{j_n'\neq j_n}k\left(\widehat{a}_{ij}^{(n)}/h_n\right)k\left(\widehat{a}_{ij'}^{(n)}/h_n\right)\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n'}^{(n)}\right\|$$

Expand $\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\|$ around the proxies for each fixed-effect term and bound using the triangle inequality as,

$$\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n}^{(n)}\right\| \leq \left\|\varphi_{i_n}^{(n)}-\widehat{\varphi}_{i_n}^{(n)}\right\| + \left\|\varphi_{j_n}^{(n)}-\widehat{\varphi}_{j_n}^{(n)}\right\| + \left\|\widehat{\varphi}_{i_n}^{(n)}-\widehat{\varphi}_{j_n}^{(n)}\right\|.$$

Then,

$$A \leq \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \times \ldots \\ \times \ldots \left(\left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\| + \left\|\varphi_{j_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\| + \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)^2$$

and by a few applications of Cauchy-Schwarz inequality this is bound as

$$A \leq \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \times \ldots \\ \times \ldots O\left(\left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\|^2 + \left\|\varphi_{j_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|^2 + \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|^2\right)$$

Take

$$\frac{1}{N_n}\sum_{i_n}\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 O\left(\left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\|^2\right) \\ = \frac{1}{N_n}\sum_{i_n} O\left(\left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\|^2\right)\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \\ \leq O\left(\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\|^2\right) O\left(N_n\right),$$

where the last inequality comes from bounded kernel functions. Hence this term is $O_p\left(C_n^{-2}N_n\right)$. By similar arguments the second term is also $O_p\left(C_n^{-2}N_n\right)$. The final term,

$$\frac{1}{N_n}\sum_{i_n}\sum_{j_n} O\left(k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|^2\right) = \frac{1}{N_n}\sum_{i_n}\sum_{j_n} O\left(h^{2\alpha}\right) = O\left(N_n h^{2\alpha}\right)$$

from Assumption 4.3.7.

This establishes

$$A \leq \frac{N_n}{P_{N_n}^2}\left(O_p\left(C_n^{-2}\right) + O\left(h^{2\alpha}\right)\right).$$

Now it is shown that $B = O(N_n A)$ such that B is the leading asymptotic term.

$$
\begin{aligned}
B &= \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{i_n}\sum_{j_n}\sum_{j_n'\neq j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right) k\left(\widehat{a}_{ij'}^{(n)}/h_n\right) \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\| \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n'}^{(n)}\right\| \\
&= \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{j_n}\sum_{j_n'\neq j_n}\left(\sum_{i_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right) k\left(\widehat{a}_{ij'}^{(n)}/h_n\right) \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\| \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n'}^{(n)}\right\|\right) \\
&\leq \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{j_n}\left(\sum_{i_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2\right)^{1/2} \times \dots \\
&\cdots \times \sum_{j_n'\neq j_n}\left(\sum_{i_n} k\left(\widehat{a}_{ij'}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n'}^{(n)}\right\|^2\right)^{1/2} \\
&\leq \frac{1}{P_{N_n}^2}\frac{1}{N_n}\sum_{j_n}\left(\sum_{i_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2\right)^{1/2} \times \dots \\
&\cdots \times \sum_{j_n}\left(\sum_{i_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2\right)^{1/2} \\
&\leq \frac{1}{P_{N_n}^2}\frac{1}{N_n}N_n^2\left(\frac{1}{N_n}\sum_{j_n}\sum_{i_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2\right)^{1/2} \times \dots \\
&\cdots \times \left(\frac{1}{N_n}\sum_{i_n}\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2\right)^{1/2} \\
&= \frac{1}{P_{N_n}^2}\frac{1}{N_n}N_n\sum_{i_n}\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)^2 \left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2 = N_n A
\end{aligned}
$$

The second line is simply rearrangement, the third line is from the Cauchy-Schwarz inequality, the fourth line simply adds an additional weakly positive term to the final sum such that the inequality is valid, and the fifth line uses Jensen's inequality.

