Target PCA: Transfer Learning Large Dimensional Panel Data*

Junting Duan[†] Markus Pelger[‡] Ruoxuan Xiong[§]

This draft: August 25, 2023
First draft: December 14, 2022

Abstract

This paper develops a novel method to estimate a latent factor model for a large target panel with missing observations by optimally using the information from auxiliary panel data sets. We refer to our estimator as target-PCA. Transfer learning from auxiliary panel data allows us to deal with a large fraction of missing observations and weak signals in the target panel. We show that our estimator is more efficient and can consistently estimate weak factors, which are not identifiable with conventional methods. We provide the asymptotic inferential theory for target-PCA under very general assumptions on the approximate factor model and missing patterns. In an empirical study of imputing data in a mixed-frequency macroeconomic panel, we demonstrate that target-PCA significantly outperforms all benchmark methods.

Keywords: Factor Analysis, Principal Components, Transfer Learning, Multiple Data Sets, Large-Dimensional Panel Data, Large N and T, Missing Data, Weak Factors, Causal Inference
JEL classification: C14, C38, C55, G12

*We thank Jose Blanchet, Kay Giesecke, Serena Ng, Neil Shephard and seminar and conference participants at Stanford, University of Rochester, Oxford, the NBER-NSF Time-Series Conference, California Econometrics Conference, SoFiE Conference, and NASMES for helpful comments.

[†]Stanford University, Department of Management Science & Engineering, Email: duanjt@stanford.edu.
[‡]Stanford University, Department of Management Science & Engineering, Email: mpelger@stanford.edu.
[§]Emory University, Department of Quantitative Theory and Methods, Email: ruoxuan.xiong@emory.edu.

1 Introduction

Panel data with a large number of units and time periods are widely available in macroeconomics, finance, and many other areas of social sciences. In many cases, these panels can be well described by an approximate factor structure, that is, a small number of common factors explain a large portion of the co-movements. A common approach is to estimate latent factors with statistical methods from the panel of interest, which we refer to as target data. In the era of big data, there often exist auxiliary panels, that contain relevant information and share some common factors with the target panel. Combining the information in multiple panels can increase the efficiency of the estimated factor model for the target. Even more importantly, using auxiliary panels can identify the factors that only affect a small subset of units or are not detectable due to the missing observations in the target. The idea of using auxiliary panel data to estimate a model that is applied to a target data set is conceptually similar to transfer learning, which has been successfully used for machine learning tasks.

There is a broad class of practical problems which our setup is relevant. A particularly important case is mixed-frequency data, which is empirically studied in this paper. For example, some macroeconomic time series are only available at a quarterly or lower frequency, for example, GDP, while other time series are available at a higher frequency, for example, stock returns. As long as some factors in stock returns are also correlated with the macroeconomic movements, they can be used to obtain higher frequency factors and imputed values for the lower frequency target panel. Another important case is casual inference on panel data with non-random treatment, where missing data in the target panel correspond to the unobserved counterfactual outcomes. The auxiliary data could be panels for another set of units or the same set of units but with different outcome variables, which are correlated with the target time series. By leveraging auxiliary data, the latent factor model can be more precisely estimated, thus improving the precision of missing data imputation in the target panel.

These examples illustrate the benefits, but also fundamental challenges, of using auxiliary data. First, the cross-sectional dimensionality and the signal-to-noise ratio of the panels can be very different. Second, the target panel might include information not contained in the auxiliary data. Therefore, naive methods, such as simply concatenating target and auxiliary data to one large panel or using them separately, can be sub-optimal or infeasible.

We propose to estimate a latent factor model for the target data by optimally combining information from auxiliary panels with the target panel. We refer to this method as target-PCA (T-PCA). This method is broadly applicable but easy to implement: It applies principal component analysis (PCA) to a weighted average of the second-moment matrices of the target and auxiliary panels. Our method can be interpreted as applying PCA to an auxiliary panel with a reward for factors that are useful for the target panel. We show the consistency and provide the inferential theory for target-PCA under general assumptions on the latent factor model and missing observations. The asymptotic distribution is essential for two reasons: First, it provides guidance on selecting efficient weights in target-PCA; second, it provides confidence intervals for missing data imputation from the estimated latent factor model.

We show two important effects of the relative weights between target and auxiliary data in target-PCA. The first effect is the consistency effect for factor identification. This matters when neither the target nor the auxiliary data alone are sufficient to estimate all the factors that we care about. Factors cannot be consistently estimated from the target panel when their signal is too weak or when the partially observed data is insufficient, for example, when certain times are unobserved for the full panel. Suppose that these weak factors are strong in the auxiliary data, but the auxiliary data might not contain the other factors for the target. In this case, target-PCA can consistently estimate all factors, if we select the target weight for the auxiliary data at the right rate to account for the different dimensions of the panels.

The second effect is the efficiency effect in the estimation of latent factors and loadings. This effect arises when target and auxiliary data are observed with different noise levels. If the weighting in target-PCA properly accounts for the noise ratio between target and auxiliary data, then we can improve the efficiency of the estimated latent factor model. Hence, after selecting the target weight in the right order to ensure consistency, we can improve the efficiency by selecting the optimal scale of the weights.

These two important effects show that the optimal selection of weights in target-PCA is a challenging problem that can depend on the relative factor strength, observation pattern, noise level, and dimensionality between target and auxiliary data. To address this problem, we develop the inferential theory for the estimated factors, loadings, and imputed values for a general weighting scheme in target-PCA. The inferential theory is then used as guidance for selecting the weights in target-PCA. The naive cases of concatenating the target and auxiliary data, or using only one of the panels, are special cases of our general method. We show that these special cases are generally less efficient and even lack identification in the worst case.

Our work contributes to four distinct areas. First, we contribute to the literature on large dimensional factor models by proposing a new setup where the latent factors can be jointly estimated from multiple panel data. Second, we provide a new solution to the problem of weak factors, which cannot be consistently estimated with conventional PCA estimators and require additional signals. We show how to leverage the information in supplementary panels to overcome the weak signal problem. Third, our paper complements the recent work on imputing missing data in large panels. We show that leveraging auxiliary data allows us to impute missing observations with higher precision and makes it possible to impute values that could otherwise not be imputed using only the target panel. Last but not least, we contribute to causal inference and can estimate heterogeneous and time-dependent treatment effects for general interventions.

Our asymptotic results are developed under the framework of an approximate latent factor structure for both target and auxiliary data, both with large cross-section and time-series dimensions. Our new setup of using multiple panels generalizes the existing factor modeling literature, which only uses one panel so far. 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, Liao, and Mincheva (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 among others adding observable factors in Bai (2009), sparse and interpretable latent factors in Pelger and Xiong (2021a), conditional loadings in Fan, Liao, and Wang (2016), Pelger and Xiong (2021b), Chen, Roussanov, and Wang (2021) and Chen (2022), time-varying, locally estimated loadings in Su and Wang (2017) and high-frequency estimation in Pelger (2019). The idea of increasing the efficiency by weighting panels differently is related to the GLS type weighting for PCA estimators suggested among others in Breitung and Tenhofen (2011) and Choi (2012). Boivin and Ng (2006) study empirically the benefits for PCA estimators when down-weighting or dropping uninformative information.

Our work complements the recent work on estimating the latent factor model from one large panel with missing observations and developing entry-wise inferential theory. Our results extend the framework of Xiong and Pelger (2023) to the new setup of multiple panels. Other closely related work includes the recent papers by Jin, Miao, and Su (2021), Bai and Ng (2021), and Cahan, Bai, and Ng (2023). These 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. Bai and Ng (2021), and Cahan, Bai, and Ng (2023) leverage a block structure in missing data. Jin, Miao, and Su (2021) focuses on the case of missing-at-random and analyzes the EM estimator considered in Stock and Watson (1998). We show that by leveraging auxiliary data, we not only increase the precision for imputing missing values, but can also accommodate more general missing patterns. The problem of missing data imputation has been actively studied in the matrix completion literature since Candès and Recht (2009). The data imputation algorithms are largely based on rank-regularized methods with worst-case guarantees. Until recently, the entry-wise inferential theory is provided by the celebrated work of Chen, Fan, Ma, and Yan (2019) under the assumption of i.i.d. sampling. In contrast, we build on the large dimensional factor modeling literature, allowing us to make progress on the inferential theory under general observation patterns.

Our proposal of using auxiliary data brings the idea of transfer learning to the estimation of latent factors. The concept of transfer learning is to apply a model estimated on auxiliary data to target data. Transfer learning has been successfully used for machine learning tasks as surveyed by Pan and Yang (2009). Related to our work, Huang, Jiang, Li, Tong, and Zhou (2022) propose to forecast one target time series by scaling each of the auxiliary predictors by its predictive power on the target. In contrast, our target-PCA allows for a large cross-sectional dimension of the target with missing observations.

Our work provides a complementary and novel solution for weak factor estimation. Onatski (2012) has shown that conventional PCA estimators cannot consistently estimate factors that are too weak. Onatski (2010), Ahn and Horenstein (2013), Bai and Ng (2023), Onatski (2022) and others have studied the properties of weak factors and test the number of factors, with a focus on one panel. Lettau and Pelger (2020) have demonstrated that including additional information from other moments of the data can overcome the problem for certain types of weak factors. Similarly, our novel use of additional information from the auxiliary data can upweight the weak signals, allowing for the identification and efficient estimation of weak factors.

Our work is also complementary to the literature on mixed-frequency data imputation. A specific application of our general framework is to impute low-frequency observations in a target panel using auxiliary panels of higher frequency. One approach to dealing with data sampled at different frequencies has emerged in work by Ghysels, Santa-Clara, and Valkanov (2004), and Andreou, Ghysels, and Kourtellos (2010) using Mi(xed) Da(ta) S(ampling) (MIDAS) regressions, and its many extensions have been surveyed among others in Ghysels, Kvedaras, and Zemlys-Balevičius (2020). The MIDAS regression relates the low-frequency time series that we wish to predict to observables at high and low frequencies. Our framework shares some similarities as it relates latent factors of higher frequency to a large panel of lower-frequency observations. Another alternative is to use state space models to deal with mixed frequency data - usually estimated using the Kalman filter (see Bai, Ghysels, and Wright (2013) for a comparison with MIDAS). State space models are parameter-driven and impose different assumptions on the data generating process.[1] Another complementary idea is based on interpolation arguments as for example in Chow and Lin (1971). The recent work of Ng and Scanlan (2023) emphasizes the importance of residual correlation in the imputation of mixed frequency data. They propose a dynamic matrix completion approach by combining the state space setup and Chow and Lin (1971) type bridge regressions with latent factor models estimated from a large partially observed panel.

In simulation and empirical studies, we show the superior performance of our target-PCA method relative to benchmarks under a variety of settings. Our comprehensive simulations compare target-PCA to the natural benchmarks of applying PCA to separate panels or simple concatenated panels. Target-PCA substantially outperforms the alternatives in- and out-of-sample under different observation patterns. Our empirical analysis shows the good performance of target-PCA for imputing missing values in popular macroeconomic panels. We demonstrate the potential of target-PCA for nowcasting macroeconomic panels by imputing unbalanced low-frequency panels with higher-frequency auxiliary data.

The rest of the paper is organized as follows. Section 2 introduces the model and target-PCA. Section 3 illustrates the two important effects of weighting auxiliary data. Section 4 formalizes the assumptions on the observation pattern and approximate factor model. Section 5 provides the consistency and asymptotic results for our estimator, which can be used as guidance to select the weight in target-PCA. Section 6 discusses the extensions of our method, whose good performance compared to other methods is demonstrated by the extensive simulations in Section 7. Section 8 shows the practical relevance of target-PCA through empirical examples of imputing missing entries in macroeconomic data. Section 9 concludes the paper.

[1] Stock and Watson (2016) discusses the issues with state space estimation of factor models with missing data.

2 Model Setup

2.1 Model

We partially observe a target panel data set Y with T time periods and N_y cross-sectional units, where both T and N_y are large. $Y \in \mathbb{R}^{T\times N_y}$ has an approximate latent factor structure with k_y common factors,

$$Y_{ti} = (F_y)_t^\top (\Lambda_y^Y)_i + (e_y)_{ti}, \quad t = 1, \cdots, T, \; i = 1, \cdots, N_y.$$

Here, Y_{ti} denotes the data for the i-th cross section at time t, $(F_y)_t$ is a $k_y \times 1$ vector of latent factors, $(\Lambda_y^Y)_i$ is its corresponding latent factor loadings, $C_{ti} = (F_y)_t^\top (\Lambda_y^Y)_i$ is the common component of Y_{ti} and $(e_y)_{ti}$ is the idiosyncratic component of Y_{ti}. This factor model can also be written in a matrix notation as

$$\underbrace{Y}_{T\times N_y} = \underbrace{F_y}_{T\times k_y} \underbrace{\Lambda_y^{Y\top}}_{k_y\times N_y} + \underbrace{e_y}_{T\times N_y}.$$

Our goal is to estimate the latent factor structure in Y. We are particularly interested in the important case, where some entries in Y can be missing. In many practical applications, the panel Y might not be informative enough to estimate the latent factor model. First, the factors in Y can be weak, that is, affect only a small subset of cross-sectional units. In this case, it is possible that the factors cannot be separated from the noise, and conventional principal component analysis (PCA) fails to estimate them. Second, the observed data might be insufficient to estimate the full factor model, either because certain times are never observed in the panel, for example with low-frequency data, or because the missingness depends on the factor structure.

Our solution is to use additional information from auxiliary panel data X. Suppose we observe auxiliary data that contain relevant information and can be helpful for the estimation of latent factors in Y. For the exposition, we focus on the case where there is only one auxiliary panel data X and study how to optimally use the information in X to estimate the factors in Y. Our results can easily be extended to the case of multiple auxiliary panels, as discussed in Section 6.3.

Suppose X has an approximate latent factor structure with T time periods and N_x cross-section units, where N_x is large. Both panels X and Y are observed for the same time periods. X has k_x factors

$$\underbrace{X}_{T\times N_x} = \underbrace{F_x}_{T\times k_x} \underbrace{\Lambda_x^{X\top}}_{k_x\times N_x} + \underbrace{e_x}_{T\times N_x},$$

where $(F_x)_t$ is a $k_x \times 1$ vector of latent factors, $(\Lambda_x^X)_i$ is its corresponding latent factor loadings and $(e_x)_{ti}$ is the idiosyncratic error of X.

The auxiliary panel data X can be useful when it has some factors in common with Y. Without loss of generality, we can use the rotation of the factor models such that the concatenated factors in F_y and in F_x are orthogonal. We denote by F the union of the factors in F_y and F_x. The total number of non-redundant latent factors in F equals the rank of the concatenated panel matrix $[F_x, F_y]$. Given our normalization, $F^\top F/T$ is a diagonal matrix. It is notationally convenient to write both the target panel data Y and auxiliary panel X in terms of F, that is the union of all the non-redundant factors. Hence, the target Y follows

$$\underbrace{Y}_{T\times N_y} = \underbrace{F}_{T\times k}\underbrace{\Lambda_y^\top}_{k\times N_y} + \underbrace{e_y}_{T\times N_y},$$

where k is the total number of non-redundant latent factors in F, and $(\Lambda_y)_i$ is a $k \times 1$ vector of loadings on F. Note that we explicitly allow the loadings to be zero for some factors. More specifically, for factors that are unique to X and do not appear in F_y, the corresponding loadings in Λ_y are zero. Similarly, we express the auxiliary panel data X as

$$\underbrace{X}_{T\times N_x} = \underbrace{F}_{T\times k}\underbrace{\Lambda_x^\top}_{k\times N_x} + \underbrace{e_x}_{T\times N_x},$$

where $(\Lambda_x)_i$ is a vector of factor loadings for F in X.

In an approximate factor model, a large part of the variation is explained by the factors, while the noise is only weakly dependent. We allow for the empirically relevant case, where not all factors in Y are strong, but only affect a subset of the panel. Let $(\Lambda_y)_{ij}$ be unit i's loading for the j-th factor of F. We refer to the j-th factor as a strong factor in Y, if it affects a large number of units in Y. Formally, factor j is strong factor if it satisfies $N_y^{-1}\sum_{i=1}^{N_y}(\Lambda_y)_{ij}^2 > 0$ for $N_y \to \infty$. On the other hand, if $\sum_{i=1}^{N_y}(\Lambda_y)_{ij}^2$ grows at a rate smaller than N_y, then we refer to the j-th factor as a weak factor in Y. The weak factor only affects a small fraction of units, with the fraction converging to zero as N_y grows. Without loss of generality, we assume that all factors F_x are strong factors in X. As our objective is to estimate the factor model for Y, we are not interested in estimating the weak factors in X. However, it is straightforward to extend our results to the case of weak factors in X.

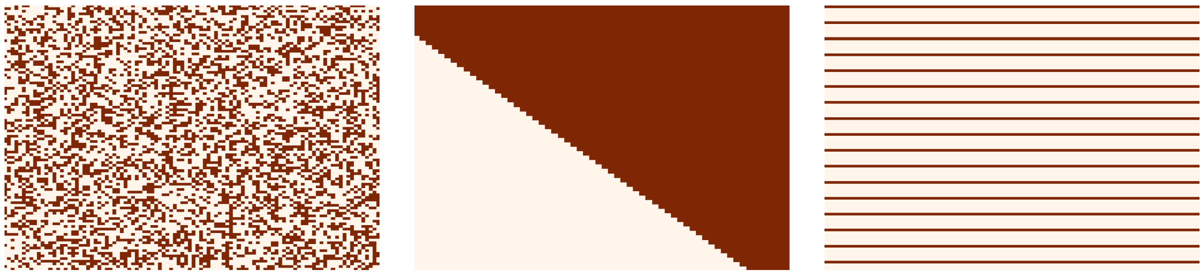
For the cross-sectional dimension of X and Y, we focus on the setting where $N_y/N_x \to c \in [0,\infty)$. This includes two cases: N_x and N_y are of the same order (i.e., $c > 0$) and N_x is much larger than N_y (i.e., $c = 0$). When $c = 0$, we consider the finite N_y case in Section 6.2. The setting of $N_y/N_x \to \infty$ is analogous and can be studied by similar arguments.

2.2 Observation Patterns

We allow the target data Y to have missing observations. Let $W^Y \in \{0,1\}^{T\times N_y}$ be the observation pattern of Y, where $W_{ti}^Y = 1$ if Y_{ti} is observed and zero otherwise. For simplicity, we assume the auxiliary data X is fully observed, but our methods and results can be easily generalized to the case where X is only partially observed as well.

We allow for very general observation patterns in Y. Whether an entry is observed or not can depend on whether other entries are observed, and on the factor model itself. The formal assumptions on the observation patterns are introduced in Section 4. To provide some intuition, Figure 1 shows three important examples of the observation patterns that we are interested in. In the first example, entries are missing completely at random, that is, whether an entry is observed

Figure 1: Examples of observation patterns



(a) Randomly missing

(b) Staggered treatment adoption

(c) Low-frequency observation

These figures show examples of observation patterns. The entries with dark color denote observed entries, while the entries with light color denote missing entries. Each row represents the observation pattern for a time period and each column represents the observation pattern for a unit.

or not does not depend on whether other entries are observed or not.

The second example is the observation pattern for control panels with staggered treatment adoption. Once a unit adopts the treatment, it stays treated afterwards, which can be modeled as missing values. This pattern is widely assumed in the literature on causal inference in panel data.

The third and attentive example shows the observation pattern of low-frequency time-series variables in Y, where variables in other data sets (including X) and applications are at a higher frequency. For example, assume that Y is a macroeconomic panel data set, where the time series are only available at a quarterly frequency, but a downstream application requires these time series as inputs at a monthly frequency. The monthly observations in between the quarters are modeled as missing observations. Importantly, in this example, there is no information in Y for the months where the quarterly variables are not reported, invalidating the existing latent factor estimation methods from partially observed Y only (Bai and Ng, 2021; Cahan, Bai, and Ng, 2023; Xiong and Pelger, 2023). In contrast, our proposed method can identify the latent factor values in these months by using auxiliary data of higher frequency.

2.3 Main Objective and Key Challenges

The main objective of this paper is to estimate a complete factor model, that is, all the relevant factors for the target Y, their loadings on the target, and the implied common component for the target, and to provide a complete inferential theory for all components of the factor model. There are three key challenges for this objective, specifically in the presence of missing observations, that invalidate the standard approaches for latent factor estimation. In particular, applying conventional PCA to either Y, X, or a simple concatenated panel of X and Y will either be infeasible, inconsistent, or inefficient.

First, the information in Y may not be sufficient to estimate the factors at all time periods with conventional PCA methods. If some factors in F_y are weak in Y, then the identification assumptions underlying PCA are violated, and a conventional PCA estimate cannot separate those factors from the noise. This problem can arise even for fully observed data. If the panel Y has missing data, then this can pose additional challenges. For only partially observed data in Y, it is possible that even the strong factors in F_y cannot be estimated for all time periods. The leading example is when Y is only observed at a lower frequency, for example, annual data. In that case, we cannot infer the factor realizations for a higher frequency, for example, monthly, even for strong factors. As another complication, missingness can "weaken" the factor signal in the observed data, as the missing pattern can depend on the factor model and affect the effective sample size. In all these three cases, we can only learn the strong factors in F_y from the target Y for the time periods with sufficiently many observations. Hence, in these situations, we need to take advantage of the additional information in X.

Second, the auxiliary panel X may not contain all the factors in Y, and hence applying PCA only to X can fail to consistently estimate the factors for Y. In applications, there is no reason to assume a priori that the strong factors in a supplementary data set, that is not specifically targeted for Y, include all the factors F_y. Hence, we cannot ignore the information in Y, and need a way to combine the two panels.

