Estimating Counterfactual Matrix Means with Short Panel Data*

Lihua Lei† Brad Ross‡

May 7, 2024

Abstract

We develop a new, spectral approach for identifying and estimating average counterfactual outcomes under a low-rank factor model with short panel data and general outcome missingness patterns. Applications include event studies and studies of outcomes of "matches" between agents of two types, e.g. workers and firms, typically conducted under less-flexible Two-Way-Fixed-Effects (TWFE) models of outcomes. Given an infinite population of units and a finite number of outcomes, we show our approach identifies all counterfactual outcome means, including those not estimable by existing methods, if a particular graph constructed based on overlaps in observed outcomes between subpopulations is connected. Our analogous, computationally efficient estimation procedure yields consistent, asymptotically normal estimates of counterfactual outcome means under fixed-T (number of outcomes), large-N (sample size) asymptotics. In a semi-synthetic simulation study based on matched employer-employee data, our estimator has lower bias and only slightly higher variance than a TWFE-model-based estimator when estimating average log-wages.

Keywords: *panel data, missing not-at-random, factor model, interactive fixed effects, event study, bipartite network data*

1 Introduction

Researchers frequently seek to estimate average counterfactual outcomes in a population using "short" panel data with outcomes that are missing "not-at-random" (Little and Rubin, 2019; Rubin, 1976), namely, observations of a small subset of the possible outcomes for each unit in a

*First version: December 11, 2023. Authors are listed alphabetically. We thank Alberto Abadie, Dmitry Arkhangelsky, Susan Athey, Mohsen Bayati, Stéphane Bonhomme, Kirill Borusyak, Jiafeng Chen, Rebecca Diamond, Matthew Gentzkow, Bryan Graham, Christian Hansen, Guido Imbens, Patrick Kline, Elena Manresa, Samuel Norris, David Ritzwoller, Raffaele Saggio, Kevin Song, Jann Spiess, Vasilis Syrgkanis, Stefan Wager, and the participants of the Stanford Econometrics Lunch, Stanford GSB Eddie Lunch, Stanford Causal Panel Data Conference, UBC econometrics seminar, and SFU ecnonometrics seminar for their thoughtful comments and valuable feedback. Code to implement the methods and simulations described in this paper can be found at https://github.com/brad-ross/apm.

†Stanford Graduate School of Business and Department of Statistics (by courtesy); Email: lihualei@stanford.edu

‡Stanford Graduate School of Business; Email: bradross@stanford.edu

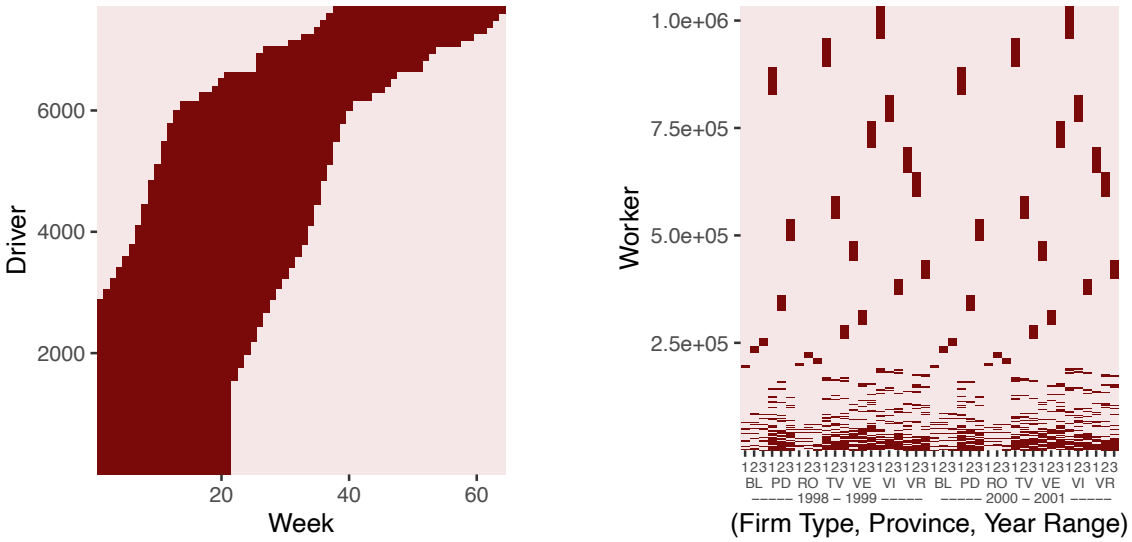
sample from that population without exogenous variation in which outcomes are observed for different units. For example, in event study settings, units receive some treatment at different times in a potentially non-random fashion, units' outcomes are observed over several time periods pre and post-treatment, and a researcher is interested in estimating average post-treatment control potential outcomes of treated units had they not been treated (Angrist and Pischke, 2009; Ashenfelter and Card, 1985; Bertrand, Duflo, and Mullainathan, 2004). In addition, several empirical literatures seek to estimate the average outcomes of counterfactual "matches" between pairs of agents of two different "types," e.g. wages of individuals when working at different firms (Abowd, Kramarz, and Margolis, 1999; Card, Heining, and Kline, 2013), test scores of students taught by different teachers (Jackson, Rockoff, and Staiger, 2014), and earnings and health outcomes of people living in different places (Card, Rothstein, and Yi, 2023; Chetty and Hendren, 2018; Finkelstein, Gentzkow, and Williams, 2016). Often in such settings, exogenous variation in which units are matched to which others is difficult to come by.

To estimate average counterfactual outcomes in these settings, researchers typically use multiple observations per unit to estimate a model with low-dimensional unobserved confounders that affect outcomes.[1] Perhaps the most canonical model of this sort is the Two-Way Fixed Effects (TWFE) model, which enables outcome means to be identified and estimated with short panel data under a myriad of outcome missingness patterns by "differencing out" unit fixed effects (Borusyak, Jaravel, and Spiess, 2024; Jochmans and Weidner, 2019). However it severely restricts how unobserved confounders can affect outcomes, as discussed in the literature on difference-in-differences methods and the "parallel trends" assumption implied by the TWFE model (Ghanem, Sant'Anna, and Wüthrich, 2022), as well as the literature on match outcomes (Bonhomme et al., 2019; Woodcock, 2015). In event study settings, a large literature has sought to allow for richer confounding than the TWFE model by using a low-rank factor model of outcomes (see Section 1.1 for references).[2] However, existing factor model-based methods cannot be applied generally, both because they only work under certain outcome missingness patterns, and because, unlike TWFE-based methods, many explicitly estimate unit-specific confounders, which requires a large number of observed outcomes per unit.

In this paper, we seek to bridge the gap between the general applicability of TWFE-based methods and the expressivity of factor-model-based methods. In particular, we develop an approach for identifying, estimating, and conducting valid inference on counterfactual outcome means under factor models in short panels with general outcome missingness patterns, including those not identified by existing methods.

[1] Usually, these models also require "strict exogeneity" (Chamberlain, 1984), namely that, conditional on low-dimensional confounders, outcomes are independent of missingness. In keeping with much of applied practice, this paper does the same. However, Ashenfelter and Card (1985) and Bonhomme, Lamadon, and Manresa (2019) discuss strict exogeneity's plausibility in event study and match outcome contexts, respectively.

[2] Factor models are also frequently called "interactive fixed effect" models since they assume outcomes are determined by the inner product of vectors of unit-specific and outcome-specific factors.



(a) Event Study With Staggered Entry

(b) Bipartite Match Outcomes

Figure 1: Each of these figures illustrates which outcomes, corresponding to columns, are observed for which units, corresponding to units, in two real-world panel datasets. A given coordinate is colored dark red if the outcome corresponding to the coordinate's column is observed for the unit corresponding to the coordinate's row and is colored light red otherwise. Figure 1a is generated using data from the longitudinal congestion pricing study studied in Ater et al. (n.d.). The coordinates in Figure 1a correspond to control potential outcomes. The outcomes from the light red upper triangular region are missing because the drivers have yet to join the study, while those from the light red lower triangular region are missing because the drivers have switched to the treatment arm after 20 weeks of monitoring. Figure 1b is constructed from the VWH Italian matched employer-employee dataset, where each coordinate corresponds to the worker's average weekly wage during a two-year range at a firm of a given type, as detailed in Section 5.1.

Figure 1 illustrates two empirical examples to which the methods developed in this paper uniquely apply on account of their complex outcome missingness patterns. As an example of an event study setting, Figure 1a displays the control potential outcome observation pattern from Ater, Shany, Ross, Turkel, and Vasserman (n.d.), which evaluates the effect of a government-run usage-based congestion pricing incentive on the driving behavior of several thousand Israeli drivers from January of 2020 through July of 2021. Unlike many event studies, here, drivers joined this study in a staggered fashion over time and received treatment after a non-random 20 weeks of monitoring. As an example of a bipartite matching setting, Figure 1b shows which of a subset of Italian workers in the Veneto Worker Histories (VWH) dataset received wages in each of seven provinces and three firm types within each province between 1998 and 2001; see Section 5.1 for details.

In both cases, it is plausible that the outcome missingness patterns are determined by potentially multiple unobserved characteristics of units that also affect outcomes. In Figure 1a's setting, it is possible that drivers who would benefit more from the treatment like remote workers joined the study earlier, while in Figure 1b's setting, workers likely live in the regions with job

opportunities that would pay them the highest wages. In addition, the TWFE model's restriction that unit-level heterogeneity does not affect how outcomes differ within units is plausibly unrealistic in both examples. In Figure 1a's setting, drivers in remote-work-compatible occupations plausibly changed their commuting patterns differently in response to COVID-19 lockdowns than drivers whose occupations required continued in-person work. In Figure 1b's setting, if industries are unequally distributed across space, workers whose skills are disproportionately valued by the industries in some region might receive higher wages there than if they lived in a region without those industries. Finally, in both examples, only a small number of outcomes are observed per unit relative to the number of units in each sample.

To identify our counterfactual outcome means in short panels with missing outcomes, we first group our population of units into subpopulations called *cohorts* that share the same sets of observed outcomes. We then use the observations from each cohort to identify collections of factor vectors corresponding to each cohort's observed outcomes up to cohort-specific bases. Our approach can accommodate any of the myriad of existing approaches for identifying factors in short panels without missing data to do so, e.g. those discussed in Section 4.3. To "align" these collections of factor vectors so that they are expressed with respect to a common basis, we aggregate these collections of cohort-specific factor vectors into a particular matrix we call an *Aggregated Projection Matrix (APM)*. Our main identification result shows that any basis for an APM's null space serves as an aligned collection of factor vectors corresponding to all outcomes so long as a particular graph is connected, where the graph's vertices correspond to cohorts, and an edge exists between two cohorts if there is sufficient overlap in the sets of observed outcomes for those two cohorts. Armed with aligned factor vectors corresponding to all outcomes, we then identify all outcome means for each cohort by learning linear relationships between the factors corresponding to the observed and missing outcomes in each target cohort.

Relative to existing methods, our approach has several desirable properties. First, it makes no assumptions about how units select into cohorts based on their low-dimensional unobserved confounders, even allowing observation patterns to be deterministic functions of unobserved confounders. Second, unlike factor-model-based approaches designed for long panels with many observed outcomes per unit, e.g. those in the "matrix completion" family (see Section 1.1 for references), our approach requires neither a known or estimable missingness mechanism nor a long panel that enables recovery of each unit's unobserved confounders exactly. As such, despite not being able to "difference out" unit-level unobserved confounders as one can under the TWFE model, our approach identifies cohort outcome means using only a finite number of observed outcomes per unit.

Third, our approach accommodates more general missingness patterns than other methods designed for short panels. In particular, it does not require the existence of a "reference" cohort of units for whom both the target outcome is observed *and* a sufficient number of observed

outcomes overlap with the target cohort's observed outcomes, as illustrated in Figure 2; such a pattern of observed and missing outcomes is often called a "block" missingness pattern (Athey, Bayati, Doudchenko, Imbens, and Khosravi, 2021). Both of the empirical examples illustrated in Figure 1 contain many cohorts and outcomes for which no reference cohort exists, as quantified in Section 5. Fourth, our method automatically stitches together different pieces of identifying information each used in isolation by existing methods (e.g. only using data from a cohort for whom all outcomes are observed to identify and estimate factors), improving sample efficiency.

We also translate this identification strategy into a plug-in estimator. In particular, we compute an estimated APM using estimates of cohort-specific factors constructed using the data corresponding to each cohort. We then use the rows of the matrix of eigenvectors corresponding to the smallest eigenvalues of the estimated APM as estimates of the factor vectors corresponding to all outcomes. As such, our estimator is simple to compute. In an asymptotic regime in which the number of outcomes remains fixed as the cross-sectional dimension of the panel and the sizes of all cohorts grow, we show that this estimator is consistent and asymptotically normal, and that a weighted bootstrap procedure provides valid asymptotic inference. These results rely on an exact, first-order expansion of the operator mapping a symmetric matrix into the projection matrix onto the space spanned by some subset of its eigenvectors. We derive this expansion using a result called Kato's integral (Kato, 1949).

Finally, we demonstrate the empirical performance of our estimator via a semi-synthetic simulation study based on the VWH dataset of wages earned by workers at different types of firms in the Veneto region of Italy. To define outcomes, we cluster firms within each province into three types based on their weekly wage distributions as in Bonhomme et al. (2019) and define a worker's outcome corresponding to a given type of firm in a given province and year range as the average weekly wage they would earn were they to work for that type of firm in that province during that year range. Importantly, the outcome missingness pattern for this setting, which we illustrate in Figure 1b, is complex enough to preclude most counterfactual outcome mean estimation using existing factor-model-based methods.

To assess the accuracy of our factor-model-based procedure relative to an estimator based on a TWFE model of counterfactual outcomes, we mask an observed outcome for some cohort of units in our data and resample units from these masked data to construct synthetic datasets. We then compute various error metrics of both estimators of the artificially hidden cohort outcome's mean across resampled synthetic datasets. Across many masked cohort outcomes, our procedure frequently delivers outcome mean estimates with lower bias and root mean squared error than the TWFE-based estimator. In line with Bonhomme et al. (2019), our simulation results suggest that complementarities between workers and firms do affect wages.

1.1 Related Work

Recently, a variety of factor-model-based methods have been proposed that can be used to estimate and conduct inference on a target cohort outcome mean by aggregating accurate imputations of individual-level factor structure estimates, e.g. Abadie, Agarwal, Dwivedi, and Shah (2024); Agarwal, Dahleh, Shah, and Shen (2023); Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021); Arkhangelsky and Hirshberg (2023); Athey et al. (2021); J. Bai (2009); J. Bai and Ng (2021); Ben-Michael, Feller, and Rothstein (2022); Chernozhukov, Hansen, Liao, and Zhu (2023); Choi, Kwon, and Liao (2023); Choi and Yuan (2023); Farias, Li, and Peng (2021); Fernández-Val, Freeman, and Weidner (2021); Freeman and Weidner (2023); Gobillon and Magnac (2016); Imbens and Viviano (2023); Moon and Weidner (2015); Xiong and Pelger (2023); Xu (2017); Yan and Wainwright (2024). These methods require the number of outcomes and maximum number of observed outcomes per cohort to grow as the number of sampled units grows, since outcome-specific factor vectors can be estimated consistently using variation across many units and unit-specific loading vectors can be estimated consistently using variation across many outcomes. However, under our asymptotic regime of interest, these methods cannot recover each unit's loadings without bias that persists as the number of units grows and does not average out to zero across units in the population, precluding averages across imputations. Further, this bias can contaminate the outcome-specific factor estimates themselves (T. Lancaster, 2000; Neyman and Scott, 1948). In addition, when the missingness pattern is complicated, a restriction on the missingness mechanism, such as missingness being at random, or knowledge of it, such as knowing the probability of jointly observing each pair of outcomes, is often needed even in long panels (Xiong and Pelger, 2023). However, this information is typically unavailable or hard to come by in our settings of interest.

Instead, we build on an approach suggested in various forms in Imbens, Kallus, and Mao (2021), Brown and Butts (2022), and Agarwal, Shah, and Shen (2023): given just factor vectors corresponding to all outcomes, we can construct a linear combination of the target cohort's observed outcome means that equals the target cohort outcome mean. Imbens et al. (2021) call any such linear combination a bridge function to highlight the connections they make between this setting and the proximal causal inference literature (Deaner, 2018; Miao, Geng, and Tchetgen Tchetgen, 2018). The bridge function-based identification strategy was originally developed in the context of settings with a block missingness pattern, as illustrated in Figure 2a. In these settings, there exists a "reference" cohort of units for whom all outcomes are observed, so any method for identifying and estimating a factor model without missing data can be applied to the data from the reference cohort to recover the factor vectors for all outcomes with respect to the same basis; we discuss several such methods in Section 4.3.

The bridge function approach can still be applied to identify some cohort outcome means under even more general missingness patterns so long as reference cohorts exist. In event study

settings with staggered treatment adoption, so long as there is a "never-treated" group of units for whom *all* control potential outcomes are observed,[3] these units can be used as a reference cohort, and the approaches developed in Callaway and Karami (2023), Brown and Butts (2022), Brown, Butts, and Westerlund (2023), Callaway and Tsyawo (2023), and Arkhangelsky and Samkov (2024) identify and yield consistent estimates of outcome means when the number of outcomes remains fixed as the cross section's size grows. Agarwal, Dahleh, et al. (2023) show that the bridge function approach can be applied to identify some cohort outcome means in more general settings like the one illustrated in Figure 2b by finding block missingness patterns embedded within the broader outcome missingness pattern.

However, several important challenges remain in estimating and conducting valid inference on cohort outcome means under factor models that are common in our empirical settings of interest. First, many embedded block missingness patterns may exist within a broader missingness pattern, and it is unclear how one should combine information gleaned from each of them in a computationally efficient manner to improve estimate precision. Second and more importantly, embedded block missingness patterns cannot always be found to identify all cohort outcome means. Our identification and estimation approach automatically combines information about factor vectors from all cohorts to identify and estimate cohort outcome means without requiring embedded block missingness patterns to exist.

The graph connectivity assumption underlying our identification argument bears resemblance to those made in several papers in the rich literature on fixed-effect-like models of bipartite match outcomes under strict exogeneity, e.g. Abowd, Creecy, and Kramarz (2002); Bonhomme et al. (2019); Hull (2018); Jochmans and Weidner (2019); see Bonhomme (2020) for a review. We show in Appendix F.1 that, while our connectivity requirement enables identification for general numbers of factors and loadings, in the special case where the factors and loadings are unidimensional, our assumption is essentially equivalent to theirs.

Our consistency and asymptotic linearity results are also related to a large literature applying perturbation-theoretic results to characterize the concentration of eigenspaces of random matrices.[4] To characterize the asymptotic distributions of eigenspaces of estimated matrices as we do in this paper, the concentration guarantees implied by canonical zero-th-order approximation error bounds like the Davis–Kahan theorem (see Yu et al. (2015) for a convenient version) are too coarse. In our paper, we instead characterize the asymptotic distributions of estimated eigenspaces and the estimators on which they depend by deriving a non-asymptotic, first-order eigenspace projection operator expansion based on Kato's integral (Kato, 1949). Several other papers have also applied Kato's integral to derive concentration results for eigenspaces of random

[3] Athey et al. (2021) also require a nontrivial share of never-treated units for their estimator to be consistent in the staggered treatment adoption setting.

[4] See Yu, Wang, and Samworth (2015) for clear and concise statements of several results from the literature useful for statistical applications, and Z. Bai and Silverstein (2010) for a textbook treatment.

matrices with (approximately) independent entries, e.g. Lei (2020); Oliveira (2010). Relatedly, Simons (2023) applies an asymptotic linearization of the eigenspace operator from J.-g. Sun (1991) to derive asymptotically valid hypothesis tests concerning eigenspaces of an estimated, potentially non-symmetric matrix.

2 Setup and Intuition

2.1 Setup

We describe our setting of interest formally as follows. Researchers observe a large, i.i.d. sample of N units, and each unit has T outcomes associated with them, where Y_{it}^* denotes unit i's outcome t. Importantly, not all outcomes are observed for each unit. To describe which outcomes are observed, we group units into C *cohorts*, where C_i denotes unit i's cohort. For the units in cohort c, we only observe the outcomes with indices t in a subset of all outcome indices $\mathcal{T}_c \subseteq \{1, \ldots, T\}$. To distinguish between observed and missing outcomes, we define $Y_{it} = Y_{it}^*$ if outcome t is observed for unit i (i.e. if $t \in \mathcal{T}_{C_i}$) and $Y_{it} = \emptyset$ otherwise.[5] Given these unbalanced panel data, researchers are often interested in estimating aggregations of cohort-level outcome means:

$$\mu_{ct} \coloneqq \mathbb{E}[Y_{it}^* \mid C_i = c].$$

We refer to μ_{ct} as a counterfactual outcome mean since outcome t may not be observed for the units in cohort c. To model the small number of observed outcomes per unit in our asymptotic theory in Section 4.2, we will focus on the case where the number of outcomes T remains bounded as the sample size N grows.

The setup described above encapsulates several kinds of causal panel data analyses. In event study settings like Figure 1a's, Y_{it}^* is unit i's control potential outcome had they not yet been treated by period t, units belong to cohort c if they first received treatment at time c, and researchers typically estimate aggregations of average treatment effects on the units in each cohort like dynamic effects across post-treatment periods.[6] In bipartite matching settings like Figure 1b's, Y_{it}^* is "row-type" unit i's outcome if matched with "column-type" unit t, row-type units belong to the same cohort if they were matched with the same set of column-type units, and researchers typically estimate aggregations of μ_{ct}s that characterize the degree to which row-type and column-type unit heterogeneity affect differences in average observed outcomes across column-type units.[7]

[5] Throughout the paper, we define $\emptyset$ such that $0 \cdot \emptyset = 0$.

[6] See L. Sun and Abraham (2021) and Callaway and Sant'Anna (2021) for more in-depth discussions of this model of event study settings with staggered treatment adoption, and De Chaisemartin and d'Haultfoeuille (2020) for treatment effect definitions relevant to settings in which treatment is not an absorbing state.

[7] In Appendix E, we discuss a variant of the decomposition proposed in Finkelstein et al. (2016) for this purpose that only requires estimates of the outcome means identified in this paper. We leave identifying the higher-order

We assume outcomes Y_{it}^* are generated according to a rank-r factor model, i.e. that outcomes are determined by the inner product of a fixed vector of outcome-specific factors $\gamma_t \in \mathbb{R}^r$ and a vector of unit-specific loadings $\lambda_i \in \mathbb{R}^r$ plus an error term ε_{it} that has zero mean given unit i's loadings and cohort membership C_i:

$$Y_{it}^* = \gamma_t' \lambda_i + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon_{it} \mid \lambda_i, C_i] = 0. \tag{1}$$

As in other fixed-effect-like models like the TWFE model, the assumption that ε_{it} is mean-independent of cohort membership C_i conditional on loadings λ_i implies that only the coordinates of λ_i can serve as unobserved confounders that affect both units' cohort memberships and their outcomes. For example, in Figure 1b's setting in which i indexes workers, t indexes groups of firms within provinces of Italy in different years, and Y_{it}^* denotes the logarithm of i's average weekly wage during their time working for firms in group t, the coordinates of λ_i could correspond to workers' multidimensional skills, and the corresponding coordinates γ_t could measure the degree to which firms in group t value those skills, akin to Lazear (2009). Insofar as workers tend to work at firms that value their skills more, workers' skills λ_i will determine both which firms they work at as well as the wages they earn at those firms (Lazear, 2009).

By considering the factor vectors γ_t to be fixed, we essentially condition our inference on the factors $\{\gamma_t : t = 1, \ldots, T\}$. If we consider the factors γ_t to be random, the residual mean condition in (1) can be read as $\mathbb{E}[\varepsilon_{it} \mid \lambda_i, C_i, \gamma_1, \ldots, \gamma_T] = 0$. Throughout this paper, we assume the rank of the factor model $r < T$ is known.[8] We suggest that researchers using our method assess the robustness of their results to different values of r, insofar as the cohort outcome means they would like to estimate are identified given that choice of r, as discussed in Section 4.1.[9]

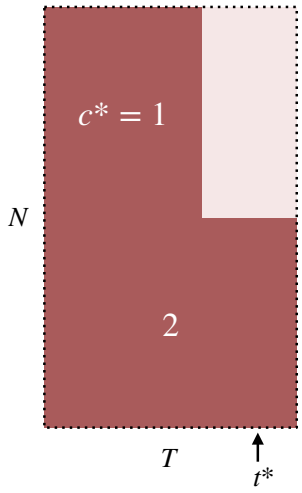
2.2 Intuition

Broadly, our approach for identifying and estimating a target cohort outcome mean $\mu_{c^*t^*}$ proceeds in two stages. First, we recover all of the factor vectors across outcomes with respect to a common basis. Then, we use the factor vectors corresponding to the target cohort's observed outcomes and the target outcome to extrapolate from the target cohort's observed outcome means to the target outcome mean. To highlight the additional challenge posed by the first stage, we proceed by first providing intuition for the second stage in a setting where the first stage happens to be

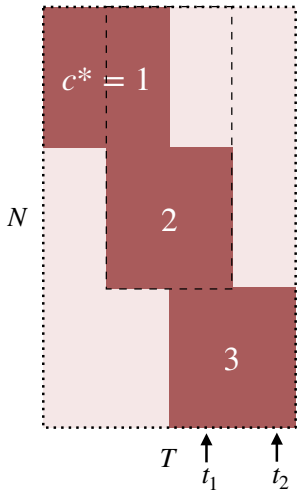
outcome moments required by outcome variance decompositions in the literature (e.g. Abowd et al. (1999)) to future work.

[8] Papers that use empirical strategies based on TWFE models or parallel trends assumptions are also making a choice about the dimension of unobserved confounders.

[9] The problem of determining the number of factors r from short panel data without assuming a the distributions of residuals ε_{it} belong to a known parametric family has only recently been studied to our knowledge (Ahn, Lee, and Schmidt, 2013; Fortin, Gagliardini, and Scaillet, 2022, 2023), unlike the well-established literature on determining r in panels for which both N and T are large (Ahn and Horenstein, 2013; J. Bai and Ng, 2002; Gagliardini, Ossola, and Scaillet, 2019; Onatski, 2009).



(a) Block Missingness Pattern



(b) Three-Cohort Staircase Pattern

Figure 2: This figure illustrates two stylized outcome missingness patterns, where, similarly to Figure 1, a coordinate is colored dark red if the outcome corresponding to the coordinate's column is observed for the unit corresponding to the coordinate's and is colored light red otherwise. We use numbers to indicate the cohorts of units for whom different blocks of outcomes are observed. In Figure 2a, cohort 2 serves as a "reference" cohort, since for those units, target outcome t^* and outcomes overlapping with the target cohort $c^* = 1$'s observed outcomes are observed. In Figure 2b, when the target cohort is $c^* = 1$ and the target outcome is $t^* = t_2$, then no such reference cohort exists.

straightforward.

In particular, we focus on a simple, three-cohort outcome missingness pattern illustrated in Figure 2b and consider estimating a target cohort outcome mean $\mu_{c^*t^*}$, where $c^* = 1$ is the target cohort and $t^* = t_1$ is the target outcome. Suppose also that we have access to the factors γ_t for all outcomes. Then, under regularity conditions, there exist many vectors of coefficients β we can construct such that a linear combination of the factor vectors for the target cohort c^*'s observed outcomes with those coefficients equals the factor vector γ_{t^*} corresponding to the target outcome t^*:

$$\sum_{t \in \mathcal{T}_{c^*}} \beta_t \gamma_t = \gamma_{t^*}. \tag{2}$$

As it happens, the same linear combination of the target cohort's observed outcome means $\mathbb{E}[Y_{it} \mid C_i = c^*]$ for $t \in \mathcal{T}_{c^*}$ will equal the target cohort's target outcome mean (Agarwal, Shah, and Shen, 2023; Brown and Butts, 2022; Imbens et al., 2021):

$$\sum_{t \in \mathcal{T}_{c^*}} \beta_t \underbrace{\mathbb{E}[Y_{it} \mid C_i = c^*]}_{= \mathbb{E}[Y^*_{it} \mid C_i = c^*]} \overbrace{=}^{\text{by (1)}} \sum_{t \in \mathcal{T}_{c^*}} \beta_t \gamma_t' \, \mathbb{E}[\lambda_i \mid C_i = c^*] \overbrace{=}^{\text{by (2)}} \gamma_{t^*}' \, \mathbb{E}[\lambda_i \mid C_i = c^*] \overbrace{=}^{\text{by (1)}} \mu_{c^*t^*}.$$

In keeping with Imbens et al. (2021), we refer to any linear transformation β' where β satisfies (2) as a *bridge function*.

