Fernández-Val, Iván; Freeman, Hugo; Weidner, Martin

Working Paper

Low-rank approximations of nonseparable panel models

cemmap working paper, No. CWP52/20

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Fernández-Val, Iván; Freeman, Hugo; Weidner, Martin (2020) : Low-rank approximations of nonseparable panel models, cemmap working paper, No. CWP52/20, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2020.5220

This Version is available at:
https://hdl.handle.net/10419/241927

cemmap
centre for microdata methods and practice

Low-rank approximations of nonseparable panel models

Iván Fernández-Val
Hugo Freeman
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP52/20

UKRI Economic and Social Research Council

Low-Rank Approximations of Nonseparable Panel Models*

Iván Fernández-Val† Hugo Freeman‡ Martin Weidner‡

October 22, 2020

Abstract

We provide estimation methods for panel nonseparable models based on low-rank factor structure approximations. The factor structures are estimated by matrix-completion methods to deal with the computational challenges of principal component analysis in the presence of missing data. We show that the resulting estimators are consistent in large panels, but suffer from approximation and shrinkage biases. We correct these biases using matching and difference-in-difference approaches. Numerical examples and an empirical application to the effect of election day registration on voter turnout in the U.S. illustrate the properties and usefulness of our methods.

1 Introduction

Nonseparable models are useful to capture multidimensional unobserved heterogeneity, which is an important feature of economic data. The presence of this heterogeneity

*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 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.

†Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215-1403, USA. Email: ivanf@bu.edu

‡Department of Economics, University College London, Gower Street, London WC1E 6BT, UK, and and CeMMAP. Email: hugo.freeman.16@ucl.ac.uk, m.weidner@ucl.ac.uk

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 individual 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 (2012); Chernozhukov, Fernández-Val, Hahn and Newey (2013).

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 square 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, Bayati, Doudchenko, Imbens and Khosravi (2017) and Amjad, Shah and Shen (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, Candès and Shen, 2010; Ma, Goldfarb and Chen, 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, Shah, Song and Yu (2017) 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, Shah, Song and Yu (2017) 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, Shah and Shen (2018) proposed a similar debiasing procedure based on the estimated factor structure, but they rely on synthetic control methods instead of

matching.

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-difference 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 describes the model and effects of interest. Section 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 4. Section 5 reports the results of the numerical examples. All the proofs of the theoretical results are gathered in the Appendix.

2 Model and Effects of Interest

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

Assumption 1 (Model)**.**

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

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 support $\mathbb{X} \subseteq \mathbb{R}^{d_x}$; $\boldsymbol{A}_i$ 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

$$\boldsymbol{U}_{it} \overset{d}{=} \boldsymbol{U}_{js} \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T, \qquad \textit{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T}, \tag{2}$$

and

$$\boldsymbol{U}_{it} \perp\!\!\!\perp \boldsymbol{X}_{js} \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \qquad \textit{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T}, \tag{3}$$

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.*

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, Massouli and Lelarge (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.

The main restriction imposed by Assumption 1 is the unit and time homogeneity in (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, 2012; Chernozhukov, Fernández-Val, Hahn and Newey, 2013). It implies that time is randomly assigned, conditional on covariates and unobserved effects. The additional restriction in (3) is an exogeneity condition of $\boldsymbol{X}_{js}$ with respect to $\boldsymbol{U}_{it}$. Given (2), it is a mild condition as time homogeneity already imposes that any relationship between $\boldsymbol{U}_{it}$ and $\boldsymbol{X}_{js}$ can only be unit and time-invariant. Taken together, (2) and (3) impose that

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

The model considered is similar to the static model in Chernozhukov, Fernández-Val, Hahn and Newey (2013), 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). For example, it can include location and scale time effects. Second, Chernozhukov, Fernández-Val, Hahn and Newey (2013) impose that Y_{it} and $\boldsymbol{X}_{it}$ are identically distributed across i, which is stronger than the unit homogeneity in (2). Thus, unit homogeneity is conditional on the treatments and unobserved effects and therefore 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

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

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 1 below.

The structural function g is generally not identified, but can be used to construct 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 $\boldsymbol{X}_{it} = \boldsymbol{x} \in \mathbb{X}$ and drawing $\boldsymbol{U}_{it}(\boldsymbol{x}) \overset{d}{=} \boldsymbol{U}_{it} \mid \boldsymbol{A}^N, \boldsymbol{B}^T$, where we impose rank similarity on $\boldsymbol{U}_{it}(\boldsymbol{x})$ across the values of $\boldsymbol{x} \in \mathbb{X}$. The main effects of interest are the average structural functions (ASFs)

$$\mu_t(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right], \quad \mu(\boldsymbol{x}) := \frac{1}{T}\sum_{t=1}^{T} \mu_t(\boldsymbol{x}), \tag{4}$$

and the conditional average structural functions (CASFs)

$$\begin{aligned}
\mu_t(\boldsymbol{x} \mid \mathbb{X}_0) &:= \frac{1}{N_t(\mathbb{X}_0)}\sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{X}_{it} \in \mathbb{X}_0\}\mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right], \quad N_t(\mathbb{X}_0) = \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{X}_{it} \in \mathbb{X}_0\}, \\
\mu(\boldsymbol{x} \mid \mathbb{X}_0) &:= \frac{1}{n(\mathbb{X}_0)}\sum_{t=1}^{T} N_t(\mathbb{X}_0)\mu_t(\boldsymbol{x} \mid \mathbb{X}_0), \quad n(\mathbb{X}_0) = \sum_{t=1}^{T} N_t(\mathbb{X}_0),
\end{aligned} \tag{5}$$

where $\mathbb{X}_0 \subseteq \mathbb{X}$. 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 (4) and (5) 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 1. The purpose is to show that Assumption 1 covers a great variety of models commonly used in empirical analysis. Our estimation methods are generic in that we do not need

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

to specify the data generating process, besides of satisfying Assumption 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 1 (Linear factor model)**.** *Consider the linear panel model with factor structure in the error terms:*

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

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

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

for some unknown function m. Here, assuming that $U_{it}(\boldsymbol{x})$ is uniform is a normalization, since m can be arbitrary. This nonparametric single index model with unobserved effects is a special case of Assumption 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 single index models for count or censored responses are also covered by Assumption 1.

Example 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*

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

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 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, Steve Ching and Ki Wan (2012), Gobillon and Magnac (2016), Athey, Bayati, Doudchenko, Imbens and Khosravi (2017), Li and Bell (2017), Xu (2017), Li (2018), Bai and Ng (2019a), Xiong and Pelger (2019), and Chan and Kwok (2020). Example 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, Fernández-Val and Weidner 2020, Zeleneev 2020) or uses stochastic blockmodels or graphon models, which typically ignore the effect of covariates (e.g. Holland, Laskey and Leinhardt (1983), Wolfe and Olhede 2013, Gao, Lu, Zhou 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.