Third, the dimensions of the panels X and Y can be very different. It is natural to assume that in many applications, the auxiliary panel X is much larger in the hope that it contains some useful information for Y. If N_x is much larger than N_y, then applying PCA to the concatenated panel of X and Y can only identify the factors F_x in X, but not the factors that are unique to Y and not included in F_x. This is because the target panel can receive a weight that is too low in a naively concatenated panel.

Therefore, the presence of these three challenges requires new methods to estimate all the factors in Y. Our target-PCA proposed in Section 2.4 below provides a solution to simultaneously address these three challenges. Target-PCA extracts the factors from X that are the most useful for explaining Y, and efficiently combines them with the information in Y.

In the following, we focus on the practically relevant and challenging case where the auxiliary data X does not contain all the factors in F_y, which is formalized in Assumption G1.

Assumption G1. *There exist some strong factors in the target Y that are not strong or not contained in the auxiliary panel X, that is,*

$$\operatorname{rank}\left(\lim_{N_x\to\infty}\frac{1}{N_x}\Lambda_{x,y_s}^\top\Lambda_{x,y_s}\right) < k_{y_s},$$

where $\Lambda_{x,y_s}\in\mathbb{R}^{N_x\times k_{y_s}}$ denotes the loadings in X that correspond to the strong factors in F_y, and k_{y_s} denotes the number of strong factors in Y.

Section 6.1 discusses the simpler case, where the strong factors in X include all factors F_y for Y, and hence the auxiliary panel X is sufficient to learn the factors in Y. This special case is a subset of our more comprehensive analysis.

2.4 Estimator

We propose a novel estimator, target-PCA, to estimate the latent factors in Y by combining the information in Y and X. Then we use the estimated factor model to impute missing observations in Y.

To illustrate the intuition of our target-PCA estimator, we start with the case where Y is fully observed. Recall, that in the conventional setup where we have only one fully-observed panel, we can estimate the latent factors that explain most variation in the data by minimizing the objective function of PCA. Our target-PCA estimator combines the PCA objective function for the auxiliary panel X and the PCA objective function for the target Y with a positive target weight γ as

$$\min_{F,\Lambda_x,\Lambda_y} \underbrace{\sum_{i=1}^{N_x}\sum_{t=1}^{T}\left(X_{ti} - F_t^\top(\Lambda_x)_i\right)^2}_{\text{auxiliary error}} + \gamma \cdot \underbrace{\sum_{j=1}^{N_y}\sum_{t=1}^{T}\left(Y_{tj} - F_t^\top(\Lambda_y)_j\right)^2}_{\text{target error}}. \tag{1}$$

This combination aims to extract the latent factors of Y, by optimally weighting the information in X and Y through the parameter γ. The target parameter γ can be interpreted as a reward for factors that help to explain the target panel Y. In this sense, γ is similar to a regularization parameter and controls the reward given to factors that reduce the error in explaining Y. As we will show next, γ can also be interpreted as a relative weight for the information in X and Y in a concatenated panel.

We can write (1) in matrix notation as

$$\min_{F,\Lambda_x,\Lambda_y} \operatorname{trace}\left(\left(\begin{bmatrix} X^\top \\ \sqrt{\gamma}Y^\top \end{bmatrix} - \begin{bmatrix} \Lambda_x \\ \sqrt{\gamma}\Lambda_y \end{bmatrix} F^\top\right)\left(\begin{bmatrix} X & \sqrt{\gamma}Y \end{bmatrix} - F\begin{bmatrix} \Lambda_x^\top & \sqrt{\gamma}\Lambda_y^\top \end{bmatrix}\right)\right). \tag{2}$$

For exposition, we introduce the notation $Z^{(\gamma)} \in \mathbb{R}^{T\times(N_x+N_y)}$ as the combination of X and Y with target weight γ, i.e.,

$$Z^{(\gamma)} := \begin{bmatrix} X & \sqrt{\gamma}Y \end{bmatrix} = F\Lambda^{(\gamma)\top} + e^{(\gamma)}, \quad \text{where } \Lambda^{(\gamma)\top} := \begin{bmatrix} \Lambda_x^\top & \sqrt{\gamma}\Lambda_y^\top \end{bmatrix} \in \mathbb{R}^{k\times(N_x+N_y)},$$
$$e^{(\gamma)} := \begin{bmatrix} e_x & \sqrt{\gamma}e_y \end{bmatrix} \in \mathbb{R}^{T\times(N_x+N_y)}.$$

With the identifying assumption $\Lambda^{(\gamma)\top}\Lambda^{(\gamma)}/(N_x+N_y) = I_k$, we can concentrate out F in the objective function (2), and obtain $\Lambda^{(\gamma)}$ from the following objective function[2]

$$\max_{\Lambda^{(\gamma)}} \operatorname{trace}\left(\Lambda^{(\gamma)\top}\big(Z^{(\gamma)\top}Z^{(\gamma)}\big)\Lambda^{(\gamma)}\right). \tag{4}$$

Hence, when Y is fully observed, we can estimate $\Lambda^{(\gamma)}$ by applying PCA to $Z^{(\gamma)\top}Z^{(\gamma)}$. Note that $Z^{(\gamma)\top}Z^{(\gamma)}/T$ is an estimator of the second moment of the weighted concatenated panel $Z^{(\gamma)}$. We denote the second population moment of $Z^{(\gamma)}$ as $\Sigma^{Z^{(\gamma)}}$, which equals the cross-sectional covariance matrix of $Z^{(\gamma)}$ in the case of demeaned data. Hence, target-PCA is equivalent to PCA on the weighted data $Z^{(\gamma)}$.

When Y has missing observations, we can also apply PCA to an estimator of $\Sigma^{Z^{(\gamma)}}$, but we need to account for the missing observations in $Z^{(\gamma)}$ in the estimation of $\Sigma^{Z^{(\gamma)}}$. Conceptually, we can use the time periods when both cross-sectional units i and j are observed to estimate $\Sigma_{ij}^{Z^{(\gamma)}}$, where $\Sigma_{ij}^{Z^{(\gamma)}}$ is the (i,j)-th entry in $\Sigma^{Z^{(\gamma)}}$. Formally, we introduce a new observation matrix W^Z, where $W_{ti}^Z = 1$ if $Z_{ti}^{(\gamma)}$ is observed and 0 otherwise. For any two cross-sectional units i and j, let $Q_{ij}^Z = \{t : W_{ti}^Z = W_{tj}^Z = 1\}$ be the set of time periods where both i and j are observed. We use the time periods in Q_{ij}^Z to estimate $\Sigma_{ij}^{Z^{(\gamma)}}$

$$\tilde{\Sigma}_{ij}^{Z^{(\gamma)}} = \frac{1}{|Q_{ij}^Z|}\sum_{t\in Q_{ij}^Z} Z_{ti}^{(\gamma)} Z_{tj}^{(\gamma)}, \quad i,j = 1,\cdots,N_x+N_y\,.$$

With the identifying assumption $\tilde{\Lambda}^{(\gamma)\top}\tilde{\Lambda}^{(\gamma)}/(N_x+N_y) = I_k$, the estimated loadings from PCA, denoted as $\tilde{\Lambda}^{(\gamma)}$, are $\sqrt{N_x+N_y}$ times the eigenvectors of the k largest eigenvalues of $\tilde{\Sigma}^{Z^{(\gamma)}}/(N_x + N_y)$, that is,

$$\frac{1}{N_x+N_y}\tilde{\Sigma}^{Z^{(\gamma)}}\tilde{\Lambda}^{(\gamma)} = \tilde{\Lambda}^{(\gamma)}\tilde{D}^{(\gamma)}\,,$$

where $\tilde{D}^{(\gamma)}$ is the diagonal matrix of the largest k eigenvalues of $\tilde{\Sigma}^{Z^{(\gamma)}}/(N_x+N_y)$. In this step, we simultaneously estimate the factor loadings in X and Y.

In a second step, we regress the observed $Z^{(\gamma)}$ on $\tilde{\Lambda}^{(\gamma)}$ to estimate the factors, that is,

$$\tilde{F}_t = \left(\sum_{i=1}^{N_x+N_y} W_{ti}^Z \tilde{\Lambda}_i^{(\gamma)}\tilde{\Lambda}_i^{(\gamma)\top}\right)^{-1}\cdot\left(\sum_{i=1}^{N_x+N_y} W_{ti}^Z Z_{ti}^{(\gamma)}\tilde{\Lambda}_i^{(\gamma)}\right), \quad t = 1,\cdots,T.$$

This step only uses the units that are observed at time t, and runs a weighted regression of the observed units' outcomes on the estimated loadings. The weight for units in Y is $W_{ti}^Y\cdot\gamma$, while the weight for units in X is 1.

[2]An alternative approach to obtain the optimal F, Λ_x and Λ_y in the objective function (1) is based on the identifying assumption $F^\top F/T = I_k$. Under this identifying assumption, we concentrate out both Λ_x and Λ_y in the objective function (2), and solve for F from the objective function

$$\max_{F} \operatorname{trace}\left(F^\top\big(Z^{(\gamma)}Z^{(\gamma)\top}\big)F\right). \tag{3}$$

Appendix D provides more details about how to adapt this alternative approach to allow for missing observations.

Next we estimate the common component of Y with the plug-in estimator $\tilde{C}_{ti} = \tilde{F}_t^\top (\tilde{\Lambda}_y)_i$, and use the estimated common components to impute the missing entries in Y. Note that our proposed estimator for the latent factor model in the partially observed Y is the same type of estimator as in Xiong and Pelger (2023). The key distinction is that our estimator is applied to the weighted concatenated panel $Z^{(\gamma)}$, while the estimator in Xiong and Pelger (2023) is applied to Y only.[3]

As shown in the original objective function (1), the key of target-PCA is the target weight γ. There are three special cases of target-PCA with different values of γ: First, when $\gamma = 0$, target-PCA degenerates to applying PCA to X. Second, when $\gamma = \infty$, target-PCA degenerates to applying PCA to Y. Third, when $\gamma = 1$, target-PCA is equivalent to applying PCA to the concatenated panel $Z^{(1)} = \begin{bmatrix} X & Y \end{bmatrix}$. The key problem is to select the target weight appropriately. Intuitively, the target weight γ should not be too small in order to ensure that the selected factors are relevant for Y, and γ cannot be too large in order to take advantage of the supplementary information in X.

The asymptotic distribution results for target-PCA require novel derivations and are not simply an application of Xiong and Pelger (2023) applied to $Z^{(\gamma)}$. A key challenge is that the target weight can go to infinity with the sample size, which would lead to exploding moments in some of the components of $Z^{(\gamma)}$ and hence violate the assumptions in existing frameworks. Therefore, we need to carefully consider the separate components of $Z^{(\gamma)}$, while allowing for joint asymptotics of γ growing with the sample size.

The main insight of target-PCA is that properly weighting the covariance matrices of multiple panels allows for the consistent and efficient estimation of latent factors. As will be shown in the next sections, the optimality of combining auxiliary data boils down to two aspects: (a) the detection of weak signals in the target panel; and (b) the efficient estimation of the factor structure. Our target-PCA estimator simultaneously achieves them in one step through the target weight γ. In fact, these two aspects correspond to two important effects of the target weight γ, which will be thoroughly discussed in Section 3.

3 Two Fundamental Effects of Target Weight γ

In this section, we illustrate and highlight the two important effects of the target weight γ in target-PCA: the consistency effect in factor identification and the efficiency effect in the estimation of factors and loadings. It is crucial to account for these two effects when choosing γ in target-PCA. In this section, we use simple models to explain these fundamental effects, and then provide the results for the general case in Section 5.

[3] It is possible to use other methods to estimate the factor model when the missing pattern has specific structures. For example, we can use Jin, Miao, and Su (2021) when observations are missing-at-random or use Bai and Ng (2021) and Cahan, Bai, and Ng (2023) for block-missing. The asymptotic results for alternative methods would require a case-by-case analysis, but we expect the general insights for the choice of the target weight to stay the same.

3.1 Effect 1: Consistency Effect of Target Weight γ

The first important effect is the consistent estimation of the factors. This effect is relevant when some factors in the observed Y are weak and cannot be identified by applying PCA to Y. For this case, if the weak factors in Y are strong factors in X, then it is possible to identify these factors with target-PCA. Specifically, if we choose $\gamma = r \cdot N_x/N_y$ for some positive constant r, then target-PCA can consistently estimate these weak factors. The intuition is as follows.

Target-PCA essentially estimates the factors from a matrix that combines $XX^\top$ and $\gamma\tilde{Y}\tilde{Y}^\top$, where $\tilde{Y} = Y \odot W^Y$ replaces the missing entries in Y with 0. Suppose that the auxiliary panel X is fully observed and the number of common observations between any two units of the target Y is proportional to T (the main setting studied in this paper). Then the top eigenvalues in $XX^\top$ and $\tilde{Y}\tilde{Y}^\top$ are at the order of N_x and N_y, respectively. If we select $\gamma = r \cdot N_x/N_y$ for some positive constant r, then the top eigenvalues in $XX^\top$ and $\gamma\tilde{Y}\tilde{Y}^\top$ are at the same order, and equivalently, the factor strengths of the strong factors in X and in Y are the same in the combined matrix of target-PCA. As the weak factors in $\tilde{Y}$ are strong factors in X, target-PCA can identify both the strong and weak factors in $\tilde{Y}$, leading to the consistency effect.

Next, we flesh out the intuition with a toy example. This two-factor example has the following key elements: (a) The dimension of the panel X is much larger than Y, i.e., $N_y/N_x \to 0$. (b) In panel Y, factor 1 is strong, but factor 2 is weak. Hence, we can only identify factor 1 from panel Y. (c) In panel X, factor 2 is strong, but factor 1 has zero exposure to the units. Hence, we can only identify factor 2 from panel X. Specifically, the loadings in Y follow

$$\begin{aligned}
&\text{first factor loadings: } (\Lambda_y)_{i1} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_{\Lambda_y}) \text{ for all } i\,,\\
&\text{second factor loadings: } (\Lambda_y)_{i2} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_{\Lambda_y}) \text{ for } i < N_y^{1/2}, \quad (\Lambda_y)_{i2} = 0 \text{ for } i \geq N_y^{1/2}\,,
\end{aligned}$$

and loadings in X follow

$$\begin{aligned}
&\text{first factor loadings: } (\Lambda_x)_{11} = \cdots = (\Lambda_x)_{N_x,1} = 0\,,\\
&\text{second factor loadings: } (\Lambda_x)_{i2} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_{\Lambda_x}) \text{ for all } i\,.
\end{aligned}$$

For simplicity, both the factors and idiosyncratic errors are drawn independently as $F_{t1}, F_{t2} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_F), (e_x)_{ti} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_{e_x})$, and $(e_y)_{ti} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2_{e_y})$ for all i and t. Furthermore, the factors and loadings have bounded fourth moments, and errors have bounded eighth moments. We assume that we can observe all the entries in Y. As we show in Section 5, the insights from this simplified setting carry over to our general model with missing observations.[4]

In this example, separate PCA on either X or Y is not able to consistently estimate both factors. However, target-PCA can identify both factors with an appropriately chosen γ. Without error terms

[4]Note that the observation pattern affects the asymptotic variance, but not the convergence rate of the estimated factor model or the top eigenvalues of $\tilde{Y}\tilde{Y}^\top$, as shown in Xiong and Pelger (2023). Therefore, the intuition for this toy example carries over to other observation patterns.

and missing observations, target-PCA estimates factors from

$$
\begin{aligned}
\frac{1}{N_x+N_y} Z^{(\gamma)} Z^{(\gamma)\top} &= \frac{1}{N_x+N_y}\begin{bmatrix} X & \sqrt{\gamma} Y \end{bmatrix}\begin{bmatrix} X^\top \\ \sqrt{\gamma} Y^\top \end{bmatrix} = \frac{1}{N_x+N_y}\begin{bmatrix} XX^\top + \gamma YY^\top \end{bmatrix} \\
&= \begin{bmatrix} F^{(1)} & F^{(2)} \end{bmatrix} \Bigg(\underbrace{\begin{bmatrix} \frac{\gamma \cdot N_y \sigma^2_{\Lambda_y}}{N_x+N_y} & \\ & \frac{N_x \sigma^2_{\Lambda_x} + \gamma \cdot N_y^{1/2} \sigma^2_{\Lambda_y}}{N_x+N_y} \end{bmatrix}}_{:=\hat{\Sigma}^{(\gamma)}_{\Lambda,t}} + o_p(1) \Bigg) \begin{bmatrix} F^{(1)\top} \\ F^{(2)\top} \end{bmatrix},
\end{aligned}
$$

where $F^{(1)} \in \mathbb{R}^{T\times 1}$ and $F^{(2)} \in \mathbb{R}^{T\times 1}$ denote the vector of the first and second factors, respectively. As $N_x, N_y \to \infty$, $\hat{\Sigma}^{(\gamma)}_{\Lambda,t}$ converges to

$$
\begin{aligned}
\lim_{N_x,N_y\to\infty} \hat{\Sigma}^{(\gamma)}_{\Lambda,t} &= \lim_{N_x,N_y\to\infty} \frac{N_x}{N_x+N_y} \left(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \sigma^2_{\Lambda_x} \end{bmatrix}}_{\Sigma_{\Lambda_x}} + \gamma \frac{N_y}{N_x} \cdot \underbrace{\begin{bmatrix} \sigma^2_{\Lambda_y} & 0 \\ 0 & \frac{\sigma^2_{\Lambda_y}}{N_y^{1/2}} \end{bmatrix}}_{\Sigma_{\Lambda_y,t}} \right) \\
&= \lim_{N_x,N_y\to\infty} \begin{pmatrix} \gamma \frac{N_y}{N_x} \cdot \sigma^2_{\Lambda_y} & 0 \\ 0 & \sigma^2_{\Lambda_x} + \gamma \frac{N_y^{1/2}}{N_x} \cdot \sigma^2_{\Lambda_y} \end{pmatrix},
\end{aligned}
$$

where Σ_{Λ_x} and $\Sigma_{\Lambda_y,t}$ are the second-moment matrices of the loadings of X and Y.[5] [6]

The key idea for selecting γ is to obtain full rank for the limit of the second-moment matrix $\hat{\Sigma}^{(\gamma)}_{\Lambda,t}$. Choosing $\gamma = r \cdot N_x/N_y$ with some positive constant r ensures that both eigenvalues in the limit of $\hat{\Sigma}^{(\gamma)}_{\Lambda,t}$ are of the same order, and therefore, both factors can be identified from $Z^{(\gamma)} Z^{(\gamma)\top}/(N_x+N_y)$. If γ is not chosen at this rate, for example, $\gamma = 1$, then only the second factor can be identified as a strong factor in the concatenated panel. We formalize the above discussion in the following proposition, which is a special case of the general Theorem 1, and provide further details in Section 5.

Proposition 1. *Under the data generating process and observation pattern described in this section, let $\delta_{N_y,T} = \min(N_y, T)$ and assume that $N_y/N_x \to 0$. Target-PCA with $\gamma = r \cdot N_x/N_y$ for some positive scaling constant r can consistently estimate the latent factors. As $T, N_x, N_y \to \infty$, there exists some rotation matrix H such that*

$$
\delta_{N_y,T} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \tilde{F}_t - H F_t \right\|^2 \right) = O_p(1).
$$

If $\gamma = O(1)$, then $\tilde{F}_t$ can be inconsistent.

[5] We use subscript t in $\Sigma_{\Lambda_y,t}$ to account for the case where there are missing observations in Y and the second-moment loading matrices of $\tilde{Y} = Y \odot W^Y$ is time-varying.

[6] If $N_y/N_x \to \infty$, we can estimate the second moment matrix $Z^{(\gamma)} Z^{(\gamma)\top}/N_y$. Selecting $\gamma = r \cdot N_x/N_y$ for a positive constant r can still ensure that $\hat{\Sigma}^{(\gamma)}_{\Lambda,t}$ converges to a full rank matrix, which allows for the consistent estimation of both factors.

3.2 Effect 2: Efficiency Effect of Target Weight γ

After we have selected γ in the right order to ensure consistency, we can improve the efficiency by selecting the optimal scale of γ. We call this the efficiency effect of γ. The essence of this effect is to use the target weight γ to balance the idiosyncratic noise levels between X and Y, thus achieving the smallest asymptotic variance in the estimation of factors, loadings, and common components.

In this section, we illustrate the efficiency effect of γ through a simple one-factor model, where factor identification is not a concern and where we can focus on the efficiency effect. The key element is that the idiosyncratic noise levels in X and Y are different.

More specifically, we assume that X and Y contain the same latent factor. Factors, loadings and idiosyncratic errors are drawn independently from $F_t \overset{\text{i.i.d.}}{\sim} (0, \sigma_F^2), (\Lambda_y)_i \overset{\text{i.i.d.}}{\sim} (0, \sigma_{\Lambda_y}^2), (\Lambda_x)_i \overset{\text{i.i.d.}}{\sim} (0, \sigma_{\Lambda_x}^2)$, $(e_x)_{ti} \overset{\text{i.i.d.}}{\sim} (0, \sigma_{e_x}^2)$, and $(e_y)_{ti} \overset{\text{i.i.d.}}{\sim} (0, \sigma_{e_y}^2)$. Furthermore, the factors and loadings have bounded fourth moments, and errors have bounded eighth moments. Suppose all entries in Y are missing at random with observation probability $\mathbb{P}(W_{ti}^Y = 1) = p$ and the number of units in X and Y are at the same order, i.e., $N_y/N_x \to c$ for some c bounded away from 0. As shown in our general model, the main conclusions do not depend on these specific assumptions.

In this setting, Proposition 2 provides the asymptotic distribution of the common components of Y. The result is a special case of the general Theorem 2 in Section 5. We optimize γ by minimizing the asymptotic variance.