Hence, the leading factor

$$B \leq \frac{N_n^2}{P_{N_n}^2}\left(O_p\left(C_n^{-2}\right) + O\left(h^{2\alpha}\right)\right).$$

Now for $P_{N_n} = \sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)$. Assumption 4.3.8 implies that

$$(1/N_n)\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)$$

converges to a bounded and strictly positive sum, where boundedness comes from having bounded kernel functions. This means the inverse $\left((1/N_n)\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^{-1}$ also converges to a bounded and strictly positive sum, hence $\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^{-2} = O_p(1/N_n^2)$, so $\frac{N_n^2}{P_{N_n}^2} = O_p(1)$. This can also be found by noting $\frac{N_n^2}{P_{N_n}^2} = (P_{N_n}/N_n)^{-2}$, which is bounded $O(1)$ because P_{N_n}/N_n converges to a strictly positive constant.

Hence,

$$\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 \leq \left(O_p\left(C_n^{-2}\right) + O\left(h^{2\alpha}\right)\right).$$

Taking the product of these terms over all dimensions then forms the statement of the result.

■

C.2 Reducing the number of estimated parameters

Analysts may be concerned with the number of parameters implied by the least squares problem (4.12). In practice, this equation implies a total of $N_1N_2g(N) + N_1g(N)N_3 + g(N)N_2N_3$ parameters, where $g(N)$ is the number of groups in each dimension that may depend on total data size $N = \prod_n N_n$. This implies the number of fixed-effects parameters with respect to total data size is

$$\frac{g(N)\sum_{n=1}^{d}\prod_{n'\neq n}N_{n'}}{\prod_n N_n} = g(N)\cdot O\left(\frac{1}{\min_{n\in\{1,\ldots,d\}}N_n}\right) \tag{A.4}$$

Hence, in the linear setting, the loss of degrees of freedom is negligible as long as the group size $g(N)$ does not grow too fast with respect to total data size. However, this makes estimation in non-linear settings like (4.5) problematic because of the incidental parameter bias, see in Chen et al. (2021). For this reason it is useful to consider versions of the within-cluster transformation that do not require so many parameters. The following is a non-exhaustive list of methods to reduce the number of estimated parameters.

The first approach to consider is to simply ensure the group sizes are small

with respect to data size. To do this consider $g_n := g(N_n)$ as the number of groups in dimension n. Take a similar calculation to (A.4) to obtain the total number of parameters $\sum_n g_n \prod_{n' \neq n} N_{n'}$. It should then be clear that as long as $g_n = o(N_n)$, the number of estimated parameters is small with respect to total data size and the incidental parameter problem is asymptotically negligible. However, the condition that $g_n = o(N_n)$ may be highly restrictive. For example, if a sample of the unobserved parameter space is very disparate then this condition restricts the analyst to make poor approximations of the fixed-effects terms as each $\varphi^{(n^*)}_{i_{n^*},\ell} - \varphi^{(n^*)}_{j_{n^*}(i_{n^*}),\ell}$ will be very large. This is why it is important to consider the alternatives provided below. As can be seen in (4.13), the approximation error is multiplicative across dimensions, which means the analyst needs only to approximate a subset of these well. This fact is utilised in the below displays.

Consider clusters along just one dimension. The within-cluster transformation associated with this is simply,

$$\widetilde{A}_{ijt} = A_{ijt} - A_{i^* jt} = \sum_{\ell=1}^{L} (\varphi^{(1)}_{i\ell} - \bar{\varphi}^{(1)}_{i^*\ell}) \varphi^{(2)}_{j\ell} \varphi^{(3)}_{t\ell}.$$