3 Estimation via Factor Structure Approximation

A natural starting point to estimate the effects in (4) and (5) is to use empirical analogs. This amounts to replace $\mathrm{E}\left[Y_{it}(\boldsymbol{x}) \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right]$ 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 under Assumption 1,

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

so that 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

$$\mathrm{E}\left[Y_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{B}_t = \boldsymbol{b}\right] =: m(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}), \tag{6}$$

where the function m does not vary with i and t, by Assumption 1. We show next how this function can be approximated and estimated using a low-rank factor structure.

3.1 Low-rank factor structure approximation

For ease of exposition we assume that the regressor domain $\mathbb{X}$ is finite in the rest of the paper. Accordingly, we denote the corresponding discrete covariate and its values by X_{it} and x instead of $\boldsymbol{X}_{it}$ and $\boldsymbol{x}$. For most of the analysis, we will focus on the binary treatment case where $\mathbb{X} = \{0, 1\}$.

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 (Sampling of $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$)**.** *(i) $\boldsymbol{A}_i$ is independent and identically distributed across $i \in \mathbb{N}$, (ii) $\boldsymbol{B}_t$ is independent and identically distributed over $t \in \mathbb{T}$, 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 imposes smoothness on $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$.

Assumption 3 (Smoothness of $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$)**.** *Let*

$$m(x, \boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{\infty} s_j(x)\, u_j(x, \boldsymbol{a})\, v_j(x, \boldsymbol{b}) \tag{7}$$

be the functional singular value decomposition of $m(x, \boldsymbol{a}, \boldsymbol{b})$*. We assume that*

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

Assumption 3 is a high-level condition on the singular values of the function $m(x, \boldsymbol{a}, \boldsymbol{b})$, which are defined by the decomposition (7). In this functional singular value decomposition the eigenfunctions $u_j(x, \boldsymbol{a}) \in \mathbb{R}$ and $v_j(x, \boldsymbol{b}) \in \mathbb{R}$ are normalized as $\mathrm{E}\, u_j(x, \boldsymbol{A}_i)^2 = 1$ and $\mathrm{E}\, v_j(x, \boldsymbol{B}_t)^2 = 1$, and they also satisfy the orthogonality conditions $\mathrm{E}\, u_j(x, \boldsymbol{A}_i) u_k(x, \boldsymbol{A}_i) = 0$ and $\mathrm{E}\, v_j(x, \boldsymbol{B}_t) v_k(x, \boldsymbol{B}_t) = 0$, for $j \neq k$. The singular values are sorted such as $s_1(x) \geq s_2(x) \geq s_3(x) \geq \ldots \geq 0$.

There is a large literature on singular value decompositions of functions, which shows that, under appropriate conditions, the singular values satisfy $s_j(x) \lesssim j^{-\alpha}$, 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 3 is therefore a high-level smoothness assumption on $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$.

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,

$$m(x, \boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{\infty} \underbrace{s_j(x)^{1/2} u_j(x, \boldsymbol{a})}_{=:\phi_j(x, \boldsymbol{a})} \underbrace{s_j(x)^{1/2} v_j(x, \boldsymbol{b})}_{=:\psi_j(x, \boldsymbol{b})} = \sum_{j=1}^{R} \phi_j(x, \boldsymbol{a}) \psi_j(x, \boldsymbol{b}) + \zeta_R(x, \boldsymbol{a}, \boldsymbol{b}). \tag{8}$$

The first term is the approximation and the second term is the approximation error. Under Assumption 3,

$$\zeta_R(x, \boldsymbol{a}, \boldsymbol{b}) \to 0 \quad \text{as} \quad R \to \infty.$$

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

$$\mathrm{E}\left[\zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)^2\right] = \mathrm{E}\left[\sum_{j=R+1}^{\infty} s_j(X_{it})\, u_j(X_{it}, \boldsymbol{A}_i)\, v_j(X_{it}, \boldsymbol{B}_t)\right]^2 \lesssim R^{1-2\alpha}.$$

Hence, $\zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)$ converges in mean square to zero.

Combining (6) and (8), we obtain the approximate factor model

$$Y_{it} = \boldsymbol{\lambda}_i(X_{it})^{\mathrm{T}} \boldsymbol{f}_t(X_{it}) + \zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t) + E_{it}, \quad E_{it} := Y_{it} - \mathrm{E}\left[Y_{it} \mid X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t\right], \tag{9}$$

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 $\nu_{it} := \zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t) + E_{it}$ contains the approximation error, $\zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)$, and the conditional expectation error, E_{it}. The factor structure $\boldsymbol{\lambda}_i(X_{it})^{\mathrm{T}} \boldsymbol{f}_t(X_{it})$ can be seen as a series approximation on unobserved individual and time effects to the function $m(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)$ if we let $R = R_{N,T}$ to grow with N and T such that $\zeta_R(X_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)$ vanishes as $N, T \to \infty$. The factor structure approximation is exact in some cases for fixed R. For instance, in Example 3

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

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

In the model (9) 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

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

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

$$\min_{\{\boldsymbol{\lambda}_i\}_{i=1}^N, \{\boldsymbol{f}_t\}_{t=1}^T} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}(x) \left(Y_{it} - \boldsymbol{\lambda}_i^{\mathrm{T}} \boldsymbol{f}_t\right)^2, \tag{10}$$

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.

3.2 Estimation by matrix completion methods

We start by expressing the program (10) 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

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

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, Bayati, Doudchenko, Imbens and Khosravi (2017) and Amjad, Shah and Shen (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 the program

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

where $\|\boldsymbol{\Gamma}\|_1$ is the nuclear norm of $\boldsymbol{\Gamma}$ and 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 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

$$\min_{\{\boldsymbol{\Gamma} \in \mathbb{R}^{N \times T}\}} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}(x) \left(Y_{it} - \Gamma_{it}\right)^2 + \rho(R_1) \|\boldsymbol{\Gamma}\|_1, \tag{11}$$

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, Hastie and Tibshirani, 2010).

Let $\widehat{\mathbf{\Gamma}}(x)$ be a solution to (11) 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, Candès and Shen, 2010; Ma, Goldfarb and Chen, 2011; Bai and Ng, 2019b). We propose two matching procedures to debias the estimator in Section 4.

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

$$E_{it}(x) := \begin{cases} E_{it} = Y_{it} - \Gamma_{it}^{\infty}(x) & \text{if } X_{it} = x, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

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 $X_{it} = x$. We assume $x \in \mathbb{X}$ throughout.

Recall that

$$\widehat{\mathbf{\Gamma}}(x) \in \underset{\mathbf{\Gamma}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} Q_{NT}(\mathbf{\Gamma}, \rho, x), \qquad Q_{NT}(\mathbf{\Gamma}, \rho, x) = \frac{1}{2}\sum_{(i,t)\in\mathbb{D}(x)} (Y_{it} - \Gamma_{it})^2 + \rho\|\mathbf{\Gamma}\|_1, \tag{13}$$

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

Assumption 4 could equivalently be replaced by the two high-level conditions

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

The first of those conditions is implied by Assumption 4 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 1. *Let the Assumptions 2, 3 and 4 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 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 1 does not restrict the assignment process that determines $\boldsymbol{X}^{NT}$. Notice also that Lemma 1 does not require Assumption 1 because $\Gamma^{\infty}(x)$ is a reduced-form parameter.