Proposition 2. *Under the data generating process and observation pattern described in this section, let $\delta_{N_y,T} = \min(N_y, T)$ and suppose $N_y/N_x \to c \in (0, \infty)$. As $T, N_x, N_y \to \infty$, the asymptotic distribution of the estimated common component of Y for any i and t is*

$$\sqrt{\delta_{N_y,T}} (\Sigma_{C,ti}^{(\gamma)})^{-1/2} \left(\tilde{C}_{ti} - C_{ti} \right) \xrightarrow{d} \mathcal{N}(0, 1),$$

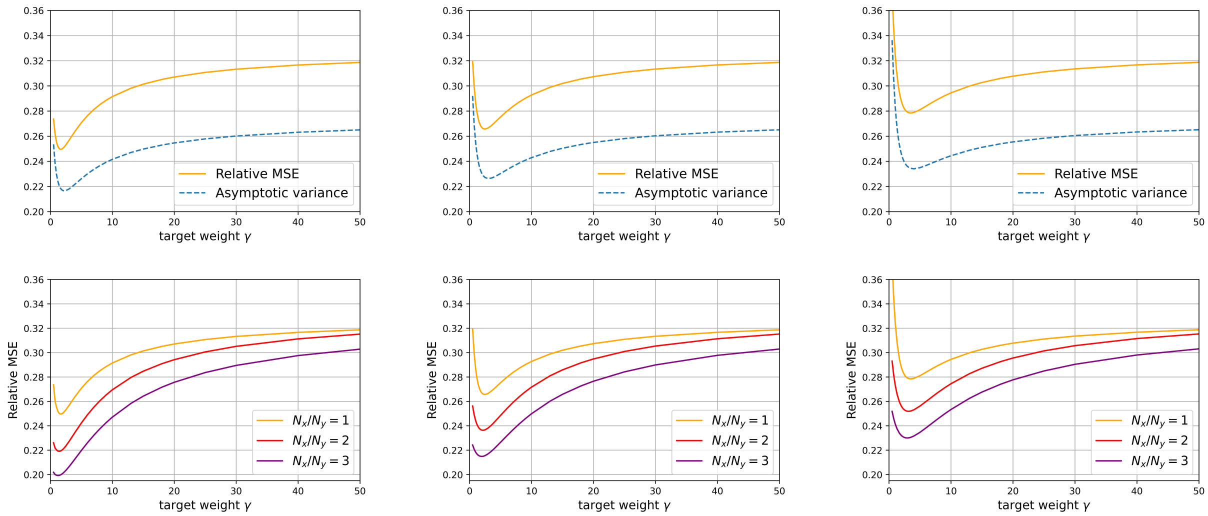
where

$$\begin{aligned} \Sigma_{C,ti}^{(\gamma)} =& \frac{\delta_{N_y T}}{T} \frac{\sigma_{e_y}^2}{p \sigma_F^2} F_t^2 + \frac{\delta_{N_y T}}{T} \left(\frac{1}{p} - 1 \right) \sigma_F^{-4} \text{Var}(F_t^2) (\Lambda_y)_i^2 F_t^2 \\ &+ \frac{\delta_{N_y T}}{N_y} (\Lambda_y)_i^2 \left(\sigma_{\Lambda_x}^2 + \gamma \frac{N_y}{N_x} p \sigma_{\Lambda_y}^2 \right)^{-2} \left(\frac{N_y}{N_x} \sigma_{\Lambda_x}^2 \sigma_{e_x}^2 + \gamma^2 \frac{N_y^2}{N_x^2} p \sigma_{\Lambda_y}^2 \sigma_{e_y}^2 \right). \end{aligned}$$

The optimized γ that minimizes $\Sigma_{C,ti}^{(\gamma)}$ is $\gamma^* = \sigma_{e_x}^2 / \sigma_{e_y}^2$ for any i and t.

The optimized γ^* minimizes $\Sigma_{C,ti}^{(\gamma)}$, which is the asymptotic variance of the common components as a function of γ. In this case, the optimized γ^* is the ratio of the variances of idiosyncratic errors $\sigma_{e_x}^2$ and $\sigma_{e_y}^2$. We up-weight the target panel Y if $\sigma_{e_y}^2 < \sigma_{e_x}^2$; otherwise, we down-weight the target panel Y. Interestingly and counter-intuitively, when observations are missing at random, the optimized γ^* is the same for all i and t, and does not depend on any of the following parameters, even though $\Sigma_{C,ti}^{(\gamma)}$ is a function of these parameters: the observation probability p, the value of factors F_t and loadings Λ_i, variance of factors σ_F^2 and loadings $\sigma_{\Lambda_x}^2$ and $\sigma_{\Lambda_y}^2$, and number of cross-section units

Figure 2: Relative MSE for different $\sigma_{e_x}/\sigma_{e_y}$ and N_x/N_y



(a) $\sigma_{e_x} = 6$ **(b)** $\sigma_{e_x} = 7$ **(c)** $\sigma_{e_x} = 8$

These figures show the relative MSE of the estimated common components of all entries in Y. The data generating process and the observation pattern follow the example described in Section 3.2. Specifically, the factors, loadings, and errors are generated from normal distributions with mean zero and $\sigma_F = \sigma_{\Lambda_y} = \sigma_\Lambda = 1$ and $\sigma_{e_y} = 4$, $N_y = T = 200$, and observation probability $p = 0.6$. For the subplots in the first row, we set $N_x = 200$. We run 200 simulations for each setup.

N_x and N_y. The intuition in this setting is the same as that for choosing the optimal weighting matrix for a weighted least squares (WLS) estimator. Specifically, suppose the heteroscedasticity of residuals does not depend on values of covariates as in this setting; then the optimal weighting matrix in WLS is proportional to the inverse variance of residuals and does not directly depend on the covariates and model parameters.

We illustrate the efficiency effect of this simple model in a simulation. Figure 2 shows the relative mean squared error (relative MSE) of the estimated common components and the theoretical asymptotic covariance $\Sigma_{C,ti}^{(\gamma)}$ averaged over all entries in Y as a function of γ. We consider different combinations of the noise ratio $\sigma_{e_x}/\sigma_{e_y}$ and the dimension ratio N_x/N_y for this one-factor example.[7]

There are three main takeaways from Figure 2. First, and most importantly, the target weight γ that minimizes the relative MSE equals the optimized γ^* that minimizes the asymptotic variance. Hence, we can use the inferential theory as a guidance for selecting the target weight γ. We will further elaborate on the optimal selection in Section 5. Second, we confirm that the optimized γ^* only varies with $\sigma_{e_x}/\sigma_{e_y}$, but not N_x/N_y in this missing at random example. Third, the effect of optimizing γ is larger if panel Y and X are more different in terms of noise and dimension. In all cases, PCA on Y only (target PCA with $\gamma = \infty$) or on X only (target PCA with $\gamma = 0$) is not optimal. The difference in relative MSE between target PCA with the optimized γ^* and the corner

[7]Section 7 presents a comprehensive simulation analysis and provides further details for the simulation setup, including the formal definition of relative MSE.

case of PCA on Y increases with N_x/N_y and $\sigma_{e_y}/\sigma_{e_x}$ (i.e., the information in X is more useful). The difference in relative MSE between target PCA with the optimized γ^* and the corner case of PCA on X only increases with N_y/N_x and $\sigma_{e_x}/\sigma_{e_y}$ (i.e., the information in X is less useful).

In summary, the consistency and efficiency effects combined provide the protocol for selecting γ. First, we select the rate of γ as $\gamma = r \cdot N_x/N_y$ to ensure consistency under general conditions. In the second step, we select the scaling positive constant r to minimize the asymptotic variance of the common component to ensure an efficient estimation.

4 Assumptions

In this section, we lay out the assumptions on the observation patterns of Y and the approximate factor models on both X and Y. First, we introduce the assumptions on the observation pattern.

Assumption G2 (Observation pattern)**.**

1. *The observation matrix W^Y is independent of the factors F and idiosyncratic errors e_y.*
2. *Let $Q^Y_{ij} = \{t : W^Y_{ti} = W^Y_{tj} = 1\}$ be the set of time periods when both units i and j of Y are observed. For any given observation matrix W^Y, there exists a positive constant q such that $|Q^Y_{ij}|/T \geq q > 0$ for all i, j. Let $q_{ij} = \lim_{T\to\infty} |Q^Y_{ij}|/T$ and $q_{ij,hl} = \lim_{T\to\infty} |Q^Y_{ij} \cap Q^Y_{hl}|/T$. For any i, j, h, l, q_{ij} and $q_{ij,hl}$ are positive constants bounded away from 0. Furthermore, the number of observed units in Y at any time period t is proportional to N_y, i.e., there exists a positive constant q' such that $N_y^{-1} \sum_{i=1}^{N_y} W^Y_{ti} \geq q' > 0$ for all t.*

Assumption G2 allows for very general observation patterns. The observation matrix W^Y can depend on the cross-sectional information, for example, the factor loadings Λ_y or time-invariant observed covariates of the units. For the purpose of identification, Assumption G2 rules out the dependence between W^Y and F. Note that the time and cross-section dimensions are symmetric in the estimation of common components. Therefore, the symmetric case where W^Y depends on F but is independent of Λ_y is allowed by swapping the roles of N and T in the estimation of the factor model. We assume that W^Y is independent of e_y, which is conceptually similar to the unconfoundedness assumption in Rosenbaum and Rubin (1983).

In Assumption G2, we assume that the number of time periods, for which any four units are simultaneously observed, grows at the order of T. This assumption implies that every entry in the covariance matrix $\Sigma^{Z^{(\gamma)}}$ can be consistently estimated at the rate $\sqrt{T}$, and ensures that the asymptotic variances of the estimated factors, loadings and common components from Target-PCA are well-defined. In Appendix C, we generalize Assumption G2 to allow $|Q^Y_{ij}|$ to grow sub-linearly in T. This does not change the conceptual arguments, but leads to a more complex notation.

We assume that both, X and Y, follow an approximate factor model similar to Bai (2003). We allow for non-trivial time-series dependency of factors, and non-trivial cross-sectional dependency of loadings in X, in Y, and between X and Y. In addition, the idiosyncratic errors can be weakly correlated in X, Y, and between X and Y, in both the time-series and cross-sectional dimensions.

The asymptotic distributions are based on general martingale central limit theorems. However, we make a key relaxation for the factor model relative to Bai (2003). Some factors in Y are allowed to be weak and only affect a small subset of units in Y. However, those factors have to be strong in X in order to identify them from $Z^{(\gamma)}$ with a properly chosen γ (as suggested in Section 3).

The assumptions for the general factor model are stated in Assumptions G3 and G4 in Appendix A. They are delegated to the Appendix as most elements are standard, but fairly technical. Our consistency result in Theorem 1 and asymptotic distribution result in Theorem 2 are derived under Assumptions G2, G3 and G4.

We use the simplified model in Example 1 to illustrate how our factor model generalizes the conventional approximate factor models. It allows us to highlight the relaxation of weak factors. Appendix B shows how our main results are simplified for this example and shows that this example is a special case of our general framework.

Example 1 (Simplified factor model)**.** *There exists constant $C < \infty$ such that*

1. *Factors: $F_t \overset{i.i.d.}{\sim} (0, \Sigma_F)$ and $\mathbb{E}\|F_t\|^4 \leq C$ for any t.*
2. *Loadings: $(\Lambda_x)_i \overset{i.i.d.}{\sim} (0, \Sigma_{\Lambda_x})$, where Σ_{Λ_x} is positive semidefinite. $(\Lambda_y^{\text{full}})_i \overset{i.i.d.}{\sim} (0, \Sigma_{\Lambda_y}^{\text{full}})$ and the loading of the j-th factor $(\Lambda_y)_{ij} = (\Lambda_y^{\text{full}})_{ij} \cdot (U_y)_{ij}$, where $\Sigma_{\Lambda_y}^{\text{full}}$ is positive definite and the Bernoulli random variable $(U_y)_{ij} \in \{0, 1\}$ is independent in i with $\mathbb{P}((U_y)_{ij} = 1) = p_j$ for some $p_j \in [0, 1]$. Furthermore, $\mathbb{E}\|(\Lambda_x)_i\|^4 \leq C$, $\mathbb{E}\|(\Lambda_y)_i\|^4 \leq C$, $N_y^{-1} \sum_{i=1}^{N_y} (\Lambda_y)_i (\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y}$, and $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y}$ is positive definite. For any t, $N_y^{-1} \sum_{i=1}^{N_y} W_{ti}^Y (\Lambda_y)_i (\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y,t}$ and $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y,t}$ is positive definite.*
3. *Idiosyncratic errors: $(e_x)_{ti} \overset{i.i.d.}{\sim} (0, \sigma_{e_x}^2), (e_y)_{ti} \overset{i.i.d.}{\sim} (0, \sigma_{e_y}^2)$, $\mathbb{E}(e_x)_{ti}^8 \leq C, \mathbb{E}(e_y)_{ti}^8 \leq C$.*
4. *Independence: $F, \Lambda_x, \Lambda_y, e_x$ and e_y are independent.*

The main difference to the conventional factor models is the loadings, while the factors and errors capture the stylized properties in a usual factor setup. The simplified model in this example assumes that all observations are i.i.d. Allowing for more complex dependencies as in our general model, does not change the arguments, but makes the notation more complex. The key element is the strength of the factors measured by their loadings. Specifically, we measure the strength of the factors by the fraction p_j of units in Y that are affected by the corresponding factor. The error terms are non-systematic with bounded eigenvalues in the covariance matrix.

The assumptions on the loadings $(\Lambda_y)_i$ account for three cases of factor strength in Y. First, if p_j is bounded away from 0 as N_y grows, then the j-th factor is a strong factor in Y. Second, if p_j decays to 0 but is nonzero as N_y grows, then the j-th factor is a weak factor in Y. Third, if p_j is 0 for all N_y, then Y does not contain the j-th factor. Note that Σ_{Λ_x} can be rank deficient, implying that the loadings of some factors can be zero for units in X. However, our assumption on $(\Lambda_x)_i$ rules out the case of weak factors in X as the estimation of weak factors in X is not our objective. We assume that $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y}$ is positive definite to ensure that each factor in F is strong in at least one of the two panels X and Y. Specifically, weak factors in Y are strong in X. Hence, all factors can be identified with target-PCA with a properly chosen γ.

The loading assumption also imposes assumptions on the missing pattern in Y to identify all factors when combining the partially observed Y and X. More specifically, the second-moment matrix $\Sigma_{\Lambda_y,t}$ does not need to be full rank in Assumption S1.2, which relaxes the full-rank assumption of $\Sigma_{\Lambda_y,t}$ in Xiong and Pelger (2023). However, we assume that $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y,t}$ is positive definite, so that target-PCA can identify all factors from X and partially observed Y.

We assume that the number of factors k can be consistently estimated. Given a consistent estimator for the number of factors, we can treat k as known. After selecting γ at the rate $r \cdot N_x/N_y$ for some positive scaling constant r, the estimation of the number of factors from the weighted concatenated panel $Z^{(\gamma)}$ is the same as in Xiong and Pelger (2023). Hence, 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 is to use cross-validation arguments. However, this is non-trivial for complex missing patterns that can depend on the factor model itself. In summary, given our analysis and selecting γ at the right rate, we can transform the problem of estimating the number of factors into a familiar setup. Furthermore, in our empirical analysis, we show that our estimator can be robust to the number of factors once γ is selected appropriately.

5 Inferential Theory

In this section, we provide the asymptotic results of the estimated factor model from target-PCA under general assumptions on the approximate factor model and missing patterns. We present the consistency result in Section 5.1 and asymptotic normality results in Section 5.2.

5.1 Consistency

The loadings and factors can be consistently estimated only if γ is properly chosen. The consistency result is an important intermediate step to show the inferential theory in Section 5.2.

Theorem 1. *Let $\delta_{N_y,T} = \min(N_y, T)$ and suppose that $N_y/N_x \to c \in [0, \infty)$. Under Assumptions G2 and G3, for $T, N_x, N_y \to \infty$:*

1. *If $\gamma = r \cdot N_x/N_y$ for some positive constant r, then $\Sigma_\Lambda^{(\gamma)} := \lim_{N_x,N_y\to\infty} N_x(N_x + N_y)^{-1}(\Sigma_{\Lambda_x} + \gamma N_y/N_x \cdot \Sigma_{\Lambda_y})$ and $\Sigma_{\Lambda,t}^{(\gamma)} := \lim_{N_x,N_y\to\infty} N_x(N_x + N_y)^{-1}(\Sigma_{\Lambda_x} + \gamma N_y/N_x \cdot \Sigma_{\Lambda_y,t})$ are positive definite. It holds that*

$$\delta_{N_y,T} \left(\frac{1}{N_x + N_y} \sum_{i=1}^{N_x+N_y} \left\| \tilde{\Lambda}_i^{(\gamma)} - H^{(\gamma)} \Lambda_i^{(\gamma)} \right\|^2 \right) = O_p(1), \tag{5}$$

$$\delta_{N_y,T} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \tilde{F}_t - (H^{(\gamma)\top})^{-1} F_t \right\|^2 \right) = O_p(1), \tag{6}$$

where $H^{(\gamma)} = T^{-1}(N_x + N_y)^{-1}(\tilde{D}^{(\gamma)})^{-1}\tilde{\Lambda}^{(\gamma)\top}\Lambda^{(\gamma)}F^\top F$. *This implies that the estimated loadings* $(\tilde{\Lambda}_y)_i$ *and common components* $\tilde{C}_{ti}$ *of* Y *are consistent.*

2. *Under Assumption G1, if* γ *and* N_x/N_y *are not of the same order, then* $\Sigma_{\Lambda,t}^{(\gamma)}$ *is not positive definite when* $N_y/N_x \to 0$. *If* $\Sigma_{\Lambda,t}^{(\gamma)}$ *is not positive definite, then* $\tilde{F}_t$ *does not converge at the rate* $\delta_{N_y,T}$ *and can be inconsistent for the factors that are strong in* Y *but not in* X.

Theorem 1 states the consistency effect of γ and generalizes Proposition 1 to general factor models. According to Theorem 1, choosing $\gamma = r \cdot N_x/N_y$ with some positive constant r ensures that factors, loadings, and common components of Y can be consistently estimated up to a rotation matrix $H^{(\gamma)}$. The convergence rate is at the smaller of $\sqrt{N_y}$ and $\sqrt{T}$. This rate is the same as the convergence rate in Bai and Ng (2002) that applies PCA to Y when all factors are strong in Y. This rate makes sense for target-PCA: When we up-weight Y by a rate of N_x/N_y, the error from Y is always a leading term in target-PCA.

If γ is not selected at the right order, the estimates of the factors that are strong in Y, but not in X, are inconsistent unless stronger assumptions are imposed. Specifically, if $N_y = N_x^\alpha$ for some $\alpha \in (0,1)$ and $\gamma = 1$, then the consistency requires $(N_x/N_x^\alpha)\cdot(1/T) \to 0$, that is, a larger T, analogous to Assumption A4 in Bai and Ng (2023). However, this assumption is not required if γ is chosen at the order of N_x/N_y.

5.2 Asymptotic Normality

Based on the consistency results in Theorem 1, we develop the inferential theory for target-PCA in this section. Theorem 2 shows the asymptotic distribution of estimated factors, estimated loadings of Y, and estimated common components of Y with target weight $\gamma = r \cdot N_x/N_y$ for every positive scaling constant r in target-PCA under general assumptions. We show the asymptotic distribution of Y because the factor model and common components of Y are of our primary interest. The asymptotic distribution of X can be shown analogously.