Under some high-level assumptions on the unobserved fixed-effects, $\bar{\varphi}^{(1)}_{i^*\ell} = \varphi^{(1)}_{i\ell} + O\left(\frac{1}{N_1}\right)$. Also, the term $\bar{\varphi}^{(1)}_{i^*\ell}$ may have to be estimated - call the estimate $\widehat{\varphi}^{(1)}_{i^*\ell}$. Again, under some high-level assumptions, this could be estimated as $\widehat{\varphi}^{(1)}_{i^*\ell} = \bar{\varphi}^{(1)}_{i^*\ell} + O_p\left(\frac{1}{\sqrt{N_2 N_3}}\right)$. Combining this leaves the estimated $\widetilde{A}_{ijt} = O_p\left(\frac{1}{\min\{N_1, \sqrt{N_2 N_3}\}}\right)$. So selection of which dimension, d^*, to cluster and difference over solves the optimisation $d^* = \text{argmax}_{d \in \{1,2,3\}} \min\{N_d, \prod_{n,m \neq d; n \neq m} \sqrt{N_n N_m}\}$. This procedure requires $N_{n \neq d^*} N_{m \notin \{d^*, n\}} \times g(d^*)$ parameters to estimate, where $g(d^*)$ is the number of groups for dimension d^*. Of course, choice of d^* may also incorporate the number of parameters required for estimation. Note that this method does not automatically project the additive terms from $\mathscr{B}$, so this should be performed after an initial within projection.

This logic can be extended to a difference across two-dimensions as,

$$\widetilde{A}_{ijt} = A_{ijt} - A_{i^*j^*t} = \sum_{\ell=1}^{L} \left(\varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} - \bar{\varphi}_{i^*\ell}^{(1)} \bar{\varphi}_{j^*\ell}^{(2)} \right) \varphi_{t\ell}^{(3)}.$$

By the same reasoning as above this leads to

$$\widehat{A}_{ijt} = O_p \left(\left(\min_{d \in \{1,2\}} \min \left\{ N_d, \sqrt{N_{\{1,2\}\setminus d} N_3} \right\} \right)^{-1} \right)$$

The optimal dimensions to cluster and difference on is

$$\{d_1^*, d_2^*\} = \underset{d_1,d_2}{\operatorname{argmax}} \min_{d \in \{d_1,d_2\}} \min \left\{ N_d, \sqrt{N_{\{d_1,d_2\}\setminus d} N_{n \notin \{d_1,d_2\}}} \right\}.$$

This requires $g(d_1^*)g(d_2^*) \times N_{n \notin \{d_1^*,d_2^*\}}$ parameters.

Take a further difference to obtain

$$\widetilde{A}_{ijt} = \left(A_{ijt} - A_{i^*j^*t}\right) - \left(A_{ijt^*} - A_{i^*j^*t^*}\right) = \sum_{\ell=1}^{L} (\varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} - \bar{\varphi}_{i^*\ell}^{(1)} \bar{\varphi}_{j^*\ell}^{(2)})(\varphi_{t\ell}^{(3)} - \bar{\varphi}_{t^*\ell}^{(3)}).$$

This reduces to

$$\widetilde{A}_{ijt} = O_p \left(\left(\min_{d \in \{1,2\}} \min \left\{ N_d, \sqrt{N_{\{1,2\}\setminus d} N_3} \right\} \min \left\{ N_3, \sqrt{N1N2} \right\} \right)^{-1} \right),$$

which is smaller than the two cluster difference. d^* can be found similarly. This requires $g(d_1^*)g(d_2^*) \times N_{n \notin \{d_1^*,d_2^*\}} + N_{n \notin \{d_1^*,d_2^*\}} \min_{m \in \{d_1^*,d_2^*\}} N_m g(d_{n \neq m}^*)$.

