Large dimensional latent factor modeling with missing observations and applications to causal inference[☆]

Ruoxuan Xiong [a], Markus Pelger [b],*

[a] *Emory University, Department of Quantitative Theory and Methods, United States of America*
[b] *Stanford University, Department of Management Science & Engineering, United States of America*

ARTICLE INFO

ABSTRACT

This paper develops the inferential theory for latent factor models estimated from large dimensional panel data with missing observations. We propose an easy-to-use all-purpose estimator for a latent factor model by applying principal component analysis to an adjusted covariance matrix estimated from partially observed panel data. We derive the asymptotic distribution for the estimated factors, loadings and the imputed values under an approximate factor model and general missing patterns. The key application is to estimate counterfactual outcomes in causal inference from panel data. The unobserved control group is modeled as missing values, which are inferred from the latent factor model. The inferential theory for the imputed values allows us to test for individual treatment effects at any time under general adoption patterns where the units can be affected by unobserved factors.

1. Introduction

Large dimensional panel data with missing entries are prevalent. In causal panel data, the main focus is to estimate the unobserved potential outcomes. In financial data, stock returns can be missing before a company is listed, after its bankruptcy, or because of illiquidity. In macroeconomic datasets, panel data might be collected at different frequencies or not for all geographical locations resulting in missing entries. In the famous Netflix challenge, a majority of users' ratings for films are missing. Estimating missing entries in panel data is a fundamental problem with applications in social science, statistics, and computer science.

This paper develops the inferential theory for latent factor models estimated from large dimensional panel data with missing observations. We propose a novel and easy-to-use approach to estimate a latent factor model by applying principal component analysis (PCA) to an adjusted covariance matrix, which is estimated from partially observed panel data. We derive the asymptotic normal distribution for the estimated factors, loadings, and imputed values. The key application is to

estimate counterfactual outcomes for causal inference. The unobserved control group is modeled as missing values, which are inferred from the latent factor model. The inferential theory for the imputed values allows us to test for individual treatment effects at a particular time. This granular test is of practical relevance because we learn not only for whom but also when a treatment is effective.

The inferential theory for latent factor models with missing data is important for a number of reasons. First, we show how to consistently impute the missing observations in a large dimensional panel data set, which can then be used as an input for other applications. Our confidence intervals for the imputed values can serve as a decision criterion if the imputed data should be used. Second, the distribution of the missing observations can actually be the object of interest itself. For example, the imputed values serve as the synthetic control in causal inference for which we need an asymptotic distribution theory. The inferential theory is key for deriving test statistics for treatment effects. Last but not least, we provide the complete inferential theory for the latent factors themselves, which is relevant when the factors are the object of interest and are used as input for other applications.

Our method is very simple to adopt and works under general assumptions. We provide an "all-purpose" estimator that performs well under all empirically relevant missing patterns and only assumes a general approximate factor model. Our estimation consists of two simple steps, where we first apply PCA to a re-weighted covariance matrix to obtain the loadings and, in a second step, run a regression on these loadings using only the observed units to obtain the factors. The missing entries are estimated by the common components of the factor model. Importantly, our estimator does not require the estimation of the observation pattern itself. In some cases, we might have additional information about the missing pattern. We provide a modification of our estimator that can take advantage of a probabilistic model of the missing pattern and use an inverse probability weight in the second step regression to obtain the factors. It is inspired by the inverse propensity weighted regression from causal inference that enjoys the doubly-robust property, meaning the estimator is robust to some form of omitted variable bias. Our probability weighted estimator also has similar desirable robustness properties when we omit latent factors, but it is generally less efficient than our all-purpose estimator. Therefore, the propensity weighted estimator can serve as a robustness check if latent factors are omitted, that is, if a researcher uses too few latent factors. If the results of the propensity weighted estimator are close to the "all-purpose" estimator, this can be interpreted as an indication that the latent factor model is correctly specified. In this case, we recommend using the more efficient all-purpose estimator.

Our framework stands out by the very general patterns of missing observations that it can accommodate. We cover the common scenarios of missing at random or a simultaneous/staggered treatment adoption, where the treatment cannot be removed once implemented. Importantly, the missing pattern can depend in a general way on the unobserved factor loadings or unit-specific features. Hence, the observations can be missing because of how the units are exposed to the latent factors. Our simple all-purpose estimator does not require us to explicitly model this relationship, but takes it automatically into account. In the case of the propensity weighted estimator, we provide feasible estimators of the probability weights that result in the same distribution as the population weights.

Deriving the inferential theory under these general conditions is a challenging problem. The missing observations have a complex effect on the asymptotic covariance matrix of the imputed entries. In particular, the asymptotic variance has an additional variance correction term compared with the fully observed panel. This term results in a larger asymptotic variance than in the fully observed case. The variance correction term arises because, in a panel with missing observations, we take averages over a different number of time periods for the different entries in the estimated covariance matrix. The variance correction term is larger if the observation pattern has many missing entries, or if it deviates more from a missing at random scheme. The propensity weighted estimator has a similar asymptotic distribution structure as our all-purpose estimator but in general a larger variance.

Our work contributes to three distinct fields: large dimensional factor modeling, matrix completion, and causal inference. First, we extend the inferential theory of latent factors to large dimensional data with general patterns in missing entries. Second, matrix completion methods impute missing entries under the assumption of a low-rank structure, which is corrupted with noise. We provide confidence intervals for the imputed values. Lastly, the key question in causal inference is the estimation of counterfactual outcomes, i.e., what would have been the outcome if a unit had not been treated or if a unit had been treated. The unobserved counterfactual outcome can naturally be formulated as a missing observation problem. We are the first to provide a test for the point-wise treatment effect that can be heterogeneous and time-dependent under general adoption patterns where the units can be affected by unobserved factors:

This paper works under the framework of an approximate latent factor structure where both the cross-section dimension and time-series dimension are large. When the data is fully observed, Bai and Ng (2002) show that the factor model can be estimated with PCA applied to the covariance matrix of the data. Bai (2003) and Fan et al. (2013) derive the consistency and asymptotic normality of the estimated factors, loadings and common components. Extensions of latent factor models with fully observed data include adding observable factors in Bai (2009), sparse and interpretable latent factors in Pelger and Xiong (2021a), time-varying loadings in Fan et al. (2016) and Pelger and Xiong (2021b), high-frequency estimation in Pelger (2019) and including additional moments to estimate weak factors as in Lettau and Pelger (2020a). When a panel has missing entries, a common approach is to estimate the factor model from a subset of the data for which a balanced panel is available. This approach has two drawbacks: First, it is, in general, less efficient as our approach makes use of all the data. Second, it can lead to a biased estimate if the data is not missing at random.

The inferential theory of large dimensional factor models with missing observations is an active area of research. Our paper is most closely related to the recent papers by Jin et al. (2021), Bai and Ng (2021), and Cahan et al. (2021). The papers

differ in the algorithms to impute the missing observations, the generality of the missing patterns, and the proportion of required observed entries relative to the missing entries. Our main results are derived under the assumption that entries are observed at the same rate as missing entries, but we show that this assumption can be considerably relaxed. Importantly, in contrast to the other papers, our framework allows the missing pattern to depend on unit-specific features and to test for an individual treatment effect at any time for any cross-section unit or a weighted treatment effect. This is exactly what we need for the main application in causal inference. Jin et al. (2021) provide the inferential theory for the estimated factor model with the expectation–maximization (EM) algorithm under the assumption of randomly missing values. This is a major advance in the literature on using the EM algorithm to impute missing values on cross-sectional data (Rubin, 1976; Dempster et al., 1977).[1] Bai and Ng (2021) provide the inferential theory for the factor-based imputed values based on the innovative idea of shuffling rows and columns such that there exist fully observed TALL and WIDE blocks for estimating the factor model. Their TALL-WIDE algorithms involves two applications of principal components on the two fully observed blocks. Cahan et al. (2021) propose the TALL-PROJECT estimator that extends the TALL-WIDE estimator by first using only the fully observed TALL block for a PCA estimation of the factors and then obtains the loadings from a time-series regression that uses all observed entries. They provide the inferential theory for this TALL-PROJECT estimator. Each of these estimators is designed for a specific observation pattern under which it performs particularly well, but might not generalize to other patterns. In contrast, we view our estimator as a simple all-purpose estimator that can reliably impute missing data and provide the correct confidence intervals for general missing patterns and factor structures, which makes it appealing for applied researchers in causal inference. In an extensive simulation study, we show that while our estimator has a similar performance as Jin et al. (2021) for data missing at random, and as Bai and Ng (2021) for missing with a block structure, our estimator can have a better performance for a staggered design or when the observation pattern depends on unit-specific features.

Our imputed values are point-wise consistent and have asymptotic normal distributions, which is relevant for the matrix completion literature that studies a similar problem. Both our paper and the matrix completion literature assume a low-rank structure in the panel data. In the matrix completion literature, the most popular method is to estimate the low-rank matrix from a convex optimization problem using a nuclear norm regularization (Mazumder et al., 2010; Negahban and Wainwright, 2011, 2012). The main results in the matrix completion literature are upper bounds for the mean-squared estimation error of the estimated matrix. However, point-wise consistency does not hold in general because the typically used nuclear norm regularization results in a bias in the estimated matrix. In their path-breaking work, Chen et al. (2019) propose de-biased estimators and provide an inferential theory under the assumption of i.i.d. sampling and i.i.d. noise. There is a trade-off in terms of the generality of the model and the required observations, where our work allows for the most general patterns in missing observations with a general approximate factor structure at the cost of observing entries at a higher rate than Chen et al. (2019). Our paper contributes to the matrix completion literature by allowing general observation patterns and dependent error structures, which is particularly relevant for applications in social science.

Our paper allows for heterogeneous and time-dependent treatment effects of an intervention and more general intervention adoption patterns compared with the synthetic control methods in causal inference. Furthermore, our paper provides a flexible test for treatment effects. In comparative case studies, a key question is to estimate the counterfactual outcomes for treated units. A valid control unit is "close" to the treatment unit except for the treatment effect. Typically synthetic controls are weighted averages of untreated units where the weights depend on unit-specific features. A popular model assumption is that the potential outcome is linear in observed covariates and unobserved common factors. Abadie et al. (2010, 2015), Doudchenko and Imbens (2016), Xu (2017), Li and Bell (2017), Li (2019) propose to match each treated unit by weighted averages of all control units using the pretreatment observations. Li and Bell (2017), Li (2019) and Masini and Medeiros (2018) show the inferential theory for the average treatment effect over time. These methods rely on the assumption that there is only one treated unit and the treatment effects are either constant or stationary. Another method is to regress the post-treatment outcomes for the control units on the pre-treatment outcomes and covariates and use the coefficients to predict the counterfactual outcome for the treated/control units. Athey et al. (2021) propose to use matrix completion methods to impute the control panel data and allow for more general treatment adoption patterns: multiple treated units and staggered treatment adoption. However, they do not provide point-wise guarantees for the imputed values. In this paper, in addition to allowing for general treatment adoption patterns, we also provide the point-wise inferential theory for the imputed counterfactual outcomes. Furthermore, we can test for treatment effects even if they are heterogeneous and time-dependent. Our approach does not require a priori knowledge about which covariates describe if treated and control units are a good match. Instead, our latent loadings capture all unit-specific information in a data-driven way. The synthetic control, that we impute, is a weighted average of the untreated units, that takes all unit-specific information into account. In causal inference, we can either model the relationship between the covariates and the outcome, or model the probabilities of missingness to estimate causal effects. Doubly robust procedures, as discussed, for example, in Kang and Schafer (2007) combine both by using a propensity weight in regressions to mitigate the selection bias. Our propensity weighted estimator builds on this intuition. Interestingly, we prove that using the estimated feasible

---

[1] Stock and Watson (2002), Bańbura and Modugno (2014), Negahban and Wainwright (2012) propose to use EM algorithms to estimate the factor model from panel data with missing observations. Giannone et al. (2008), Doz et al. (2011), Jungbacker et al. (2011), Stock and Watson (2016) propose to use the state-space framework and Kalman Filtering to estimate the factor model with missing observations. Gagliardini et al. (2019) propose a simple diagnostic criterion for an approximate factor structure in large (unbalanced) panel data sets.

(a) Randomly missing   (b) Simultaneous treatment adoption   (c) Staggered treatment adoption

Fig. 1. Examples of patterns for missing observations.
These figures show examples of patterns of missing observations. The shaded entries indicate the missing entries for different $N \times T$ observation matrices W.

propensity instead of the population weights does not affect the asymptotic distribution. This observation is aligned with the results for the classical inverse propensity weighted estimator in Hirano et al. (2003).

The rest of the paper is organized as follows. Section 2 introduces the model and provides the simple all-purpose estimator for factors, loadings, and common components. Section 3 states the necessary assumptions for the asymptotic distribution results that are presented in Section 4. Sections 5 and 6 extend the results to the propensity weighted estimator. Section 7 shows how to apply our model to test treatment effects. We discuss the feasible estimation in Section 8 and how to relax the rate conditions in Section 9. The extensive simulation in Section 10 shows the good finite sample properties, the strong performance relative to other methods, and robustness results under misspecification. The Internet Appendix collects additional simulation results and all proofs.

2. Model and estimation

2.1. Model

Assume we partially observe a panel data set Y with T time periods and N cross-sectional units. $Y \in \mathbb{R}^{N \times T}$ has a factor structure with r common factors. We denote by $F_t \in \mathbb{R}^r$ the latent factors, $\Lambda_i \in \mathbb{R}^r$ the factor loadings, $C_{it} = \Lambda_i^\top F_t$ the common component, and e_{it} the idiosyncratic error:

$$Y_{it} = \Lambda_i^\top F_t + e_{it} \quad \text{for } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T$$

or in vector notation,

$$\underbrace{Y_t}_{N \times 1} = \underbrace{\Lambda}_{N \times r} \underbrace{F_t}_{r \times 1} + \underbrace{e_t}_{N \times 1} \quad \text{for } t = 1, 2, \ldots, T.$$

In an asymptotic setup where N and T are both large, we randomly observe some entries in Y. Let $W_{it} \in \{0, 1\}$ be a binary variable, where $W_{it} = 1$ indicates that the (i, t)-th entry is observed and $W_{it} = 0$ otherwise. In this paper, we will estimate the latent factors F and loadings Λ from the partially observed Y, impute the missing values, and provide the inferential theory for all estimators.

2.2. Missing observations

We allow for very general patterns in the missing observations. Fig. 1 shows three important examples widely seen in empirical applications. The first one is a randomly missing pattern, that is, whether an entry is observed or not does not depend on other entries or observable covariates. For example, the observational pattern of the Netflix challenge is usually modeled as entries missing at random. The second and third ones are the observation patterns for control panels in simultaneous and staggered treatment adoptions. Once a unit adopts the treatment, it stays treated afterward, which will be modeled as missing values. These two patterns are widely assumed in the literature on causal inference in panel data.[2]

$Q_{ij} = \{t : W_{it} = 1 \text{ and } W_{jt} = 1\}$ denotes the set of time periods t when both units i and j are observed. $|Q_{ij}|$ is the cardinality of the set Q_{ij}. Assumption S1 states the conditions on the observation pattern.

---

[2] See Candès and Recht (2009), Zhou et al. (2008) for the Netflix challenge and Athey et al. (2021), Athey and Imbens (2021) for missing patterns used in causal inference.

Assumption S1 (*Observation Pattern*)**.**

1. W is independent of F and e.
2. For a given observation matrix W, $\frac{|\mathcal{Q}_{ij}|}{T} \geq q > 0$ and there exist constants q_{ij} and $q_{ij,kl}$ for all i, j, k, l such that $q_{ij} = \lim_{T \to \infty} \frac{|\mathcal{Q}_{ij}|}{T}$ and $q_{ij,kl} = \lim_{T \to \infty} \frac{|\mathcal{Q}_{ij} \cap \mathcal{Q}_{kl}|}{T}$.

Assumption S1 allows very general observation patterns that can vary over time and depend on unit-specific features. In particular, the observation pattern can depend on the factor loadings that capture cross-sectional information. For the purpose of identification, we assume that the observation pattern is independent of the factors. Note that the estimator of the common components is "symmetric" in N and T, and therefore we could switch the roles of N and T in the above assumptions. In that case, the observation pattern would be independent of the loadings but can depend on the factors. The assumption that the observation pattern is independent of the errors is closely related to the unconfoundedness assumption in Rosenbaum and Rubin (1983). Assumption S1 implicitly assumes that for any two units, the number of time periods when both are observed is proportional to T. This simplifies the presentation of our results and is sufficient for most empirically relevant cases, but we will also discuss how this assumption can be relaxed.

Our framework allows for the following important examples:

1. *Missing at random:* $P(W_{it} = 1) = p$ for all i and t. In this case all units and times are equally likely to be observed.
2. *Cross-section missing at random:* $P(W_{it} = 1) = p_t$. For each t each cross-sectional unit is equally likely to miss.
3. *Time-series missing at random:* $P(W_{it} = 1) = p_i$. For each i each time observation is equally likely to miss.
4. *Cross-section and time-series dependency:* $P(W_{it} = 1) = p_{it}$, which allows for different probabilities for each unit and time.
5. *Staggered treatment adoption:* If $W_{it} = 0$ then $W_{it'} = 0$ for all $t' \geq t$. This is a special case of 4. with $P(W_{it} = 1) = p_{it}$. For the special case that the probability does not depend on i, the staggered design is a special case of cross-section missing at random $P(W_{it} = 1) = p_t$.
6. *Mixed frequency observations:* Each cross-section unit has a fixed known observation pattern over time. This can be modeled as one random draw for each cross-section unit to assign it to a specific pattern. A feasible model approach uses $P(W_{it} = 1) = p_t$ as this is another special case of cross-section missing at random.

We provide an "all-purpose" estimator without the need to explicitly model the probability distribution $P(W_{it} = 1)$. However, in some cases, we might have additional information about the missing pattern. In Section 5, we provide a modification of our estimator that can take advantage of a model for $P(W_{it} = 1)$. More specifically, we allow the cross-sectional observation pattern $P(W_{it} = 1|S_i)$ to depend on observed cross-sectional features $S = [S_i]_{i \in \{1, \cdots, N\}} \in \mathbb{R}^{N \times K}$. These covariates S_i are assumed to be time-invariant. They can be discrete, for example, an indicator variable for gender in the evaluation of a drug treatment or continuous, for example, standardized past test scores in the evaluation of an educational policy change. The cross-sectional features S_i can actually be the estimated latent loadings Λ_i themselves. We discuss how the observation probability $P(W_{it} = 1|S_i) = p_t(S_i)$ can be estimated with parametric or non-parametric estimators. While this modified estimator requires some changes to Assumption S1, it provides the same level of generality for the missing pattern, as discussed in Section 5.

2.3. Estimator

There are two steps to estimate the latent factor model from the partially observed panel data: First, we need to estimate the covariance matrix of the data, and second we estimate the latent factors and loadings based on the eigenvectors of the estimated covariance matrix. The conventional latent factor estimator without missing values applies principal component analysis to the sample covariance matrix. A natural way to deal with the missing values is to set these entries to zero. However, the conventional PCA estimator will then be biased. Our estimator correctly re-weights the entries in the covariance matrix before applying PCA.

We first impute the missing entries by 0 and denote the imputed matrix as $\tilde{Y}$[3]:

$$\tilde{Y}_{it} = Y_{it} W_{it}, \quad \text{for } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T.$$

When some entries in Y are missing, the conventional sample covariance estimator $\frac{1}{T} \tilde{Y} \tilde{Y}^{\top}$ is biased because the actual realizations of the missing values are not equal to zero. We propose a natural estimator of the covariance matrix, where for each entry we only use the time periods when both units are observed. This is equivalent to estimating the sample covariance matrix with $\tilde{Y}$, but reweighting the entries. Table 1 is a simple example to illustrate the covariance matrix estimation if the entries are partly missing in the second half of the data. More generally, our sample covariance matrix estimator equals

$$\tilde{\Sigma}_{ij} = \frac{1}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} Y_{it} Y_{jt}. \tag{1}$$

---

[3] In matrix notation, we have $\tilde{Y} = Y \odot W$, where $\odot$ denotes the Hadamard product.

Table 1

Example of covariance matrix estimation with missing entries.

(a) Observation pattern for Y: Shaded entries are missing.							(b) Sample covariance matrix $\tilde{\Sigma}$: Shaded entries are estimated using observations up to time T_0	
$\mathbf{Y}_{1,1}$	$\cdots$	$\mathbf{Y}_{1,T_0}$	$\mathbf{Y}_{1,T_0+1}$	$\cdots$	$\mathbf{Y}_{1,T}$		$\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbf{Y}_{1,t}\mathbf{Y}_{1,t}^{\top}$	$\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbf{Y}_{1,t}\mathbf{Y}_{2,t}^{\top}$
$\mathbf{Y}_{2,1}$	$\cdots$	$\mathbf{Y}_{2,T_0}$	$\mathbf{Y}_{2,T_0+1}$	$\cdots$	$\mathbf{Y}_{2,T}$		$\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbf{Y}_{2,t}\mathbf{Y}_{1,t}^{\top}$	$\frac{1}{T}\sum_{t=1}^{T}\mathbf{Y}_{2,t}\mathbf{Y}_{2,t}^{\top}$

This table shows an illustrative example for the covariance matrix estimation for Y with missing entries. The missing entries follow a simultaneous treatment adoption pattern. For $t = T_0 + 1, \ldots, T$ the first N_0 cross section units are missing, while the elements $N_0 + 1, \ldots, N$ are observed for all t, i.e. $\mathbf{Y}_{1,t} = \begin{pmatrix} Y_{1,t} & \cdots & Y_{N_0,t} \end{pmatrix}^{\top}$ and $\mathbf{Y}_{2,t} = \begin{pmatrix} Y_{N_0+1,t} & \cdots & Y_{N,t} \end{pmatrix}^{\top}$.