Theorem 2. *Define* $\delta_{N_y,T} = \min(N_y, T)$. *Suppose that* $N_y/N_x \to c \in [0,\infty)$ *and* $\gamma = r \cdot N_x/N_y$ *for some positive constant* r. *If the eigenvalues of* $\Sigma_F\left(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\right)$ *are distinct, then under Assumptions G2, G3 and G4, as* $T, N_x, N_y \to \infty$ *we have for each* i *and* t:

1. *For* $\sqrt{T}/N_y \to 0$, *the asymptotic distribution of the estimated loadings of* Y *is*

$$\sqrt{T}(\Sigma_{\Lambda_y,i}^{(\gamma)})^{-1/2}\left((H^{(\gamma)})^{-1}(\tilde{\Lambda}_y)_i - (\Lambda_y)_i\right) \xrightarrow{d} \mathcal{N}(0, I_k), \tag{7}$$

where

$$\Sigma_{\Lambda_y,i}^{(\gamma)} = \Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}} + \Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{miss}},$$

$\Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}} = \Sigma_F^{-1}(\Sigma_\Lambda^{(\gamma)})^{-1}\Gamma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}}(\Sigma_\Lambda^{(\gamma)})^{-1}\Sigma_F^{-1}$, $\Gamma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}}$ *is defined in Assumption G4.6,* $\Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{miss}} = \Sigma_F^{-1}(\Sigma_\Lambda^{(\gamma)})^{-1}h_{i+N_x}^{(\gamma)}((\Lambda_y)_i)(\Sigma_\Lambda^{(\gamma)})^{-1}\Sigma_F^{-1}$, *and the function* $h_i^{(\gamma)}(\cdot)$ *is defined in Assumption G4.8.*

2. *For* $\sqrt{T}/N_y \to 0$ *and* $\sqrt{N_y}/T \to 0$,

- *Case 1: If all the factors in F_y are strong factors in Y, then the asymptotic distribution of the estimated factors is*

$$\sqrt{\delta_{N_y,T}}(\Sigma_{F,t}^{(\gamma)})^{-1/2}(H^{(\gamma)\top}\tilde{F}_t - F_t) \xrightarrow{d} \mathcal{N}(0, I_k), \tag{8}$$

where

$$\Sigma_{F,t}^{(\gamma)} = \frac{\delta_{N_y,T}}{N_y}\Sigma_{F,t}^{(\gamma),\mathrm{obs}} + \frac{\delta_{N_y,T}}{T}\Sigma_{F,t}^{(\gamma),\mathrm{miss}},$$

$\Sigma_{F,t}^{(\gamma),\mathrm{obs}} = (\Sigma_{\Lambda,t}^{(\gamma)})^{-1}\Gamma_{F,t}^{(\gamma),\mathrm{obs}}(\Sigma_{\Lambda,t}^{(\gamma)})^{-1}$, $\Gamma_{F,t}^{(\gamma),\mathrm{obs}}$ *is defined in Assumption G4.7,* $\Sigma_{F,t}^{(\gamma),\mathrm{miss}} = (\Sigma_{\Lambda,t}^{(\gamma)})^{-1}\cdot g_t^{(\gamma)}\big((\Sigma_{\Lambda}^{(\gamma)})^{-1}\Sigma_F^{-1}F_t\big)\cdot(\Sigma_{\Lambda,t}^{(\gamma)})^{-1}$*, and the function* $g_t^{(\gamma)}(\cdot)$ *is defined in Assumption G4.8;*

- *Case 2: Suppose some factors in F_y are weak factors in Y. Let $F_{t,w}$ be the weak factors in Y, and let $F_{t,s} = F_t \backslash F_{t,w}$ be the remaining strong factors in F_t. For simplicity of notation, we assume that the loadings of the weak factors $F_{t,w}$ are asymptotically orthogonal to the loadings of the strong factors $F_{t,s}$. The asymptotic distribution of the estimated weak factors $(H^{(\gamma)\top}\tilde{F}_t)_w$ corresponding to $F_{t,w}$ is*

$$\sqrt{\delta_{N_w,T}}(\Sigma_{F_w,t}^{(\gamma)})^{-1/2}((H^{(\gamma)\top}\tilde{F}_t)_w - F_{t,w}) \xrightarrow{d} \mathcal{N}(0, I_k), \tag{9}$$

where $\delta_{N_w,T} = \min(N_w, T)$, $N_w = \min(N_y^2/g(N_y), N_x)$, $g(N_y)$ *is the rate at which* $\sum_{i=1}^{N_y}(\Lambda_y)_{i,w}^2$ *grows,*

$$\Sigma_{F_w,t}^{(\gamma)} = \frac{\delta_{N_w,T}}{N_w}\Sigma_{F_w,t}^{(\gamma),\mathrm{obs}} + \frac{\delta_{N_w,T}}{T}\Sigma_{F_w,t}^{(\gamma),\mathrm{miss}},$$

$\Sigma_{F_w,t}^{(\gamma),\mathrm{obs}} = (\Sigma_{\Lambda,t,w}^{(\gamma)})^{-1}\Gamma_{F_w,t}^{(\gamma),\mathrm{obs}}(\Sigma_{\Lambda,t,w}^{(\gamma)})^{-1}$, $\Gamma_{F_w,t}^{(\gamma),\mathrm{obs}}$ *is defined in Assumption G4.7,* $\Sigma_{\Lambda,t,w}^{(\gamma)}$ *and* $\Sigma_{F_w,t}^{(\gamma),\mathrm{miss}}$ *are respectively the diagonal blocks of* $\Sigma_{\Lambda,t}^{(\gamma)}$ *and* $\Sigma_{F,t}^{(\gamma),\mathrm{miss}}$ *corresponding to the weak factors.*[8]

3. *For $\sqrt{T}/N_y \to 0$ and $\sqrt{N_y}/T \to 0$, the asymptotic distribution of the estimated common components of Y is*

$$\sqrt{\delta_{N_y,T}}(\Sigma_{C,ti}^{(\gamma)})^{-1/2}(\tilde{C}_{ti} - C_{ti}) \xrightarrow{d} \mathcal{N}(0,1), \tag{10}$$

where

$$\begin{aligned}\Sigma_{C,ti}^{(\gamma)} =& \frac{\delta_{N_y,T}}{T}F_t^\top\left(\Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}} + \Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{miss}}\right)F_t + \frac{\delta_{N_y,T}}{N_y}(\Lambda_y)_i^\top\Sigma_{F,t}^{(\gamma),\mathrm{obs}}(\Lambda_y)_i \\ &+ \frac{\delta_{N_y,T}}{T}(\Lambda_y)_i^\top\Sigma_{F,t}^{(\gamma),\mathrm{miss}}(\Lambda_y)_i - 2\frac{\delta_{N_y,T}}{T}(\Lambda_y)_i^\top\Sigma_{\Lambda_y,F,i,t}^{(\gamma),\mathrm{miss,cov}}F_t,\end{aligned}$$

$\Sigma_{\Lambda_y,F,i,t}^{(\gamma),\mathrm{miss,cov}} = (\Sigma_{\Lambda,t}^{(\gamma)})^{-1}\cdot g_{i,t}^{(\gamma),\mathrm{cov}}\big((\Lambda_y)_i, (\Sigma_{\Lambda}^{(\gamma)})^{-1}\Sigma_F^{-1}F_t\big)\cdot(\Sigma_{\Lambda}^{(\gamma)})^{-1}\Sigma_F^{-1}$*, and function* $g_{i,t}^{(\gamma),\mathrm{cov}}(\cdot,\cdot)$ *is defined in Assumption G4.8.*

[8] The asymptotic distribution of the estimated strong factors $(H^{(\gamma)\top}\tilde{F}_t)_s$ is $\sqrt{\delta_{N_y,T}}(\Sigma_{F_s,t}^{(\gamma)})^{-1/2}((H^{(\gamma)\top}\tilde{F}_t)_s - F_{t,s}) \xrightarrow{d} \mathcal{N}(0, I_k)$, where $\Sigma_{F_s,t}^{(\gamma)}$ is the diagonal block of $\Sigma_{F,t}^{(\gamma)}$ corresponding to the strong factors $F_{t,s}$.

The factors, loadings, and common components are asymptotically normally distributed. The asymptotic variance differs from the conventional PCA in Bai (2003) in three aspects. First, the asymptotic variances $\Sigma^{(\gamma)}_{\Lambda_y,i}$, $\Sigma^{(\gamma)}_{F,t}$ and $\Sigma^{(\gamma)}_{C,ti}$ depend on γ (or equivalently r when $\gamma = r \cdot N_x/N_y$). This will be important, as we will use it as the criteria to select the scale of γ. Second, in the case of missing data, we have additional correction terms to capture the additional uncertainty due to missingness. These correction terms follow the same structure and arguments as in Xiong and Pelger (2023). Without missing data, the correction matrices in the variance disappear. Third, as shown in Theorem 2.2, the strong and weak factors in Y have different convergence rates.[9] Importantly, this separation between strong and weak factors does not affect the asymptotic distribution of the estimated common components of Y. Hence, the asymptotic variance of the common components allows us to select an efficient target weight γ independent of the factor strength.

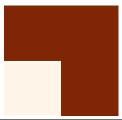
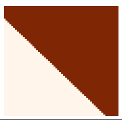
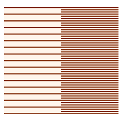
The distribution results of Theorem 2 simplify under Example 1, and we can provide explicit expressions for the asymptotic variances. Appendix B shows the analytical expression of the asymptotic variances under the simplified factor model, which allows us to gain intuition on how γ affects the efficiency of the estimation. The asymptotic variances depend on the following key quantities: the target weight γ, the noise ratio (NR) $\sigma_{e_x}/\sigma_{e_y}$, the dimension ratio (DR) N_x/N_y and the dependency structure in the missing pattern. We illustrate this dependency and the effect on the optimal choice of γ with a simulation example based on the model in Example 1 and different observation patterns.

We aim to select the optimized γ^* as the efficient target weight for the common components. The numerical examples in Table 1 illustrate how the optimized γ^* that minimizes $\sum_{t,i} \Sigma^{(\gamma)}_{C,ti}$ depends on the observation pattern, noise ratio and dimension ratio, in a one-factor model. The fraction of observed data p in Y varies between 60%, 75% and 90% with the four missing patterns: missing-at-random, block-missing, staggered-missing and mixed-frequency. The dimension ratio is either $N_x/N_y = 1$ or 4, and the noise ratio $\sigma_{e_x}/\sigma_{e_y}$ varies between 0.25, 1 and 4.

This numerical example illustrates three points. First, the optimized γ^* can substantially deviate from the naive concatenating weight 1, even when the dimensions of X and Y are the same. In particular, for complex missing patterns, the optimized target weight can deviate substantially from equally weighting the panels. Second, when observations are missing at random, the optimized γ^* only depends on the noise NR, but not on the dimension ratio N_x/N_y or fraction of observed entries p, confirming Proposition 2 in the illustration of the efficiency effect. However, for more complex observation patterns, the optimized γ^* can depend on N_x/N_y, p, and other quantities related to the observation pattern. Specifically, the optimized γ^* generally increases with N_x/N_y and $1-p$,

[9] The estimated weak factors $F_{t,w}$ have faster convergence rates than the estimated strong factors $F_{t,s}$. This is a consequence of $g(N_y)$ growing at a smaller rate than N_y, which implies that $\delta^{1/2}_{N_w,T}$ grows at a larger rate than $\delta^{1/2}_{N_y,T}$. This result might seem to be counterintuitive at first glance, but makes sense after a careful analysis of the factor estimation errors. Note that the weak factors $F_{t,w}$ are essentially estimated from X. Their estimation errors are dominated by the cross-sectional average of the errors of X. In contrast, the strong factors $F_{t,s}$ are mainly estimated from Y and their estimation errors are dominated by those of Y, which then converge to zero at a slower rate as Y has fewer units than X. The asymptotic distribution of the estimated common components is dominated by that of the strong factors in Y, as strong factors have a slower convergence rate than weak factors. Therefore, the two cases in Theorem 2.2 lead to the same case in Theorem 2.3.

Table 1: Optimized target weight γ^* for different missing patterns and noise ratios

	p	$N_x/N_y = 1$			$N_x/N_y = 4$		
		NR=0.25	NR=1	NR=4	NR=0.25	NR=1	NR=4
	60%	0.25	1.00	4.00	0.25	1.00	4.00
	75%	0.25	1.00	4.00	0.25	1.00	4.00
	90%	0.25	1.00	4.00	0.25	1.00	4.00
	60%	0.61	1.75	4.25	1.95	5.09	7.00
	75%	0.42	1.53	4.35	1.06	3.62	6.12
	90%	0.28	1.15	4.18	0.40	1.62	4.61
	60%	0.55	1.96	4.66	1.69	5.52	7.84
	75%	0.39	1.47	4.35	0.93	3.26	5.87
	90%	0.28	1.13	4.13	0.40	1.58	4.51
	60%	0.35	1.72	4.41	0.99	4.43	6.89
	75%	0.35	1.41	4.26	0.79	3.00	5.64
	90%	0.30	1.15	4.10	0.48	1.78	4.60

This table reports the optimized γ^* for different missing patterns, dimension ratios N_x/N_y and noise ratios (NR) $\sigma^2_{e_x}/\sigma^2_{e_y}$. In this table, the optimized γ^* equals N_x/N_y multiply by the efficient scaling r^* obtained by minimizing $\sum_{t,i} \Sigma^{(\gamma)}_{C,ti}$ in Corollary 1. The figures on the left show the observation patterns, with the shaded entries indicating the missing entries. We set $N_y = T = 50$ and the fraction of observed entries p to $60\%, 75\%$ and 90%. We generate a one-factor model where factors, loadings, and errors are drawn from normal distributions with $\Sigma_F = \Sigma_{\Lambda_x} = \Sigma_{\Lambda_y} = 1$. We let $\sigma^2_{e_x} = 1$ and $\sigma^2_{e_y} = 4$ for NR $= 0.25$, $\sigma^2_{e_x} = \sigma^2_{e_y} = 1$ for NR $= 1$, and $\sigma^2_{e_x} = 4$ and $\sigma^2_{e_y} = 1$ for NR $= 4$. The missing patterns in this table are generated as follows: (a) Missing uniformly at random: Entries are independently observed with probability p. (b) Block-missing pattern: $2(1-p)$ fraction of randomly selected units are missing from time $0.5 \cdot T$. (c) Staggered treatment pattern: All units are in the control group for $t < c \cdot T$. Starting from time $t = c \cdot T$, $(t/T - c)$ fraction of randomly selected units are in the treated group at time t. (d) Mixed-frequency observation: Entries in the first half of the units are simultaneously observed at every t_1 time period and entries in the second half of the units are simultaneously observed at every t_2 time period. For $p = 60\%$, $t_1 = 1/35\%$ and $t_2 = 1/85\%$; for $p = 75\%$, $t_1 = 1/60\%$ and $t_2 = 1/90\%$; for $p = 90\%$, $t_1 = 1/80\%$ and $t_2 = 1$.

implying that when the number of observations (or effective sample size) on Y is small compared to X, the optimized γ^* increases to balance the relative contributions of the two panels. Third, the optimized γ^* grows with a larger dependency in the missing pattern. For the four observation patterns considered in Table 1, the dependency between the entries of W^Y is generally the highest for the block-missing pattern, followed by the staggered treatment pattern, then the mixed-frequency pattern, while missing-at-random has the lowest dependency. A higher dependency implies a smaller effective sample size in Y and hence a larger γ^*.

One 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. The same arguments as in Xiong and Pelger (2023) for how to use the results for causal inference apply to target-PCA. We can test in an analogous way for point-wise treatment effects that can be heterogeneous and time-dependent under general adoption patterns where the units can be affected by unobserved factors. Importantly, by optimally leveraging auxiliary data, we allow for more general adoption patterns and more precise estimates of the counterfactual outcomes.

5.3 Selection of Target Weight γ

As we have seen, selecting γ appropriately is crucial for the consistent and efficient estimation of the latent factor model on Y. We suggest a two-stage approach for choosing γ based on our inferential theory.

In the first stage, we select $\gamma^{\text{first-stage}} = N_x/N_y$ to consistently estimate the latent factor model. Based on Theorem 1, we can consistently estimate all the factors and loadings using target-PCA with $\gamma^{\text{first-stage}}$. In the second stage, we estimate the asymptotic variance of the common component $\Sigma_{C,ti}^{(\gamma)}$ using the estimated factor model from the first stage and the inferential theory in Theorem 2 and Corollary 1. We then select the scaling constant to achieve efficiency by minimizing a linear combination of the estimated $\hat{\Sigma}_{C,ti}^{(\gamma)}$. If our goal is to estimate all common components as precisely as possible, then the objective function is to minimize $\sum_{t,i} \Sigma_{C,ti}^{(\gamma)}$. This is the objective that we use in our applications. If we want to impute the missing entries in Y as precisely as possible, then an appropriate objective function would be to minimize $\sum_{t,i}(1 - W_{ti}^Y) \cdot \Sigma_{C,ti}^{(\gamma)}$.

Generally, cross-validation could be an alternative approach for selecting γ. The idea of a cross-validation approach is to mask some observed entries in Y and select the value of γ that can most precisely estimate these masked out-of-sample entries. The challenge lies in how to mask the observed entries, and an appropriate implementation is more complicated than what it might appear to be. In particular, some naive masking schemes, such as random masking, may not be appropriate if the actual missing pattern is more complex. Intuitively, the masking should replicate the missing pattern, in order to ensure that the imputation errors of the masked entries can (unbiasedly) estimate the imputation errors of the missing entries in Y (if imputing missing entries in Y is our main goal). However, estimating the propensity of missingness is challenging and can be sensitive to the specification of a propensity model. In addition, the observation pattern can depend on the latent factor model itself, which further complicates the estimation. Our selection criterion based on the inferential theory avoids these issues.

6 Extensions

6.1 Auxiliary Data X Sufficient to Estimate All Factors

So far, we have focused on the important setting where the auxiliary data X is not sufficient to estimate all the factors for the target Y, which is assumed in Assumption G1. A simpler case is where X already contains all the information to learn the factors in Y, that is, all the factors in Y are strong in X. This is a special case of our more general analysis, where we can consistently estimate the factors by applying PCA to X and we would only use target-PCA to achieve higher efficiency. Theorem 2 applies with minor modifications.

For this case, if $N_y/N_x \to 0$, then choosing $\gamma = r$ with any constant r is equivalent to applying PCA on X, which has the convergence rate of N_x. Here, choosing $\gamma = r \cdot N_x/N_y$ is sub-optimal, since it up-weights Y and thus slows down the convergence rate of the estimated factors to N_y. If $N_y/N_x \to c$ with c bounded away from 0, then it is beneficial to include the target panel Y in the estimation of factors to improve the efficiency. Specifically, we can select the optimized $\gamma^* = r$ based on the asymptotic normality results similar to Theorem 2.

This setting is relevant when the target panel Y has a low-frequency observation pattern with no available information in Y for some time periods. If X contains all the necessary information to estimate the latent factors in Y, then we can accurately impute the value of Y in the periods with no observations and, hence, obtain an imputed target panel with higher frequency observations.

We recommend the choice of target weight $\gamma = r \cdot N_x/N_y$ with a positive constant r from the main setting for robustness. Even in the case where all relevant factors are strong in X, selecting $\gamma = r \cdot N_x/N_y$ ensures consistent estimation. Importantly, this rate also guarantees consistency, when not all factors can be estimated from X. In practice, we do not know if all factors for Y are strong factors in X. In many applications, X might not be selected in a targeted way, and hence it is likely that factors needed for Y can be missing or weak in X. The selection procedure in Section 5.3 provides a robust solution.

6.2 Finite Cross-Sectional Dimension of Target Data

So far, we have focused on the case where $N_y \to \infty$. In some practical applications, N_y may be finite (Huang, Jiang, Li, Tong, and Zhou, 2022). Our results can be extended to the case of finite N_y with minor technical modifications.[10] Specifically, we consider the choice of γ in two different settings depending on the nature of units in Y.

In the first setting, the units in Y are similar to the units in X, and the idiosyncratic noise level in X and Y are at the same scale, that is, $\mathbb{E}[(e_y)^2_{ti}]/\mathbb{E}[(e_x)^2_{ti}] = O(1)$. For this case, if all the factors in Y can be identified by applying PCA to X, then selecting $\gamma = 1$ (i.e., PCA on X) is optimal. This is the degenerate case that asymptotically does not require Y for target-PCA.

In the second setting, units in Y are (weighted) averages of M units, where M is much larger than N_y and can be of the same order as N_x. An important example is when Y are the principal components from another panel. In this case, if $\mathbb{E}[(e_y)^2_{ti}]/\mathbb{E}[(e_x)^2_{ti}] = O(1/M)$, then we should choose $\gamma = r \cdot N_x$, such that target-PCA can identify the factors in Y that are either weak or nonexistent in X.[11]

The following two propositions formalize the above discussion about choosing γ in each of the two different settings.

[10] We need to modify the definition of $\Sigma_{\Lambda_y,t}$ in Assumption S2.3 to $\Sigma_{\Lambda_y,t} = \frac{1}{N_y}\sum_{i=1}^{N_y} W^Y_{ti}(\Lambda_y)_i(\Lambda_y)_i^\top$, when N_y is finite.

[11] This case is the same as choosing $\gamma = O(N_x/N_y) = O(N_x)$ for the consistency effect, that is, our general rule for selecting the rate of γ still applies.

Proposition 3. *Suppose $\mathbb{E}[(e_y)_{ti}^2]/\mathbb{E}[(e_x)_{ti}^2] = O(1)$. If all the factors can be identified in X and $\gamma = r$ for some positive constant r, then Theorem 1 holds with convergence rate $\delta_{N_x,T} = \min(N_x, T)$, Theorem 2.2 holds with convergence rate $\sqrt{\delta_{N_x,T}}$, and the asymptotic variance of the factors is independent of Y.*

Proposition 4. *Suppose $\mathbb{E}[(e_y)_{ti}^2]/\mathbb{E}[(e_x)_{ti}^2] = O(1/M)$ for $M \to \infty$. If $\gamma = r \cdot N_x$ for some positive constant r, then Theorem 1 holds with convergence rate $\delta_{N_x,T,M} = \min(\delta_{N_x,T}, M)$, and Theorem 2.2 holds with convergence rate $\sqrt{\delta_{N_x,T,M}}$.*

6.3 Multiple Panels

Target-PCA can be generalized to the setting with multiple auxiliary panels $X_1, \cdots, X_{m_x}$ or/and multiple target panels $Y_1, \cdots, Y_{m_y}$, where m_x and m_y are the numbers of auxiliary and target panels.

In the case of multiple auxiliary panels $X_1, \cdots, X_{m_x}$, we can combine all auxiliary panels and the target panel into one panel with weights $\eta_1, \cdots, \eta_{m_x}$ for the auxiliary panels

$$Z^{(\eta_1,\cdots,\eta_{m_x})} = \begin{bmatrix} \eta_1 X_1 & \eta_2 X_2 & \cdots & \eta_{m_x} X_m & Y \end{bmatrix}$$

and apply our proposed estimator in Section 2.4 to $Z^{(\eta_1,\cdots,\eta_{m_x})}$. If $m_x = 1$, then the problem collapses to the setup in the main setting, and selecting the target weight as γ is the same as choosing the source weight as $\eta_1 = 1/\gamma$. Based on our previous discussion, we should choose $\eta_1 = r_1 \cdot N_y/N_{x,1}$ for some positive constant r_1, where $N_{x,1}$ is the number of units in X_1. If there are multiple auxiliary panels, then the natural generalization is to choose $\eta_j = r_j \cdot N_y/N_{x,j}$ for any $j \in \{1, \cdots, m_x\}$ and some positive constant r_j. Such a choice of η_j accounts for the case where different auxiliary panels may have different sets of factors that are useful for identifying the factors in Y (consistency effect), and may have different idiosyncratic noise levels (efficiency effect).