The above parameter reduction exercises and specifically the choice of which dimension(s) to cluster on are also subject to the proxies used for clustering. For example, along some dimensions there may exist observable characteristics that provide a good signal of individual unobserved fixed-effect cluster. Diagnostics discussed in Section 4.3.1 also uncover which dimension exhibits low-rank variation, making it a good candidate for single dimension clustering. The d^*'s above are given as guides in applications where there is no obvious dimension to concentrate on when parameter reduction is required. It should also be clear that more esti-

mated parameters can lead to tighter asymptotic rates of decay in the unobserved remainder term, which becomes obvious in the asymptotic results discussed later. One last consideration when choosing from these reduced parameter options is the implication on the additive fixed-effects terms, where not all additive terms are automatically projected with each of these reduction methods.

Bibliography

AHN, S. C. AND A. R. HORENSTEIN (2013): "Eigenvalue ratio test for the number of factors," *Econometrica*, 81, 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, 219–255.

——— (2013): "Panel data models with multiple time-varying individual effects," *Journal of Econometrics*, 174, 1–14.

ALTONJI, J. G. AND R. L. MATZKIN (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053–1102.

AMJAD, M., D. SHAH, AND D. SHEN (2018): "Robust Synthetic Control," *Journal of Machine Learning Research*, 19, 1–51.

ATHEY, S., M. BAYATI, N. DOUDCHENKO, G. IMBENS, AND K. KHOSRAVI (2017): "Matrix Completion Methods for Causal Panel Data Models," .

AUERBACH, E. (2019): "Identification and estimation of a partially linear regression model using network data," *arXiv preprint arXiv:1903.09679*.

BAI, J. (2003): "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

——— (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77, 1229–1279.

BAI, J. AND S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2019a): "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data," .

——— (2019b): "Rank regularized estimation of approximate factor models," *Journal of Econometrics*, 212, 78–96.

BAI, J. AND P. WANG (2016): "Econometric analysis of large factor models," *Annual Review of Economics*, 8, 53–80.

BAI, Z. D., J. W. SILVERSTEIN, AND Y. Q. YIN (1988): "A note on the largest eigenvalue of a large dimensional sample covariance matrix," *J. Multivar. Anal.*, 26, 166–168.

BEYHUM, J. AND E. GAUTIER (2019): "Square-root nuclear norm penalized estimator for panel data models with approximately low-rank unobserved heterogeneity," .

——— (2022): "Factor and Factor Loading Augmented Estimators for Panel Regression With Possibly Nonstrong Factors," *Journal of Business & Economic Statistics*, 1–12.

BODELET, J. AND J. SHAN (2020): "Nonparametric additive factor models," *arXiv preprint arXiv:2003.13119*.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2021): "Discretizing unobserved heterogeneity," *Econometrica (Forthcoming)*.

BORDENAVE, C., S. COSTE, AND R. R. NADAKUDITI (2020): "Detection thresholds in very sparse matrix completion," *arXiv preprint arXiv:2005.06062*.

BREIMAN, L. AND J. H. FRIEDMAN (1985): "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, 80, 580–598.

CAI, J.-F., E. J. CANDÈS, AND Z. SHEN (2010): "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, 20, 1956–1982.

CANDÈS, E. J. AND B. RECHT (2009): "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 9, 717.

CANDES, E. J. AND T. TAO (2010): "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Transactions on Information Theory*, 56, 2053–2080.

CHAMBERLAIN, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

CHAN, M. K. AND S. KWOK (2020): "The PCDID Approach: Difference-in-Differences when Trends are Potentially Unparallel and Stochastic," Working Papers 2020-03, University of Sydney, School of Economics.

CHATTERJEE, S. (2015): "Matrix estimation by Universal Singular Value Thresholding," *Ann. Statist.*, 43, 177–214.

CHEN, M., I. FERNÁNDEZ-VAL, AND M. WEIDNER (2020): "Nonlinear factor models for network and panel data," *Journal of Econometrics*.

——— (2021): "Nonlinear factor models for network and panel data," *Journal of Econometrics*, 220, 296–324.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND A. GALICHON (2010): "Quantile and probability curves without crossing," *Econometrica*, 78, 1093–1125.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2013a): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

——— (2013b): "Average and quantile effects in nonseparable panel models," *Econometrica*, 81, 535–580.