Lemma 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

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)\right]^2 = o_P(1), \tag{14}$$

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 (14). This approach shows that under certain conditions on a 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, Bayati, Doudchenko, Imbens and Khosravi (2017). Thus, if $M_{it} = \widehat{\Gamma}_{it}(x) - \Gamma_{it}^{\infty}(x)$ and $\boldsymbol{X}^{NT}$ satisfy restricted strong convexity, then (14) would follow from Lemma 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 1 is equal to $\mu(x)$ defined in (4). 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 (14), 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 5.1, $W_{it}(x) = n(1)^{-1}X_{it}$.

Theorem 1. *Let the Assumptions 1, 2, 3 and 4 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 $\rho_{NT}/\sqrt{NT} \to 0$. Define $P_{it}(x) := \Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right)$, and assume that $\min_{i,t} P_{it}(x) > 0$ and 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*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} P_{it}^{-1}(x)\, G_{it}(x) = o_P(1), \qquad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \Gamma_{it}^{\infty}(x)\, G_{it}(x) = o_P(1). \tag{15}$$

Then,

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

To interpret the conditions in Theorem 1, 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 (15) 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, Silverstein and Yin (1988), Yin, Bai and Krishnaiah (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 1 in the appendix, where 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.

An important restriction on the treatment design that is imposed by Theorem 1 is that $\Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right) > 0$ for all i and t. However, the key technical step in our proof of the theorem is Proposition 1 in the appendix, which does not necessarily require that strong condition.[3] We will not explore deviations from that assumption here, because we think that that $P_{it}(x) > 0$ is a plausible assumption in many applications. For example, in our empirical application in Section 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 $P_{it}(x) > 0$ is satisfied.

[3]This is because $P_{it}(x)$ need not be chosen equal to $\Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right)$ in that proposition, but verifying the conditions of the proposition is harder if $P_{it}(x)$ is chosen differently.

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 (8). 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 (13). 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 (13) 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 $\boldsymbol{\Gamma}(x)$, $\widehat{\boldsymbol{\Gamma}}(x)$ 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, Candès and Shen, 2010; Ma, Goldfarb and Chen, 2011; Bai and Ng, 2019b). Because of these biases, we explore alternative estimates for $\mu(x)$ in Section 4.

3.4 Covariates and fixed effects

As we mentioned in Section 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. 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, Bayati, Doudchenko, Imbens and Khosravi, 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 (11) by

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

Chernozhukov, Hansen, Liao and Zhu (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 $\widehat{\boldsymbol{\Gamma}}(x)$ 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

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

4 Debiasing Using Matching Methods

The matrix completion estimator of the ASF is generally biased. As we explained in Section 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 5.2.

To construct the estimators of $\boldsymbol{\Gamma}^{\infty}(x)$, we start by extracting the factor structure of $\widehat{\boldsymbol{\Gamma}}(x)$ in (13). 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

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

We also consider a two-way matching procedure that combines matching with a difference-in-difference 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

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

(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-difference 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,

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

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

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

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 5.2.

We develop the theory for a debised estimator that allows for multiple matches and estimated factors and loadings. Multiple matches are expected 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

$$\widetilde{Y}_{it}(x) = \begin{cases} Y_{it} & \text{if } X_{it} = x, \\ \dfrac{1}{n_{it}} \displaystyle\sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\} \left(Y_{is} + Y_{jt} - Y_{js}\right) & \text{if } X_{it} \neq x \text{ and } n_{it} > 0, \\ \frac{1}{n(x)} \sum_{(j,s) \in \mathbb{D}(x)} Y_{js} & \text{if } n_{it} = 0, \end{cases} \tag{16}$$

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 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 re-

[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.

spect 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 5 (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* $v_{NT}^2 = O_P(\xi_{NT})$.

(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*

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

Remark 1 (Assumption 5)**.** *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. *Under Assumptions 1 and 5,*

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

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 5.2, where $\widetilde{\mu}(x)$ is found to typically have much smaller bias than $\widehat{\mu}(x)$.

5 Numerical Examples

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 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 at 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-difference 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

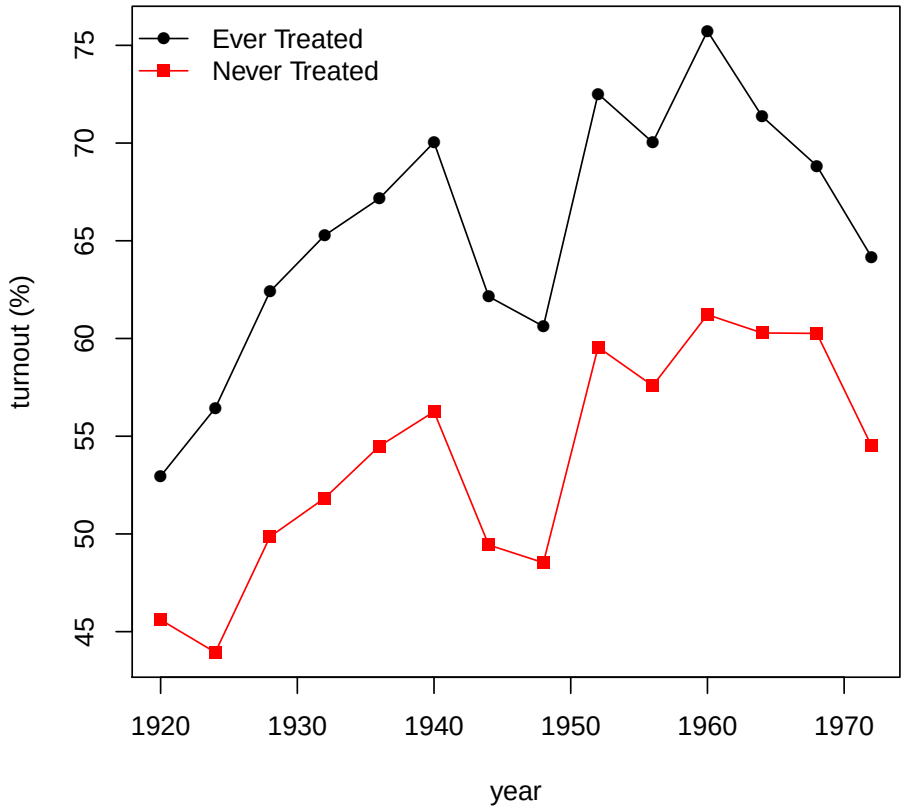

Figure 1: Pretrends in turnout rate before EDR by future treatment status

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 5.2.

Figure 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 estimates of the effect than the naive comparison of means between treated and nontreated states.

[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.

This finding agrees with the pre-EDR differences found in fig. 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.