In the case of multiple target panels $Y_1, \cdots, Y_{m_x}$, we can use a sequential approach for estimating the factor model and imputing missing values in each target panel. Concretely, we first apply target-PCA to X and Y_1, and impute the missing observations in Y_1. Second, we treat X and the imputed panel Y_1 as two auxiliary panels, and combine them with Y_2 to estimate the factor model and impute missing values in Y_2. We repeat this procedure until the factor model is estimated and missing values are imputed for all the targets. Essentially, this sequential approach is conceptually the same as target-PCA with only one target panel, but with a more complicated notation.

6.4 Anchored Time Series

So far, all information for the factor model and the data imputation was based on the cross-sectional dependency in the contemporaneous outcomes in Y and X. However, in the case of persistent time series, the prior realizations in Y can provide useful information for imputing missing entries. We propose an extension of target-PCA that takes advantage of prior realizations and the contemporaneous dependency structure. It can be interpreted as anchoring the imputed values around the prediction of a time-series model and correcting them with the innovations around the time-series model estimated from the contemporaneous auxiliary data.

We focus on the practically relevant case of low-frequency data in Y, which is combined with high-frequency supplementary data X. In this case, the contemporaneous low-frequency outcomes in Y are not sufficient to impute the higher frequency missing values. As a concrete example, we refer to our empirical study, where we consider a panel Y of annually observed macroeconomic time series, while X is a panel of monthly observed supplementary data. Many of these macroeconomic time series are highly persistent, and hence their differenced time series fluctuate around their mean value. Hence, the prior realizations of these time series can serve as anchor points.

A simple modification of target-PCA allows us to include prior values as anchor points, while the formal theoretical results continue to hold. Concretely, we replace missing entries in Y by their most recent observed values, and then apply target-PCA as before. For example, for yearly observed Y, we only observe entries in January, while observations from February to December are completely missing. In this case, we construct the anchored time series by filling in the missing entries from February to December with the observations in January of the corresponding year. This is a valid approach, if the common component based on prior factor realizations is an unbiased estimator of the common component in the next period. This is the case under the assumption that the factors have constant means, while errors have zero means. Under this assumption, the consistency result of Theorem 1 and the asymptotic distribution of Theorem 2 continue to hold. In particular, the optimal choice of γ follows the same arguments as for the conventional target-PCA.

Intuitively, the target weight implies a weighted average between the estimate of the common component of Y from the last period and the common component estimated on X from the current period. Without the contemporaneous observations in X or past values of Y, we cannot make any statements about the higher frequency observations in Y. The prior value of the common component of Y can be a noisy estimate of the next period's value, and hence benefits from the cross-sectional contemporaneous information in X to correct the variation and reduce the variance. The weight on the anchored Y panel could be interpreted as a prior for the past observation. If we put no weight on X, we would simply use the prior common component to impute the missing values. In the other extreme case, where we put all weight on X, we only use the factor model in X for imputation, but ignore the prior values. The optimal choice of γ minimizes the variance of the weighted average of these two extreme estimators.

The extension to more complex time-series models is beyond the scope of this paper, but the general logic still applies. The weight γ would imply a weighted average of a time-series model forecast of the common component and the contemporaneous realization of the common component based on auxiliary data.

7 Simulation

In simulations, we show the superior performance of our target-PCA method relative to benchmarks under a variety of settings. For comparison, we include the three natural benchmark methods that apply PCA either to Y, a simple concatenated panel of X and Y, or X and Y separately and combine those factors. These are naive estimation methods for a target panel with auxiliary data. In more detail, we compare the following estimators:

1. **T-PCA**: Target-PCA with optimized γ^* selected as $\gamma = r \cdot N_y/N_x$ with r minimizing $\sum_{t,i} \sum_{C,ti}^{(\gamma)}$.
2. $\mathbf{XP}_Y$: PCA estimator of Xiong and Pelger (2023) applied only to Y (special case of target-PCA with $\gamma = \infty$).
3. $\mathbf{XP}_{Z^{(1)}}$: PCA estimator of Xiong and Pelger (2023) applied to the concatenated panel $Z^{(1)} = [X, Y]$ (special case of target-PCA with $\gamma = 1$).
4. **SE-PCA**: Separately estimate factors from X and Y with the method of Xiong and Pelger (2023), combine the two sets of factors and estimate loadings of the combined factors to impute missing values on Y.[12]

We generate data from a two-factor model $Y_{ti} = F_t^\top (\Lambda_y)_i + (e_y)_{ti}$ and $X_{ti} = F_t^\top (\Lambda_x)_i + (e_x)_{ti}$, where $F_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2), (\Lambda_y)_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2), (\Lambda_x)_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2), (e_x)_{ti} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{e_x}^2)$ and $(e_y)_{ti} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{e_y}^2)$. We consider three missing patterns for target Y:

1. *Missing-at-random:* Entries of Y are missing uniformly at random.
2. *Low-frequency observation:* Entries in Y are observed at a lower frequency and only every second time-series observation is available.
3. *Missingness depends on loadings:* Entries of Y are missing conditional on a unit-specific characteristic $S_i = \mathbb{1}(|(\Lambda_y)_{i2}| > \text{threshold})$. This means that units that are more exposed to the second factor are more likely to be missing in Y. In this case, we assume $(\Lambda_x)_{i1} = 0$, that is, factor 1 is not included in X.

The detailed description of observation patterns and data-generating processes is in Table 2.

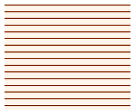
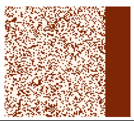
Table 2 compares the performance of the four methods in estimating the common components of the target Y. Specifically, it reports the relative mean squared error (relative MSE) of the estimated common components of the observed, missing, and all entries in Y, which is defined as

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

where $\mathcal{M}$ denotes the set of either observed, missing, or all entries in Y. The MSE for the observed values can be interpreted as an in-sample evaluation, while the MSE for the imputed values serves as an out-of-sample evaluation as it evaluates the model on data that was not used in the estimation.

[12] Note that there is no simple way to determine the number of factors extracted from X and Y respectively. When the factor number is k for other methods, we simply combine k factors from X and k factors from Y in SE-PCA. Note that SE-PCA uses $2k$ factors in total to estimate the common components and impute missing entries in Y. Therefore, SE-PCA is more likely to identify all the factors, but at the cost of an efficiency loss compared to other methods.

Table 2: Relative MSE for different estimators

Observation Pattern	$\mathcal{M}$	T-PCA	XP_Y	$XP_{Z^{(1)}}$	SE-PCA
	obs	**0.184**	0.408	0.224	0.530
	miss	**0.182**	0.414	0.220	0.564
	all	**0.183**	0.411	0.222	0.547
	obs	**0.291**	-	0.846	1.059
	miss	**1.029**	-	1.119	1.104
	all	**0.656**	-	0.979	1.080
	obs	**0.219**	0.238	0.262	0.280
	miss	**0.252**	0.293	0.287	0.356
	all	**0.244**	0.280	0.281	0.338

This table reports the relative MSE of T-PCA (our benchmark method), XP_Y (PCA on Y), $XP_{Z^{(1)}}$ (PCA on concatenated panel) and SE-PCA (separate PCA). The figures on the left show patterns of missing observations with each row representing the observation pattern for a specific time period, and the shaded entries indicating the observed entries. Bold numbers indicate the best relative model performance. We generate a two-factor model and the observation patterns are generated as follows: (a) Missing uniformly at random: $\sigma_{e_x} = 1$, $\sigma_{e_y} = 4$, and entries of Y are missing independently with observation probability $p = \mathbb{P}(W_{ti}^Y = 1) = 0.5$. (b) Low-frequency observation: $\sigma_{e_x} = 16, \sigma_{e_y} = 4$, and entries in Y are only observed every second time period. (c) Missingness depends on loadings: $(\Lambda_x)_{i1} = 0, \sigma_{e_y} = 2, \sigma_{e_x} = 4$ and define a unit-specific characteristic $S_i = \mathbb{1}(|(\Lambda_y)_{i,2}| > 0.1)$. Entries are missing independently with observation probability $p = 0.2$ if $S_i = 1$, and $p = 1$ if $S_i = 0$. We assume $N_x = N_y = T = 200$ and run 200 simulations for each setup.

As shown in Table 2, target-PCA performs well under different observation patterns and dominates the other benchmarks. Our estimator has the smallest relative MSEs as compared to the three benchmark methods, whereas the three benchmark methods are either infeasible or inefficient in different settings. These results hold for the observed and imputed observations for all types of missing patterns.

In the case of missing-at-random, target-PCA has the smallest relative MSEs as compared to the three benchmark methods. This is the setting where all benchmark methods can identify all the factors in Y, but target-PCA is more efficient by appropriately using the information in the auxiliary panel. The gain is particularly large relative to using only Y or estimating the factors separately from X and Y. Both cases have a much smaller effective sample size than that of target-PCA, leading to more than double of the relative MSE.

In the setting of low-frequency observations, target-PCA continues to dominate the benchmark methods. This is a particularly interesting case, as the panel Y is not sufficient for estimating the full factor model, and hence XP_Y is not feasible. As estimating latent factors on Y is not feasible in some periods, SE-PCA degenerates to PCA on X, and therefore the performance of SE-PCA solely depends on the auxiliary panel X. When X has a low signal-to-noise ratio (i.e., σ_{e_x} is large), SE-PCA can perform poorly. XP_Z simultaneously uses the information in both X and Y, and therefore performs the best among the three benchmark methods. Target-PCA further improves upon XP_Z by efficiently weighting the two panels X and Y.

Target-PCA also has the smallest relative MSE when missingness depends on the loadings. This setting can be viewed as endogenously missing data: The second factor has a weak signal on the observed entries of Y, but is important to model the missing data in Y. This setting is similar to, but more complicated than, our toy example in Section 3.1. The auxiliary panel X only contains the second factor, but not the first one; target Y contains both of the two factors, but the second factor is relatively weak on the observed entries of Y because the missing pattern depends on the factor loadings of Y. In this setting, XP_Y performs worse than target-PCA because XP_Y can hardly detect the second factor using the observed entries in Y. $\mathrm{XP}_{Z^{(1)}}$ performs worse than target-PCA mainly because $\mathrm{XP}_{Z^{(1)}}$ does not properly weight the two panels to account for their differences in the idiosyncratic noise levels. SE-PCA also performs worse than target-PCA because each separate estimation is noisier than our combined estimation.

Our results are robust to modifying the parameters of the simulations. The Internet Appendix collects extensive robustness results, where we vary the noise variances and the fraction of observed entries for the models. Target-PCA continues to perform well and to dominate the benchmark methods. We conclude that target-PCA with its more comprehensive use of the target and auxiliary panels is better than the conventional benchmarks under various settings.

8 Empirical Results

8.1 Data

In our empirical study, we show the good performance of target-PCA for imputing missing values in popular macroeconomic panels. Our empirical analysis uses two standard data sets from the Federal Reserve Economic Data (FRED) of the St. Louis Fed.

Our first macroeconomic panel is the FRED-MD macroeconomic database introduced by McCracken and Ng (2016).[13] This dataset consists of 127 monthly macroeconomic variables which are classified into 8 groups: (a) output and income, (b) labor market, (c) housing, (d) consumption, orders, and inventories, (e) money and credit, (f) bond and exchange rates, (g) prices, and (h) stock market. Among them, we use the 120 time series that are fully observed over the time window from 01/1960 to 12/2020. We obtain stationary time series by applying the standard data transformation suggested by McCracken and Ng (2016) to each time series and then normalize them to have zero mean and unit standard deviation. In Section 8.2, we use this data set to evaluate the precision of different imputation methods.

Our second macroeconomic panel contains 58 quarterly-observed macroeconomic time series from two categories, the national income & product accounts category and the flow of funds category of the FRED database. These 58 time series are represented as percentage changes relative to the prior year. Table IA.2 in the Internet Appendix provides a complete description of the data. We normalize the time series to have zero mean and unit standard deviation. In Section 8.3, we combine this low-frequency panel with the higher-frequency panel of monthly-observed FRED-MD data. This

[13]We use the version of December 2021.

is an example where the higher frequency data is not available and our imputation results provide a nowcasting time series of higher frequency.

8.2 Comparison with Benchmark Methods

In this section, we compare the imputation accuracy of target-PCA with the previously considered benchmark methods using the 120 monthly macroeconomic variables from the FRED-MD macroeconomic database. We take the 19 variables in the interest and exchange rates group as our target panel Y, and the remaining 101 variables in the other 7 groups as the auxiliary panel X. Table IA.1 in the Internet Appendix provides a list of the variables in Y. Having a good model to explain and impute interest and exchange rate time series is of economic interest.

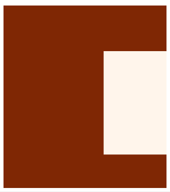
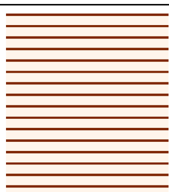
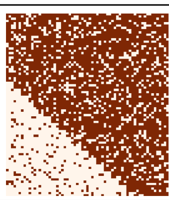
We compare the out-of-sample imputation results for the target panel Y. We mask some entries in Y as missing values with different missing patterns, and compare for various methods the imputation accuracy of the masked entries. Specifically, we consider the following four masking/missing patterns:

1. *Missing-at-random:* The entries of Y are missing uniformly at random.
2. *Block-missing:* A subset of the macroeconomic variables in Y is completely missing during some time periods.
3. *Low-frequency observation:* All variables in Y are observed at a lower than the desired frequency (annually instead of monthly).
4. *Censoring:* The entries of Y are missing if their values exceed a certain threshold.

Table 3 provides a detailed description of the masking mechanism and the percentage of observed values. This table compares the imputation accuracy of our target-PCA estimator with the three benchmark estimators XP_Y (using only Y), $\text{XP}_{Z^{(1)}}$ (naive concatenation), and SE-PCA (separate PCA). Target-PCA selects the optimized γ^* based on our theory as $\gamma = r \cdot N_y/N_x$ with r minimizing $\sum_{t,i} \sum_{C,ti}^{(\gamma)}$. We report the relative MSE of using estimated common components to impute masked entries in Y for various estimators. The relative MSE in this case represents the out-of-sample imputation performance.

Target-PCA method dominates the benchmark methods for all missing patterns. As shown in Table 3, target-PCA achieves the smallest MSEs compared to other methods. In the case of missing-at-random and block-missing patterns, our target-PCA is more efficient since it can appropriately combine the information from the target and auxiliary panel. This is also supported by our theoretical results in Section 5. In the low-frequency observation setting, target-PCA performs better by leveraging the information from the auxiliary panel. For the low-frequency masked data, we use the time series anchored at the most recent observed entries in Y as described in Section 6.4. In more detail, for this specific case, we fill the missing entries in Y (February to December) with the observations in January of the corresponding year before applying target-PCA. Similarly, we also use these anchored time series for the XP_Y and SE-PCA estimators, which otherwise would not be applicable. The anchoring improves the performance for the low-frequency case, but the relative

Table 3: Out-of-sample relative MSE for different methods on FRED-MD

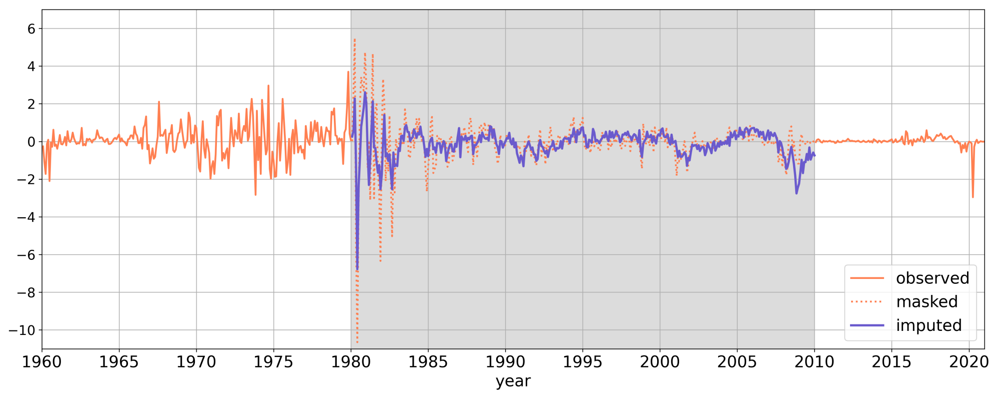
	Observation Pattern (missing ratio)	Factor Number	T-PCA	XP_Y	$XP_{Z^{(1)}}$	SE-PCA
	missing-at-random (40%)	$k=1$	**0.785**	0.789	0.986	0.800
		$k=2$	**0.488**	0.503	0.968	0.500
		$k=3$	**0.485**	0.683	0.926	0.675
		$k=4$	**0.491**	0.813	0.797	0.795
		$k=5$	**0.483**	1.363	0.615	1.355
	block-missing (19%)	$k=1$	**0.958**	1.018	0.971	1.003
		$k=2$	**0.710**	0.805	0.961	0.852
		$k=3$	**0.713**	0.796	0.974	0.803
		$k=4$	**0.778**	0.783	0.974	0.781
		$k=5$	**0.792**	2.601	0.935	2.584
	low-frequency (92%)	$k=1$	**0.942**	0.949	1.019	1.009
		$k=2$	**0.927**	1.140	0.931	1.149
		$k=3$	**0.926**	1.213	0.936	1.223
		$k=4$	**0.910**	1.212	1.095	1.234
		$k=5$	**1.017**	1.251	1.092	1.280
	Censoring (40%)	$k=1$	**0.927**	-	0.996	0.995
		$k=2$	**0.881**	-	0.996	0.994
		$k=3$	**0.892**	-	0.993	0.992
		$k=4$	**0.885**	-	0.990	0.987
		$k=5$	**0.869**	-	0.984	0.981

This table reports the relative out-of-sample MSEs of the target panel Y on the FRED-MD macroeconomic panel for three benchmark methods with different numbers of latent factors. The target panel Y contains the 19 variables in the interest and exchange rates group, and the remaining 101 variables from X from 01/1960 to 12/2020. The masking/ missing patterns in this table are generated as follows. (a) Missing uniformly at random: We randomly mask each entry in Y with probability 0.4. We repeat this random masking 100 times and report the average relative MSE. (b) Block-missing: we mask the period from 01/1980 to 12/2009 for the 7 time series in Y related to bond prices: TB3MS, TB6MS, GS1, GS5, GS10, AAA, and BAA. (c) Low-frequency observation: We mask the observations of Y from February to December each year. Only for this case, the input target data for the latent factor model estimation is anchored at the most recent observed value as described in Section 6.4. This means we augment Y with the anchored time series that use the observation in January of the corresponding year for February to December for estimating the various latent factor models. (d) Censoring: We mask entries whose absolute value exceeds a threshold in Y. We set the threshold to 0.6, in order to mask approximately 40% of the entries. The plots on the left column of this table illustrate these four missing patterns. The entries with dark color denote observed entries, and the entries with light color indicate that entries are missing.

qualitative results are the same without anchoring. Our target-PCA estimator is also the best for the case of censoring, which could cause weak signals in the observed entries. In this case, XP_Y is not applicable since there are time periods where all units have large absolute values and thus are missing. For this censored masking, the missing pattern can depend on the noise and factor realization, and hence violate our assumptions. Nevertheless, our method still performs very well and better than the benchmark methods.

Our target-PCA estimator is robust to the number of latent factors. Table 3 compares the

Figure 3: Imputed time series of 6-month treasury yield for block-missing pattern

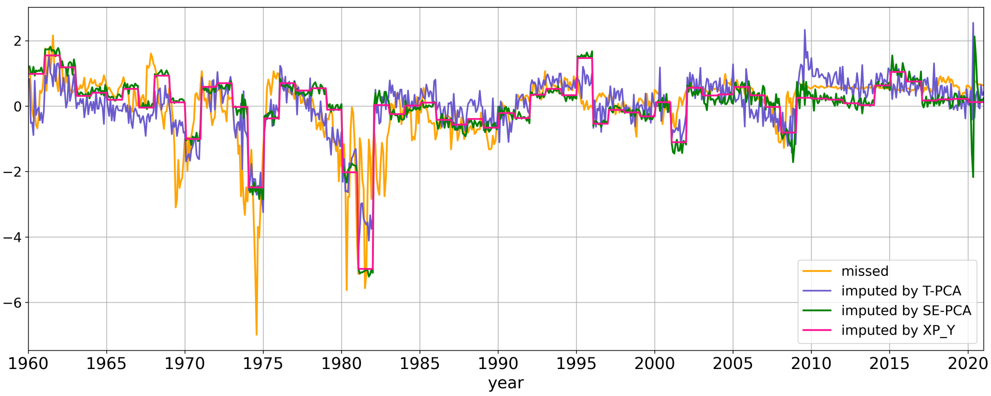


This figure shows the time series of the 6-month treasury yield for the block-missing pattern. The orange line shows the actual time series and the true masked values. The purple line denotes the imputed values with target-PCA with $k = 3$ latent factors. The gray block indicates the missing period for the out-of-sample imputation.