CHERNOZHUKOV, V., C. HANSEN, Y. LIAO, AND Y. ZHU (2018): "Inference for Heterogeneous Effects using Low-Rank Estimation of Factor Slopes," .

——— (2020): "Inference for Low-Rank Models," .

CHUDIK, A. AND M. H. PESARAN (2013): "Large panel data models with cross-sectional dependence: a survey," *CAFE Research Paper*.

CHUDIK, A., M. H. PESARAN, AND E. TOSETTI (2011a): "Weak and strong cross-section dependence and estimation of large panels," *The Econometrics Journal*, 14, C45–C90.

——— (2011b): "Weak and strong cross-section dependence and estimation of large panels," *Econometrics Journal*, 14, 45–90.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883–931.

DE SILVA, V. AND L.-H. LIM (2008): "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM Journal on Matrix Analysis and Applications*, 30, 1084–1127.

DHAENE, G. AND K. JOCHMANS (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82, 991–1030.

DZEMSKI, A. (2019): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," *Review of Economics and Statistics*, 101, 763–776.

ELDEN, L. AND B. SAVAS (2011): "Perturbation theory and optimality conditions for the best multilinear rank approximation of a tensor," *SIAM journal on matrix analysis and applications*, 32, 1422–1450.

EVDOKIMOV, K. (2010): "Identification and estimation of a nonparametric panel data model with unobserved heterogeneity," *Department of Economics, Princeton University*.

FAZEL, S. M. (2003): "Matrix rank minimization with applications." .

FERNÁNDEZ-VAL, I., H. FREEMAN, AND M. WEIDNER (2021): "Low-rank approximations of nonseparable panel models," *The Econometrics Journal*, 24, C40–C77.

FERNÁNDEZ-VAL, I. AND M. WEIDNER (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192, 291–312.

FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000): "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and statistics*, 82, 540–554.

——— (2005): "The generalized dynamic factor model: one-sided estimation and forecasting," *Journal of the American Statistical Association*, 100, 830–840.

FREEMAN, H. AND M. WEIDNER (2022): "Linear panel regressions with two-way unobserved heterogeneity," *arXiv preprint arXiv:2109.11911*.

FREYBERGER, J. (2017): "Non-parametric Panel Data Models with Interactive Fixed Effects," *The Review of Economic Studies*, 85, 1824–1851.

GALVAO, A. F. AND K. KATO (2014): "Estimation and inference for linear panel data models under misspecification when both n and t are large," *Journal of Business & Economic Statistics*, 32, 285–309.

GAO, C., Y. LU, H. H. ZHOU, ET AL. (2015): "Rate-optimal graphon estimation," *The Annals of Statistics*, 43, 2624–2652.

GEMAN, S. (1980): "A limit theorem for the norm of random matrices," *Annals of Probability*, 8, 252–261.

GIGLIO, S., M. MAGGIORI, AND J. STROEBEL (2016): "No-bubble condition: Model-free tests in housing markets," *Econometrica*, 84, 1047–1091.

GIGLIO, S. AND D. XIU (2021): "Asset pricing with omitted factors," *Journal of Political Economy*, 129, 000–000.

GOBILLON, L. AND T. MAGNAC (2016a): "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *The Review of Economics and Statistics*, 98, 535–551.

——— (2016b): "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 98, 535–551.

GRAF, S. AND H. LUSCHGY (2002): "Rates of convergence for the empirical quantization error," *The Annals of Probability*, 30, 874–897.

GRAHAM, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

GRAHAM, B. S. AND J. L. POWELL (2012): "Identification and estimation of average partial effects in ñirregularî correlated random coefficient panel data models," *Econometrica*, 80, 2105–2152.

GRIEBEL, M. AND H. HARBRECHT (2013): "Approximation of bi-variate functions: singular value decomposition versus sparse grids," *IMA Journal of Numerical Analysis*, 34, 28–54.