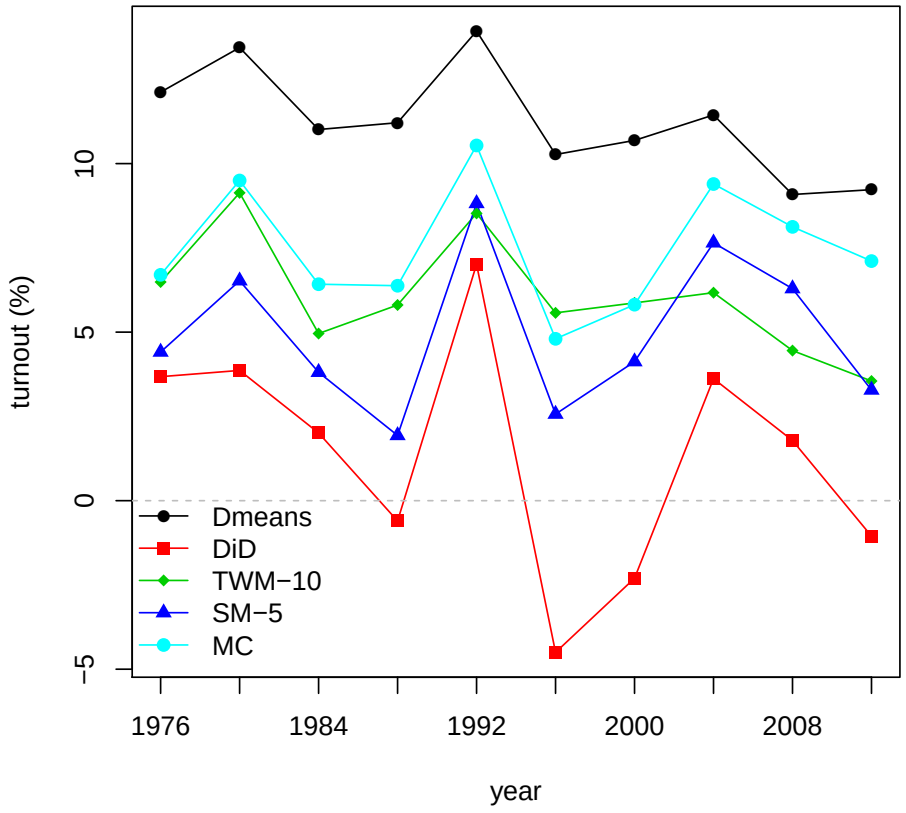

Figure 2: Average treatment effect on the treated of EDR on turnout rate at each election

Figure 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 (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 for states with low voter turnouts. Comparing with the Dquantiles estimates, we find that the sign of the selection bias switches from positive to negative around the middle of the turnout distribution.

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.,

$$g(a, b) = \frac{1}{\sqrt{2\pi}\theta} \exp\left(-\frac{(a-b)^2}{\theta^2}\right). \tag{17}$$

This design is similar to that used in Bordenave, Coste and Nadakuditi (2020), with kernel function specification from the numerical simulations in Griebel and Harbrecht (2010).[7] The parameter θ controls the decay of the singular values of g and can be calibrated to make sure the singular values decay slowly. Smaller values for θ lead to greater dispersion

[6] We rearrange the estimates of the distribution to guarantee that they are increasing with respect to y (Chernozhukov, Fernández-Val and Galichon, 2010).

[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.

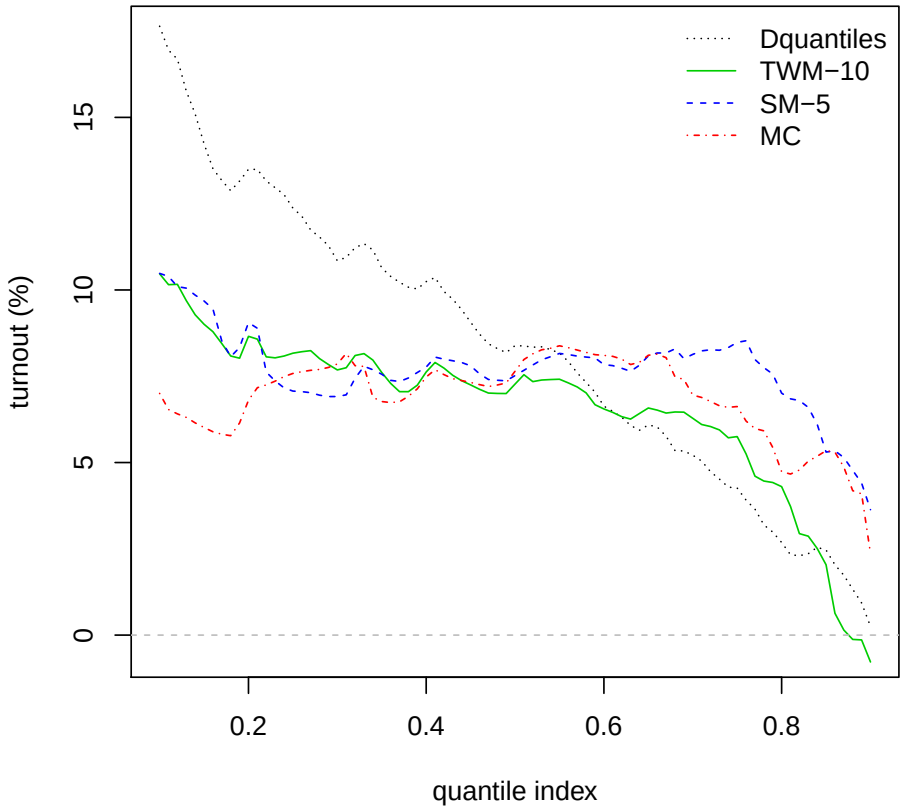

Figure 3: Time-averaged QTT of EDR on turnout rate

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, $i \in \{1, \ldots, 15\}$, and the second half of the panel, $t \in \{15, \ldots, 30\}$, 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 $\Pr(g(A_i, B_t) \geq c) = .5$.

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

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

	Bias	St. Dev.	RMSE
	Additive design		
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 5.1 to estimate the CASFs $\mu_t(0 \mid 1)$, $t \in \{15, \ldots, 30\}$, 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 values of $U_{it}(x)$ and hold A_i, B_t and X_{it} fixed. Table 1 reports the results for the time-aggregated CASF, $\mu(0 \mid 1)$, and Figure 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 3.3 and 4.

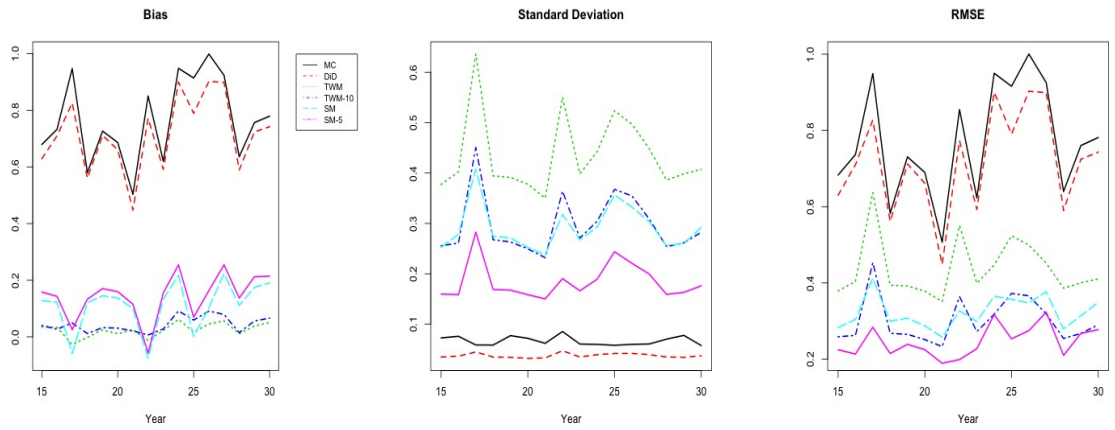

Figure 4: Results for $t \mapsto \mu_t(0 \mid 1)$.

References

Amjad, M., D. Shah, and D. Shen (2018). Robust synthetic control. *Journal of Machine Learning Research 19*(22), 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. and S. Ng (2019a). Matrix completion, counterfactuals, and factor analysis of missing data.

Bai, J. and S. Ng (2019b). Rank regularized estimation of approximate factor models. *Journal of Econometrics 212*(1), 78–96.

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*(2), 166–168.

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

Bordenave, C., S. Coste, and R. R. Nadakuditi (2020). Detection thresholds in very sparse matrix completion. *arXiv preprint arXiv:2005.06062*.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization 20*(4), 1956–1982.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics 9*(6), 717.

Candes, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory 56*(5), 2053–2080.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics 18*(1), 5–46.

Chan, M. K. and S. Kwok (2020, March). 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*(1), 177–214.

Chen, M., I. Fernández-Val, and M. Weidner (2020). Nonlinear factor models for network and panel data. *Journal of Econometrics*.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2018). Inference for heterogeneous effects using low-rank estimation of factor slopes.

Dzemski, A. (2019). An empirical model of dyadic link formation in a network with unobserved heterogeneity. *Review of Economics and Statistics 101*(5), 763–776.

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.

Freyberger, J. (2017, 09). Non-parametric Panel Data Models with Interactive Fixed Effects. *The Review of Economic Studies 85*(3), 1824–1851.

Gao, C., Y. Lu, H. H. Zhou, et al. (2015). Rate-optimal graphon estimation. *The Annals of Statistics 43*(6), 2624–2652.

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

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics 98*(3), 535–551.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica 85*(4), 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*(5), 2105–2152.

Griebel, M. and H. Harbrecht (2010). *Approximation of two-variate functions: Singular value decomposition versus regular sparse grids*. SFB 611.

Griebel, M. and H. Harbrecht (2013, 05). Approximation of bi-variate functions: singular value decomposition versus sparse grids. *IMA Journal of Numerical Analysis 34*(1), 28–54.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

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*(3), 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*(5), 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*.

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*(1), 282–303.

Latała, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society 133*(5), 1273–1282.

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*(1), 65 – 75.

Li, Y., D. Shah, D. Song, and C. L. Yu (2017). Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model.

Ma, S., D. Goldfarb, and L. Chen (2011). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming 128*(1-2), 321–353.

Manski, C. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica: Journal of the Econometric Society 55*(2), 357–362.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research 11*(80), 2287–2322.

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1), 158–195.

Moon, H. R. and M. Weidner (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*(1), 1665–1697.

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*(2), 437–461.

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*, ICML Õ05, New York, NY, USA, pp. 713–719. Association for Computing Machinery.

Silverstein, J. W. (1989). On the eigenvectors of large dimensional sample covariance matrices. *J. Multivar. Anal. 30*(1), 1–16.

Srebro, N. and T. Jaakkola (2003). Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 720–727.

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, 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*(1), 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.

A Proofs

A.1 Proof of Lemma 1

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 2. *Let Assumptions 2 and 3 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 2 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 1 and of Theorem 1. The proof of Lemma 2 is provided in Appendix A.4.

The following technical lemma provides the key step in the proof of Lemma 1 in the main text.

Lemma 3. *Under Assumptions 2 and 3,*

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

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

Notice that Lemma 3 is a non-stochastic finite sample result, which only requires that $E_{it}(x)$ and $\widehat{\mathbf{\Gamma}}(x)$ are as defined in (12) and (13). The proof of Lemma 3 is provided in Appendix A.4.

We are now ready to provide the proof of the lemma in the main text.

Proof of Lemma 1. The definition of $E_{it}(x)$ in (12) guarantees that $\mathrm{E}\left[E_{it}(x) \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \boldsymbol{X}^{NT}\right] = 0$, and Assumption 4 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, $\Gamma_{it}^{\infty}(x) = m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$ only depends on $\boldsymbol{A}^N$ and $\boldsymbol{B}^T$. We therefore find

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

where b is the constant from Assumption 4. From this we conclude that

$$\frac{1}{n(x)}\sum_{(i,t)\in\mathbb{D}(x)}\Gamma_{it}^{\infty}(x)\,E_{it} = O_P\left(\frac{1}{n^{1/2}(x)}\right) = o_P(1). \tag{18}$$

Next, applying Assumption 4 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 3 to find that, with probability approaching one, we have

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

where we applied (18) and Lemma 2, as well as the condition $\rho_{NT}\sqrt{NT}/n(x) \to 0$. ■

A.2 Proof of Theorem 1

In the following consider a generic reduced form parameter

$$\nu(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(x)\, \Gamma_{it}^{\infty}(x), \tag{19}$$

with corresponding estimator

$$\widehat{\nu}(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(x)\, \widehat{\Gamma}_{it}(x), \tag{20}$$

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 1. *Let the Assumptions 2, 3 and 4 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

$$|\widehat{\nu}(x) - \nu(x)| \leq c_4.$$

The proof of Proposition 1 is provided in Appendix A.4. Proposition 1 is the key step required for the proof of Theorem 1. However, before proving this main text result we want to provide an informal remark on the usefulness of Proposition 1 more generally.

Remark 2 (Consistency of $\widehat{\nu}(x)$)**.** *Proposition 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

$$\begin{aligned}
c_1 &= q + O_P(1/\sqrt{NT}),\\
c_2 &= O_P(1/\sqrt{NT})\\
c_3 &= \frac{2\,\rho}{q\,NT}\|\boldsymbol{\Gamma}^\infty(x)\|_1 + O_P(1/\sqrt{NT}),\\
c_4 &= \sqrt{\frac{2\,\rho}{q\,NT}\|\boldsymbol{\Gamma}^\infty(x)\|_1} + \textit{smaller order terms.}
\end{aligned}$$

Thus, if, like in Lemma 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

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

The following proof formalizes this heuristic argument for the case that $W_{it}(x) = 1$.

Proof of Theorem 1. Let $W_{it}(x) = 1$, and let $\nu(x)$ and $\widehat{\nu}(x)$ be as defined in (19) and (20) above. We then have

$$\begin{aligned}
\mu(x) &= \nu(x),\\
\widehat{\mu}(x) &= \widehat{\nu}(x) + \frac{1}{NT}\sum_{(i,t)\in\mathbb{D}(x)} E_{it}(x) - \frac{1}{NT}\sum_{(i,t)\in\mathbb{D}(x)}\left[\widehat{\Gamma}_{it}(x) - \Gamma_{it}^\infty(x)\right]. \qquad (21)
\end{aligned}$$

We drop all the arguments x in the rest of this proof. We want to apply Proposition 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 1, 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 2 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 (18) 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 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 1 to find that with probability approach one we have

$$|\widehat{\nu} - \nu| \leq c_4 = o_P(1).$$

We have thus shown that $\widehat{\nu} = \nu + o_P(1)$.

Furthermore, analogous to the result in (18) 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 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 (21) we find $\widehat{\mu}(x) = \mu(x) + o_P(1)$. ■

A.3 Proof of Theorem 2

In this section we present and prove a more general version of Theorem 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 (19) 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 (16). 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 5 in the main text is generalized as follows.

Assumption 6. *There exists a sequence $\xi_{NT} > 0$ such that $\xi_{NT} \to 0$ as $N, T \to \infty$, and*

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

(ii) Y_{it} *and* $W_{it}(x)$ *are 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 twice continuously differentiable with uniformly bounded second derivatives.*

(v) 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* $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{B}$.

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

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

(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*

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

The generalized version of Theorem 2 is given in the following.

Theorem 3. *Under Assumptions 1 and 6,*

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

Proof of Theorem 3 (containing Theorem 2 as a special case). Define $m_{it}(x) := m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$. We decompose

$$\widetilde{\nu}(x) - \nu(x) = e_0(x) + e_1(x) + e_2(x), \tag{22}$$

where

$$e_0(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(x)\,\mathbb{1}\{X_{it}\neq x\,\&\,n_{it}=0\}\,[m_{it}(X_{it}) - m_{it}(x)],$$

and

$$e_1(x) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{1}\{X_{it}\neq x\,\&\,n_{it}>0\}\;W_{it}(x)\,e_{1,it}(x),$$

$$e_{1,it}(x) := \frac{\sum_{j\in\mathbb{N}_i}\sum_{s\in\mathbb{T}_t}\mathbb{1}\{X_{is}=X_{jt}=X_{js}=x\}\,[m_{is}(x)+m_{jt}(x)-m_{js}(x)-m_{it}(x)]}{\sum_{j\in\mathbb{N}_i}\sum_{s\in\mathbb{T}_t}\mathbb{1}\{X_{is}=X_{jt}=X_{js}=x\}},$$

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 6(i) and (ii) guarantee that

$$|e_0(x)| \leq \left(\max_{it} |m_{it}(X_{it}) - m_{it}(x)|\right) \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(x) \, \mathbb{1}\left\{X_{it} \neq x \,\&\, n_{it} = 0\right\}$$
$$= O_P\left(\xi_{NT}\right). \tag{23}$$

Bound on $e_1(x)$: Assumption 6(iv) guarantees that there exists a constant $b > 0$ such that

$$\left| m(x, \boldsymbol{a}, \boldsymbol{b}) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t) - (\boldsymbol{a} - \boldsymbol{A}_i)' \frac{\partial m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)}{\partial \boldsymbol{A}_i} - (\boldsymbol{b} - \boldsymbol{B}_t)' \frac{\partial m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)}{\partial \boldsymbol{B}_t} \right|$$
$$\leq b \left(\|\boldsymbol{a} - \boldsymbol{A}_i\|^2 + \|\boldsymbol{b} - \boldsymbol{B}_t\|^2 \right).$$

Using this we find that

$$m_{is}(x) + m_{jt}(x) - m_{js}(x) - m_{it}(x) \leq 2\, b \, \left(\|\boldsymbol{A}_i - \boldsymbol{A}_j\|^2 + \|\boldsymbol{B}_t - \boldsymbol{B}_s\|^2 \right),$$

and therefore

$$|e_{1,it}(x)| \leq \frac{2\, b \sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\} \left(\|\boldsymbol{A}_i - \boldsymbol{A}_j\|^2 + \|\boldsymbol{B}_t - \boldsymbol{B}_s\|^2 \right)}{\sum_{j \in \mathbb{N}_i} \sum_{s \in \mathbb{T}_t} \mathbb{1}\{X_{is} = X_{jt} = X_{js} = x\}}$$
$$\leq 2\, b \left(\max_{j \in \mathbb{N}_i} \|\boldsymbol{A}_i - \boldsymbol{A}_j\|^2 + \max_{s \in \mathbb{T}_t} \|\boldsymbol{B}_t - \boldsymbol{B}_s\|^2 \right).$$

We thus find

$$|e_1(x)| \leq 2\, b \left(\max_{ij} |W_{it}(x)| \right) \left(\frac{1}{N} \sum_{i=1}^{N} \max_{j \in \mathbb{N}_i} \|\boldsymbol{A}_i - \boldsymbol{A}_j\|^2 + \frac{1}{T} \sum_{t=1}^{T} \max_{s \in \mathbb{T}_t} \|\boldsymbol{B}_t - \boldsymbol{B}_s\|^2 \right)$$
$$\leq 2\, b\, c \left(\max_{ij} |W_{it}(x)| \right) \left(\frac{1}{N} \sum_{i=1}^{N} \max_{j \in \mathbb{N}_i} \|\boldsymbol{\phi}(\boldsymbol{A}_i) - \boldsymbol{\phi}(\boldsymbol{A}_j)\|^2 + \frac{1}{T} \sum_{t=1}^{T} \max_{s \in \mathbb{T}_t} \|\boldsymbol{\psi}(\boldsymbol{B}_t) - \boldsymbol{\psi}(\boldsymbol{B}_s)\|^2 \right)$$
$$= 2\, b\, c \left(\max_{ij} |W_{it}(x)| \right) \left(\frac{1}{N} \sum_{i=1}^{N} \max_{j \in \mathbb{N}_i} \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|^2 + \frac{1}{T} \sum_{t=1}^{T} \max_{s \in \mathbb{T}_t} \|\boldsymbol{\psi}_t - \boldsymbol{\psi}_s\|^2 \right).$$

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

$$\max_{j \in \mathbb{N}_i} \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|^2 \leq \max_{j \in \mathbb{N}_i} \left(\left\| \widehat{\boldsymbol{\phi}}_i - \widehat{\boldsymbol{\phi}}_j \right\| + \left\| \widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \right\| + \left\| \widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j \right\| \right)^2$$
$$\leq \max_{j \in \mathbb{N}_i} \left(\tau_{NT} + \left\| \widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \right\| + \left\| \widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j \right\| \right)^2$$
$$\leq 3\tau_{NT}^2 + 3 \left\| \widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \right\|^2 + 3 \max_{j \in \mathbb{N}_i} \left\| \widehat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j \right\|^2.$$

Analogously we find

$$\max_{s\in\mathbb{T}_t}\|\boldsymbol{\psi}_t-\boldsymbol{\psi}_s\|^2\le 3\upsilon_{NT}^2+3\left\|\widehat{\boldsymbol{\psi}}_t-\boldsymbol{\psi}_t\right\|^2+3\max_{s\in\mathbb{T}_t}\left\|\widehat{\boldsymbol{\psi}}_s-\boldsymbol{\psi}_s\right\|^2.$$

We thus obtain

$$\begin{aligned}|e_1(x)|\le 6\,b\,c\,\left(\max_{ij}|W_{it}(x)|\right)\Bigg\{&\tau_{NT}^2+\upsilon_{NT}^2\\&\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\widehat{\boldsymbol{\phi}}_i-\boldsymbol{\phi}_i\right\|^2+\max_{j\in\mathbb{N}_i}\left\|\widehat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right\|^2\right)\\&\frac{1}{T}\sum_{t=1}^{T}\left(\left\|\widehat{\boldsymbol{\psi}}_t-\boldsymbol{\psi}_t\right\|^2+\max_{s\in\mathbb{T}_t}\left\|\widehat{\boldsymbol{\psi}}_s-\boldsymbol{\psi}_s\right\|^2\right)\Bigg\}\\=O_P\left(\xi_{NT}\right).&\end{aligned}\tag{24}$$

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

$$\begin{aligned}T_0&:=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{it}^2(\boldsymbol{Y}^{NT})\,E_{it}^2,\\T_1&:=\frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\left\{(i,t)\neq(j,s)\right\}\\&\qquad\times\left[\omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT})-\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\,\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\right]\,E_{it}\,E_{js},\\T_2&:=\frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{1}\left\{(i,t)\neq(j,s)\right\}\omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\,\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\,E_{it}\,E_{js}.\end{aligned}$$

We have

$$\begin{aligned}\mathrm{E}\left[T_0\,\middle|\,\boldsymbol{X}^{NT},\,\boldsymbol{A}^N,\,\boldsymbol{B}^T\right]&\le\left(\max_{i,t}|E_{it}|\right)^2\frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{E}\left[\omega_{it}^2\left(\boldsymbol{Y}^{NT}\right)\,\middle|\,\boldsymbol{X}^{NT},\,\boldsymbol{A}^N,\,\boldsymbol{B}^T\right]\\&=O_P(\xi_{NT}^2),\end{aligned}$$

and

$$\begin{aligned}
&\left|\mathrm{E}\left[T_1 \,\middle|\, \boldsymbol{X}^{NT},\, \boldsymbol{A}^N,\, \boldsymbol{B}^T\right]\right| \\
&\leq \left(\max_{i,t}|E_{it}|\right)^2 \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T} \mathbb{1}\left\{(i,t)\neq(j,s)\right\} \\
&\quad \times \mathrm{E}\left[\left|\omega_{it}(\boldsymbol{Y}^{NT})\,\omega_{js}(\boldsymbol{Y}^{NT}) - \omega_{it}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\,\omega_{js}\left(\boldsymbol{Y}^{NT}_{-(i,t),-(j,s)}\right)\right| \,\middle|\, \boldsymbol{X}^{NT},\, \boldsymbol{A}^N,\, \boldsymbol{B}^T\right] \\
&= O_P(\xi_{NT}^2).
\end{aligned}$$

where we used that Y_{it} (and thus E_{it}) is uniformly bounded, together with Assumption 6(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 (22), (23), and (24) this gives the statement of the theorem. ■

A.4 Proof of Intermediate Results

Proof of Lemma 2. 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

$$
\begin{aligned}
\left\|\boldsymbol{\Gamma}^{\infty}(x)\right\|_1 &\leq \sum_{j=1}^{\infty} s_j(x)\ \left\|\boldsymbol{u}_j(x)\right\|\ \left\|\boldsymbol{v}_j(x)\right\| \\
&= \sqrt{NT}\sum_{j=1}^{\infty} s_j(x)\sqrt{\frac{1}{N}\sum_{i=1}^{N}[u_j(x,\boldsymbol{A}_i)]^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}[v_j(x,\boldsymbol{B}_t)]^2} \\
&\leq \sqrt{NT}\sum_{j=1}^{\infty} s_j(x)\left(1+\frac{\frac{1}{N}\sum_{i=1}^{N}[u_j(x,\boldsymbol{A}_i)]^2-1}{2}\right)\left(1+\frac{\frac{1}{T}\sum_{t=1}^{T}[v_j(x,\boldsymbol{B}_t)]^2-1}{2}\right) \\
&= \sqrt{NT}\sum_{j=1}^{\infty} s_j(x)+\sqrt{NT}\,R_{NT} \\
&= \sqrt{NT}\ \left\|m(x,\cdot,\cdot)\right\|_* + \sqrt{NT}\,R_{NT},
\end{aligned}
$$

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 3 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 and the WLLN therefore guarantees that $R_{NT}=o_P(1)$. We have thus shown that $\left\|\boldsymbol{\Gamma}^{\infty}(x)\right\|_1 \leq \sqrt{NT}\ \left\|m(x,\cdot,\cdot)\right\|_* + o_P(\sqrt{NT})$, and since $\left\|m(x,\cdot,\cdot)\right\|_*$ is finite and non-random we also have $\left\|\boldsymbol{\Gamma}^{\infty}(x)\right\|_1 = O_P(\sqrt{NT})$. ■

Proof of Lemma 3. The nuclear norm (or trace norm) can be defined by

$$
\left\|\boldsymbol{\Gamma}\right\|_1 = \max_{\{\boldsymbol{M}\in\mathbb{R}^{N\times T}:\left\|\boldsymbol{M}\right\|_{\infty}\leq 1\}} \underbrace{\mathrm{Tr}\left(\boldsymbol{M}'\boldsymbol{\Gamma}\right)}_{=\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}\Gamma_{it}}. \tag{25}
$$

Our assumption $\rho\geq\left\|\boldsymbol{E}(x)\right\|_{\infty}$ guarantees that a possible choice in this maximization is $\boldsymbol{M}=\rho^{-1}\boldsymbol{E}(x)$, and we therefore have

$$
\rho\left\|\boldsymbol{\Gamma}\right\|_1 \geq \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(x)\,E_{it}(x)\,\Gamma_{it}.
$$

Using this and the model $Y_{it} = \Gamma_{it}^{\infty}(x) + E_{it}(x)$ 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 1. In this proof we drop the argument x everywhere, and we define $\theta = NT\nu$ and $\theta_0 = NT\nu$. Define the NT-vectors $\boldsymbol{\gamma} = \text{vec}(\mathbf{\Gamma})$, $\mathbf{\Gamma}^{\infty} = \text{vec}(\mathbf{\Gamma}^{\infty})$, $\boldsymbol{w} = \text{vec}(W_{it} : i \in \mathbb{N}, t \in \mathbb{T})$, $\boldsymbol{d} = \text{vec}(D_{it} : i \in \mathbb{N}, t \in \mathbb{T})$, 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

$$
\widehat{\theta} := \operatorname*{argmin}_{\theta \in \mathbb{R}} L_{NT}(\theta, \rho) = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\widehat{\Gamma}_{it}
$$

is close to $\theta := \boldsymbol{w}'\mathbf{\Gamma}^{\infty} = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\Gamma_{it}^{\infty}$. Using the definition of $Q_{NT}(\mathbf{\Gamma}, \rho)$ and $Y_{it} = \Gamma_{it}^{\infty} + E_{it}$ we find that

$$
L_{NT}(\theta, \rho) \leq Q_{NT}(\mathbf{\Gamma}^{\infty}, \rho) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\,E_{it}^2 + \rho\|\mathbf{\Gamma}^{\infty}\|_1. \tag{26}
$$

If for a given value of $\theta = \boldsymbol{w}'\boldsymbol{\gamma}$ we have that the matrix $\boldsymbol{M}(\theta)$ with elements $M_{it}(\theta) := D_{it}\,E_{it} - \frac{\boldsymbol{w}'(\boldsymbol{\gamma} - \mathbf{\Gamma}^{\infty})}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}\frac{(D_{it} - P_{it})W_{it}}{P_{it}}$ satisfies $\|\boldsymbol{M}(\theta)\|_{\infty} \leq \rho$, then by the definition of $\|\cdot\|_1$ in

(25) 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 $Y_{it} = \Gamma_{it}^\infty + E_{it}$ we find that

$$
\begin{aligned}
Q_{NT}(\mathbf{\Gamma}, \rho) &= \frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T D_{it}\left(Y_{it} - \Gamma_{it}\right)^2 + \rho\|\mathbf{\Gamma}\|_1 \\
&\geq \frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T D_{it}\left[(\Gamma_{it}^\infty - \Gamma_{it}) + E_{it}\right]^2 + \sum_{i=1}^N\sum_{t=1}^T \left\{ D_{it}\, E_{it} - \frac{[(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\,\boldsymbol{w}]}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} \frac{(D_{it} - P_{it})W_{it}}{P_{it}} \right\} \Gamma_{it} \\
&= \underbrace{\frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T D_{it}\left(\Gamma_{it} - \Gamma_{it}^\infty\right)^2 - \frac{[(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\,\boldsymbol{w}]}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} \sum_{i=1}^N\sum_{t=1}^T \frac{(D_{it} - P_{it})W_{it}}{P_{it}}\left(\Gamma_{it} - \Gamma_{it}^\infty\right)}_{=:Q_{NT}^{(\text{low},1)}(\mathbf{\Gamma})} \\
&\qquad + \underbrace{\sum_{i=1}^N\sum_{t=1}^T M_{it}(\theta)\,\Gamma_{it}^\infty + \frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T D_{it}\,E_{it}^2}_{=:Q_{NT}^{(\text{low},2)}},
\end{aligned}
$$

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 multiplied out $\left[(\Gamma_{it}^\infty - \Gamma_{it}) + E_{it}\right]^2$, 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 $\mathbf{P} = \frac{\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}\,\boldsymbol{w}'}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}$ and $\mathbf{R} = \mathbf{I}_{NT} - \mathbf{P}$. We then have

$$
\begin{aligned}
&Q_{NT}^{(\text{low},1)}(\mathbf{\Gamma}) \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\,\text{diag}(\boldsymbol{d})\,(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) - \frac{[(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\,\boldsymbol{w}]}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}\left[\boldsymbol{w}'\,\text{diag}(\boldsymbol{p})^{-1}\,\text{diag}(\boldsymbol{d} - \boldsymbol{p})\,(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)\right] \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\left(\mathbf{P}' + \mathbf{R}'\right)\text{diag}(\boldsymbol{d})\,(\mathbf{P} + \mathbf{R})\,(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) - (\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\mathbf{P}'\text{diag}(\boldsymbol{d} - \boldsymbol{p})\,(\mathbf{P} + \mathbf{R})\,(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\mathbf{R}'\text{diag}(\boldsymbol{d})\mathbf{R}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) + \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\mathbf{P}'\text{diag}\,(2\boldsymbol{p} - \boldsymbol{d})\,\mathbf{P}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty), \\
&= \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\mathbf{R}'\text{diag}(\boldsymbol{d})\mathbf{R}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) + \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\mathbf{P}'\text{diag}\,(\boldsymbol{p} - \boldsymbol{d})\,\mathbf{P}(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) + \frac{1}{2}\frac{[(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\boldsymbol{w}]^2}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}
\end{aligned}
$$

where all the "mixed terms" (that involve both $\mathbf{P}$ and $\mathbf{R}$) cancel because we have $\mathbf{P}'\,\text{diag}(\boldsymbol{p})\,\mathbf{R} = 0$, and in the last step we used that $\mathbf{P}'\,\text{diag}\,(\boldsymbol{p})\,\mathbf{P} = \frac{\boldsymbol{w}\,\boldsymbol{w}'}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}$. We have

$$
\min_{\{\mathbf{\Gamma} \in \mathbb{R}^{N\times T}\,:\,\theta = \boldsymbol{w}'\boldsymbol{\gamma}\}} (\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty)'\,\mathbf{R}'\,\text{diag}(\boldsymbol{d})\,\mathbf{R}\,(\boldsymbol{\gamma} - \mathbf{\Gamma}^\infty) = 0,
$$

because $\boldsymbol{\gamma}^* = \mathbf{R}\mathbf{\Gamma}^\infty + \theta\frac{\text{diag}(\boldsymbol{p})^{-1}w}{w'\text{diag}(\boldsymbol{p})^{-1}w}$ 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

$$\begin{aligned}
&\min_{\{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}\,:\,\theta=\boldsymbol{w}'\boldsymbol{\gamma}\}} Q_{NT}^{(\text{low},1)}(\boldsymbol{\Gamma}) \\
&= \frac{1}{2}\,(\theta-\theta_0)^2\left(\frac{1}{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}} + \frac{\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\text{diag}\,(\boldsymbol{p}-\boldsymbol{d})\,\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w}}{(\boldsymbol{w}'\text{diag}(\boldsymbol{p})^{-1}\boldsymbol{w})^2}\right) \\
&= \frac{1}{2}\,(\theta-\theta_0)^2\left(\frac{1}{\sum_{i=1}^N\sum_{t=1}^T W_{it}^2 P_{it}^{-1}} + \frac{\sum_{i=1}^N\sum_{t=1}^T W_{it}^2 P_{it}^{-2}(P_{it}-D_{it})}{(\sum_{i=1}^N\sum_{t=1}^T W_{it}^2 P_{it}^{-1})^2}\right) \\
&= \frac{NT}{2}\,c_1\,(\nu-\nu_0)^2,
\end{aligned}$$

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 (26) 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

$$\begin{aligned}
\frac{2\,[L_{NT}(NT\nu,\rho) - L_{NT}(NT\nu_0,\rho)]}{c_1\,NT} &\geq (\nu-\nu_0)^2 - \frac{2\,c_2}{c_1}(\nu-\nu_0) + \left(\frac{c_2}{c_1}\right)^2 - c_3 \\
&= \left(\nu - \nu_0 - \frac{c_2}{c_1}\right)^2 - c_3. \qquad (27)
\end{aligned}$$

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 (25).

The inequality in (27) 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 (27) 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 $Q_{NT}(\boldsymbol{\Gamma}, \rho)$. 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. ■