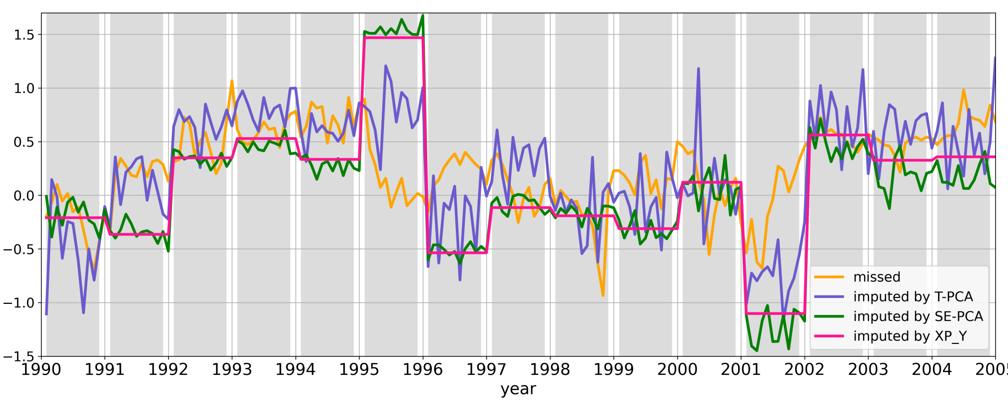
out-of-sample results for different numbers of factors. First, target-PCA dominates the benchmarks uniformly in the number of factors, that is, for the same number of factors it always performs better. Second, the results of target-PCA are very robust to the choice of k. Target-PCA with two factors is close to optimal and outperforms all other estimators even if they use more factors. The benchmark methods are not stable in the number of factors. For a larger number of factors ($k = 5$), XP_Y and SE-PCA can perform substantially worse than the other two methods, especially for the missing-at-random and block-missing patterns. A plausible reason is that both XP_Y and SE-PCA estimate the factors from regressing the observed Y on estimated loadings, which can be very noisy for higher-order factors in Y which are weaker. Both target-PCA and $\text{XP}_{Z^{(1)}}$ do not have this problem as factors are estimated from the regression on the combined panel of X and Y. Target-PCA further improves upon $\text{XP}_{Z^{(1)}}$ by choosing an appropriate γ for the consistency and efficiency effects.

Target-PCA results in a meaningful time series of imputed values. Figures 3 and 4 illustrate the precise imputation of target-PCA for representative examples. Figure 3 plots the imputed time series with target-PCA and the actual values of the 6-month treasury bill for the block-missing pattern. Our target-PCA imputation is quite accurate for the masked/missing time period as it captures well the fluctuation in the real time series. Figure 4 plots the imputed time series with target-PCA and the actual time series of the spread between the 3-month treasury and Fed Funds rate for the low-frequency observation pattern. As a reference, we also include the time series imputed by the benchmark methods XP_Y and SE-PCA. The lower panel zooms in to show the time period between 1990 and 2005. Target-PCA imputes the real time series reasonably well and visibly better than the benchmarks. In particular, a proper weight on X seems to be essential to capture the contemporaneous fluctuations in the missing blocks.

Figure 4: Imputed time series of rate spread for low-frequency missing pattern



(a) Full time period (01/1960 - 12/2020)



(b) Zoom-in (01/1990 - 12/2005)

This figure shows the time series of the spread between the 3-month treasury and Fed Funds rate for the low-frequency missing pattern. The orange line shows the actual time series and the true masked values. Observations are only observed in January and missing from February to December each year. The gray blocks in the zoom-in plot indicate the missing periods for the out-of-sample imputation with different methods. The number of factors equals $k = 4$ for each imputation method.

8.3 Nowcasting with Target-PCA

An important practical problem is nowcasting macroeconomic panel data. We use target-PCA to impute unbalanced low-frequency macroeconomic panel data that is not available at a high frequency. The imputed high-frequency data represents a nowcasted panel, that is of interest to itself and also for downstream applications that require higher frequency data.[14]

In this section, we illustrate the time series of imputed nowcast data that are not available at

[14] A widely used approach for nowcasting is based on state-space models. Those models are complementary to our approach. State-space models are usually dynamic models for low dimensions that impose distributional assumptions. In contrast, our method is designed for large panels and focuses on leveraging cross-sectional information. For example, when a panel is of mixed frequency, our method can take advantage of high-frequency observations to "nowcast" the low-frequency variables. An interesting direction for future work is to extend our framework to dynamic factor models that optimally use auxiliary data.

a higher frequency. We impute the monthly values of only quarterly-observed time series of the FRED database. Our target panel consists of 58 macroeconomic time series of the national income and product accounts category and the flow of funds category. It includes several fundamental macroeconomic time series needed in economic research. To impute the monthly values of our target, we construct our auxiliary data by the 120 monthly macroeconomic indicators of FRED-MD from the first empirical study.

Target-PCA imputation performs well for the observed quarterly time periods and results in economically meaningful imputed values. Figure 5 illustrates the imputation with target-PCA with the representative Gross Domestic Product (GDP) time series. We only observe quarterly-frequency GDP data, which is denoted by the orange "+". The monthly observed auxiliary panel X, allows us to use target-PCA to impute the monthly values of the GDP time series, which is plotted as the purple curve in Figure 5.[15] Target-PCA captures the unknown and unobserved variation between two quarterly GDP observations using the monthly observed auxiliary data. Figure A.1 in the Appendix collects further examples, which all demonstrate the good performance of target-PCA for meaningful imputed values between the quarters. Our results use $k = 5$ latent factors. As discussed in Section 8.2, the target-PCA estimator is robust to the number of latent factors, and the results are very similar for different numbers of factors.[16]

9 Conclusion

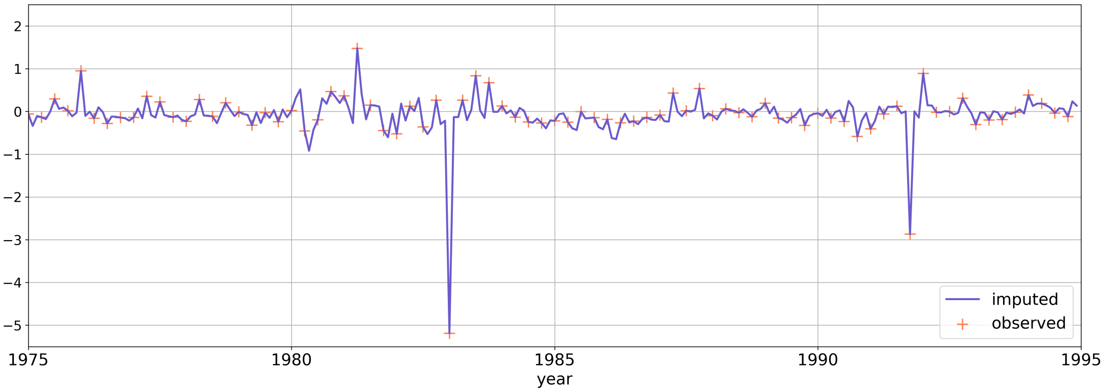
This paper proposes our novel method target-PCA to estimate a latent factor model for target data by optimally weighting and combining the information from auxiliary data that is relevant to our target. This method is broadly applicable but easy to implement: It applies principal component analysis to a weighted average of the covariance matrices of the target and auxiliary panels. Target-PCA is particularly beneficial when the target data have missing observations, which could handle some of the scenarios that conventional methods cannot solve. A leading example is the mixed-frequency observation pattern, where conventional PCA cannot estimate the factors in times without any observations.

To optimally combine the auxiliary data, we need to overcome the differences in dimensionality and noise ratio between the two panels. Target-PCA tackles these problems by introducing a target weight γ and combining information from the auxiliary data with the weighted target data. We show two essential effects of this target weight for target-PCA: the consistency and efficiency effects.

[15]Our imputation can be used for economic stock and flow variables. In the case of a flow variable, the sum of monthly observations has to add up to lower frequencies like quarterly and annual values. For example, for GDP growth, the quarterly growth is the sum of the monthly growths within the quarter. Our imputed monthly values represent an appropriate moving average of the GDP growth; that is, the imputed values represent the growth over the previous three months. Obviously, the imputed values can always be transformed into monthly GDP growth. In summary, in our imputation process, we do not need to distinguish between flow and stock variables, as we are not directly imputing the monthly GDP growth. In practice, it is important to consider the interpretation of imputed values.

[16]The factor model is estimated only once on the unbalanced panel. Applying target-PCA on an expanding window to avoid using future data gives essentially identical results. The results are available upon request.

Figure 5: Quarterly observed GDP vs. monthly imputed GDP with target-PCA



This figure shows the time series of observed GDP and monthly imputed GDP with target-PCA. The orange "+" denotes the quarterly observed values of the GDP time series, and the purple curve denotes the imputed values of the GDP time series with target-PCA for $k = 5$ latent factors. We only impute for the months when the observation is missing and take the actual values when the data is observed. The time series represent percentage changes relative to the prior year.

First, by selecting the target weight at the right rate, we can ensure the consistent estimation of all factors in the target panel, including weak factors. Second, by selecting the scaling of the target weight to account for the noise ratios between the panels, we improve the efficiency of the estimated factor model.

We develop the inferential theory for the estimated factors, loadings, and imputed values of target-PCA under very general assumptions on the approximate factor model and missing patterns. The asymptotic results are used to construct confidence intervals and provide guidance on choosing the optimal target weight for target-PCA. In an empirical analysis, we illustrate the benefit of our approach with the imputation of unbalanced macroeconomic panel data.

References

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

Andreou, E., E. Ghysels, and A. Kourtellos (2010): "Regression models with mixed sampling frequencies," *Journal of Econometrics*, 158(2), 246–261.

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

——— (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77(4), 1229–1279.

Bai, J., E. Ghysels, and J. H. Wright (2013): "State space models and MIDAS regressions," *Econometric Reviews*, 32(7), 779–813.

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

——— (2021): "Matrix completion, counterfactuals, and factor analysis of missing data," *Journal of the American Statistical Association*, 116(536), 1746–1763.

——— (2023): "Approximate factor models with weaker loadings," *Journal of Econometrics*, 235(2), 1893–1916.

Boivin, J., and S. Ng (2006): "Are more data always better for factor analysis?," *Journal of Econometrics*, 132(1), 169–194.

Breitung, J., and J. Tenhofen (2011): "GLS estimation of dynamic factor models," *Journal of the American Statistical Association*, 106(495), 1150–1166.

Cahan, E., J. Bai, and S. Ng (2023): "Factor-based imputation of missing values and covariances in panel data of large dimensions," *Journal of Econometrics*, 233(1), 113–131.

Candès, E. J., and B. Recht (2009): "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 9(6), 717–772.

Chen, Q. (2022): "A unified framework for estimation of high-dimensional conditional factor models," *arXiv preprint arXiv:2209.00391*.

Chen, Q., N. Roussanov, and X. Wang (2021): "Semiparametric conditional factor models: Estimation and inference," *arXiv preprint arXiv:2112.07121*.

Chen, Y., J. Fan, C. Ma, and Y. Yan (2019): "Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences*, 116(46), 22931–22937.

Choi, I. (2012): "Efficient estimation of factor models," *Econometric Theory*, 28(2), 274–308.

Chow, G. C., and A.-l. Lin (1971): "Best linear unbiased interpolation, distribution, and extrapolation of time series by related series," *The Review of Economics and Statistics*, 53(4), 372–375.

Fan, J., Y. Liao, and M. Mincheva (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4), 603–680.

Fan, J., Y. Liao, and W. Wang (2016): "Projected principal component analysis in factor models," *Annals of Statistics*, 44(1), 219–254.

Ghysels, E., V. Kvedaras, and V. Zemlys-Balevičius (2020): "Mixed data sampling (MIDAS) regression models," *Handbook of Statistics*, 42(4), 117–153.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2004): "The MIDAS touch: Mixed data sampling regression models," Cirano working papers, CIRANO.

Huang, D., F. Jiang, K. Li, G. Tong, and G. Zhou (2022): "Scaled PCA: A new approach to dimension reduction," *Management Science*, 68(3), 1678–1695.

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

Lettau, M., and M. Pelger (2020): "Estimating latent asset-pricing factors," *Journal of Econometrics*, 218(1), 1–31.

McCracken, M. W., and S. Ng (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34(4), 574–589.

Ng, S., and S. Scanlan (2023): "Constructing high frequency economic indicators by imputation," *arXiv preprint arXiv:2303.01863*.

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

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

——— (2022): "Uniform asymptotics for weak and strong factors," *Working paper*.

Pan, S. J., and Q. Yang (2009): "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.

Pelger, M. (2019): "Large-dimensional factor modeling based on high-frequency observations," *Journal of Econometrics*, 208(1), 23–42.

Pelger, M., and R. Xiong (2021a): "Interpretable sparse proximate factors for large dimensions," *Journal of Business & Economic Statistics*, 40(4), 1642–1664.

——— (2021b): "State-varying factor models of large dimensions," *Journal of Business & Economic Statistics*, 40(3), 1315–1333.

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

Stock, J. H., and M. W. Watson (1998): "Diffusion indexes," *NBER Working paper 6702*.

——— (2016): "Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics," vol. 2 of *Handbook of Macroeconomics*, chap. 8, pp. 415–525. Elsevier.

Su, L., and X. Wang (2017): "On time-varying factor models: Estimation and testing," *Journal of Econometrics*, 198(1), 84–101.

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

A General Assumptions

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

Assumption G3 (Factor model)**.**

1. *Factors:* $\forall t$, $\mathbb{E}\|F_t\|^4 \leq C$. *There exists a positive definite* $k \times k$ *matrix* Σ_F *such that* $\frac{1}{T}\sum_{t=1}^T F_t F_t^\top \xrightarrow{p} \Sigma_F$ *and* $\mathbb{E}\left\|\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^T F_t F_t^\top - \Sigma_F\right)\right\|^2 \leq C$. *Furthermore, for any* Q_{ij}^Y, $\frac{1}{|Q_{ij}^Z|}\sum_{t\in Q_{ij}^Z} F_t F_t^\top \xrightarrow{p} \Sigma_F$ *and* $\mathbb{E}\left\|\sqrt{|Q_{ij}^Z|}\left(\frac{1}{|Q_{ij}^Z|}\sum_{t\in Q_{ij}^Z} F_t F_t^\top - \Sigma_F\right)\right\|^2 \leq C$.
2. *Loadings: loadings are independent of factors and errors, and* Λ_x *is independent with* Λ_y. $\mathbb{E}\|(\Lambda_x)_i\|^4 \leq C$, $\mathbb{E}\|(\Lambda_y)_i\|^4 \leq C$, $\frac{1}{N_x}\sum_{i=1}^{N_x}(\Lambda_x)_i(\Lambda_x)_i^\top \xrightarrow{p} \Sigma_{\Lambda_x}$, $\frac{1}{N_y}\sum_{i=1}^{N_y}(\Lambda_y)_i(\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y}$, *and* $\Sigma_{\Lambda_x}+\Sigma_{\Lambda_y}$ *is a positive definite matrix. Furthermore, for any* t, $\frac{1}{N_y}\sum_{i=1}^{N_y} W_{ti}^Y(\Lambda_y)_i(\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y,t}$, *and* $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y,t}$ *is a positive definite matrix.*
3. *Idiosyncratic errors:*

   (a) $\mathbb{E}(e_x)_{ti} = \mathbb{E}(e_y)_{ti} = 0, \mathbb{E}(e_x)_{ti}^8 \leq C$ *and* $\mathbb{E}(e_y)_{ti}^8 \leq C$.

(b) $\mathbb{E}\left[(e_x)_{ti}(e_x)_{si}\right]=\gamma_{st,i}^{(e_x)}$ *with* $|\gamma_{st,i}^{(e_x)}|\leq\gamma_{st}$ *and* $\mathbb{E}\left[(e_y)_{ti}(e_y)_{si}\right]=\gamma_{st,i}^{(e_y)}$ *with* $|\gamma_{st,i}^{(e_y)}|\leq\gamma_{st}$ *for some* γ_{st} *and all* i, $\sum_{s=1}^{T}\gamma_{st}\leq C$ *for all* t. *Furthermore,* $\gamma_{tt,i}^{(e_x)}/\gamma_{tt,i}^{(e_y)}$ *is bounded away from zero for all* t,i.

(c) $\mathbb{E}\left[(e_x)_{ti}(e_x)_{tj}\right]=\tau_{ij,t}^{(e_x)}$ *with* $|\tau_{ij,t}^{(e_x)}|\leq\tau_{ij}^{(e_x)}$ *for some* $\tau_{ij}^{(e_x)}$ *and all* t, $\sum_{j=1}^{N_x}\tau_{ij}^{(e_x)}\leq C$ *for all* i; $\mathbb{E}\left[(e_y)_{ti}(e_y)_{tj}\right]=\tau_{ij,t}^{(e_y)}$ *with* $|\tau_{ij,t}^{(e_y)}|\leq\tau_{ij}^{(e_y)}$ *for some* $\tau_{ij}^{(e_y)}$ *and all* t, $\sum_{j=1}^{N_y}\tau_{ij}^{(e_y)}\leq C$ *for all* i; $\mathbb{E}\left[(e_y)_{ti}(e_x)_{tj}\right]=\tau_{ij,t}^{(e_y,e_x)}$ *with* $|\tau_{ij,t}^{(e_y,e_x)}|\leq\tau_{ij}^{(e_y,e_x)}$ *for some* $\tau_{ij}^{(e_y,e_x)}$ *and all* t, $\sum_{i=1}^{N_y}\tau_{ij}^{(e_y,e_x)}\leq C$ *for all* j, *and* $\sum_{j=1}^{N_x}\tau_{ij}^{(e_y,e_x)}\leq C$ *for all* i.

(d) $\mathbb{E}\left[(e_x)_{ti}(e_x)_{sj}\right]=\tau_{ij,ts}^{(e_x)}$, *and* $\sum_{j=1}^{N_x}\sum_{s=1}^{T}|\tau_{ij,ts}^{(e_x)}|\leq C$ *for all* i *and* t; $\mathbb{E}\left[(e_y)_{ti}(e_y)_{sj}\right]=\tau_{ij,ts}^{(e_y)}$, *and* $\sum_{j=1}^{N_y}\sum_{s=1}^{T}|\tau_{ij,ts}^{(e_y)}|\leq C$ *for all* i *and* t; $\mathbb{E}\left[(e_x)_{ti}(e_y)_{sj}\right]=\tau_{ij,ts}^{(e_y,e_x)}$, *for all* i *and* t, $\sum_{i=1}^{N_x}\sum_{s=1}^{T}|\tau_{ij,ts}^{(e_y,e_x)}|\leq C$.

(e) $\mathbb{E}\Big[\frac{1}{|Q_{ij}^Y|^{1/2}}\sum_{t\in Q_{ij}^Y}\left((e_y)_{ti}(e_y)_{tj}-\mathbb{E}\left[(e_y)_{ti}(e_y)_{tj}\right]\right)\Big]^4\leq C$, $\mathbb{E}\Big[\frac{1}{T^{1/2}}\sum_{t=1}^{T}\big((e_x)_{ti}(e_x)_{tj}-\mathbb{E}[(e_x)_{ti}(e_x)_{tj}]\big)\Big]^4\leq C$ *and* $\mathbb{E}\Big[\frac{1}{|Q_{jj}^Y|^{1/2}}\sum_{t\in Q_{jj}^Y}\left((e_x)_{ti}(e_y)_{tj}-\mathbb{E}\left[(e_x)_{ti}(e_y)_{tj}\right]\right)\Big]^4\leq C$.

4. *Weak dependence between factor and idiosyncratic errors:* $\mathbb{E}\left\|\frac{1}{\sqrt{|Q_{ij}^Y|}}\sum_{t\in Q_{ij}^Y}F_t(e_y)_{tj}\right\|^2\leq C$ *for any* $i,j=1,\cdots,N_y$. $\mathbb{E}\left\|\frac{1}{\sqrt{|Q_{ii}^Y|}}\sum_{t\in Q_{ii}^Y}F_t(e_x)_{tj'}\right\|^2\leq C$ *and* $\mathbb{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t(e_x)_{tj'}\right\|^2\leq C$ *for any* $j'=1,\cdots,N_x$.

Assumption G4 (Moment conditions and central limit theorems)**.**

1. *For any* $i=1,\cdots,N_y$, *there is* $\mathbb{E}\left\|\sqrt{\frac{T}{N_x}}\sum_{j=1}^{N_x}(\Lambda_x)_j\frac{1}{|Q_{ii}^Y|}\sum_{s\in Q_{ii}^Y}F_s^\top(e_x)_{sj}\right\|^2\leq C$, *and* $\mathbb{E}\left\|\sqrt{\frac{T}{N_y}}\sum_{j=1}^{N_y}(\Lambda_y)_j\frac{1}{|Q_{ij}^Y|}\sum_{s\in Q_{ij}^Y}F_s^\top(e_y)_{sj}\right\|^2\leq C$.
2. $\mathbb{E}\left\|\frac{\sqrt{N_xT}}{N_x^2}\sum_{i,j=1}^{N_x}(\Lambda_x)_j(\Lambda_x)_j^\top\frac{1}{T}\sum_{s=1}^{T}F_s(\Lambda_x)_i^\top(e_x)_{si}\right\|^2\leq C$,
$\mathbb{E}\left\|\frac{\sqrt{N_yT}}{N_y^2}\sum_{i,j=1}^{N_y}(\Lambda_y)_j(\Lambda_y)_j^\top\frac{1}{|Q_{ij}^Y|}\sum_{s\in Q_{ij}^Y}F_s(\Lambda_y)_i^\top(e_y)_{si}\right\|^2\leq C$,
$\mathbb{E}\left\|\frac{\sqrt{N_xT}}{N_yN_x}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}(\Lambda_y)_j(\Lambda_y)_j^\top\frac{1}{|Q_{jj}^Y|}\sum_{s\in Q_{jj}^Y}F_s(\Lambda_x)_i^\top(e_x)_{si}\right\|^2\leq C$,
$\mathbb{E}\left\|\frac{\sqrt{N_yT}}{N_yN_x}\sum_{i=1}^{N_y}\sum_{j=1}^{N_x}(\Lambda_x)_j(\Lambda_x)_j^\top\frac{1}{|Q_{ii}^Y|}\sum_{s\in Q_{ii}^Y}F_s(\Lambda_y)_i^\top(e_y)_{si}\right\|^2\leq C$.
3. *For any* $i=1,\cdots,N_y$, $\mathbb{E}\left\|\sqrt{\frac{T}{N_x}}\sum_{j=1}^{N_x}(\Lambda_x)_j\frac{1}{|Q_{ii}^Y|}\sum_{s\in Q_{ii}^Y}\left((e_x)_{sj}(e_y)_{si}-\mathbb{E}\left[(e_x)_{sj}(e_y)_{si}\right]\right)\right\|^2\leq C$, *and* $\mathbb{E}\left\|\sqrt{\frac{T}{N_y}}\sum_{j=1}^{N_y}(\Lambda_y)_j\frac{1}{|Q_{ij}^Y|}\sum_{s\in Q_{ij}^Y}\left((e_y)_{si}(e_y)_{sj}-\mathbb{E}\left[(e_y)_{si}(e_y)_{sj}\right]\right)\right\|^2\leq C$.
4. *For any* $t=1,\cdots,T$, $\mathbb{E}\left\|\sqrt{\frac{T}{N_x^3}}\sum_{i,j=1}^{N_x}\frac{1}{T}\sum_{s=1}^{T}\phi_{ij,st}\left((e_x)_{si}(e_x)_{sj}-\mathbb{E}\left[(e_x)_{si}(e_x)_{sj}\right]\right)\right\|^2\leq C$ *with* $\phi_{ij,st}=(\Lambda_x)_j(\Lambda_x)_i^\top$, $(\Lambda_x)_j(\Lambda_x)_i^\top F_s$, $(\Lambda_x)_j(e_x)_{ti}$, $\mathbb{E}\left\|\sqrt{\frac{T}{N_y^3}}\sum_{i,j=1}^{N_y}\frac{1}{|Q_{ij}^Y|}\sum_{s\in Q_{ij}^Y}\phi_{ij,st}((e_y)_{si}(e_y)_{sj}-\mathbb{E}\left[(e_y)_{si}(e_y)_{sj}\right])\right\|^2\leq C$ *with* $\phi_{ij,st}=(\Lambda_y)_j(\Lambda_y)_i^\top$, $W_{ti}^Y(\Lambda_y)_j(\Lambda_y)_i^\top F_s$, $W_{ti}^Y(\Lambda_y)_j(e_y)_{ti}$, *and moreover,* $\mathbb{E}\left\|\sqrt{\frac{T}{N_x^2N_y}}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\frac{1}{|Q_{jj}^Y|}\sum_{s\in Q_{jj}^Y}\phi_{ij,st}\left((e_x)_{si}(e_y)_{sj}-\mathbb{E}\left[(e_x)_{si}(e_y)_{sj}\right]\right)\right\|^2\leq C$ *with* $\phi_{ij,st}=(\Lambda_y)_j(\Lambda_x)_i^\top$, $(\Lambda_y)_j(\Lambda_x)_i^\top F_s$, $W_{tj}^Y(\Lambda_x)_i(\Lambda_y)_j^\top F_s$, $(\Lambda_y)_j(e_x)_{ti}$, $W_{tj}^Y(\Lambda_x)_i(e_y)_{tj}$.
5. *For any* t *and* $j=1,\cdots,N_y$, $\mathbb{E}\left\|\sqrt{\frac{T}{N_x}}\sum_{i=1}^{N_x}\left(\frac{1}{|Q_{jj}^Y|}\sum_{s\in Q_{jj}^Y}F_sF_s^\top-\frac{1}{T}\sum_{s=1}^{T}F_sF_s^\top\right)(\Lambda_x)_i(e_x)_{ti}\right\|^4\leq$

C, and $\mathbb{E}\left\|\sqrt{\frac{T}{N_y}}\sum_{i=1}^{N_y}\left(\frac{1}{|(Q_{ij}^*|}\sum_{s\in Q_{ij}^*}F_sF_s^\top-\frac{1}{T}\sum_{s=1}^{T}F_sF_s^\top\right)W_{ti}^Y(\Lambda_y)_i(e_y)_{ti}\right\|^4\leq C$ with $Q_{ij}^*=Q_{ii}^Y$ or Q_{ij}^Y.

6. $\frac{N_x}{N_x+N_y}\left(\frac{\sqrt{T}}{N_x}\sum_{j=1}^{N_x}(\Lambda_x)_j(\Lambda_x)_j^\top\frac{1}{|Q_{ii}^Y|}\sum_{t\in Q_{ii}^Y}F_t(e_y)_{ti}+\gamma\cdot\frac{\sqrt{T}}{N_x}\sum_{j=1}^{N_y}(\Lambda_y)_j(\Lambda_y)_j^\top\frac{1}{|Q_{ij}^Y|}\sum_{t\in Q_{ij}^Y}F_t(e_y)_{ti}\right)$ $\xrightarrow{d}\mathcal{N}(0,\Gamma_{\Lambda_y,i}^{(\gamma),\text{obs}})$ for any $\gamma=r\cdot N_x/N_y$ and $i=1,\cdots,N_y$.

7. For any t, $\frac{1}{\sqrt{N_x}}\sum_{i=1}^{N_x}(\Lambda_x)_i(e_x)_{ti}\xrightarrow{d}\mathcal{N}(0,\Sigma_{\Lambda_xe_x,t})$, $\frac{1}{\sqrt{N_y}}\sum_{i=1}^{N_y}W_{ti}^Y(\Lambda_y)_i(e_y)_{ti}\xrightarrow{d}\mathcal{N}(0,\Sigma_{\Lambda_ye_y,t})$. For any $\gamma=r\cdot N_x/N_y$, $\Gamma_{F,t}^{(\gamma),\text{obs}}:=\lim\frac{N_xN_y}{(N_x+N_y)^2}\Sigma_{\Lambda_xe_x,t}+(\gamma\frac{N_y}{N_x+N_y})^2\Sigma_{\Lambda_ye_y,t}$. If there is some weak factor F_w in Y whose loading $\sum_{i=1}^{N_y}(\Lambda_y)_{i,w}^2$ grows at the rate $g(N_y)$, then let $N_w=\min(N_y^2/g(N_y),N_x)$. For F_w, $\frac{1}{\sqrt{g(N_y)}}\sum_{i=1}^{N_y}W_{ti}^Y(\Lambda_y)_{i,w}(e_y)_{ti}\xrightarrow{d}\mathcal{N}(0,\Sigma_{\Lambda_{y,w}e_y,t})$, and for any $\gamma=r\cdot N_x/N_y$, $\Gamma_{F_w,t}^{(\gamma),\text{obs}}:=\lim\frac{N_xN_w}{(N_x+N_y)^2}\Sigma_{\Lambda_xe_x,t}+\gamma^2\frac{g(N_y)N_w}{(N_x+N_y)^2}\Sigma_{\Lambda_{y,w}e_y,t}$.

8. Define the filtration $\mathcal{G}^t=\sigma(\cup_{s=1}^T\mathcal{G}_s^t)$ with $\mathcal{G}_s^t=\sigma(\{W_{ij}^Y,i\leq s,\text{all }j\},\Lambda_y,v_t^{(\gamma)})$ generated by $\{W_{ij}^Y,i\leq s,\text{all }j\},\Lambda_y$ and $v_t^{(\gamma)}=(\Sigma_\Lambda^{(\gamma)})^{-1}\Sigma_F^{-1}F_t$. For every i, t, and $u_i=(\Lambda_y)_i$, it holds

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

When $i=1,\cdots,N_x$, $X_i^{(\gamma)}=\gamma\cdot\frac{1}{N_x+N_y}\sum_{j=1}^{N_y}(\Lambda_y)_j(\Lambda_y)_j^\top\left(\frac{1}{|Q_{jj}^Y|}\sum_{t\in Q_{jj}^Y}F_tF_t^\top-\frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top\right)$, and when $i=N_x+1,\cdots,N_x+N_y$, $X_i^{(\gamma)}=\frac{1}{N_x+N_y}\sum_{j=1}^{N_x}(\Lambda_x)_j(\Lambda_x)_j^\top\left(\frac{1}{|Q_{i'i'}^Y|}\sum_{t\in Q_{i'i'}^Y}F_tF_t^\top-\frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top\right)+\gamma\cdot\frac{1}{N_x+N_y}\sum_{j=1}^{N_y}(\Lambda_y)_j(\Lambda_y)_j^\top\left(\frac{1}{|Q_{i'j}^Y|}\sum_{t\in Q_{i'j}^Y}F_tF_t^\top-\frac{1}{T}\sum_{t=1}^{T}F_tF_t^\top\right)$ with $i'=i-N_x$. For any t, $\mathbf{X}_t^{(\gamma)}=\frac{1}{N_x+N_y}\sum_{i=1}^{N_x}X_i^{(\gamma)}(\Lambda_x)_i(\Lambda_x)_i^\top+\gamma\cdot\frac{1}{N_x+N_y}\sum_{i=1}^{N_y}X_{i+N_x}^{(\gamma)}W_{ti}^Y(\Lambda_y)_i(\Lambda_y)_i^\top$.

The assumptions on the factor structure in Y are on the same level of generality as in Xiong and Pelger (2023) with the generalization that factors in Y can be weak. The approximate factor model in X is also on a similar level of generality as in Bai (2003) with the generalization that not all factors in Y are included in X. The assumptions essentially assume that for a properly selected γ the combined weighted panel $Z^{(\gamma)}$ satisfies the assumptions in Xiong and Pelger (2023).

In more detail, Assumption G3 describes an approximate factor structure and is at a similar level of generality as Bai (2003): (1) Assumption G3.1 states that each factor has a nontrivial variation for the observed time periods. (2) We assume loadings are random but independent of factors and errors in Assumption G3.2. This is a standard assumption. Assumption G3.2 ensures that the loadings are systematic for both the full observed weighted matrix $Z^{(\gamma)}$ and the partially observed $\tilde{Z}^{(\gamma)}$ with our proposed target weight $\gamma=r\cdot N_x/N_y$. This assumption is needed to show the consistency of loadings and factors. This assumption deviates from the usual factor model assumptions as neither Σ_{Λ_x} nor Σ_{Λ_y} has to be full rank. (3) Assumption G3.3 is a standard weak dependency assumption in the noise and allows the errors to be time-series and cross-sectionally weakly correlated. (4) Assumption G3.4 allows the factors and idiosyncratic errors to be weakly correlated.

Assumption G4 is not necessary to show the consistency of loadings and factors but is only used to show the asymptotic normality of the estimators. The assumptions are closely related to the moment and CLT assumptions in Bai (2003). Assumptions G4.1-G4.5 bound the second moments of certain averages. Assumption G4.6 and G4.7 state the necessary central limit theorems. Assumption G4.8 is specific to the missing value problem and introduces the correction terms that appear in the asymptotic distribution. These terms emerge because our estimator averages over different numbers of observations for different entries in the covariance matrix. Assumption G4.8 assumes a central limit theorem for $X_i^{(\gamma)}$ and $\mathbf{X}_t^{(\gamma)}$. The conventional CLT of the form

$$\sqrt{T}\begin{bmatrix}\mathrm{vec}(X_i^{(\gamma)})\\ \mathrm{vec}(\mathbf{X}_t^{(\gamma)})\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix}\Psi_i & (\Psi_{i,t}^{\mathrm{cov}})^\top \\ \Psi_{i,t}^{\mathrm{cov}} & \Psi_t\end{bmatrix}\right)$$

would not be sufficient as $X_i^{(\gamma)}$ and $\mathbf{X}_t^{(\gamma)}$ are multiplied with the random variables u_i and $v_t^{(\gamma)}$ in $\sqrt{T}\begin{bmatrix}(X_i^{(\gamma)}u_i)^\top & (\mathbf{X}_t^{(\gamma)}v_t^{(\gamma)})^\top\end{bmatrix}$. Hence, the asymptotic variances of these products are quadratic functions in the elements of those random variables given by $h_i^{(\gamma)}(u_i)$ and $g_t^{(\gamma)}(v_t^{(\gamma)})$ and take the form of $h_i^{(\gamma)}(u_i) = (u_i^\top \otimes I_k)\Psi_i(u_i \otimes I_k)$ and $g_t^{(\gamma)}(v_t^{(\gamma)}) = (v_t^{(\gamma)\top} \otimes I_k)\Psi_t(v_t^{(\gamma)} \otimes I_k)$.

In Assumption G4.8 we require a central limit theorem for stable convergence in law which is stronger than the conventional central limit theorem for convergence in distribution. This is because the asymptotic variance in Assumption G4.8 depends on $(\Lambda_y)_i$, $(\Lambda_x)_i$ and F_t, which are random variables. Thus, 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 $(\Lambda_y)_i$, $(\Lambda_x)_i$, and F_t. The stable convergence in law result implies that the estimated factors and common components normalized by their random standard deviation converge to a standard normal distribution. More specifically, Assumption G4.8 implies that $X_i^{(\gamma)}$ and $\mathbf{X}_t^{(\gamma)}$ jointly converge $\mathcal{G}^t$-stably for $(N_x, N_y, T) \to \infty$ to a mixed normal distribution, whose asymptotic variance is random but measurable with respect to the sigma-field $\mathcal{G}^t$. Assumption G4.8 is needed for the asymptotic distribution of the variance correction term in Theorem 2 as its asymptotic variance is random. Our simplified factor model specified by Assumptions S1 satisfies a central limit theorem for stable convergence in law.

B A Simplified Factor Model

We present a simplified factor model with the stronger Assumptions S1 and S2, which substantially simplifies the notation but conveys the main conceptual insights of the general model. It allows us to highlight the effect of the target weight, the relaxation of weak factors and missing observations. In particular, under these assumptions, we can provide explicit expressions for the asymptotic variances in Theorem 2.

The consistency results are based on the simplified Assumption S1 that assumes that all observations are i.i.d. The key element is the strength of the factors measured by their loadings. Specifically, we measure the strength of the factors by the fraction p_j of units in Y that are affected by the corresponding factor. The error terms are non-systematic with bounded eigenvalues in the covariance matrix. Allowing for more complex dependency as in our general model does not change the arguments, but makes the notation more complex.

Assumption S1 (Simplified factor model). *There exists constant $C < \infty$ such that*

1. *Factors:* $F_t \overset{i.i.d.}{\sim} (0, \Sigma_F)$ *and* $\mathbb{E}\|F_t\|^4 \leq C$ *for any* t.
2. *Loadings:* $(\Lambda_x)_i \overset{i.i.d.}{\sim} (0, \Sigma_{\Lambda_x})$*, where* Σ_{Λ_x} *is positive semidefinite.* $(\Lambda_y^{\text{full}})_i \overset{i.i.d.}{\sim} (0, \Sigma_{\Lambda_y}^{\text{full}})$ *and the loading of the* j*-th factor* $(\Lambda_y)_{ij} = (\Lambda_y^{\text{full}})_{ij} \cdot (U_y)_{ij}$*, where* $\Sigma_{\Lambda_y}^{\text{full}}$ *is positive definite and the Bernoulli random variable* $(U_y)_{ij} \in \{0, 1\}$ *is independent in* i *with* $\mathbb{P}((U_y)_{ij} = 1) = p_j$ *for some* $p_j \in [0, 1]$*. Furthermore,* $\mathbb{E}\|(\Lambda_x)_i\|^4 \leq C$*,* $\mathbb{E}\|(\Lambda_y)_i\|^4 \leq C$*,* $N_y^{-1} \sum_{i=1}^{N_y} (\Lambda_y)_i (\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y}$*, and* $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y}$ *is positive definite. For any* t*,* $N_y^{-1} \sum_{i=1}^{N_y} W_{ti}^Y (\Lambda_y)_i (\Lambda_y)_i^\top \xrightarrow{p} \Sigma_{\Lambda_y,t}$ *and* $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y,t}$ *is positive definite.*
3. *Idiosyncratic errors:* $(e_x)_{ti} \overset{i.i.d.}{\sim} (0, \sigma_{e_x}^2), (e_y)_{ti} \overset{i.i.d.}{\sim} (0, \sigma_{e_y}^2)$*,* $\mathbb{E}(e_x)_{ti}^8 \leq C, \mathbb{E}(e_y)_{ti}^8 \leq C$.
4. *Independence:* $F, \Lambda_x, \Lambda_y, e_x$ *and* e_y *are independent.*

Our generative model for $(\Lambda_y)_i$ accounts for three cases of factor strength on Y. First, if p_j is bounded away from 0 as N_y grows, then the j-th factor is a strong factor in Y. Second, if p_j decays to 0 but is nonzero as N_y grows, then the j-th factor is a weak factor in Y. Third, if p_j is 0 for all N_y, then Y does not contain the j-th factor. Note that Σ_{Λ_x} can be rank deficient, implying that the loadings of some factors can be zero for units in X. However, our assumption on $(\Lambda_x)_i$ rules out the case of weak factors in X as the estimation of weak factors in X is not our objective. We assume that $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y}$ is positive definite to ensure that each factor in F is strong in at least one of the two panels X and Y. Specifically, weak factors in Y are strong in X. Hence, all factors can be identified with target-PCA with a properly chosen γ.

Furthermore, Assumption S1.2 imposes assumptions on the missing pattern in Y to identify all factors when combining the partially observed Y and X. More specifically, the second-moment matrix $\Sigma_{\Lambda_y,t}$ does not need to be full rank in Assumption S1.2, which relaxes the full-rank assumption of $\Sigma_{\Lambda_y,t}$ in Xiong and Pelger (2023). However, we assume that $\Sigma_{\Lambda_x} + \Sigma_{\Lambda_y,t}$ is positive definite, so that target-PCA can identify all factors from X and partially observed Y.

The asymptotic results require additional restrictions on the observation pattern, as stated in Assumption S2 below.

Assumption S2 (Moment conditions under partial observations). *For any* i*, there exist constants* $\omega_i^{(1)}$*,* $\omega_i^{(2,1)}$*,* $\omega_i^{(2,2)}$*,* $\omega_i^{(2,3)}$ *and* $\omega_i^{(3)}$*, such that* $N_y^{-1} \sum_{j=1}^{N_y} \frac{q_{ij}}{q_{ii} q_{jj}} \xrightarrow{p} \omega_i^{(1)}$*,* $N_y^{-2} \sum_{j,l=1}^{N_y} \frac{q_{ii,jl}}{q_{ii} q_{jl}} \xrightarrow{p} \omega_i^{(2,1)}$*,* $N_y^{-2} \sum_{j,l=1}^{N_y} \frac{q_{jj,il}}{q_{jj} q_{il}} \xrightarrow{p} \omega_i^{(2,2)}$*,* $N_y^{-2} \sum_{j,l=1}^{N_y} \frac{q_{ij,il}}{q_{ij} q_{il}} \xrightarrow{p} \omega_i^{(2,3)}$*, and* $N_y^{-3} \sum_{j,l,h=1}^{N_y} \frac{q_{il,jh}}{q_{il} q_{jh}} \xrightarrow{p} \omega_i^{(3)}$*. Furthermore, there exist constants* $\omega^{(1)}$*,* $\omega^{(2)}$ *and* $\omega^{(3)}$*, such that* $N_y^{-1} \sum_{i=1}^{N_y} \omega_i^{(1)} \xrightarrow{p} \omega^{(1)}$*,* $N_y^{-1} \sum_{i=1}^{N_y} \omega_i^{(2,1)} = N_y^{-1} \sum_{i=1}^{N_y} \omega_i^{(2,2)} \xrightarrow{p} \omega^{(2)}$*, and* $N_y^{-1} \sum_{i=1}^{N_y} \omega_i^{(3)} \xrightarrow{p} \omega^{(3)}$.

Assumption S2 introduces interpretable parameters that fully capture the information in the observation pattern, analogous to Xiong and Pelger (2023). These key parameters are $\omega_i^{(1)}$, $\omega_i^{(2,1)}$, $\omega_i^{(2,2)}$, $\omega_i^{(2,3)}$, $\omega_i^{(3)}$, $\omega^{(1)}$, $\omega^{(2)}$, and $\omega^{(3)}$, and they are relevant for the asymptotic distribution. All these quantities are the averages of $\frac{q_{il,jh}}{q_{il} q_{jh}}$, where $\frac{q_{il,jh}}{q_{il} q_{jh}}$ roughly measures the correlation in the observation patterns for unit i, j, h and l. The superscript $m \in \{1,2,3\}$ in $\omega_i^{(m)}$, $\omega_i^{(m,m')}$ and $\omega^{(m)}$ refers to the number of indices over which $\frac{q_{il,jh}}{q_{il}q_{jh}}$ is averaged. Roughly speaking, $\omega_i^{(m)}$ and $\omega_i^{(m,m')}$ measure the average correlation in the observation patterns between unit i and any other m units. $\omega^{(m)}$ is the average of $\omega_i^{(m)}$ or $\omega_i^{(m,m')}$ over i. Note that $\omega_i^{(2,1)}$, $\omega_i^{(2,2)}$ and $\omega_i^{(2,3)}$ are closely connected, but they are slightly different in that the two indices over which $\frac{q_{il,jh}}{q_{il}q_{jh}}$ are averaged are different.

In the special case, when observations are missing at random with observed probability p, we have $\frac{q_{il,jh}}{q_{il}q_{jh}} = 1$ for four distinct units i, j, h and l, and we can show that $\omega_i^{(2,3)} = 1/p$, $\omega_i^{(1)} = \omega_i^{(2,1)} = \omega_i^{(2,2)} = \omega_i^{(3)} = 1$, and $\omega^{(1)} = \omega^{(2)} = \omega^{(3)} = 1$. For other observation patterns, $\omega_i^{(\cdot)}$, $\omega_i^{(\cdot,\cdot)}$ and $\omega^{(\cdot)}$ tend to increase if there are stronger correlations in whether entries are observed across units and across time. In addition, these quantities tend to decrease with the fraction of observed entries.[17] In Corollary 1 below, we show that these quantities can summarize all the information in the observation pattern that is relevant for the efficiency of the estimated factor model.

The simplified factor model specified by Assumptions S1 and S2 satisfies the general Assumptions G3 and G4, as stated in Proposition 5.

Proposition 5. *The simplified factor model specified in Assumptions S1 and S2 is a special case of the general approximate factor model assumed in Assumptions G3 and G4. Specifically, Assumptions G2 and S1 imply Assumption G3, and Assumptions G2, S1 and S2 imply Assumption G4.*

The distribution results of Theorem 2 simplify under Assumptions S1 and S2, and we can provide explicit expressions for the asymptotic variances. Corollary 1 shows the analytical expression of the asymptotic variances under the simplified factor model, which allows us to gain intuition on how γ affects the efficiency of the estimation.

Corollary 1. *Under Assumptions G2, S1 and S2, and for $\gamma = r \cdot N_x/N_y$ with some positive constant r, the asymptotic distributions in Theorem 2 hold as $T, N_x, N_y \to \infty$. In addition, if we assume that q_{ij} and $q_{ij,hl}$ are independent of $(\Lambda_x)_m(\Lambda_x)_m^\top$ and $(\Lambda_y)_m(\Lambda_y)_m^\top$ for any i,j,h,l,m, then the asymptotic variances in Theorem 2 can be explicitly written as follows:*

1. *The asymptotic variance of the estimated loadings of Y in (7) is*

$$\Sigma_{\Lambda_y,i}^{(\gamma)} = \Sigma_{\Lambda_y,i}^{(\gamma),\text{obs}} + \Sigma_{\Lambda_y,i}^{(\gamma),\text{miss}},$$

where

$$\Sigma_{\Lambda_y,i}^{(\gamma),\text{obs}} = \frac{1}{q_{ii}}\sigma_{e_y}^2 \Sigma_F^{-1},$$

$$\Sigma_{\Lambda_y,i}^{(\gamma),\text{miss}} = \left(\frac{1}{q_{ii}} - 1\right)\Sigma_F^{-1}((\Lambda_y)_i^\top \otimes I_k)\Xi_F((\Lambda_y)_i \otimes I_k)\Sigma_F^{-1} + \left(\omega_i^{(2,3)} - \frac{1}{q_{ii}}\right)\Sigma_F^{-1}\left(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\right)^{-1} \left[\sigma_{e_y}^2 r^2 \Sigma_{\Lambda_y}\Sigma_F\Sigma_{\Lambda_y} + \left((\Lambda_y)_i^\top \otimes r\Sigma_{\Lambda_y}\right)\Xi_F\left((\Lambda_y)_i \otimes r\Sigma_{\Lambda_y}\right)\right]\left(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\right)^{-1}\Sigma_F^{-1},$$

[17] See Table 5 in Xiong and Pelger (2023) for the values of these quantities under different observation patterns and fractions of observed entries.

and $\Xi_F = \mathrm{Var}(vec(F_t F_t^\top))$.

2. *For case 1 in Theorem 2.2, the asymptotic variance of the estimated factors in* (8) *is*

$$\Sigma_{F,t}^{(\gamma)} = \frac{\delta_{N_y,T}}{N_y} \cdot \Sigma_{F,t}^{(\gamma),\mathrm{obs}} + \frac{\delta_{N_y,T}}{T} \cdot \Sigma_{F,t}^{(\gamma),\mathrm{miss}},$$

where

$$\begin{aligned}
\Sigma_{F,t}^{(\gamma),\mathrm{obs}} &= (\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})^{-1}\Big(\frac{N_y}{N_x}\sigma_{e_x}^2 \Sigma_{\Lambda_x} + r^2 \sigma_{e_y}^2 \Sigma_{\Lambda_y,t}\Big)(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})^{-1}, \\
\Sigma_{F,t}^{(\gamma),\mathrm{miss}} &= (\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})^{-1}\left(I_k \otimes F_t^\top \Sigma_F^{-1}\big(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\big)^{-1}\right)\Big[\big(\Sigma_{\Lambda_x} \otimes r\Sigma_{\Lambda_y} + r\Sigma_{\Lambda_y,t} \otimes \Sigma_{\Lambda_x}\big) \\
&\quad \Xi_F\Big[(\omega^{(1)} - 1)\big(\Sigma_{\Lambda_x} \otimes r\Sigma_{\Lambda_y} + r\Sigma_{\Lambda_y,t} \otimes \Sigma_{\Lambda_x}\big) + (\omega^{(2)} - 1)\big(r\Sigma_{\Lambda_y,t} \otimes r\Sigma_{\Lambda_y}\big)\Big] + \\
&\quad \big(r\Sigma_{\Lambda_y,t} \otimes r\Sigma_{\Lambda_y}\big)\,\Xi_F\Big[(\omega^{(2)} - 1)\big(\Sigma_{\Lambda_x} \otimes r\Sigma_{\Lambda_y} + r\Sigma_{\Lambda_y,t} \otimes \Sigma_{\Lambda_x}\big) + (\omega^{(3)} - 1) \\
&\quad \big(r\Sigma_{\Lambda_y,t} \otimes r\Sigma_{\Lambda_y}\big)\Big]\Big]\big(I_k \otimes (\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y})^{-1}\Sigma_F^{-1} F_t\big)\big(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t}\big)^{-1}.
\end{aligned}$$

For case 2 in Theorem 2.2, the asymptotic variance of the estimated weak factors in (9) *is*

$$\Sigma_{F_w,t}^{(\gamma)} = \frac{\delta_{N_w,T}}{N_w} \cdot \Sigma_{F_w,t}^{(\gamma),\mathrm{obs}} + \frac{\delta_{N_w,T}}{T} \cdot \Sigma_{F_w,t}^{(\gamma),\mathrm{miss}},$$

where

$$\Sigma_{F_w,t}^{(\gamma),\mathrm{obs}} = (\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})_w^{-1}\Big(\frac{N_w}{N_x}\sigma_{e_x}^2 \Sigma_{\Lambda_x,w} + r^2 \frac{p_w N_w}{N_y}\sigma_{e_y}^2 \Sigma_{\Lambda_y,t,w}\Big)(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})_w^{-1},$$

$N_w = \min(N_y/p_w, N_x)$, p_w *is defined in Assumption S1.2 as the fraction of units in* Y *that are affected by the weak factors,* $(N_y p_w)^{-1}\sum_{i=1}^{N_y} W_{ti}^Y (\Lambda_y)_{i,w}(\Lambda_y)_{i,w}^\top \xrightarrow{p} \mathcal{N}(0, \Sigma_{\Lambda_y,t,w})$, $(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})_w^{-1}$, $\Sigma_{\Lambda_x,w}$ *and* $\Sigma_{F_w,t}^{(\gamma),\mathrm{miss}}$ *are respectively the diagonal block of* $(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t})^{-1}$, Σ_{Λ_x} *and* $\Sigma_{F,t}^{(\gamma),\mathrm{miss}}$ *corresponding to the weak factors.*

3. *The asymptotic variance of the estimated common components of* Y *in* (10) *is*

$$\begin{aligned}
\Sigma_{C,ti}^{(\gamma)} = &\frac{\delta_{N_y,T}}{T} F_t^\top \left(\Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{obs}} + \Sigma_{\Lambda_y,i}^{(\gamma),\mathrm{miss}}\right) F_t + \frac{\delta_{N_y,T}}{N_y}(\Lambda_y)_i^\top \Sigma_{F,t}^{(\gamma),\mathrm{obs}}(\Lambda_y)_i \\
&+ \frac{\delta_{N_y,T}}{T}(\Lambda_y)_i^\top \Sigma_{F,t}^{(\gamma),\mathrm{miss}}(\Lambda_y)_i - 2\frac{\delta_{N_y,T}}{T}(\Lambda_y)_i^\top \Sigma_{\Lambda_y,F,i,t}^{(\gamma),\mathrm{miss,cov}} F_t,
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{\Lambda_y,F,i,t}^{(\gamma),\mathrm{miss,cov}} &= \big(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y,t}\big)^{-1}\left(I_k \otimes F_t^\top \Sigma_F^{-1}\big(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\big)^{-1}\right)\Big[\big(\Sigma_{\Lambda_x} \otimes r\Sigma_{\Lambda_y} + r\Sigma_{\Lambda_y,t} \otimes \Sigma_{\Lambda_x}\big) \\
&\quad \Xi_F\left((\omega_i^{(1)} - 1)((\Lambda_y)_i \otimes \Sigma_{\Lambda_x}) + (\omega_i^{(2,2)} - 1)((\Lambda_y)_i \otimes r\Sigma_{\Lambda_y})\right) + (r\Sigma_{\Lambda_y,t} \otimes r\Sigma_{\Lambda_y})\Xi_F \\
&\quad \left((\omega_i^{(2,1)} - 1)((\Lambda_y)_i \otimes \Sigma_{\Lambda_x}) + (\omega_i^{(3)} - 1)\big((\Lambda_y)_i \otimes r\Sigma_{\Lambda_y}\big)\right)\Big]\big(\Sigma_{\Lambda_x} + r\Sigma_{\Lambda_y}\big)^{-1}\Sigma_F^{-1}.
\end{aligned}$$

The simplified model provides a clear interpretation of how the asymptotic variance is affected by missingness and the target weight. The explicit expressions in Corollary 1 show how the asymptotic variances depend on the following key quantities: the target weight γ, the noise ratio (NR) $\sigma_{e_x}/\sigma_{e_y}$, the dimension ratio (DR) N_x/N_y and the dependency structure in the missing pattern captured by the parameters defined in Assumption S2. Without missing data, the correction matrices in the variance disappear. Some correction terms also disappear when data is missing at random, as in this case it holds that $\omega_i^{(2,3)} = 1/p$, $\omega_i^{(1)} = \omega_i^{(2,1)} = \omega_i^{(2,2)} = \omega_i^{(3)} = 1$, and $\omega^{(1)} = \omega^{(2)} = \omega^{(3)} = 1$.

C Sparser Observation Pattern

In the main text, we have considered the case where the number of observed entries in Y is of the order N_yT. In this section, we generalize our results to the case where the target Y has even more missing observations. For this case, we can still estimate the factor model using our target-PCA method, but we have a slower convergence rate due to more missingness. Specifically, we assume that the number of observed time periods in Y can grow sublinearly in T.

Assumption G1′. *(Sparser Observation Pattern)*

1. *The observation matrix W^Y is independent of the factors F and idiosyncratic errors e_y.*
2. *For any given observation matrix W^Y, there exist positive constants q_1, q_2 and $\alpha \in (0,1]$ such that $q_1 \geq |Q_{ij}^Y|/T^\alpha \geq q_2 > 0$ for all i,j. Furthermore, let $q_{ij} = \lim_{T\to\infty} |Q_{ij}^Y|/T^\alpha$ and $q_{ij,hl} = \lim_{T\to\infty} |Q_{ij}^Y \cap Q_{hl}^Y|/T^\alpha$. For any i,j,h,l, q_{ij} and $q_{ij,hl}$ are positive constants bounded away from 0.*

Assumption G1′ generalizes Assumption G2. More specifically, Assumption G1′.1 is the same as Assumption G2.1. Assumption G1′.2 is more general than Assumption G2.2, which is the special case for $\alpha = 1$. We assume that the number of time periods when any two and any four units are observed are both proportional to T^α. This assumption includes the important example of low-frequency observations in Figure 1(c), where the observed time periods are the same for all units and are at the order of T^α. Under Assumption G1′, if X contains all the factors in Y, the latent factor model on Y can still be consistently estimated from target-PCA and the estimators are asymptotically normal, but with a slower convergence rate.

Proposition 6. *Define $\delta_{N_y,T^\alpha} = \min(N_y, T^\alpha)$ and select γ such that $\Sigma_{\Lambda,t}^{(\gamma)}$ is a positive definite matrix and the eigenvalues of $\Sigma_F \left(\Sigma_{\Lambda_x} + \gamma N_y/N_x \cdot \Sigma_{\Lambda_y}\right)$ are distinct. Suppose Assumptions G1′ and G3 hold, and Assumption G4 holds with $\sqrt{T}$ replaced by $\sqrt{T^\alpha}$. As $T, N_x, N_y \to \infty$ we have for each i and t:*

1. *For $\sqrt{T^\alpha}/N_y \to 0$, the asymptotic distribution of the loadings of Y in Theorem 2.1 continues to hold with convergence rate $\sqrt{T^\alpha}$.*
2. *For $\sqrt{T^\alpha}/N_y \to 0$ and $\sqrt{N_y}/T^\alpha \to 0$, the asymptotic distribution of the factors in Theorem 2.2 and the asymptotic distribution of the common components of Y in Theorem 2.3 continue to hold with convergence rate $\sqrt{\delta_{N_y,T^\alpha}}$.*

Note that the convergence rate of the estimated loadings for each unit in Y is $\sqrt{T^{\alpha}}$, and the convergence rate of the estimated factors and common components equals the smaller value of $\sqrt{T^{\alpha}}$ and $\sqrt{N_y}$. This convergence rate makes sense, as for each unit in Y, we only have T^{α} time-series observations to estimate its factor loadings.

Choosing γ under Assumption G1$'$ is conceptually similar to choosing γ in our base case. Target-PCA essentially estimates the latent factor model from a matrix that combines the sample covariance matrix of X (that is $X^{\top}X/T$) and the sample covariance matrix of Y weighted by γ (that is $\gamma \cdot \tilde{Y}^{\top}\tilde{Y}/T^{\alpha}$ if Y has a low-frequency observation pattern). As the top eigenvalues in $X^{\top}X/T$ and of $\gamma \cdot \tilde{Y}^{\top}\tilde{Y}/T^{\alpha}$ are at the order of N_x and γN_y,[18] selecting $\gamma = r \cdot N_x/N_y$ ensures that the top eigenvalues in two sample covariance matrices are of the same order, and the strong factors in either X or Y can be identified with target-PCA. In summary, the general logic and selection rules from the main setting carry over to this more general case.

D An Alternative Estimator

An alternative approach to estimate F, Λ_x and Λ_y is based on the identification assumption $F^{\top}F/T = I_k$, which allows us solve for F from the objective function

$$\max_{F} \operatorname{trace}\left(F^{\top}\left(Z^{(\gamma)}Z^{(\gamma)\top}\right)F\right), \tag{11}$$

when Y is fully observed. The solution to this optimization problem is obtained by applying PCA to $Z^{(\gamma)}Z^{(\gamma)\top}$. This is essentially the same as applying PCA to the $T \times T$ time-series second moment matrix of $Z^{(\gamma)}$, denoted by $\Sigma^{Z(\gamma),\mathsf{time}}$, with the (s,t)-th entry (suppose zero idiosyncratic noise for simplicity)

$$\Sigma^{Z(\gamma),\mathsf{time}}_{s,t} = F_s^{\top}\left(\lim_{N_x,N_y\to\infty} \frac{N_x}{N_x+N_y}\Sigma_{\Lambda_x} + \gamma\frac{N_y}{N_x+N_y}\Sigma_{\Lambda_y}\right)F_t.$$

Without missing values, the estimation of the latent factor model by applying PCA to either the cross-sectional or time-series second-moment matrix of $Z^{(\gamma)}$ is equivalent. However, this changes in the case of only partially observed data.

When Y has missing observations, we can estimate $\Sigma^{Z(\gamma),\mathsf{time}}_{(s,t)}$ using an approach analogous to the one in Section 2.4:

$$\hat{\Sigma}^{Z(\gamma),\mathsf{time}}_{s,t} = \frac{N_x}{N_x+N_y}\cdot \underbrace{\frac{1}{N_x}\sum_{i=1}^{N_x} X_{si}X_{ti}}_{\text{estimate } F_s^{\top}\Sigma_{\Lambda_x}F_t} + \gamma\frac{N_y}{N_x+N_y}\underbrace{\frac{1}{|Q_{st}|}\sum_{j\in Q_{st}} Y_{sj}Y_{tj}}_{\text{estimate } F_s^{\top}\Sigma_{\Lambda_y}F_t},$$

where Q_{st} is the set of units in Y that are observed at both time s and t. Then, we can estimate F by applying PCA to $\hat{\Sigma}^{Z(\gamma),\mathsf{time}}$.

[18] If $\alpha = 1$, then this is equivalent to selecting γ such that the top eigenvalues of $X^{\top}X$ and of $\gamma \cdot \tilde{Y}^{\top}\tilde{Y}$ are at the same order (or equivalently, $XX^{\top}$ and $\gamma \cdot \tilde{Y}\tilde{Y}^{\top}$ given that for any matrix A, the nonzero eigenvalues of $AA^{\top}$ are the same as the nonzero eigenvalues of $A^{\top}A$).

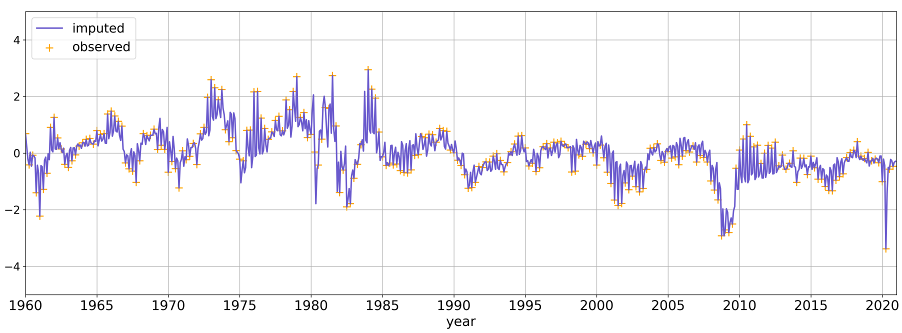
In order to ensure consistency of the PCA factors based on $\hat{\Sigma}^{Z(\gamma),\mathsf{time}}$, we need to require that the following holds for any s and t,

$$\frac{1}{|Q_{st}|}\sum_{j\in Q_{st}} Y_{sj}Y_{tj} = F_t^\top\left(\frac{1}{|Q_{st}|}\sum_{j\in Q_{st}} \left(\Lambda_y\right)_{tj}\left(\Lambda_y\right)_{tj}^\top\right)F_s + O_P(1) = F_s^\top \Sigma_{\Lambda_y} F_t + O_P(1),$$

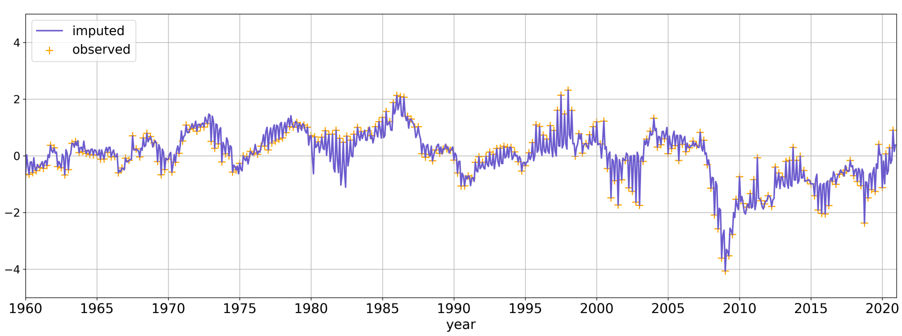
where the dominating eigenvalues have to come from $F_s^\top \Sigma_{\Lambda_y} F_t$, that is, the eigenvectors of the largest eigenvalues of $|Q_{st}|^{-1}\sum_{j\in Q_{st}} Y_{sj}Y_{tj}$ converge to the factors up to a rotation. This implies that the observation pattern of a unit has to be independent of its loadings. For causal inference applications, this assumption is less benign than Assumption G2 that the observation pattern of a unit is independent of the factors. Specifically, the treatment adoption pattern is commonly allowed to depend on the characteristics of units (e.g. the loadings) in observational studies.

E Empirical Study

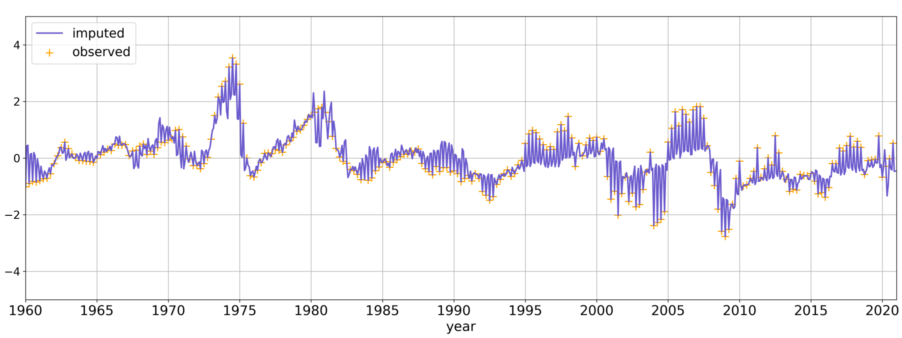
Figure A.1: Examples of quarterly observed vs. monthly imputed time series with target-PCA



(a) Goods



(b) Total liabilities and equity in domestic financial sectors



(c) Net worth (IMA) of state and local governments

This figure shows examples of observed monthly imputed time series with target-PCA. The quarterly observed target consists of macroeconomic time series of the national income & product accounts category and the flow of funds category from the FRED database. The orange "+" denotes the quarterly observed values of the macroeconomic time series, and the purple curve displays the imputed values of the time series with target-PCA for $k = 5$ latent factors. The time series are percentage changes relative to the prior year.