——— (2014): "Approximation of bi-variate functions: singular value decomposition versus sparse grids," *IMA journal of numerical analysis*, 34, 28–54.

GUGGENBERGER, P. (2010): "The impact of a Hausman pretest on the size of a hypothesis test: The panel data case," *Journal of Econometrics*, 156, 337–343.

GUNSILIUS, F. AND S. M. SCHENNACH (2019): "Independent nonlinear component analysis," Tech. rep., cemmap working paper.

HALLIN, M. AND R. LIŠKA (2007): "Determining the number of factors in the general dynamic factor model," *Journal of the American Statistical Association*, 102, 603–617.

HARDING, M. AND C. LAMARCHE (2011): "Least squares estimation of a panel data model with multifactor error structure and endogenous covariates," *Economics Letters*, 111, 197–199.

HAUSMAN, J., G. LEONARD, AND J. D. ZONA (1994): "Competitive analysis with differenciated products," *Annales d'Economie et de Statistique*, 159–180.

HODERLEIN, S. AND H. WHITE (2012a): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics*, 168, 300–314.

——— (2012b): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Journal of Econometrics*, 168, 300–314.

HOLLAND, P. W., K. B. LASKEY, AND S. LEINHARDT (1983): "Stochastic blockmodels: First steps," *Social networks*, 5, 109–137.

HOLTZ-EAKIN, D., W. NEWEY, AND H. S. ROSEN (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–95.

HONORÉ, B. (1992): "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects," *Econometrica: Journal of the Econometric Society*, 60, 533–565.

HSIAO, C., H. STEVE CHING, AND S. KI WAN (2012): "A PANEL DATA APPROACH FOR PROGRAM EVALUATION: MEASURING THE BENEFITS OF POLITICAL AND ECONOMIC INTEGRATION OF HONG KONG WITH MAINLAND CHINA," *Journal of Applied Econometrics*, 27, 705–740.

IMAI, K. AND I. S. KIM (2019): "On the use of two-way fixed effects regression models for causal inference with panel data," *Unpublished paper: Harvard University*.

JUODIS, A. (2020): "This Shock is Different: Estimation and Inference in Misspecified Two-Way Fixed Effects Panel Regressions," *Working Paper*.

JUODIS, A. AND V. SARAFIDIS (2018): "Fixed T dynamic panel data estimators with multifactor errors," *Econometric Reviews*, 37, 893–929.

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

KAPETANIOS, G., L. SERLENGA, AND Y. SHIN (2019): "Testing for Correlated Factor Loadings in Cross Sectionally Dependent Panels," .

——— (2021): "Estimation and inference for multi-dimensional heterogeneous panel datasets with hierarchical multi-factor error structure," *Journal of Econometrics*, 220, 504–531.

KARABIYIK, H., F. C. PALM, AND J.-P. URBAIN (2019): "Econometric analysis of panel data models with multifactor error structures," *Annual Review of Economics*, 11, 495–522.

KIM, D. AND T. OKA (2014): "DIVORCE LAW REFORMS AND DIVORCE RATES IN THE USA: AN INTERACTIVE FIXED-EFFECTS APPROACH," *Journal of Applied Econometrics*.

KLOPP, O. ET AL. (2014): "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, 20, 282–303.

KOLDA, T. G. AND B. W. BADER (2009): "Tensor decompositions and applications," *SIAM review*, 51, 455–500.

KRUSKAL, J. B. (1989): "Rank, decomposition, and uniqueness for 3-way and N-way arrays," in *Multiway data analysis*, 7–18.

LATAŁA, R. (2005): "Some Estimates of Norms of Random Matrices," *Proceedings of the American Mathematical Society*, 133, 1273–1282.

LATALA, R. (2005): "Some estimates of norms of random matrices," *Proc. Amer. Math. Soc.*, 133, 1273–1282.