When the data is fully observed, we can apply PCA to $\frac{1}{NT}YY^{\top}$ to estimate the loadings. Up to rescaling, the eigenvectors of the largest eigenvalues estimate the loadings. Then, we regress Y on the estimated loadings to obtain an estimate of the factors.[4] Similarly, for the partially observed data, we apply PCA to $\frac{1}{N}\tilde{\Sigma}$ to estimate the loadings.[5] Under the standard identification assumption $\tilde{\Lambda}^{\top}\tilde{\Lambda}/N = I_r$, we estimate the loadings $\tilde{\Lambda}$ as $\sqrt{N}$ times the eigenvectors of the r largest eigenvalues of the sample covariance matrix, that is

$$\frac{1}{N}\tilde{\Sigma}\tilde{\Lambda} = \tilde{\Lambda}\tilde{D}, \tag{2}$$

where $\tilde{D}$ is a diagonal matrix. Then, for every time period t, we regress the observed Y_t on $\tilde{\Lambda}$ to estimate the factors:

$$\tilde{F}_t = \left(\sum_{i=1}^{N} W_{it}\tilde{\Lambda}_i\tilde{\Lambda}_i^{\top}\right)^{-1}\left(\sum_{i=1}^{N} W_{it}\tilde{\Lambda}_i Y_{it}\right). \tag{3}$$

Interestingly, this very simple estimator automatically corrects for the impact of general observation patterns. If we have additional information that allows us to model the observation pattern as $P(W_{it} = 1|S_i)$, we propose an alternative weighted regression:

$$\tilde{F}_t^S = \left(\sum_{i=1}^{N} \frac{W_{it}}{P(W_{it} = 1|S_i)}\tilde{\Lambda}_i\tilde{\Lambda}_i^{\top}\right)^{-1}\left(\sum_{i=1}^{N} \frac{W_{it}}{P(W_{it} = 1|S_i)}\tilde{\Lambda}_i Y_{it}\right). \tag{4}$$

This conditional estimator uses the weights $\frac{1}{P(W_{it}=1|S_i)}$ in the cross-sectional regression. The estimator for $\tilde{F}_t^S$ is motivated by the inverse propensity score estimator, which is widely used in causal inference.[6] The rationale is that the re-weighted observations correspond to a model where the data is cross-sectionally missing at random. More specifically, after re-weighting the observed data, the loadings should follow the same distribution as in the complete panel without missing observations. This could be relevant if units, that are exposed to specific factors, are more likely to miss. In the special case for cross-sectional missing at random, i.e., $P(W_{it} = 1) = p_t$, the two estimators coincide.[7] We will first study the simple all-purpose estimator $\tilde{F}_t$ and extend it to the propensity weighted estimator $\tilde{F}_t^S$ in Section 5. We show that both estimators are consistent and asymptotically normal. In most cases, $\tilde{F}_t$ is more efficient than the propensity score estimator, but $\tilde{F}_t^S$ can have desirable robustness properties under misspecification. The last step is to estimate the common component $C_{it} = \Lambda_i^{\top}F_t$. We use the plug-in estimator, $\tilde{C}_{it} = \tilde{\Lambda}_i^{\top}\tilde{F}_t$ respectively $\tilde{C}_{it}^S = \tilde{\Lambda}_i^{\top}\tilde{F}_t^S$. If Y_{it} is not observed, we impute the missing values with $\tilde{C}_{it}$ or $\tilde{C}_{it}^S$.

2.4. Illustration

We illustrate in a simple example how missing observations change the conventional PCA estimator with fully observed data. Assume that we have only one factor, and the factor, loading and residual component are i.i.d. normally distributed with $F_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_F^2)$, $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $e_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2)$. We assume that the observations for units $1, \ldots, N_0$ and for

---

[4] Alternatively, we can apply PCA to $\frac{1}{NT}YY^{\top}$ to estimate the loadings and then regress $Y^{\top}$ on the estimated loadings to estimate the factors. Bai and Ng (2002), Bai (2003) show that the estimated factors and loadings from this approach are consistent and asymptotically normally distributed for a fully observed panel.

[5] We assume that the true number of factors is r, and has been consistently estimated as in Bai (2003). See Section 3 for further discussions.

[6] Since Horvitz and Thompson (1952), weighting observations by inverse probability has been frequently used to account for missing data in mean estimation (David et al., 1983; Little, 1986, 1988), regression analysis (Robins et al., 1994; Robins and Rotnitzky, 1995), and causal inference (Hirano et al., 2003).

[7] The cross-sectional reweighting is only affected by cross-sectional differences in the probability p_{it} at time t. Therefore we model only the dependency on a set of cross-sectional covariates S_i, but allow this dependency to be time-varying. We could without loss of generality consider the dependency on cross-sectional and time-varying covariates S_{it}. However, this would not change the cross-sectional reweighting at a specific time t, but would come at the cost of a more complicated notation.

the times $T_0 + 1, \ldots, T$ are missing according to the simultaneous adoption pattern of Table 1. We separate the vector of factor realizations into its first $\mathbf{F}_1 = \begin{pmatrix} F_1 & \cdots & F_{T_0} \end{pmatrix}^\top$ and second part $\mathbf{F}_2 = \begin{pmatrix} F_{T_0+1} & \cdots & F_T \end{pmatrix}^\top$ and similarly for the loadings $\mathbf{\Lambda}_1 = \begin{pmatrix} \Lambda_1 & \cdots & \Lambda_{N_0} \end{pmatrix}^\top$ and $\mathbf{\Lambda}_2 = \begin{pmatrix} \Lambda_{N_0+1} & \cdots & \Lambda_N \end{pmatrix}^\top$. Note that in this simple example $\tilde{F}$ and $\tilde{F}^S$ coincide.

We start with the simplest case without error terms e_t to illustrate the logic of reweighting entries. In this case the conventional covariance matrix equals

$$
\frac{1}{T}\tilde{Y}\tilde{Y}^\top = \frac{1}{T}\begin{pmatrix} \mathbf{\Lambda}_1\mathbf{F}_1^\top & 0 \\ \mathbf{\Lambda}_2\mathbf{F}_1^\top & \mathbf{\Lambda}_2\mathbf{F}_2^\top \end{pmatrix}\begin{pmatrix} \mathbf{F}_1\mathbf{\Lambda}_1^\top & \mathbf{F}_1\mathbf{\Lambda}_2^\top \\ 0 & \mathbf{F}_2\mathbf{\Lambda}_2^\top \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{T_0}{T}}\mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{pmatrix}\left(\sigma_F^2 + o_P(1)\right)\begin{pmatrix} \sqrt{\frac{T_0}{T}}\mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_2^\top \end{pmatrix}.
$$

Obviously, the eigenvector of this matrix is a biased estimate of the loadings. In contrast, the eigenvector of the correctly weighted sample covariance matrix consistently estimates the loadings:

$$
\tilde{\Sigma} = \begin{pmatrix} \mathbf{\Lambda}_1\frac{\mathbf{F}_1^\top\mathbf{F}_1}{T_0}\mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_1\frac{\mathbf{F}_1^\top\mathbf{F}_1}{T_0}\mathbf{\Lambda}_2^\top \\ \mathbf{\Lambda}_2\frac{\mathbf{F}_1^\top\mathbf{F}_1}{T_0}\mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_2\frac{\mathbf{F}_1^\top\mathbf{F}_1+\mathbf{F}_2^\top\mathbf{F}_2}{T}\mathbf{\Lambda}_2^\top \end{pmatrix} = \begin{pmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{pmatrix}\left(\sigma_F^2 + o_P(1)\right)\begin{pmatrix} \mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_2^\top \end{pmatrix}.
$$

The same logic carries over to the estimator of the factors. Assume that we know the population loadings, which we use here instead of the estimated loadings in the regression to estimate the factors:

$$
\frac{\tilde{Y}^\top\Lambda}{N}\left(\frac{\Lambda^\top\Lambda}{N}\right)^{-1} = \frac{1}{N}\begin{pmatrix} \mathbf{F}_1\mathbf{\Lambda}_1^\top & \mathbf{F}_2\mathbf{\Lambda}_2^\top \\ 0 & \mathbf{F}_2\mathbf{\Lambda}_2^\top \end{pmatrix}\begin{pmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{pmatrix} + o_P(1) = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2\frac{N-N_0}{N} \end{pmatrix} + o_P(1),
$$

which is a biased estimator for the second time period. The regression in Eq. (3) corresponds to a weighted least square regression which provides the correct estimator:

$$
\tilde{\mathbf{F}}_1 = \mathbf{F}_1\frac{\mathbf{\Lambda}_1^\top\mathbf{\Lambda}_1 + \mathbf{\Lambda}_2^\top\mathbf{\Lambda}_2}{N}\left(\frac{\mathbf{\Lambda}_1^\top\mathbf{\Lambda}_1 + \mathbf{\Lambda}_2^\top\mathbf{\Lambda}_2}{N}\right)^{-1} = \mathbf{F}_1 + o_P(1)
$$

$$
\tilde{\mathbf{F}}_2 = \mathbf{F}_2\frac{\mathbf{\Lambda}_2^\top\mathbf{\Lambda}_2}{N - N_0}\left(\frac{\mathbf{\Lambda}_2^\top\mathbf{\Lambda}_2}{N - N_0}\right)^{-1} = \mathbf{F}_2 + o_P(1)
$$

The proper reweighting in the loading and factor estimation leads to an additional correction term in the asymptotic variance of the estimator. As an illustration of this additional challenge, we add the i.i.d. error term e_{it} to our example. In our simplified setup our consistent estimator for the loadings $\tilde{\Lambda}$ has the following expansion for $i = 1, \ldots, N_0$[8]:

$$
\sqrt{T}\left(\tilde{\Lambda}_i - \Lambda_i\right) = \sqrt{\frac{T}{T_0}}\left(\frac{\tilde{F}^\top\tilde{F}}{T}\right)^{-1}\frac{1}{\sqrt{T_0}}\sum_{t=1}^{T_0}F_t e_{it} + \sqrt{T}\left(\frac{\tilde{F}^\top\tilde{F}}{T}\right)^{-1}\left(\frac{\mathbf{F}_1^\top\mathbf{F}_1}{T_0} - \frac{\mathbf{F}^\top\mathbf{F}}{T}\right)\Lambda_i + o_P(1),
$$

which results in the asymptotic normal distribution

$$
\sqrt{T}\left(\tilde{\Lambda}_i - \Lambda_i\right) \xrightarrow{d} \mathcal{N}\left(0, \text{plim}\left(\frac{T}{T_0}\frac{\sigma_e^2}{\sigma_F^2} + 2\frac{T - T_0}{T_0}\Lambda_i^2\right)\right) \qquad \text{for } i = 1, \ldots, N_0.
$$

The second term in the asymptotic expansion is due to averaging over different number of units for different elements of the loadings. This additional variance correction term vanishes for $T_0 \to T$. Similar terms appear in the distribution of

---

[8] The results are similar for $i > N_0$ with the expansion $\sqrt{T}\left(\tilde{\Lambda}_i - \Lambda_i\right) = \sqrt{T}\left(\frac{\tilde{F}^\top\tilde{F}}{T}\right)^{-1}\frac{1}{N}\left[\mathbf{\Lambda}_1^\top\mathbf{\Lambda}_1\frac{1}{T_0}\sum_{t=1}^{T_0}F_t e_{it} + \mathbf{\Lambda}_2^\top\mathbf{\Lambda}_2\frac{1}{T}\sum_{t=1}^{T}F_t e_{it}\right]$ $+\sqrt{T}\left(\frac{\tilde{F}^\top\tilde{F}}{T}\right)^{-1}\frac{1}{N}\mathbf{\Lambda}_1^\top\mathbf{\Lambda}_1\left(\frac{\mathbf{F}_1^\top\mathbf{F}_1}{T_0} - \frac{\mathbf{F}^\top\mathbf{F}}{T}\right)\Lambda_i + o_P(1)$ and asymptotic distribution $\sqrt{T}\left(\tilde{\Lambda}_i - \Lambda_i\right) \xrightarrow{d} \mathcal{N}\left(0, \text{plim}\left(\left(\frac{T-T_0}{T_0}\frac{N_0^2}{N^2} + 1\right)\frac{\sigma_e^2}{\sigma_F^2} + 2\frac{N_0^2}{N^2}\frac{T-T_0}{T_0}\Lambda_i^2\right)\right)$. Note that the asymptotic variance can be random and depends on the limit of the various sequences. The asymptotic results in Section 4 provide a rigorous formulation of the distribution statements.

the estimators of the factors and common components. We show under general conditions how these correction terms arise in the asymptotic distribution and how to take them into account for the inferential theory.

3. Assumptions

We assume an approximate factor structure at the same level of generality as in Bai (2003). The factors and loadings have non-trivial time-series and cross-sectional dependency. We allow the errors to be weakly correlated in the time-series and cross-sectional dimension. The asymptotic distributions are based on general martingale central limit theorems. The general Assumptions G2 and G3 are collected in Appendix A. In the main text, we present a simplified factor model with the stronger Assumptions S2 and S3, which substantially simplifies the notation but conveys the main conceptual insights of the general model. It allows us to highlight the effect of missing observations.

Assumption Assumption S2 is sufficient for the consistency results. This simplified assumption assumes that all observations are i.i.d. The key elements are that the factors and loadings are systematic in the sense that they lead to exploding eigenvalues, while the error terms are non-systematic with bounded eigenvalues in the covariance matrix of Y. These are standard factor model assumptions. The asymptotic distribution results require additional restrictions on the missing patterns, as stated in Assumption S3.

Assumption S2 (*Simplified Factor Model*)**.** There exists a positive constant $M < \infty$ such that:

1. Factors: $F_t \overset{\text{i.i.d.}}{\sim} (0, \Sigma_F)$, $\mathbb{E}[\|F_t\|^4] \leq M$, and $\mathbb{E}\left\|F_t F_t^\top - \Sigma_F\right\|^{2+\epsilon} \leq M$ for some $\epsilon \in (0, 1)$.
2. Factor loadings: $\Lambda_i \overset{\text{i.i.d.}}{\sim} (0, \Sigma_\Lambda)$ and $\mathbb{E}[\|\Lambda_i\|^4] \leq M$.
3. Errors: $e_{it} \overset{\text{i.i.d.}}{\sim} (0, \sigma_e^2)$, $\mathbb{E}[e_{it}^8] \leq M$.
4. Independence: F, Λ and e are independent.
5. Eigenvalues: The eigenvalues of $\Sigma_\Lambda \Sigma_F$ are distinct.

Assumption S3 (*Moments of Simplified Factor Model*)**.**

1. Systematic loadings: $\frac{1}{N} \sum_{i=1}^N \Lambda_i \Lambda_i^\top W_{it} \overset{P}{\to} \Sigma_{\Lambda,t}$ for some positive definite matrix $\Sigma_{\Lambda,t}$ for any t.
2. Dependency in missing pattern: $\frac{1}{N^2} \sum_{i=1}^N \sum_{l=1}^N \frac{q_{ij,lj}}{q_{ij}q_{lj}} \overset{P}{\to} \omega_{jj}$, $\lim_{N\to\infty} \frac{1}{N^3} \sum_{i=1}^N \sum_{l=1}^N \sum_{k=1}^N \frac{q_{li,kj}}{q_{li}q_{kj}} \overset{P}{\to} \omega_j$ and $\lim_{N\to\infty} \frac{1}{N^4} \sum_{i=1}^N \sum_{l=1}^N \sum_{j=1}^N \sum_{k=1}^N \frac{q_{li,kj}}{q_{li}q_{kj}} \overset{P}{\to} \omega$ for all j and some constants $\omega_{jj}, \omega_j, \omega$.

Assumption S3 has two key elements. First, the full rank assumption of $\Sigma_{\Lambda,t}$ captures that the factor loadings are systematic for the observed entries. Second, the number of observed units at every time period t is proportional to N and different units share a number of observed entries that is proportional to T. The impact of the missing pattern on the asymptotic variances of the estimators is captured by the three key parameters ω, ω_j and ω_{jj}. Note that by construction these constants satisfy $\omega_{jj}, \omega_j, \omega \geq 1$. These parameters tend to increase with a larger fraction of missing observations, or with stronger dependency between the rows of the observation patterns. If observations are missing at random with probability $1 - p$, then $\omega_{jj} = \frac{1}{1-p}$, $\omega_j = 1$ and $\omega = 1$. Intuitively, given the same proportion of missing entries, these key parameters increase as the missing patterns deviates from missing at random, and hence can be interpreted as a measure of complexity in the missing pattern. In fact, these parameters encapsulate all information about the missing patterns for the simplified model. Two missing patterns that have the same parameters $\omega_{jj}, \omega_j, \omega$ result in the same asymptotic distribution for the all-purpose estimator. We show that larger values of these parameters will increase the asymptotic variance for the estimators of the latent factor model.

As stated in Proposition 3 in Appendix A, the simplified model is just a special case of the general approximate factor model specified by Assumptions G2 and G3. The simplified Assumption S2 implies the general Assumption G2, while Assumption S3 combined with the other simplified assumptions implies the general Assumption G3. In the next sections, all the theorems are derived under the general assumptions. Each theorem has a corollary, which states the more transparent results under the simplified assumptions. In particular, Corollary 1 shows the asymptotic distribution of the latent factor model for the simplified model, and highlights the role of the missing pattern parameters ω, ω_j and ω_{jj}.

We assume that the number of factors r is consistently estimated. Given a consistent estimator for the number of factors, we can treat r as known. Jin et al. (2021) provide a consistent estimator in the case of missing at random based on cross-validation arguments. Given the various bounds and expansions derived in this paper, it seems possible to extend the estimator for the number of factors developed in Bai and Ng (2002) to our case of general missing values. A promising alternative would be to extend the cross-validation estimator of Jin et al. (2021) or an eigenvalue ratio argument as in Ahn and Horenstein (2013) to general missing patterns. However, the formal development and analysis of tests for the number of factors under general missing patterns are a challenging and interesting problem left for future work.

4. Asymptotic results

4.1. Consistency

We first show the consistency of our estimators. Our analysis starts with plugging $\tilde{Y} = (\Lambda^\top F + e) \odot W$ into Eq. (2) which yields the following decomposition:

$$
\tilde{\Lambda}_j = \underbrace{\frac{1}{NT} \tilde{D}^{-1} \sum_{i=1}^{N} \tilde{\Lambda}_i \Lambda_i^\top F^\top \operatorname{diag}(W_i \odot W_j) F \Lambda_j / q_{ij}}_{H_j \Lambda_j} + \underbrace{\frac{1}{NT} \tilde{D}^{-1} \sum_{i=1}^{N} \tilde{\Lambda}_i e_i^\top \operatorname{diag}(W_i \odot W_j) F \Lambda_j / q_{ij}}_{(a)}
$$

$$
+ \underbrace{\frac{1}{NT} \tilde{D}^{-1} \sum_{i=1}^{N} \tilde{\Lambda}_i \Lambda_i^\top F^\top \operatorname{diag}(W_i \odot W_j) e_j / q_{ij}}_{(b)} + \underbrace{\frac{1}{NT} \tilde{D}^{-1} \sum_{i=1}^{N} \tilde{\Lambda}_i e_i^\top \operatorname{diag}(W_i \odot W_j) e_j / q_{ij}}_{(c)}.
$$

Similar to Bai and Ng (2002) this decomposition relates the estimated loadings to the population loadings, $\tilde{\Lambda}_j = H_j \Lambda_j + (a) + (b) + (c)$, up to a rotation matrix $H_j = \frac{1}{NT} \tilde{D}^{-1} \sum_{i=1}^{N} \tilde{\Lambda}_i \Lambda_i^\top F^\top \operatorname{diag}(W_i \odot W_j) F / q_{ij}$. The key difference to factor analysis with fully observed data is that this rotation matrix can be different for different units j. However, the estimation of the factors is based on a projection on the loading space and hence implicitly requires the same rotation matrix for all loadings.

We consider for all units a unified rotation matrix defined as $H = \frac{1}{NT} \tilde{D}^{-1} \tilde{\Lambda}^\top \Lambda F^\top F$ which is essentially the same conventional rotation matrix as in Bai and Ng (2002). This yields the decomposition

$$
\tilde{\Lambda}_j - H \Lambda_j = \tilde{\Lambda}_j - H_j \Lambda_j + (H_j - H) \Lambda_j = (a) + (b) + (c) + (H_j - H) \Lambda_j.
$$

We show that the cross-section averages of the square of (a), (b) and (c) converge to 0 at the rate $O_P\left(\min\left(\frac{1}{N}, \frac{1}{T}\right)\right)$. The key difference compared with the fully observed factor analysis is the last term. If $\frac{1}{T} F^\top F \xrightarrow{P} \Sigma_F$ and $\frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} F_t F_t^\top \xrightarrow{P} \Sigma_F$, we can show that $H_j - H = O_P\left(\min\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{T}}\right)\right)$. This rate is sufficiently fast to obtain consistency, but will contribute to the asymptotic normal distribution. Note that the correction term $H_j - H$ is a fundamental problem for any estimator that makes use of all observations.[9]

The next theorem shows the consistency of the estimated loadings.

Theorem 1. *Define $\delta_{NT} = \min(N, T)$. Under Assumptions S1 and G2 it holds that*

$$
\delta_{NT} \left(\frac{1}{N} \sum_{j=1}^{N} \left\| \tilde{\Lambda}_j - H \Lambda_j \right\|^2 \right) = O_P(1), \tag{5}
$$

where $H = \frac{1}{NT} \tilde{D}^{-1} \tilde{\Lambda}^\top \Lambda F^\top F$.

Theorem 1 states that the complete loading matrix can be consistently estimated up to an appropriate rotation as $N, T \to \infty$ even if we only observe an incomplete panel matrix. The convergence rate is the same rate as for the fully observed panel in Bai and Ng (2002). Theorem 1 is based on the assumption that the observed entries are representative of the missing entries and hence provide a consistent estimation. Theorem 1 is a critical intermediate step to show the asymptotic normality of the estimated factor model in the next section.

4.2. Asymptotic normality

The factors, loadings, and common components are asymptotically normally distributed. Indeed, Theorem 2 states that the asymptotic distributions have two parts: First, we recover the asymptotic variance that is identical to the conventional PCA in Bai (2003) under the same rate conditions. These are the expression when we set the additional correction terms $\Gamma_{\Lambda,j}^{\mathrm{miss}}$ and $\Gamma_{F,t}^{\mathrm{miss}}$ to zero. However, in the presence of missing values, these correction terms are necessary to capture the additional uncertainty. Theorem 2 also includes the asymptotic expansions that lead to the normal distributions. As stated in the previous section, the difference between the unit-specific rotation H_j and the "global" rotation matrix H

---

[9] The estimator in Bai and Ng (2021) can avoid this term by neglecting partially observed entries, which means that in general, they are using less information. The estimator of Bai and Ng (2021) is optimized for the block structure of a simultaneous adoption pattern. It runs two PCA estimates for the block with full cross-sectional observations and the block with full time-series observations. Hence, they can infer the "local" rotation matrices for each block and rotate the estimates to avoid the correction term $H_j - H$. Cahan et al. (2021) leverage the block structure, and in the first step run PCA to estimate factors on the block with full-series observations, so that the correction term $H_j - H$ is avoided. If we seek to use full observations in the first step, which is what we propose in this paper, then the correction term $H_j - H$ cannot be avoided.

contributes to the distribution and leads to the variance correction terms $\Gamma_{\Lambda,j}^{\text{miss}}$ and $\Gamma_{F,t}^{\text{miss}}$. As expected, this variance correction is increasing in the number of missing observations. We want to emphasize again that this type of variance correction is a conceptual issue that cannot be avoided when making use of all observed entries.[10]

Theorem 2. *Under Assumptions S1, G2 and G3 and for $N, T \to \infty$ we have for each j and t:*

1. *For $\sqrt{T}/N \to 0$ the asymptotic distribution of the loadings is*

$$
\sqrt{T}\,\Sigma_{\Lambda,j}^{-1/2}(H^{-1}\tilde{\Lambda}_j - \Lambda_j) = \Sigma_{\Lambda,j}^{-1/2}\left(\frac{1}{T}F^{\top}F\right)^{-1}\left(\frac{1}{N}\Lambda^{\top}\Lambda\right)^{-1}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^{\top}\sqrt{\frac{T}{|Q_{ij}|}}\frac{1}{\sqrt{|Q_{ij}|}}\sum_{t\in Q_{ij}}F_t e_{jt}\right)\right.
$$
$$
\left. + \left(\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^{\top}\sqrt{T}\left(\frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}}F_t F_t^{\top} - \frac{1}{T}F^{\top}F\right)\right)\Lambda_j\right] + o_P(1) \xrightarrow{d} \mathcal{N}(0, I_r), \tag{6}
$$

where

$$
\Sigma_{\Lambda,j} = \Sigma_F^{-1}\Sigma_{\Lambda}^{-1}\left[\Gamma_{\Lambda,j}^{\text{obs}} + \Gamma_{\Lambda,j}^{\text{miss}}\right]\Sigma_{\Lambda}^{-1}\Sigma_F^{-1}
$$

with $\Gamma_{\Lambda,j}^{\text{miss}} = h_j(\Lambda_j)$. $\Gamma_{\Lambda,j}^{\text{obs}}$ and the function $h_j(\cdot)$ are defined in Assumptions G3.3 and G3.5.

2. *For $\sqrt{N}/T \to 0$ and $\sqrt{T}/N \to 0$, the asymptotic distribution of the factors is*

$$
\sqrt{\delta_{NT}}\,\Sigma_{F,t}^{-1/2}(H^{\top}\tilde{F}_t - F_t) = \Sigma_{F,t}^{-1/2}\left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\Lambda_i\Lambda_i^{\top}\right)^{-1}\left(\sqrt{\frac{\delta_{NT}}{N}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_{it}\Lambda_i e_{it}\right)
$$
$$
+ \Sigma_{F,t}^{-1/2}\left(\frac{1}{N}\sum_{i=1}^{N}W_{it}\Lambda_i\Lambda_i^{\top}\right)^{-1}\left(\frac{\sqrt{\delta_{NT}}}{N}\sum_{i=1}^{N}W_{it}\left(H^{-1}\tilde{\Lambda}_i - \Lambda_i\right)\Lambda_i^{\top}F_t\right) + o_P(1) \xrightarrow{d} \mathcal{N}(0, I_r), \tag{7}
$$

where

$$
\Sigma_{F,t} = \Sigma_{\Lambda,t}^{-1}\left[\frac{\delta_{NT}}{N}\Gamma_{F,t}^{\text{obs}} + \frac{\delta_{NT}}{T}\Gamma_{F,t}^{\text{miss}}\right]\Sigma_{\Lambda,t}^{-1},
$$

with $\Gamma_{F,t}^{\text{miss}} = g_t(F_t)$. $\Gamma_{F,t}^{\text{obs}}$ and the function $g_t(\cdot)$ are defined in Assumptions G3.4 and G3.5.

3. *The asymptotic distribution of the common component is*

$$
\sqrt{\delta_{NT}}\,\Sigma_{C,it}^{-1/2}(\tilde{C}_{jt} - C_{jt}) = \sqrt{\delta_{NT}}\,\Sigma_{C,it}^{-1/2}\left(H^{-1}\tilde{\Lambda}_j - \Lambda_j\right)^{\top}F_t + \sqrt{\delta_{NT}}\,\Sigma_{C,it}^{-1/2}\Lambda_j^{\top}\left(H^{\top}\tilde{F}_t - F_t\right) + o_P(1)
$$
$$
\xrightarrow{d} \mathcal{N}(0, 1), \tag{8}
$$

where

$$
\Sigma_{C,it} = \frac{\delta_{NT}}{T}F_t^{\top}\Sigma_F^{-1}\Sigma_{\Lambda}^{-1}\left(\Gamma_{\Lambda,j}^{\text{obs}} + \Gamma_{\Lambda,j}^{\text{miss}}\right)\Sigma_{\Lambda}^{-1}\Sigma_F^{-1}F_t + \frac{\delta_{NT}}{T}\Lambda_j^{\top}\Sigma_{\Lambda,t}^{-1}\Gamma_{F,t}^{\text{miss}}\Sigma_{\Lambda,t}^{-1}\Lambda_j
$$
$$
+ \frac{\delta_{NT}}{N}\Lambda_j^{\top}\Sigma_{\Lambda,t}^{-1}\Gamma_{F,t}^{\text{obs}}\Sigma_{\Lambda,t}^{-1}\Lambda_j - 2\frac{\delta_{NT}}{T}\Lambda_j^{\top}\Sigma_{\Lambda,t}^{-1}\Gamma_{\Lambda,F,j,t}^{\text{miss, cov}}\Sigma_{\Lambda}^{-1}\Sigma_F^{-1}F_t.
$$

$\Gamma_{\Lambda,F,j,t}^{\text{miss, cov}} = g_{j,t}^{\text{cov}}(\Lambda_j, F_t)$, *and the function $g_{j,t}^{\text{cov}}(\cdot, \cdot)$ is defined in Assumption G3.5.*

Importantly, the estimator for the factors has a different convergence rate compared to the conventional estimator on fully observed data. The asymptotic distribution of the factors is determined by two terms with different convergence rates, $\frac{\delta_{NT}}{N}\Gamma_{F,t}^{\text{obs}} + \frac{\delta_{NT}}{T}\Gamma_{F,t}^{\text{miss}}$. With a fully observed panel $\Gamma_{F,t}^{\text{miss}}$ would disappear, and the factors would converge at a rate of $\sqrt{N}$. However, with observations that are not missing at random, the difference between H_j and H, that appears in the loading expansion $H^{-1}\tilde{\Lambda}_i - \Lambda_i$ and has a convergence rate of $\sqrt{T}$, also contributes to the asymptotic distribution of estimated factors, which results in the overall rate $\sqrt{\delta_{NT}}$.

The asymptotic distribution of common components depends on the estimation error of the estimated loadings and factors. In the asymptotic distribution of the estimated loadings and factors, the conventional part with asymptotic variances $\Gamma_{\Lambda,j}^{\text{obs}}$ and $\Gamma_{F,t}^{\text{obs}}$ is asymptotically independent as argued in Bai (2003). However, the second part with the asymptotic variances $\Gamma_{\Lambda,j}^{\text{miss}}$ and $\Gamma_{F,t}^{\text{miss}}$ that captures the difference between H_j and H is in general correlated, and hence their covariance $\Gamma_{\Lambda,F,j,t}^{\text{miss, cov}}$ contributes to the asymptotic variance of common components as stated in Eq. (8).

---

[10] In the asymptotic distribution, we apply the rotation matrices to the estimated loadings and factors instead of their population values as in Bai (2003). Obviously, these two representations are equivalent and can be easily transformed into each other. Our choice of representation was made for exposition purposes only.

The correction terms in the asymptotic variances are determined by the functions $h_j(.)$, $g_t(.)$ and $g_{j,t}^{\text{cov}}(.,.)$. These are quadratic functions in the elements of Λ_i and F_t, which depend on the moments of the factor model. As the loadings and factors are random, it implies that the asymptotic variances themselves are random. This complicates the analysis, but our assumptions ensure that the normalized estimates converge to a standard normal distribution.

The distribution results of Theorem 2 simplify under Assumptions S2 and S3, and we can provide explicit expressions for the asymptotic variances. If we assume in addition that the proportions of observed time-series (q_{ij} and $q_{ij,kl}$) are independent of the second moment of the loadings $\Lambda_i \Lambda_i^\top$, we can further separate the effect of missing patterns from the properties of the factor model.

Corollary 1. *Suppose Assumptions S1, S2 and S3 hold and $N, T \to \infty$. Then Theorem 2 holds. If in addition, q_{ij} and $q_{ij,kl}$ are independent of $\Lambda_m \Lambda_m^\top$ for all i, j, k, l, m, then the asymptotic variances simplify as follows with the weights ω, ω_j and ω_{jj} defined in Assumption S3:*

1. The asymptotic variance of the loadings in formula (6) simplifies to

$$\Sigma_{\Lambda,j} = \omega_{jj} \cdot \Sigma_\Lambda^{\text{obs}} + (\omega_{jj} - 1)\Sigma_{\Lambda,j}^{\text{miss}},$$

where

$$\Sigma_\Lambda^{\text{obs}} = \Sigma_F^{-1}\sigma_e^2, \qquad \Sigma_{\Lambda,j}^{\text{miss}} = \Sigma_F^{-1}\Sigma_\Lambda^{-1}\big(\Lambda_j^\top \otimes \Sigma_\Lambda\big)\Xi_F\big(\Lambda_j \otimes \Sigma_\Lambda\big)\Sigma_\Lambda^{-1}\Sigma_F^{-1},$$

and $\mathbb{E}[\text{vec}(F_t F_t^\top - \Sigma_F)\text{vec}(F_t F_t^\top - \Sigma_F)^\top] = \Xi_F$.
2. The asymptotic variance of the factors in formula (7) simplifies to

$$\Sigma_{F,t} = \frac{\delta_{NT}}{N}\Sigma_{F,t}^{\text{obs}} + \frac{\delta_{NT}}{T}(\omega - 1)\Sigma_{F,t}^{\text{miss}},$$

where

$$\Sigma_{F,t}^{\text{obs}} = \Sigma_{\Lambda,t}^{-1}\sigma_e^2, \qquad \Sigma_{F,t}^{\text{miss}} = \Sigma_{\Lambda,t}^{-1}\big(I_r \otimes (F_t^\top \Sigma_F^{-1}\Sigma_\Lambda^{-1})\big)(\Sigma_{\Lambda,t} \otimes \Sigma_\Lambda)\Xi_F(\Sigma_{\Lambda,t} \otimes \Sigma_\Lambda)\big(I_r \otimes (\Sigma_\Lambda^{-1}\Sigma_F^{-1}F_t)\big)\Sigma_{\Lambda,t}^{-1}.$$

3. The asymptotic variance of the common component in formula (8) simplifies to

$$\Sigma_{C,it} = \frac{\delta_{NT}}{T}\left[F_t^\top\big(\omega_{jj} \cdot \Sigma_\Lambda^{\text{obs}} + (\omega_{jj} - 1) \cdot \Sigma_{\Lambda,j}^{\text{miss}}\big)F_t + (\omega - 1)\Lambda_j^\top \Sigma_{F,t}^{\text{miss}}\Lambda_j - 2(\omega_j - 1)\Lambda_j^\top \Sigma_{\Lambda,F,j,t}^{\text{miss, cov}}F_t\right]$$
$$+ \frac{\delta_{NT}}{N}\Lambda_j^\top \Sigma_{F,t}^{\text{obs}}\Lambda_j,$$

where

$$\Sigma_{\Lambda,F,j,t}^{\text{miss, cov}} = \Sigma_{\Lambda,t}^{-1}\big(I_r \otimes (F_t^\top \Sigma_F^{-1}\Sigma_\Lambda^{-1})\big)(\Sigma_{\Lambda,t} \otimes \Sigma_\Lambda)\Xi_F(\Lambda_j \otimes \Sigma_\Lambda)\Sigma_\Lambda^{-1}\Sigma_F^{-1}.$$

The simplified model provides a clear interpretation of the effect of missing data. Importantly, the parameters ω, ω_j and ω_{jj}, that depend only on the missing pattern, but not on the factor model, determine the weights of correction terms. The asymptotic covariance of the loadings is a weighted combination of the variance of an OLS regression of the population factors F on Y_j and the correction term. The weight $\omega_{jj} \geq 1$ depends on the number of the observed entries and the similarities in observation patterns for different units. Without missing data, it equals $\omega_{jj} = 1$ and the correction term disappears. If the data is observed uniformly at random with probability p, the weight equals $\omega_{jj} = 1/p$ which is increasing in the proportion of missing observations.

Similarly, the asymptotic variance of the factors has two components: the variance of an OLS regression of the population loadings on Y_t using only observed entries, and the correction term. The weight $\omega \geq 1$ increases the scale of the correction term. When all entries are observed, or all entries are observed cross-sectionally at random (with either the same or different probabilities), then $\omega = 1$, the correction term vanishes, and the asymptotic variance only depends on $\Sigma_{F,t}^{\text{obs}}$. If the missing pattern does not depend on the loadings, then $\Sigma_{\Lambda,t} = p_t\Sigma_\Lambda$ and $\Sigma_{F,t}^{\text{obs}}$ simplifies to $\frac{1}{p_t}\Sigma_\Lambda^{-1}\sigma_e^2$ which is the variance of an OLS regression of the population loadings on Y_t scaled by the inverse proportion of observed entries at time t.

The distribution of the common component depends on all three parameters ω, ω_j and ω_{jj}. If all entries are observed at random, then $\omega_j = 1$ and the contribution of the loading and factor distribution to the common component are separated similar to the conventional PCA setup in Bai (2003). In this case, only the three terms $\omega_{jj}F_t^\top \Sigma_\Lambda^{\text{obs}}F_t$, $(\omega_{jj} - 1)F_t^\top \Sigma_{\Lambda,j}^{\text{miss}}F_t$, and $\Lambda_j^\top \Sigma_{F,t}^{\text{obs}}\Lambda_j$ remain in the asymptotic variance.

Remark. If all entries are observed at random with equal probability, we can use the approach of Jin et al. (2021) to estimate the factor model and impute the missing entries. We compare the efficiency of our approach with the one of Jin et al. (2021). For a direct comparison, we follow the order of estimation in Jin et al. (2021) and switch the role of factors and loadings in our all-purpose estimator: We first estimate the factors from the time-series sample covariance matrix, and then estimate the loadings from a time-series regression of the observed outcomes on the estimated factors.

Proposition 1. *Suppose Assumptions S1, S2 and S3 hold and that every entry is randomly missing with observed probability p. We switch the role of factors and loadings in the all-purpose estimator. As $N, T \to \infty$, it holds that:*

1. *The estimated factors $\tilde{F}_t$ are asymptotically the same as the initial estimates of factors in Jin et al. (2021), but are asymptotically less efficient than the iterated estimates of factors in Jin et al. (2021).*
2. *The estimated loadings $\tilde{\Lambda}_i$ are asymptotically more efficient than the initial estimates of loadings in Jin et al. (2021), but are asymptotically the same as the iterated estimates of loadings in Jin et al. (2021).*

5. Propensity weighted estimator

We provide the assumptions and general distribution theory for the propensity weighted estimator for the factors $\tilde{F}_t^S$ defined in Eq. (4). This conditional estimator uses the weights $\frac{1}{P(W_{it}=1|S_i)}$ in the cross-sectional regression to obtain the factors. We allow the observation probability to depend on observed cross-sectional features $S = [S_i] \in \mathbb{R}^{N \times K}$ that explain why certain units are more likely to be observed than other units. This conditional setup requires some modifications of the previous assumptions. In addition to Assumption S1 we require the following assumption:

Assumption C1 (*Conditional Observational Pattern*).

1. W is independent of Λ conditional on S.
2. For any i and j satisfying $i \neq j$, and for any t and s, W_{it} is independent of W_{js} conditional on S_i and S_j where t and s can be the same. The probability of $W_{it} = 1$ depends on S_i and satisfies $P(W_{it} = 1|S_i) \geq \underline{p} > 0$.

We assume S contains all the information in Λ that is predictive for the observation pattern. In other words, W is independent of Λ conditional on S, as stated in Assumption C1.1. This is closely related to the unconfoundedness assumption in causal inference. It also assumes that the conditional probability $P(W_{it} = 1|S_i)$ is bounded away from 0, which implies that the number of observed cross-sectional and time-series entries is proportional to N and T, respectively. This corresponds to the overlap assumption in causal inference.[11] Note that it is straightforward to include the covariates of "neighbor units" in S_i to allow for network effects.

We replace Assumptions G2 and G3 by their conditional counterpart Assumptions GC2 and GC3 which have a similar level of generality. These are required for the asymptotic normality of $\tilde{F}_t^S$ and $\tilde{C}_{it}^S$. As before, we collect the Assumptions GC2 and GC3 for a general approximate factor model in Appendix A and present the assumptions for a simplified factor model in the main text, which are sufficient to convey all conceptual insights.

Assumption C2 (*Conditional Factor Model*).

1. S is independent of F and e.
2. For any i, Λ_i is independent of S_j conditional on S_i for $j \neq i$. Moreover, for any i and j satisfying $i \neq j$, Λ_i is independent of Λ_j conditional on S_i and S_j.

Assumption C3 (*Moments of Conditional Factor Model*).

1. $\mathbb{E}[\|\Lambda_i\|^8 |S] \leq \overline{\Lambda} < \infty$.
2. Systematic loadings: $\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{P(W_{it}=1|S_i)} \mathbb{E}[\Lambda_i \Lambda_i^\top |S_i] \overset{P}{\to} \Sigma_{\Lambda,S,t}$ for every t for some positive definite matrix $\Sigma_{\Lambda,S,t}$.

Under Assumption C2, $\frac{1}{N} \sum_{i=1}^{N} \frac{W_{it}}{P(W_{it}=1|S_i)} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top$ converges in probability to an identity matrix which is the same limit as the loading estimates in conventional PCA without missing data. The assumption that S is independent of F and e is conceptually similar to the assumption that Λ is independent of F and e, where the latter is standard in the literature on large dimensional factor modeling. The additional moment conditions in Assumption C3 are required for the asymptotic distribution, where $\Sigma_{\Lambda,S,t}$ appears in the asymptotic covariances of $\tilde{F}_t^S$ and $\tilde{C}_{it}^S$.

Proposition 3 in Appendix A shows that the simplified model is just a special case of the general approximate factor model specified by Assumptions GC2 and GC3. The simplified Assumption S2 combined with Assumptions S1, C1 and S2 imply the general conditional Assumption G2, while Assumption S3 combined with the other simplified Assumptions S1, C1, S2, S3.2 and C2 imply the general Assumption GC3.

5.1. Asymptotic normality

The propensity weighted estimator only differs in the distribution of the factors and common components. Both $\tilde{F}_t^S$ and $\tilde{C}_{it}^S$ follow a normal distribution, but in most cases have a larger asymptotic variance than the estimators $\tilde{F}_t$ and $\tilde{C}_{it}$. The loadings are not affected by the propensity score weighting.

---

[11] See Rosenbaum and Rubin (1983) for the connection to unconfoundedness and the overlap assumption. We assume $P(W_{it} = 1|S)$ is bounded away from 0, such that $\frac{1}{P(W_{it}=1|S)}$ does not diverge, which is equivalent to the overlap assumption in causal inference.

Theorem 3. *Under Assumptions S1, C1, G2, GC2 and GC3 and for $N, T \to \infty$ we have for each j and t:*

1. *The asymptotic distribution of the loadings is the same as in Theorem 2.*
2. *For $\sqrt{N}/T \to 0$ and $\sqrt{T}/N \to 0$, the asymptotic distribution of the factors is*

$$\sqrt{\delta_{NT}} (\Sigma_{F,t}^S)^{-1/2} (H^\top \tilde{F}_t^S - F_t) \xrightarrow{d} \mathcal{N}(0, 1), \tag{9}$$

where

$$\Sigma_{F,t}^S = \Sigma_\Lambda^{-1} \left[\frac{\delta_{NT}}{N} \Gamma_{F,t}^{\mathrm{obs},S} + \frac{\delta_{NT}}{T} \Gamma_{F,t}^{\mathrm{miss},S} \right] \Sigma_\Lambda^{-1},$$

with $\Gamma_{F,t}^{\mathrm{miss},S} = g_t^S(F_t)$. $\Gamma_{F,t}^{\mathrm{obs},S}$ and $g_t^S(\cdot)$ are defined in Assumptions GC3.4 and GC3.5.

3. *The asymptotic distribution of the common components is*

$$\sqrt{\delta_{NT}} (\Sigma_{C,it}^S)^{-1/2} (\tilde{C}_{jt}^S - C_{jt}) \xrightarrow{d} \mathcal{N}(0, 1), \tag{10}$$

where

$$\Sigma_{C,it}^S = \frac{\delta_{NT}}{T} F_t^\top \Sigma_F^{-1} \Sigma_\Lambda^{-1} (\Gamma_{\Lambda,j}^{\mathrm{obs}} + \Gamma_{\Lambda,j}^{\mathrm{miss}}) \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t + \frac{\delta_{NT}}{T} \Lambda_j^\top \Sigma_\Lambda^{-1} \Gamma_{F,t}^{\mathrm{miss},S} \Sigma_\Lambda^{-1} \Lambda_j$$

$$+ \frac{\delta_{NT}}{N} \Lambda_j^\top \Sigma_\Lambda^{-1} \Gamma_{F,t}^{\mathrm{obs},S} \Sigma_\Lambda^{-1} \Lambda_j - 2 \cdot \frac{\delta_{NT}}{T} \Lambda_j^\top \Sigma_\Lambda^{-1} \Gamma_{\Lambda,F j,t}^{\mathrm{miss,\ S,\ cov}} \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t,$$

with $\Gamma_{\Lambda,F j,t}^{\mathrm{miss,\ S,\ cov}} = g_{j,t}^{\mathrm{cov},S}(\Lambda_j, F_t)$, and the function $g_{j,t}^{\mathrm{cov},S}(\cdot, \cdot)$ is defined in Assumption GC3.5.

The distribution results have the same general structure as in Theorem 2. However, there are two key differences. First, the outer matrices in the variance of $\tilde{F}_t^S$ are Σ_Λ^{-1} while they depend on the observational pattern in $\Sigma_{\Lambda,t}^{-1}$ in Eq. (7). Second, the middle terms $\Gamma_{F,t}^{\mathrm{obs},S}$ and $\Gamma_{F,t}^{\mathrm{miss},S}$ may depend on $P(W_{it} = 1|S_i)$. The same structure carries over to the common component. In the case of generalized least squares regressions, it is straightforward to compare the asymptotic covariances for different weights and to determine an efficient estimator. With missing observations, the problem becomes more challenging as the asymptotic covariances depend on two matrices for the factor estimates and three terms for the common components. For the general models in Theorems 2 and 3 we cannot state which estimator is more efficient without imposing additional structure. However, for the simplified model, we can rank the efficiency of the two estimators.

Corollary 2. *Suppose Assumptions S1, C1, S2, S3.2, C2, and C3 hold and $N, T \to \infty$. Then Theorem 3 holds. If in addition, q_{ij} and $q_{ij,kl}$ are independent of $\Lambda_m \Lambda_m^\top$ for all i, j, k, l, m, then the asymptotic variances simplify as follows with the weights ω, ω_j and ω_{jj} defined in Assumption S3:*

1. *The asymptotic variance of the factors in formula (9) simplifies to*

$$\Sigma_{F,t}^S = \frac{\delta_{NT}}{N} \Sigma_F^{\mathrm{obs},S} + \frac{\delta_{NT}}{T} (\omega - 1) \Sigma_{F,t}^{\mathrm{miss},S},$$

where

$$\Sigma_{F,t}^{\mathrm{obs},S} = \Sigma_\Lambda^{-1} \Sigma_{\Lambda,S,t} \Sigma_\Lambda^{-1} \sigma_e^2,$$
$$\Sigma_{F,t}^{\mathrm{miss},S} = \Sigma_\Lambda^{-1} (I_r \otimes (F_t^\top \Sigma_F^{-1} \Sigma_\Lambda^{-1})) (\Sigma_\Lambda \otimes \Sigma_\Lambda) \Xi_F (\Sigma_\Lambda \otimes \Sigma_\Lambda) (I_r \otimes (\Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t)) \Sigma_\Lambda^{-1}.$$

2. *The asymptotic variance of the common component in formula (10) simplifies to*

$$\Sigma_{C,it}^S = \frac{\delta_{NT}}{T} \left[F_t^\top \left(\omega_{jj} \Sigma_\Lambda^{\mathrm{obs}} + (\omega_{jj} - 1) \Sigma_{\Lambda,j}^{\mathrm{miss}} \right) F_t + (\omega - 1) \Lambda_j^\top \Sigma_{F,t}^{\mathrm{miss},S} \Lambda_j - 2(\omega_j - 1) \Lambda_j^\top \Sigma_{\Lambda,F j,t}^{\mathrm{miss,\ S,\ cov}} F_t \right]$$

$$+ \frac{\delta_{NT}}{N} \Lambda_j^\top \Sigma_{F,t}^{\mathrm{obs},S} \Lambda_j,$$

where

$$\Sigma_{\Lambda,F,t}^{\mathrm{miss,\ S,\ cov}} = \Sigma_\Lambda^{-1} (I_r \otimes (F_t^\top \Sigma_F^{-1} \Sigma_\Lambda^{-1})) (\Sigma_\Lambda \otimes \Sigma_\Lambda) \Xi_F (\Lambda_j \otimes \Sigma_\Lambda) \Sigma_\Lambda^{-1} \Sigma_F^{-1}.$$

3. *$\tilde{F}_t^S$ is weakly less efficient than $\tilde{F}_t$, if S is independent of Λ. In the case of only one factor, i.e. $r = 1$, $\tilde{F}_t^S$ is weakly less efficient than $\tilde{F}_t$ for any S.*

An interesting observation is that $\Sigma_{F,t}^{\mathrm{miss},S}$ and $\Sigma_{\Lambda,F j,t}^{\mathrm{miss,\ S,\ cov}}$ depend neither on the observation pattern nor on S. This is because $\frac{1}{P(W_{it}=1|S_i)}$ removes the asymptotic dependency between W_{it} and Λ_i. Hence, this part of the asymptotic distribution has a complete separation between the missing observation pattern captured by the weights ω, ω_j and ω_{jj} and distribution terms that depend only on the factor model. However, $\Sigma_{F,t}^{\mathrm{obs},S}$ depends on $P(W_{it} = 1|S_i)$ as this component comes from a

probability weighted least square regression of the population loadings on the observed entries in Y, which is different from the corresponding OLS regression in Corollary 1.1.

The key observation is that $\tilde{F}_t^S$ and as a consequence also $\tilde{C}_{it}^S$ seem to be in many cases less efficient than $\tilde{F}_t$ and $\tilde{C}_{it}$, which means that the asymptotic variances of the all-purpose estimator are less than or equal to those of the propensity weighted estimator. In the case of only one factor it holds that $\Sigma_{F,t}^{\text{miss},S} = \Sigma_{F,t}^{\text{miss}}$. Not surprisingly, it holds that $\Sigma_{F,t}^{\text{obs},S} \geq \Sigma_{F,t}^{\text{obs}}$ as in the case of i.i.d. errors, an OLS regression is the most efficient linear estimator. This result can also be derived from $\Sigma_{A,S,t}/\Sigma_A^2 \geq 1/\Sigma_{A,t}$, which follows from the Cauchy–Schwartz inequality. In the case of multiple factors, we take advantage of the concavity of the average weighted by $1/P(W_{it} = 1|S_i)$ to prove the efficiency relationship. In simulations we confirm that when the loadings depend on S, it is possible that $\Sigma_{F,t}^{\text{miss},S} < \Sigma_{F,t}^{\text{miss}}$, which can result in minor efficiency gains for $\tilde{F}_t^S$. For a general residual covariance matrix, the efficiency results are more complex. Pelger and Xiong (2021a) show that for fully observed data under certain assumptions, the optimal weight in the factor regression is the inverse residual covariance matrix or equivalently PCA, applied to a covariance matrix re-weighted by the square-root of the inverse residual covariance matrix, is the most efficient estimator. Hence, if the propensity weight is close to the inverse residual covariance matrix, it lowers the first term $\Sigma_{F,t}^{\text{obs},S}$. However, the effect on the second term is more complex, and hence there are in general cases where $\tilde{F}_t^S$ can be more efficient than $\tilde{F}_t$. In simulations, we show that for a correctly specified model, the estimates of $\tilde{F}_t$ and $\tilde{F}_t^S$ are close, but $\tilde{F}_t$ is generally more precise. However, the "doubly-robust" estimator $\tilde{F}_t^S$ seems to be less affected by various forms of misspecification, e.g., omitted factors, weak factors, or a nonlinear factor model. Hence, $\tilde{F}_t^S$ might be appealing because it is more robust but not based on efficiency arguments.

5.2. Robustness to model misspecification

The propensity weighted regressions can be robust to the selection bias from omitting factors. In the causal inference literature regressions weighted by propensity scores have been used in the estimation of causal effects to reduce the bias that arises from omitting regressors or misspecifying the outcome model. However, as the propensity weighted regressions have a larger variance, regressions without the propensity weights seem to be preferred for correctly specified models.[12] In this section, we illustrate that this logic carries over to our latent factor model setup.

Our setup differs from classical causal inference as we estimate the covariates as latent factors from the data. However, we can have a situation similar to omitted variables if we estimate too few latent factors, the factors are weak, or the population model is nonlinear. In the previous section we have shown that the propensity weighted estimator is in general less efficient than our all-purpose estimator when we use the correct number of factors r. However, when we use too few latent factors, our weighted estimator can have a smaller selection bias, and hence be preferable. We illustrate the general logic of this result with an example and confirm it with extensive simulations in Section 10.4.

We assume a two-factor model $Y_{it} = \Lambda_{i1}F_{t1} + \Lambda_{i2}F_{t2} + e_{it}$, where

$$\begin{bmatrix} F_{t1} \\ F_{t2} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_F^2 & 0 \\ 0 & \sigma_F^2 \end{bmatrix}\right), \qquad \begin{bmatrix} \Lambda_{i1} \\ \Lambda_{i2} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_\Lambda \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\Lambda^2 & 0 \\ 0 & \sigma_\Lambda^2 \end{bmatrix}\right), \qquad e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2).$$

The key assumption is that the observation pattern depends on the loadings. In order to have a transparent example, we assume the following pattern. The first T_0 time periods are fully observed. After time T_0, whether a unit is observed or not depends on an indicator variable S_i, defined as $S_i = \mathbf{1}_{\Lambda_{i1} > c_1, \Lambda_{i2} > c_2}$ for some $c_1 > 0$ and $c_2 > 0$. Suppose $P(W_{it} = 1|S_i = 1) = p$ and $P(W_{it} = 1|S_i = 0) = 1 - p$ for some $p > 0.5$. In other words, units with large loadings are more likely to be observed. We would get similar results if large loadings are more likely to be missing. Without loss of generality, we can shuffle the units and obtain the observation pattern in Table 1(a). In the following, we estimate the factor model from the shuffled outcome matrix, where the outcomes for the first N_0 units after time T_0 are missing, and the outcomes for the last $N - N_0$ units are fully observed.

Assume that we omit one factor and estimate only a one-factor model with both, the simple and propensity weighted, estimators. In this case, our factor model is misspecified. Without loss of generality, we can set $\mu_\Lambda^2 + \sigma_\Lambda^2 = 1$. The estimated loading vector is consistent and the same for both estimators:

$$\tilde{\Lambda}_1 = \Lambda_1 + o_p(1).$$

We compare for both approaches the estimates of the first factor and the common components from time $T_0 + 1$ to T (the two approaches coincide from time 1 to T_0 as all units are fully observed). For the simple regression and for $T_0 < t < T$, we can show that

$$\tilde{F}_{t1} = \left(\sum_{i=N_0+1}^{N} Y_{it}\tilde{\Lambda}_{i1}\right)\left(\sum_{i=N_0+1}^{N} \tilde{\Lambda}_{i1}^2\right)^{-1} = F_{t1} + \gamma F_{t2} + o_p(1),$$

---

[12] Robins et al. (1994), Robins and Rotnitzky (1995) among others discuss the reduction of bias from omitting regressors. Robins and Wang (2000), Kang and Schafer (2007), Robins et al. (2007) show the large variance of propensity weighted regressions, and Freedman and Berk (2008) suggest unweighted regressions for correctly specified models.

where $\gamma = \lim_{N_0, N \to \infty} \frac{\sum_{i=N_0+1}^{N} \Lambda_{i1} \Lambda_{i2}}{\sum_{i=N_0+1}^{N} \Lambda_{i1}^2}$. The key element is that $\gamma \neq 0$ because of the dependence of the missing pattern on Λ. In our example, both Λ_{i1} and Λ_{i2} tend to have large values on the observed units, and hence are not asymptotically orthogonal on the subset of observed data. A non-zero γ creates a selection bias similar to the conventional omitted variable bias.[13]

Since the common component is $C_{it} = \Lambda_{i1}F_{t1} + \Lambda_{i2}F_{t2}$, the estimated common components have an error of $\tilde{C}_{it} - C_{it} = (\gamma \Lambda_{i1} + \Lambda_{i2})F_{t2} + o_p(1)$. Hence, the mean squared error from $T_0 + 1$ to T equals

$$\frac{1}{N(T - T_0)} \sum_{i=1}^{N} \sum_{t=T_0+1}^{T} (\tilde{C}_{it} - C_{it})^2 = \frac{1}{N(T - T_0)} \sum_{i=1}^{N} \sum_{t=T_0+1}^{T} \left(\gamma \Lambda_{i1} + \Lambda_{i2} \right)^2 F_{t2}^2 + o_p(1)$$

$$= (\gamma^2 + \sigma_\Lambda^2)\sigma_F^2 + o_p(1)$$

as $\frac{1}{N} \sum_{i=1}^{N} \Lambda_{i1} \Lambda_{i2} \to 0$. In contrast, the propensity weighted regression for $T_0 < t < T$ equals

$$\tilde{F}_{t1}^S = \left(\sum_{i=N_0+1}^{N} \frac{1}{P(W_{it} = 1|S_i)} Y_{it} \tilde{\Lambda}_{i1} \right) \left(\sum_{i=N_0+1}^{N} \frac{1}{P(W_{it} = 1|S_i)} \tilde{\Lambda}_{i1}^2 \right)^{-1} = F_{t1} + o_p(1).$$

Thus, the propensity weighted regression corrects for the selection bias and is the same as if γ would have been zero. The estimated common components have an error of $\tilde{C}_{it}^S - C_{it} = \Lambda_{i2}F_{t2} + o_p(1)$, which implies that the mean squared error from time $T_0 + 1$ to T is

$$\frac{1}{N(T - T_0)} \sum_{i=1}^{N} \sum_{t=T_0+1}^{T} (\tilde{C}_{it}^S - C_{it})^2 = \frac{1}{N(T - T_0)} \sum_{i=1}^{N} \sum_{t=T_0+1}^{T} \Lambda_{i2}^2 F_{t2}^2 + o_p(1) = \sigma_\Lambda^2 \sigma_F^2 + o_p(1).$$

Importantly, $\tilde{C}_{it}^S$ has a smaller mean squared error than $\tilde{C}_{it}$. In summary, when the estimated factor model has omitted factor(s) and is misspecified, the propensity weighted estimator could adjust for the selection bias and reduce the estimation error. We conclude that the propensity weighted estimator can serve as a valid robustness check for model misspecification. Estimating the number of latent factors is a challenging problem. Most existing estimators for the number of latent factors explicitly or implicitly depend on choice parameters, which implies that in practice it is possible to use too few factors.[14] Given an estimate for the number of factors, we recommend to compare the results of our all-purpose and propensity weighted estimators. A large difference between the results can indicate that the number of latent factors should be increased.

6. Feasible estimator of the probability weighting

We provide feasible estimators for $P(W_{it} = 1|S_i)$ which we need in Eq. (4) to estimate the factors, and we show that the asymptotic distribution of factors is not affected by using the estimated weights instead of their population counterpart. While in (stratified) randomized experiments, researchers decide and therefore know the treatment assignment probability given covariates, $P(W_{it} = 1|S_i)$, the probability distribution of the missing pattern in observational studies generally needs to be estimated, which can affect the distribution theory for the latent factor model. Here we provide conditions under which the previously derived results continue to hold with a feasible estimator. To simplify notation denote by $p_{it} = P(W_{it} = 1|S_i)$ the propensity score and its estimate by $\hat{p}_{it} = \hat{P}(W_{it} = 1|S_i)$. The feasible estimator for the factors $\hat{F}_t^S$ replaces p_{it} by $\hat{p}_{it}$ in Eq. (4), which yields the following decomposition:

$$\hat{F}_t^S = \left(\sum_{i=1}^{N} \frac{W_{it}}{\hat{p}_{it}} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right)^{-1} \left(\sum_{i=1}^{N} \frac{W_{it}}{\hat{p}_{it}} Y_{it} \tilde{\Lambda}_i \right) = \tilde{F}_t^S + \left(\frac{1}{N} \sum_{i=1}^{N} \frac{W_{it}}{p_{it}} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{p_{it} - \hat{p}_{it}}{\hat{p}_{it}} \frac{W_{it}}{p_{it}} Y_{it} \tilde{\Lambda}_i \right)$$

$$+ \left(\frac{1}{N} \sum_{i=1}^{N} \frac{W_{it}}{p_{it}} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\hat{p}_{it} - p_{it}}{\hat{p}_{it}} \frac{W_{it}}{p_{it}} \tilde{\Lambda}_i \tilde{\Lambda}_i^\top \right) \cdot \hat{F}_t^S.$$

Under weak assumptions on $\hat{p}_{it}$, that are satisfied for feasible estimators of the most empirically relevant observation patterns, the additional term $\hat{F}_t^S - \tilde{F}_t^S$ can be neglected in the asymptotic distribution.

Theorem 4. *We replace the propensity score in p_{it} in Eq. (4) by its estimate $\hat{p}_{it}$.*

---

[13] Simon (1954) refers to this type of selection bias as a spurious correlation.

[14] For example, Pelger (2019) and Lettau and Pelger (2020b) estimate the number of latent factors in panels of stock returns, and show that the estimates are affected by tuning parameters, the relative strength of the factors to each other and also relative to the noise.

Table 2

Examples of feasible estimators of the probability weight.

Description	$P(W_{it} = 1\|S_i)$	Estimator	Asymptotic distribution of $\hat{p}_{it} - p_{it}$	Effect on distribution
Time-series missing at random (parametric)	$p(S_i)$	Logit on full panel W	$O_P\left(\frac{1}{\sqrt{NT}}\right)$	No
Time-series missing at random (non-parametric)	$p(S_i)$	Kernel on full panel W	$O_P\left(\frac{1}{\sqrt{NTh}}\right)$	No
Cross-section and time-series dependency (parametric)	$p_t(S_i)$	Logit on W_t	$O_P\left(\frac{1}{\sqrt{N}}\right)$	No
Cross-section and time-series dependency (discrete S)	$p_t(s)$	$\hat{p}_t(s) = \frac{\|\mathcal{O}_{s,t}\|}{N_s}$	$\frac{1}{\sqrt{N_s}}\mathcal{N}\left(0, p_t(s)(1 - p_t(s))\right)$	No
Staggered treatment adoption with S (parametric)	$p_t(S_i)$	Hazard rate model	$O_P\left(\frac{1}{\sqrt{N}}\right)$	No

This table shows feasible estimators for the probability of the most important cases of missing patterns. We propose examples of feasible estimators. The asymptotic distribution for $\hat{p}_{it} - p_{it}$ is given under suitable assumptions and for S_i being i.i.d and sub-Gaussian. The main text includes additional details. The effect on distribution refers to the asymptotic distribution of the factors and common components in Theorem 3. The exact details are described in Theorem 4.

1. *The estimates of the loadings do not depend on the propensity score. Hence, Theorem 1 and the asymptotic distribution of the loadings in Theorem 3 continue to hold independently of $\hat{p}_{it}$.*
2. *The following holds for the distribution of the factors and common components.*
   (a) *If $\max_i |\hat{p}_{it} - p_{it}| = o_P(1)$, then the factors and common components are estimated consistently pointwise under the assumptions of Theorem 3.*
   (b) *If $\max_i |\hat{p}_{it} - p_{it}| = o_P\left(\frac{1}{N^{1/4}}\right)$, then Theorem 3 continues to hold as it is.*

We discuss feasible estimators for the most important cases of missing patterns which are summarized in Table 2. Obviously, we only need to consider the case where p_{it} varies for different cross-sectional units as otherwise the estimator simplifies to our estimator in Eq. (2). For simplicity these examples assume that S_i are *i.i.d.* and sub-Gaussian but can be generalized to weak dependency patterns. The simplest case is missing at random only in the time-series dimension, that is $P(W_{it} = 1|S_i) = p(S_i)$ for some parametric or non-parametric function $p(.)$. A relevant example is the estimation of $p(S_i)$ with a logit model on the full panel W which has the convergence rate $\hat{p}(S_i) = p(S_i) + O_P\left(\frac{1}{\sqrt{NT}}\right)$ and a uniform bound of order $\frac{\log(NT)}{\sqrt{NT}}$. Hence, Theorem 4.2(b) applies. If $p(S_i)$ is estimated non-parametrically with a kernel with bandwidth h, the convergence rate is typically $\sqrt{NTh}$ with a uniform bound of order $\frac{\log(NTh)}{\sqrt{NTh}}$, which does not change the distribution results if Th is sufficiently large. In the more complex model $P(W_{it} = 1|S_i) = p_t(S_i)$ the observations probability depends on the cross-section and time-series information. A relevant example for a parametric model is a logit model estimated on W_t for each t separately with a convergence rate of $\sqrt{N}$. Under weak assumptions on S_i, the uniform convergence bound in Theorem 4.2(b) holds.

An important special case is discrete values for S, that is, the covariates S take only finitely many values. An example for a binary variable S would be gender, when male or female individuals have different probabilities to be treated. If the probabilities for the different discrete outcomes of S are bounded away from zero, then the estimator $P(W_{it} = 1|S_i = s) = p_t(s)$ simplifies to p_t, but just averaged over the cross-section units for which $S_i = s$. In more detail, consider the estimator $\hat{p}_t(s) = \frac{|\mathcal{O}_{s,t}|}{N_s}$ where $N_s = \sum_{i=1}^N \mathbf{1}(S = s)$ and $\mathcal{O}_{s,t} = \{i : W_{it} = 1 \text{ and } S = s\}$. Then, $\sqrt{N_s}\left(\hat{p}_t(s) - p_t(s)\right) \xrightarrow{d} \mathcal{N}\left(0, p_t(s)(1 - p_t(s))\right)$. If N_s is sufficiently large, for example proportional to N, then the feasible estimator does not change the distribution results in Theorem 3. These estimators directly carry over to staggered treatment adoption. The staggered design can also be modeled with a parametric hazard model $P(W_{it} = 1|S_i) = p(t, S_i)$, which under appropriate assumptions converges at the rate $\sqrt{N}$ as well. In summary, for all these cases the distribution results are not affected by using a feasible estimator for the propensity score.

As previously mentioned, we allow $S_i = \Lambda_i$. This is appealing as Λ is by construction capturing the unit-specific features and hence should account for the differences in cross-sectional observation patterns. As the estimator $\hat{\Lambda}$ does not depend on the probability weights, it can be used in the estimation of $P(W_{it}|\Lambda_i)$. Theorem 3 states that the estimation error of $\hat{\Lambda}_i$ is of the order $O_P\left(\frac{1}{\sqrt{N}}\right)$. While the consistency results for the factors and common components continue to hold, we need some additional weak assumptions on the tail behavior of the loadings and error terms to satisfy the uniform condition in Theorem 4.2(b).

7. Tests of treatment effects

The key application of our asymptotic distribution theory is to test causal effects. The fundamental problem in causal inference is that we observe an outcome either for the control or the treated data, but not for both at the same time. The unknown counterfactual of what the treated observations could have been without treatment can be naturally modeled as a data imputation problem. In this section, we consider the case where once a unit adopts the treatment, it stays treated afterward, for example, the simultaneous and staggered treatment designs illustrated in Fig. 1. Given the general missing patterns that we allow for, the generalization to more complex adoption patterns is straightforward. We denote by $T_{0,i}$ and $T_{1,i}$ the number of control and treated time periods for unit i where their sum adds up to $T_{0,i} + T_{1,i} = T$. The superscripts (0) and (1) denote the observations for control and treated observations.

The individual treatment effect measures the difference between the treated and control outcomes:

$$\tau_{it} = Y_{it}^{(1)} - Y_{it}^{(0)} \qquad \text{for } t > T_{0,i}, i = 1, \dots, N,$$

where by construction for a specific time t and unit i we only observe either $Y_{it}^{(1)}$ or $Y_{it}^{(0)}$, but not both. Average treatment effects can be estimated by an average over time or the cross-section of the individual treatment effects. We assume that the data has a factor structure which results in a model of the form

$$Y_{it} = \tau_{it} D_{it} + \Lambda_i^\top F_t + e_{it}, \tag{11}$$

where $D_{it} = 1$ is a treatment indicator. Note that this model is very general and captures many relevant models as special cases. The factor structure includes interactive fixed effects as in Bai (2009). Simple time- and cross-sectional fixed effects are a special case for constant loadings respectively factors. The factors can be either observed covariates or latent factors. One of the main challenges in estimating a treatment effect is to control for all relevant covariates. Failure in doing so results in an omitted variable bias in the treatment effect estimation as discussed among others in Gobillon and Magnac (2016). The strength of our latent factor model is that we can avoid this problem by automatically including all relevant covariates in a data driven way. Note that our latent factor model can also account for some uncertainty in the functional form of the dependency on the factors. For example, if Y_{it} is a polynomial function of a factor, this could be captured by including additional latent factors as described for example in Pelger and Xiong (2021b). A generalization of Eq. (11) adds additional observed covariates $X_{it} \in \mathbb{R}^K$ to Y_{it} which yields $Y_{it} = \tau_{it} D_{it} + \Lambda_i^\top F_t + X_{it}^\top b + e_{it}$. If these observed covariates follow a factor structure $X_{it} = \Lambda_i^{X\top} F_t^X + e_{it}^X$, it puts us back into the framework of Eq. (11). Otherwise it is straightforward to include general observable covariates X_{it} by studying the residual $Y_{it} - X_{it}\hat{b}$, where $\hat{b}$ is estimated by a regression on the control group.[15]

We only observe $Y_{it}^{(1)}$ for the treated group and could obtain the counterfactual outcome $Y_{it}^{(0)}$ from the imputed value $\hat{Y}_{it}^{(0)} = \hat{C}_{it}^{(0)}$, where $\hat{C}_{it}^{(0)}$ is the common component estimated only from the untreated control data. This is the same setup as in Bai and Ng (2021). Given our asymptotic distribution theory for the common component, we can provide the asymptotic distribution of the individual and average treatment effects analogously to Bai and Ng (2021). A shortcoming of estimating the individual treatment effect by $Y_{it}^{(1)} - \hat{Y}_{it}^{(0)}$ is that the observed treated observations $Y_{it}^{(1)}$ contain an idiosyncratic error e_{it}. Hence, it is not possible to test for individual treatment effects without imposing very strong additional assumptions on the error. For sufficiently large $T - T_0$, this error component can be averaged out in the average treatment effect.

We impose slightly stronger assumptions on the structure of the treatment effect which will allow us to derive substantially stronger results. Assume that the treatment effect has also a factor structure, that is $\tau_{it} = \left(\Lambda_i^\tau\right)^\top F_t^\tau$. In this case we can represent the problem as

$$Y_{it}^{(1)} = \left(\Lambda_i^{(1)}\right)^\top F_t^{(1)} + e_{it} \qquad Y_{it}^{(0)} = \left(\Lambda_i^{(0)}\right)^\top F_t^{(0)} + e_{it}, \tag{12}$$

where the factor structure subsumes the treatment effect. Hence, the individual treatment effect is equivalent to the difference in the common components between the treated and control:

$$\tau_{it} = Y_{it}^{(1)} - Y_{it}^{(0)} = C_{it}^{(1)} - C_{it}^{(0)} \qquad \text{for } t > T_{0,i}, i = 1, \dots, N.$$

Fundamentally, we are testing if the treatment changes the underlying factor structure. Hence, we can test if the treatment changes interactive fixed effects. This is a very general setup that allows for time and cross-sectional heterogeneity in the treatment effect, while the treatment itself can depend on the latent cross-sectional covariates modeled by the loadings.

In the following we consider three different treatment effects:

1. Individual treatment effect: $\tau_{it} = C_{it}^{(1)} - C_{it}^{(0)}$
2. Average treatment effect over time: $\tau_i = \frac{1}{T_{1,i}} \sum_{t=T_{0,i}+1}^{T} \tau_{it}$

---

[15] Using the residuals $Y_{it} - X_{it}\hat{b}$ for the factor analysis and treatment effect analysis with our method generally adds another covariance term to the asymptotic covariance matrix. This term comes from the regression to obtain $\hat{b}$ and is straightforward to include. Here we focus on the conceptually more challenging problem of dealing with the unobserved factors.

3. Weighted average treatment effect: $\tau_{\beta,i} = \beta_i^{(1)} - \beta_i^{(0)}$ where β_i are the regression coefficients on some covariates Z:

$$\beta_i^{(0)} = (Z^\top Z)^{-1} Z^\top C_{i,(T_{0,i}+1):T}^{(0)} \quad \text{and} \quad \beta_i^{(1)} = (Z^\top Z)^{-1} Z^\top C_{i,(T_{0,i}+1):T}^{(1)}.$$

Here, $C_{i,(T_{0,i}+1):T}^{(0)} = \begin{bmatrix} C_{i,T_{0,i}+1}^{(0)} & \cdots & C_{iT}^{(0)} \end{bmatrix}^\top \in \mathbb{R}^{T_{1,i}}$ denotes the observations for $t > T_{0,i}$. The weighted average treatment effect $\tau_{\beta,i}$ generalizes the average treatment effect τ_i, which is a special case for $Z = \vec{1}$. Both tests for the individual treatment effect and the weighted average treatment effect cannot be obtained with conventional estimators, but are important to answer economic questions. For example, in our companion paper Pelger and Xiong (2020), we test if pricing anomalies of investment strategies as measured by their pricing errors persist after these strategies have been published in academic journals. In this problem the treatment is the publication of an investment strategy in a journal and the treatment effect is measured by a change in regression coefficients. More specifically, the pricing error corresponds to the intercept in a regression of the excess returns of the strategies on a set of benchmark risk factors. A simple average treatment effect would not be sufficient to study this question.

For each of the three treatment effects we derive the asymptotic distribution under the null-hypothesis of no effect, which allows us to run one-sided or two-sided hypothesis tests. For example, the two-sided hypothesis test for the weighted average treatment effect takes the form

$$\mathcal{H}_0 : \tau_{\beta,i} = 0, \qquad \mathcal{H}_1 : \tau_{\beta,i} \neq 0. \tag{13}$$

This is the hypothesis we test in our simulation and the empirical companion paper. The problem formulated in Eq. (12) can be solved by applying our latent factor model estimation twice: First, we estimate $C_{it}^{(1)}$ from the treated data with the control observations as missing values. Second, we estimate $C_{it}^{(0)}$ from the control data, while the treated observations are viewed as missing. The inferential theory follows readily from Theorems 2 and 3. The asymptotic variance for the individual treatment effect τ_{it} is the sum of the asymptotic variances of $\hat{C}_{it}^{(0)}$ and $\hat{C}_{it}^{(1)}$ and a covariance term based on the correction terms for the control and treated. While the calculations are tedious, they are a direct consequence of the distribution results that we have derived. The average treatment effects follow then from the results of the individual treatment effects. In this section, we want to focus on a special case, which we consider the most relevant from a practical perspective.

In most causal inference applications, such as the empirical study in our companion paper and Abadie et al. (2010, 2015), the majority of observations are control observations. Hence, it might be infeasible to estimate a latent factor model only from the treated data as required in Eq. (12). For example in the simultaneous treatment case in Table 1, we can estimate a latent factor for the control, but not for the treated. Hence, we impose the additional assumption that the control and treated panel share the same underlying factors, while the loadings can be different, that is,

$$Y_{it}^{(0)} = (\Lambda_i^{(0)})^\top F_t + e_{it}, \qquad Y_{it}^{(1)} = (\Lambda_i^{(1)})^\top F_t + e_{it}. \tag{14}$$

This implies that the treatment can only affect the loadings. This is still a very general setup as the loadings and factors are latent. For example, a model based on Eq. (12) with one factor that changes on the treated data, can be captured in Eq. (14) by a two-factor model where the corresponding loadings change on the treated data.

First, we estimate the factor model from the incomplete control panel $Y^{(0)}$ and obtain $\tilde{C}^{(0)} = (\tilde{\Lambda}^{(0)})^\top \tilde{F}$. Second, we use an ordinary least squares regression to estimate the loadings for the treated $\tilde{\Lambda}_i^{(1)}$,

$$\tilde{\Lambda}_i^{(1)} = \left(\sum_{t=T_{0,i}+1}^{T} \tilde{F}_t \tilde{F}_t^\top \right)^{-1} \sum_{t=T_{0,i}+1}^{T} \tilde{F}_t Y_{it}^{(1)}, \tag{15}$$

which yields an estimate for the common components for the treated panel $\tilde{C}_{it}^{(1)} = (\tilde{\Lambda}_i^{(1)})^\top \tilde{F}_t$.[16]

The following theorem shows the asymptotic distributions for $\tilde{C}_{it}^{(1)}$, the individual treatment effect, and the weighted average treatment effect. The asymptotic distributions allow us to construct test statistics for various treatment effects.

Theorem 5. *Suppose Assumptions S1, G2, G3 and G4 hold and the control and treated panel share the same factors. For $\delta_{NT_i} = \min(N, T_{1,i})$, as $\delta_{NT_i} \to \infty$ the following holds:*

1. *The asymptotic distribution for the common component is*

$$\sqrt{\delta_{NT_i}} (\Sigma_{C,it}^{(1)})^{-1/2} (\tilde{C}_{it}^{(1)} - C_{it}^{(1)}) \xrightarrow{d} \mathcal{N}(0, 1),$$

where

$$\Sigma_{C,it}^{(1)} = F_t^\top \Sigma_F^{-1} \left[\frac{\delta_{NT_i}}{T_{1,i}} \Gamma_{\Lambda,i}^{obs,(1)} + \frac{\delta_{NT_i}}{T} \Gamma_{\Lambda,i}^{miss,(1)} \right] \Sigma_F^{-1} F_t + (\Lambda_i^{(1)})^\top \Sigma_{\Lambda,t}^{-1} \left[\frac{\delta_{NT_i}}{N} \Gamma_{F,t}^{obs} + \frac{\delta_{NT_i}}{T} \Gamma_{F,t}^{miss} \right] \Sigma_{\Lambda,t}^{-1} \Lambda_i^{(1)}$$

---

[16] If units switch between treatment and control, we can modify Eq. (15) to $\tilde{\Lambda}_i^{(1)} = \left(\sum_{t \in \mathcal{S}_i} \tilde{F}_t \tilde{F}_t^\top \right)^{-1} \sum_{t \in \mathcal{S}_i} \tilde{F}_t Y_{it}^{(1)}$, where $\mathcal{S}_i$ is the set of indices for the treated observations.

$$- 2 \cdot \frac{\delta_{NT_i}}{T} F_t^\top \Sigma_F^{-1} \Gamma_{\Lambda,F,i,t}^{\text{miss, cov},(0),(1)} \Sigma_{\Lambda,t}^{-1} \Lambda_i^{(1)},$$

with $\Gamma_{F,t}^{\text{obs}}$ and $\Gamma_{F,t}^{\text{miss}}$ given in Theorem 2, $\Gamma_{\Lambda,i}^{\text{obs},(1)} = \Sigma_{F,e_i}$, $\Gamma_{\Lambda,i}^{\text{miss},(1)} = \Sigma_\Lambda^{-1} \left[\frac{1}{T_{1,i}^2} \sum_{u,s=T_{0,i}+1}^{T} g_{u,s}(\Sigma_{\Lambda,u}^{-1} \Lambda_i^{(1)}, \Sigma_{\Lambda,s}^{-1} \Lambda_i^{(1)}) \right]$
Σ_Λ^{-1}, $\Gamma_{\Lambda,F,i,t}^{\text{miss, cov},(0),(1)} = \Sigma_\Lambda^{-1} \left[\frac{1}{T_{1,i}} \sum_{u=T_{0,i}+1}^{T} g_{u,s}(\Sigma_{\Lambda,u}^{-1} \Lambda_i^{(1)}, \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t) \right]$, and the function $g_{u,s}(\cdot, \cdot)$ is defined in Assumption G4.

2. The asymptotic distribution for the individual treatment effect is

$$\sqrt{\delta_{NT_i}} (\Sigma_{\tau,it})^{-1/2} \left((\tilde{C}_{it}^{(1)} - C_{it}^{(1)}) - (\tilde{C}_{it}^{(0)} - C_{it}^{(0)}) \right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{16}$$

where

$$\Sigma_{\tau,it} = F_t^\top \Sigma_F^{-1} \Gamma_{\Lambda,i}^{\text{obs,miss}} \Sigma_F^{-1} F_t + \left(\Lambda_i^{(1)} - \Lambda_i^{(0)} \right)^\top \Gamma_{F,t}^{\text{obs,miss}} \left(\Lambda_i^{(1)} - \Lambda_i^{(0)} \right) + 2 \cdot F_t^\top \Sigma_F^{-1} \Gamma_{\Lambda,F,i,t}^{\text{miss,cov,diff}} \left(\Lambda_i^{(1)} - \Lambda_i^{(0)} \right)$$

with $\Gamma_{\Lambda,i}^{\text{obs}}$, $\Gamma_{\Lambda,i}^{\text{miss}}$ and $\Gamma_{\Lambda,F,i,t}^{\text{miss, cov}}$ given in Theorem 2, and $\Gamma_{F,t}^{\text{obs,miss}} = \Sigma_{\Lambda,t}^{-1} \left[\frac{\delta_{NT_i}}{N} \Gamma_{F,t}^{\text{obs}} + \frac{\delta_{NT_i}}{T} \Gamma_{F,t}^{\text{miss}} \right] \Sigma_{\Lambda,t}^{-1}$,

$\Gamma_{\Lambda,i}^{\text{obs,miss}} = \frac{\delta_{NT_i}}{T} \Sigma_\Lambda^{-1} \left[\Gamma_{\Lambda,i}^{\text{obs}} + \Gamma_{\Lambda,i}^{\text{miss}} \right] \Sigma_\Lambda^{-1} + \frac{\delta_{NT_i}}{T_{1,i}} \Gamma_{\Lambda,i}^{\text{obs},(1)} + \frac{\delta_{NT_i}}{T} \Gamma_{\Lambda,i}^{\text{miss},(1)} - \frac{\delta_{NT_i}}{T} \left(\Gamma_{\Lambda,\Lambda,i}^{\text{miss, cov},(0),(1)} + (\Gamma_{\Lambda,\Lambda,i}^{\text{miss, cov},(0),(1)})^\top \right)$,
$\Gamma_{\Lambda,F,i,t}^{\text{miss,cov,diff}} = \frac{\delta_{NT_i}}{T} \left(\Sigma_\Lambda^{-1} \Gamma_{\Lambda,F,i,t}^{\text{miss, cov}} - \Gamma_{\Lambda,F,i,t}^{\text{miss, cov},(0),(1)} \right) \Sigma_{\Lambda,t}^{-1}$, $\Gamma_{\Lambda,\Lambda,i}^{\text{miss, cov},(0),(1)} = \Sigma_\Lambda^{-1} \left[\frac{\delta_{NT_i}}{T_{1,i}} \sum_{s=T_{0,i}+1}^{T} g_{i,s}^{\text{cov}}(\Lambda_i^{(0)}, \Sigma_{\Lambda,s}^{-1} \Lambda_i^{(1)}) \right] \Sigma_\Lambda^{-1}$,
and the function $g_{i,s}^{\text{cov}}(\cdot, \cdot)$ is defined in Assumption G3.5.

3. If, in addition, it holds for $Z \in \mathbb{R}^{T_{1,i} \times L}$, $Z^\top Z / T_{1,i} \xrightarrow{P} \Sigma_Z$ and $\frac{1}{T_{1,i}} \sum_{t=T_{0,i}+1}^{T} Z_t F_t^\top \xrightarrow{P} \Sigma_{F,Z}$, then the asymptotic distribution for the weighted average treatment effect is

$$\sqrt{\delta_{NT_i}} (\Sigma_{\beta,t})^{-1/2} \left((\tilde{\beta}_i^{(1)} - \beta_i^{(1)}) - (\tilde{\beta}_i^{(0)} - \beta_i^{(0)}) \right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{17}$$

with

$$\begin{aligned}
\Sigma_{\beta,t} =& \Sigma_Z^{-1} \Sigma_{F,Z} \Sigma_F^{-1} \Gamma_{\Lambda,i}^{\text{obs,miss}} \Sigma_F^{-1} \Sigma_{F,Z}^\top \Sigma_Z^{-1} + \Sigma_Z^{-1} \Gamma_{Z,i}^{\text{miss},\Delta} \Sigma_Z^{-1} \\
&+ \frac{\delta_{NT_i}}{T} \Sigma_Z^{-1} \left[\Sigma_{F,Z} \Sigma_F^{-1} \Sigma_\Lambda^{-1} \Gamma_{\Lambda,Z,i}^{\text{miss, cov},(0),\Delta} + (\Gamma_{\Lambda,Z,i}^{\text{miss, cov},(0),\Delta})^\top \cdot \Sigma_\Lambda^{-1} \Sigma_F^{-1} \Sigma_{F,Z}^\top \right] \Sigma_Z^{-1} \\
&- \frac{\delta_{NT_i}}{T} \Sigma_Z^{-1} \left[\Sigma_{F,Z} \Sigma_F^{-1} \cdot \Gamma_{\Lambda,Z,i}^{\text{miss, cov},(1),\Delta} + (\Gamma_{\Lambda,Z,i}^{\text{miss, cov},(1),\Delta})^\top \cdot \Sigma_F^{-1} \Sigma_{F,Z}^\top \right] \Sigma_Z^{-1},
\end{aligned}$$

with $\Gamma_{\Lambda,Z,i}^{\text{miss, cov},(0),\Delta} = \left[\frac{1}{T_{1,i}} \sum_{s=T_{0,i}+1}^{T} g_{i,s}^{\text{cov}}(\Lambda_i^{(0)}, \Sigma_{\Lambda,s}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})) \right] \Sigma_\Lambda^{-1} \Sigma_F^{-1} \Sigma_{F,Z}$,
$\Gamma_{\Lambda,Z,i}^{\text{miss, cov},(1),\Delta} = \Sigma_\Lambda^{-1} \left[\frac{1}{T_{1,i}^2} \sum_{u,s=T_{0,i}+1}^{T} g_{u,s}(\Sigma_{\Lambda,u}^{-1} \Lambda_i^{(1)}, \Sigma_{\Lambda,s}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})) \right] \Sigma_\Lambda^{-1} \Sigma_F^{-1} \Sigma_{F,Z}$,
and $\Gamma_{\Lambda,Z,i}^{\text{miss},\Delta} = \Sigma_{F,Z}^\top \Sigma_F^{-1} \Sigma_\Lambda^{-1} \left[\frac{1}{T_{1,i}^2} \sum_{u,s=T_{0,i}+1}^{T} g_{u,s}(\Sigma_{\Lambda,u}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)}), \Sigma_{\Lambda,s}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})) \right] \Sigma_\Lambda^{-1} \Sigma_F^{-1} \Sigma_{F,Z}$, $g_{i,s}^{\text{cov}}(\cdot, \cdot)$ and the functions $g_{u,s}(\cdot, \cdot)$ are defined in Assumptions G3.5 and G4.

Suppose Assumptions S1, C1, G2, GC2, GC3 and GC4 hold. The above three results hold for the propensity weighted estimator after replacing $\Gamma_{F,t}^{\text{obs}}$, $\Gamma_{\Lambda,i}^{\text{obs}}$, $\Gamma_{F,t}^{\text{miss}}$, $\Gamma_{\Lambda,i}^{\text{miss}}$, $\Sigma_{\Lambda,t}$, $g_{i,s}^{\text{cov}}(\cdot, \cdot)$ and $g_{u,s}(\cdot, \cdot)$ with $\Gamma_{F,t}^{\text{obs},S}$, $\Gamma_{\Lambda,i}^{\text{obs},S}$, $\Gamma_{F,t}^{\text{miss},S}$, $\Gamma_{\Lambda,i}^{\text{miss},S}$, Σ_Λ, $g_{i,s}^{\text{cov},S}(\cdot, \cdot)$ and $g_{u,s}^S(\cdot, \cdot)$.

The results of Theorem 5 are a consequence of Theorems 2 and 3. The challenge arises from correctly capturing the asymptotic covariance between the estimated treated and control common components. This additional covariance term is due to the correction terms from the missing observations. In Theorem 5, we impose the additional Assumption G4 for the general estimator and Assumption GC4 for the probability-weighted estimator. Both simply state that the conventional central limit theorems based on the weak dependencies in the errors apply to the subset of treated time periods. These conditions are automatically satisfied in our simplified model and thus can be neglected, as stated in Proposition 3 in Appendix A. As expected, the rate of the asymptotic distribution results depends on the number of treated time periods $T_{1,i}$, as the loadings are estimated from time-series regressions on the treated panel. Note that Assumption S3 implies that the number of control time series observations $T_{0,i}$ is of the order T.

Our analysis focuses on treatment tests for individual units. An interesting extension left for future research would be aggregated test statistics for the joint null hypothesis, that the treatment effect for all cross-sectional units is zero. While our results lay the foundation for this extension, there are a number of additional challenges that need to be overcome. As the number of cross-sectional units is asymptotically large, there are different possible joint hypotheses. The strongest statement is based on a sup norm of τ_i or τ_{it} and tests if every single unit is not affected by treatment. Such a unified test statistic requires uniform distribution results, which are stronger than our pointwise results. A weaker statement would take the form of a Wald type test statistic and allows a finite number of units to have non-zero treatment effects.[17]

---

[17] Fan et al. (2015) provide a discussion on the challenges of testing large dimensional hypotheses.

8. Feasible estimation and testing

Theorems 2, 3 and 5 are formulated with respect to the asymptotic covariances based on the population model. In order to use them in practice we need feasible estimators for the covariance terms. We propose to use the plug-in estimators $\tilde{F}_t$, $\tilde{\Lambda}_i$ and $\tilde{e}_{it} = Y_{it} - \tilde{\Lambda}_i^\top \tilde{F}_t$ for $(H^{-1})^\top F_t$, $H\Lambda_i$ and e_{it}. All moments are based on these three objects. For example $\hat{\Sigma}_F := \frac{1}{T}\tilde{F}^\top \tilde{F}$ consistently estimates $(H^{-1})^\top \Sigma_F (H^{-1})$. The rotation matrix H can be ignored in the estimated covariances of the common components and the treatment effects as it cancels out. It is only the distribution of the loadings and factors that are estimated up to a rotation matrix. The challenge is to deal with the time and cross-sectional dependency in the residuals. We impose the additional assumption that the time-series and cross-section covariance matrices of the errors e_{it} are sparse in the sense that only a finite number of row elements are non-zero and we know the indices of the non-zero elements. More specifically we define

$$\mathcal{E}_t = \{i, j : \mathbb{E}[e_{it}e_{jt}] \neq 0\} \qquad \mathcal{E} = \{i, j, s, t : \mathbb{E}[e_{it}e_{js}] \neq 0\}$$

and assume that $|\mathcal{E}_t| = O(N)$ and $|\mathcal{E}| = O(NT)$. The estimator for $H\Gamma^{\text{obs}}_{\Lambda,j}H^\top$ and $H\Gamma^{\text{obs}}_{F,t}H^\top$ depends on the dependency structure in the residuals and we propose the plug-in estimator based on only the non-zero moments of the residuals:

$$\widehat{\Gamma^{\text{obs}}}_{F,t} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{it}W_{jt}\tilde{\Lambda}_i\tilde{\Lambda}_j^\top \tilde{e}_{it}\tilde{e}_{jt}\mathbb{1}_{\{i,j\in\mathcal{E}_t\}}$$

$$\widehat{\Gamma^{\text{obs}}}_{\Lambda,j} = \frac{T}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\tilde{\Lambda}_i\tilde{\Lambda}_i^\top \frac{1}{|\mathcal{Q}_{ij}||\mathcal{Q}_{kj}|}\sum_{t,s\in\mathcal{Q}_{ij}}\tilde{F}_t\tilde{F}_s^\top \tilde{\Lambda}_k\tilde{\Lambda}_k^\top \tilde{e}_{it}\tilde{e}_{ks}\mathbb{1}_{\{i,k,s,t\in\mathcal{E}\}}.$$

The estimators are analogous for the probability-weighted estimator. A special case is the estimation approach in Bai (2003) that assumes independence of the residuals over time and the cross-section and hence only uses the diagonal entries of the residual covariance and autocovariance matrix. Instead of assuming knowledge of the non-zero entries, it is possible to generalize the estimator similar to Fan et al. (2013) and estimate the non-zero entries with a thresholding estimation approach. We propose a HAC estimator for $\Gamma^{\text{miss}}_{\Lambda,j}$, $\Gamma^{\text{miss}}_{F,t}$ and $\Gamma^{\text{miss, cov}}_{\Lambda,F,j,t}$ to account for the time-series dependency in the factors similar to Bai (2003).

Proposition 2. *Suppose that the assumptions of Theorems 2, 3 or 5 hold. In addition, we assume that the time-series and cross-section covariance matrices of the errors e_{it} are sparse in the sense that $|\mathcal{E}_t| = O(N)$ and $|\mathcal{E}| = O(NT)$ and we know the non-zero elements. Then, the plug-in estimators of the asymptotic covariances in Theorems 2, 3 and 5 are consistent and the asymptotic statements in the respective theorems continue to hold with the estimated covariance matrices.*

Hence, the treatment effects normalized by their estimated standard deviations follow asymptotically a standard normal distribution, and we obtain feasible test statistics for the various treatment effects.

9. Generalization of the missing patterns

Our results can be generalized to the case where the number of observed entries is not proportional to N or T but grows at a strictly smaller rate. The general arguments of the proofs stay the same but we need to carefully account for the convergence rates of each term based on the set $\mathcal{Q}_{ij}$. The mean squared consistency of the estimated loadings in Theorem 1 generalizes to

$$\frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j - H\Lambda_j\right\|^2 = O_p\left(\max\left(\frac{1}{N},\ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{|\mathcal{Q}_{ij}|}\right)\right).$$

Moreover, we can show the asymptotic normality of the estimated loadings $\tilde{\Lambda}$, factors $\tilde{F}$ from the equally weighted regression (3), and common components $\tilde{C}$ under similar assumptions as those in Theorem 2. The estimated loadings $\tilde{\Lambda}_j$ are asymptotically normal with convergence rate

$$H^{-1}\tilde{\Lambda}_j - \Lambda_j = O_p\left(\left[\max\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{Q}_{ij}|},\ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{lj}|}{|\mathcal{Q}_{ij}||\mathcal{Q}_{lj}|}\right)\right]^{1/2}\right),$$

where the second term is closely related to ω_{jj} defined in Assumption S3. The estimated factors $\tilde{F}_t$ are asymptotically normal with convergence rate

$$H^\top\tilde{F}_t - F_t = O_p\left(\left[\max\left(\frac{1}{\sum_{i=1}^{N}W_{it}},\ \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{kl}|}{|\mathcal{Q}_{ij}||\mathcal{Q}_{kl}|}\right)\right]^{1/2}\right),$$

where the second term is closely related to ω defined in Assumption S3. Similarly, by combining the rates of estimated factors and loadings, the estimated common components $\tilde{C}_{it}$ have an asymptotic normal distribution with rate

$$
\tilde{C}_{jt} - C_{jt} = O_p\left(\left[\max\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{Q}_{ij}|}, \; \frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{lj}|}{|\mathcal{Q}_{ij}||\mathcal{Q}_{lj}|}, \; \frac{1}{\sum_{i=1}^{N}W_{it}}, \right.\right.\right.
$$
$$
\left.\left.\left. \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{kl}|}{|\mathcal{Q}_{ij}||\mathcal{Q}_{kl}|}, \; \frac{1}{N^3}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{kl}|}{|\mathcal{Q}_{ij}||\mathcal{Q}_{kl}|}\right)\right]^{1/2}\right).
$$

The last term is closely related to ω_j defined in Assumption S3. The expression for the asymptotic covariances of the estimators become more complex. The proofs for the consistency and asymptotic normality for the general case, when observed entries are not proportional to N and T, are very similar to the proofs of Theorems 1 and 2, but just require carefully keeping track of the convergence rates of each term.[18]

We illustrate the more general convergence rates in the simultaneous treatment observation pattern in Table 1(a), where we can provide explicit expressions for the different rates. The mean square consistency result of the loadings simplifies to

$$
\frac{1}{N}\sum_{j=1}^{N}\left\|\tilde{\Lambda}_j - H\Lambda_j\right\|^2 = O_p\left(\max\left(\frac{1}{N}, \; \frac{N_0}{NT_0}, \; \frac{1}{T}\right)\right).
$$

We obtain two different convergence rates for the estimated loadings:

$$
H^{-1}\tilde{\Lambda}_j - \Lambda_j = \begin{cases} O_P\left(\frac{1}{\sqrt{T_0}}\right) & j \le N_0 \\ O_P\left(\max\left(\sqrt{\frac{N_0}{NT_0}}, \frac{1}{\sqrt{T}}\right)\right) & j > N_0. \end{cases}
$$

Similarly, the estimated factors have two different convergence rates depending on which time block we consider:

$$
H^\top\tilde{F}_t - F_t = \begin{cases} O_P\left(\max\left(\frac{1}{\sqrt{N}}, \frac{N_0}{N\sqrt{T_0}}, \frac{1}{\sqrt{T}}\right)\right) & t \le T_0 \\ O_P\left(\max\left(\frac{1}{\sqrt{N-N_0}}, \frac{N_0}{N\sqrt{T_0}}, \frac{1}{\sqrt{T}}\right)\right) & t > T_0. \end{cases}
$$

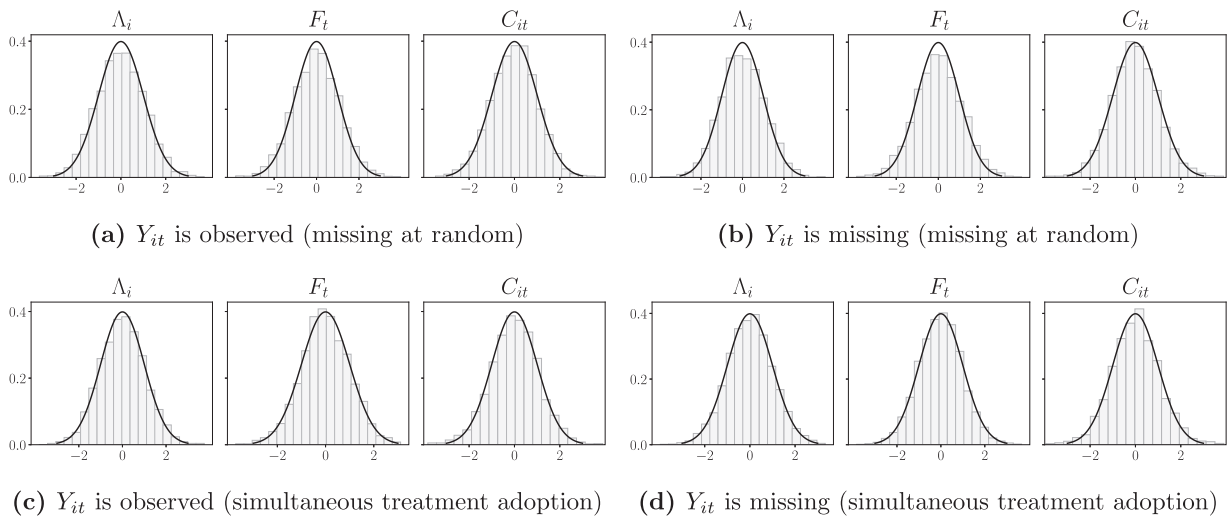This results in four different convergence rates for each block for the estimated common components:

$$
\tilde{C}_{jt} - C_{jt} = \begin{cases} O_P\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{T_0}}\right)\right) & j \le N_0, \; t \le T_0 \\ O_P\left(\max\left(\frac{1}{\sqrt{N}}, \sqrt{\frac{N_0}{NT_0}}, \frac{1}{\sqrt{T}}\right)\right) & j > N_0, \; t \le T_0 \\ O_P\left(\max\left(\frac{1}{\sqrt{N-N_0}}, \frac{1}{\sqrt{T_0}}\right)\right) & j \le N_0, \; t > T_0 \\ O_P\left(\max\left(\frac{1}{\sqrt{N-N_0}}, \sqrt{\frac{N_0}{NT_0}}, \frac{1}{\sqrt{T}}\right)\right) & j > N_0, \; t > T_0. \end{cases}
$$

10. Simulation

10.1. Asymptotic distributions

In this section, we demonstrate the finite sample properties of our asymptotic results for both the observed entries and the missing entries. We confirm the theoretical distribution results for the estimated factor, loadings, common components, and treatment effects. We generate the data from a one-factor model $X_{it} = \Lambda_i F_t + e_{it}$, where $F_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $\Lambda_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The observation pattern depends on unit-specific characteristics $S_i = \mathbb{1}(\Lambda_i \ge 0)$, which are a function of the factor loadings. We study two observation patterns which are illustrated in Figs. 1(a) and 1(b):

1. *Missing at random*: Entries are observed independently with probability 0.75 if $S_i = 1$, and 0.5 if $S_i = 0$.
2. *Simultaneous treatment adoption*: Once a unit adopts treatment, it stays treated afterward. For the units with $S_i = 1$, 25% randomly selected units adopt the treatment from time $0.75 \cdot T$ and the remaining 75% units stay in the control group until the end. For the units with $S_i = 0$, 62.5% randomly selected units adopt the treatment from time $0.375 \cdot T$ and the remaining 37.5% units stay in the control group until the end. We model the treated data as missing.

(a) Y_{it} is observed (missing at random)   **(b)** Y_{it} is missing (missing at random)

(c) Y_{it} is observed (simultaneous treatment adoption)   **(d)** Y_{it} is missing (simultaneous treatment adoption)

Fig. 2. Histograms of standardized loadings, factors, and common components.
These figures show the histograms of estimated standardized loadings, factors, and common components normalized by their estimated standard divisions, where $N = 100$ and $T = 150$. The normal density function is superimposed on the histograms. The results are based on 2000 Monte Carlo simulations. The Internet Appendix collects the histograms for other specifications of N and T.

(a) Treatment effect $\tau = 0$

(b) Treatment effect $\tau = 0.25$
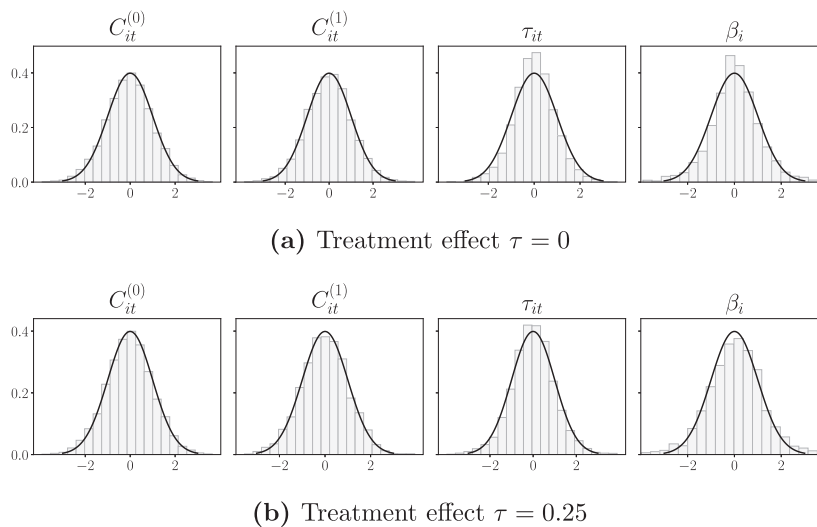
Fig. 3. Histograms of standardized control and treated common components, individual and average treatment effects.
These figures show the histograms of estimated control and treated common components, individual and average treatment effect ($Z = \vec{1}$) normalized by their estimated standard deviations, where $N = 100$ and $T = 150$. The normal density function is superimposed on the histograms. The observation pattern follows the simultaneous treatment adoption pattern. The results are based on 2000 Monte Carlo simulations. The Internet Appendix collects the histograms for other specifications of N and T.

To conserve space, we report here the distribution results for the regression based estimator based on Eq. (3), but the results extend to the propensity-weighted estimator. Fig. 2 shows the histograms of standardized factors, loadings, and common components for randomly selected observed entries and missing entries based on Theorem 2. The histograms match the standard normal density function very well and support the validity of our asymptotic results in finite samples.

Fig. 3 confirms that our treatment test in Theorem 5 has the correct size. The control data follows our benchmark one-factor model. We assume a constant treatment effect, i.e., $\Lambda_i^{(1)} = \Lambda_i^{(0)} + \tau$, where τ is set to 0 or 0.25. Fig. 3 shows the histograms of standardized common components for treated and control, the individual treatment effect, and an equally weighted treatment effect for randomly selected units and times. As expected, the histograms support the validity of our asymptotic results in finite samples.

---

[18] The proofs are available upon request.

Table 3
Statistical power of treatment effect tests.

N	T	$\Lambda_i^{(1)} - \Lambda_i^{(0)}$	$\tilde{C}_{it}^{(1)} - \tilde{C}_{it}^{(0)}$				$\tilde{\beta}_i^{(1)} - \tilde{\beta}_i^{(0)}$			
		μ_F	0.25	0.50	1.00	2.00	0.25	0.50	1.00	2.00
100	100	0.900	0.171	0.450	0.902	0.991	0.198	0.440	0.864	0.968
		0.000	0.271	0.660	0.946	0.996	0.271	0.654	0.939	0.996
		0.500	0.347	0.835	0.981	1.000	0.345	0.831	0.979	0.998
		0.000	0.466	0.906	0.991	1.000	0.481	0.910	0.991	1.000
250	100	0.900	0.175	0.464	0.896	0.994	0.165	0.456	0.866	0.981
		0.000	0.273	0.722	0.959	0.998	0.271	0.731	0.954	0.998
	500	0.900	0.572	0.954	1.000	1.000	0.558	0.931	0.983	0.989
		0.000	0.764	0.970	1.000	1.000	0.772	0.979	1.000	1.000
500	500	0.900	0.610	0.975	1.000	1.000	0.591	0.958	0.998	1.000
		0.000	0.805	0.987	1.000	1.000	0.809	0.987	1.000	1.000
	1000	0.900	0.860	0.992	1.000	1.000	0.848	0.983	0.998	0.998
		0.000	0.959	1.000	1.000	1.000	0.957	1.000	1.000	1.000

This table shows the proportion of test statistics of the treatment effect that reject the null hypotheses $\mathcal{H}_0 : C_{it}^{(1)} - C_{it}^{(0)} = 0$ or $\mathcal{H}_0 : \beta_i^{(1)} - \beta_i^{(0)} = 0$, where $\beta_i^{(1)} = \frac{1}{T_{1,i}} \sum_{T_{0,i}+1}^{T} C_{it}^{(1)}$ and $\beta_i^{(0)} = \frac{1}{T_{1,i}} \sum_{T_{0,i}+1}^{T} C_{it}^{(0)}$. We consider a 95% confidence level (the test statistics are within $[-1.96, 1.96]$) over 500 Monte Carlo simulations. The test statistics normalize $\tilde{C}_{it}^{(1)} - \tilde{C}_{it}^{(0)}$ and $\tilde{\beta}_i^{(1)} - \tilde{\beta}_i^{(0)}$ with their estimated standard deviation from Eqs. (16) and (17). The estimated standard deviations are estimated under the null hypothesis of $\Lambda_i^{(1)} - \Lambda_i^{(0)} = 0$. The observation pattern follows the simultaneous treatment adoption pattern. The proportion of acceptance decreases with N, T, μ_F and $\tilde{\beta}_i^{(1)} - \tilde{\beta}_i^{(0)}$, implying that the statistical power increases with the data dimensionality and the scale of the treatment effect. The Internet Appendix collects additional robustness tests confirming the same findings for different specifications and also showing that the statistical power increases with the proportion of observed entries in the data.

Table 3 demonstrates the statistical power of our tests for individual and average treatment effects, where the null hypotheses are $\mathcal{H}_0 : \beta_i^{(1)} - \beta_i^{(0)} = 0$ with equal weights for all time periods, i.e., $\beta_i^{(1)} = \tau_i^{(1)}$ and $\beta_i^{(0)} = \tau_i^{(0)}$. The power increases with the data dimensionality (N and T) and the scale of treatment effect that is determined by the mean of the factor μ_F and the difference between the control and treated loadings $\Lambda_i^{(1)} - \Lambda_i^{(0)}$. The null hypothesis implies $\Lambda_i^{(1)} - \Lambda_i^{(0)} = 0$, which we use in the estimation of the asymptotic variance. This slightly improves the power, but the results in the Internet Appendix show that we also have good power properties without imposing the null hypothesis in the estimation of the asymptotic covariances. Moreover, the statistical power increases with the proportion of observed entries, as shown in the comparison between Tables 3 and A.IX in the Internet Appendix.

10.2. Robustness to missing patterns

In this section, we show that our benchmark regression-based estimator (denoted as XP) and propensity-weighted estimator (denoted as XP$_{\text{PROP}}$) perform well under a variety of missing patterns. As reference we also include the estimators of Jin et al. (2021) (denoted as JMS) and Bai and Ng (2021) (denoted as BN). Each of the two estimators is designed for a specific observation pattern and hence provides a natural reference level for that specific pattern. Jin et al. (2021) assume that observations are missing at random, while Bai and Ng (2021) is tailored to an observation pattern with a block structure after proper reshuffling. These are the four estimation approaches that provide an inferential theory for imputed common components in an approximate factor model and were available at the time of submission of this paper.

We generate the data from a two-factor model $X_{it} = \Lambda_i^\top F_t + e_{it}$, where $F_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2)$, $\Lambda_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2)$ and $e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We consider six different observation patterns. The first three cases are (1) missing uniformly at random, (2) simultaneous treatment adoption, and (3) staggered treatment adoption. Then, we allow the observation pattern for these three cases to depend on a unit-specific characteristic defined as $S_i = \mathbb{1}(\Lambda_{i,2} \geq 0)$. Hence, case four to six are (4) missing at random conditional on S_i, (5) simultaneous treatment adoption conditional on S_i, (6) staggered treatment adoption conditional on S_i. Table 4 contains figures showing the observation patterns and their detailed descriptions. Note that these are all practically relevant patterns, in particular the staggered treatment adoption that appears in our empirical companion paper and is prevalent in empirical applications.

Table 4 compares the performance of estimating the common components. We report the normalized mean squared error (MSE) of the four methods for observed, missing and all units defined as follows:

$$\text{relative MSE}_\mathcal{S} = \frac{\sum_{(i,t) \in \mathcal{S}} (\tilde{C}_{it} - C_{it})^2}{\sum_{(i,t) \in \mathcal{S}} C_{it}^2},$$

where $\mathcal{S}$ is either the set observed, missing or all observations.

First, and most importantly, our benchmark estimator shows excellent performance for all observation patterns. Our estimator has the smallest or at least a very similar small MSE compared to the other methods, as indicated by the

Table 4
Robustness to missing patterns.

	Observation pattern	W_{it}	XP	XP$_{PROP}$	JMS	BN
	Random	obs	**0.015**	**0.015**	0.023	–
		miss	**0.015**	**0.015**	0.021	–
		all	**0.015**	**0.015**	0.023	–
	Simultaneous	obs	**0.012**	**0.012**	0.124	**0.012**
		miss	0.020	0.020	0.184	**0.017**
		all	0.014	0.014	0.139	**0.013**
	Staggered	obs	**0.017**	**0.017**	0.366	0.073
		miss	**0.043**	**0.043**	0.318	0.087
		all	**0.027**	**0.027**	0.347	0.078
	Random W depends on S	obs	**0.019**	0.020	0.077	–
		miss	**0.024**	0.024	0.067	–
		all	**0.021**	0.021	0.073	–
	Simultaneous W depends on S	obs	**0.032**	0.040	0.703	0.141
		miss	**0.231**	0.256	0.521	0.279
		all	**0.129**	0.145	0.615	0.209
	Staggered W depends on S	obs	**0.016**	0.018	0.272	0.117
		miss	**0.064**	0.069	0.346	0.186
		all	**0.033**	0.036	0.299	0.142

This table reports the relative MSE of XP (our benchmark method), XP$_{PROP}$ (our propensity-weighted method), JMS (Jin et al., 2021), and BN (Bai and Ng, 2021) on observed, missing and all entries, $N = 250$, $T = 250$. The figures on the left show patterns of missing observations with the shaded entries indicating the missing entries. Bold numbers indicate the best relative model performance. We generate a two-factor model and a unit-specific characteristic $S_i = \mathbb{1}(\Lambda_{i,2} \geq 0)$. The observation patterns are generated as follows. (1) *Missing uniformly at random*: Entries are observed independently with probability $p = 0.75$. (2) *Simultaneous treatment adoption*: 50% randomly selected units adopt the treatment from time $0.5 \cdot T$ and the remaining 50% units stay in the control group until the end. (3) *Staggered treatment adoption*: All units are in the control group for $t < 0.1 \cdot T$. At time $0.1 \cdot T \leq t \leq T$, $\frac{t - 0.1 \cdot T}{T}$% units are in the treated group. The remaining 10% units stay in the control group until the end. (4) *Missing at random conditional on S_i*: Entries are observed independently with probability $p_{it} = 0.75$ $S_i = 1$, and $p_{it} = 0.5$ if $S_i = 0$. (5) *Simultaneous treatment adoption conditional on S_i*: For the units with $S_i = 1$, 95% units adopt the treatment from time $0.5 \cdot T$ and 5% units stay in the control group until the end. For the units with $S_i = 0$, 50% units adopt the treatment from time $0.02 \cdot T$ and 50% units stay in the control group until the end. (6) *Staggered treatment adoption conditional on S_i*: All units are in the control group for $t < 0.02 \cdot T$. For the units with $S_i = 1$, at time $0.02 \cdot T \leq t \leq T$, $\frac{t - 0.02 \cdot T}{T}$% units are in the treated group with the remaining 2% staying in the control group until the end. For the units with $S_i = 0$, at time $0.02 \cdot T \leq t \leq T$, $\frac{t - 0.02 \cdot T}{1.96T}$% units are in the treated group with the remaining 50% staying in the control group until the end. We run 100 Monte Carlo simulations. The Internet Appendix collects additional robustness results with the same findings.

bold numbers. Hence, we view our approach as a simple and reliable all-purpose estimator. Our propensity-weighted estimator is very close to the benchmark estimator but performs slightly worse. This is in line with our theoretical result that propensity weighting is generally less efficient.

In the case of missing at random conditional or unconditional on S_i, our methods have the smallest MSE. Jin et al. (2021) also have a small MSE as long as the observation pattern does not depend on S_i as their method is designed for missing uniformly at random. Missing at random violates the assumptions of Bai and Ng (2021) and their estimator not applicable as there not sufficiently large blocks of fully observed entries.

In the case of simultaneous treatment adoption, Bai and Ng (2021) have the smallest MSE as their method is tailored to this case. Interestingly, our method as an all-purpose estimator is very close to Bai and Ng (2021). When the observation pattern depends on S, it can shrink the size of fully observed blocks, which increases the importance of using all observed entries resulting in the smallest MSE for our method. In the case of simultaneous treatment adoption, the assumptions in Jin et al. (2021) are violated, which is reflected in the larger MSE.

Our methods have the smallest MSE for the case of staggered treatment adoption that is prevalent in empirical applications (Athey and Imbens, 2021). This holds whether the observation pattern depends or does not depend on S. In contrast to Bai and Ng (2021), we use all observed entries in the estimation, which provides a more efficient estimator. Note that in this simulation example the fully observed blocks are very small, and hence, similar to the missing at random case, the assumptions in Bai and Ng (2021) might not be satisfied. As the assumptions of Jin et al. (2021) are violated, their imputation results in larger errors.

The Internet Appendix shows that the findings are robust to the size of the panel and the parameters of the observation patterns. We also compare the MSE of the various methods after iterations in Tables A.I–A.III in the Internet Appendix. In more detail, we first impute the missing values with different methods. In the second step, we apply PCA to the full panel with imputed values to estimate the factor model and update the imputed values with the estimated common components. The observed entries stay the same. This process is repeated for multiple iterations. Note that this iterated estimation approach is actually a different estimation approach by itself. The four methods provide different starting values for the same iterative estimation approach that is based on a fixed-point argument. Importantly, there is no

Table 5
Key parameters for different observation patterns.

	p	ω	ω_j		ω_{jj}			p	ω	ω_j		ω_{jj}	
			ω_1	ω_N	ω_{11}	ω_{NN}				ω_1	ω_N	ω_{11}	ω_{NN}
	60%	1.00	1.00	1.00	1.67	1.67		60%	1.00	1.00	1.00	1.36	2.16
	75%	1.00	1.00	1.00	1.33	1.33		75%	1.00	1.00	1.00	1.19	1.52
	90%	1.00	1.00	1.00	1.11	1.11		90%	1.00	1.00	1.00	1.07	1.17
	60%	2.27	1.94	2.49	1.68	2.72		60%	7.20	4.69	4.90	3.16	3.54
	75%	1.55	1.38	1.76	1.25	2.00		75%	3.18	2.27	2.39	1.69	1.92
	90%	1.13	1.08	1.25	1.05	1.46		90%	1.36	1.24	1.29	1.12	1.21
	60%	2.10	1.76	3.17	1.51	9.08		60%	2.05	1.73	2.95	1.50	7.29
	75%	1.45	1.30	1.86	1.19	3.41		75%	1.43	1.30	1.72	1.20	2.65
	90%	1.11	1.07	1.26	1.04	1.81		90%	1.12	1.08	1.21	1.05	1.46

This table shows the values of ω, ω_j, and ω_{jj} (defined in Assumption S3) for different missing patterns with the same proportion of missing values. The subplots show the observation patterns, which follow the same structure as in Table 4. The fraction of observed entries p takes the values 60%, 75% and 90%. We report ω_j and ω_{jj} for $j = 1$ and $j = N$, where the unit $j = N$ has more missing values. The parameters ω, ω_j, and ω_{jj} capture the weights of the variance correction terms in Corollary 1.

inferential theory for iterative estimators under general patterns.[19] Hence, if the goal is to estimate treatment effects, these iterative estimators cannot be used. Since our methods start with a value that has a smaller MSE, our methods, in general, converge faster (often already after three iterations) and also have a small MSE for a fixed number of iterations. Our results are robust to the choice of N and T and we present the corresponding results for $N = 100$ and $T = 150$ in Tables A.IV–A.VII in the Internet Appendix. In summary, if the goal is to only minimize the imputation error without an inferential theory, the iterative estimation generally improves the results, but the relative performance of the different estimation approaches without iteration carries over to the iteration setup.

10.3. The effect of different missing patterns

The observation patterns affect the asymptotic variance of the estimated models. Our distribution theory can capture this effect in a transparent and intuitive way. The impact of the missing pattern on the asymptotic variances of the estimators is captured by the three key parameters ω, ω_j and ω_{jj}. In the simplified model, these parameters are sufficient to capture all the information about the observation pattern. Corollary 1 shows that the asymptotic distribution of the all-purpose estimator is identical if different missing patterns have the same parameters $\omega_{jj}, \omega_j, \omega$.

Table 5 shows the values of ω, ω_j, and ω_{jj} for different missing patterns with the same proportion of missing values. The subplots show the observation patterns, which follow the same structure as in Table 4. The fraction of observed entries p takes the values 60%, 75% and 90%. The parameters ω, ω_j, and ω_{jj} increase with a larger fraction of missing observations. We confirm, that when observations are missing at random with probability $1 - p$, we have $\omega_{jj} = \frac{1}{1-p}$, $\omega_j = 1$ and $\omega = 1$.

The asymptotic variance of the factors in Corollary 1 depends only on ω. We observe that ω increases with a stronger dependency between the rows of the observation patterns. In the special case of $\omega = 1$ the correction term in the asymptotic factor variance vanishes. Hence, if the missing pattern is close to missing at random or the proportion of observed values is sufficiently large, the variance correction terms due to missingness becomes less relevant and using only the conventional asymptotic variance based on fully observed data might provide a sufficiently good approximation.

The asymptotic variance of the loadings and common components in Corollary 1 depend on ω_j and ω_{jj}, which are unit specific. In our simulation the unit $j = 1$, which is the first row in the subplots, has the most observed entries, while unit $j = N$, which is the last row, has less observed values. The effect of missingness is heterogeneous, and, as expected, has a stronger effect on units with less observed values. This effect can be substantial as in the case of a staggered treatment design. We conclude that even for the same proportion of observed values, the type of missing pattern has a crucial impact on the asymptotic distribution of the estimated latent factor model.

10.4. Misspecification and robustness of propensity-weighted estimator

In this section, we show that the propensity-weighted estimator can have desirable robustness properties under misspecification. Our results are motivated by insights from causal inference that propose doubly robust estimation procedures for missing values, as discussed, for example, in Kang and Schafer (2007). In causal inference, we can either model the relationship between the covariates and the outcome or model the probabilities of missingness to estimate causal effects. Doubly robust procedures combine both by using a propensity weight in regressions to mitigate the selection bias. Their potential advantage is that they can provide reliable estimates in the case of omitted variables. Our

---

[19] While Jin et al. (2021) consider iterations, their asymptotic results only hold for missing at random. Bai and Ng (2021) provide distribution results for a different iteration that is not making use of all observations and therefore only has a minor effect.

Table 6
Benchmark and propensity-weighted estimator for weak and missing factors.

k estimated factors	1				2			
$[\mu_{F,1}, \mu_{F,2}]$	[1,1]		[5,0.5]		[1,1]		[5, 0.5]	
$[\sigma_{F,1}, \sigma_{F,2}]$	[1,1]		[5,0.5]		[1,1]		[5, 0.5]	
Method	XP	XP$_{\text{PROP}}$	XP	XP$_{\text{PROP}}$	XP	XP$_{\text{PROP}}$	XP	XP$_{\text{PROP}}$
obs $C_{it}^{(0)}$	**0.227**	0.251	0.011	0.011	0.014	0.014	**0.002**	0.003
miss $C_{it}^{(0)}$	0.478	**0.288**	0.007	0.007	**0.044**	0.045	0.026	**0.023**
all $C_{it}^{(0)}$	0.314	**0.264**	0.009	0.009	**0.024**	0.025	0.014	**0.012**
obs $C_{it}^{(0)}(S=1)$	**0.184**	0.254	**0.755**	0.761	0.013	0.013	**0.122**	0.125
miss $C_{it}^{(0)}(S=1)$	**0.046**	0.261	**0.751**	0.769	0.019	0.019	**0.123**	0.132
obs $C_{it}^{(0)}(S=0)$	0.304	**0.268**	0.001	**0.000**	0.016	0.016	0.001	**0.001**
miss $C_{it}^{(0)}(S=0)$	0.721	**0.308**	0.003	**0.002**	0.059	0.059	0.025	**0.022**
obs $C_{it}^{(1)}$	0.402	**0.278**	0.007	**0.006**	0.037	**0.036**	**0.002**	0.003
$C_{it}^{(1)} - C_{it}^{(0)}$	0.481	**0.294**	0.008	**0.007**	0.052	0.052	0.026	**0.023**
$\beta_i^{(1)} - \beta_i^{(0)}$	0.168	**0.032**	0.002	0.002	**0.012**	0.013	0.008	**0.007**
ATE	0.090	**0.026**	**0.006**	0.007	0.009	**0.008**	0.012	**0.011**

This table compares the percentage errors for various estimates with the benchmark estimator (XP) and the propensity weighted estimator XP$_{\text{PROP}}$ for omitted and weak factors. The data is simulated with a two-factor model and a simultaneous treatment adoption for different means and variances of the latent factors. **For $k = 1$ one factor is omitted in the estimation as the population model is a two-factor model. For $[\sigma_{F,1}, \sigma_{F,2}] = [5, 0.5]$ the second factor is weak**. In more detail: $Y_{it}^{(0)} = \Lambda_{i,1}^{(0)} F_{t,1} + \Lambda_{i,2}^{(0)} F_{t,2} + e_{it}^{(0)}$ and $Y_{it}^{(1)} = \Lambda_{i,1}^{(1)} F_{t,1} + \Lambda_{i,2}^{(1)} F_{t,2} + e_{it}^{(1)}$. The first half of the cross-section depends on the first factor, while the second half depends on the second factor: For $i = 1, \ldots, N/2$, $\Lambda_{i,1}^{(0)} \sim \mathcal{N}(0, 1)$, $\Lambda_{i,1}^{(1)} = \Lambda_{i,1}^{(0)} + \mathcal{N}(0.2, 1)$ and $\Lambda_{i,2}^{(1)} = \Lambda_{i,2}^{(0)} = 0$, and for $i = N/2+1, \ldots, N$, $\Lambda_{i,1}^{(1)} = \Lambda_{i,1}^{(0)} = 0$, $\Lambda_{i,2}^{(0)} \sim \mathcal{N}(0, 1)$ and $\Lambda_{i,2}^{(1)} = \Lambda_{i,2}^{(0)} + \mathcal{N}(0.2, 1)$. Let $N = 250$, $T = 250$ and $e_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. The observation pattern depends on an observed unit specific variable defined as $S_i = \mathbb{1}(\Lambda_{i,2}^{(0)} \neq 0)$ which only depends on the loadings of the second factor. Once a unit adopts treatment, it stays treated afterward. For the units with $S_i = 1$, 50% randomly selected units adopt the treatment from time $0.5 \cdot T$ and the remaining 50% units stay in the control group until the end. For the units with $S_i = 0$, 90% randomly selected units adopt the treatment from time $0.5 \cdot T$ and the remaining 10% units stay in the control group until the end. We report the relative MSE for common components for observed and unobserved treated and control common components. We also report the results conditional on the characteristic S_i and the relative MSE of $\beta_i^{(1)} - \beta_i^{(0)}$ capturing the average treatment effect over time for each unit and ATE which is the relative MSE of the overall average treatment effect $\sum_{(i,t):W_{it}=0} (\hat{C}_{it}^{(1)} - \hat{C}_{it}^{(0)})$. The results are generated from 1000 Monte Carlo simulations. The results show that XP$_{\text{PROP}}$ can be a more robust estimator for missing observations under misspecification (omitted or weak factors).

setup differs from classical causal inference as we estimate the covariates as latent factors from the data. However, we can have a situation similar to omitted variables if we estimate too few latent factors, the factors are weak, or the population model is nonlinear.

We compare our benchmark estimator (XP) and propensity-weighted estimator (XP$_{\text{PROP}}$) under two types of model misspecification. In Table 6, we consider the case of omitted factors. The population model is generated by a two-factor model, but we only estimate one latent factor. In this case, the propensity-weighted estimator can perform better than the benchmark estimator. However, when the model is correctly specified, and we estimate two factors, the benchmark estimator dominates. When the second factor is weak in the sense that its variance and corresponding eigenvalue are very small, the situation is similar to an omitted factor. In this case, it is possible that the propensity-weighted estimator performs better even if we estimate the correct number of latent factors. Note that weak factors are also a form of misspecification, as discussed in Onatski (2012). In this simulation, observations are more likely to miss if they are exposed to the omitted or weak second factor. Hence, the robustness gains of the propensity-weighted estimator arise for the missing data and the treatment effects.

The case of omitted latent factors shares similarity with the case of a misspecified functional form. In Table A.XIV in the Online Appendix we generate the data from a non-linear one-factor model. Under certain assumptions it is possible to approximate a non-linear transformation as a linear function of appropriate basis functions. Such an approximation can be formulated as a linear latent multi-factor model, where the additional factors are non-linear transformations of the underlying one-factor model. Hence, some form of functional model misspecification can be corrected by using more latent factors. In our example, the non-linearity is very well approximated by three latent factors for the simple and propensity-weighted estimator. However, if we use only one or two latent factors, the propensity-weighted estimator is more robust to the misspecification. While we do not provide a formal non-parametric theory, our simulation suggests that a non-linear misspecification can share similar features with the case of omitted factors. Therefore, if a researcher

suspects some form of model misspecification, the propensity-weighted estimator can be a useful alternative and serves as a valid robustness check.[20]

11. Conclusion

This paper develops the inferential theory for latent factor models estimated from large dimensional panel data with missing observations. Our paper stands out by the generality of the missing patterns that we allow for. We propose two estimators for the latent factor model: a simple all-purpose estimator and an extension to a probability-weighted estimator. Our all-purpose estimator is easy to use while it performs well under a variety of missing patterns. The propensity weighted estimator is an alternative that is less efficient for correctly specified models but can be more robust to certain forms of misspecification. The key application of our asymptotic distribution theory is to test causal treatment effects. We provide a test for the point-wise treatment effect that can be heterogeneous and time-dependent under general adoption patterns where the units can be affected by unobserved factors.

Appendix A

Notation. Let $M < \infty$ denote a generic constant. Let $\|v\|$ denote the vector norm and $\|A\| = trace(A^\top A)^{1/2}$ the Frobenius norm of matrix A.

General assumptions

Assumption G2 (*Factor Model*).

1. Factors: $\forall t$, $\mathbb{E}[\|F_t\|^4] \leq \bar{F} < \infty$. There exists some positive definite $r \times r$ matrix Σ_F, such that $\frac{1}{T} \sum_{t=1}^T F_t F_t^\top \xrightarrow{P} \Sigma_F$ and $\mathbb{E} \left\| \sqrt{T} \left(\frac{1}{T} \sum_{t=1}^T F_t F_t^\top - \Sigma_F \right) \right\|^2 \leq M$. Furthermore, for any $\mathcal{Q}_{ij}$, $\frac{1}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t F_t^\top \xrightarrow{P} \Sigma_F$ and $\mathbb{E} \left\| \sqrt{|\mathcal{Q}_{ij}|} \left(\frac{1}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t F_t^\top - \Sigma_F \right) \right\|^2 \leq M$.

2. Factor loadings: loadings are random, independent of factors and errors. $\forall t$, $\mathbb{E}[\|\Lambda_i\|^4] \leq \bar{\Lambda} < \infty$. There exists some positive definite $r \times r$ matrix Σ_Λ such that $\frac{1}{N} \sum_{i=1}^N \Lambda_i \Lambda_i^\top \xrightarrow{P} \Sigma_\Lambda$ and $\mathbb{E} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^N \Lambda_i \Lambda_i^\top - \Sigma_\Lambda \right) \right\| \leq M$.

3. Time and cross-section dependence and heteroskedasticity of errors: There exists a positive constant $M < \infty$, such that for all N and T:

    (a) $\mathbb{E}[e_{it}] = 0$, $\mathbb{E}|e_{it}|^8 \leq M$.
    (b) $\mathbb{E}[e_{is}e_{it}] = \gamma_{st,i}$ with $|\gamma_{st,i}| \leq \gamma_{st}$ for some γ_{st} and all i. For all t, $\sum_{s=1}^T \gamma_{st} \leq M$.
    (c) $\mathbb{E}[e_{it}e_{jt}] = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq \tau_{ij}$ for some τ_{ij} and all t. For all i, $\sum_{j=1}^N \tau_{ij} \leq M$.
    (d) $\mathbb{E}[e_{it}e_{js}] = \tau_{ij,ts}$ and $\sum_{j=1}^N \sum_{s=1}^T |\tau_{ij,ts}| \leq M$ for all i and t.
    (e) For all i and j, $\mathbb{E} \left| \frac{1}{|\mathcal{Q}_{ij}|^{1/2}} \sum_{t \in \mathcal{Q}_{ij}} \left(e_{it}e_{jt} - \mathbb{E}[e_{it}e_{jt}] \right) \right|^4 \leq M$.

4. Weak dependence between factor and idiosyncratic errors: for every (i, j),

$$\mathbb{E} \left\| \frac{1}{\sqrt{|\mathcal{Q}_{ij}|}} \sum_{i \in \mathcal{Q}_{ij}} F_t e_{it} \right\|^4 \leq M.$$

5. Eigenvalues: The eigenvalues of $\Sigma_\Lambda \Sigma_F$ are distinct.

Assumption G3 (*Moments and Central Limit Theorems*). For all N and T,

1. $\mathbb{E} \left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \frac{1}{|\mathcal{Q}_{ij}|} \sum_{s \in \mathcal{Q}_{ij}} \phi_{i,st} \left(e_{is}e_{js} - \mathbb{E}[e_{is}e_{js}] \right) \right\|^2 \right] \leq M$, where $\phi_{i,st} = W_{it}F_s$, Λ_i, $W_{it}\Lambda_i$, for every j and t.

2. $\mathbb{E} \left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \frac{\phi_{it}}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t^\top e_{it} \right\|^2 \right] \leq M$ for every t and for $\phi_{it} = \Lambda_i$, W_{it}, $W_{it}\Lambda_i$.

3. $\frac{\sqrt{T}}{N} \sum_{i=1}^N \Lambda_i \Lambda_i^\top \frac{1}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t e_{it} \xrightarrow{d} \mathcal{N}(0, \Gamma_{\Lambda,j}^{\text{obs}})$ for every j.