Turning to the first stage of our approach, we now discuss how to recover the factors γ_t used to construct bridge functions β' via solutions to (2). A well known fact in the rich literature on factor models is that, even in settings where all outcomes are observed, only a common linear transformation $Q\gamma_t$ of each factor vector γ_t can be identified, where Q is an *unknown*, $r \times r$ invertible matrix we refer to as a basis.[10]

Luckily, the bridge function approach renders this basis indeterminacy immaterial when our target cohort is $c^* = 1$ and our target outcome is $t^* = t_1$. To see why, note that in this case, there exists a "reference" cohort 2 of units for whom we observe *both* the target outcome t_1 and some of the outcomes also observed for the units in the target cohort (Agarwal, Dahleh, et al., 2023; Brown and Butts, 2022). In other words, the set of units and outcomes inside the black dashed rectangle in Figure 2b constitutes a block missingness pattern like the one shown in Figure 2a embedded within the broader outcome missingness pattern in Figure 2b. Then, under a variety of additional assumptions discussed in Section 4.3, a myriad of methods can still be applied using just the data from cohort 2 to recover transformed factor vectors $Q\gamma_{t_1}$ corresponding to the target outcome and $Q\gamma_t$ for $t \in \mathcal{T}_1 \cap \mathcal{T}_2$ corresponding to the overlapping observed outcomes between cohorts 1 and 2, where Q is a common basis. Since the set of valid bridge functions β' satisfying (2) is invariant to multiplying all of the factor vectors γ_t by the *same* basis matrix Q, the second stage of our identification approach remains unaffected.

Unfortunately, when our target outcome is instead $t^* = t_2$, no reference cohort exists for whom we observe both the target outcome t_2 and any observed outcomes for our target cohort $c^* = 1$. As such, there is no subset of the data we can directly use to recover the factors corresponding to the target outcome and the target cohort's observed outcomes with respect to the same basis.[11] Instead, again under different sets of additional assumptions (see Section 4.3), we can only use the observed outcomes of units in each cohort c to recover transformed factor vectors $\tilde{\gamma}_{ct} \coloneqq Q_c\gamma_t$ corresponding to cohort c's observed outcomes $t \in \mathcal{T}_c$ with respect to a *cohort-specific* basis Q_c.[12] In other words, at best, we can recover cohort-specific factor vectors that are "misaligned." To be able to find bridge functions that satisfy (2), we need to construct an aligned set of factor vectors expressed with respect to a common basis that correspond to both the target outcome t_2 and the target cohort's observed outcomes $t \in \mathcal{T}_1$.

To describe our solution to this factor vector misalignment problem, we define some additional notation. First, stack the factor vectors γ_t row-wise into a $T \times r$ matrix Γ, and stack the cohort-specific transformed factor vectors $\tilde{\gamma}_{ct}$ recovered from the data for cohort c into the $T \times r$ matrix $\tilde{\Gamma}_c$, where the t-th row of $\tilde{\Gamma}_c$ equals $\tilde{\gamma}_{ct} = Q_c\gamma_t$ if outcome t is observed for the units in cohort c and

[10] This basis indeterminacy is inevitable because, for any invertible Q, Y_{it}^* and $\gamma_t' Q Q^{-1}\lambda_i + \varepsilon_{it}$ have the same distribution; see e.g. Anderson (2009); Anderson and Rubin (1956).

[11] The lack of a reference cohort also precludes the imposition of a common normalization that ensures the factor vectors are uniquely defined; see J. Bai and Ng (2013) for a detailed discussion of such normalizations.

[12] As will be made clear in Section 4, our approach will not actually require the identification of $\tilde{\Gamma}_c = \Gamma Q_c'$ for some fixed basis matrices Q_c; we introduce them here for simplicity of exposition.

a vector of zeros otherwise. In addition, let E_c be the $T \times T$ diagonal matrix whose t-th diagonal entry is one if outcome t is observed for cohort c, i.e. $t \in \mathcal{T}_c$, and zero otherwise. Finally, for any matrix M, we denote the projection matrix onto M's column space by $\Pi(M) \coloneqq M(M'M)^{+}M'$, where $(M'M)^{+}$ is the Moore-Penrose pseudo-inverse of $M'M$.

Armed with this notation, we can now describe the two insights that underpin our solution to the factor vector alignment problem introduced above. First, since the column space of $\tilde{\Gamma}_c = E_c \Gamma Q_c'$ does not depend on Q_c, the projection matrix $\Pi(\tilde{\Gamma}_c)$ is the same as the projection matrix $\Pi(E_c\Gamma)$ onto the column space of the matrix $E_c\Gamma$ whose non-zero rows are exactly the true factor vectors corresponding to cohort c's observed outcomes.[13] Thus, the projection matrix $\Pi(\tilde{\Gamma}_c)$ derived from cohort c's data provides a unique representation of the available information about the factor vectors corresponding to cohort c's observed outcomes free from contamination by the cohort-specific Q_c.

Given the cohort-specific projection matrices $\Pi(\tilde{\Gamma}_c)$, we then must aggregate them in such a way so as to recover the factor vectors γ_t defined with respect to some common basis. Along these lines, our second key observation is that the column space of Γ is a subset of the null space of the matrix $E_c - \Pi(\tilde{\Gamma}_c) = E_c - \Pi(E_c\Gamma)$ for each cohort c.[14] To leverage this insight, we define the *Aggregated Projection Matrix* (APM) operator as follows for any C matrices $\Gamma_1, \ldots, \Gamma_C \in \mathbb{R}^{T \times r}$:

$$A(\Gamma_1, \ldots, \Gamma_C) \coloneqq \sum_{c=1}^{C} \left(E_c - \Pi(\Gamma_c)\right). \tag{3}$$

Despite its name, $A(\Gamma_1, \ldots, \Gamma_C)$ is not itself a projection matrix in general. However, since $A(\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_c)$ is by definition the sum of the matrices $E_c - \Pi(\tilde{\Gamma}_c)$ across cohorts c, the column space of Γ must lie in the null space of $A(\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_c)$ as well. In Section 4.1, we show that, perhaps surprisingly, so long as there is sufficient overlap between the observed outcomes across pairs of cohorts, the null space of $A(\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_c)$ is in fact *exactly* the column space of Γ. Thus, the rows of any basis matrix for the null space of A can serve as valid factor vectors when applying the bridge function identification strategy.[15]

[13] By definition, we have $\Pi(\tilde{\Gamma}_c) = E_c\Gamma Q_c'(Q_c\Gamma' E_c \Gamma Q_c)^{+} Q_c \Gamma' E_c = E_c \Gamma(\Gamma' E_c' \cdot E_c \Gamma)^{-1}\Gamma' E_c = \Pi(E_c\Gamma)$.

[14] $\Pi(E_c\Gamma) = E_c\Pi(E_c\Gamma) = \Pi(E_c\Gamma)E_c$ and $E_c^2 = E_c$, so $(E_c - \Pi(E_c\Gamma))\Gamma = E_c(I - \Pi(E_c\Gamma))E_c\Gamma = \mathbf{0}_{T\times r}$.

[15] One might imagine instead explicitly recovering the matrix $Q_{c_1}Q_{c_2}^{-1}$ that aligns cohort c_2's factors with cohort c_1's by regressing cohort c_1's factor vectors $\tilde{\gamma}_{c_1 t}$ on cohort c_2's factor vectors $\tilde{\gamma}_{c_2 t}$ corresponding to the overlapping observed outcomes $t \in \mathcal{T}_{c_1} \cap \mathcal{T}_{c_2}$ between the two cohorts as in J. Bai and Ng (2021). One could then chain multiplications of these "pairwise aligning" matrices together to align cohort-specific factor vectors with each other, evocative of the identification and estimation strategy proposed in Hull (2018) under a TWFE-like model of outcomes. However, constructing these aligning chains is nontrivial in realistic applications with more than a few cohorts like those illustrated in Figure 1, and it is not clear how one would aggregate information about factors across potentially large numbers of possible aligning chains. In Section 4.1, we discuss how the procedure we propose next makes use of information from all available aligning chains to identify the column space of Γ without needing to enumerate and weight them explicitly.

3 Estimation and Inference Procedure

Having provided intuition for the conceptual underpinnings of our approach, we now introduce our general estimation and inference procedure, which has four high-level steps. First, we consider the data corresponding to units in each cohort separately and use them to estimate the factors corresponding to the observed outcomes for each cohort. Second, we aggregate these cohort-specific factor estimates into an estimated APM whose eigendecomposition yields valid estimates of the factors corresponding to all outcomes. Third, given those factor estimates, for each cohort, we construct the minimum norm bridge function that consistently extrapolates from that cohort's observed outcome means to all target outcome means. Finally, we conduct valid inference on cohort outcome means by taking advantage of the asymptotic normality of our estimator. Code to implement our procedure can be found at `https://github.com/brad-ross/apm`.

We now describe our procedure in more detail, beginning with step two outlined above. We assume for now that for each cohort c, we have access to estimates $\hat{\gamma}_{ct}$ of the factor vectors corresponding to each outcome $t \in \mathcal{T}_c$ observed for cohort c. In Section 4.3, we discuss a variety of methods that exist for constructing valid cohort-specific factor vector estimates under different sets of additional assumptions beyond the factor model (1). Importantly, most estimators cohort-specific factor vectors $\tilde{\gamma}_{ct}$ require minimal computational overhead for small T, and they can be computed in parallel across cohorts using subsets of the data associated with each cohort. We then stack these cohort-specific estimates into the $T \times r$ matrix $\hat{\Gamma}_c$ whose t-th row is $\hat{\gamma}_{ct}$ if $t \in \mathcal{T}_c$, i.e. if outcome t is observed for the units in cohort c and a vector of zeros otherwise.

Given cohort-specific estimated factor matrices $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_C$, the second step of our procedure consists of constructing the estimated APM $\hat{A} \coloneqq A(\hat{\Gamma}_1, \ldots, \hat{\Gamma}_C)$ and computing its eigendecomposition. Because T is typically small in settings in which we envision our method will be applied, this step can be done extremely quickly using the myriad of optimized eigendecomposition routines available in most programming languages.[16] Let $\hat{\Gamma}$ be the $T \times r$ matrix whose columns are eigenvectors corresponding to the r smallest eigenvalues of $\hat{A}$;[17] as we show in Section 4.2, row t of $\hat{\Gamma}$ is a consistent estimate of the factor vector γ_t with respect to a particular error metric.

Describing the third step of our procedure requires several more definitions. Collect unit i's observed outcomes into the vector $Y_i \coloneqq (Y_{i1}, \ldots, Y_{iT})'$, and note that although Y_i has undefined entries corresponding to unobserved outcomes, all entries of the vector $E_{C_i} Y_i$ are well-defined since the entries corresponding to the unobserved outcomes are zero. We also let $N_c \coloneqq \sum_{i=1}^{N} \mathbb{1}\{C_i = c\}$ denote the number of units in cohort c. We can then define our estimator of the vector $\mu_{c^*} \in \mathbb{R}^T$

[16] Examples include the `eigvecs` function in `Julia`'s `LinearAlgebra` module and the `eigen` function available in base `R`.

[17] Recall that eigenvectors are unique up to magnitudes, signs, and permutations between indices corresponding to repeated eigenvalues.

whose t-th entry is the target cohort c^*'s mean value of outcome t, μ_{c^*t}:

$$\hat{\mu}_{c^*} \coloneqq \hat{\Gamma}\left(E_{c^*}\hat{\Gamma}\right)^+ \cdot \frac{1}{N_{c^*}} \sum_{i=1}^{N} \mathbb{1}\{C_i = c^*\} E_{C_i} Y_i. \tag{4}$$

Given that $\hat{\mu}_{c^*}$ is simply a vector of least-norm solutions to underdetermined linear equations, it can also be computed efficiently in most programming languages.[18]

Often, researchers are actually interested in estimating a known, vector-valued function $\theta \coloneqq h(\mu, \eta)$ of cohort outcome means $\mu \coloneqq (\mu_1', \ldots, \mu_C')'$, along with a vector of nuisance parameters η that are estimable from the data. For example, in event study settings, it is common to report average effects of a treatment across different numbers of time periods relative to units' treatment times. As introduced in Callaway and Sant'Anna (2021) and L. Sun and Abraham (2021) and reviewed in Example E.1 of Appendix E, we can express these estimands as weighted averages of the differences between each cohort's average control potential outcome means μ_c and their observed, treated potential outcome means (part of η), where the weights are determined by the relative sizes of each cohort (also part of η).

In the context of bipartite match data on patients' health outcome when living in different geographic areas, Finkelstein et al. (2016) suggest an approach to attributing shares of the differences in average health outcomes across regions to people and places that can be expressed in the form $h(\mu, \eta)$, albeit based on a TWFE model of match outcomes. In Example E.2 of Appendix E, we discuss nonparametric analogs of their estimands that do not depend on a particular model of outcomes but can be estimated under the factor model (1) using our procedure.

Given θ's relevance in applied contexts, as an extension of the third step of our procedure, we compute a plug-in estimate $\hat{\theta} \coloneqq h(\hat{\mu}, \hat{\eta})$ of θ, where the vector $\hat{\mu} \coloneqq (\hat{\mu}_1', \ldots, \hat{\mu}_C')'$ collects the cohort outcome mean vector estimates $\hat{\mu}_c$ defined in (4) across cohorts, and $\hat{\eta}$ is an estimate of the nuisance parameter vector η. For convenience, we summarize the three steps required to implement our estimator in Algorithm 1.

For the fourth and final step of our procedure, we construct simultaneous $1-\alpha$ confidence intervals for the p coordinates of θ based on a Bayesian bootstrap (Rubin, 1981).[19] Since our Bayesian bootstrap algorithm is similar to other simultaneous inference procedures (see Chernozhukov, Fernández-Val, and Melly (2013), for example), we defer a detailed description to Appendix E.2 for brevity.

[18] Given that T is small, the simplest method is to compute the singular value decomposition $E_{c^*}\hat{\Gamma} = USV'$, e.g. using the `svd` function available in `Julia`'s `LinearAlgebra` module or base `R`, and then compute $\hat{\mu}_{c^*} = \hat{\Gamma} V S^+ U' \frac{1}{N_{c^*}} \sum_{i=1}^{N} \mathbb{1}\{C_i = c^*\} E_{C_i} Y_i$, where S^+ is simply the diagonal matrix S with its positive diagonal entries replaced by their inverses.

[19] A set of confidence intervals for the coordinates of θ has simultaneous coverage $1-\alpha$ if the probability that all coordinates of θ lie inside their respective intervals simultaneously is at least $1-\alpha$. In principle, any bootstrap procedure that weights or resamples units could also be used, but to avoid pathological cases where no units from some cohort are sampled, we prefer weighted bootstrap procedures.

Algorithm 1: Estimation

Data: $\{(C_i, Y_i)\}_{i=1}^N$

for $c \in \{1, \ldots, C\}$ **do**
  Compute cohort-specific factor estimates $\hat{\gamma}_{ct}$ for $t \in \mathcal{T}_c$ (see Section 4.3)
  Construct a $T \times r$ matrix $\hat{\Gamma}_c$ with row t equalling $\hat{\gamma}_{ct}$ if $t \in \mathcal{T}_c$, and $\mathbf{0}_r$ otherwise
end
Construct estimated APM $\hat{A} = P(\hat{\Gamma}_1, \ldots, \hat{\Gamma}_C)$ as in (3)
Compute $T \times r$ matrix $\hat{\Gamma}$ of eigenvectors corresponding to $\hat{A}$'s r smallest eigenvalues
for $c \in \{1, \ldots, C\}$ **do**
  Compute outcome mean estimate vector $\hat{\mu}_c$ for cohort c as in (4)
end
Compute estimate $\hat{\eta}$ of nuisance parameters η necessary for estimating θ
Compute plug-in target parameter estimate $\hat{\theta} = h(\hat{\mu}, \hat{\eta})$, where $\hat{\mu} = (\hat{\mu}_1', \ldots, \hat{\mu}_C')'$

4 Theoretical Properties

4.1 Identification Given Cohort-Specific Factors

To highlight the contributions of this paper, we assume for now that we have identified the column spaces of the cohort-specific factor matrices $E_c\Gamma$:

Assumption 1. The projection matrices $\Pi(E_c\Gamma)$ are identified for all cohorts $c = 1, \ldots, C$.

In Section 4.3, we will discuss several sets of additional assumptions standard in the rich literature on factor models under which Assumption 1 holds.

To characterize which outcomes our approach can and cannot identify given identified cohort specific factor projection matrices $\Pi(E_c\Gamma)$, we define an object we call the *Observed Outcome Overlap Graph* $\mathcal{G}_r$. The graph $\mathcal{G}_r$ consists of C vertices, one corresponding to each cohort, and an edge between two distinct cohorts c_1 and c_2 if the space spanned by the factor vectors corresponding to the two cohorts' overlapping observed outcomes $\mathcal{T}_{c_1} \cap \mathcal{T}_{c_2}$ has at least r dimensions, or formally, if

$$\operatorname{rank}\left(E_{c_1} E_{c_2} \Gamma\right) = r. \tag{5}$$

While (5) is a requirement on the unobserved factor matrix Γ, we note that a necessary condition that only depends on the known sets of observed outcomes $\mathcal{T}_c$ across cohorts is that cohorts c_1 and c_2 have at least r overlapping observed outcomes: $|\mathcal{T}_{c_1} \cap \mathcal{T}_{c_2}| \geq r$. Such a condition becomes sufficient when we also impose a general position-type requirement on the collection of factor vectors, namely that the members of every size-r subset of factor vectors are linearly independent; this general position requirement becomes vacuously true when $r = 1$. In Figure G.1, we illustrate $\mathcal{G}_1$ for the two empirical examples introduced in Section 1.

Our key requirement to identify all of the outcome means for a given cohort can be stated succinctly in terms of the connectedness of $\mathcal{G}_r$:

Assumption 2. The observed outcome overlap graph $\mathcal{G}_r$ is connected.

Before continuing, three remarks concerning Assumption 2 are in order. First, Assumption 2 implies that every cohort must have at least r observed outcomes, and that every factor must affect at least one outcome in every cohort (a proof is provided in Appendix A.1):

Lemma 1. *Under Assumption 2,* $\text{rank}(E_c\Gamma) = r$ *for all cohorts* $c = 1, \ldots, C$.

A consequence of Lemma 1 is that our identification, estimation, and inference results hold for the subset of units for whom at least r outcomes are observed, i.e. conditional on $|\mathcal{T}_{C_i}| \geq r$.[20]

Second, if $\mathcal{G}_r$ is not connected, our identification, estimation, and inference results apply instead to the subset of cohorts belonging to the connected component of $\mathcal{G}_r$ that contains the target cohort c^* and the subset of outcomes observed for at least one of those cohorts. Third, we show in Appendix F.1 that when $r = 1$, Assumption 2 is equivalent to typical connectedness assumptions made in the literature on fixed-effect-type models of bipartite network data under strict exogeneity (see Abowd et al. (2002); Bonhomme et al. (2019); Hull (2018); Jochmans and Weidner (2019) for examples, and Bonhomme (2020) for a review).

Given identification of the cohort-specific factor projection matrices $\Pi(E_c\Gamma)$, we denote the population APM by substituting the population cohort-specific factor matrices $E_c\Gamma$ into the APM definition (3) as follows:

$$A \coloneqq A(E_1\Gamma, \ldots, E_c\Gamma).$$

We are now equipped to state our main identification result:

Theorem 2. *Suppose Assumptions 1 and 2 hold; then the column space of* Γ *is identified by the null space of the APM* A.

The containment of the column space of Γ in the null space of the APM A can be shown succinctly even without Assumption 2 (see Footnote 14). However, the containment of the null space of the APM A in the column space of Γ under Assumption 2 requires a more involved argument, which we provide in Appendix A.2.

Given identification of the column space of the factor matrix Γ, we can then identify the target cohort's outcome means via the bridge function approach articulated in Section 3:

Corollary 3. *Suppose Assumptions 1 and 2 hold, and let* $\tilde{\Gamma}$ *be any basis for the null space of the APM* A. *Then the vector* μ_{c^*} *of outcome means for cohort* c^* *is identified as follows:*

$$\mu_{c^*} = \tilde{\Gamma}\left(E_{c^*}\tilde{\Gamma}\right)^{+}\mathbb{E}[E_{c^*}Y_i \mid C_i = c^*].$$

[20] Such an assumption is analogous to restrictions of the samples in empirical work on bipartite matching settings to the row-type units who are matched with at least two column-type units.

We provide a proof of Corollary 3 in Appendix A.3.

Before continuing, we note that, because we impose the factor model functional form (1), our identification argument for μ_c does not rely on any support assumptions about the cohort-specific distributions of loadings, unlike some other approaches to estimating counterfactual outcome means in the presence of unobserved heterogeneity. For example, in the context of event study settings, methods in the Synthetic Control family like Abadie, Diamond, and Hainmueller (2010) and Arkhangelsky et al. (2021) typically require the average target cohort's loadings to lie in the convex hull of the loadings of the units in the donor pool. In the context of bipartite match outcomes, Bonhomme et al. (2019) require the number of distinct latent types to be finite and that every latent type of row-type unit is matched with every latent type of column-type unit with positive probability. In Section 4.3, none of the additional sets of assumptions we discuss require restricted or overlapping support in the cohort-specific loading distributions for Assumption 1 to hold either. As such, our approach is robust to large discrepancies in the unobserved confounding variables across cohorts, so long as the low-rank factor model (1) is correctly specified.

4.2 Asymptotic Linearity of $\hat{\mu}_c$ Given Cohort-Specific Factors

Having shown how to identify the column space of Γ and, by extension, μ_c, we now turn to showing that the plug-in estimator described in Section 3 yields asymptotically linear estimates of these parameters. First, we introduce some convenient notation. Let $\text{vec}(M)$ denote the vectorization of the matrix M, i.e. the vector containing the columns of the matrix M stacked in order, and let $\hat{\mathbb{E}}_N$ denote the expectation operator with respect to the empirical measure $\hat{P}_N$ with sample size N.

To highlight the contributions of this paper, as in Section 4.1, we will assume in this section that we are equipped with asymptotically linear estimators $\Pi(\hat{\Gamma})$ of the column spaces of $E_c\Gamma$, represented uniquely by $\Pi(E_c\Gamma)$:

Assumption 3. For each cohort $c = 1, \ldots, C$, there exists a function ϕ_c of C_i and Y_i such that the vectorization of the estimated cohort-specific factor projection matrix $\Pi(\hat{\Gamma}_c)$ satisfies the following asymptotic expansion as $N \to \infty$:

$$\sqrt{N}\text{vec}\left(\Pi(\hat{\Gamma}_c) - \Pi(E_c\Gamma)\right) = \sqrt{N}\hat{\mathbb{E}}_N[\phi_c(C_i, Y_i)] + o_p\left(1\right), \tag{6}$$

$\mathbb{E}[\phi_c(C_i, Y_i)] = \mathbf{0}_T$, and $\mathbb{E}\left[\|\phi_c(C_i, Y_i)\|_2^2\right] < \infty$.

In Section 4.3, we will discuss several sets of low-level assumptions standard in the rich literature on factor models under which estimators satisfying Assumption 3 exist.

Next, we can show that the plug-in estimator $\hat{\Gamma}$ of Γ is asymptotically linear in the following sense, where $\otimes$ denotes the Kronecker product:

Theorem 4. *Suppose that Assumptions 1, 2, and 3 hold. Then as* $N \to \infty$,

$$\sqrt{N}\text{vec}\left(\Pi(\hat{\Gamma}) - \Pi(\Gamma)\right) = \sqrt{N}\hat{\mathbb{E}}_N\left[H\sum_{c=1}^{C}\phi_c(C_i, Y_i)\right] + o_p(1) \tag{7}$$

where H *is the following* $T^2 \times T^2$ *matrix:*

$$H \coloneqq A^+ \otimes \Pi(\Gamma) + \Pi(\Gamma) \otimes A^+,$$

$\mathbb{E}[H\sum_{c=1}^{C}\phi_c(C_i, Y_i)] = \mathbf{0}_{T^2}$, *and* $\mathbb{E}[\|H\sum_{c=1}^{C}\phi_c(C_i, Y_i)\|_2^2] < \infty$.

We provide a proof of Theorem 4 in Appendix C.1. Our proof relies on an exact, first-order expansion of the operator mapping a symmetric matrix into the projection matrix onto the space spanned by some subset of its eigenvectors derived using Kato's integral (Kato, 1949). Since the projection matrix onto the space spanned by a matrix's columns is a unique representation of that matrix's column space, this expansion allows us to directly bound the error incurred by $\Pi(\hat{\Gamma})$ as an estimator of the column space of Γ under minimal assumptions on the eigenvalues of A. Because it is exact, our expansion appears to be new as far as we are aware. Since this expansion may be of independent interest, we provide a self-contained description in Appendix B.

Having shown that our plug-in estimator of the column space of Γ is asymptotically linear, we now turn to showing that $\hat{\mu}_c$ is therefore also an asymptotically linear estimator of μ_c. To do so, we require the following additional regularity conditions on cohort sizes and outcome moments, where $Y_i^* \coloneqq (Y_{i1}^*, \ldots, Y_{iT}^*)'$ denotes the vector that collects unit i's outcomes:

Assumption 4. For all $c = 1, \ldots, C$, $\mathbb{P}(C_i = c) > 0$ and $\mathbb{E}\left[\|Y_i^*\|_2^2 \mid C_i = c\right] < \infty$.

As it happens, we can express $(E_c\Gamma)^+$ as a known function of E_c and $\Pi(\Gamma)$, which enables us to derive a first-order asymptotic expansion of our plug-in estimator $\hat{\mu}_c$ with respect to $\Pi(\hat{\Gamma})$.[21] This result implies $\hat{\mu}_c$ has exactly our desired properties:

Theorem 5. *Suppose that Assumptions 1, 2, 3, and 4 hold. Then as* $N \to \infty$,

$$\begin{aligned}
\sqrt{N}(\hat{\mu}_{c^*} - \mu_{c^*}) &= \sqrt{N}\hat{\mathbb{E}}_N\left[\psi_{c^*}(C_i, Y_i)\right] + o_p(1), \\
\psi_{c^*}(C_i, Y_i) &\coloneqq \left((\mu_{c^*}' E_{c^*} R_{c^*}) \otimes R_{c^*}\right) H \sum_{c=1}^{C}\phi_c(C_i, Y_i) \\
&\quad + \frac{\mathbb{1}\{C_i = c^*\}}{\mathbb{P}(C_i = c^*)}\left(\Gamma\left(E_{c^*}\Gamma\right)^+ E_{c^*}Y_i - \mu_{c^*}\right),
\end{aligned} \tag{8}$$

where

$$R_{c^*} \coloneqq I + \Gamma(\Gamma' E_{c^*}\Gamma)^{-1}\Gamma'(I - E_{c^*}), \tag{9}$$

[21] Such a result may be no surprise given that the Moore-Penrose pseudo-inverse of a tall matrix is invariant to an invertible transformation of the matrix's rows.

$\mathbb{E}[\psi_{c^*}(C_i, Y_i)] = \mathbf{0}_T$, *and* $\mathbb{E}\left[\|\psi_{c^*}(C_i, Y_i)\|_2^2\right] < \infty$.

We provide a proof of Theorem 5 in Appendix C.2. We note that R_{c^*} is invariant to changes of Γ's basis.

Given $\hat{\mu}_c$'s asymptotic linearity, we can derive the asymptotic properties of the plug-in estimator $\hat{\theta}$ of some target estimand $\theta = h(\theta, \eta)$ and the inference procedure for θ introduced in Section 3 as corollaries, where h is a known, smooth function of θ and a vector of nuisance parameters η that are consistently estimable at a parametric rate. In particular, in Appendix E.3, we prove that $\hat{\theta}$ is asymptotically normal and that our Bayesian-bootstrap-based simultaneous confidence intervals described in Section E.2 have valid simultaneous coverage of the coordinates of θ under minimal additional regularity conditions.

In Appendix E, we also describe two examples of target parameters that satisfy those regularity conditions. For event study settings, we show how the dynamic treatment effects discussed in Callaway and Sant'Anna (2021) and L. Sun and Abraham (2021) fit into our framework. For studying match outcomes, we introduce a decomposition of the difference analogous to the target parameter in Finkelstein et al. (2016).

4.3 Identifying and Estimating Cohort-Specific Factors

Having established that, given sufficiently accurate estimates of cohort-specific factor projection matrices $\Pi(E_c\Gamma)$, our approach yields consistent estimates of and allows us to conduct valid inference on functions of outcome means, we now turn to constructing such estimates of $\Pi(E_c\Gamma)$. Importantly, without additional assumptions, $\Pi(E_c\Gamma)$ cannot be identified (Anderson and Rubin, 1956). As such, in this section, we provide a non-exhaustive discussion of several common sets of assumptions from the rich literature on estimating factor models without missing outcomes that enable identification and estimation of $\Pi(E_c\Gamma)$ when the number of observed outcomes $|\mathcal{T}_c|$ for cohort c remains fixed as N_c grows.

Uncorrelated, Homoskedastic Outcomes. First, we discuss a minimal set of assumptions that allows the canonical Principal Components (PC) estimator to yield consistent estimates of $\Pi(E_c\Gamma)$ when we require T to stay finite as N grows. In particular, let V_c denote the $T \times T$ matrix of cohort c's observed outcomes' second moments:

$$V_c \coloneqq \mathbb{E}[E_c Y_i Y_i' E_c \mid C_i = c], \tag{10}$$

and let $\tilde{\Gamma}_{c,\text{PC}}$ be any matrix whose columns are eigenvectors of V_c corresponding to V_c's r largest eigenvalues. Further, let $\hat{V}_c$ denote the empirical counterpart of V_c:

$$\hat{V}_c \coloneqq \hat{\mathbb{E}}_N[E_c Y_i Y_i' E_c \mid C_i = c], \tag{11}$$

and define the PC estimator of $E_c\Gamma$ to be any matrix $\hat{\Gamma}_{c,\text{PC}}$ whose columns are eigenvectors of $\hat{V}_c$ corresponding to $\hat{V}_c$'s r largest eigenvalues.[22]

Importantly, the PC estimator does not require any auxiliary data like covariates or instruments to be consistent, and, since it is equivalent to an eigendecomposition, efficient algorithms exist for its computation. The costs of its simplicity are the assumptions it requires. First, all loadings must have nontrivial variances, which rules out the case where all units have the same loadings. Second, the residuals ε_{it} must be uncorrelated across units i and outcomes t, and, while they can have arbitrarily varying variances across units i, they must be constant across outcomes t.

While these assumptions are weaker than the requirement from the classical factor model literature that ε_{it} be drawn independently from the same mean-zero Gaussian distribution (see e.g. Anderson (1963)), Theorem 4 in J. Bai (2003) proves that they are both sufficient (as shown in Connor and Korajczyk (1986)) and necessary for the PC estimator to be consistent when T remains fixed as N grows. In Appendix F.2, we state these assumptions formally and prove directly that $\Pi(\hat{\Gamma}_{c,\text{PC}})$ is an asymptotically linear estimator of $\Pi(E_c\Gamma)$ as required by Assumption 3 under slightly weaker conditions than are typically imposed in the literature. To do so, we again rely on our first-order expansion of the eigenspace projection operator described in Appendix B.

Uncorrelated, Heteroskedastic Outcomes. If we are instead only willing to believe that the residuals ε_{it} are uncorrelated across units i and outcomes t but can have arbitrarily heterogeneous variances across both units and outcomes, then the PC estimator $\Pi(\hat{\Gamma}_{c,\text{PC}})$ will be inconsistent for $\Pi(E_c\Gamma)$ if T remains fixed as N grows (J. Bai, 2003; Chamberlain and Rothschild, 1983).[23] Instead, the literature suggests estimating $\Pi(E_c\Gamma)$ via an optimization-based approach called Factor Analysis (FA) (see e.g. Chapter 14 of Anderson (2009)). While the global maximizers of the FA objective yield asymptotically linear estimators of $\Pi(E_c\Gamma)$ assuming the population parameters maximize the population objective (Anderson, 2009), the objective is non-concave, and even state-of-the-art algorithms for verifiably computing its global maximum are difficult to scale for even moderate r and T (Bertsimas, Copenhaver, and Mazumder, 2017; Khamaru and Mazumder, 2019).

Instead, in Appendix F.3, we sketch a new, computationally efficient procedure for estimating $\Pi(E_c\Gamma)$ that can be applied when residuals ε_{it} are uncorrelated but heteroskedastic so long as at least $2r+1$ outcomes are observed per cohort, like internal-instruments-based approaches to identifying factors under uncorrelated but heteroskedastic residuals (Freyberger, 2018; Harding,

[22] As discussed in Footnote 16 in Section 3, eigendecompositions can be computed efficiently using optimized routines available in most programming languages.

[23] In fact, the PC estimator is numerically equivalent to minimizing a least squares objective $N^{-1}\sum_{i=1}^{N}\|E_c(Y_i-\Gamma\lambda_i)\|_2^2$ with respect to Γ and every λ_i, so, as shown in Chamberlain and Rothschild (1983), $\hat{\Gamma}_{c,\text{PC}}$ suffers from the incidental parameter problem under residual heteroskedasticity (T. Lancaster, 2000; Neyman and Scott, 1948).

Lamarche, and Muris, 2022; Heckman and Scheinkman, 1987; Madansky, 1964). The procedure is based on the fact that, if we split the observed outcomes for a given cohort into two disjoint subsets, the matrix of covariances between pairs of observed outcomes from the two subsets is determined solely by the factor structure. As such, the top r left singular vectors of this matrix yield estimates of the factor vectors corresponding to each subset of outcomes with respect to the same basis. Repeating this procedure for a particular sequence of partitions of a cohort's observed outcomes yields factor vector estimates that can then be "stitched together" via another application of our APM-based factor estimation procedure. This procedure is guaranteed to recover $\Pi(E_c\Gamma)$ with only $r+1$ eigendecompositions.

Other Identifying Assumptions Based on Auxiliary Data. With access to auxiliary data, we can also relax the assumption of uncorrelatedness of residuals across outcomes using several different approaches. One approach is to assume that some vector of at least r unit-and-outcome-specific covariates are also outcome-specific linear functions of λ_i plus covariate residuals that are uncorrelated with λ_i and ε_{it}. Then one can apply the Common Correlated Effects method (Pesaran, 2006; Westerlund, Petrova, and Norkute, 2019), Essential Regression (Bing, Bunea, and Wegkamp, 2022), transfer learning (Duan, Pelger, and Xiong, 2023), or the Diversified Projections method (Fan and Liao, 2022) to estimate $\Pi(E_c\Gamma)$. Alternatively, given access to at least r unit-and-outcome-specific external instruments that are correlated with the coordinates of λ_i but not ε_{it}, one can apply the GMM estimators proposed in Ahn et al. (2013) and Robertson and Sarafidis (2015) to estimate $\Pi(E_c\Gamma)$.

5 Empirical Illustration

5.1 Setting

Our empirical illustration of our method is based on data from the Veneto Worker Histories (VWH) dataset,[24] which is a dataset derived from the Italian social security administration containing the full history of weekly wages for every person who ever lived or worked in any of the seven provinces in the Veneto region of Italy from 1975 to 2001.[25] Each observation corresponds to a worker working for a firm in a given year, and contains information about the number of weeks that person worked at that firm in that year as well as the total wages they were paid for doing so.

We focus on characterizing the performance of our procedure as an alternative to TWFE-

[24]The VWH dataset was developed by the Economics Department in Università Ca' Foscari Venezia under the supervision of Giuseppe Tattara, and can be accessed at https://www.frdb.org/en/dati/dati-inps-carriere-lavorative-in-veneto/.

[25]The code used to conduct our empirical illustration can be found at https://github.com/brad-ross/apm.

based methods for assessing to what degree worker's locations causally affect their wages as opposed to differences in workers' observed wages across locations being driven by purely worker sorting, in the spirit of Card et al. (2023). As discussed in Example E.2 that is inspired by Finkelstein et al. (2016), important inputs to any such analysis are the predicted counterfactual wages for workers had they instead worked in locations we do not observe them working at in the data. The workhorse methods for constructing these predicted average counterfactual wages are based on the TWFE model (Abowd et al., 1999; Card et al., 2023; Finkelstein et al., 2016), which rules out any complementarities between workers and locations, unlike predicted average counterfactual wages given by our factor-model-based method.[26]

To assess how well our method performs at predicting counterfactual mean wages for workers across provinces relative to counterfactual mean wages predicted by a TWFE regression, we construct a panel dataset that fits within our setup of interest as follows. First, we restrict our attention to the subsample of 116,814 firms that existed in the Veneto region from 1998 through 2001 and the 1,109,551 unique workers who always worked for firms in this subsample during the same period; this restriction diminishes the impact of long-run secular trends on our results (Lachowska, Mas, Saggio, and Woodbury, 2023). To avoid understating the degree to which wages are determined by workers' provinces of employment by ignoring within-province firm heterogeneity (Card et al., 2023), we then cluster the between 5,000 and 22,000 firms located in each province into three types using the k-means-based procedure proposed in Bonhomme et al. (2019); details can be found in Appendix G.1. We then let t correspond to a triple of a two-year range (either 1998 - 1999 or 2000 - 2001), a province, and a firm's type within that province, meaning $T = 42$. Finally, to limit the numerical instability of our estimator, we drop units belonging to cohorts of fewer than 75 units, leaving us with 1,033,144 workers in our final sample.

Having defined our panel data structure, we now define the observed outcome Y_{it} to be the natural logarithm of worker i's average weekly wage earned while working at all firms in t's province and firm type during t's two-year range, reminiscent of the firm-by-time fixed effects in Lachowska et al. (2023). The outcome Y_{it}^* can then be defined analogously, but it corresponds to the potentially counterfactual log-average weekly wage worker i would have earned at firms in t's firm group that they never actually worked at during t's year range. Importantly, because we restrict our sample to workers who worked at Veneto-region firms in both 1998 - 1999 and 2000 - 2001, we observe at least two average weekly wages for each worker; as such, we can estimate the factors γ_t using both the typical mover population discussed in the literature on matched

[26] Bonhomme et al. (2019) note that typical empirical tests in the literature that claim to not detect match effects in fact have no power under a variety of models that allow for worker-firm complementarities (see Card et al. (2013) for an example deploying such tests). Card et al. (2013) and Woodcock (2015) find evidence for the existence of match effects using repeated measurements of the same matches over time, but such estimates are noisy and/or require random-effects-like assumptions. Bonhomme et al. (2019) estimate a model of bipartite match outcomes with discrete worker and firm types and also find evidence of complementarities.

employer-employee data, as well as the "stayer" population of workers who don't change firm groups during our relevant time period.

For simplicity, we use the PC estimator described in Section 4.3 to estimate cohort-specific factors in this application. After examining the output of the PC estimator applied to the data for each cohort, the assumption that $r = 1$ appears to be quite plausible in this setting; see Figure G.2 and its caption for a visualization and more detailed discussion. As such, we focus on evaluating the $r = 1$ version of our estimator. While assuming $r = 1$ rules out the existence of multidimensional unobserved confounders, our approach still allows for complementarity between worker and year-range-by-province-by-firm-type effects in determining log-wages, unlike the TWFE model.

Figure 1b illustrates the outcome missingness pattern in our sample. Importantly, no embedded block missingness pattern exists with which we can identify most cohorts' outcome means (see Figure G.3 for a quantification),[27] precluding the use of methods that rely on the existence of reference cohorts, as discussed in Section 2.2. In contrast, the observed outcome overlap graph $\mathcal{G}_1$ (which is illustrated in Figure G.1b) is connected, so Assumption 2 holds and $\Pi(\Gamma)$ and μ_c are identified. As such, we focus on comparing our method to cohort outcome means estimated based on TWFE regressions in our empirical evaluation below.

5.2 Semi-Synthetic Simulation Study of Estimator Performance

To evaluate the performance of our method at predicting counterfactual outcome means, we conduct a semi-synthetic simulation study based on the dataset whose construction we described in Section 5.1. First, we choose a target cohort c^* of workers with at least three observed outcomes and "mask" one of their outcomes t^*, treating it as if it were unobserved.[28] Next, we resample with replacement from each cohort in this masked dataset 100 times, constructing 100 synthetic datasets drawn from the same distribution and with the same cohort sizes as our masked dataset but without the target outcome t^* observed for the units in the target cohort c^*. For each masked synthetic dataset, we compute $\hat{\mu}_{c^*t^*}$ using our method, as well as a TWFE-based analog. Finally, we evaluating the accuracy of each estimator by computing the absolute bias, standard error, and root mean squared error over the 100 estimates constructed using each estimator, treating the actual sample mean in our original, unmasked dataset of the target outcome t^* for the units in the target cohort c^* as the ground truth. To give a representative sense of each estimator's accuracy,

[27] For example, over 50% of units belong to cohorts for whom only 20 out of 42 cohort outcome means are identifiable using embedded block missingness patterns.

[28] This masking exercise is analogous to the network cross-validation procedure introduced in Li, Levina, and Zhu (2020), which can be used for model comparison and selection without "double-dipping." In our setting, it could be applied to choose our assumed rank r and/or a method for estimating cohort-specific factors from the menu of options in Section 4.3 rigorously. However, we leave a thorough investigation of its applicability for future research.

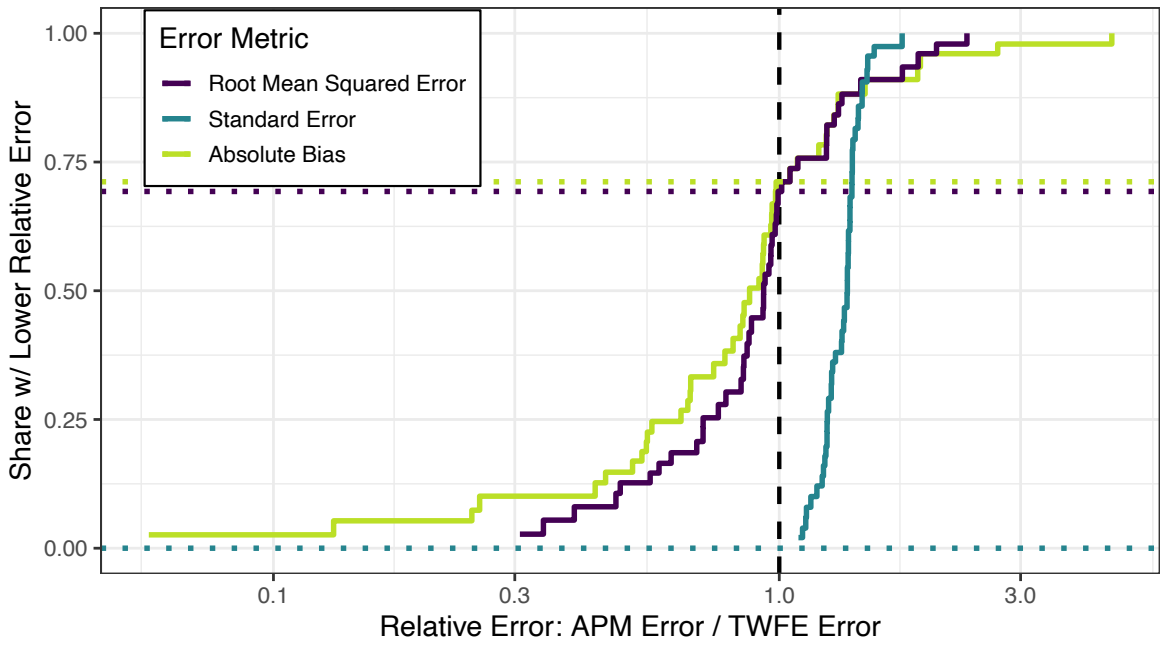


Figure 3: This figure plots the CDFs of error metric ratios between our estimator and the TWFE-based estimator across masked target parameters. The vertical, black, dashed line denotes the threshold above which our estimator performs worse than the TWFE estimator with respect to a given error metric, while the horizontal dotted lines correspond to the shares of target parameters for which our estimator performs better than the TWFE estimator on the error metric corresponding to its color, as labeled in the plot's legend.

we repeat this bootstrapped estimator evaluation procedure across every observed outcome t^* for the 15 largest cohorts c^* of workers with at least three observed outcomes.

We illustrate the distributions of our estimation error metrics across target cohorts c^* and outcomes t^* in Figure 3. In particular, Figure 3 plots the cumulative distribution functions (CDF) of error metric ratios between our estimator and the TWFE-based estimator across masked target parameters weighted by cohort size.[29] For 69% of target cohorts and outcomes, our factor-model-based method attains smaller root mean squared error than the TWFE-based estimator. This frequent better performance is due to our estimator having lower bias than the TWFE-based estimator for 71% of target cohorts and outcomes. The price of using our estimator over a TWFE-based estimator tends to be slightly higher variance, as indicated by the fact that the TWFE-based estimator had lower variance for every target cohort and outcome we considered. However, the increased variance of our estimator is not enough to outweigh our estimator's smaller bias relative to a TWFE-based estimator when comparing the root mean squared errors of the two methods, as discussed above. Overall, this semi-synthetic simulation study provides more evidence that accounting for complementarities between workers and firms using a method like

[29] In Figure G.4, we provide scatter plots illustrating the estimators' absolute error metric values across target cohorts and outcomes.

ours can yield more accurate estimates of average counterfactual match outcomes in bipartite matching settings.

6 Conclusion

In this paper, we develop a new approach for identifying and estimating average counterfactual outcomes with short panel data applicable in settings like event studies and studies of bipartite match outcomes. Relying only on an eigendecomposition of a new spectral operator, our method produces consistent, asymptotically normal estimates of means of outcomes generated by a latent factor model under general outcome missingness patterns as only the cross-sectional dimension of a panel grows large. Importantly, our procedure accommodates more general missingness patterns than other approaches for short panels, and it automatically stitches together different pieces of identifying information each used in isolation by existing methods. Through our simulation study based on the Veneto Worker Histories dataset, we also demonstrate the practicality of our approach in real-world "short" panel data settings.

References

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly robust inference in causal latent factor models. *arXiv preprint arXiv:2402.11652*. [6]

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505. [17]

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002). *Computing person and firm effects using linked longitudinal employer-employee data* (Tech. Rep.). Center for Economic Studies, US Census Bureau. [7, 16, 59, 60]

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, *67*(2), 251–333. [2, 9, 22]

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023). Causal matrix completion. In *The thirty sixth annual conference on learning theory* (pp. 3821–3826). [6, 7, 11]

Agarwal, A., Shah, D., and Shen, D. (2023). Synthetic interventions. *arXiv preprint arXiv:2006.07691*. [6, 10]

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227. [9]

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, *174*(1), 1–14. [9, 21]

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, *34*(1), 122–148. [20]

Anderson, T. W. (2009). *An introduction to multivariate statistical analysis, 3rd edition.* Wiley-Interscience. [11, 20]

Anderson, T. W., and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third berkeley symposium on mathematical statistics and probability: Held at the statistical laboratory, university of california, december, 1954, july and august, 1955* (Vol. 1, p. 111). [11, 19]

Angrist, J. D., and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press. [2]

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, *111*(12), 4088–4118. [6, 17]

Arkhangelsky, D., and Hirshberg, D. (2023). Large-sample properties of the synthetic control method under selection on unobservables. *arXiv preprint arXiv:2311.13575*. [6]

Arkhangelsky, D., and Samkov, A. (2024). Sequential synthetic difference in differences. *arXiv preprint arXiv:2404.00164*. [7]

Ashenfelter, O., and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, *67*(4), 648–660. [2]

Ater, I., Shany, A., Ross, B., Turkel, E., and Vasserman, S. (n.d.). Can usage-based pricing reduce traffic congestion? *Work in progress*. [3, 64]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. W., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, *116*(536), 1716–1730. [5, 6, 7]

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171. [20, 62]

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279. [6]

Bai, J., and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221. [9, 62]

Bai, J., and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, *176*(1), 18–29. [11]

Bai, J., and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, *116*(536), 1746–1763. [6, 12]

Bai, J., and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*. [62]

Bai, Z., and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices* (Vol. 20). Springer. [7]

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298. [55]

Ben-Michael, E., Feller, A., and Rothstein, J. (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B*, *84*(2), 351–381. [6]

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275. [2]

Bertsimas, D., Copenhaver, M. S., and Mazumder, R. (2017). Certifiably optimal low rank factor

analysis. *The Journal of Machine Learning Research*, *18*(1), 907–959. [20]

Bing, X., Bunea, F., and Wegkamp, M. (2022). Inference in latent factor regression with clusterable features. *Bernoulli*, *28*(2), 997–1020. [21]

Bonhomme, S. (2020). Econometric analysis of bipartite networks. In *The econometric analysis of network data* (pp. 83–121). Elsevier. [7, 16, 59]

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A distributional framework for matched employer employee data. *Econometrica*, *87*(3), 699–739. [2, 5, 7, 16, 17, 22, 60, 63, 64]

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, rdae007. [2]

Brown, N., and Butts, K. (2022). Generalized imputation estimators for factor models. [6, 7, 10, 11]

Brown, N., Butts, K., and Westerlund, J. (2023). Difference-in-differences via common correlated effects. *arXiv preprint arXiv:2301.11358*. [7]

Callaway, B., and Karami, S. (2023). Treatment effects in interactive fixed effects models with a small number of time periods. *Journal of Econometrics*, *233*(1), 184–208. [7]

Callaway, B., and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. [8, 14, 19, 52]

Callaway, B., and Tsyawo, E. S. (2023). Treatment effects in staggered adoption designs with non-parallel trends. *arXiv preprint arXiv:2308.02899*. [7]

Card, D., Heining, J., and Kline, P. (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly Journal of Economics*, *128*(3), 967–1015. [2, 22]

Card, D., Rothstein, J., and Yi, M. (2023). *Location, location, location* (Tech. Rep.). National Bureau of Economic Research. [2, 22]

Chamberlain, G. (1984). Panel data. *Handbook of econometrics*, *2*, 1247–1318. [2]

Chamberlain, G., and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*(5), 1281–1304. [20]

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205–2268. [14, 55]

Chernozhukov, V., Hansen, C., Liao, Y., and Zhu, Y. (2023). Inference for low-rank models. *The Annals of statistics*, *51*(3), 1309–1330. [6]

Chetty, R., and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, *133*(3), 1163–1228. [2]

Choi, J., Kwon, H., and Liao, Y. (2023). Inference for low-rank completion without sample splitting with application to treatment effect estimation. *arXiv preprint arXiv:2307.16370*. [6]

Choi, J., and Yuan, M. (2023). Matrix completion when missing is not at random and its applications in causal panel data models. *arXiv preprint arXiv:2308.02364*. [6]

Connor, G., and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of financial economics*, *15*(3), 373–394. [20]

Deaner, B. (2018). Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*. [6]

De Chaisemartin, C., and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996. [8]

Duan, J., Pelger, M., and Xiong, R. (2023). Target pca: Transfer learning large dimensional panel data. *Journal of Econometrics*, 105521. [21]

Fan, J., and Liao, Y. (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, *117*(538), 909–924. [21]

Farias, V., Li, A., and Peng, T. (2021). Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems*, *34*, 14001–14013. [6]

Fernández-Val, I., Freeman, H., and Weidner, M. (2021). Low-rank approximations of nonseparable panel models. *The Econometrics Journal*, *24*(2), C40–C77. [6]

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, *131*(4), 1681–1726. [2, 8, 14, 19, 22, 53]

Fortin, A.-P., Gagliardini, P., and Scaillet, O. (2022). Eigenvalue tests for the number of latent factors in short panels. *arXiv preprint arXiv:2210.16042*. [9]

Fortin, A.-P., Gagliardini, P., and Scaillet, O. (2023). Latent factor analysis in short panels. *arXiv preprint arXiv:2306.14004*. [9]

Freeman, H., and Weidner, M. (2023). Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics*, *237*(1), 105498. [6]

Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *The Review of Economic Studies*, *85*(3), 1824–1851. [20]

Gagliardini, P., Ossola, E., and Scaillet, O. (2019). A diagnostic criterion for approximate factor structure. *Journal of Econometrics*, *212*(2), 503–521. [9]

Ghanem, D., Sant'Anna, P. H., and Wüthrich, K. (2022). Selection and parallel trends. *arXiv preprint arXiv:2203.09001*. [2]

Gobillon, L., and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, *98*(3), 535–551. [6]

Harding, M., Lamarche, C., and Muris, C. (2022). Estimation of a factor-augmented linear model with applications using student achievement data. *arXiv preprint arXiv:2203.03051*. [20]

Heckman, J., and Scheinkman, J. (1987). The importance of bundling in a gorman-lancaster model of earnings. *The Review of Economic Studies*, *54*(2), 243–255. [21]

Hull, P. (2018). Estimating treatment effects in mover designs. *arXiv preprint arXiv:1804.06721*. [7, 12, 16, 53, 60]

Imbens, G. W., Kallus, N., and Mao, X. (2021). Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv preprint arXiv:2108.03849*. [6, 10]

Imbens, G. W., and Viviano, D. (2023). *Identification and inference for synthetic controls with confounding*. [6]

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, *6*(1), 801–825. [2]

Jochmans, K., and Weidner, M. (2019). Fixed-effect regressions on network data. *Econometrica*, *87*(5),

1543–1560. [2, 7, 16, 59, 60]

Kato, T. (1949). On the convergence of the perturbation method. i. *Progress of Theoretical Physics*, *4*(4), 514–523. [5, 7, 18, 32, 39]

Kato, T. (1980). *Perturbation theory for linear operators* (Vol. 132). Springer Science & Business Media. [32, 39]

Khamaru, K., and Mazumder, R. (2019). Computation of the maximum likelihood estimator in low-rank factor analysis. *Mathematical Programming*, *176*, 279–310. [20]

Lachowska, M., Mas, A., Saggio, R., and Woodbury, S. A. (2023). Do firm effects drift? evidence from washington administrative data. *Journal of Econometrics*, *233*(2), 375–395. [22]

Lancaster, P., and Farahat, H. K. (1972). Norms on direct sums and tensor products. *mathematics of computation*, *26*(118), 401–414. [36]

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics*, *95*(2), 391–413. [6, 20]

Lazear, E. P. (2009). Firm-specific human capital: A skill-weights approach. *Journal of political economy*, *117*(5), 914–940. [9]

Lei, L. (2020). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*. [8]

Li, T., Levina, E., and Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, *107*(2), 257–276. [23]

Little, R. J., and Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons. [1]

Madansky, A. (1964). Instrumental variables in factor analysis. *Psychometrika*, *29*(2), 105–113. [21]

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, *105*(4), 987–993. [6]

Montiel Olea, J. L., and Plagborg-Møller, M. (2019). Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics*, *34*(1), 1–17. [59]

Moon, H. R., and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, *83*(4), 1543–1579. [6]

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1–32. [6, 20]

Oliveira, R. I. (2010). Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*. [8]

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, *77*(5), 1447–1479. [9]

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, *74*(4), 967–1012. [21]

Robertson, D., and Sarafidis, V. (2015). Iv estimation of panels with factor residuals. *Journal of Econometrics*, *185*(2), 526–541. [21]

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. [1]

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, 130–134. [14]

Simons, J. R. (2023). Inference on eigenvectors of non-symmetric matrices. *arXiv preprint arXiv:2303.18233*. [8, 32]

Sun, J.-g. (1991). Perturbation expansions for invariant subspaces. *Linear Algebra and its Applications*, *153*, 85–97. [8, 32]

Sun, L., and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. [8, 14, 19, 52]

van der Vaart, A. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press. [56]

van der Vaart, A., and Wellner, J. A. (1996). Empirical processes. In *Weak convergence and empirical processes: With applications to statistics.* Springer. [54, 58, 59]

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press. [38]

Westerlund, J., Petrova, Y., and Norkute, M. (2019). Cce in fixed-t panels. *Journal of Applied Econometrics*, *34*(5), 746–761. [21]

Woodcock, S. D. (2015). Match effects. *Research in Economics*, *69*(1), 100–121. [2, 22]

Xiong, R., and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, *233*(1), 271–301. [6]

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76. [6]

Yan, Y., and Wainwright, M. J. (2024). Entrywise inference for causal panel data: A simple and instance-optimal approach. *arXiv preprint arXiv:2401.13665*. [6]

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, *102*(2), 315–323. [7]

A Proofs of Results in Section 4.1

A.1 Proof of Lemma 1

By Assumption 2, there must exist at least one other cohort c' that is a neighbor of cohort c in the observed outcome overlap graph $\mathcal{G}_r$. Further, since $E_{c'}$ is a square matrix, the column space of $\Gamma' E_c E_{c'}$ must be a subset of the column space of $\Gamma' E_c$, which, along with (5), implies that

$$\begin{aligned}
\operatorname{rank}(\underbrace{E_c\Gamma}_{T\times r}) &= \operatorname{rank}(\Gamma' E_c) && (T > r, \text{ rank is min. of row rank and column rank}) \\
&\geq \operatorname{rank}(\Gamma' E_c E_{c'}) \\
&= r && (c \text{ and } c' \text{ are neighbors, so (5) holds}).
\end{aligned}$$

Since $E_c\Gamma$ is $T \times r$ with $T > r$, $\operatorname{rank}(E_c\Gamma) \leq r$ must also hold, completing the proof.

A.2 Proof of Theorem 2

First, as in Footnote 14, we show that for any cohort c, the columns of Γ lie in the null space of $E_c - \Pi(E_c\Gamma)$, in which case $\text{col}(\Gamma) \subseteq \text{null}(A)$, where $\text{col}(\Gamma)$ denotes the column space of Γ. Since $E_c^2 = E_c$ and $E_c\Pi(E_c\Gamma) = \Pi(E_c\Gamma)E_c = \Pi(E_c\Gamma)$, we have that

$$(E_c - \Pi(E_c\Gamma))\Gamma = (E_c^2 - E_c\Pi(E_c\Gamma)E_c)\Gamma = E_c \cdot (I - \Pi(E_c\Gamma))E_c\Gamma = \mathbf{0}_{T\times r},$$

as required.

Next, we show that $\text{null}(A) \subseteq \text{col}(\Gamma)$ also holds. Define the matrix

$$\tilde{A} := \begin{bmatrix} E_1 - \Pi(E_1\Gamma) \\ \vdots \\ E_C - \Pi(E_C\Gamma) \end{bmatrix},$$

and note that since

$$(E_c - \Pi(E_c\Gamma))^2 = E_c^2 - E_c\Pi(E_c\Gamma) - \Pi(E_c\Gamma)E_c - \Pi(E_c\Gamma)^2 = E_c - \Pi(E_c\Gamma)$$

and $E_c - \Pi(E_c\Gamma)$ is symmetric, $A = \tilde{A}'\tilde{A}$, implying that $\text{null}(A) = \text{null}(\tilde{A})$.[30]

Consider any $v \in \text{null}(\tilde{A})$; by definition, it must be that for any cohort c, $(E_c - \Pi(E_c\Gamma))v = \mathbf{0}_T$, which in turn implies that $E_c v \in \text{col}(E_c\Gamma)$. Thus, there exists a vector $\omega_c \in \mathbb{R}^r$ such that $E_c v = E_c\Gamma\omega_c$. Then, for any edge (c_1, c_2) in the observed outcome overlap graph $\mathcal{G}_r$, since diagonal matrices commute,

$$E_{c_1}E_{c_2}v = E_{c_2}E_{c_1}v = E_{c_2}E_{c_1}\Gamma\omega_{c_1} = E_{c_1}E_{c_2}\Gamma\omega_{c_1}, \quad E_{c_1}E_{c_2}v = E_{c_1}E_{c_2}\Gamma\omega_{c_2},$$

implying that $E_{c_1}E_{c_2}\Gamma(\omega_{c_1} - \omega_{c_2}) = \mathbf{0}_T$. Since (5) implies that the null space of $E_{c_1}E_{c_2}\Gamma$ is trivial and $\omega_{c_1} - \omega_{c_2}$ lies in it, $\omega_{c_1} = \omega_{c_2}$ must also hold. Since $\mathcal{G}_r$ is connected, there exists a path in $\mathcal{G}_r$ between any two cohorts, so $\omega_1 = \ldots = \omega_C = \omega$ for some common vector $\omega \in \mathbb{R}^r$. For the t-th entry v_t of v, since every outcome is observed for at least one cohort, there exists at least one cohort c_t for whom $t \in \mathcal{T}_{c_t}$. Letting e_t denote the t-th standard basis vector, we have that

$$v_t = e_t'E_{c_t}\Gamma\omega = \gamma_t'\omega,$$

which, stacking across $t = 1, \ldots, T$, implies that $v = \Gamma\omega$, i.e. $v \in \text{col}(\Gamma)$. Putting everything together, we have that $\text{null}(A) = \text{null}(\tilde{A}) \subseteq \text{col}(\Gamma)$, as required.

[30] It is straightforward to check that if $v \in \text{null}(\tilde{A})$, then $v \in \text{null}(A) = \text{null}(\tilde{A}'\tilde{A})$. To check the other direction of inclusion, for any $v \in \text{null}(A)$, we have that $\mathbf{0} = Av = \tilde{A}'\tilde{A}v$. Further, we have that $0 = v'\,\mathbf{0} = v'\tilde{A}'\tilde{A}v = (\tilde{A}v)'(\tilde{A}v)$, so it must be that $\tilde{A}v = \mathbf{0}$, as required.

A.3 Proof of Corollary 3

First, we show that Assumption 2 is sufficient to ensure that $E_{c^*}\tilde{\Gamma}$ is has full rank r (since $T > r$ by assumption), and thus the $r \times r$ matrix $\tilde{\Gamma}' E_{c^*} \tilde{\Gamma} = \left(E_{c^*}\tilde{\Gamma}\right)' \cdot E_{c^*}\tilde{\Gamma}$ is invertible. Since $\tilde{\Gamma}$ is a basis for column space of Γ by Theorem 2, there exists some invertible basis matrix $Q \in \mathbb{R}^{r \times r}$ for which $\tilde{\Gamma} Q' = \Gamma$, so $\tilde{\Gamma} = \Gamma \left(Q'\right)^{-1}$. Then

$$\begin{aligned}
\text{rank}(E_{c^*}\tilde{\Gamma}) &= \text{rank}\left(E_{c^*}\Gamma\left(Q'\right)^{-1}\right) \\
&= \text{rank}(\underbrace{E_{c^*}\Gamma}_{T \times r}) && (\left(Q'\right)^{-1} \text{ is full-rank}) \\
&= r. && (\text{Lemma 1})
\end{aligned}$$

Next, since for full rank matrices M with more rows than columns, $M^+ = (M'M)^{-1}M'$, we have that

$$\begin{aligned}
\tilde{\Gamma}\left(E_{c^*}\tilde{\Gamma}\right)^+ \mathbb{E}[E_{c^*}Y_i \mid C_i = c^*] &= \Gamma Q' \cdot (Q\Gamma' E_{c^*}\Gamma Q')^{-1} Q\Gamma' E_{c^*} \cdot \mathbb{E}[E_{c^*}Y_i \mid C_i = c^*] \\
&= \Gamma \cdot (\Gamma' E_{c^*}\Gamma)^{-1}\Gamma' E_{c^*}\Gamma\, \mathbb{E}[\lambda_i \mid C_i = c^*] \\
&= \Gamma \cdot \mathbb{E}[\lambda_i \mid C_i = c^*] = \mu_{c^*},
\end{aligned}$$

as required.

B An Exact, First-Order Expansion of the Eigenspace Operator

In this section, we derive an exact, first-order expansion of the operator mapping a symmetric matrix into the projection matrix onto the space spanned by some subset of its eigenvectors. Our expansion is based on Kato's integral, which characterizes the projection matrix onto the space spanned by some subset of a matrix's eigenvectors as a contour integral of that matrix's resolvent Kato (1949). We characterize the approximation error in our expansion up to exact constants, unlike the asymptotic expansions given in e.g. Kato (1980) and J.-g. Sun (1991) (which is applied in Simons (2023) to construct hypothesis tests concerning eigenspaces). As such, our result below may be of independent interest.

To describe our expansion, we first introduce some notation. Let $\mathbb{S}^d$ denote the set of real-valued, d-dimensional symmetric matrices, and for any $M \in \mathbb{S}^d$, let $\lambda_j(M)$ be the jth eigenvalue of A, where

$$\lambda_0(M) \coloneqq -\infty < \lambda_1(M) \leq \cdots \leq \lambda_d(M) < \lambda_{d+1}(M) \coloneqq \infty,^{31}$$

[31] The eigenvalues of any symmetric matrix are all real-valued.

and we denote a generic eigendecomposition of M as follows:

$$M = U(M)\Lambda(M)U(M)' = \sum_{j=1}^{d} \lambda_j(M)\Pi(u_j(M)),$$

where $U(M)$ is any matrix with jth column $u_j(M)$ being an eigenvector corresponding to the jth eigenvalue $\lambda_j(M)$, and $\Pi(u_j(M))$ is the projection onto the span of $u_j(M)$.[32] For notational convenience, for any integers $1 \leq j, k \leq d$, we let

$$U_{j:k}(M) \coloneqq \begin{bmatrix} u_j(M) & \cdots & u_k(M) \end{bmatrix}$$

denote the matrix whose columns are the eigenvectors corresponding to $\lambda_j(M)$ through $\lambda_k(M)$.

Armed with this notation, we construct an exact bound on the error incurred by a first-order expansion of the difference between the projection matrices onto eigenspaces of two matrices M and $\hat{M}$ in terms of the magnitude of the difference between M and $\hat{M}$:

Theorem B.1. *Consider any integers s, r such that $1 \leq s+1 \leq s+r \leq d$ and any real-valued, d-dimensional symmetric matrix M satisfying the following eigen-gap condition:*

$$\Delta(M) > 0, \quad \Delta(M) \coloneqq 4^{-1} \min\left\{\lambda_{s+1}(M) - \lambda_s(M), \lambda_{s+r+1}(M) - \lambda_{s+r}(M)\right\}, \tag{B.1}$$

and define the following neighborhood of M:

$$\mathcal{B}(M) \coloneqq \left\{\hat{M} \in \mathbb{S}^d \ : \ \|\hat{M} - M\|_{\text{op}} \leq \Delta(M)\right\}.$$

Then for any $\hat{M} \in \mathcal{B}(M)$, the following first-order approximation holds:

$$\begin{aligned} &\Big\| \Big[\Pi(U_{(s+1):(s+r)}(\hat{M})) - \Pi(U_{(s+1):(s+r)}(M))\Big] \\ &\quad - \sum_{j=s+1}^{s+r} \sum_{k \notin [s+1, s+r]} \frac{1}{\lambda_k(M) - \lambda_j(M)} \Big[\Pi(u_j(M))(\hat{M} - M)\Pi(u_k(M)) \\ &\qquad\qquad + \Pi(u_k(M))(\hat{M} - M)\Pi(u_j(M))\Big] \Big\|_{op} \\ &\leq \frac{2}{\pi\Delta(M)^2}\|\hat{M} - M\|_{op}^2. \end{aligned} \tag{B.2}$$

We provide a proof of Theorem B.1 in Appendix D.1.

[32] Even though $U(M)$ is required to be orthonormal, the columns of $U(M)$ are only unique up to signs and permutations any eigenvectors corresponding to repeated eigenvalues.

C Proofs of Results in Section 4.2

C.1 Proof of Theorem 4

First, we show that $\|\hat{A} - A\|_{\text{op}} = O_p\left(N^{-1/2}\right)$:

$$\begin{aligned}
\|\hat{A} - A\|_{\text{op}} &\leq \|\hat{A} - A\|_F \\
&= \|\text{vec}(\hat{A} - A)\|_2 \\
&\leq \sum_{c=1}^{C} \|\text{vec}(\Pi(\hat{\Gamma}_c) - \Pi(E_c\Gamma))\|_2. \qquad \text{(Definitions of } \hat{A} \text{ and } A\text{)}
\end{aligned}$$

By Assumption 3, $\text{vec}(\Pi(\hat{\Gamma}_c) - \Pi(E_c\Gamma)) = O_p\left(N^{-1/2}\right)$. Since $T = O(1)$, it must be that $C = O(1)$ and hence $\|\hat{A} - A\|_{\text{op}} = O_p\left(N^{-1/2}\right)$ as well.

Next, let λ_j be the jth smallest eigenvalue of the population APM A, and denote a generic eigendecomposition of A as follows:

$$A = \sum_{j=1}^{T} \lambda_j \Pi(u_j),$$

where u_j is an eigenvector corresponding to the jth eigenvalue λ_j, and $\Pi(u_j)$ is the projection onto the span of u_j.[33] We note that all of A's eigenvalues must be non-negative.[34]

Given this notation, we now apply Theorem B.1 with $M = A$, $\hat{M} = \hat{A}$ as defined in Section 3, $s = 0$, and $r = r$. We also note that $\lambda_j = \lambda_j(A)$ using the notation defined in Appendix B. Since Theorem 2 implies that

$$0 = \lambda_1 = \ldots = \lambda_r < \lambda_{r+1},$$

we have that $\Delta(A) = 4^{-1}\lambda_{r+1} > 0$, satisfying (B.1). Further, Theorem 2 implies that

$$\Pi(\Gamma) = \Pi(U_{1:r}(A)),$$

and we define $\hat{\Gamma}$ in Section 3 such that

$$\Pi(\hat{\Gamma}) = \Pi(U_{1:r}(\hat{A})).$$

Then so long as

$$\|\hat{A} - A\|_{\text{op}} \leq 4^{-1}\lambda_{r+1}, \tag{C.1}$$

[33] Even though the matrix of eigenvectors whose columns are u_j is required to be orthonormal, its columns are only unique up to signs. Furthermore, if A has repeated eigenvalues, then any of the eigenvectors corresponding to those repeated eigenvalues are interchangeable.

[34] $E_c - \Pi(E_c\Gamma) = E_c[I - \Pi(E_c\Gamma)]E_c$, so since E_c and $I - \Pi(E_c\Gamma)$ are both projection matrices, their product must be positive semidefinite. The sum of positive semidefinite matrices must also be positive semidefinite, so then A must be positive semidefinite as well by (3).

Theorem B.1 implies that

$$\left\| \Pi(\hat{\Gamma}) - \Pi(\Gamma) - \sum_{j=1}^{r} \sum_{k=r+1}^{T} \lambda_k^{-1} \left[\Pi(u_j)(\hat{A} - A)\Pi(u_k) + \Pi(u_k)(\hat{A} - A)\Pi(u_j) \right] \right\|_{\text{op}} \leq \frac{32}{\pi \lambda_{r+1}^2} \|\hat{A} - A\|_{\text{op}}^2. \tag{C.2}$$

The fact that $\|\hat{A} - A\|_{\text{op}} = O_p\left(N^{-1/2}\right)$ implies (C.1) holds with probability approaching one as $N \to \infty$, so by (C.2) and the definitions of $\hat{A}$ and A,

$$\begin{aligned}
&\Pi(\hat{\Gamma}) - \Pi(\Gamma) \\
&= \sum_{j=r+1}^{T} \frac{1}{\lambda_j} \sum_{k=1}^{r} \left\{ \Pi(u_j)(\hat{A} - A)\Pi(u_k) + \Pi(u_k)(\hat{A} - A)\Pi(u_j) \right\} + o_p\left(N^{-1/2}\right). \\
&= \left(\sum_{j=r+1}^{T} \frac{1}{\lambda_j} \Pi(u_j) \right) (\hat{A} - A) \left(\sum_{k=1}^{r} \Pi(u_k) \right) \\
&\quad + \left(\sum_{k=1}^{r} \Pi(u_k) \right) (\hat{A} - A) \left(\sum_{j=r+1}^{T} \frac{1}{\lambda_j} \Pi(u_j) \right) + o_p\left(N^{-1/2}\right).
\end{aligned} \tag{C.3}$$

Next, since $\lambda_1 = \ldots = \lambda_r = 0$, it must be that

$$\sum_{j=r+1}^{T} \frac{1}{\lambda_j} \Pi(u_j) = \sum_{j=1}^{T} \frac{\mathbb{1}\{j > r\}}{\lambda_j} \Pi(u_j) = A^+,$$

and since $\text{rank}(\Gamma) = r$ and $\text{null}(A) = \text{col}(\Gamma)$,

$$\sum_{k=1}^{r} \Pi(u_k) = \Pi(\Gamma).$$

Substituting these two simplifications back into (C.3) and applying the definitions of $\hat{A}$ and A, we have

$$\begin{aligned}
&\Pi(\hat{\Gamma}) - \Pi(\Gamma) \\
&= \sum_{c=1}^{C} \left\{ A^+ [\Pi(\hat{\Gamma}_c) - \Pi(E_c\Gamma)]\Pi(\Gamma) + \Pi(\Gamma)[\Pi(\hat{\Gamma}_c) - \Pi(E_c\Gamma)]A^+ \right\} + o_p\left(N^{-1/2}\right).
\end{aligned} \tag{C.4}$$

Next, we note that the so-called mixed product property of the Kronecker product implies that, for any matrices $A \in \mathbb{R}^{d \times m}$, $B \in \mathbb{R}^{m \times p}$, and $C \in \mathbb{R}^{p \times q}$,

$$\text{vec}(ABC) = (C' \otimes A)\text{vec}(B). \tag{C.5}$$

Applying this fact to the vectorization of $\sqrt{N}$ times both sides of (C.4), since any projection

matrix $\Pi(\cdot)$ is symmetric, we can write the vectorization of our scaled and centered statistic $\sqrt{N}(\Pi(\hat{\Gamma}) - \Pi(\Gamma))$ as follows:

$$\begin{aligned}&\sqrt{N}\text{vec}(\Pi(\hat{\Gamma}) - \Pi(\Gamma)) \\ &= \left(A^+ \otimes \Pi(\Gamma) + \Pi(\Gamma) \otimes A^+\right) \left(\sum_{c=1}^{C} \sqrt{N}\text{vec}\left(\Pi(\hat{\Gamma}_{c^*}) - \Pi(E_{c^*}\Gamma)\right)\right) + o_p\left(1\right).\end{aligned} \tag{C.6}$$

Combining (C.6) and (6) yields (7), as required.

By the triangle inequality and the product property of Kronecker product spectra (P. Lancaster and Farahat, 1972),

$$\|H\|_{\text{op}} \leq \left\|A^+ \otimes \Pi(\Gamma)\right\|_{\text{op}} + \left\|\Pi(\Gamma) \otimes A^+\right\|_{\text{op}} = 2\left\|A^+\right\|_{\text{op}} \|\Pi(\Gamma)\|_{\text{op}} = \frac{2}{\lambda_{r+1}} < \infty.$$

The zero mean and bounded squared norm properties of $H\sum_{c=1}^{C}\phi_c(C_i, Y_i)$ follow from Assumption 3 and the fact that the operator norm of H is bounded.

C.2 Proof of Theorem 5

First, we show that with probability approaching one as $N \to \infty$, $\hat{\Gamma}' E_{c^*} \hat{\Gamma}$ is invertible. To do so, we note that by Theorem 4, $\Pi(\hat{\Gamma}) - \Pi(\Gamma) = o_p\left(1\right)$, so there must be a random, $r \times r$ invertible basis matrix $\hat{Q}$ such that $\hat{\Gamma}\hat{Q}' = \Gamma + o_p\left(1\right)$; if not, then for all random, $r \times r$ invertible basis matrices $\hat{Q}$, $\Pi(\hat{\Gamma}) = \Pi(\hat{\Gamma}\hat{Q}') \neq \Pi(\Gamma) + o_p\left(1\right)$, which would be a violation of Theorem 4. Since Lemma 1 implies that that $\Gamma' E_{c^*}\Gamma$ is invertible and $\hat{\Gamma}' E_{c^*}\hat{\Gamma} = (\hat{Q})^{-1}\Gamma E_{c^*}\Gamma(\hat{Q}')^{-1} + o_p\left(1\right)$, it must be that $\hat{\Gamma}' E_{c^*}\hat{\Gamma}$ is invertible with probability approaching one as $N \to \infty$. As such, in the remainder of the proof below, we condition on the event that $\hat{\Gamma}' E_{c^*}\hat{\Gamma}$ is invertible.

Next, we let $B_{c^*} \coloneqq \Gamma(E_{c^*}\Gamma)^+$ and $\hat{B}_{c^*} \coloneqq \hat{\Gamma}(E_{c^*}\hat{\Gamma})^+$. By Lemma 1, $E_{c^*}\Gamma$ is full-rank, in which case we have that $B_{c^*} = \Gamma(\Gamma' E_{c^*}\Gamma)^{-1}\Gamma' E_{c^*}$ by the definition of the Moore-Penrose pseudo-inverse. Further, since we've conditioned on the probability-one event that $\hat{\Gamma}' E_{c^*}\hat{\Gamma}$ is invertible and the rank of any matrix M must be equal to the rank of $M'M$, $E_{c^*}\hat{\Gamma}$ must have full rank, in which case we also have that $\hat{B}_{c^*} = \hat{\Gamma}(\hat{\Gamma}' E_{c^*}\hat{\Gamma})^{-1}\hat{\Gamma}' E_{c^*}$, again by the definition of the Moore-Penrose pseudo-inverse.

Having defined these quantities, we can decompose $\hat{\mu}_{c^*} - \mu_{c^*}$ as follows:

$$\begin{aligned}\hat{\mu}_{c^*} - \mu_{c^*} &= \hat{B}_{c^*}\hat{\mathbb{E}}_N[E_{c^*}Y_i \mid C_i = c^*] - B_{c^*}\,\mathbb{E}[E_{c^*}Y_i \mid C_i = c^*] \\ &= (\hat{B}_{c^*} - B_{c^*})E_{c^*}\mu_{c^*} + B_{c^*}(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*] \\ &\quad + (\hat{B}_{c^*} - B_{c^*})(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*].\end{aligned} \tag{C.7}$$

To approximate $\hat{B}_{c^*} - B_{c^*}$, we will rely on Lemma C.1 proved in Appendix D.2:

Lemma C.1. *For any two $T \times r$ matrices $\hat{\Gamma}$ and Γ such that $E_{c^*}\hat{\Gamma}$ and $E_{c^*}\Gamma$ are full-rank, let $\hat{\Delta}_\Gamma \coloneqq \Pi(\hat{\Gamma}) - \Pi(\Gamma)$. If*

$$\|\hat{\Delta}_\Gamma\|_{\text{op}} \leq \left(2(1 + \|R_{c^*}\|_{\text{op}})\right)^{-1} \tag{C.8}$$

1/, then the following first-order approximation holds:

$$\left\|(\hat{B}_{c^*} - B_{c^*}) - R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\right\|_{\text{op}} \leq b\|\hat{\Delta}_\Gamma\|^2_{\text{op}},$$

where $b \coloneqq 12\left(1 + \|R_{c^*}\|_{\text{op}}\right)^3 < \infty$ and R_{c^*} is defined in (9).

Since $\|\hat{\Delta}_\Gamma\|_{\text{op}} = O_p\left(N^{-1/2}\right)$, the probability that $\|\hat{\Delta}_\Gamma\|_{\text{op}} \leq 1/\left(2(1 + \|R_{c^*}\|_{\text{op}})\right)$ converges to one, so for the rest of the proof we also condition on this event. By Lemma C.1,

$$\|\hat{B}_{c^*} - B_{c^*}\|_{\text{op}} \leq \|R_{c^*}\|^2_{\text{op}}\|\hat{\Delta}_\Gamma\|_{\text{op}} + b\|\hat{\Delta}_\Gamma\|^2_{\text{op}} \leq b\left(\|\hat{\Delta}_\Gamma\|_{\text{op}} + \|\hat{\Delta}_\Gamma\|^2_{\text{op}}\right). \tag{C.9}$$

We can also expand (C.7) as follows:

$$\begin{aligned}\hat{\mu}_{c^*} - \mu_{c^*} &= R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\mu_{c^*} + B_{c^*}(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*] \\ &\quad + (\hat{B}_{c^*} - B_{c^*} - R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*})E_{c^*}\mu_{c^*} + (\hat{B}_{c^*} - B_{c^*})(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*]\end{aligned}$$

Rearranging the display above and taking the norm, we have that, again by Lemma C.1,

$$\begin{aligned}&\left\|[\hat{\mu}_{c^*} - \mu_{c^*}] - R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\mu_{c^*} - B_{c^*}(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*]\right\|_2 \\ &\leq \|\hat{B}_{c^*} - B_{c^*} - R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\|_{\text{op}}\|E_{c^*}\mu_{c^*}\|_2 + \|\hat{B}_{c^*} - B_{c^*}\|_{\text{op}}\|(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*]\|_2 \\ &\leq b\|\hat{\Delta}_\Gamma\|^2_{\text{op}}\|E_{c^*}\mu_{c^*}\|_2 \qquad \text{(by Lemma C.1)} \\ &\quad + b\left(\|\hat{\Delta}_\Gamma\|_{\text{op}} + \|\hat{\Delta}_\Gamma\|^2_{\text{op}}\right)\|(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*]\|_2. \qquad \text{(by (C.9))}\end{aligned} \tag{C.10}$$

Next, Theorem 4 implies $\|\hat{\Delta}_\Gamma\|_{\text{op}} = O_p\left(N^{-1/2}\right)$, and under Assumption 4,

$$\|(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*]\|_2 \leq \|(\hat{\mathbb{E}}_N - \mathbb{E})[Y_i^* \mid C_i = c^*]\|_2 = O_p\left(N^{-1/2}\right).$$

Therefore, by (C.10),

$$\sqrt{N}(\hat{\mu}_{c^*} - \mu_{c^*}) = \sqrt{N}R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\mu_{c^*} + \sqrt{N}B_{c^*}(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*] + o_p(1). \tag{C.11}$$

Applying (C.5) to the first term in the expansion (C.11), we have that

$$\begin{aligned}\sqrt{N}\text{vec}(R_{c^*}\hat{\Delta}_\Gamma R'_{c^*}E_{c^*}\mu_{c^*}) &= ((\mu'_{c^*}E_{c^*}R_{c^*}) \otimes R_{c^*}) \cdot \sqrt{N}\text{vec}(\Pi(\hat{\Gamma}) - \Pi(\Gamma)) \\ &= \sqrt{N}\hat{\mathbb{E}}_N\left[((\mu'_{c^*}E_{c^*}R_{c^*}) \otimes R_{c^*})\, H \sum_{c=1}^{C}\phi_c(C_i, Y_i)\right] + o_p(1).\end{aligned}$$

The following lemma, which we prove in Appendix D.3, will help us complete the proof:

Lemma C.2. *Let $V_1, V_2, \ldots$ be a sequence of i.i.d. random variables such that $\mathbb{E}[\|V_1\|_2^2] < \infty$, and let D_i be a binary random variable such that $\mathbb{P}(D_i = 1) > 0$. Then*

$$\sqrt{N}(\hat{\mathbb{E}}_N[V_i \mid D_i = 1] - \mathbb{E}[V_i \mid D_i = 1]) = \sqrt{N}\hat{\mathbb{E}}_N\left[\frac{D_i}{\mathbb{P}(D_i = 1)}(V_i - \mathbb{E}[V_i \mid D_i = 1])\right] + o_p(1).$$

By Lemma C.2, we have that

$$(\hat{\mathbb{E}}_N - \mathbb{E})[E_{c^*}Y_i \mid C_i = c^*] = \hat{\mathbb{E}}_N\left[\frac{\mathbb{1}\{C_i = c^*\}}{\mathbb{P}(C_i = c^*)}(E_{c^*}Y_i - E_{c^*}\mu_{c^*})\right].$$

Substituting the expressions in the two displays above back into (C.11) yields the asymptotically linear expansion in (8). Further, under (6) and (4), $\mathbb{E}[\psi_{c^*}(C_i, Y_i)] = \mathbf{0}_T$ and

$$\mathbb{E}[\|\psi_{c^*i}\|_2^2] \leq \|\mu_{c^*}\|_2 \|R_{c^*}\|_{\text{op}}^2 \|H\|_{\text{op}}^2 \sum_{c=1}^{C} \mathbb{E}[\|\phi_c(C_i, Y_i)\|_2^2] + 2\,\mathbb{E}\left[\|Y_i^*\|_2^2 \mid C_i = c^*\right] < \infty.$$

D Proofs of Intermediate Results in Appendices B and C

D.1 Proof of Theorem B.1

Consider any $\hat{M} \in \mathcal{B}(M)$. By Weyl's inequality (see e.g. Vershynin (2018, Theorem 4.5.3)),

$$\max_j |\lambda_j(M) - \lambda_j(\hat{M})| \leq \|M - \hat{M}\|_{\text{op}} \leq \Delta(M). \tag{D.1}$$

Let

$$a(M) := \begin{cases} \frac{\lambda_{s+1}(M) + \lambda_s(M)}{2}, & s > 0 \\ \lambda_1(M) - 2\Delta(M), & s = 0 \end{cases}, \quad b(M) := \begin{cases} \frac{\lambda_{s+r+1}(M) + \lambda_{s+r}(M)}{2}, & s + r < d \\ \lambda_d(M) + 2\Delta(M), & s + r = d \end{cases}.$$

When $s > 0$, it must be that

$$a(M) - \lambda_s(\hat{M}) = \underbrace{\frac{\lambda_{s+1}(M) - \lambda_s(M)}{2}}_{\geq 2\Delta(M)} + \underbrace{\lambda_s(M) - \lambda_s(\hat{M})}_{\geq -\Delta(M) \text{ by (D.1)}} \geq \Delta(M),$$

and

$$\lambda_{s+1}(\hat{M}) - a(M) = \underbrace{\lambda_{s+1}(\hat{M}) - \lambda_{s+1}(M)}_{\geq -\Delta(M) \text{ by (D.1)}} + \underbrace{\frac{\lambda_{s+1}(M) - \lambda_s(M)}{2}}_{\geq 2\Delta(M)} \geq \Delta(M).$$

When $s = 0$, $a(M) - \lambda_s(\hat{M}) = \infty \geq \Delta(M)$, and

$$\lambda_{s+1}(\hat{M}) - a(M) = \underbrace{\lambda_1(\hat{M}) - \lambda_1(M)}_{\geq -\Delta(M) \text{ by (D.1)}} + 2\Delta(M) \geq \Delta(M).$$

Thus, in both cases, by the definition of $\Delta(M)$ in (B.1),

$$\min\left\{a(M) - \max\{\lambda_s(M), \lambda_s(\hat{M})\}, \min\{\lambda_{s+1}(M), \lambda_{s+1}(\hat{M})\} - a(M)\right\} \geq \Delta(M) > 0. \tag{D.2}$$

Using similar logic, we can also show that

$$\begin{aligned}\min\Big\{&b(M) - \max\{\lambda_{s+r}(M), \lambda_{s+r}(\hat{M})\},\\ &\min\{\lambda_{s+r+1}(M), \lambda_{s+r+1}(\hat{M})\} - b(M)\Big\} \geq \Delta(M) > 0.\end{aligned} \tag{D.3}$$

Next, let $\mathcal{C}$ be some closed, bounded, positively oriented curve in $\mathbb{C}$ that intersects the real line only at $(\lambda_{s+1}(M) + \lambda_s(M))/2$ and $(\lambda_{s+r}(M) + \lambda_{s+r+1}(M))/2$. By (D.2) and (D.3), we have that the $(s+1)$th through $(s+r)$th eigenvalues of both M and $\hat{M}$ are strictly inside $\mathcal{C}$, while all other eigenvalues are strictly outside $\mathcal{C}$. Problem I-5.9 in Kato (1980) (a result first shown in Kato (1949)) then dictates that for $\tilde{M} \in \{M, \hat{M}\}$,

$$\Pi(U_{(s+1):(s+r)}(\tilde{M})) = -\frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} \left(\tilde{M} - \zeta I_d\right)^{-1} d\zeta,$$

so

$$\begin{aligned}&\Pi(U_{(s+1):(s+r)}(\hat{M})) - \Pi(U_{(s+1):(s+r)}(M))\\ &= \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} \left[(M - \zeta I_d)^{-1} - \left(\hat{M} - \zeta I_d\right)^{-1}\right] d\zeta.\end{aligned} \tag{D.4}$$

Considering the integrand in (D.4) in more detail, we can apply the following lemma, which we prove in Appendix D.4:

Lemma D.1. *For any two invertible matrices $M, \hat{M} \in \mathbb{R}^{d\times d}$, we have that*

$$M^{-1} - \hat{M}^{-1} = M^{-1}(\hat{M} - M)M^{-1} - M^{-1}(\hat{M} - M)M^{-1}(\hat{M} - M)\hat{M}^{-1}.$$

Then

$$\begin{aligned}&(M - \zeta I_d)^{-1} - \left(\hat{M} - \zeta I_d\right)^{-1}\\ &= (M - \zeta I_d)^{-1}(\hat{M} - M)(M - \zeta I_d)^{-1}\\ &\quad - (M - \zeta I_d)^{-1}(\hat{M} - M)(M - \zeta I_d)^{-1}(\hat{M} - M)\left(\hat{M} - \zeta I_d\right)^{-1}.\end{aligned} \tag{D.5}$$

Returning to the expression (D.4), expanding the integrand via (D.5), rearranging terms, and

taking the operator norm, we have that

$$\begin{aligned}
&\Big\|\Pi(U_{(s+1):(s+r)}(\hat{M})) - \Pi(U_{(s+1):(s+r)}(M)) \\
&\quad - \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} (M - \zeta I_d)^{-1} (\hat{M} - M) (M - \zeta I_d)^{-1} \, d\zeta \Big\|_{\text{op}} \\
&= \left\| \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} (M - \zeta I_d)^{-1} (\hat{M} - M) (M - \zeta I_d)^{-1} (\hat{M} - M) \left(\hat{M} - \zeta I_d\right)^{-1} d\zeta \right\|_{\text{op}}.
\end{aligned} \tag{D.6}$$

Take $\mathcal{C}$ to be the boundary of a positively oriented rectangular contour on the complex plane with the following corners for some $\upsilon > 0$:

$$a(M) \pm \upsilon\sqrt{-1}, \quad b(M) \pm \upsilon\sqrt{-1}.$$

Then we can bound the operator norm of the integral in (D.6) along each side of $\mathcal{C}$ separately.

To do so, for any $z \in \mathbb{C}$, we define the following decomposition of z into its real and imaginary parts: $z = \text{re}(z) + \text{im}(z)\sqrt{-1}$. We then state the following convenient lemma, which we prove in Appendix D.5:

Lemma D.2. *For any real, symmetric matrix B,*

$$\left\|(B - \zeta I_d)^{-1}\right\|_{op} = \max_j \left((\lambda_j(B) - \text{re}(\zeta))^2 + \text{im}(\zeta)^2\right)^{-1/2}.$$

For notational convenience, we write a and b for $a(M)$ and $b(M)$, respectively. On the horizontal segment $\{x + \upsilon\sqrt{-1} : x \in [a, b]\}$,

$$\begin{aligned}
&\left\| \frac{1}{2\pi\sqrt{-1}} \int_a^b \left(M - (x + \upsilon\sqrt{-1})I_d\right)^{-1} (\hat{M} - M) \left(M - (x + \upsilon\sqrt{-1})I_d\right)^{-1} \right. \\
&\qquad \left. \cdot (\hat{M} - M) \left(\hat{M} - (x + \upsilon\sqrt{-1})I_d\right)^{-1} dx \right\|_{\text{op}} \\
&\leq \frac{\|\hat{M} - M\|_{\text{op}}^2}{2\pi} \cdot \int_a^b \left\|\left(M - (x + \upsilon\sqrt{-1})I_d\right)^{-1}\right\|_{\text{op}}^2 \left\|\left(\hat{M} - (x + \upsilon\sqrt{-1})I_d\right)^{-1}\right\|_{\text{op}} dx \\
&\leq \frac{\|\hat{M} - M\|_{\text{op}}^2}{2\pi} \cdot \int_a^b \underbrace{\left((\lambda_j(M) - x)^2 + \upsilon^2\right)^{-1}}_{\leq \upsilon^{-2}} \underbrace{\left((\lambda_j(\hat{M}) - x)^2 + \upsilon^2\right)^{-1/2}}_{\leq \upsilon^{-1}} dx \text{ (by Lemma D.2)} \\
&\leq \frac{\|\hat{M} - M\|_{\text{op}}^2}{2\pi} \cdot \frac{b - a}{\upsilon^3}.
\end{aligned} \tag{D.7}$$

Similarly, on the horizontal segment $\{x - \upsilon\sqrt{-1} : x \in [a, b]\}$,

$$\left\| \frac{1}{2\pi\sqrt{-1}} \int_b^a \left(M - (x - \upsilon\sqrt{-1})I_d\right)^{-1} (\hat{M} - M) \left(M - (x - \upsilon\sqrt{-1})I_d\right)^{-1} \right.$$

$$\cdot (\hat{M} - M)\left(\hat{M} - (x - v\sqrt{-1})I_d\right)^{-1} dx\Big\|_{\mathrm{op}}$$
$$\leq \frac{\|\hat{M} - M\|_{\mathrm{op}}^2}{2\pi} \cdot \frac{b-a}{v^3}. \tag{D.8}$$

Now we turn to the vertical segment $\{a + y\sqrt{-1} : y \in [-v, v]\}$. By (D.2) and (D.3),

$$|\lambda_j(M) - a| \geq \min\Big\{|\lambda_s(M) - a|, |\lambda_{s+1}(M) - a|\Big\} \geq \Delta(M),$$

so by Lemma D.2,

$$\left\|\left(M - (a + y\sqrt{-1}))I_d\right)^{-1}\right\|_{\mathrm{op}} = \max_j \left((\lambda_j(M) - a)^2 + y^2\right)^{-1/2} \leq \left(\Delta(M)^2 + y^2\right)^{-1/2}.$$

Since (D.2) and (D.3) also imply the bound $|\lambda_j(\hat{M}) - a| \geq \Delta(M)$, Lemma D.2 also implies

$$\left\|\left(\hat{M} - (a + y\sqrt{-1}))I_d\right)^{-1}\right\|_{\mathrm{op}} = \max_j \left((\lambda_j(\hat{M}) - a)^2 + y^2\right)^{-1/2} \leq \left(\Delta(M)^2 + y^2\right)^{-1/2}.$$

Then, we can upper bound the operator norm of the integral on the vertical segment as follows:

$$\begin{aligned}
&\left\|\frac{1}{2\pi\sqrt{-1}}\int_{-v}^{v} \left(M - (a + y\sqrt{-1})I_d\right)^{-1}(\hat{M} - M)\left(M - (a + y\sqrt{-1})I_d\right)^{-1}\right. \\
&\qquad \left. \cdot (\hat{M} - M)\left(\hat{M} - (a + y\sqrt{-1})I_d\right)^{-1} dx\right\|_{\mathrm{op}} \\
&\leq \frac{\|\hat{M} - M\|_{\mathrm{op}}^2}{2\pi}\int_{-v}^{v} \left(\Delta(M)^2 + y^2\right)^{-3/2} dy \\
&= \frac{\|\hat{M} - M\|_{\mathrm{op}}^2}{2\pi}\frac{1}{\Delta(M)^2}\int_{-\infty}^{\infty} \left(1 + u^2\right)^{-3/2} du \\
&= \frac{\|\hat{M} - M\|_{\mathrm{op}}^2}{\pi} \cdot \frac{1}{\Delta(M)^2},
\end{aligned} \tag{D.9}$$

where the equality (D.9) applies the following identity:

$$\int_{-\infty}^{\infty} \left(1 + u^2\right)^{-3/2} du = \int_{-\pi/2}^{\pi/2} \left(1 + \tan^2\theta\right)^{-3/2} d\tan\theta = \int_{-\pi/2}^{\pi/2} \cos\theta d\theta = 2.$$

Similarly, on the vertical segment $\{b + y\sqrt{-1} : y \in [-v, v]\}$,

$$\begin{aligned}
&\left\|\frac{1}{2\pi\sqrt{-1}}\int_{-v}^{v} \left(M - (b + y\sqrt{-1})I_d\right)^{-1}(\hat{M} - M)\left(M - (b + y\sqrt{-1})I_d\right)^{-1}\right. \\
&\qquad \left. \cdot (\hat{M} - M)\left(\hat{M} - (b + y\sqrt{-1})I_d\right)^{-1} dx\right\|_{\mathrm{op}} \\
&\leq \frac{\|\hat{M} - M\|_{\mathrm{op}}^2}{\pi} \cdot \frac{1}{\Delta(M)^2}.
\end{aligned} \tag{D.10}$$

Combining (D.7), (D.8), (D.9), and (D.10), we obtain that

$$\begin{aligned}
&\left\| \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} (M - \zeta I_d)^{-1} (\hat{M} - M) (M - \zeta I_d)^{-1} (\hat{M} - M) \left(\hat{M} - \zeta I_d\right)^{-1} d\zeta \right\|_{\text{op}} \\
&\leq \frac{\|\hat{M} - M\|_{\text{op}}^2}{\pi} \left\{ \frac{b-a}{v^3} + \frac{2}{\Delta(M)^2} \right\}.
\end{aligned}$$

Taking $v \to \infty$, we can combine the bound in the above display with (D.6) to obtain the following bound:

$$\begin{aligned}
&\Bigg\| \Pi(U_{(s+1):(s+r)}(\hat{M})) - \Pi(U_{(s+1):(s+r)}(M)) \\
&\quad - \frac{1}{2\pi\sqrt{-1}} \oint_{\mathcal{C}} (M - \zeta I_d)^{-1} (\hat{M} - M) (M - \zeta I_d)^{-1} d\zeta \Bigg\|_{\text{op}} \\
&\leq \frac{2\|\hat{M} - M\|_{\text{op}}^2}{\pi \Delta(M)^2}.
\end{aligned} \tag{D.11}$$

Next, we consider the first-order error term inside the operator norm in (D.11):

$$\begin{aligned}
&\oint_{\mathcal{C}} (M - \zeta I_d)^{-1} (\hat{M} - M) (M - \zeta I_d)^{-1} d\zeta \\
&= \oint_{\mathcal{C}} U(M)(\Lambda(M) - \zeta I_d)^{-1} U(M)'(\hat{M} - M) U(M)(\Lambda(M) - \zeta I_d)^{-1} U(M)' d\zeta \\
&= \oint_{\mathcal{C}} \left[\sum_{j=1}^{d} (\lambda_j(M) - \zeta)^{-1} \Pi(u_j(M)) \right] (\hat{M} - M) \left[\sum_{j=1}^{d} (\lambda_j(M) - \zeta)^{-1} \Pi(u_j(M)) \right] d\zeta \\
&= \sum_{j=1}^{d} \sum_{k=1}^{d} \Pi(u_j(M))(\hat{M} - M)\Pi(u_k(M)) \oint_{\mathcal{C}} (\lambda_j(M) - \zeta)^{-1} (\lambda_k(M) - \zeta)^{-1} d\zeta
\end{aligned} \tag{D.12}$$

When $j = k$, by Cauchy's integral formula,

$$\oint_{\mathcal{C}} (\lambda_j(M) - \zeta)^{-1} (\lambda_k(M) - \zeta)^{-1} d\zeta = \oint_{\mathcal{C}} \frac{1}{(\lambda_j(M) - \zeta)^2} d\zeta = 0,$$

and when $j \neq k$, we have that

$$\begin{aligned}
&\oint_{\mathcal{C}} (\lambda_j(M) - \zeta)^{-1} (\lambda_k(M) - \zeta)^{-1} d\zeta \\
&= (\lambda_j(M) - \lambda_k(M))^{-1} \\
&\quad \cdot \left(\oint_{\mathcal{C}} (\lambda_k(M) - \zeta)^{-1} d\zeta - \oint_{\mathcal{C}} (\lambda_j(M) - \zeta)^{-1} d\zeta \right) \qquad \text{(Partial Fraction Decomposition)} \\
&= 2\pi\sqrt{-1} (\lambda_j(M) - \lambda_k(M))^{-1} \\
&\quad \cdot (\mathbb{1}\{k \in [s+1, s+r]\} - \mathbb{1}\{j \in [s+1, s+r]\}). \qquad \text{(Residue Theorem)}
\end{aligned}$$

Thus, we can rewrite (D.12) as follows:

$$
\begin{aligned}
&\frac{1}{2\pi\sqrt{-1}}\oint_{\mathcal{C}}(M-\zeta I_d)^{-1}(\hat{M}-M)(M-\zeta I_d)^{-1}\,d\zeta \\
&=\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{\mathbb{1}\{k\in[s+1,s+r]\}-\mathbb{1}\{j\in[s+1,s+r]\}}{\lambda_j(M)-\lambda_k(M)}\Pi(u_j(M))(\hat{M}-M)\Pi(u_k(M)) \\
&=\sum_{j\notin[s+1,s+r]}\sum_{k=s+1}^{s+r}\frac{1}{\lambda_j(M)-\lambda_k(M)}\Pi(u_j(M))(\hat{M}-M)\Pi(u_k(M)) \\
&\quad+\sum_{j=s+1}^{s+r}\sum_{k\notin[s+1,s+r]}\frac{1}{\lambda_k(M)-\lambda_j(M)}\Pi(u_j(M))(\hat{M}-M)\Pi(u_k(M)) \\
&=\sum_{j=s+1}^{s+r}\sum_{k\notin[s+1,s+r]}\frac{1}{\lambda_k(M)-\lambda_j(M)}\Big[\Pi(u_j(M))(\hat{M}-M)\Pi(u_k(M)) \\
&\qquad\qquad+\Pi(u_k(M))(\hat{M}-M)\Pi(u_j(M))\Big].
\end{aligned}
$$

Putting together (D.11) and the final expression in the display above yields (B.2).

D.2 Proof of Lemma C.1

We begin by stating a helpful lemma that expresses our target quantity that is a function of the matrix Γ as a function of $\Pi(\Gamma)$:

Lemma D.3. *For any $T\times r$ matrix $\tilde{\Gamma}$ such that $E_{c^*}\tilde{\Gamma}$ is full-rank,*

$$
\tilde{\Gamma}\left(\tilde{\Gamma}'E_{c^*}\tilde{\Gamma}\right)^{-1}\tilde{\Gamma}'E_{c^*}=\left(I+E_{c^*}\Pi(\tilde{\Gamma})(1-E_{c^*})\right)G_{c^*}(\tilde{\Gamma})\Pi(\tilde{\Gamma})E_{c^*}, \tag{D.13}
$$

where

$$
\begin{aligned}
G_{c^*}(\tilde{\Gamma}) &\coloneqq (I-(I-E_{c^*})\Pi(\tilde{\Gamma})(I-E_{c^*}))^{-1} \\
&= I+(I-E_{c^*})\tilde{\Gamma}(\tilde{\Gamma}'E_{c^*}\tilde{\Gamma})^{-1}\tilde{\Gamma}'(I-E_{c^*}) \\
&=: I+D_{c^*}(\tilde{\Gamma}).
\end{aligned} \tag{D.14}
$$

We provide a proof in Appendix D.6.

Applying Lemma D.3, we can expand the error expression $\hat{\Delta}_B\coloneqq\hat{B}_{c^*}-B_{c^*}$ as follows:

$$
\begin{aligned}
\hat{\Delta}_B &= \hat{\Gamma}(\hat{\Gamma}'E_{c^*}\hat{\Gamma})^{-1}\hat{\Gamma}'E_{c^*}-\Gamma(\Gamma'E_{c^*}\Gamma)^{-1}\Gamma'E_{c^*} \\
&= \left(I+E_{c^*}\Pi(\hat{\Gamma})(1-E_{c^*})\right)G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma})E_{c^*}-\left(I+E_{c^*}\Pi(\Gamma)(1-E_{c^*})\right)G_{c^*}(\Gamma)\Pi(\Gamma)E_{c^*} \\
&= E_{c^*}\hat{\Delta}_\Gamma(I-E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma)E_{c^*}
\end{aligned}
$$

$$+ (I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*}))\,(G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma))E_{c^*}$$
$$+ E_{c^*}\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)(G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma))E_{c^*}. \tag{D.15}$$

We now proceed to construct close approximations to the second and third terms of (D.15), beginning by approximating the expression $G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma)$ that appears in both terms. To do so, we first state the following lemma that provides an expansion for $G_{c^*}(\hat{\Gamma}) - G_{c^*}(\Gamma)$:

Lemma D.4. *Suppose $E_{c^*}\Gamma$ and $E_{c^*}\hat{\Gamma}$ are both full-rank. Let $\hat{\Delta}_G = G_{c^*}(\hat{\Gamma}) - G_{c^*}(\Gamma)$, and let $\ell_{c^*} \coloneqq \|G_{c^*}(\Gamma)\|_{\mathrm{op}}$. If $\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} \le 1/(2\ell_{c^*})$, then*

$$\|\hat{\Delta}_G\|_{\mathrm{op}} \le 2\ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}, \tag{D.16}$$

and

$$\|\hat{\Delta}_G - G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\|_{\mathrm{op}} \le 2\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2. \tag{D.17}$$

We provide a proof in Appendix D.7.

To apply Lemma D.4, we apply (D.14) and the definition of R_{c^*} in (9) as follows:

$$G_{c^*}(\Gamma) = I + (I - E_{c^*})(R_{c^*} - I) = E_{c^*} + (I - E_{c^*})R_{c^*}.$$

By the triangle inequality,

$$\|G_{c^*}(\Gamma)\|_{\mathrm{op}} = \ell_{c^*} \le \|E_{c^*}\|_{\mathrm{op}} + \|R_{c^*}(I - E_{c^*})\|_{\mathrm{op}} \le 1 + \|R_{c^*}\|_{\mathrm{op}}.$$

Thus the condition $\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} \le 1/(2\ell_{c^*})$ in Lemma D.4 holds:

$$\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} \le (2(1 + \|R_{c^*}\|_{\mathrm{op}}))^{-1} \Longrightarrow \|\hat{\Delta}_\Gamma\|_{\mathrm{op}} \le (2\ell_{c^*})^{-1}.$$

Next, since $E_{c^*}\Gamma$ and $E_{c^*}\hat{\Gamma}$ are both full-rank, $G_{c^*}(\Gamma)$ and $G_{c^*}(\hat{\Gamma})$ are well-defined, so

$$G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma) = G_{c^*}(\Gamma)\hat{\Delta}_\Gamma + \hat{\Delta}_G\Pi(\Gamma) + \hat{\Delta}_G\hat{\Delta}_\Gamma. \tag{D.18}$$

Note that $G_{c^*}(\Gamma) \succeq I$ since from (D.14), $G_{c^*}(\tilde{\Gamma})$ takes the form $I + M_1M_2^{-1}M_1$ where $M_1 \coloneqq (I - E_{c^*})\tilde{\Gamma}$ and $M_2 \coloneqq \tilde{\Gamma}'E_c\tilde{\Gamma} = (\tilde{\Gamma}'E_c)(E_c\tilde{\Gamma}) \succeq 0$, meaning $M_2^{-1} = (\tilde{\Gamma}'E_c\tilde{\Gamma})^{-1} \succeq 0$ as well. As such, $\ell_{c^*} \ge 1$. By Lemma D.4 then,

$$\begin{aligned}
&\|G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma)\|_{\mathrm{op}} \\
&\le \ell_{c^*}\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} + \|\hat{\Delta}_G\|_{\mathrm{op}}\|\Pi(\Gamma)\|_{\mathrm{op}} + \|\hat{\Delta}_G\|_{\mathrm{op}}\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} && \text{(by (D.18))} \\
&\le (\ell_{c^*} + 2\ell_{c^*}^2 + 2\ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\mathrm{op}})\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} && \text{(by (D.16))} \\
&\le \ell_{c^*}(1 + 3\ell_{c^*})\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} && \text{(by (C.8))} \\
&\le 4\ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}, && (\ell_{c^*} \ge 1)
\end{aligned} \tag{D.19}$$

and

$$\begin{aligned}
&\Big\| G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma) \\
&\quad - \Big(G_{c^*}(\Gamma)\hat{\Delta}_\Gamma + G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma) \Big) \Big\|_{\mathrm{op}} \\
&\le \|\hat{\Delta}_G\|_{\mathrm{op}}\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} + 2\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2\|\Pi(\Gamma)\|_{\mathrm{op}} && \text{(by (D.17))} \\
&\le 2\ell_{c^*}^2(1+\ell_{c^*})\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2 && \text{(by (D.16))} \\
&\le 4\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2., && (\ell_{c^*} \ge 1) \qquad \text{(D.20)}
\end{aligned}$$

We are now equipped to reconsider the second and third terms in (D.15). First, by (D.20), we can approximate the second term of (D.15) as follows:

$$\begin{aligned}
&\Big\| \left(I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*})\right) \left(G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma)\right)E_{c^*} \\
&\quad - \left(I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*})\right) \Big(G_{c^*}(\Gamma)\hat{\Delta}_\Gamma + G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma) \Big) E_{c^*} \Big\|_{\mathrm{op}} \\
&\le \|I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*})\|_{\mathrm{op}}\|E_{c^*}\|_{\mathrm{op}} \cdot 4\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2 \\
&\le 8\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2. && \text{(D.21)}
\end{aligned}$$

Second, by (D.19), we can bound the third term of (D.15) as follows:

$$\begin{aligned}
&\|E_{c^*}\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)(G_{c^*}(\hat{\Gamma})\Pi(\hat{\Gamma}) - G_{c^*}(\Gamma)\Pi(\Gamma))E_{c^*}\|_{\mathrm{op}} \\
&\le \|\hat{\Delta}_\Gamma\|_{\mathrm{op}} \cdot \ell_{c^*} \cdot 4\ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\mathrm{op}} = 4\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2. && \text{(D.22)}
\end{aligned}$$

Putting together (D.15), (D.21), and (D.22), we obtain the following approximation of $\hat{\Delta}_B$:

$$\begin{aligned}
&\Big\| \hat{\Delta}_B - \Big(E_{c^*}\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma)E_{c^*} \\
&\quad + \left(I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*})\right) G_{c^*}(\Gamma) \Big(\hat{\Delta}_\Gamma + (I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma) \Big) E_{c^*} \Big) \Big\|_{\mathrm{op}} \\
&\le 12\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\mathrm{op}}^2. && \text{(D.23)}
\end{aligned}$$

We now simplify the approximation of $\hat{\Delta}_B$ in (D.23). By (D.14),

$$G_{c^*}(\Gamma)E_{c^*} = (I + D_{c^*}(\Gamma))E_{c^*} = E_{c^*} = E_{c^*}(I + D_{c^*}(\Gamma)) = E_{c^*}G_{c^*}(\Gamma).$$

Thus,

$$\begin{aligned} G_{c^*}(\Gamma)(I - E_{c^*}) = (I - E_{c^*})G_{c^*}(\Gamma) = G_{c^*}(\Gamma) - E_{c^*}, \\ (I - E_{c^*})G_{c^*}(\Gamma)(I - E_{c^*}) = (I - E_{c^*})G_{c^*}(\Gamma). \end{aligned} \tag{D.24}$$

Let

$$\begin{aligned} K_{c^*}(\Gamma) &\coloneqq G_{c^*}(\Gamma)(I - E_{c^*})\Pi(\Gamma)E_{c^*} \\ &= (I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma)E_{c^*} && \text{(by (D.24))} \\ &= (I - E_{c^*})G_{c^*}(\Gamma)(I - E_{c^*})\Pi(\Gamma)E_{c^*} && \text{(by (D.24))} \\ &= (I - E_{c^*})K_{c^*}(\Gamma). \end{aligned} \tag{D.25}$$

Then the approximation of $\hat{\Delta}_B$ in (D.23) simplifies as follows since $G_{c^*}(\Gamma)$ is symmetric:

$$\begin{aligned} &E_{c^*}\hat{\Delta}_\Gamma K_{c^*}(\Gamma) + (G_{c^*}(\Gamma) + K_{c^*}(\Gamma)')\left(\hat{\Delta}_\Gamma E_{c^*} + (I - E_{c^*})\hat{\Delta}_\Gamma K_{c^*}(\Gamma)\right) \\ &= E_{c^*}\hat{\Delta}_\Gamma K_{c^*}(\Gamma) + (G_{c^*}(\Gamma) + K_{c^*}(\Gamma)')\hat{\Delta}_\Gamma E_{c^*} \\ &\quad + (G_{c^*}(\Gamma) - E_{c^*} + K_{c^*}(\Gamma)')\hat{\Delta}_\Gamma K_{c^*}(\Gamma) && \text{(by (D.24) and (D.25))} \\ &= E_{c^*}\hat{\Delta}_\Gamma K_{c^*}(\Gamma) + (G_{c^*}(\Gamma) + K_{c^*}(\Gamma)')\hat{\Delta}_\Gamma E_{c^*} \\ &\qquad + (I + E_{c^*}\Pi(\Gamma)(1 - E_{c^*}))\,(I - E_{c^*})G_{c^*}(\Gamma)\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\Pi(\Gamma)E_{c^*} && \text{(by (D.24))} \\ &= E_{c^*}\hat{\Delta}_\Gamma K_{c^*}(\Gamma) + (G_{c^*}(\Gamma) + K_{c^*}(\Gamma)')\hat{\Delta}_\Gamma E_{c^*} \\ &\quad + (G_{c^*}(\Gamma) - E_{c^*} + K_{c^*}(\Gamma)')\,\hat{\Delta}_\Gamma K_{c^*}(\Gamma) && \text{(by (D.24))} \\ &= (G_{c^*}(\Gamma) + K_{c^*}(\Gamma)')\hat{\Delta}_\Gamma(E_{c^*} + K_{c^*}(\Gamma)). \end{aligned} \tag{D.26}$$

By (D.30) and (D.33),

$$K_{c^*}(\Gamma) = (I - E_{c^*})\Gamma(\Gamma' E_{c^*}\Gamma)^{-1}\Gamma' E_{c^*} = (R_{c^*} - I)' E_{c^*}.$$

By (D.29) and (D.33),

$$G_{c^*}(\Gamma) = I + (I - E_{c^*})\Gamma(\Gamma' E_{c^*}\Gamma)^{-1}\Gamma'(I - E_{c^*}) = I + (I - E_{c^*})(R_{c^*} - I).$$

Thus,

$$G_{c^*}(\Gamma) + K_{c^*}(\Gamma)' = R_{c^*}, \quad E_{c^*} + K_{c^*}(\Gamma) = R'_{c^*}E_{c^*}.$$

Plugging the expressions in the display above into (D.26) and then plugging that expression in turn into (D.23) completes the proof.

D.3 Proof of Lemma C.2

First, note that $\hat{\mathbb{E}}_N[D_iV_i] = \mathbb{E}[D_iV_i] + O_p\left(N^{-1/2}\right)$ and $\hat{\mathbb{E}}_N[D_i] = \mathbb{P}(D_i = 1) + O_p\left(N^{-1/2}\right)$ by a classical multivariate Central Limit Theorem. Then

$$
\begin{aligned}
&\hat{\mathbb{E}}_N[V_i \mid D_i = 1] - \mathbb{E}[V_i \mid D_i = 1] \\
&= \frac{1}{\hat{\mathbb{E}}_N[D_i]}\hat{\mathbb{E}}_N[D_iV_i] - \mathbb{E}[V_i \mid D_i = 1] \\
&= \left(\frac{1}{\hat{\mathbb{E}}_N[D_i]} - \frac{1}{\mathbb{P}(D_i = 1)}\right)\hat{\mathbb{E}}_N[D_iV_i] + \left(\frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} - \mathbb{E}[V_i \mid D_i = 1]\right) \\
&= \frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\hat{\mathbb{E}}_N[D_i]\,\mathbb{P}(D_i = 1)}\hat{\mathbb{E}}_N[D_iV_i] + \left(\frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} - \mathbb{E}[V_i \mid D_i = 1]\right) \\
&= \frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)^2}\hat{\mathbb{E}}_N[D_iV_i] + \left(\frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} - \mathbb{E}[V_i \mid D_i = 1]\right) \\
&\quad + \left(\frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\hat{\mathbb{E}}_N[D_i]\,\mathbb{P}(D_i = 1)} - \frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)^2}\right)\hat{\mathbb{E}}_N[D_iV_i] \\
&= \frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)^2}\,\mathbb{E}[D_iV_i] + \left(\frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} - \mathbb{E}[V_i \mid D_i = 1]\right) \\
&\quad + \underbrace{\frac{\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)^2}\left(\hat{\mathbb{E}}_N[D_iV_i] - \mathbb{E}[D_iV_i]\right)}_{O_p(N^{-1})} \\
&\quad + \left(\frac{1}{\hat{\mathbb{E}}_N[D_i]\,\mathbb{P}(D_i = 1)} - \frac{1}{\mathbb{P}(D_i = 1)^2}\right)\left(\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]\right)\hat{\mathbb{E}}_N[D_iV_i] \\
&= \left(1 - \frac{\hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)}\right)\mathbb{E}[V_i \mid D_i = 1] + \left(\frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} - \mathbb{E}[V_i \mid D_i = 1]\right) \\
&\quad + \underbrace{\left(\mathbb{P}(D_i = 1) - \hat{\mathbb{E}}_N[D_i]\right)^2}_{O_p(N^{-1})}\underbrace{\frac{1}{\hat{\mathbb{E}}_N[D_i]\,\mathbb{P}(D_i = 1)^2}\hat{\mathbb{E}}_N[D_iV_i]}_{O_p(1)} + o_p\left(N^{-1/2}\right) \\
&= -\frac{\hat{\mathbb{E}}_N[D_i]}{\mathbb{P}(D_i = 1)}\,\mathbb{E}[V_i \mid D_i = 1] + \frac{\hat{\mathbb{E}}_N[D_iV_i]}{\mathbb{P}(D_i = 1)} + o_p\left(N^{-1/2}\right) \\
&= \hat{\mathbb{E}}_N[\varphi(D_i, V_i)] + o_p\left(N^{-1/2}\right).
\end{aligned}
$$

Since

$$
\mathbb{E}\left[\varphi(D_i, V_i)\right] = \frac{1}{\mathbb{P}(D_i = 1)}(\mathbb{E}[D_iV_i] - \mathbb{E}[D_iV_i]) = 0,
$$

a classical multivariate Central Limit Theorem implies that

$$\sqrt{N}(\hat{\mathbb{E}}_N[V_i \mid D_i = 1] - \mathbb{E}[V_i \mid D_i = 1]) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\varphi(D_i, V_i)\varphi(D_i, V_i)'\right]\right),$$

and

$$\begin{aligned}
&\mathbb{E}\left[\varphi(D_i, V_i)\varphi(D_i, V_i)'\right] \\
&= \mathbb{E}\left[\frac{D_i}{\mathbb{P}(D_i = 1)^2}\left(V_i - \mathbb{E}[V_i \mid D_i = 1]\right)\left(V_i - \mathbb{E}[V_i \mid D_i = 1]\right)'\right] \\
&= \frac{1}{\mathbb{P}(D_i = 1)}\mathbb{E}\left[\left(V_i - \mathbb{E}[V_i \mid D_i = 1]\right)\left(V_i - \mathbb{E}[V_i \mid D_i = 1]\right)' \;\middle|\; D_i = 1\right] \\
&= \frac{1}{\mathbb{P}(D_i = 1)}\operatorname{Var}(V_i \mid D_i = 1),
\end{aligned}$$

as required.

D.4 Proof of Lemma D.1

$$\begin{aligned}
M^{-1} - \hat{M}^{-1} &= M^{-1}\hat{M}\hat{M}^{-1} - M^{-1}M\hat{M}^{-1} \\
&= M^{-1}(\hat{M} - M)\hat{M}^{-1} \\
&= M^{-1}(\hat{M} - M)\hat{M}^{-1} + M^{-1}(\hat{M} - M)M^{-1} - M^{-1}(\hat{M} - M)M^{-1} \\
&= M^{-1}(\hat{M} - M)M^{-1} - M^{-1}(\hat{M} - M)\left(M^{-1} - \hat{M}^{-1}\right).
\end{aligned} \tag{D.27}$$

Applying (D.27) to the last expression in the display above yields the desired result.

D.5 Proof of Lemma D.2

Since B is symmetric and real, $\lambda_j(B)$ must be real, in which case, letting z^* denote the complex conjugate of $z \in \mathbb{C}$,

$$\begin{aligned}
&\left\|(B - \zeta I_d)^{-1}\right\|_{\text{op}} \\
&= \max_j \left((\lambda_j(B) - \zeta)^*(\lambda_j(B) - \zeta)\right)^{-1/2} \\
&= \max_j \left(\left(\lambda_j(B) - \operatorname{re}(\zeta) - \operatorname{im}(\zeta)\sqrt{-1}\right)\left(\lambda_j(B) - \operatorname{re}(\zeta) + \operatorname{im}(\zeta)\sqrt{-1}\right)\right)^{-1/2} \\
&= \max_j \left(\left(\lambda_j(B) - \operatorname{re}(\zeta)\right)^2 - \operatorname{im}(\zeta)^2 \cdot (-1)\right)^{-1/2},
\end{aligned}$$

which equals the desired result.

D.6 Proof of Lemma D.3

First, we note that for any matrix $\tilde{\Gamma} \in \mathbb{R}^{T\times r}$, we can divide $\Pi(\tilde{\Gamma})$ up into four disjoint sets of entries via left (to select rows) and right (to select columns) multiplications by E_{c^*} and $(I - E_{c^*})$ as follows:

$$\Pi(\tilde{\Gamma}) = (E_{c^*}\tilde{\Gamma} + (I - E_{c^*})\tilde{\Gamma}) \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma} + \tilde{\Gamma}'(I - E_{c^*})\tilde{\Gamma}\right)^{-1} (E_{c^*}\tilde{\Gamma} + (I - E_{c^*})\tilde{\Gamma})'. \tag{D.28}$$

Recall that for any invertible M, the Sherman-Morrison-Woodbury matrix identity states that

$$(M + C'C)^{-1} = M^{-1} - M^{-1}C'(I + CM^{-1}C')^{-1}CM^{-1}.$$

Letting $M = \tilde{\Gamma}' E_{c^*}\tilde{\Gamma}, C = (I - E_{c^*})\tilde{\Gamma}$, we have that

$$\begin{aligned}
&\left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma} + \tilde{\Gamma}'(I - E_{c^*})\tilde{\Gamma}\right)^{-1} \\
&= \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \\
&\quad - \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}'(I - E_{c^*}) \\
&\quad \cdot \left(I + \underbrace{(I - E_{c^*})\tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}'(I - E_{c^*})}_{D_{c^*}(\tilde{\Gamma})} \right)^{-1} (I - E_{c^*})\tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1}.
\end{aligned} \tag{D.29}$$

For the rest of the proof, we write $D_{c^*}(\tilde{\Gamma})$ as D for notational convenience. Then using (D.29) and (D.28), we can write an equation with our quantity of interest on one side:

$$\begin{aligned}
E_{c^*}\Pi(\tilde{\Gamma})E_{c^*} &= \tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}' E_{c^*} \\
&\quad - \underbrace{(I - E_{c^*})\tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}' E_{c^*}}_{K_{c^*}(\tilde{\Gamma})} \\
&\quad - E_{c^*}\tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}'(I - E_{c^*}) (I + D)^{-1} (I - E_{c^*})\tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}' E_{c^*} \\
&= \tilde{\Gamma} \left(\tilde{\Gamma}' E_{c^*}\tilde{\Gamma}\right)^{-1} \tilde{\Gamma}' E_{c^*} - K_{c^*}(\tilde{\Gamma}) - K_{c^*}(\tilde{\Gamma})'(I + D)^{-1} K_{c^*}(\tilde{\Gamma}).
\end{aligned} \tag{D.30}$$

For the rest of the proof, we also write $K_{c^*}(\tilde{\Gamma})$ as K for notational convenience. Next, note that, again by (D.29) and (D.28),

$$\begin{aligned}
(I - E_{c^*})\Pi(\tilde{\Gamma})(I - E_{c^*}) &= D - D(I + D)^{-1}D \\
&= (I + D)(I + D)^{-1}D - D(I + D)^{-1}D \\
&= (I + D - D)(I + D)^{-1}D
\end{aligned}$$

$$
\begin{aligned}
&= (I+D)^{-1}D \\
&= (I+D)^{-1}(I+D-I) \\
&= I-(I+D)^{-1},
\end{aligned} \tag{D.31}
$$

and

$$
E_{c^*}\Pi(\tilde{\Gamma})(I-E_{c^*}) = K' - K'(I+D)^{-1}D = K'(I-(I+D)^{-1}D) = K'(I+D)^{-1}. \tag{D.32}
$$

Therefore,

$$
\begin{aligned}
(I+D)^{-1} &= I-(I-E_{c^*})\Pi(\tilde{\Gamma})(I-E_{c^*}) = G_{c^*}(\tilde{\Gamma})^{-1}, \\
K &= (I+D)(I-E_{c^*})\Pi(\tilde{\Gamma})E_{c^*} \\
&= G_{c^*}(\tilde{\Gamma})(I-E_{c^*})\Pi(\tilde{\Gamma})E_{c^*}.
\end{aligned} \tag{D.33}
$$

Substituting (D.32) and the components of (D.33) into (D.30) and rearranging, we have that

$$
\begin{aligned}
&\tilde{\Gamma}\left(\tilde{\Gamma}'E_{c^*}\tilde{\Gamma}\right)^{-1}\tilde{\Gamma}'E_{c^*} \\
&= E_{c^*}\Pi(\tilde{\Gamma})E_{c^*} + G_{c^*}(\tilde{\Gamma})(I-E_{c^*})\Pi(\tilde{\Gamma})E_{c^*} + E_{c^*}\Pi(\tilde{\Gamma})(I-E_{c^*})G_{c^*}(\tilde{\Gamma})(I-E_{c^*})\Pi(\tilde{\Gamma})E_{c^*},
\end{aligned}
$$

which simplifies to (D.13). By (D.24),

$$
\begin{aligned}
&\tilde{\Gamma}\left(\tilde{\Gamma}'E_{c^*}\tilde{\Gamma}\right)^{-1}\tilde{\Gamma}'E_{c^*} \\
&= E_{c^*}\Pi(\tilde{\Gamma})E_{c^*} + (G_{c^*}(\tilde{\Gamma})-E_{c^*})\Pi(\tilde{\Gamma})E_{c^*} + E_{c^*}\Pi(\tilde{\Gamma})(I-E_{c^*})G_{c^*}(\tilde{\Gamma})\Pi(\tilde{\Gamma})E_{c^*} \\
&= \left(I + E_{c^*}\Pi(\tilde{\Gamma})(I-E_{c^*})\right)G_{c^*}(\tilde{\Gamma})\Pi(\tilde{\Gamma})E_{c^*}.
\end{aligned}
$$

D.7 Proof of Lemma D.4

By (D.27) with $M = G_{c^*}(\Gamma)^{-1}$ and $\hat{M} = G_{c^*}(\hat{\Gamma})^{-1}$,

$$
\hat{\Delta}_G = -G_{c^*}(\Gamma)(G_{c^*}(\hat{\Gamma})^{-1} - G_{c^*}(\Gamma)^{-1})G_{c^*}(\hat{\Gamma}).
$$

By (D.31),

$$
G_{c^*}(\hat{\Gamma})^{-1} = (I + D_{c^*}(\hat{\Gamma}))^{-1} = I - (I-E_{c^*})\Pi(\hat{\Gamma})(I-E_{c^*}),
$$

and

$$
G_{c^*}(\Gamma)^{-1} = (I + D_{c^*}(\Gamma))^{-1} = I - (I-E_{c^*})\Pi(\Gamma)(I-E_{c^*}).
$$

Combining the above expressions, we have that

$$
\begin{aligned}
G_{c^*}(\hat{\Gamma})^{-1} - G_{c^*}(\Gamma)^{-1} &= (I + D_{c^*}(\hat{\Gamma}))^{-1} - (I + D_{c^*}(\Gamma))^{-1} \\
&= -(I-E_{c^*})\hat{\Delta}_\Gamma(I-E_{c^*}).
\end{aligned}
$$

Thus,

$$\hat{\Delta}_G = G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\hat{\Gamma}). \tag{D.34}$$

Next, since $\ell_{c^*} = \|G_{c^*}(\Gamma)\|_{\text{op}}$ from the statement of Lemma D.4,

$$\|G_{c^*}(\hat{\Gamma})\|_{\text{op}} = \|G_{c^*}(\Gamma) + \hat{\Delta}_G\|_{\text{op}} \leq \|G_{c^*}(\Gamma)\|_{\text{op}} + \|\hat{\Delta}_G\|_{\text{op}} = \ell_{c^*} + \|\hat{\Delta}_G\|_{\text{op}}, \tag{D.35}$$

in which case

$$\begin{aligned}
\|\hat{\Delta}_G\|_{\text{op}} &\leq \|G_{c^*}(\Gamma)\|_{\text{op}}\|G_{c^*}(\hat{\Gamma})\|_{\text{op}}\|\hat{\Delta}_\Gamma\|_{\text{op}} && \text{(by (D.34))} \\
&\leq \ell_{c^*}(\ell_{c^*} + \|\hat{\Delta}_G\|_{\text{op}})\|\hat{\Delta}_\Gamma\|_{\text{op}} && \text{(by (D.35))} \\
&\leq \ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\text{op}} + \frac{1}{2}\|\hat{\Delta}_G\|_{\text{op}}.
\end{aligned}$$

where the last line uses the condition that $\|\hat{\Delta}_\Gamma\|_{\text{op}} \leq 1/(2\ell_{c^*})$ from the statement of Lemma D.4. Rearranging the display above implies (D.16).

As a further consequence of this line of reasoning, we have that

$$\begin{aligned}
\|G_{c^*}(\hat{\Gamma})\|_{\text{op}} &\leq \ell_{c^*} + \|\hat{\Delta}_G\|_{\text{op}} && \text{(by (D.35))} \\
&\leq \ell_{c^*} + 2\ell_{c^*}^2\|\hat{\Delta}_\Gamma\|_{\text{op}} && \text{(by (D.16))} \\
&\leq 2\ell_{c^*}. && (\|\hat{\Delta}_\Gamma\|_{\text{op}} \leq 1/(2\ell_{c^*}))
\end{aligned} \tag{D.36}$$

Our desired result then follows from the following display:

$$\begin{aligned}
&\|\hat{\Delta}_G - G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)\|_{\text{op}} \\
&\leq \|G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})\underbrace{(G_{c^*}(\hat{\Gamma}) - G_{c^*}(\Gamma))}_{\hat{\Delta}_G}\|_{\text{op}} && \text{(by (D.34))} \\
&\leq \|G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\Gamma)(I - E_{c^*})\hat{\Delta}_\Gamma(I - E_{c^*})G_{c^*}(\hat{\Gamma})\|_{\text{op}} && \text{(by (D.34))} \\
&\leq \ell_{c^*}^2\|G_{c^*}(\hat{\Gamma})\|_{\text{op}}\|\hat{\Delta}_\Gamma\|_{\text{op}}^2 && (\|G_{c^*}(\Gamma)\|_{\text{op}} = \ell_{c^*}) \\
&\leq 2\ell_{c^*}^3\|\hat{\Delta}_\Gamma\|_{\text{op}}^2. && \text{(by (D.36))}
\end{aligned}$$

E Target Parameters: Examples, Estimation, and Inference

In this appendix, we provide several examples of empirically relevant target parameters θ that researchers can estimate and conduct inference on using our estimator. Then, we describe our Bayesian-bootstrap-based simultaneous inference procedure for θ in detail. Finally, we establish that our plug-in estimator $\hat{\theta}$ described in Section 3 is a consistent and asymptotically normal estimator of θ, and that the Bayesian-bootstrap-based simultaneous confidence intervals described in Section E.2 have valid simultaneous coverage.

E.1 Examples of Target Parameters

Here, we introduce two examples of target parameters $\theta = h(\mu, \eta)$ that aggregate cohort outcome means μ and nuisance parameters η nonlinearly through h as introduced in Section 3:

Example E.1 (Dynamic Treatment Effects)**.** In event study settings, researchers are often interested in reporting dynamic treatment effect paths, i.e. average effects of a treatment across different numbers of time periods relative to units' treatment times. Recall that, in our notation, outcome Y_{it}^* refers to unit i's potential outcome in period t had they not yet been treated by period t (often denoted by $Y_{it}(\infty)$ in this literature), C_i refers to unit i's first treatment period, and the set of observed outcomes for the units treated in period c is $\mathcal{T}_c \subset \{1, \ldots, c-1\}$. To express dynamic treatment effects in the form $h(\mu, \eta)$, we first define Z_{it}^* as unit i's potential outcome in period t had they been treated in period C_i, and, to distinguish between observed and missing outcomes, we let $Z_{it} = Z_{it}^*$ if unit i was treated by period t (i.e. $C_i \leq t$) and $Z_{it} = \emptyset$ otherwise.

Assuming we observe either Y_{it}^* or Z_{it}^* in every period for simplicity,[35] we can write a coordinates of the p-dimensional parameter vector whose entries correspond to the dynamic treatment effects from b periods before treatment through $p - b$ treated periods in the form of Equation (3.4) in Callaway and Sant'Anna (2021) or Equation (26) in L. Sun and Abraham (2021):

$$\theta_{\text{dyn},j} \coloneqq \sum_{t=1}^{T} \sum_{c=1}^{C} \mathbb{1}\{t - c = j - b - 1\} \, \mathbb{P}(C_i = c) \cdot \left(\underbrace{\mathbb{E}[\mathbb{1}\{t \geq c\} Z_{it} + \mathbb{1}\{t < c\} Y_{it} \mid C_i = c]}_{m_{ct}} - \mu_{ct} \right), \tag{E.1}$$

where $j \in \{1, \ldots, p\}$ indexes the coordinates of θ_{dyn}, $t - c$ reflects the index of the current period relative to cohort c's treatment time (with zero corresponding to a cohort's first treated period in relative time), and m_{ct} denotes the average observed outcome for units in cohort c in period t; when period t is before cohort c's treatment time c, the observed outcome is Y_{it} which equals the control potential outcome Y_{it}^*, and when period t is after cohort c's treatment time c, the observed outcome is Z_{it} which equals the treated potential outcome Z_{it}^*. We note that $\theta_{\text{dyn},j}$ can be defined without any restrictions on heterogeneity in treatment effects $Z_{it}^* - Y_{it}^*$ across units or time periods (Callaway and Sant'Anna, 2021; L. Sun and Abraham, 2021).

Example E.2 (Match Outcome Attribution)**.** In the context of bipartite match outcomes, a common exercise is to attribute differences between the outcomes of row-type units matched to

[35] In many settings like Figure 1a's, we do not observe control or treated potential outcome in every period. Callaway and Sant'Anna (2021) and L. Sun and Abraham (2021) discuss several ways of defining estimands that take this additional unbalancedness into account; for brevity, we simply note that these estimands can also be written in the form $h(\mu, \eta)$ and refer the interested reader to Section 3.1.1 of Callaway and Sant'Anna (2021) for details.

two different column-type units to either differences in how column-type units affect match outcomes on average across all row-type units or selection of different row-type units into matches with different column-type units. Analogous to Finkelstein et al. (2016), we note that the difference in average observed match outcomes for row-type units matched to column-type units t_1 and t_2 can be decomposed additively as follows:

$$\begin{aligned}
&\mathbb{E}[Y_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y_{it_1} \mid t_2 \in \mathcal{T}_{C_i}] \\
&= \underbrace{\mathbb{E}[Y^*_{it_1}] - \mathbb{E}[Y^*_{it_2}]}_{\text{Average effect of column-type unit differences on match outcomes}} \\
&+ \underbrace{(\mathbb{E}[Y^*_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y^*_{it_1}]) - (\mathbb{E}[Y^*_{it_2} \mid t_2 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y^*_{it_2}])}_{\text{Average effect on outcomes of differential row-type unit selection into matches with column-type units } t_1 \text{ and } t_2}.
\end{aligned}$$

In line with Finkelstein et al. (2016), we can then define

$$\theta_{\text{col},t_1,t_2} \coloneqq \frac{\mathbb{E}[Y^*_{it_1}] - \mathbb{E}[Y^*_{it_2}]}{\mathbb{E}[Y_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y_{it_1} \mid t_2 \in \mathcal{T}_{C_i}]},$$

as the "share" of the difference in average observed match outcomes between column-type unit t_1 and column-type unit t_2 that can be attributed to differences in how column-type units t_1 and t_2 affect match outcomes on average across all units, and we can define

$$\theta_{\text{row},t_1,t_2} \coloneqq \frac{(\mathbb{E}[Y^*_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y^*_{it_1}]) - (\mathbb{E}[Y^*_{it_2} \mid t_2 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y^*_{it_2}])}{\mathbb{E}[Y_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y_{it_1} \mid t_2 \in \mathcal{T}_{C_i}]}$$

as the "share" of the difference in average observed match outcomes between column-type unit t_1 and column-type unit t_2 that can be attributed to differences in which row-type units select into matches with column-type unit t_1 versus column-type unit t_2. We put "share" in quotes because, like in Finkelstein et al. (2016), θ_{col} and θ_{row} need not lie between zero and one.

We note that the definitions of $\theta_{\text{col},t_1,t_2}$ and $\theta_{\text{row},t_1,t_2}$ above are only defined in terms of outcomes, which is beneficial for interpretation (Hull, 2018), but they coincide exactly with S_{place} and S_{pat} in Finkelstein et al. (2016) when Y^*_{it} is determined by the TWFE model $\mathbb{E}[Y^*_{it} \mid \lambda_i, C_i] = \lambda_i + \gamma_t$:

$$\begin{aligned}
\theta_{\text{col},t_1,t_2} &\propto \gamma_{t_1} - \gamma_{t_2} \\
\theta_{\text{row},t_1,t_2} &\propto \mathbb{E}[\lambda_i \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[\lambda_i \mid t_2 \in \mathcal{T}_{C_i}].
\end{aligned}$$

As discussed in Finkelstein et al. (2016), one could also extend this decomposition by studying analogous differences in average observed match outcomes across both units and groups of outcomes $\mathcal{T} \subset \{1, \ldots, T\}$, i.e. differences in

$$\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{E}[Y_{it} \mid t \in \mathcal{T}_{C_i}]$$

between two groups of column-type units $\mathcal{T} = \check{\mathcal{T}}$ and $\mathcal{T} = \tilde{\mathcal{T}}$. For simplicity however, we restrict our attention to $\theta_{\text{col},t_1,t_2}$ and $\theta_{\text{row},t_1,t_2}$ in our remaining discussion.

E.2 A Bootstrap-Based Simultaneous Inference Procedure for θ

To describe how we construct our Bayesian-bootstrap-based simultaneous confidence intervals, we introduce more notation. Given a vector of N non-negative weights $W \coloneqq (W_1, \ldots, W_N)'$ that sum to one, we will assume that both our cohort-specific factor matrix estimators $\hat{\Gamma}_c$ and our nuisance parameter estimator $\hat{\eta}$ can be adapted to accommodate non-uniform sampling weights W, which we denote $\hat{\Gamma}_c(W)$ and $\hat{\eta}(W)$, respectively. Typically, when $W_1 = \ldots = W_N = \frac{1}{N}$, we have that $\hat{\Gamma}_c(W) = \hat{\Gamma}_c$ and $\hat{\eta}(W) = \hat{\eta}$. For example, to define a weighted version $\hat{\Gamma}_{c,\text{PC}}(W)$ of the PC estimator discussed in Section 4.3, we let $\hat{V}_c(W)$ denote a weighted counterpart of the cohort-specific second moment matrix $\hat{V}_c$ defined in (11) for some vector of weights $W = (W_1, \ldots, W_N)'$ that are non-negative and sum to one:

$$\hat{V}_c(W) \coloneqq N\hat{\mathbb{E}}_N[W_i E_c Y_i Y_i' E_c \mid C_i = c].$$

We then define the weighted PC estimator of $E_c\Gamma$ to be any matrix $\hat{\Gamma}_{c,\text{PC}}(W)$ whose columns are eigenvectors of $\hat{V}_c(W)$ corresponding to $\hat{V}_c(W)$'s r largest eigenvalues.

Next, we let $\hat{A}(W) \coloneqq A(\hat{\Gamma}_1(W), \ldots, \hat{\Gamma}_C(W))$ denote the APM constructed from the weighted, estimated cohort-specific factor matrices $\hat{\Gamma}_1(W), \ldots, \hat{\Gamma}_C(W)$, we let $\hat{\Gamma}(W)$ denote the equivalent of $\hat{\Gamma}$ constructed from the weighted estimated APM $\hat{A}(W)$, and we let $\hat{\mu}_c(W)$ denote the weighted equivalent of $\hat{\mu}_c$ defined as follows:

$$\hat{\mu}_c(W) \coloneqq \hat{\Gamma}(W)\left(E_c\hat{\Gamma}(W)\right)^+ \sum_{i=1}^{N} \frac{W_i \mathbb{1}\{C_i = c\}}{\sum_{j=1}^{N} W_j \mathbb{1}\{C_j = c\}} E_{C_i} Y_i. \tag{E.2}$$

A weighted version of our plug-in estimator of $\hat{\theta}$ is $\hat{\theta}(W) \coloneqq h(\hat{\mu}(W), \hat{\eta}(W))$, where $\hat{\mu}(W)$ is the weighted analog of $\hat{\mu}$. Before continuing, we note that, again, under uniform weights $W_1 = \ldots = W_N = \frac{1}{N}$, all of the quantities defined previously in this paragraph equal their unweighted counterparts.

Given this notation, our inference procedure proceeds as follows. First, for each iteration m of a large number M of repetitions,[36] we take N i.i.d. draws $\xi_{m1}, \ldots, \xi_{mN}$ from the Exponential(1) distribution,[37] construct a vector $W_m \coloneqq (W_{m1}, \ldots, W_{mN})'$ of N normalized weights $W_{mi} \coloneqq \xi_{mi} / \sum_{j=1}^{N} \xi_{mj}$, and compute a weighted target parameter estimate $\hat{\theta}_m^* \coloneqq \hat{\theta}(W_m)$.[38] Next, for

[36] We recommend $M \geq 500$.

[37] Other non-negative distributions are also possible as long as they satisfy regularity conditions; see Section 3.6.2 in van der Vaart and Wellner (1996) for more examples.

[38] Since T is small, each computation of $\hat{\theta}(W_m)$ should be quite fast, as discussed when describing Algorithm 1 above. Further, $\hat{\theta}(W_m)$ can be computed in parallel across iterations m, boosting computational efficiency further.

Algorithm 2: Bayesian Bootstrap Inference

Data: $\{(C_i, Y_i)\}_{i=1}^N$, number of bootstrap samples M.

for $m \in \{1, \ldots, M\}$ **do**
  Sample $\xi_{m1}, \ldots, \xi_{mN} \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$
  Construct weight vector $W_m := (W_{m1}, \ldots, W_{mN})'$ with $W_{mi} := \xi_{mi} / \sum_{j=1}^N \xi_{mj}$
  for $c \in \{1, \ldots, C\}$ **do**
    Compute weighted cohort outcome mean estimate vector $\hat{\mu}_c(W_m)$ as in (E.2) (using Algorithm 1 with weighted analogs)
  end
  Compute weighted nuisance parameter estimates $\hat{\eta}(W_m)$
  Compute weighted target parameter estimate $\hat{\theta}^*_m = h(\hat{\mu}(W_m), \hat{\eta}(W_m))$
end
for $j \in \{1, \ldots, p\}$ **do**
  Compute estimate $\hat{\sigma}_j$ of $\hat{\theta}$'s standard error, e.g. as in (E.3)
end
Compute estimated critical value $\hat{q}_{1-\alpha}$ as in (E.4)
Compute simultaneous $1-\alpha$ confidence intervals $\hat{\mathcal{C}}_j$ as in (E.5)

each coordinate $j \in \{1, \ldots, p\}$ of θ, we compute an estimate $\hat{\sigma}_j$ of the standard error of $\hat{\theta}_j$, e.g.

$$\hat{\sigma}_j := \frac{q_{0.75}(\hat{\theta}^*_{1j}, \ldots, \hat{\theta}^*_{Mj}) - q_{0.25}(\hat{\theta}^*_{1j}, \ldots, \hat{\theta}^*_{Mj})}{q_{0.75}(Z) - q_{0.25}(Z)}, \tag{E.3}$$

where we let $q_\zeta(x_1, \ldots, x_M)$ denote the ζth quantile across scalars $x_1, \ldots, x_M$, and we let $q_\zeta(Z)$ denote the ζth quantile of the standard Gaussian distribution.[39] Finally, we let $\hat{q}_{1-\alpha}$ denote the following estimated critical value:

$$\hat{q}_{1-\alpha} := q_{1-\alpha}(z^*_1, \ldots, z^*_M), \quad z^*_m := \max_{j \in \{1,\ldots,p\}} \left| \hat{\theta}^*_{mj} / \hat{\sigma}_j \right|, \tag{E.4}$$

and we define our simultaneous $1-\alpha$ confidence intervals $\hat{\mathcal{C}}_j$ for θ_j across $j = 1, \ldots, p$ as follows:

$$\hat{\mathcal{C}}_j := \left[\hat{\theta}_j - \hat{q}_{1-\alpha} \hat{\sigma}_j, \hat{\theta}_j + \hat{q}_{1-\alpha} \hat{\sigma}_j \right]. \tag{E.5}$$

For convenience, we summarize the steps of our inference procedure in Algorithm 2.

However, in settings where $T \ll N$ but N and T are both very large, the simplicity of this weighted bootstrap procedure may be outweighed its computational burden. In such cases, one could instead construct a multiplier bootstrap inference procedure like the one proposed in Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017) using the influence function expression given in Theorem 4, Theorem 5, and Corollary E.1.

[39] We suggest this interquartile-range-based estimate of estimator standard errors because it is more robust to outliers than other standard error estimators like the standard deviation over weighted bootstrap draws; see the discussion in Remark 3.2 in Chernozhukov et al. (2013) for details.

E.3 Plug-in Estimator Asymptotic Linearity and Confidence Interval Validity

Having established in Section 4.2 that $\hat{\mu}_c$ is an asymptotically linear estimator of the vector of cohort outcome means μ_c, we now establish asymptotic linearity of our plug-in estimator $\hat{\theta}$ of our target estimand $\theta = h(\mu, \eta)$ and validity of our simultaneous confidence intervals $\hat{\mathcal{C}}_j$. To do so, we assume that nuisance parameter estimator $\hat{\eta}$ is also asymptotically linear:

Assumption E.1. There exists an additional observed random vector $\varrho_i \in \mathbb{R}^q$ measurable with respect to the same probability space as (C_i, Y_i^*) that satisfies $\mathbb{E}[\varrho_i] = \mathbf{0}_q$ and $\mathbb{E}[\|\varrho_i\|_2^2] < \infty$ such that the following expansion holds as $N \to \infty$:

$$\sqrt{N}(\hat{\eta} - \eta) = \sqrt{N}\hat{\mathbb{E}}_N\left[\varrho_i\right] + o_p\left(1\right).$$

In addition, we assume that the function h that defines $\theta = h(\mu, \eta)$ is sufficiently smooth:

Assumption E.2. $h\colon \mathbb{R}^{CT} \times \mathbb{R}^q \to \mathbb{R}^p$ is differentiable at (μ, η).

Under the additional Assumptions E.1 and E.2, our desired result holds:

Corollary E.1. *Suppose that Assumptions 1, 2, 3, 4, E.1, and E.2 hold. Then*

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta) &= \sqrt{N}\hat{\mathbb{E}}_N\left[\varphi(C_i, Y_i, \varrho_i)\right] + o_p\left(1\right), \\ \varphi(C_i, Y_i, \varrho_i) &\coloneqq \sum_{c=1}^{C} \frac{\partial h}{\partial \mu_c'}\psi_c(C_i, Y_i) + \frac{\partial h}{\partial \eta'}\varrho_i, \end{aligned} \tag{E.6}$$

$\mathbb{E}[\varphi(C_i, Y_i, \varrho_i)] = \mathbf{0}_p$, *and* $\mathbb{E}[\|\varphi(C_i, Y_i, \varrho_i)\|_2^2] < \infty$. *Further, as* $N \to \infty$,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_p, \Sigma_\theta\right), \quad \Sigma_\theta \coloneqq \mathbb{E}[\varphi(C_i, Y_i, \varrho_i)\varphi(C_i, Y_i, \varrho_i)']. \tag{E.7}$$

Corollary E.1 follows from a straightforward application of the Delta Method and a classical multivariate Central Limit Theorem.[40]

Before continuing, we note that Σ_θ is not guaranteed to be strictly positive definite. For example, if the target parameter $\theta = a'\mu_{c^*}$ for some T-dimensional vector a and target cohort c^*, and the target cohort's average loadings $\mathbb{E}[\lambda_i \mid C_i = c^*] = \mathbf{0}_r$, then by (8) and (E.6), the asymptotic variance Σ_θ of $\hat{\theta}$ is given by the following quadratic form of at most rank r:

$$\Sigma_\theta = a' \operatorname{Var}(\psi_{c^*}(C_i, Y_i))a = a' \underbrace{\Gamma}_{T\times r} \underbrace{(E_{c^*}\Gamma)^+ \operatorname{Var}\left(E_{c^*}Y_i \mid C_i = c^*\right)\left((E_{c^*}\Gamma)^+\right)'}_{r\times r} \underbrace{\Gamma'}_{r\times T} a.$$

Since the rank r of $\operatorname{Var}(\psi_{c^*}(C_i, Y_i))$ is smaller than its dimension T, there are non-zero vectors a that lie in its non-trivial null space, and thus for those target parameters, $\hat{\theta}$ will have zero

[40] See e.g. Theorem 3.1 in van der Vaart (2000).

asymptotic variance. An implication of this limiting distribution degeneracy is that the confidence intervals $\hat{\mathcal{C}}_j$ defined in (E.5) based on the limiting Gaussian distribution (E.7) will have zero width for some coordinates and thus zero coverage of those coordinates. The following assumption rules out such knife-edge cases:

Assumption E.3. The diagonal entries of Σ_θ are all strictly positive.

To state our result on the validity of the bootstrap-based inference procedure described in Section 3 under this assumption, we define the random vector $\hat{\theta}^* = \hat{\theta}(W)$, where $\hat{\theta}(\cdot)$ is defined in (E.2), W is a random vector of weights with ith coordinate given by $W_i = \xi_i / \sum_{j=1}^N \xi_j$, and $\xi_1, \ldots, \xi_N$ are draws from Exponential(1) independent of both each other and $\hat{\theta}(\cdot)$. Then, under some slight generalizations of Assumptions 3 and E.1 stated in Appendix E.4 for brevity, our inference procedure satisfies the following validity guarantee:

Theorem E.2. *Suppose that Assumptions 1, 2, 4, E.2, and E.3 hold, along with Assumptions E.4 and E.5 stated in Appendix E.4. Then the confidence intervals $\hat{\mathcal{C}}_j$ defined in (E.5) for the coordinates of θ have asymptotic simultaneous coverage at least $1 - \alpha$:*

$$\liminf_{N \to \infty} \mathbb{P}\left(\theta \in \bigtimes_{j=1}^{p} \hat{\mathcal{C}}_j\right) \geq 1 - \alpha.$$

We provide a proof of Theorem E.2 in Appendix E.4.

We now return to Examples E.1 and E.2 and verify that Assumptions E.2 and E.5 (and therefore Assumption E.1) hold:

Example E.1 (continued)**.** From (E.1), we can see that the cohort sizes $\mathbb{P}(C_i = c)$ and cohort-specific observed outcome means m_{ct} are not included in μ and thus form the components of η we must also estimate. Of course, $\mathbb{P}(C_i = c)$ and m_{ct} are identified and consistently estimable via simple averages of the observables $\mathbb{1}\{C_i = c\}$, $\mathbb{1}\{t \geq C_i\} Z_{it}$, and $\mathbb{1}\{t < C_i\} Y_{it}$. Thus, by standard arguments, the components of η expressed as a map from distributions to nuisance parameter values (see Appendix E.4 for details) must be Hadamard differentiable with respect to the distribution over which the expectations in their definitions are taken, implying Assumption E.5 and, by extension, the weaker Assumption E.1 hold. Since $\theta_{\text{dyn},j}$ is linear in products of the components of μ and η, Assumption E.2 immediately holds as well.

Example E.2 (continued)**.** To verify that our target parameters $\theta_{\text{col},t_1,t_2}$ and $\theta_{\text{row},t_1,t_2}$ satisfy the assumptions required for Theorem E.2 to hold, we first note that they are functions of $\mathbb{E}[Y_{it} \mid t \in \mathcal{T}_{C_i}]$, which can be expressed as follows:

$$\mathbb{E}[Y_{it} \mid t \in \mathcal{T}_{C_i}] = \frac{\mathbb{E}[\mathbb{1}\{t \in \mathcal{T}_{C_i}\} Y_{it}]}{\mathbb{E}[\mathbb{1}\{t \in \mathcal{T}_{C_i}\}]} = \frac{\sum_{c=1}^{C} \mathbb{1}\{t \in \mathcal{T}_c\} \mathbb{P}(C_i = c) \mu_{ct}}{\sum_{c=1}^{C} \mathbb{1}\{t \in \mathcal{T}_c\} \mathbb{P}(C_i = c)},$$

as well as outcome means $\mathbb{E}[Y_{it}^*]$, which can be expressed as weighted averages of cohort outcome means:

$$\mathbb{E}[Y_{it}^*] = \sum_{c=1}^{C} \mathbb{P}(C_i = c)\mu_{ct}.$$

As such, the components of the nuisance parameter vector η in this example consist of means of $\mathbb{1}\{t \in \mathcal{T}_{C_i}\}Y_{it}$, means of $\mathbb{1}\{t \in \mathcal{T}_{C_i}\}$, and shares of units in each cohort $\mathbb{P}(C_i = c)$. By standard arguments, these components of η expressed as a map from distributions to nuisance parameter values (again, see Appendix E.4 for details) are Hadamard differentiable with respect to the distributions over which the means are taken, so Assumption E.5 and the weaker Assumption E.1 hold. Further since $\theta_{\text{col},t_1,t_2}$ and $\theta_{\text{row},t_1,t_2}$ can be expressed as ratios of linear combinations of products of the components of μ and η, it is clearly differentiable so long as

$$\mathbb{E}[Y_{it_1} \mid t_1 \in \mathcal{T}_{C_i}] - \mathbb{E}[Y_{it_1} \mid t_2 \in \mathcal{T}_{C_i}] \neq 0.$$

Thus, under this additional condition, Assumption E.2 holds as well.

E.4 Preliminaries for and Proof of Theorem E.2

Preliminaries. Let $\tilde{P}$ denote some distribution over (C_i, Y_i^*), and let $\tilde{P}_{\text{obs}}$ denote the distribution over (C_i, Y_i) implied by $\tilde{P}$. In addition, let $\Gamma(\tilde{P})$ denote the parameter Γ implied by the distribution $\tilde{P}$. Assumption 1 implies that, for each cohort $c = 1, \ldots, C$, there exists a known function $\Pi(\tilde{\Gamma}_c(\tilde{P}_{\text{obs}}))$ mapping distributions over observables $\tilde{P}_{\text{obs}}$ to $\Pi(E_c\Gamma(\tilde{P}))$. We now slightly strengthen Assumption 3 in the following manner:

Assumption E.4. For each cohort $c = 1, \ldots, C$, the map $\Pi(\tilde{\Gamma}_c(\cdot))$ is Hadamard differentiable and satisfies $\Pi(\tilde{\Gamma}_c(\hat{P}_N)) = \Pi(\hat{\Gamma}_c)$.

We note that Assumption E.4 implies Assumption 3 by the Delta method (see e.g. Theorem 3.9.4 in van der Vaart and Wellner (1996)), the fact that $\hat{P}_N$ converges to P, and the fact that ϕ_c is the Riesz representer of the linear Hadamard derivative of $\Pi(\tilde{\Gamma}_c(\cdot))$. As we discuss in Appendix F.2, the PC estimator satisfies Assumption E.4.

Next, we let $\eta(\tilde{P})$ denote the value of the nuisance parameter implied by the distribution $\tilde{P}$, we let $\tilde{\eta}(\tilde{P}_{\text{obs}})$ be the representation of the consistent estimator $\hat{\eta}$ as a map from observable data distributions to estimate values. We can then also slightly strengthen Assumption E.1 in a similar manner:

Assumption E.5. The map $\tilde{\eta}(\cdot)$ is Hadamard differentiable and satisfies $\tilde{\eta}(\hat{P}_N) = \hat{\eta}$.

Finally, we let $\tilde{\theta}(\tilde{P}_{\text{obs}})$ denote the mapping from observed data distributions $\tilde{P}_{\text{obs}}$ to target parameter values such that $\tilde{\theta}(\hat{P}_N) = \hat{\theta}$. Given these definitions and assumptions, we are now equipped to prove Theorem E.2.

Proof. First, we note that Assumption E.4, (C.2) from the proof of Theorem 4, (C.10) from the proof of Theorem 5, and Assumptions E.5 and E.2 allow us to apply the chain rule (see e.g. Lemma 3.9.3 in van der Vaart and Wellner (1996)) to say that the map $\tilde{\theta}(\cdot)$ is itself Hadamard differentiable.

Next, we note that the random weight vector W satisfies Equation (3.6.8) in van der Vaart and Wellner (1996) with $c = 1$ by Example 3.6.9 in the same book. Thus, Theorem 3.6.13 in van der Vaart and Wellner (1996) implies that the Bayesian bootstrap yields a consistent estimate $\hat{P}_N^*$ of the true data-generating distribution P_{obs}, while the bootstrap delta method given in van der Vaart and Wellner (1996)'s Theorem 3.9.11 implies that the asymptotic distribution of $\hat{\theta}^* = \tilde{\theta}(\hat{P}_N^*)$ converges to that of $\hat{\theta}$ in probability. Since quantile functions are also Hadamard differentiable (see Example 3.9.21 in van der Vaart and Wellner (1996)), we have that

$$\sqrt{N}\hat{\sigma}_j = \sqrt{\Sigma_{\theta,jj}} + o_p(1).$$

Given the bootstrap and standard error estimator consistency results above, we can appeal to Proposition 3 and Lemma 1 in the supplemental appendix of Montiel Olea and Plagborg-Møller (2019) to show that our confidence intervals have simultaneous coverage, as required.

F Supplementary Discussions and Results

F.1 Equivalence of Graph-Based Identification Criteria in Panel Data Models and Assumption 2

Several papers in the literature on bipartite match data develop approaches for identifying μ_{ct} (or functions of it) under other models of how unobserved confounders affect outcomes Y_{it}^* that still impose strict exogeneity and fixed-effect-like assumptions (Bonhomme, 2020). To prove identification of μ_{ct} under their alternative models, these papers appeal to the connectedness of different graphs with nodes that represent units and/or outcomes in the nomenclature of this paper. In this section, we show that two types of graphs and accompanying assumptions about their connectedness made in this literature are both equivalent to Assumption 2 when $r = 1$.

First, we consider the identification arguments in Abowd et al. (2002) and Jochmans and Weidner (2019), which study TWFE models of outcomes $\mathbb{E}[Y_{it}^* \mid \lambda_i, C_i] = \gamma_t + \lambda_i$, where γ_t and λ_i are both one-dimensional. These papers condition on a sample of units and the set of their uni-dimensional fixed effects $\{\lambda_i\}_{i=1}^N$, and they show that under appropriate normalizations of the fixed effects, the TWFE regression estimator $\hat{\gamma}_t$ converges to a normalized instance of the corresponding outcome fixed effect $\tilde{\gamma}_t$ at a $N^{-1/2}$ rate as $N \to \infty$ under two assumptions. First, they assume that a finite-population variant of Assumption 4 holds so that a non-vanishing fraction of units have outcome t observed. Second, they define the bipartite graph $\tilde{G}$ consisting

of nodes corresponding to units and outcomes and an edge between any unit i and outcome t for which outcome Y_{it}^* is observed and make the following assumption:

Assumption F.1. The bipartite graph $\tilde{\mathcal{G}}$ is connected.

As it turns out, Assumptions F.1 and 2 are essentially equivalent:

Proposition F.1. *When $r = 1$ and every entry of Γ is non-zero, Assumption F.1 implies Assumption 2. If, in addition, Assumption 4 holds, then Assumption 2 implies Assumption F.1 holds with high probability.*

We provide a proof of Proposition F.1 in Appendix H.1. We note that since when $r = 1$, Γ is a vector, the assumption that every entry of Γ is non-zero is just assuming that the systematic component $\gamma_t \lambda_i$ in the factor model (1) matters for determining every potential outcome.

Second, we consider the identification arguments in Bonhomme et al. (2019), which identifies the distributions of Y_{it} under a model of outcomes with discrete unobserved heterogeneity, and Hull (2018), which identifies differences in μ_{ct} across t under a TWFE-like model of outcomes. Unlike Abowd et al. (2002) and Jochmans and Weidner (2019) and similarly to this paper, these papers model an infinite population of units and, due to their focus on bipartite match outcomes, take seriously the fact that units form one match at a time sequentially. Abstracting away from the time dimension to match the setup in this paper, at their core, Bonhomme et al. (2019) and Hull (2018) base their identification arguments off of variants of the assumption about the connectivity of the graph $\check{\mathcal{G}}$ whose nodes correspond to outcomes and whose edges between two outcomes t_1 and t_2 exist if a positive measure of units have both outcomes t_1 and t_2 observed:

Assumption F.2. The graph $\check{\mathcal{G}}$ is connected.

Similarly to Proposition F.1, we can also show that Assumptions F.2 and 2 are essentially equivalent:

Proposition F.2. *When $r = 1$ and every entry of Γ is non-zero, Assumption F.2 implies Assumption 2. If, in addition, Assumption 4 holds, then Assumption 2 implies Assumption F.2.*

We provide a proof of Proposition F.2 in Appendix H.2.

F.2 Asymptotic Linearity of the Principal Components Estimator

As discussed intuitively in Section 4.3, the Principal Components (PC) estimator of cohort-specific factors can be shown to identify the column space of $E_c\Gamma$ under two assumptions, which we state formally below:

Assumption F.3. For every cohort $c = 1, \ldots, C$, the cohort-specific loading covariance matrix $\mathrm{Var}(\lambda_i \mid C_i = c)$ is positive definite, and $\mathbb{E}[\|Y_i^*\|_4^4 \mid C_i = c] < \infty$.

To introduce the next assumption, let $\varepsilon_i \coloneqq (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ denote the vector of outcome residuals ε_{it}.

Assumption F.4. There exist positive random variables σ_i^2 measurable with respect to the same probability space as (C_i, Y_i^*) such that $\mathbb{E}[\varepsilon_i \varepsilon_i' \mid \sigma_i^2, C_i] = \sigma_i^2 I$ almost surely and $\sigma_c^2 \coloneqq \mathbb{E}[\sigma_i^2 \mid C_i = c] < \infty$ for every cohort $c = 1, \ldots, C$.

To state our formal identification, consistency, and asymptotic linearity results, let

$$s_{1c}^2 \leq \ldots \leq s_{rc}^2 \tag{F.1}$$

denote the smallest through largest eigenvalues of the matrix

$$\Gamma' E_c \Gamma \, \mathbb{E}[\lambda_i \lambda_i' \mid C_i = c].$$

We note that $s_{1c}^2 > 0$ by Assumption F.3 and the fact that $E_c\Gamma$ is full-rank, as shown in Lemma 1. We are now equipped to state our identification result:

Lemma F.3. *Suppose Assumptions 2, F.3, and F.4 hold, and let $\tilde{\Gamma}_{c,PC}$ be any $T \times r$ matrix whose columns are eigenvectors corresponding to the r largest eigenvalues of V_c, defined in (10). Then $\Pi(\tilde{\Gamma}_{c,PC}) = \Pi(E_c\Gamma)$, meaning Assumption 1 holds. Further, V_c's eigenvalues ordered from smallest to largest are $T - |\mathcal{T}_c|$ zeros followed by $|\mathcal{T}_c| - r$ repetitions of σ_c^2 followed by $s_{c1}^2 + \sigma_c^2, \ldots, s_{cr}^2 + \sigma_c^2$.*

We provide a proof of Lemma F.3 in Appendix H.3.

Next, we show that $\Pi(\hat{\Gamma}_{c,\text{PC}})$ is a consistent and asymptotically linear estimator of $\Pi(E_c\Gamma)$. To introduce this result, we let $\tilde{\gamma}_{cj}$ denote an eigenvector of V_c corresponding to the jth smallest eigenvalue of V_c.

Proposition F.4. *Suppose Assumptions 2, 4, F.3, and F.4, hold. Then as $N \to \infty$,*

$$\begin{aligned} \sqrt{N}\text{vec}\left(\Pi(\hat{\Gamma}_{c,\text{PC}}) - \Pi(E_c\Gamma)\right) &= \sqrt{N}\hat{\mathbb{E}}_N\left[\phi_{c,\text{PC}}(C_i, Y_i)\right] + o_p(1), \\ \phi_{c,\text{PC}}(C_i, Y_i) &\coloneqq \frac{\mathbb{1}\{C_i = c\}}{\mathbb{P}(C_i = c)} H_{c,\text{PC}}\text{vec}\left(E_c Y_i Y_i' E_c - V_c\right), \end{aligned} \tag{F.2}$$

where

$$H_{c,\text{PC}} \coloneqq \sum_{j=T-r+1}^{T} \sum_{k=1}^{T-r} \frac{-1}{s_{(j-(T-r))c}^2 + \sigma_c^2 \mathbb{1}\{k \leq T - |\mathcal{T}_c|\}} \left[\Pi(\tilde{\gamma}_{cj}) \otimes \Pi(\tilde{\gamma}_{ck}) + \Pi(\tilde{\gamma}_{ck}) \otimes \Pi(\tilde{\gamma}_{cj})\right],$$

$\mathbb{E}[\phi_{c,\text{PC}}(C_i, Y_i)] = \mathbf{0}_{T^2}$, *and* $\mathbb{E}[\|\phi_{c,\text{PC}}(C_i, Y_i)\|_2^2] < \infty$. *Thus, Assumption 3 holds.*

We provide a proof of Proposition F.4 in Appendix H.4. We note that as a consequence of (H.6), Assumption E.4 holds as well, although for brevity, we do not introduce the additional notation necessary to state such a result formally here.

We conclude this section with a more detailed discussion of why the assumptions needed for Proposition F.4 to hold are slightly weaker than existing consistency and asymptotic linearity results for the PC estimator in the literature of which we are aware. It has been shown that the PC estimator is consistent and asymptotic linear when T remains finite as N grows under Assumptions F.3 and F.4 in the sense that

$$\sqrt{N}\left(\hat{\Gamma}_{c,\mathrm{PC}} - E_c\Gamma\hat{Q}\right) = \hat{\mathbb{E}}_N\left[\tilde{\phi}_{c,\mathrm{PC}}(C_i, Y_i)\right] + o_p(1) \tag{F.3}$$

for some mean-zero influence function $\tilde{\phi}_{c,\mathrm{PC}}$ and random matrix $\hat{Q}$ such that $\hat{Q} \xrightarrow{p} Q$ for some deterministic matrix Q (see e.g. Theorem 5 in J. Bai (2003)).[41] Such results require the eigenvalues $s_{1c}^2, \ldots, s_{rc}^2$ to be distinct.[42]

However, our theory in Section 4.2 only requires assumptions that guarantee the consistency of $\Pi(\hat{\Gamma}_{c,\mathrm{PC}})$ as an estimator of $\Pi(E_c\Gamma)$, which in turn only requires the existence of a random basis matrix $\hat{Q}$ such that $\hat{\Gamma}_{c,\mathrm{PC}}\hat{Q}' = E_c\Gamma + O_p\left(N^{-1/2}\right)$, not that $\hat{Q}$ has a deterministic probability limit, as in (F.3).[43] As such, we do not require distinctness of the eigenvalues $s_{1c}^2, \ldots, s_{rc}^2$ to show that the population equivalent of $\hat{\Gamma}_{c,\mathrm{PC}}$ satisfies Assumption 1, and that $\hat{\Gamma}_{c,\mathrm{PC}}$ itself satisfies Assumptions 3 and E.4.

F.3 A New Approach to Factor Estimation With Uncorrelated, Heteroskedastic Outcomes

If we are only willing to believe that the residuals ε_{it} are uncorrelated across units i and outcomes t but can have arbitrarily heterogeneous variances, then

$$V_c = E_c(\Gamma\,\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]\Gamma' + \Sigma_{\varepsilon,c})E_c,$$

where $\Sigma_{\varepsilon,c} \coloneqq \mathbb{E}[\varepsilon_i\,\varepsilon_i' \mid C_i = c]$ is a diagonal matrix. As such, the argument used to prove Lemma F.3 breaks down, and the PC estimator is inconsistent. Instead, in this appendix, we sketch a computationally efficient spectral procedure for estimating $\Pi(E_c\Gamma)$ that can be applied so long as at least $2r+1$ outcomes are observed per cohort. The procedure is novel, at least to our knowledge.

To describe the approach, for any set of outcome indices $\mathcal{T} \subseteq \{1, \ldots, T\}$, we let $E_{\mathcal{T}}$ denote

[41] Theorem 1 in J. Bai and Ng (2002) does show a similar result to the statement $\sqrt{N}\|\hat{\Gamma}_{c,\mathrm{PC}} - E_c\Gamma\hat{Q}\|_F^2 = O_p(1)$ without requiring $\hat{Q}$ to have a probability limit, but they prove it in the more general case where ε_{it} is allowed to be heteroskedastic and weakly dependent, so in their theorem statement the $\sqrt{N}$ factor is replaced with $\min\{\sqrt{N}, \sqrt{T}\}$ and they require $\min\{N, T\} \to \infty$.

[42] See Assumption G in J. Bai (2003) and Assumption A2(iii) in J. Bai and Ng (2023) for examples of such eigenvalue uniqueness conditions. After introducing Assumption G, J. Bai (2003) notes that such an assumption is not necessary to show that consistent estimators exist for identifiable quantities derived from the factor model.

[43] To see why, note that if such a $\hat{Q}$ did not exist, then for all potentially random basis matrices $\hat{Q}$, $\Pi(\hat{\Gamma}_{c,\mathrm{PC}}) = \Pi(\hat{\Gamma}_{c,\mathrm{PC}}\hat{Q}') \neq \Pi(E_c\Gamma) + O_p\left(N^{-1/2}\right)$, which would be a violation of Proposition F.4.

the diagonal matrix with ones in the diagonal entries corresponding to the indices in $\mathcal{T}$ and zeros elsewhere; for intuition, we note that $E_{\mathcal{T}_c} = E_c$. Given this notation, our approach relies on the fact that, for some size-r "holdout" subset of cohort c's outcomes $\tilde{\mathcal{T}}_1 \subset \mathcal{T}_c$, we can write the matrix $E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} V_c E_{\tilde{\mathcal{T}}_1}$ whose at least $r+1$ non-zero rows correspond to the rows of V_c indexed by $\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1$ and whose r non-zero columns correspond to the columns of V_c indexed by $\tilde{\mathcal{T}}_1$ as follows:

$$E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} V_c E_{\tilde{\mathcal{T}}_1} = E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} \Gamma \, \mathbb{E}[\lambda_i \lambda_i' \mid C_i = c] \Gamma' E_{\tilde{\mathcal{T}}_1}. \tag{F.4}$$

We note that the diagonal residual variance matrix $\Sigma_{\varepsilon,c}$ does not appear in (F.4) because $\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1$ and $\tilde{\mathcal{T}}_1$ are disjoint by construction, meaning $E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} V_c E_{\tilde{\mathcal{T}}_1}$ must have zeros along the diagonal. As such, the column space of the left singular vectors corresponding to the top r singular values of $E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} V_c E_{\tilde{\mathcal{T}}_1}$ identifies $\Pi(E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} \Gamma)$.

Next, we can construct $\tilde{\mathcal{T}}_2$ by swapping one of the outcomes in the holdout set $\tilde{\mathcal{T}}_1$ for an outcome in $\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1$. We can then construct the matrix $E_{\tilde{\mathcal{T}}_2} V_c E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_2}$ and use the span of its left singular vectors corresponding to its r largest singular values to identify $\Pi(E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_2} \Gamma)$ in a similar fashion. Repeating this process $r-2$ more times, we can identify $\Pi(E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_1} \Gamma), \ldots, \Pi(E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_r} \Gamma)$. By construction, $\mathcal{T}_c \setminus \tilde{\mathcal{T}}_j$ and $\mathcal{T}_c \setminus \tilde{\mathcal{T}}_{j+1}$ have r overlapping outcomes for $j = 1, \ldots, r-1$. As such, under an assumption akin to Assumption 2, the null space of another APM $\tilde{A}_c$ defined below identifies the column space of $E_c \Gamma$ by the same logic underlying Theorem 2:

$$\tilde{A}_c \coloneqq \sum_{j=1}^{r} \left[E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_j} - \Pi(E_{\mathcal{T}_c \setminus \tilde{\mathcal{T}}_j} \Gamma) \right].$$

A plug-in estimator based on this strategy should also be asymptotically linear under additional regularity conditions like Assumption F.3 using the same logic underlying Proposition F.4 and Theorem 4, and it can be computed using only $r+1$ eigendecompositions. For brevity, we defer formal proofs of this approach's properties to future work.[44]

G Empirical Illustration: More Details and Results

G.1 Clustering Firms into Types within Each Province

As discussed in Section 5.1, to account for within-province firm heterogeneity in a flexible way, we cluster the between 5,000 and 22,000 firms located in each province into $K = 3$ types using the k-means-based procedure proposed in Bonhomme et al. (2019). In particular, we let F be the number of firms in our sample, we let N_f denote the number of workers who ever worked for firm f in 1998 and 1999, we let $\mathcal{F}_p \subset \{1, \ldots, F\}$ denote the subset of firms located in province p,

[44] Besides formal identification and asymptotic linearity proofs, one avenue to explore could be improving statistical efficiency by aggregating this procedure across multiple holdout set sequences $\tilde{\mathcal{T}}_1, \ldots, \tilde{\mathcal{T}}_r$ at the cost of increased computation.

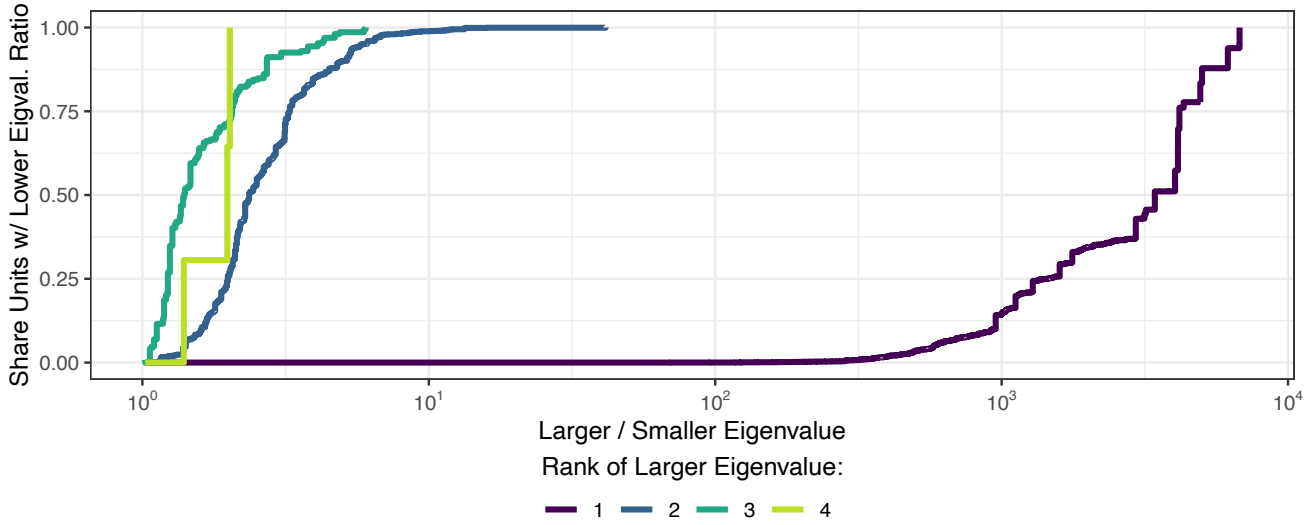


Figure G.2: This figure provides evidence that, in the setting of our empirical illustration in Section 5, the assumption that $r = 1$ is plausible. We compute the empirical outcome second moment matrices $\hat{V}_c$ for every cohort and compute their eigenvalues; in accordance with Lemma F.3, so long as ε_{it} are homoskedastic across outcomes, these eigenvalues are sums of "signals" s^2_{jc} defined in (F.1) and homoskedastic noise variances σ^2_c, where we let $s^2_{jc} = 0$ for $j < 1$. For each eigenvalue indexed from largest to smallest by k, we compute the consecutive eigenvalue ratio $(s^2_{(r-k+1)c} + \sigma^2_c)/(s^2_{(r-k)c} + \sigma^2_c)$ and plot its CDF across cohorts weighted by cohort size. From the figure, it is clear that the first eigenvalue always dominates the others by several orders of magnitude, leading us to assume $r = 1$.

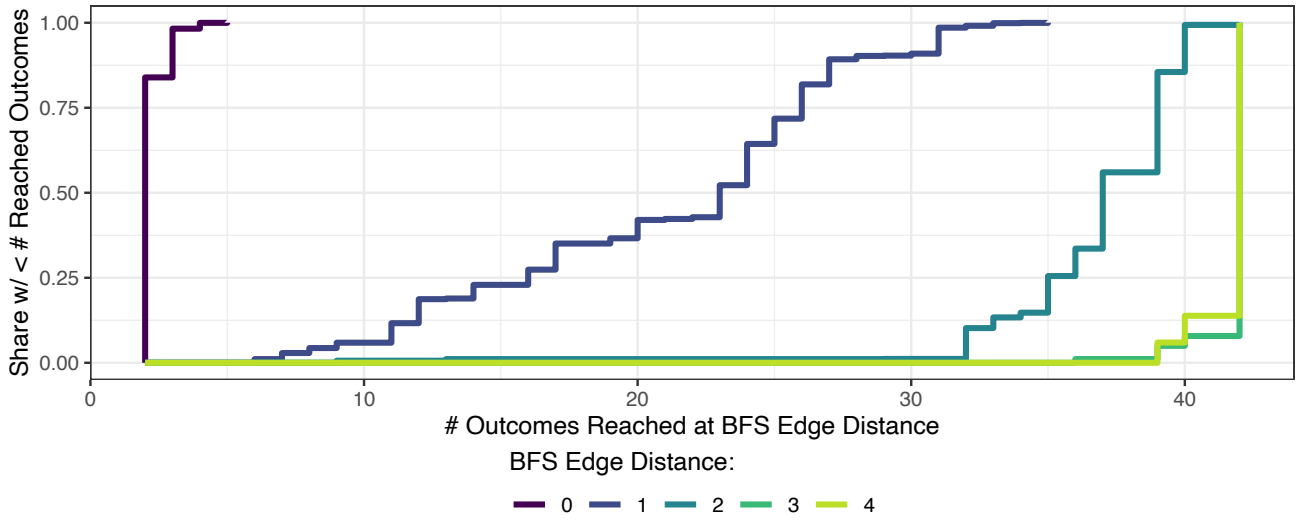


Figure G.3: This figure provides evidence for our empirical application in Section 5 that few embedded block outcome missingness patterns exist in our panel. For each cohort c, we conduct a breadth-first search (BFS) through $\mathcal{G}_1$ starting from node c, counting the edge distances required to reach some other cohort for whom each other outcome is observed. For each edge distance reached during the BFS searches, we plot the cohort-size-weighted distribution across cohorts of the number of unique observed outcomes reached up to that edge distance. Importantly, if an outcome t is not observed for a cohort neighboring c in $\mathcal{G}_1$, then no reference cohort exists for cohort c's outcome t. As a result, according to the CDF for edge distance 1, for 50% of units, half of their cohorts' outcome means cannot be identified using reference-cohort-based methods.

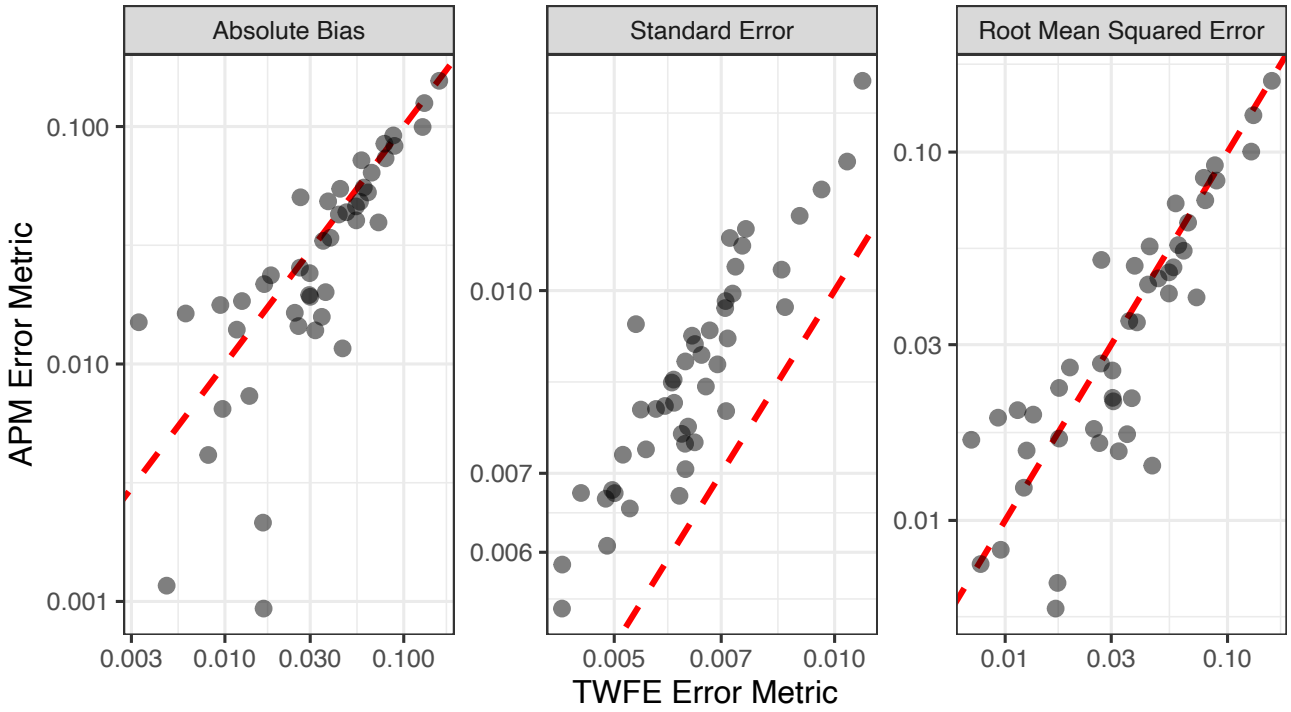


Figure G.4: This figure provides scatter plots of the values of an error metric for our estimator and the TWFE estimator in the context of our empirical illustration in Section 5, where each dot corresponds to one cohort outcome mean target parameter. If a dot in any panel of either subfigure lies below the 45-degree red dashed line, our estimator performs better on that error metric, and if not, then the TWFE estimator performs better on that error metric.

H Proofs of Supplementary Results

H.1 Proof of Proposition F.1

First, we show that when $r = 1$ and every entry of Γ is non-zero, (5) reduces to the requirement that $\mathcal{T}_{c_1} \cap \mathcal{T}_{c_2} \neq \emptyset$, i.e. that cohorts c_1 and c_2 share at least one observed outcome. To see why, note that $E_{c_1}E_{c_2} \neq \mathbf{0}$ whenever cohorts c_1 and c_2 share an observed outcome. Further, when $r = 1$, the factor matrix Γ is a column vector, in which case since, by assumption, every entry of Γ is non-zero, $E_{c_1}E_{c_2}\Gamma \neq \mathbf{0}$ if and only if cohorts c_1 and c_2 share an observed outcome. Thus, the rank of the column vector $E_{c_1}E_{c_2}\Gamma$ is exactly one if and only if cohorts c_1 and c_2 share an observed outcome. Thus, $\mathcal{G}_1$ has an edge between two cohorts if and only if they share at least one observed outcome.

Next, we will show that Assumption F.1 implies that $\mathcal{G}_1$ is connected. Under Assumption F.1, the connectedness of the bipartite $\tilde{\mathcal{G}}$ implies there must exist a sequence of edges with some length ℓ denoted

$$((i_1, t_1), (i_2, t_1), \ldots, (i_{\ell-1}, t_{\ell/2}), (i_\ell, t_{\ell/2})) \tag{H.1}$$

between any two unit i_1 and i_ℓ in $\tilde{\mathcal{G}}$. Consider any pair of edges $(i_j, t), (i_{j+1}, t)$ in the path (H.1) that connect to the same outcome t. If $C_{i_j} \neq C_{i_{j+1}}$, i.e. that units i_j and i_{j+1} belong to different cohorts. Since the edges (i_j, t) and (i_{j+1}, t) both belong to $\tilde{\mathcal{G}}$, by definition, outcome t is observed for i_j and i_{j+1}, in which case cohorts C_{i_j} and $C_{i_{j+1}}$ must share an edge in $\mathcal{G}_1$. If on the other hand $C_{i_j} = C_{i_{j+1}}$, then the same logic would imply there is a self-edge connected to C_{i_j} if not for the fact that (5) only holds for pairs of distinct cohorts.

Based on the argument above, we can iteratively construct a length ℓ path in $\mathcal{G}_1$ corresponding to the path (H.1) in $\tilde{\mathcal{G}}$ that connects cohort C_{i_1} to cohort C_{i_ℓ}. Since connectedness of $\tilde{\mathcal{G}}$ implies that a path between any two units i_1 and i_2 exists in $\tilde{\mathcal{G}}$ and every cohort c must have at least one unit belonging to it, any two cohorts in $\mathcal{G}_1$ must have a path between them. Thus, $\mathcal{G}_1$, must also be connected, so Assumption 2 must also hold.

Finally, we will show that if Assumption 4 also holds, then Assumption 2 implies Assumption F.1 holds with probability approaching one as $N \to \infty$. First, we note that since Assumption 4 requires a unit to belong to each cohort with positive probability and there are a finite number of cohorts, it must be that $\tilde{\mathcal{G}}$ contains a unit belonging to every cohort with probability approaching one as $N \to \infty$. As such, for the remainder of the proof, we shall condition on this event.

Next, under Assumption 2, $\mathcal{G}_1$ is connected, so there exists a sequence of edges with some length ℓ denoted

$$((c_1, c_2), \ldots, (c_{\ell-1}, c_\ell))$$

that connects any two cohorts $c_1 \neq c_\ell$. For a given edge (c_j, c_{j+1}), consider any two units i_j in cohort c_j and i_{j+1} in cohort c_{j+1}. Since the edge (c_j, c_{j+1}) exists in $\mathcal{G}_1$, there must be at least one

outcome t_j that is observed for the units in both cohort c_j and cohort c_{j+1}. As such, there must exist edges (i_j, t_j) and (i_{j+1}, t_j) in $\tilde{\mathcal{G}}$. Based on the argument above, we can iteratively construct a length ℓ path in $\tilde{\mathcal{G}}$ between any unit i_1 in cohort c_1 and any unit i_ℓ in cohort c_ℓ.

To show that any two units i_1 and i_2 in the same cohort c are connected in $\tilde{\mathcal{G}}$, we note that if the units are in the same cohort, the same set of outcomes is observed for both of them, meaning there exists at least one outcome t such that the edges (i_1, t) and (i_2, t) exist. As such, there exists a length two path in $\tilde{\mathcal{G}}$ connecting any two units in the same cohort. In addition, since without loss of generality, every outcome is observed for at least one unit, for any outcome t, there exists at least one edge (i, t) connecting it to some unit i. Putting everything together, we know that there exists a path in $\tilde{\mathcal{G}}$ connecting any two units, regardless of whether they belong to the same or different cohorts, and every outcome is connected to at least one unit. Thus, $\tilde{\mathcal{G}}$ must be connected, as required.

H.2 Proof of Proposition F.2

First, we will show that Assumption F.2 implies Assumption 2. Since $\check{\mathcal{G}}$ is connected, there exists a path in $\check{\mathcal{G}}$ of some length ℓ denoted

$$((t_1, t_2), \ldots, (t_\ell, t_{\ell+1})) \tag{H.2}$$

between any two outcomes t_1 and t_ℓ. If $\ell = 1$, that directly implies the existence of a cohort of units for whom outcomes t_1 and $t_\ell = t_2$ are both observed.

If $\ell \geq 2$ on the other hand, consider any two adjacent edges (t_j, t_{j+1}) and (t_{j+1}, t_{j+2}) in the path (H.2). Since the edge (t_j, t_{j+1}) exists in $\check{\mathcal{G}}$, there must be some cohort of units c_j for whom both outcomes t_j and t_{j+1} are observed. By similar logic, there must be some cohort of units c_{j+1} for whom both outcomes t_{j+1} and t_{j+2} are observed. Since outcome t_{j+1} is observed for the units in cohorts c_j and c_{j+1}, they must then share an edge in $\mathcal{G}_1$ (see the proof of Proposition F.1 in Appendix H.1 to see why we need to assume all entries of Γ are non-zero). Applying this logic iteratively along the path in (H.2), we can construct a path in $\mathcal{G}_1$ from any cohort for whom outcome t_1 is observed to any cohort for whom outcome t_ℓ is observed. Since without loss, at least one outcome is observed for the units in every cohort, we can therefore construct a path in $\mathcal{G}_1$ between any pair of cohorts, meaning $\mathcal{G}_1$ is connected.

Next, we show that Assumption 2 implies Assumption F.2. Since $\mathcal{G}_1$ is connected, there exists a path in $\mathcal{G}_1$ of some length ℓ denoted

$$((c_1, c_2), \ldots, (c_{\ell-1}, c_\ell)) \tag{H.3}$$

between any two cohorts c_1 and c_ℓ. For a given edge (c_j, c_{j+1}) in $\mathcal{G}_1$, by definition, there must be some outcome t_j that is observed for both of the positive measures of units in cohorts c_j and

c_{j+1} by Assumption 4. Thus, for any outcome observed for the units in c_j, there must be an edge in $\check{\mathcal{G}}$ between it and t_j, and similarly, for any outcome observed for the units in c_{j+1}, there must be an edge in $\check{\mathcal{G}}$ between it and t_j. Thus, a path of length two exists in $\check{\mathcal{G}}$ between any outcome observed for the units in c_j and any outcome observed for the units in c_{j+1}. Applying this same logic iteratively along the path (H.3), a path can be constructed in $\check{\mathcal{G}}$ connecting any outcome observed for the units in cohort c_1 to any outcome observed for the units in cohort c_ℓ. Since without loss of generality, every outcome is observed for the units in at least one cohort, a path in $\check{\mathcal{G}}$ can be constructed between any two outcomes. Thus, $\check{\mathcal{G}}$ is connected, as required.

H.3 Proof of Lemma F.3

Let $U_\Gamma S_\Gamma V_\Gamma'$ be a compact singular value decomposition of $E_c\Gamma$, and recall that, by Lemma 1, $E_c\Gamma$ is rank r. Next, note that

$$
\begin{aligned}
V_c &= E_c\,\mathbb{E}[Y_i^*(Y_i^*)' \mid C_i = c]E_c \\
&= E_c\big(\Gamma\,\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]\Gamma' + \mathbb{E}[\varepsilon_i\,\varepsilon_i' \mid C_i = c] \\
&\qquad + \Gamma\,\mathbb{E}[\lambda_i\,\varepsilon_i' \mid C_i = c] + \mathbb{E}[\varepsilon_i\,\lambda_i' \mid C_i = c]\Gamma'\big)E_c && \text{(by (1))} \\
&= E_c\left(\Gamma\,\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]\Gamma' + \mathbb{E}[\varepsilon_i\,\varepsilon_i' \mid C_i = c]\right)E_c && (\mathbb{E}[\varepsilon_i \mid \lambda_i, C_i] = \mathbf{0} \text{ by (1)}) \\
&= E_c\left(\Gamma\,\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]\Gamma' + \underbrace{\mathbb{E}[\sigma_i^2 \mid C_i = c]}_{\sigma_c^2} I\right)E_c && \text{(by Assumption F.4)} \\
&= U_\Gamma \underbrace{S_\Gamma V_\Gamma'\,\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]V_\Gamma S_\Gamma}_{M} U_\Gamma' + \sigma_c^2 E_c. && (E_c\Gamma = U_\Gamma S_\Gamma V_\Gamma')
\end{aligned}
$$

Now, let $U_M S_M^2 U_M'$ denote an eigendecomposition of the $r \times r$ matrix M in the display above, and note that since the $r \times r$ matrix $\mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]$ is rank r by Assumption F.3 and so is $E_c\Gamma$ by Lemma 1, S_M^2 has r non-zero eigenvalues. Then we have that

$$
V_c = U_\Gamma U_M S_M^2 U_M' U_\Gamma' + \sigma_c^2 E_c. \tag{H.4}
$$

Since U_M is an orthonormal matrix and the columns of U_Γ are orthonormal to one another, $U_\Gamma U_M$ must also have r orthonormal columns. Further, we will show that $E_c U_\Gamma = U_\Gamma$, in which case $E_c U_\Gamma U_M = U_\Gamma U_M$. To see why, note that

$$
\begin{aligned}
E_c\Gamma\Gamma'E_c \cdot E_c U_\Gamma &= E_c \cdot E_c\Gamma\Gamma'E_c \cdot U_\Gamma && (E_c \text{ is idempotent}) \\
&= E_c U_\Gamma S_\Gamma V_\Gamma' V_\Gamma S_\Gamma U_\Gamma' U_\Gamma && (E_c\Gamma = U_\Gamma S_\Gamma V_\Gamma') \\
&= E_c U_\Gamma S_\Gamma^2. && (V_\Gamma' V_\Gamma = U_\Gamma' U_\Gamma = I)
\end{aligned}
$$

Thus, $E_c U_\Gamma$ are eigenvectors corresponding to the r non-zero eigenvalues of $E_c\Gamma\Gamma'E_c$, so they

must also be left singular vectors of $E_c\Gamma$.

Next, let $U_{Y1} \coloneqq E_c U_\Gamma U_M$, let U_{Y2} be a $T \times (|\mathcal{T}_c| - r)$ matrix with orthonormal columns such that $U_{Y1}'U_{Y2} = \mathbf{0}$ and $E_c U_{Y2} = U_{Y2}$, i.e. the columns of U_{Y2} are orthogonal to the columns of U_{Y1}, and U_{Y2} only has non-zero entries in the indices $\mathcal{T}_c$, and let U_{Y3} be a $T \times (T - |\mathcal{T}_c|)$ matrix such that

$$U_Y \coloneqq \begin{bmatrix} U_{Y1} & U_{Y2} & U_{Y3} \end{bmatrix} =: \begin{bmatrix} U_{Yc} & U_{Y3} \end{bmatrix}$$

is orthonormal. We note that the existence of the aforementioned matrices is guaranteed constructively by applications of the Gram-Schmidt process. By construction, since $U_Y U_Y' = I$ and U_{Yc} has orthonormal columns with non-zero entries only in the indices $\mathcal{T}_c$, it must be that $U_{Yc}U_{Yc}' = E_c$ and $U_{Y3}U_{Y3}' = I - E_c$. Then, expanding the right side of (H.4) using these matrices, we have that

$$\begin{aligned} V_c &= \begin{bmatrix} U_{Y1} & \begin{bmatrix} U_{Y2} & U_{Y3} \end{bmatrix} \end{bmatrix} \begin{bmatrix} S_M^2 & \mathbf{0}_{r\times(T-r)} \\ \mathbf{0}_{(T-r)\times r} & \mathbf{0}_{(T-r)\times(T-r)} \end{bmatrix} \begin{bmatrix} U_{Y1}' \\ \begin{bmatrix} U_{Y2}' \\ U_{Y3}' \end{bmatrix} \end{bmatrix} \\ &\quad + \begin{bmatrix} U_{Yc} & U_{Y3} \end{bmatrix} \begin{bmatrix} \sigma_c^2 I_{|\mathcal{T}_c|} & \mathbf{0}_{|\mathcal{T}_c|\times(T-|\mathcal{T}_c|)} \\ \mathbf{0}_{(T-|\mathcal{T}_c|)\times|\mathcal{T}_c|} & \mathbf{0}_{(T-|\mathcal{T}_c|)\times(T-|\mathcal{T}_c|)} \end{bmatrix} \begin{bmatrix} U_{Yc}' \\ U_{Y3}' \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} U_{Y1} & U_{Y2} & U_{Y3} \end{bmatrix}}_{U_Y} \begin{bmatrix} S_M^2 + \sigma_c^2 I_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_c^2 I_{|\mathcal{T}_c|-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \underbrace{\begin{bmatrix} U_{Y1}' \\ U_{Y2}' \\ U_{Y3}' \end{bmatrix}}_{U_Y'}. \end{aligned}$$

Since U_Y is an orthonormal matrix, the center matrix in the last line of the display above is diagonal, and the diagonal entries of $S_M^2 + \sigma_c^2 I_r$ are strictly larger than those of $\sigma_c^2 I_{|\mathcal{T}_c|-r}$ from the fact that S_M is positive definite, the expression in the display above must be an eigendecomposition of V_c.

Because the eigenvalues of products of nonsingular square matrices are invariant to cyclic permutations of the product terms,[45] the eigenvalues of the matrices $S_\Gamma V_\Gamma' \mathbb{E}[\lambda_i\lambda_i' \mid C_i = c] V_\Gamma S_\Gamma$ and $V_\Gamma S_\Gamma^2 V_\Gamma' \mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]$ are the same. Since S_M^2 is the diagonal matrix whose non-zero entries are the ordered eigenvalues of $S_\Gamma V_\Gamma' \mathbb{E}[\lambda_i\lambda_i' \mid C_i = c] V_\Gamma S_\Gamma$, and $V_\Gamma S_\Gamma^2 V_\Gamma' = \Gamma' E_c \cdot E_c \Gamma = \Gamma' E_c \Gamma$, we equivalently have that S_M^2 is the diagonal matrix whose non-zero entries are the ordered eigenvalues of $\Gamma' E_c \Gamma \, \mathbb{E}[\lambda_i\lambda_i' \mid C_i = c]$. The results stated in the statement of the lemma then follow from the fact that $\Pi(U_{Y1}) = \Pi(E_c U_\Gamma U_M) = \Pi(U_\Gamma) = \Pi(E_c\Gamma)$ and inspection of the eigendecomposition.

[45] For any square matrices $A, B \in \mathbb{R}^{d\times d}$, if λ is an eigenvalue of AB with corresponding eigenvector v, then since $ABv = \lambda v$, we have that $BA(Bv) = B(ABv) = B(\lambda v) = \lambda(Bv)$. Thus, λ is also an eigenvalue of BA with corresponding eigenvector Bv.

H.4 Proof of Proposition F.4

First, we note that under Assumptions 4 and F.3, a classical multivariate Central Limit Theorem dictates that $\text{vec}(\hat{V}_c - V_c) = O_p\left(N^{-1/2}\right)$, so

$$\begin{aligned}\|\hat{V}_c - V_c\|_{\text{op}} &\leq \|\hat{V}_c - V_c\|_F \\ &= \|\text{vec}(\hat{V}_c - V_c)\|_2 \\ &= O_p\left(N^{-1/2}\right).\end{aligned}$$

Next, we apply Theorem B.1 with $M = V_c$, $\hat{M} = \hat{V}_c$, $s = T - r$, and $r = r$. Lemma F.3 implies that, using the notation from Appendix B,

$$\lambda_{T-r}(V_c) = \sigma_c^2 < s_{1c}^2 + \sigma_c^2 = \lambda_{T-r+1}(V_c) \leq \ldots s_{rc}^2 + \sigma_c^2 = \lambda_T(V_c) < \lambda_{T+1}(V_c) = \infty.$$

As such, we have that $\Delta(V_c) = 4^{-1}s_{1c}^2 > 0$, satisfying (B.1). Lemma F.3 also implies that, for $j \in \{T - r + 1, \ldots, T\}$ and $k \in \{1, \ldots, T - r\}$,

$$\lambda_j - \lambda_k = s_{(j-(T-r))c}^2 + \sigma_c^2 \mathbb{1}\{k \leq T - |\mathcal{T}_c|\}.$$

Further, Lemma F.3 implies that

$$\Pi(E_c\Gamma) = \Pi(U_{(T-r+1):T}(V_c))$$

and we define $\hat{\Gamma}_{c,\text{PC}}$ such that

$$\Pi(\hat{\Gamma}_{c,\text{PC}}) = \Pi(U_{(T-r+1):T}(\hat{V}_c)).$$

Then so long as

$$\|\hat{V}_c - V_c\|_{\text{op}} \leq 4^{-1}s_{1c}^2, \tag{H.5}$$

Theorem B.1 implies that

$$\begin{aligned}&\Big\|\Pi(\hat{\Gamma}_{c,\text{PC}}) - \Pi(E_c\Gamma) \\ &\quad - \sum_{j=T-r+1}^{T}\sum_{k=1}^{T-r} \frac{-1}{s_{(j-(T-r))c}^2 + \sigma_c^2\mathbb{1}\{k \leq T - |\mathcal{T}_c|\}} \\ &\qquad\qquad \cdot \Big[\Pi(\tilde{\gamma}_{cj})(\hat{V}_c - V_c)\Pi(\tilde{\gamma}_{ck}) + \Pi(\tilde{\gamma}_{ck})(\hat{V}_c - V_c)\Pi(\tilde{\gamma}_{cj})\Big]\Big\|_{\text{op}} \\ &\leq \frac{64}{\pi s_{1c}^4}\|\hat{V}_c - V_c\|_{\text{op}}^2.\end{aligned} \tag{H.6}$$

The fact that $\|\hat{V}_c - V_c\|_{\text{op}} = O_p\left(N^{-1/2}\right)$ implies (H.5) holds with probability approaching one

as $N \to \infty$, so by the display above,

$$\begin{aligned}
&\Pi(\hat{\Gamma}_{c,\mathrm{PC}}) - \Pi(E_c\Gamma) \\
&= \sum_{j=T-r+1}^{T} \sum_{k=1}^{T-r} \frac{-1}{s^2_{(j-(T-r))c} + \sigma_c^2 \mathbb{1}\{k \le T - |\mathcal{T}_c|\}} \Big\{ \Pi(\tilde{\gamma}_{cj})(\hat{V}_c - V_c)\Pi(\tilde{\gamma}_{ck}) \\
&\qquad\qquad + \Pi(\tilde{\gamma}_{ck})(\hat{V}_c - V_c)\Pi(\tilde{\gamma}_{cj}) \Big\} \\
&\quad + o_p\left(N^{-1/2}\right).
\end{aligned} \tag{H.7}$$

Applying (C.5) to $\sqrt{N}$ times the vectorization of both sides of (H.7), since $\Pi(\cdot)$ must be symmetric,

$$\begin{aligned}
&\sqrt{N}\mathrm{vec}\left(\Pi(\hat{\Gamma}_{c,\mathrm{PC}}) - \Pi(E_c\Gamma)\right) \\
&= \sum_{j=T-r+1}^{T} \sum_{k=1}^{T-r} \frac{-1}{s^2_{(j-(T-r))c} + \sigma_c^2 \mathbb{1}\{k \le T - |\mathcal{T}_c|\}} \left[\Pi(\tilde{\gamma}_{cj}) \otimes \Pi(\tilde{\gamma}_{ck}) + \Pi(\tilde{\gamma}_{ck}) \otimes \Pi(\tilde{\gamma}_{cj})\right] \\
&\quad \cdot \sqrt{N}\mathrm{vec}\left(\hat{V}_c - V_c\right) + o_p\left(N^{-1/2}\right).
\end{aligned}$$

(F.2), $\mathbb{E}[\phi_c(C_i, Y_i)] = \mathbf{0}_{T^2}$ and the boundedness of the expected squared norm of ϕ_c then follow from the expansion of $\hat{V}_c - V_c$ implied by Lemma C.2, the fact that H_c certainly has bounded operator norm (since $s^2_{(j-(T-r))c} + \sigma_c^2 \mathbb{1}\{k \le T - |\mathcal{T}_c|\} > 0$), and Assumption 4.