LATAŁA, R. (2005): "Some estimates of norms of random matrices," *Proceedings of the American Mathematical Society*, 133, 1273–1282.

LEE, N., H. R. MOON, AND M. WEIDNER (2012): "Analysis of interactive fixed effects dynamic linear panel regression with measurement error," *Economics Letters*, 117, 239–242.

LI, H., Q. LI, AND Y. SHI (2017a): "Determining the number of factors when the number of factors can increase with sample size," *Journal of Econometrics*, 197, 76–86.

LI, K. (2018): "Inference for factor model based average treatment effects," *Available at SSRN 3112775*.

LI, K. T. AND D. R. BELL (2017): "Estimation of average treatment effects with panel data: Asymptotic theory and implementation," *Journal of Econometrics*, 197, 65 – 75.

LI, X., A. WANG, J. LU, AND Z. TANG (2019): "Statistical performance of convex low-rank and sparse tensor recovery," *Pattern Recognition*, 93, 193–203.

LI, Y., D. SHAH, D. SONG, AND C. L. YU (2017b): "Nearest Neighbors for Matrix Estimation Interpreted as Blind Regression for Latent Variable Model," .

LU, X. AND L. SU (2016): "Shrinkage estimation of dynamic panel data models with interactive fixed effects," *Journal of Econometrics*, 190, 148–175.

MA, S., D. GOLDFARB, AND L. CHEN (2011): "Fixed point and Bregman iterative methods for matrix rank minimization," *Mathematical Programming*, 128, 321–353.

MANSKI, C. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica: Journal of the Econometric Society*, 55, 357–362.

MATYAS, L. (2017): "The econometrics of multi-dimensional panels," *Advanced studies in theoretical and applied econometrics. Berlin: Springer*.

MAZUMDER, R., T. HASTIE, AND R. TIBSHIRANI (2010): "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322.

MENZEL, K. (2018): "Bootstrap with cluster-dependence in two or more dimensions," *ArXiv eprints, New York University*.

——— (2021): "Bootstrap with Cluster-Dependence in Two or More Dimensions," *Econometrica*, 89, 2143–2188.

MOON, H. R., M. SHUM, AND M. WEIDNER (2018): "Estimation of random coefficients logit demand models with interactive fixed effects," *Journal of Econometrics*, 206, 613–644.

MOON, H. R. AND M. WEIDNER (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83, 1543–1579.

——— (2017): "Dynamic linear panel regression models with interactive fixed effects," *Econometric Theory*, 33, 158–195.

——— (2018): "Nuclear Norm Regularized Estimation of Panel Regression Models," .

NEGAHBAN, S. AND M. J. WAINWRIGHT (2012): "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, 13, 1665–1697.

NORKUTĖ, M., V. SARAFIDIS, T. YAMAGATA, AND G. CUI (2021): "Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure," *Journal of Econometrics*, 220, 416–446.

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

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

ORBANZ, P. AND D. M. ROY (2015): "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 437–461.

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

PESARAN, M. H. AND E. TOSETTI (2011): "Large panels with common factors and spatial correlation," *Journal of Econometrics*, 161, 182–202.

RABANSER, S., O. SHCHUR, AND S. GÜNNEMANN (2017): "Introduction to tensor decompositions and their applications in machine learning," *arXiv preprint arXiv:1711.10781*.

RENNIE, J. D. M. AND N. SREBRO (2005): "Fast Maximum Margin Matrix Factorization for Collaborative Prediction," in *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, USA: Association for Computing Machinery, ICML Õ05, 713–719.

RICHARDS, T. J. AND B. J. RICKARD (2021): "Dynamic model of beer pricing and buyouts," *Agribusiness*, 37, 685–712.

ROBERTSON, D. AND V. SARAFIDIS (2015): "IV estimation of panels with factor residuals," *Journal of Econometrics*, 185, 526–541.