4. $\frac{1}{\sqrt{N}} \sum_{i=1}^N W_{it} \Lambda_i e_{it} \xrightarrow{d} \mathcal{N}(0, \Gamma_{F,t}^{\text{obs}})$ for every t.

---

[20] For a non-linear factor model, Feng (2020) proposes a local PCA method that uses a linear model approximation in a local neighborhood. Our argument is based on a global approximation of the non-linear functional relationship, where the additional latent factors serve as additional basis functions.

5. We define the filtration $\mathcal{G}^t = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts}^t)$ with $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda, v_t)$ generated by $\{W_{ij}, j \leq s, \text{all } i\}$, Λ and v_t, which is given by $v_t = \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t$. For every i and t, and $u_i = \Lambda_i$, it holds

$$\sqrt{T} \begin{bmatrix} X_i u_i \\ \mathbf{X}_t v_t \end{bmatrix} \to \mathcal{N}\left(0, \begin{bmatrix} h_i(u_i) & g_{i,t}^{\text{cov}}(u_i, v_t)^\top \\ g_{i,t}^{\text{cov}}(u_i, v_t) & g_t(v_t) \end{bmatrix}\right) \quad \mathcal{G}^t - \text{stably},$$

where $X_i = \frac{1}{N} \sum_{l=1}^N \Lambda_l \Lambda_l^\top \left(\frac{1}{|\mathcal{Q}_{li}|} \sum_{s \in \mathcal{Q}_{li}} F_s F_s^\top - \frac{1}{T} \sum_{s=1}^T F_s F_s^\top \right)$ and $\mathbf{X}_t = \frac{1}{N} \sum_{i=1}^N W_{it} X_i \Lambda_i \Lambda_i^\top$.

6. $\mathbb{E}\left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \left(\frac{1}{|\mathcal{Q}_{li}|} \sum_{s \in \mathcal{Q}_{li}} F_s F_s^\top - \frac{1}{T} \sum_{s=1}^T F_s F_s^\top \right) \Lambda_i W_{it} e_{it} \right\|^2 \right] \leq M$ for every l.

Assumption G4 (*Additional Assumptions on Factor Model*). As $T_{1,i} \to \infty$, it holds

1. $\frac{1}{\sqrt{T_{1,i}}} \sum_{T_{0,i}+1}^T F_t e_{it} \xrightarrow{d} \mathcal{N}(0, \Sigma_{F,e_i})$.

2. $\mathbb{E}\left[\left\| \frac{1}{\sqrt{NT_{1,i}}} \sum_{t=T_{0,i}+1}^T \sum_{j=1}^N W_{jt} \Lambda_j e_{jt} \right\|^2 \right] \leq M$ and $\mathbb{E}\left[\left\| \frac{1}{\sqrt{NT_{1,i}}} \sum_{t=T_{0,i}+1}^T \sum_{j=1}^N Z_t F_t^\top W_{jt} \Lambda_j e_{jt} \right\|^2 \right] \leq M$ for every i, $Z \in \mathbb{R}^{T_{1,i} \times L}$ and $\|Z_t\| \leq M$.

3. [Assumption G3.5](#) holds for v_t equal to $\Sigma_{\Lambda,t}^{-1} \Lambda_i^{(1)}$ and $\Sigma_{\Lambda,t}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})$ under the filtration $\mathcal{G} = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts})$ with $\mathcal{G}_{Ts} = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda)$ generated by $\{W_{ij}, j \leq s, \text{all } i\}$ and Λ.

4. For v_t equal to $\Sigma_{\Lambda,t}^{-1} \Lambda_i^{(1)}$, $\Sigma_{\Lambda,t}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})$ or $\Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t$, the elements of the random vector $\mathbf{X}_{T_{0,i}+1} v_{T_{0,i}+1}, \ldots, \mathbf{X}_T v_T$ are jointly $\mathcal{G}^t$-stably normal with $\text{ACov}(\mathbf{X}_t v_t, \mathbf{X}_s v_s) = g_{t,s}^{\text{cov}}(v_t, v_s)$ for all $T_{0,i} \leq t, s \leq T$, where $\mathcal{G}^t$ is defined as $\mathcal{G}^t = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts}^t)$, $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda)$ if $v_t \neq \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t$, and $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda, v_t)$ otherwise.

Assumption GC2 (*Conditional Factor Model*).

1. Factor loadings: $\mathbb{E}[\|\Lambda_i\|^4 | S] \leq \bar{\Lambda} < \infty$. There exists some positive definite $r \times r$ matrix Σ_Λ such that $\frac{1}{N} \sum_{i=1}^N \frac{W_{it}}{P(W_{it}=1|S_i)} \Lambda_i \Lambda_i^\top \xrightarrow{P} \Sigma_\Lambda$ and $\mathbb{E}\left\| \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{P(W_{it}=1|S_i)} W_{it} \Lambda_i \Lambda_i^\top - \Sigma_\Lambda \right) \right\| \leq M$.

Assumption GC3 (*Conditional Moments and Central Limit Theorems*). S is independent of F and e and $\mathbb{E}[\|\Lambda_i\|^6 | S] \leq \bar{\Lambda}$. For all N and T,

1. $\mathbb{E}\left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \frac{1}{|\mathcal{Q}_{ij}|} \sum_{s \in \mathcal{Q}_{ij}} \phi_{i,st} \left(e_{is} e_{js} - \mathbb{E}[e_{is} e_{js}] \right) \right\|^2 \right] \leq M$ for every j and t, where $\phi_{i,st} = \frac{W_{it} F_s}{P(W_{it}=1|S_i)}$, Λ_i, $\frac{W_{it}}{P(W_{it}=1|S_i)} \Lambda_i$.

2. $\mathbb{E}\left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \frac{\phi_{it}}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t^\top e_{it} \right\|^2 \right] \leq M$ for every t and for $\phi_{it} = \Lambda_i$, $\frac{W_{it}}{P(W_{it}=1|S_i)} \Lambda_i$.

3. $\frac{\sqrt{T}}{N} \sum_{i=1}^N \Lambda_i \Lambda_i^\top \frac{1}{|\mathcal{Q}_{ij}|} \sum_{t \in \mathcal{Q}_{ij}} F_t e_{it} \xrightarrow{d} \mathcal{N}(0, \Gamma_{\Lambda,j}^{\text{obs}})$ for every j.

4. $\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{W_{it}}{P(W_{it}=1|S_i)} \Lambda_i e_{it} \xrightarrow{d} \mathcal{N}(0, \Gamma_{F,t}^{\text{obs},S})$ for every t.

5. We define the filtration $\mathcal{G}^t = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts}^t)$ with $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda, v_t)$ generated by $\{W_{ij}, j \leq s, \text{all } i\}$, Λ and v_t, which is given by $v_t = \Sigma_\Lambda^{-1} \Sigma_F^{-1} F_t$. For every i and t, and $u_i = \Lambda_i$, it holds

$$\sqrt{T} \begin{bmatrix} X_i u_i \\ \mathbf{X}_t^S v_t \end{bmatrix} \to \mathcal{N}\left(0, \begin{bmatrix} h_i(u_i) & g_{i,t}^{\text{cov},S}(u_i, v_t)^\top \\ g_{i,t}^{\text{cov},S}(u_i, v_t) & g_t^S(v_t) \end{bmatrix}\right) \quad \mathcal{G}^t - \text{stably},$$

where $X_i = \frac{1}{N} \sum_{l=1}^N \Lambda_l \Lambda_l^\top \left(\frac{1}{|\mathcal{Q}_{li}|} \sum_{s \in \mathcal{Q}_{li}} F_s F_s^\top - \frac{1}{T} \sum_{s=1}^T F_s F_s^\top \right)$ and $\mathbf{X}_t = \frac{1}{N} \sum_{i=1}^N \frac{W_{it}}{P(W_{it}=1|S_i)} X_i \Lambda_i \Lambda_i^\top$.

6. $\mathbb{E}\left[\left\| \sqrt{\frac{T}{N}} \sum_{i=1}^N \left(\frac{1}{|\mathcal{Q}_{li}|} \sum_{s \in \mathcal{Q}_{li}} F_s F_s^\top - \frac{1}{T} \sum_{s=1}^T F_s F_s^\top \right) \frac{W_{it}}{P(W_{it}=1|S_i)} \Lambda_i e_{it} \right\|^2 \right] \leq M$ for every l.

Assumption GC4 (*Additional Assumptions on Factor Model*). As $T_{1,i} \to \infty$, it holds

1. $\frac{1}{\sqrt{T_{1,i}}} \sum_{T_{0,i}+1}^T F_t e_{it} \xrightarrow{d} \mathcal{N}(0, \Sigma_{F,e_i})$.

2. $\mathbb{E}\left[\left\| \frac{1}{\sqrt{NT_{1,i}}} \sum_{t=T_{0,i}+1}^T \sum_{j=1}^N \frac{W_{jt}}{P(W_{it}=1|S_i)} \Lambda_j e_{jt} \right\|^2 \right] \leq M$ and $\mathbb{E}\left[\left\| \frac{1}{\sqrt{NT_{1,i}}} \sum_{t=T_{0,i}+1}^T \sum_{j=1}^N Z_t F_t^\top \frac{W_{jt}}{P(W_{it}=1|S_i)} \Lambda_j e_{jt} \right\|^2 \right] \leq M$ for every i, $Z \in \mathbb{R}^{T_{1,i} \times L}$ and $\|Z_t\| \leq M$.

3. [Assumption GC3.5](#) holds for v_t equal to $\Sigma_{\Lambda,t}^{-1} \Lambda_i^{(1)}$ and $\Sigma_{\Lambda,t}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})$ under the filtration $\mathcal{G} = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts})$ with $\mathcal{G}_{Ts} = \sigma(\{W_{ij}, j \leq s, \text{all } i\}, \Lambda)$ generated by $\{W_{ij}, j \leq s, \text{all } i\}$ and Λ.

4. For v_t equal to $\Sigma_{\Lambda,t}^{-1}\Lambda_i^{(1)}$, $\Sigma_{\Lambda,t}^{-1}(\Lambda_i^{(1)} - \Lambda_i^{(0)})$ or $\Sigma_\Lambda^{-1}\Sigma_F^{-1}F_t$, the elements of the random vector $\mathbf{X}_{T_{0,i}+1}^S v_{T_{0,i}+1}, \ldots, \mathbf{X}_T^S v_T$ are jointly $\mathcal{G}^t$-stably normal with $\mathrm{ACov}(\mathbf{X}_t^S v_t, \mathbf{X}_s^S v_s) = g_{t,s}^S(v_t, v_s)$ for all $T_{0,i} \le t, s \le T$, where $\mathcal{G}^t$ is defined as $\mathcal{G}^t = \sigma(\cup_{s=1}^T \mathcal{G}_{Ts}^t)$, $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \le s, \text{all } i\}, \Lambda)$ if $v_t \ne \Sigma_\Lambda^{-1}\Sigma_F^{-1}F_t$, and $\mathcal{G}_{Ts}^t = \sigma(\{W_{ij}, j \le s, \text{all } i\}, \Lambda, v_t)$ otherwise.

Assumption G2 describes an approximate factor structure and is at a similar level of generality as (Bai, 2003): (1) Assumption G2.1 ensures that each factor has a nontrivial contribution to the variation in X. (2) We assume loadings are random but independent of factors and errors in Assumption G2.2. We could study a factor model conditioned on some particular realization of the loadings, and the analysis would essentially be equivalent to that under the assumption that loadings are nonrandom. (3) Assumption G2.3 allows errors to be time-series and cross-sectionally weakly correlated. (4) Assumption G2.4 allows factors and idiosyncratic errors to be weakly correlated. (5) Assumption G2.5 guarantees that each loading and factor can be uniquely identified up to some rotation matrix. Additionally, we assume that these aspects also hold if we look at a subset of all time periods (the subset is denoted as $\mathcal{Q}_{ij}$ in Assumption G2). Together with Assumption C1.2, our covariance matrix estimator (1) using incomplete observations has similar properties as the conventional covariance matrix estimator $\frac{1}{T}XX^\top$ using full observations. For example, both $\frac{1}{|\mathcal{Q}_{ij}|}\sum_{t \in \mathcal{Q}_{ij}} X_{it}X_{jt}$ and $\frac{1}{T}\sum_{t=1}^T X_{it}X_{jt}$ are consistent estimators for Σ_{ij}. Moreover, the top r eigenvalues estimated from both matrices are consistent as shown in Lemma 4 in the Internet Appendix, which is the foundation for developing the inferential theory of the factor model estimated from Eq. (1).

Assumption G3 is not required to show the consistency of loadings and factors but is only used to show the asymptotic normality of the estimators. Assumption G3.1–4 are closely related to the moment and CLT assumptions in Bai (2003). The first two parts in Assumption G3 restrict the second moments of certain averages. The 3rd and 4th points state the necessary central limit theorems. $\frac{\sqrt{T}}{N}\sum_{i=1}^N \Lambda_i\Lambda_i^\top \frac{1}{|\mathcal{Q}_{ij}|}\sum_{t \in \mathcal{Q}_{ij}} F_t e_{it} \xrightarrow{d} \mathcal{N}(0, \Phi_j)$ is one of the leading terms in the asymptotic distribution of the estimated loadings $\tilde{\Lambda}_j$. However, $\frac{1}{|\mathcal{Q}_{ij}|}\sum_{t \in \mathcal{Q}_{ij}} F_t e_{it}$ varies with j so we cannot separately average over the cross-sectional and time dimension as in the conventional framework. Point 5 is specific to the missing value problem and introduces the correction terms that appear in the asymptotic distribution. They are due to the fact that our estimator averages over different number of observations for different entries in the covariance matrix. Assumption G3.5 assumes a central limit theorem for X_i and $\mathbf{X}_t$. The usual CLT of the form

$$\sqrt{T}\begin{bmatrix} \mathrm{vec}(X_i) \\ \mathrm{vec}(\mathbf{X}_t) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \Phi_i & (\mathbf{\Phi}_{i,t}^{\mathrm{cov}})^\top \\ \mathbf{\Phi}_{i,t}^{\mathrm{cov}} & \Phi_t \end{bmatrix}\right),$$

is not sufficient as X_i and $\mathbf{X}_t$ are multiplied with the random variables u_i and v_t in $\sqrt{T}\begin{bmatrix} (X_i u_i)^\top & (\mathbf{X}_t^S v_t)^\top \end{bmatrix}$. The asymptotic variances of these products are quadratic functions in the elements of those random variables given by $h_i(u_i)$ and $g_t(v_t)$ and take the form of $h_i(u_i) = (u_i^\top \otimes I_r)\Phi_i(u_i \otimes I_r)$ and $g_t(v_t) = (v_t^\top \otimes I_r)\Phi_t(v_t \otimes I_r)$ respectively.

Assumption G3.5 requires a central limit theorem for stable convergence in law which is stronger than the conventional central limit theorem for convergence in distribution. The reason is that the asymptotic variance in Assumption G3.5 depends on both Λ_i and F_t, which are random variables. Hence, we deal with a mixed normal limit and stable convergence in law ensures that the normal distribution of the central limit theorem will be independent of Λ_i and F_t. Because of the stable convergence in law result, the estimated factors and common components normalized by their random standard deviation will converge to a standard normal distribution. In more detail, Assumption G3.5 implies that X_i and $\mathbf{X}_t$ jointly converge $\mathcal{G}^t$-stably for $(N, T) \to \infty$ to a mixed normal distribution, whose asymptotic variance is random but measurable with respect to the sigma-field $\mathcal{G}^t$. Assumption G3.5 is used in Theorem 2 to show the asymptotic distribution of the variance correction term whose asymptotic variance is random. Our simplified factor model specified by Assumption S2 is sufficient to guarantee a central limit theorem for stable convergence in law. Proposition 3 shows that the simplified model implies Assumption G3.5.

Assumptions GC2 and GC3 are the corresponding assumptions for the propensity-weighted estimator with a similar level of generality. The additional Assumptions G4 and GC4 are only needed for the treatment effect tests. The simplified assumptions imply the general assumptions as stated in Proposition 3.

Proposition 3. *The simplified model is a special case of the general model:*

1. *Assumptions G2 and G3 are satisfied in the simplified model:*

   (a) *Assumptions S1 and S2 imply Assumption G2.*
   (b) *Assumptions S1, S2 and S3 imply Assumption G3.*

2. *Assumptions GC2 and GC3 are satisfied in the simplified conditional model:*

   (a) *Assumptions S1, C1, S2 and C2 imply GC2.*
   (b) *Assumptions S1, C1, S2, S3.2, C2 and C3 imply Assumption GC3.*

3. *Assumptions G4 and GC4 are satisfied in the simplified model. Specifically,*

   (a) *Assumptions S1, S2 and S3 imply Assumption G4.*
   (b) *Assumptions S1, C1, S2, S3.2, C2 and C3 imply Assumption GC4.*

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.04.005.

References

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J. Amer. Statist. Assoc. 105 (490), 493–505.

Abadie, A., Diamond, A., Hainmueller, J., 2015. Comparative politics and the synthetic control method. Am. J. Political Sci. 59 (2), 495–510.

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

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., Khosravi, K., 2021. Matrix completion methods for causal panel data models. J. Amer. Statist. Assoc. 1–15.

Athey, S., Imbens, G.W., 2021. Design-based analysis in difference-in-differences settings with staggered adoption. J. Econometrics.

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

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

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

Bai, J., Ng, S., 2021. Matrix completion, counterfactuals, and factor analysis of missing data. J. Amer. Statist. Assoc. 1–18.

Bańbura, M., Modugno, M., 2014. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. J. Appl. Econometrics 29 (1), 133–160.

Cahan, E., Bai, J., Ng, S., 2021. Factor-based imputation of missing values and covariances in panel data of large dimensions. arXiv preprint arXiv:2103.03045.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9 (6), 717.

Chen, Y., Fan, J., Ma, C., Yan, Y., 2019. Inference and uncertainty quantification for noisy matrix completion. Proc. Natl. Acad. Sci. 116 (46), 22931–22937.

David, M., Little, R.J., Samuhel, M., Triest, R., 1983. Nonrandom nonresponse models based on the propensity to respond. In: Proceedings of the Business and Economic Statistics Section. American Statistical Association, pp. 168–173.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–22.

Doudchenko, N., Imbens, G.W., 2016. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: a Synthesis. Discussion paper, National Bureau of Economic Research.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econometrics 164 (1), 188–205.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 603–680.

Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. Ann. Statist. 44 (1), 219.

Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 4, 1497–1541.

Feng, Y., 2020. Causal inference in possibly nonlinear factor models. arXiv preprint arXiv:2008.13651.

Freedman, D.A., Berk, R.A., 2008. Weighting regressions by propensity scores. Eval. Rev. 32 (4), 392–409.

Gagliardini, P., Ossola, E., Scaillet, O., 2019. A diagnostic criterion for approximate factor structure. J. Econometrics.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. J. Monetary Econ. 55 (4), 665–676.

Gobillon, L., Magnac, T., 2016. Regional policy evaluation: Interactive fixed effects and synthetic controls. Rev. Econ. Stat. 98 (3), 535–551.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47 (260), 663–685.

Jin, S., Miao, K., Su, L., 2021. On factor models with random missing: EM estimation, inference, and cross validation. J. Econometrics 222 (1), 745–777.

Jungbacker, B., Koopman, S.J., Van der Wel, M., 2011. Maximum likelihood estimation for dynamic factor models with missing data. J. Econom. Dynam. Control 35 (8), 1358–1368.

Kang, J.D.Y., Schafer, J.L., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 523–539.

Lettau, M., Pelger, M., 2020a. Estimating latent asset pricing factors. J. Econom. 218 (1), 1–31.

Lettau, M., Pelger, M., 2020b. Factors that fit the time series and cross-section of stock returns. Rev. Financial Stud. 33 (5), 2274–2325.

Li, K.T., 2019. Statistical inference for average treatment effects estimated by synthetic control methods. J. Amer. Statist. Assoc. 1–16.

Li, K.T., Bell, D.R., 2017. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. J. Econometrics 197 (1), 65–75.

Little, R.J., 1986. Survey nonresponse adjustments for estimates of means. Int. Stat. Rev./Revue Int. Stat. 139–157.

Little, R.J., 1988. Missing-data adjustments in large surveys. J. Bus. Econom. Statist. 6 (3), 287–296.

Masini, R., Medeiros, M.C., 2018. Counterfactual analysis with artificial controls: Inference, high dimensions and nonstationarity. High Dimensions and Nonstationarity (September 1, 2018).

Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11 (Aug), 2287–2322.

Negahban, S., Wainwright, M.J., 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 1069–1097.

Negahban, S., Wainwright, M.J., 2012. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Mach. Learn. Res. 13 (May), 1665–1697.

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

Pelger, M., 2019. Large-dimensional factor modeling based on high-frequency observations. J. Econometrics 208 (1), 23–42.

Pelger, M., Xiong, R., 2020. The causal effect of publication on the cross-section of stock returns. Working Paper.

Pelger, M., Xiong, R., 2021a. Interpretable sparse proximate factors for large dimensions. J. Bus. Econom. Statist. 1–23.

Pelger, M., Xiong, R., 2021b. State-varying factor models of large dimensions. J. Bus. Econom. Statist. 1–50.

Robins, J.M., Rotnitzky, A., 1995. Semiparametric efficiency in multivariate regression models with missing data. J. Amer. Statist. Assoc. 90 (429), 122–129.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89 (427), 846–866.

Robins, J., Sued, M., Lei-Gomez, Q., Rotnitzky, A., 2007. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. Statist. Sci. 22 (4), 544–559.

Robins, J.M., Wang, N., 2000. Inference for imputation estimators. Biometrika 87 (1), 113–124.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

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

Simon, H.A., 1954. Spurious correlation: A causal interpretation. J. Amer. Statist. Assoc. 49 (267), 467–479.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20 (2), 147–162.

Stock, J.H., Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In: Handbook of Macroeconomics, Vol. 2. Elsevier, pp. 415–525.

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

Zhou, Y., Wilkinson, D., Schreiber, R., Pan, R., 2008. Large-scale parallel collaborative filtering for the netflix prize. In: International Conference on Algorithmic Applications in Management. Springer, pp. 337–348.