SALEH, J. C. (2014): "Simple estimators for cross price elasticity parameters with product differentiation: panel data methods and testing," *Revista de la Competencia y la Propiedad Intelectual*, 10, 15–40.

SARAFIDIS, V. AND D. ROBERTSON (2009): "On the impact of error cross-sectional dependence in short dynamic panel estimation," *The Econometrics Journal*, 12, 62–81.

SILVERSTEIN, J. W. (1989): "On the eigenvectors of large dimensional sample covariance matrices," *J. Multivar. Anal.*, 30, 1–16.

SREBRO, N. AND T. JAAKKOLA (2003): "Weighted low-rank approximations," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 720–727.

STOCK, J. H. AND M. W. WATSON (2002): "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

SU, L. AND Q. CHEN (2013): "Testing homogeneity in panel data models with interactive fixed effects," *Econometric Theory*, 29, 1079–1135.

SU, L., Z. SHI, AND P. C. PHILLIPS (2016): "Identifying latent structures in panel data," *Econometrica*, 84, 2215–2264.

SU, L. AND X. WANG (2017): "On time-varying factor models: Estimation and testing," *Journal of Econometrics*, 198, 84–101.

SU, L., X. WANG, AND S. JIN (2019): "Sieve estimation of time-varying panel data models with latent structures," *Journal of Business & Economic Statistics*, 37, 334–349.

TOMIOKA, R., K. HAYASHI, AND H. KASHIMA (2010): "Estimation of low-rank tensors via convex optimization," *arXiv preprint arXiv:1010.0789*.

TOTTY, E. (2017): "The effect of minimum wages on employment: A factor model approach," *Economic Inquiry*, 55, 1712–1737.

TREMBLAY, C. H. AND V. J. TREMBLAY (1995): "Advertising, price, and welfare: evidence from the US brewing industry," *Southern Economic Journal*, 367–381.

VANNIEUWENHOVEN, N., R. VANDEBRIL, AND K. MEERBERGEN (2012): "A new truncation strategy for the higher-order singular value decomposition," *SIAM Journal on Scientific Computing*, 34, A1027–A1052.

WESTERLUND, J., Y. PETROVA, AND M. NORKUTE (2019): "CCE in fixed-T panels," *Journal of Applied Econometrics*, 34, 746–761.

WESTERLUND, J. AND J.-P. URBAIN (2013): "On the estimation and inference in factor-augmented panel regressions with correlated loadings," *Economics Letters*, 119, 247–250.

WHITE, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica: journal of the Econometric Society*, 817–838.

WOLFE, P. J. AND S. C. OLHEDE (2013): "Nonparametric graphon estimation," *arXiv preprint arXiv:1309.5936*.

XIONG, R. AND M. PELGER (2019): "Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference," .

XU, A.-B. (2020): "Tensor completion via a low-rank approximation pursuit," *arXiv preprint arXiv:2004.08872*.

XU, J., L. MASSOULI, AND M. LELARGE (2014): "Edge Label Inference in Generalized Stochastic Block Models: from Spectral Theory to Impossibility Results," .

XU, Y. (2017): "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76.

YIN, Y. Q., Z. D. BAI, AND P. KRISHNAIAH (1988): "On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix," *Probability Theory Related Fields*, 78, 509–521.

ZELENEEV, A. (2020): "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," .

Statement of Conjoint Work

Note on the joint work in Hugo Stuart Harold Freeman's thesis "Essays in the Econometric Theory of Panel and Multidimensional Data".

Chapter 2, "Low-Rank Approximations of Nonseparable Panel Models", was undertaken as joint work with Iván Fernández-Val and Martin Weidner.
Chapter 3, "Linear Panel Regression with Two-Way Unobserved Heterogeneity", was undertaken as joint work with Martin Weidner.
Chapter 4, "Multidimensional Interactive Fixed-Effects", is single authored work by Hugo Stuart Harold Freeman.

Contributions from the authors were equal in the case of the coauthored chapters. Signatures from coauthors: