Counterfactual Forecasting For Panel Data

Navonil Deb
Cornell University

Raaz Dwivedi
Cornell Tech

Sumanta Basu
Cornell University

Abstract

We address the challenge of forecasting counterfactual outcomes in a panel data with missing entries and temporally dependent latent factors—a common scenario in causal inference, where estimating unobserved potential outcomes ahead of time is essential. We propose Forecasting Counterfactuals under Stochastic Dynamics (FOCUS), a method that extends traditional matrix completion methods by leveraging time series dynamics of the factors, thereby enhancing the prediction accuracy of future counterfactuals. Building upon a PCA estimator, our method accommodates both stochastic and deterministic components within the factors, and provides a flexible framework for various applications. In case of stationary autoregressive factors and under standard conditions, we derive error bounds and establish asymptotic normality of our estimator. Empirical evaluations demonstrate that our method outperforms existing benchmarks when the latent factors have an autoregressive component. We illustrate FOCUS to the HeartSteps, a mobile health study, illustrating its effectiveness in forecasting step counts for users receiving activity prompts, thereby leveraging temporal patterns in user behavior.

1 Introduction

Counterfactual estimation is a central challenge in panel data settings, with wide-ranging applications in personalized healthcare, economics, recommendation systems, and policy evaluation. An additional key task is forecasting counterfactual outcomes under hypothetical interventions even when the future interventions are not assigned yet– especially relevant to disciplines that require prospective future decision-making. The difficulty particularly arises when the outcomes are noisy and temporally dependent across units. A powerful approach models outcome trajectories via low-dimensional latent factors that capture shared variation over time and units. This idea underpins prominent methods in causal inference, including synthetic controls [1], difference-in-differences [7, 26] and matrix completion methods [3, 8, 24], which rely on low-rank assumptions. Incorporating the temporal evolution of latent factors enables forecasting counterfactual outcomes beyond the observed panel.

To formalize the problem, we consider a potential outcome model with N units and T time points. For unit i and time t, the potential outcome under treatment $w \in \{0, 1\}$ is denoted by $Y_{i,t}(w)$ and modeled as

$$Y_{i,t}(w) = \theta_{i,t}(w) + \varepsilon_{i,t}, \tag{1}$$

where $\theta_{i,t}(w)$ is called a *common component*, and $\varepsilon_{i,t}$ is zero mean *idiosyncratic noise*. We consider a linear factor model on the common components as follows:

$$\theta_{i,t}(w) = \Lambda_i^\top(w) F_t(w) \tag{2}$$

where $F_t(w)$ are the underlying factors and $\Lambda_i(w)$ are the loadings, both having dimension $r \le \min\{N, T\}$. Our goal is to forecast the mean potential outcome $\theta_{i,T+h}(w)$ for a given horizon h, under time series structure on $F_t(w)$. A special case of the factors dynamics illustrated in this work is a vector autoregressive model of order 1 (VAR(1)) [38, Sec. 2.1.1]

$$F_t(w) = A(w) F_{t-1}(w) + \eta_t(w), \tag{3}$$

where $A_t(w) \in \mathbb{R}^{r \times r}$ is the coefficient matrix associated with the VAR(1) process. $\eta_t(w)$ is a stationary noise process, independent across t, with mean 0 and covariance matrix $\Sigma_\eta(w)$.

One motivation for counterfactual forecasting arises in policy-making and mobile health (mHealth) applications, where timely decision-making is critical. In the HeartSteps V1 study [35, 36], users received five daily prompts encouraging walking activities. Walking behaviors of each user often exhibit a clear temporal structure across users: step counts in consecutive decision slots are negatively correlated, with higher activity in one slot typically followed by lower activity in the next (see Fig. 5 and 6(a)-(c) in Appendix). When modeled with a linear factor model, the latent factors obtained from the steps under the intervention exhibit strong correlations across some consecutive slots (Fig. 2(a) and in Appendix). Capturing this dependencies in the factor level enables accurate forecasting of future steps under intervention (or control), thereby facilitating prospective evaluation of the intervention's effectiveness.

Related works. Recent advances address matrix completion and treatment effect estimation via low-rank factor models, however these works do not exploit temporal dynamics of the factors, and do not conduct forecasting [12, 32, 18, 48, 4, 20]. Goldin et al. [25] propose SyNBEATS– a neural network-based method with synthetic controls that performs post-treatment counterfactual estimation. However they do not provide theoretical guarantees, and do not leverage the latent dynamics to forecast out-of-sample. Another limitation of SyNBEATS is its reliance on a fully observed pre-treatment panel, which prevents it from addressing missing entries in control units before the intervention. Pang et al. [40] propose a Bayesian alternative to synthetic control methods and consider autoregressive factors in the model, however do not address the out-of-sample forecast performance and lacks theoretical guarantees. Ben-Michael et al. [14], Chen et al. [19] incorporate multi-task Gaussian process to model the potential outcome dynamics, however do not conduct forecasting. Agarwal et al. [2], Alomar et al. [6] address forecasting in panel data under missing observations with Multivariate singular spectrum analysis (mSSA) that addresses low rank assumption and time varying deterministic factors with singular spectrum, but does not accommodate stochastic and serially correlated factors with non-singular spectra.

Several works show that smoothing the estimated factors improves forecast accuracy by exploiting their temporal structure [22, 41]. Often in context of factor-augmented VAR models [15, 37], dynamic factor models use *collapsing techniques* for capturing the latent factor dynamics [34, 16]. However, the existing forecast methods in dynamic factor models assume that the extent of missingness in the panel is limited. In such settings, the proportion of missing entries is typically finite and smaller than that in counterfactual estimation frameworks– invoking the need of a suitable counterfactual forecast method equipped to handle wide range of factor dynamics and missing entries.

Our contributions. We introduce a method namely Focus (*forecasting counterfactuals under stochastic dynamics*) in Sec.3. To the best of our knowledge, this is the first work to address forecasting counterfactuals out-of-sample. Equipped with a reliable factor estimation algorithm, Focus has two steps- (i) restricts the observed panel under treatment (or control) and estimates the latent factors from the treatment panel by considering the control (or treated) observations as missing, (ii) fits an appropriate time series model on the estimated factors to construct a forecast estimator $\hat{\theta}_{i,T:T+h}$ (in (12)) of the forecast estimand $\theta_{i,T:T+h}$ (in (6)) for unit i horizon h. On a general note, our method integrates smoothing and filtering techniques from time series analysis with matrix completion and counterfactual estimation methods. Allowing for stochastic dynamics in the factors makes Focus more flexible to handling non-singular spectra. Additionally, Focus is equipped to capture latent temporal dependencies by accommodating serial correlation of the factors. We overcome the limitation of restricted missingness of panel forecasting methods in dynamic factor model literature by proposing algorithm equipped to handle wide range of missing patterns. When the factors are estimated with the PCA algorithm of Xiong and Pelger [48], under a VAR(1) assumption on the true factors with VAR coefficient matrix A, and standard assumptions on the loadings, the noise in the outcome model and observation mechanism, we establish high probability error bounds associated with $\hat{\theta}_{i,T:T+h}$ (Thm.4.1). The error bounds carry two components due to (i) estimating the factors from the partially observed panel, and (ii) forecasting with the factors. Under additional regularities on the observation pattern, we establish $\min\left\{\sqrt{N}, \sqrt{T}\right\}$-consistency and asymptotic normality for $\hat{\theta}_{i,T:T+h}$ (Thm.4.2) and provide valid confidence intervals of $\theta_{i,T:T+h}$ (Cor.4.1). In several simulation settings, Focus yields more accurate forecast trajectories than our benchmarks mSSA and SyNBEATS, supporting the importance of leveraging the stochastic temporal latent dynamics. In HeartSteps V1, we leverage the temporal association in the factors across consecutive suggestion slots. Focus, by capturing this temporal pattern, yields more accurate forecast of step counts under intervention than mSSA.

Organization. We set up the problem under a simple setting and define our forecast estimand in Sec. 2. We illustrate our forecast algorithm Focus in Sec. 3. The main results are stated in Sec. 4. We evaluate the performance of Focus on both simulated data and the real data set HeartSteps, as presented in Sec. 5. Finally the proofs, the general assumptions and the additional experiments are deferred in the supplement.

Notation. For $n \in \mathbb{N}$, we denote $[n] = 1, \dots, n$. For $x \in \mathbb{R}^n$, $\|x\|_2$ denotes the Eucledian ℓ_2-norm of x. For any event E, $\mathbb{I}[E] = 1$ if E occurs, 0 otherwise. For a matrix M, $\|M\| = \sup_{\|x\|_2=1} |Mx|_2$; $\rho(M)$ is its spectral radius. We abbreviate almost surely to a.s., and independently and identically distributed to i.i.d.. Convergence almost surely, in probability and distribution are denoted by $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{d}$. For r.v.'s X_n and sequence a_n, $X_n = \mathcal{O}_{\mathbb{P}}(a_n)$ means X_n/a_n is bounded in probability, i.e. for every $\delta > 0$ there exists c_δ and n_δ such that $\mathbb{P}\Big(\frac{|X_n|}{a_n} > c_n\Big) < \delta$ for $n > n_\delta$. We use the terms *unit* and *time* interchangeably with *row* and *column* respectively.

2 Problem set-up

We translate the counterfactual forecast problem into a forecast problem with missing entries in the panel. We build our forecast target based on the later.

2.1 Dynamic latent factors

The observed data is denoted by $(Y_{i,t}, W_{i,t})$, $(i,t) \in [N] \times [T]$, where $Y_{i,t}$ and $W_{i,t} \in \{0,1\}$ are the observed outcome and the binary treatment variable for unit i and time t. When $W_{i,t} = 1$, we observe the treated outcome $Y_{i,t}(1)$, but not the control outcome $Y_{i,t}(0)$. Hence, $W_{i,t}$ indicates the observed entries in the potential outcomes $Y_{i,t}(1)$. Similarly, $1 - W_{i,t}$ is the observation indicator of $Y_{i,t}(0)$. We fix $w = 1$ and drop the notation w from $Y_{i,t}(w), \theta_{i,t}(w), F_{i,t}(w)$ and $\Lambda_{i,t}(w)$. Restricting to the treatment panel with $W_{i,t} = 1$, we rewrite (1) as

$$Y_{i,t} = \begin{cases} \theta_{i,t} + \varepsilon_{i,t} = \Lambda_i^\top F_t + \varepsilon_{i,t} & \text{if } W_{i,t} = 1, \\ \star & \text{if } W_{i,t} = 0. \end{cases} \tag{4}$$

Here $\star$ denotes a missing entry. We impose a stationary vector autoregressive structure on the factors as follows:

Assumption 1 (Stable VAR(1) factors)**.** *The latent factors F_t follow a stable r-dimensional VAR(1) model*

$$F_t = AF_{t-1} + \eta_t, \tag{5}$$

where $\rho(A) < 1$ and η_t is a noise process with mean 0 and covariance matrix Σ_η.

The VAR(1) structure simplifies our analysis, and can be generalized to more complex stationary time series models. The coefficient matrix A governs the structure of the autocorrelated factors. Additionally Wold's decomposition theorem Lütkepohl [38, Prop. 2.1] guarantees that under mild regularity conditions, every stationary and purely nondeterministic process is well-approximated by a finite order VAR process. The stationary VAR(1) process (5) is stable if $\det(I_r - Az) \neq 0$ for all $|z| \leq 1$, which is ensured if the eigenvalues of A are smaller than 1, i.e. $\rho(A) < 1$.

2.2 Forecast estimand

The mean outcome at forecast horizon h is $\theta_{i,T+h} = \Lambda_i^\top F_{T+h}$ that involves future realizations of the dynamic factor. Under the VAR representation (5), we express the factor at horizon h as $F_{T+h} = A^h F_T + \sum_{j=1}^h A^{h-j}\eta_{T+j}$. Since the future errors $\{\eta_{T+j} : j = 1, \ldots, h\}$ are unobserved at time T, exact recovery of F_{T+h} is not feasible. Under Assum. 1, we calculate the *best linear predictor* of F_{T+h} given the information until T encoded by the filtration $\mathcal{F}_T = \sigma(\{F_t : t \leq T\})$. This best linear predictor is the conditional expectation $\mathbb{E}[F_{T+h} \mid \mathcal{F}_T] = A^h F_T$ that minimizes mean squared error among all $\mathcal{F}_T$-measurable estimators of F_{T+h}. Accordingly the best linear predictor of the outcome variable is denoted by

$$\theta_{i,T:T+h} \triangleq \mathbb{E}[\theta_{i,T+h} \mid \mathcal{F}_T] = \Lambda_i^\top A^h F_T. \tag{6}$$

We refer to $\theta_{i,T:T+h}$ as the *forecast target* that is our estimand. Our method in Sec. 3, constructs an estimator of this target parameter. We discuss more on identifiability of $\theta_{i,T:T+h}$ under rotation in App. E.

3 Forecast method

Our method namely *forecasting counterfactuals under stochastic dynamics* (Focus) addresses the task of estimating $\theta_{i,T:T+h}$ entry-by-entry by leveraging the dynamics of F_t. We illustrate the method under the simplified factor model (4) and Assum. 1.

The inputs to Focus are the observed panel of outcome variables $Y \triangleq (Y_{i,t})_{(i,t)\in[N]\times[T]}$ and the matrix of observation indicators $W \triangleq (W_{i,t})_{(i,t)\in[N]\times[T]}$. For estimating the factors $\{F_t\}_{t\in[T]}$ and the loadings $\{\Lambda_i\}_{i\in[N]}$, we use the PCA method of Xiong and Pelger [48] interchanging the roles of unit and time. This method can be replaced by any algorithm that consistently estimates the factors and loadings under general observation pattern W. The steps of our proposed algorithm are as follows:

Focus(Y, W):

1. **Estimating the $\{F_t\}_{t\in[T]}$ and $\{\Lambda_i\}_{i\in[N]}$:**

   For each pair of columns $s, t \in [T]$, define the set

   $$\mathcal{Q}_{s,t} \triangleq \{i \in [N] : W_{i,s} = 1 \text{ and } W_{i,t} = 1\}. \tag{7}$$

   In words, $\mathcal{Q}_{s,t}$ contains the rows for which both columns s and t are observed. We calculate the sample covariance matrix $\hat{\Sigma}$ for PCA with entries

   $$\hat{\Sigma}_{s,t} \triangleq \frac{1}{|\mathcal{Q}_{s,t}|} \sum_{i\in\mathcal{Q}_{s,t}} Y_{i,t} Y_{i,s}, \quad |\mathcal{Q}_{s,t}| > 0. \tag{8}$$

Estimated factors are $\hat{F} \triangleq \left[\hat{F}_1 : \ldots : \hat{F}_T\right]^\top$ where

$$\hat{F} = \sqrt{T} \times \text{First } r \text{ eigenvectors of } \hat{\Sigma}/T, \quad (9)$$

Estimated loadings are $\hat{\Lambda} \triangleq \left[\hat{\Lambda}_1 : \ldots : \hat{\Lambda}_N\right]^\top$ where

$$\hat{\Lambda}_i = \left(\sum_{t=1}^{T} W_{i,t}\hat{F}_t\hat{F}_t^\top\right)^{-1}\left(\sum_{t=1}^{T} W_{i,t}\hat{F}_t Y_{i,t}\right). \quad (10)$$

2. **Forecasting with $\hat{F}$:**

The ordinary least squares (OLS) estimator $\hat{A}$ of the VAR(1) coefficient matrix A is calculated with $\{\hat{F}_t\}_{t\in[T]}$ as

$$\hat{A} \triangleq \left(\sum_{t=1}^{T-1} \hat{F}_{t+1}\hat{F}_t^\top\right)\left(\sum_{t=1}^{T-1} \hat{F}_t\hat{F}_t^\top\right)^{-1}. \quad (11)$$

Then the plug-in estimator of $\theta_{i,T:T+h}$ is

$$\hat{\theta}_{i,T:T+h} = \hat{\Lambda}_i^\top \hat{A}^h \hat{F}_T. \quad (12)$$

If the true factors $\{F_t\}_{t\in[T]}$ are observed, learning the autocorrelation of F_t's provide an estimator of the best linear predictor of the future F_{T+h} conditioned on their true past. Focus replaces F_t's with their PCA estimators $\hat{F}_t$'s and the associated autocorrelation are learned from the estimates. The PCA algorithm of Xiong and Pelger [48] provides consistent estimators $\hat{\Lambda}_i$'s and $\hat{F}_t$'s under a wide range of observation mechanism. We leverage their estimation approach in our forecasting method. In step 1, we assume that r i.e. the rank of the latent process is known to the algorithm. If r is unknown, several works address the rank estimation methods [10, 11, 21, 47] that can be applied here. Additionally $\hat{\Sigma}_{s,t}$ in (8) requires existence of at least one row with both columns s and t observed (see Rem.4.3 for more details). If $\mathcal{Q}_{s,t}$ is empty, we can set $\hat{\Sigma}_{s,t} = 0$ and consistency of the e estimated factors [33] can still be ensured.

Remark 3.1 (Observed covariates). The algorithm Focus can be viewed as operating on a panel that has been denoised with respect to any observed covariates X_{it}. Covariates may be handled through several methods– including interactive fixed effects [12, 20] and inverse propensity weighting of the entries of $\hat{\Sigma}$ [48, Sec. 2]. Extensions incorporating covariates are left for future work.

4 Main results

Under regularity conditions on the factor model (4) and the observation matrix W, we establish probabilistic error bounds and asymptotic normality of $\hat{\theta}_{i,T:T+h}$ obtained from Focus. The matrix of idiosyncratic noises is denoted by $\varepsilon \triangleq (\varepsilon_{i,t})_{(i,t)\in[N]\times[T]}$. Since the factors are stationary and stable, the factor covariance matrix is denoted by $\Sigma_F \triangleq \mathbb{E}[F_tF_t^\top]$.

4.1 Forecast error bound

We start with some sufficient regularity conditions on the factor model (4) for stating our main results.

Assumption 2 (Factor model)**.** *In* (5)*, the factor noises* $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_\eta)$ *for some positive definite* Σ_η*. The loadings* Λ_i *are i.i.d. with mean 0 and positive definite covariance matrix* Σ_Λ*, with* $\mathbb{E}[\|\Lambda_i\|_2^4] < \infty$*. Furthermore, the eigenvalues of* $\Sigma_F\Sigma_\Lambda$ *are distinct.*

Remark 4.1. The i.i.d. loadings in Assum. 2 also appears in Xiong and Pelger [48]. The distinct eigenvalues of $\Sigma_F\Sigma_\Lambda$ is a standard condition in the factor model literature [48, 9, 12]. The Gaussian assumption on η_t ensures existence of 4^{th} moments of F_t (Lem. D.3) that we later use to show our main results.

Assumption 3 (Idiosyncratic noise)**.** $\varepsilon_{i,t}$ *are i.i.d. with mean* 0*, variance* $\sigma_\varepsilon^2 > 0$ *and* $\mathbb{E}[|\varepsilon_{it}|^8] < \infty$*.*

Assumption 4 (Mutual independence)**.** Λ_i*,* η_t *and* $\varepsilon_{i,t}$ *are mutually independent for all* i *and* t*.*

Remark 4.2. Assum. 3 and 4 are imposed to establish our results and hold under broader conditions (Sec.D). The results in Thm.4.1 follows under these general assumptions, while the asymptotic normality result in Thm.4.2 is shown under Assum. 4 and can be extended with more elaborate variance formulas.

Next we impose some regularity conditions on the observation pattern of the panel.

Assumption 5 (Observation pattern)**.** *The observation matrix* W *is independent of the factor matrix* F *and noise process* ε*, and there exists* $\underline{q} \in (0, 1]$ *such that for* $s, t \in [T]$*,*

$$|\mathcal{Q}_{s,t}| \geq N\underline{q} \quad \textit{almost surely.} \quad (13)$$

Furthermore for $s, t, s', t' \in [T]$*, there exist* $\alpha_{s,t}, \beta_{s,t,s',t'} \in (0, 1]$ *such that for* $N \to \infty$*,*

$$\frac{1}{N}|\mathcal{Q}_{s,t}| \xrightarrow{\text{a.s}} \alpha_{s,t}, \;\; \textit{and} \;\; \frac{1}{N}|\mathcal{Q}_{s,t} \cap \mathcal{Q}_{s',t'}| \xrightarrow{\text{a.s}} \beta_{s,t,s',t'}. \quad (14)$$

Assum. 5 is similar to Xiong and Pelger [48, Assum. S1]. As N grows, the number of observed rows across any pair or any quadruple of columns concentrates around $N\alpha_{s,t}$ and $N\beta_{s,t,s',t'}$, respectively. The condition that W is independent of ε parallels the unconfoundedness assumption commonly used in treatment effect identification [44]. Requiring independence of W from F and ε does not confine W to static treatment assignments; W can also capture time-varying treatment policies. We leave more exploration of sequential policies as future work.

Remark 4.3 (Validity of the assumption). The assumption requires $|\mathcal{Q}_{s,t}| > 0$, i.e., at least one treated unit at any two time points. Under missing completely at random (MCAR; see Sec. 4.3), this condition holds almost surely for large N. Similar arguments can be developed for other observation patterns, such as sequential MAR [23, Assum. 1] and staggered adoption. The requirement $|\mathcal{Q}_{s,t}| > N\underline{q}$ and proportionality conditions (14) can be relaxed to $|\mathcal{Q}_{s,t}| > f(N)\underline{q}$ with $f(N) = o(N)$, without proportionality, at the cost of slower convergence rates for $\hat{\theta}_{i,T:T+h}$ [48, Sec. 9].

We define a shorthand $\delta_{NT} \triangleq \min\left\{\sqrt{N}, \sqrt{T}\right\}$ that controls the rates of convergence of our proposed forecast estimators. Equipped with the regularity conditions, we state our first result.

Theorem 4.1 (Error bound for $\hat{\theta}_{i,T:T+h}$)**.** *Consider a factor model* (4) *with N units and T time points satisfying Assum. 1 to 5. Then for the* Focus *estimator $\hat{\theta}_{i,T:T+h}$ in* (12)*, any fixed unit $i \in [N]$ and forecast horizon $h \geq 1$, the absolute error associated with $\hat{\theta}_{i,T:T+h}$ is bounded as*

$$\left|\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h}\right| = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big) + \mathcal{O}_{\mathbb{P}}\Big(h\|A\|^{h-1}(N^{-1} + T^{-1/2})\Big). \quad (15)$$

The proof is deferred in App. A. The result addresses high probability error bound for unit × horizon level estimates. $\mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big)$ is similar to the unit × time-level rate for estimating the mean outcome [10, 12, 48]. The second term $\mathcal{O}_{\mathbb{P}}\big(h\|A\|^{h-1}(N^{-1} + T^{-1/2})\big)$ reflects the estimation error of the coefficient matrix $\hat{A}$ with the fitted factors and forecasting with it.

Remark 4.4 (Role of A in the convergence rate). As $\rho(A)$ approaches 1, F_t exhibits slower noise decay in its MA representation (43), leading to longer memory and reduced stability [13, Prop. 2.2]. Since $\|A\| \geq \rho(A)$, the convergence rate in (15) deteriorates accordingly.

Remark 4.5 (Role of h). In practice, $\hat{\theta}_{i,T:T+h}$ is more informative for small to moderate h. Lem. D.2 shows–

$$|\theta_{i,T:T+h}| < \|\Lambda_i\|_2\|F_T\|_2\Big(\tfrac{1+\rho(A)}{2}\Big)^h, \quad h \geq \underline{N},$$

where $\underline{N}$ depends on A. Since $\|\Lambda_i\|_2$ and $\|F_T\|_2$ are $\mathcal{O}_{\mathbb{P}}(1)$ by Assum. 2 and Lem. D.3 respectively, $\theta_{i,T:T+h}$ decays exponentially. Hence the second term in (15), which captures forecast error in the estimated factors, vanishes with h. For large horizons, both $\hat{\theta}_{i,T:T+h}$ and $\theta_{i,T:T+h}$ are small, and (15) primarily reflects factor model estimation error.

4.2 Asymptotic distribution of $\hat{\theta}_{i,T:T+h}$

Before stating the result with asymptotic distribution for our proposed estimator in (12), we make additional assumptions.

Assumption 6 (Time limits of observation pattern)**.** *For every $i \in [N]$, there exists a positive definite matrix $\Sigma_{F,i}$ such that*

$$\hat{\Sigma}_{F,i} \triangleq \frac{1}{T}\sum_{t=1}^{T} W_{it}F_tF_t^\top \xrightarrow{\mathbb{P}} \Sigma_{F,i} \quad \text{as } T \to \infty. \quad (16)$$

For every $s, t \in [T]$, $\alpha_{s,T} \xrightarrow{\mathbb{P}} \nu_s \in (0,1]$ and $\beta_{s,T,t,T} \xrightarrow{\mathbb{P}} \rho_{s,t} \in (0,1]$ as $T \to \infty$. Furthermore, there exist $\omega_i \in (0,1)$, $i = 1,2,3$ such that the following limits exist as $T \to \infty$,

$$\frac{1}{T^2}\sum_{s,t=1}^{T}\frac{\beta_{s,T,t,T}}{\alpha_{s,T}\alpha_{s',T}} \xrightarrow{\mathbb{P}} \omega_1, \quad \frac{1}{T^3}\sum_{s,s',t=1}^{T}\frac{\beta_{s,t,s',T}}{\alpha_{s,t}\alpha_{s',T}} \xrightarrow{\mathbb{P}} \omega_2,$$

$$\frac{1}{T^4}\sum_{s,t,s',t'=1}^{T}\frac{\beta_{s,t,s',t'}}{\alpha_{s,t}\alpha_{s',t'}} \xrightarrow{\mathbb{P}} \omega_3.$$

(16) captures that the factors are systematic over the observed entries. The impact of missing entries on the variance is captured by ω_i, $i = 1,2,3$. Similar conditions also appear in Xiong and Pelger [48, Assum.S3].

Equipped with the regularity condition, our second result involves the asymptotic distribution of $\hat{\theta}_{i,T:T+h}$.

Theorem 4.2 (Asymptotic normality of $\hat{\theta}_{i,T:T+h}$)**.** *Consider the setup from Thm.4.1 and suppose Assum.6 holds. Then for any fixed unit $i \in [N]$ and forecast horizon $h \geq 1$, $\hat{\theta}_{i,T:T+h}$ in* (12) *satisfies*

$$\frac{\delta_{NT}}{\sigma_{i,T,h}}\left(\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h}\right) \xrightarrow{d} \mathcal{N}(0,1), \quad (17)$$

where the asymptotic variance is denoted by $\sigma_{i,T,h}^2 \triangleq \xi_{i,T,h}^2 + \tau_{i,T,h}^2$, with $\xi_{i,T,h}^2$ and $\tau_{i,T,h}^2$ defined in (25) *and* (26) *in the Appendix.*

The proof is deferred in Appendix B. Thm. 4.2 guarantees that under the regularity conditions, $\hat{\theta}_{i,T:T+h}$ is δ_{NT}-consistent and asymptotically normal. We note that $\sigma_{i,T,h}^2$ has two components– (a) The asymptotic variance due to estimating the factors from the panel with missing entries is captured by $\xi_{i,T,h}^2$, and also appears in Xiong and Pelger [48, Cor. 1] for the variance of unit × time-level outcome estimation, with the roles of rows and columns interchanged. If all the entries are observed, $\xi_{i,T,h}^2$ has a similar form with the asymptotic variance in Bai [9, Thm. 3] where the mean outcome $\theta_{i,t}$ is estimated with $\hat{\theta}_{i,t} \triangleq \hat{\Lambda}_i^\top \hat{F}_t$. (b) $\tau_{i,T,h}^2$ reflects the uncertainty arising solely from forecasting, and is asymptotically independent of other leading terms (App. B). If factors were fully observed, the VAR coefficients would be $\sqrt{T}$-consistent and asymptotically normal [38, Prop. 3.1]–yielding a similar asymptotic variance formula. When $N/T \to 0$, $\tau^2_{i,T,h} \to 0$ and uncertainty arises solely from factor estimation.

4.2.1 Forecast confidence interval

Confidence intervals are essential for quantifying the uncertainty associated with point forecasts. If we replace $\sigma^2_{i,T,h}$ in Thm. 4.2 with its consistent estimator, we obtain confidence interval for $\theta_{i,T:T+h}$. Similar to Bai [9], Xiong and Pelger [48], we can estimate $\xi^2_{i,T,h}$ with $\hat{\xi}^2_{i,T,h}$, a consistent estimator obtained via plug-in approach where the individual components in the variance formula is estimated with HAC estimators [39]. For estimating $\tau^2_{i,T,h}$, we can again use a plug-in approach similar to Lütkepohl [38, Eq. 3.2.17, Prop. 3.2] by replacing the true unknown parameters in (26) with corresponding plug-in estimators, and obtain a consistent variance estimator $\hat{\tau}^2_{i,T}$. The detailed formulas of the variance estimators are deferred in App. B.2.

Equipped with the variance estimators, we state the result for forecast confidence intervals without proof.

Corollary 4.1 (Asymptotic C.I. for $\hat{\theta}_{i,T:T+h}$)**.** *Consider the setup of Thm. 4.2. Let $\hat{\sigma}^2_{i,T,h} \triangleq \hat{\xi}^2_{i,T,h} + \hat{\tau}^2_{i,T}$ be the asymptotic variance estimator. Then given a level of significance $\alpha \in (0,1)$, the following holds*

$$\lim_{N,T\to\infty} \mathbb{P}\left(\theta_{i,T:T+h} \in \left[\hat{\theta}_{i,T:T+h} \ \mp \ z_{1-\alpha/2} \cdot \frac{\hat{\sigma}_{i,T,h}}{\delta_{NT}}\right]\right) = 1-\alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution.

Hence, we provide a $100(1-\alpha)\%$ confidence interval for $\theta_{i,T:T+h}$ under standard regularity assumptions.

4.3 Example: one factor model with MCAR

For demonstrating our results with a simple setting, we consider a *one-factor model* ($r=1$) for the outcome variable. Assum. 1 for F_t in Assum. 1 is simply translated to a first order autoregressive (AR(1)) process with coefficient a. We retain the outcome model structure of Assum. 2, 3 and 4 in the one-factor model with $\sigma^2_\eta \triangleq \mathrm{Var}(\eta_t), \sigma^2_F \triangleq \mathrm{Var}(F_t)$ and $\sigma^2_\Lambda \triangleq \mathrm{Var}(\Lambda_i)$.

Definition 4.1 (Missing complete at random)**.** *W has entries missing completely at random (MCAR) if $W_{i,t} \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(p), p \in (0,1]$ and $W \perp\!\!\!\perp \Lambda, F, \varepsilon$.*

Since $|\mathcal{Q}_{s,t}|/N$ is average of Bernoulli(p^2) random variables, MCAR can be shown to satisfy $|\mathcal{Q}_{s,t}| > N\underline{q}$ almost surely for large N (App. F). Thus the condition (13) is a reasonable assumption in the MCAR setting. The notations in Assum. 5 and 6 also translate for MCAR as in App. C.

Corollary 4.2. *Consider a one-factor model with AR(1) factors with coefficient a with $|a| < 1$, and MCAR(p) observation pattern. Then under Assum. 2 to 4 and for any unit $i \in [N]$ and horizon $h \geq 1$,* (15) *holds. Furthermore under* (13), (17) *holds with the asymptotic variance given in* (31) *and* (32)*.*

The proof is deferred to App. C. When the AR(1) factors are fully observed, $\tau^2_{i,T,h}$ reflects the uncertainty of forecasting with information up to time T [45, Sec. 3.4]. As p decreases from 1 to 0, $\xi^2_{i,T,h}$ in (31) grows and the asymptotic variance of $\hat{\theta}_{i,T:T+h}$ increases with the probability of missing observations.

5 Experimental results

We conduct experiments on simulated data and the HeartSteps data to demonstrate that our proposed method Focus outperforms the benchmark methods when the factors have temporal correlation, particularly an autoregressive component. The simulation experiments are conducted in the BioHPC server hosted by Cornell University. The HeartSteps experiments are performed in a Macbook Pro M1 with 16 GB RAM.

5.1 Simulation studies

We compare h-step Mean Squared Forecast Error for first 32 rows (for computational efficiency) denoted by

$$\mathrm{MSFE} = \frac{1}{32}\sum_{i=1}^{32}(\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h})^2,$$

with $\hat{\theta}_{i,T:T+h}$ of different benchmarks. The performance of Focus is compared with mSSA [2] and SyNBEATS [25]. We highlight that mSSA performs forecasting by imposing a deterministic latent dynamics, and SyNBEATS requires a fully observed panel up to T and post treatment observations for other units.

Setup. The outcomes are generated from a one-factor model $Y_{i,t} = \Lambda_i F_t + \varepsilon_{i,t}$, $\Lambda_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.5^2)$ and $\varepsilon_{i,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.1^2)$. We consider three data generating processes (DGP)– (a) DGP-1 involves factors from an AR(1) process and the observation pattern W is MCAR(0.7), (b) DGP-2 has factors as a sum of AR(1) process and a quadratic trend, with the observation pattern being simultaneous adoption, and (c) DGP-3 has factors with ARMA(3,1) process with a quadratic

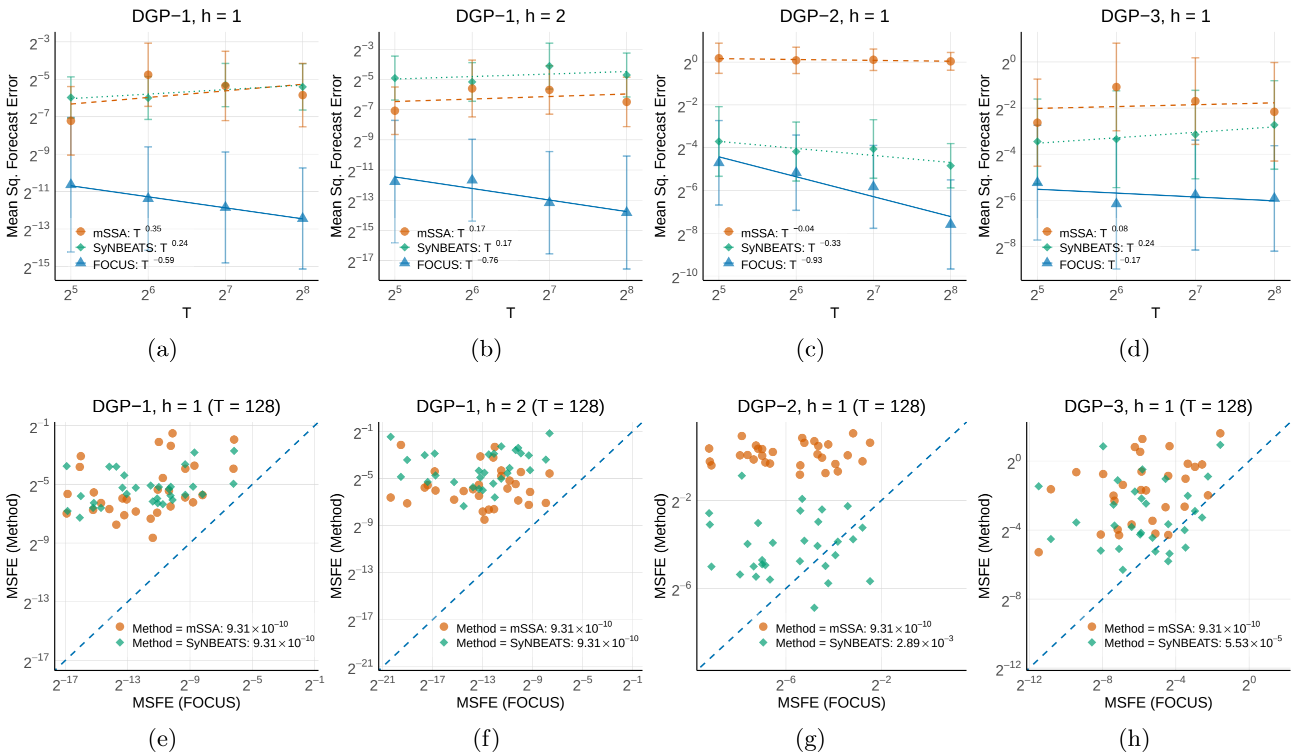


Figure 1: **Mean squared forecast error (MSFE, averaged over 30 trials) across the benchmarks for $N = 64$ and three generative models.** Panels (a)-(d) present the average MSFE of Focus (blue triangle), mSSA (orange circle) and SyNBEATS (green diamond) across $T \in \{2^5, 2^6, 2^7, 2^8\}$, and the vertical lines mark the one standard deviation error bars. As comapared to SyNBEATS and mSSA, Focus has lower average MSFE that decreases faster with T (empirical rates in the legends). Panels (e)-(h) present scatter plots of difference of MSFE (Focus Benchmark method) for $T = 128$. The errors of Focus are significantly lower (p-values of Wilcoxon's one-sided pairwise test in legends < 0.01), resulting the scatter plots concentrated in $y > x$ region.

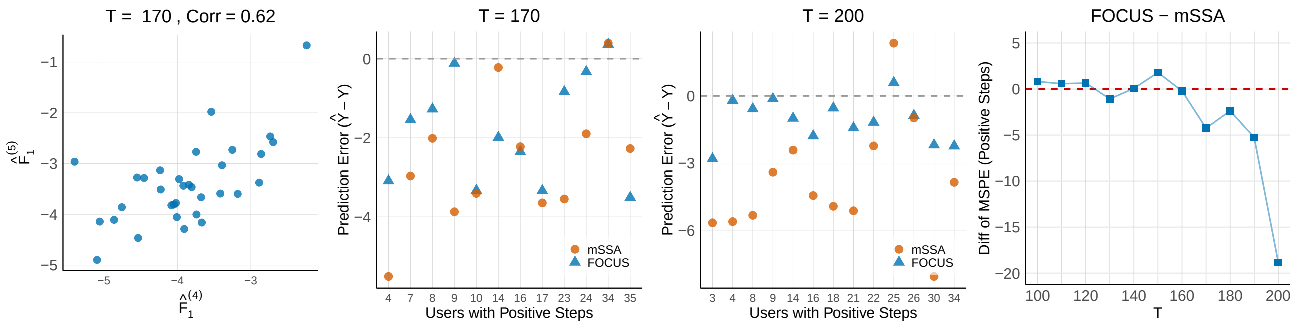
component and MCAR(0.7) observation pattern. We encounter DGP-3 to demonstrate the performance of Focus under model misspecification.

We implement our algorithm Focus in Sec. 3 with the AR order selected with *Akaike information criterion* (AIC). For DGP-2 and 3, the deterministic component in $\hat{F}_t$ is estimated with *penalized smoothing splines* [29, 27] and the residuals are used for capturing the stochastic part. For selecting the tuning parameter of the splines, we use a 10-fold block cross-validation [42, 43]. The experiments are conducted for $N = 64$, $T \in \{2^5, \ldots, 2^8\}$ with 30 trials. Additional details on the simulation setup are deferred in App. G.

Results. As shown in Fig. 1(a)–(d), Focus consistently achieves the lowest MSFE among all benchmarks across varying T, data-generating processes, and observation patterns. Moreover, Focus exhibits the fastest empirical decay rates for MSFE: under DGP-1 and DGP-2 with AR(1) factors, the observed rates are even sharper than the theoretical $\mathcal{O}(T^{-1/2})$ bound in Thm. 4.1. To assess statistical significance, Fig. 1(e)–(h) reports pairwise comparisons at $N = 64$ across experiment trials, showing that Focus outperforms all benchmarks with Wilcoxon's test p-values $< 10^{-2}$. Importantly, Focus maintains its advantage under model misspecification (DGP-3). Even when SyNBEATS is granted additional information by including pre-treatment and post-treatment outcomes, Focus still delivers significantly better forecasts.

5.2 HeartSteps V1 case study

We illustrate our method with the HeartSteps V1 dataset, a 6-week mHealth intervention study involving 37 sedentary adults. Participants received walking notifications at five daily decision points via fitness trackers (e.g., Jawbone or Google Fit) that are considered as interventions. The outcome of our interest is the log-step counts 30 minutes after each suggestion, denoted by `jbsteps30`. In particular for a user i and decision time t, $Y_{i,t} = \log(1 + \texttt{jbsteps30}_{i,t})$. The observation matrix W has entries $W_{i,t} = 1$ if and only



(a) (b) (c) (d)

Figure 2: **Results of Focus and mSSA for HeartSteps data.** Panel (a) presents a scatter plot of estimated factors for slot pair (4, 5) that shows strong temporal correlation 0.62 for $T = 170$, highlighting the predictive power leveraged by Focus. Panels (b) and (c) present the point-wise prediction error for mSSA and Focus at $T = 170, 200$ and for users with positive steps at $T + h$. For most users, Focus outputs (blue triangle) are closer to the horizontal line at 0 than mSSA (orange circle), yielding more accurate prediction. We note that for both methods, most users have negative prediction errors. Panel (c) presents difference of MSPE (Focus - mSSA) for users with positive steps at $T + h$ and across different T. As T grows, the difference stays under the horizontal line ar 0 and decreases with T–indicating an empirically better performance of Focus with increasing T.

if a user is available and receives a nudge, and 0 otherwise. More details regarding the data preprocessing and implementation are deferred in App. H.

We evaluate forecast performance by training on the first T time points and testing on periods $T + 1$ to $T+5$. For each user i, we estimate $\theta_{i,T:T+5}$ i.e. 5-step-ahead mean potential steps under intervention with Focus. Performance is measured by the Mean squared prediction error (MSPE) denoted by

$$\text{MSPE} = \sum_{i:Y_{i,T+5}>0} (\hat{\theta}_{i,T:T+5} - Y_{i,T+5})^2 / Y_{i,T+5}^2,$$

computed for users with positive steps. Since SyN-BEATS require observations for post-treatment period and fully-observed panel for pre-treatment period, we only consider mSSA as our feasible benchmark.

Informative slot pairs. We designate slot pair $(4, 5)$ as *informative* due to strong correlations in the estimated factors across consecutive suggestion slots (Fig. 2). Certain factor components also show high cross-temporal correlation between slots 4 and 5 (Fig. 6(a)-(c) in Appendix). We exploit this association by predicting the outcome under a nudge at slot 5 using the estimated factors at slot 4. In line with autoregressive frameworks, where predictions at a given time are based on the immediately preceding time, we model the outcome at slot 5 under a nudge as depending on the outcome at slot 4. For simplicity, we assume that steps under a nudge at slot 4 on a day are independent of steps under a nudge at slot 5 on the previous day.

Results. Figure 2 shows that Focus consistently outperforms mSSA in forecasting future steps under intervention. For $T = 170, 200$, Focus achieves lower MSPE, and the performance gap widens as T increases (Figure 2(d)). Additional results for $T = 190$ (Fig. 6 in Appendix) confirm the same pattern. Taken together, these findings demonstrate that explicitly leveraging temporal latent dynamics is critical for accurate counterfactual forecasting, and that Focus provides a clear and scalable advantage over the benchmark mSSA.

6 Discussion

We propose a novel counterfactual forecasting method for panel data with low-rank structure and stochastic, time-varying latent factors. Under standard assumptions, we derive error bounds for the forecast estimator and construct valid confidence intervals. Empirically, our method consistently outperforms benchmarks when factors follow autoregressive dynamics, and a proof-of-concept on the HeartSteps V1 dataset highlights its practical utility in capturing temporal patterns in walking behavior.

A key limitation is the reliance on stationarity. Extending the framework to more flexible non-stationary settings, such as state space models with hidden Markov or Markov switching dynamics, is an important direction for future work. Further extensions include incorporating covariates, adapting the method to dynamic treatment regimes and sequential policies, and developing a doubly robust variant that can more effectively address missing observations and the latent factor dynamics.

References

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

[2] Anish Agarwal, Abdullah Alomar, and Devavrat Shah. On multivariate singular spectrum analysis and its variants. *arXiv preprint arXiv:2006.13448*, 2020.

[3] Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

[4] Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *The thirty sixth annual conference on learning theory*, pages 3821–3826. PMLR, 2023.

[5] Awad H Al-Mohy and Nicholas J Higham. Computing the fréchet derivative of the matrix exponential, with an application to condition number estimation. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1639–1657, 2009.

[6] Abdullah Alomar, Munther Dahleh, Sean Mann, and Devavrat Shah. Samossa: Multivariate singular spectrum analysis with stochastic autoregressive noise. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

[8] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

[9] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[10] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[11] Jushan Bai and Serena Ng. Rank regularized estimation of approximate factor models. *Journal of Econometrics*, 212(1):78–96, 2019.

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

[13] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4): 1535 – 1567, 2015. doi: 10.1214/15-AOS1315. URL `https://doi.org/10.1214/15-AOS1315`.

[14] Eli Ben-Michael, David Arbour, Avi Feller, Alexander Franks, and Steven Raphael. Estimating the effects of a california gun control program with multitask gaussian processes. *The Annals of Applied Statistics*, 17(2):985–1016, 2023.

[15] Ben S Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics*, 120(1):387–422, 2005.

[16] Falk Bräuning and Siem Jan Koopman. Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, 30(3):572–584, 2014.

[17] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer science & business media, 1991.

[18] Ercument Cahan, Jushan Bai, and Serena Ng. Factor-based imputation of missing values and covariances in panel data of large dimensions. *Journal of Econometrics*, 233(1):113–131, 2023.

[19] Yehu Chen, Annamaria Prati, Jacob Montgomery, and Roman Garnett. A multi-task gaussian process model for inferring time-varying treatment effects in panel data. In *International Conference on Artificial Intelligence and Statistics*, pages 4068–4088. PMLR, 2023.

[20] Jungjun Choi and Ming Yuan. Matrix completion when missing is not at random and its applications in causal panel data models. *Journal of the American Statistical Association*, pages 1–15, 2024.

[21] Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, pages 2590–2617, 2017.

[22] Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, 164 (1):188–205, 2011.

[23] Raaz Dwivedi, Katherine Tian, Sabina Tomkins, Predrag Klasnja, Susan Murphy, and Devavrat Shah. Counterfactual inference for sequential experiments. *arXiv preprint arXiv:2202.06891*, 2022.

[24] Raaz Dwivedi, Katherine Tian, Sabina Tomkins, Predrag Klasnja, Susan Murphy, and Devavrat Shah. Doubly robust nearest neighbors in factor models. *arXiv preprint arXiv:2211.14297*, 2022.

[25] Jacob Goldin, Julian Nyarko, and Justin Young. Forecasting algorithms for causal inference with panel data. *arXiv preprint arXiv:2208.03489*, 2022.

[26] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277, 2021.

[27] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

[28] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[29] Trevor J Hastie. Generalized additive models. *Statistical models in S*, pages 249–307, 2017.

[30] Erich Häusler and Harald Luschgy. *Stable convergence and stable limit theorems*, volume 74. Springer, 2015.

[31] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[32] Sainan Jin, Ke Miao, and Liangjun Su. On factor models with random missing: Em estimation, inference, and cross validation. *Journal of Econometrics*, 222(1):745–777, 2021.

[33] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[34] B.M.J.P. Jungbacker and S.J. Koopman. Likelihood-based analysis for dynamic factor models. WorkingPaper 08-007/4, Tinbergen Instituut (TI), 2008.

[35] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

[36] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

[37] Jiahe Lin and George Michailidis. Regularized estimation of high-dimensional factor-augmented vector autoregressive (favar) models. *Journal of machine learning research*, 21(117):1–51, 2020.

[38] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[39] Whitney K Newey and Kenneth D West. Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787, 1987.

[40] Xun Pang, Licheng Liu, and Yiqing Xu. A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2): 269–288, 2022.

[41] Pilar Poncela, Esther Ruiz, and Karen Miranda. Factor extraction using kalman filter and smoothing: This is not just another survey. *International Journal of Forecasting*, 37(4):1399–1425, 2021.

[42] Jeff Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of econometrics*, 99(1):39–61, 2000.

[43] David R Roberts, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[44] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1):41–55, 1983.

[45] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

[46] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[47] Jie Wei and Hui Chen. Determining the number of factors in approximate factor models by twice k-fold cross validation. *Economics Letters*, 191: 109149, 2020.

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

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix for "Counterfactual Forecasting For Panel Data"

Table of Contents for the Appendix

Additional notation. For any matrix M, $\mathrm{vec}(M)$ denotes the column-wise vectorization of M. For matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, $A \otimes B \in \mathbb{R}^{mp\times nq}$ denotes the Kronecker product defined blockwise as $A \otimes B = [A_{i,j}B]_{i\in[m],j\in[n]}$. We write $X_n = o_{\mathbb{P}}(a_n)$ if $\frac{X_n}{a_n} \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

A Proof of Thm. 4.1: Error bound for $\hat{\theta}_{i,T:T+h}$

We denote $H \triangleq \frac{1}{NT}\hat{D}^{-1}\hat{F}^\top \hat{F}\Lambda^\top\Lambda$. In words, H is a non-singular matrix that aligns $\hat{F}_t$, $\hat{\Lambda}_i$ with F_t and Λ_i respectively. For $i \in [N]$ and $h \geq 1$, we also denote:

$$\begin{aligned}
\Delta_{\Lambda,i} &\triangleq \hat{\Lambda}_i - (H^\top)^{-1}\Lambda_i,\\
\Delta_{F,T} &\triangleq \hat{F}_T - HF_T,\\
\Delta_{A,h} &\triangleq \hat{A}^h - HA^hH^{-1} = \hat{A}^h - (HAH^{-1})^h.
\end{aligned}$$

The h-step forecast error is decomposed as

$$\begin{aligned}
\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h} =\ & \hat{\Lambda}_i^\top \hat{A}^h\hat{F}_T - {\Lambda_i}^\top A^hF_T\\
=\ & \left[\Delta_{\Lambda,i} + (H^\top)^{-1}\Lambda_i\right]^\top\left[\Delta_{A,h} + HA^hH^{-1}\right]\left[\Delta_{F,T} + HF_T\right] - {\Lambda_i}^\top A^hF_t\\
=\ & \Delta_{\Lambda,i}^\top\Delta_{A,h}\Delta_{F,T} + \Delta_{\Lambda,i}^\top\Delta_{A,h}(HF_T) + \Delta_{\Lambda,i}(HAH^{-1})^h\Delta_{F,T}\\
& + ((H^\top)^{-1}\Lambda_i)^\top\Delta_{A,h}\Delta_{F,T} + \Delta_{\Lambda,i}^\top(HAH^{-1})^h(HF_T)\\
& + ((H^\top)^{-1}\Lambda_i)^\top\Delta_{A,h}(HF_T) + ((H^\top)^{-1}\Lambda_i)^\top(HAH^{-1})^h\Delta_{F,T}.
\end{aligned} \tag{18}$$

By Lem. D.1, Assum. 1 to Assum. 5 imply that the general assumptions Assum. 7 to 9 hold. From Bai [9, Lemma A.2], under Assum. 5 and Assum. 7 to 9 we can show:

$$\begin{aligned}
\|\Delta_{F,T}\|_2 =\ & \|\hat{F}_T - HF_T\|_2\\
=\ & \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{T}\delta_{NT}}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N}\delta_{NT}}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N}}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N}\delta_{NT}}\right)\\
=\ & \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big).
\end{aligned}$$

Additionally, similar to Xiong and Pelger [48, Thm. 2.2], we can show that $\|\Delta_{\Lambda,i}\|_2 = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big)$. Lem. A.2 implies, $\|\Delta_{A,1}\| = \mathcal{O}_{\mathbb{P}}\big(N^{-1}\big) + \mathcal{O}_{\mathbb{P}}\big(T^{-1/2}\big)$. For any horizon h, Taylor's expansion implies

$$\begin{aligned}
\left\|\hat{A}^h - HA^hH^{-1}\right\| =\ & \left\|(\Delta_{A,h} + HAH^{-1})^h - (HAH^{-1})^h\right\|\\
\leq\ & \left\|h\Delta_{A,h}HA^{h-1}H^{-1}\right\| + o_{\mathbb{P}}(1)\\
=\ & \mathcal{O}_{\mathbb{P}}\big(h\|A\|^{h-1}N^{-1}\big) + \mathcal{O}_{\mathbb{P}}\Big(h\|A\|^{h-1}T^{-1/2}\Big).
\end{aligned}$$

From Bai [9, Lemma A.3], $\|H\| = \mathcal{O}_{\mathbb{P}}(1)$. By Assum. 7 $\|\Lambda_i\|_2 = \mathcal{O}_{\mathbb{P}}(1)$, and $\|F_T\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. Therefore, the h-step forecast error is

$$\left|\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h}\right| = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big) + \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-1}\big) + \mathcal{O}_{\mathbb{P}}\big(h\|A\|^{h-1}N^{-1}\big) + \mathcal{O}_{\mathbb{P}}\Big(h\|A\|^{h-1}T^{-1/2}\Big),$$

and we are done. □

The proof invokes the following lemmata.

Lemma A.1 (Error bound on empirical autocovariance)**.** *Consider a factor model* (4) *with N units and T time points satisfying Assum. 5 and Assum. 7 to 9. Define the following matrices:*

$$\tilde{\Gamma}(\ell) \triangleq \frac{1}{T-\ell}\sum_{t=\ell+1}^{T} HF_t(HF_{t-\ell})^\top, \quad \hat{\Gamma}(\ell) \triangleq \frac{1}{T-\ell}\sum_{t=\ell+1}^{T}\hat{F}_t\hat{F}_{t-\ell}^\top, \quad \ell \in \{0,1\}.$$

Then we have $\left\|\hat{\Gamma}(\ell) - \tilde{\Gamma}(\ell)\right\| = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big)$.

Lemma A.2 (Error bound on $\hat{A}$)**.** *Consider the setup of Lem. A.2. Then the following holds:*

$$\|\hat{A} - HAH^{-1}\| = \mathcal{O}_{\mathbb{P}}\big(N^{-1}\big) + \mathcal{O}_{\mathbb{P}}\Big(T^{-1/2}\Big).$$

Remark A.1. Similar to Lem. A.2, $\mathcal{O}_{\mathbb{P}}\big(N^{-1} + T^{-1/2}\big)$ rate is obtained when $\hat{A}$ is estimated from fully observed panel with PCA-estimated factors under VAR(1) assumption[22, Prop. 3].

We now prove Lem. A.1 and A.2.

A.1 Proof of Lem. A.1: Error bound on empirical autocovariance

We have the following:

$$\begin{aligned}
&\left\|\hat{\Gamma}(\ell) - \tilde{\Gamma}(\ell)\right\| \\
\leq\ &\left\|\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} \hat{F}_t\hat{F}_{t-\ell}^{\top} - \frac{1}{T-\ell}\sum_{t=\ell+1}^{T} HF_t(HF_{t-\ell})^{\top}\right\| \\
\leq\ &\left\|\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} (\hat{F}_t - HF_t)(\hat{F}_{t-\ell} - HF_{t-\ell})^{\top}\right\| + 2\left\|\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} (\hat{F}_t - HF_t)F_{t-\ell}^{\top}\right\| \\
\leq\ &\left[\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} \left\|\hat{F}_t - HF_t\right\|_2^2\right]^{1/2}\left[\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} \left\|\hat{F}_{t-\ell} - HF_{t-\ell}\right\|_2^2\right]^{1/2} + 2\|H\|\left\|\frac{1}{T-\ell}\sum_{t=\ell+1}^{T} (\hat{F}_t - HF_t)F_{t-\ell}^{\top}\right\|.
\end{aligned}$$

Using Bai [9, Lemma A.3], $\|H\| = \mathcal{O}_{\mathbb{P}}(1)$. From Bai [9, Lemma A.1],

$$\frac{1}{T-\ell}\sum_{t=\ell+1}^{T}\left\|\hat{F}_t - HF_t\right\|_2^2 = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big), \quad \frac{1}{T-\ell}\sum_{t=\ell+1}^{T}\left\|\hat{F}_{t-\ell} - HF_{t-\ell}\right\|_2^2 = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big),$$

and Bai [9, Lemma A.2] implies $\frac{1}{T-\ell}\sum_{t=\ell+1}^{T}(\hat{F}_t - HF_t)F_{t-\ell}^{\top} = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big)$.Hence combining the three rates, we complete the proof. □

A.2 Proof of Lem. A.2: Error bound on $\hat{A}$

We use the fact that for any two invertible matrices A and B of the same dimension,

$$A^{-1} - B^{-1} = -A^{-1}(A-B)B^{-1} \quad \Longrightarrow \quad \|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\ \|A-B\|\ \|B^{-1}\|.$$

We use the notation of Lem. A.1. Denote

$$\tilde{A} \triangleq \tilde{\Gamma}(1)\tilde{\Gamma}(0)^{-1} = \left(\sum_{t=1}^{T-1} F_{t+1}F_t^{\top}\right)\left(\sum_{t=1}^{T-1} F_tF_t^{\top}\right)^{-1}. \tag{19}$$

We have

$$\begin{aligned}
\left\|\hat{A} - H\tilde{A}H^{-1}\right\| &= \left\|\hat{\Gamma}(1)\hat{\Gamma}(0)^{-1} - \tilde{\Gamma}(1)\tilde{\Gamma}(0)^{-1}\right\| \\
&= \left\|\hat{\Gamma}(1)\hat{\Gamma}(0)^{-1} - \tilde{\Gamma}(1)\hat{\Gamma}(0)^{-1} + \tilde{\Gamma}(1)\hat{\Gamma}(0)^{-1} - \tilde{\Gamma}(1)\tilde{\Gamma}(0)^{-1}\right\| \\
&\leq \left\|\hat{\Gamma}(1) - \tilde{\Gamma}(1)\right\|\left\|\hat{\Gamma}(0)^{-1}\right\| + \left\|\tilde{\Gamma}(1)\right\|\left\|\hat{\Gamma}(0)^{-1} - \tilde{\Gamma}(0)\right\| \\
&\leq \left\|\hat{\Gamma}(1) - \tilde{\Gamma}(1)\|\|\hat{\Gamma}(0)^{-1}\| + \|\tilde{\Gamma}(1)\|\|\hat{\Gamma}(0)^{-1}\|\|\hat{\Gamma}(0) - \tilde{\Gamma}(0)\|\|\tilde{\Gamma}(0)^{-1}\right\|.
\end{aligned}$$

Assum. 7 implies that there exists a positive definite matrix $\Sigma_F^{(\ell)}$ such that $\tilde{\Gamma}(\ell) = H\Sigma_F^{(\ell)}H^{\top} + o_{\mathbb{P}}(1)$. In addition, using Lem. A.1 we have

$$\hat{\Gamma}(\ell) = \tilde{\Gamma}(\ell) + \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big) = H\Sigma_F^{(h)}H^{\top} + o_{\mathbb{P}}(1).$$

Thus under Assum. 7, $\|\tilde{\Gamma}(h)^{-1}\| = \mathcal{O}_{\mathbb{P}}(1)$ and $\|\hat{\Gamma}(h)^{-1}\| = \mathcal{O}_{\mathbb{P}}(1)$, which implies

$$\left\|\hat{A} - H\tilde{A}H^{-1}\right\| = \mathcal{O}_{\mathbb{P}}\big(\delta_{NT}^{-2}\big).$$

From Lütkepohl [38, Prop. 3.1], $\|\tilde{A} - A\| = \mathcal{O}_{\mathbb{P}}\big(T^{-1/2}\big)$. Therefore, combining all the terms and using $H = \mathcal{O}_{\mathbb{P}}(1)$ [9, Lem. A.1], the result holds. □

B Proof of Thm. 4.2: Asymptotic normality of $\hat{\theta}_{i,T:T+h}$

Before we stat the proof, following we introduce some necessary shorthands for formulating the asymptotic variance of $\hat{\theta}_{i,T:T+h}$. For $i \in [N]$, we denote

$$\begin{aligned}
\mathcal{V}_\Lambda &\triangleq \mathbb{E}[\mathrm{vec}(\Lambda_i\Lambda_i^\top - \Sigma_\Lambda)\ \mathrm{vec}(\Lambda_i\Lambda_i^\top - \Sigma_\Lambda)^\top],\\
\Psi_T &\triangleq (F_T \otimes \Sigma_F)\ (\Sigma_F^{-1}\Sigma_\Lambda^{-1}),\\
\Upsilon_i &\triangleq (\Sigma_{F,i} \otimes \Sigma_\Lambda^{-1}\Lambda_i)\Sigma_{F,i}^{-1}.
\end{aligned}$$

We define the following covariance matrices that will be needed to express the asymptotic variance of $\hat{\theta}_{i,T:T+h}$:

$$\Sigma_F^{\mathrm{obs}} \triangleq \sigma_\varepsilon^2\Sigma_\Lambda^{-1}, \quad \Sigma_{F,T}^{\mathrm{miss}} \triangleq \Psi_T^\top\mathcal{V}_\Lambda\Psi_T, \quad \Sigma_{\Lambda,i}^{\mathrm{obs}} \triangleq \sigma_\varepsilon^2\Sigma_{F,i}^{-1}, \tag{20}$$

$$\Sigma_{\Lambda,i}^{\mathrm{miss}} \triangleq \Upsilon_i^\top\mathcal{V}_\Lambda\Upsilon_i, \quad \Sigma_{F,\Lambda,T,i}^{\mathrm{miss,cov}} \triangleq \Psi_T^\top\mathcal{V}_\Lambda\Upsilon_i. \tag{21}$$

Similar to (18), we write $\delta_{NT}(\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h})$ where each term is a product of at most 3 error terms. The product terms containing only one error term dominates the distribution and the rest of the terms are stochastically negligible. Therefore,

$$\begin{aligned}
\delta_{NT}(\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h}) = &\underbrace{\delta_{NT}((H^\top)^{-1}\Lambda_i)^\top(\hat{A}^h - (HAH^{-1})^h)(HF_T)}_{\text{(I)}}\\
&+ \underbrace{\delta_{NT}((H^\top)^{-1}\Lambda_i)^\top(HAH^{-1})^h(\hat{F}_T - HF_T)}_{\text{(II)}}\\
&+ \underbrace{\delta_{NT}(\hat{\Lambda}_i - (H^\top)^{-1}\Lambda_i)^\top(HAH^{-1})^h(HF_T)}_{\text{(III)}} \quad + o_{\mathbb{P}}(1).
\end{aligned}$$

From Lem. B.1,

$$\text{(I)} = \ \delta_{NT}\Lambda_i^\top H^{-1}\Big(\hat{A}^h - (HAH^{-1})^h\Big)HF_T = \delta_{NT}\Lambda_i^\top\big(H^{-1}\Delta_{A,h}H\big)F_T + o_{\mathbb{P}}(1). \tag{22}$$

From Lem. B.2,

$$\text{(II)} = \ \delta_{NT}\Lambda_i^\top A^h(H^{-1}\hat{F}_T - F_T) = \delta_{NT}((A^\top)^h\Lambda_i)^\top(e_{F,T}^{(1)} + e_{F,T}^{(2)}), \tag{23}$$

and

$$\text{(III)} = \ \delta_{NT}(H^\top\hat{\Lambda}_i - \Lambda_i)^\top(A^hF_T) = \delta_{NT}(e_{\Lambda,i}^{(1)} + e_{\Lambda,i}^{(2)})^\top(A^hF_T) + o_{\mathbb{P}}(1). \tag{24}$$

Additionally by Lem. B.2, every term in (I), (II) and (III) are asymptotically independent except $e_{F,T}^{(2)}$ and $e_{\Lambda,i}^{(2)}$. By Lem. B.3, the limiting asymptotic covariance term is $-2(\omega_2 - 1)\Sigma_{F,\Lambda,T,i}^{\mathrm{miss,cov}}$. Therefore, combining B.1, Lem. B.4 and an asymptotic argument similar to Bai [9, Eq. (C.1)], we obtain

$$\delta_{NT}\Omega_{i,T,h}^{-1}(\hat{\theta}_{i,T:T+h} - \theta_{i,T:T+h}) \xrightarrow{d} \mathcal{N}(0,1),$$

where the asymptotic variance term is $\Omega_{i,T,h} = \xi_{i,T,h}^2 + \tau_{i,T,h}^2$ with

$$\begin{aligned}\xi_{i,T,h}^2 \triangleq & \Lambda_i^\top A^h \Omega_T^{(F)} (A^\top)^h \Lambda_i + F_T^\top (A^\top)^h \Omega_i^\Lambda A^h F_T - 2(\omega_2 - 1)\Sigma_{F,\Lambda,T,i}^{\text{miss,cov}} \\ = & \frac{\delta_{NT}^2}{N} \Bigg[\Lambda_i^\top A^h \left\{\omega_1 \, \Sigma_F^{\text{obs}} + (\omega_1 - 1) \, \Sigma_{F,T}^{\text{miss}}\right\} (A^\top)^h \Lambda_i 2(\omega_2 - 1) \, (\Lambda_i^\top A^h) \, \Sigma_{F,\Lambda,T,i}^{\text{miss,cov}},\end{aligned} \tag{25}$$

and

$$\tau_{i,T,h}^2 \triangleq \Omega_{\Lambda_i,F_T}^{(A)} = \frac{\delta_{NT}^2}{T} \sum_{k,\ell=0}^{h-1} \big(\Lambda_i^\top (A^{h-1-k})^\top \Sigma_F^{-1} A^{h-1-\ell} \Lambda_i\big)\big(F_T^\top A^k \Sigma_\eta (A^\ell)^\top F_T\big), \tag{26}$$

where Σ_F^{obs}, $\Sigma_{F,T}^{\text{miss}}$, $\Sigma_{\Lambda,i}^{\text{obs}}$, $\Sigma_{\Lambda,i}^{\text{miss}}$ and $\Sigma_{F,\Lambda,T,i}^{\text{miss,cov}}$ are defined in (20) and (21). Hence the proof is complete. $\square$

Below we state some useful asymptotic normality results for the proof of Thm 4.2.

Lemma B.1 (Asymptotic normality of $\hat{A}$)**.** *Consider a factor model* (4) *with N units and T time points satisfying Assum. 1 to 5. Then for $\sqrt{T}/N \to 0$, $h \in \mathbb{N}$ and $u, v \in \mathbb{R}^r$,*

$$\delta_{NT} {\Omega_{u,v}^{(A)}}^{-1/2} \; u^\top (H^{-1} \hat{A}^h H - A^h) H v \xrightarrow{d} \mathcal{N}(0,1),$$

where $\Omega_{u,v}^{(A)} := \frac{\delta_{NT}^2}{T} \sum_{k,\ell=0}^{h-1} \big(v^\top [A^{h-1-k}]^\top \Sigma_F^{-1} A^{h-1-\ell} v\big) \cdot \big(u^\top A^k \Sigma_\eta [A^\ell]^\top u\big)$.

Lemma B.2 (Asymptotic independence)**.** *Consider a factor model* (4) *with N units and T time points satisfying Assum. 7 to 10. Then for $\sqrt{N}/T \to 0$ and $\sqrt{T}/N \to 0$,*

$$H^{-1} F_T = e_{F,T}^{(1)} + e_{F,T}^{(2)}, \quad H^\top \hat{\Lambda}_i - \Lambda_i = e_{\Lambda,i}^{(1)} + e_{\Lambda,i}^{(2)},$$

with $e_{F,T}^{(1)}$ and $e_{F,T}^{(2)}$ being asymptotically independent, $e_{\Lambda,i}^{(1)}$ and $e_{\Lambda,i}^{(2)}$ being asymptotically independent, and

$$\delta_{NT}^2 \mathrm{Cov}(e_{\Lambda,i}^{(2)}, e_{F,T}^{(2)}) \to \Sigma_\Lambda^{-1} \Sigma_F^{-1} [g_{i,T}^{\text{cov}}(u_i)]^\top \Sigma_{F,i}^{-1},$$

where $g_{i,T}^{\text{cov}}(u_i)$ is defined in (42) *in Assum. 10. Furthermore, each of $e_{F,T}^{(1)}$, $e_{F,T}^{(2)}$, $e_{\Lambda,i}^{(1)}$ and $e_{\Lambda,i}^{(2)}$ is asymptotically independent with $\Delta_{A,h}$.*

Lemma B.3 (Simplified form of asymptotic covariance)**.** *Under Assum. 1 to 6,*

$$\Sigma_\Lambda^{-1} \Sigma_F^{-1} [g_{i,T}^{\text{cov}}(u_i)]^\top \Sigma_{F,i}^{-1} = -2(\omega_2 - 1) \Sigma_{F,\Lambda,T,i}^{\text{miss,cov}}.$$

Lemma B.4 (Asymptotic normality of $\hat{F}_T$ and $\hat{\Lambda}_i$)**.** *Under the assumptions of Lem. B.2,*

$$\delta_{NT} {\Omega_T^{(F)}}^{-1/2} (H^{-1} \hat{F}_T - F_T) \xrightarrow{d} \mathcal{N}(0, I_r) \quad \textit{and} \quad \delta_{NT} {\Omega_i^{(\Lambda)}}^{-1/2} (H^\top \hat{\Lambda}_i - \Lambda_i) \xrightarrow{d} \mathcal{N}(0, I_r)$$

where $\Omega_T^{(F)} := \frac{\delta_{NT}^2}{N} \Sigma_\Lambda^{-1} \Sigma_F^{-1} [\Gamma_T^{\text{obs}} + \Gamma_T^{\text{miss}}] \Sigma_F^{-1} \Sigma_\Lambda^{-1}$ *and* $\Omega_i^{(\Lambda)} := \Sigma_{F,i}^{-1} \Big[\frac{\delta_{NT}^2}{T} \Phi_i^{\text{obs}} + \frac{\delta_{NT}^2}{N} \Phi_i^{\text{miss}}\Big] \Sigma_{F,i}^{-1}$. *Under Assum. 1 to 6, $\Omega_t^{(F)}$ and $\Omega_i^{(\Lambda)}$ are simplified as*

$$\Omega_T^{(F)} = \frac{\delta_{NT}^2}{T} \big[\omega_1 \Sigma_F^{\text{obs}} + (\omega_1 - 1) \Sigma_{F,T}^{\text{miss}}\big], \quad \Omega_i^{(\Lambda)} = \frac{\delta_{NT}^2}{T} \Sigma_{\Lambda,i}^{\text{obs}} + \frac{\delta_{NT}^2}{N} (\omega_3 - 1) \Sigma_{\Lambda,i}^{\text{miss}}, \tag{27}$$

where Σ_F^{obs}, $\Sigma_{F,T}^{\text{miss}}$, $\Sigma_{\Lambda,i}^{\text{obs}}$ and $\Sigma_{\Lambda,i}^{\text{miss}}$ are defined in Thm. 4.2.

The proofs of Lemmata B.2, B.3 and B.4 follow from Bai [9, Thm. 1 and Thm. 2] and Xiong and Pelger [48, Thm. 2, Corollary 1]. The proof techniques involve the roles of T and N exchanged in order to maintain the structure of Bai [9, Thm. 1 and Thm. 2] and conduct our theoretical analysis. The final statement of Lem. B.2 follows since the randomness in $\Delta_{A,h}$ is generated by the time average of $\eta_t, t = 2, \ldots, T$ and hence independent from the source of randomness of the other error terms. Hence a mutual asymptotic independence of the three terms hold. We only provide a proof of Lem. B.1.

B.1 Proof of Lemma B.1: Asymptotic normality of $\hat{A}$

Denote the following matrices $F^{(-T)} \triangleq [F_1 : \ldots : F_{T-1}]^\top_{(T-1)\times r}$ and $U \triangleq [\eta_2 : \ldots : \eta_T]$. From Lütkepohl [38, Prop. 3.1], we have

$$\tilde{A} - A = \left[\left(\frac{1}{T-1} F^{(-T)^\top} F^{(-T)}\right) \otimes I_r\right] \frac{1}{\sqrt{T-1}} \left(F^{(-T)^\top} \otimes I_r\right) \mathrm{vec}(U)$$

and the following holds:

$$\sqrt{T}\ \mathrm{vec}(\tilde{A} - A) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_F^{-1} \otimes \Sigma_\eta\big), \tag{28}$$

where $\tilde{A}$, defined in (19), is the OLS estimator of A with the true rotated factors F_t up to a rotation with H.

First we fix $h = 1$. Following A.2,

$$\mathrm{vec}(H^{-1}\hat{A}H - A) = \mathrm{vec}(H^{-1}\hat{A}H - \tilde{A}) + \mathrm{vec}(\tilde{A} - A) = \mathrm{vec}(\tilde{A} - A) + \mathcal{O}_{\mathbb{P}}\big(N^{-1}\big) + \mathcal{O}_{\mathbb{P}}\Big(T^{-1/2}\Big).$$

If $\sqrt{T}/N \to 0$, the second term dominates and contributes to the asymptotic distribution i.e.

$$\sqrt{T}\ \mathrm{vec}(H^{-1}\hat{A}H - A) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_F^{-1} \otimes \Sigma_\eta\big).$$

Next we consider the matrix valued operator $f(A) = \mathrm{vec}(A^h)$, $A \in \mathbb{R}^{r\times r}$. Following Al-Mohy and Higham [5, Thm. 3.1], $A \mapsto A^h$ has a Fréchet derivative in the directional matrix E as $D_{A^h}(A; E) = \sum_{k=0}^{h-1} A^k E A^{h-1-k}$, and hence the vectorized gradient operator is $\nabla f(A) = \sum_{k=0}^{h-1}\big((A^{h-1-k})^\top \otimes A^k\big)$. Therefore, applying multivariate Delta-method, the asymptotic variance term becomes $\Big[\sum_{k=0}^{h-1}\big((A^{h-1-k})^\top \otimes A^k\big)\Big](\Sigma_F^{-1} \otimes \Sigma_\eta)\Big[\sum_{k=0}^{h-1}\big((A^{h-1-k})^\top \otimes A^k\big)\Big]^\top$, and the rest follows from the standard properties of Kronecker products. □

B.2 HAC estimators of $\sigma^2_{i,T,h}$

Estimation of $\xi^2_{i,T,h}$. We estimate $\hat{\xi}^2_{i,T,h}$, a consistent estimator of $\xi^2_{i,T,h}$, via the plug-in approach. In (25), the plug-in estimators $\hat{\Lambda}_i$, $\hat{F}_T$, and $\hat{A}$ are obtained from $\textsc{Focus}(Y, W)$ as described in Sec. 3. Consistent estimators of $\hat{\Sigma}^{\mathrm{obs}}_F$ and $\hat{\Sigma}^{\mathrm{obs}}_{\Lambda,i}$ are constructed using HAC estimators [39], while $\hat{\Sigma}^{\mathrm{miss}}_{F,T}$, $\hat{\Sigma}^{\mathrm{miss}}_{\Lambda,i}$, $\hat{\Sigma}^{\mathrm{miss,cov}}_{F,\Lambda,T,i}$, and $\hat{\omega}_i$ rely on the same plug-in principle. Details can be found in Xiong and Pelger [48, Sec. 8]

Estimation of $\tau^2_{i,T,h}$. We can again use a plug-in approach for estimating $\tau^2_{i,T,h}$. We can use the estimators $\hat{\Sigma}_F = \frac{1}{T}\sum_{t=1}^T \hat{F}_t\hat{F}_t^\top$, $\hat{\Sigma}_\eta = \frac{1}{T-1}\sum_{t=2}^T(\hat{F}_t - \hat{A}\hat{F}_{t-1})(\hat{F}_t - \hat{A}\hat{F}_{t-1})^\top$, and the variance estimator can be calculated as

$$\hat{\tau}^2_{i,T} := \frac{\delta^2_{NT}}{T}\sum_{k,\ell=0}^{h-1}\Big(\hat{\Lambda}_i^\top(\hat{A}^{h-1-k})^\top\hat{\Sigma}_F^{-1}\hat{A}^{h-1-\ell}\hat{\Lambda}_i\Big)\Big(\hat{F}_T^\top\hat{A}^k\hat{\Sigma}_\eta(\hat{A}^\ell)^\top\hat{F}_T\Big).$$

A detailed discussion of consistency of the estimators $\hat{\Sigma}_F$ and $\hat{\Sigma}_\eta$ and the translation of asymptotic normality with the estimated variance can be found in Lütkepohl [38, Prop. 3.2, Cor. 3.2.1].

C Proof of Cor. 4.2

For one factor model $Y_{i,t} = \Lambda_i F_t + \varepsilon_{i,t}$ with AR(1) factors F_t with AR coefficient a and $|a| < 1$, Assum. 1 is satisfied. Next we verify that Assum. 5 and 6 hold under MCAR.

Implication of Assum. 5. For any $i \in [N]$ and $s, t \in [T]$, $W_{i,s}W_{i,t} \sim \mathrm{Bernoulli}(p^2)$ and are independent across i. Therefore $\frac{1}{N}|\mathcal{Q}_{s,t}| = \frac{1}{N}\sum_{i=1}^N W_{i,s}W_{i,t}$ is an average of N Bernoulli(p^2) random variables. Applying strong law of large numbers, Assum. 5 is satisfied with

$$\alpha_{s,t} = \begin{cases} p & \text{if } s = t, \\ p^2 & \text{otherwise.} \end{cases} \tag{29}$$

Similarly, $\frac{1}{N}|\mathcal{Q}_{s,t} \cap \mathcal{Q}_{s',t'}| = \frac{1}{N}\sum_{i=1}^{N} W_{i,s}W_{i,t}W_{i,s'}W_{i,t'}$. Again applying strong law of large numbers,

$$\beta_{s,t,s',t'} = p^{\#\text{distinct}(\{s,t,s',t'\})}, \tag{30}$$

where for any set S, $\#\text{distinct}(S) \triangleq$ Number of distinct elements in S.

Implication of Assum. 6. We fix any $s,t \in [T]$. Using (29) and (30), Assum. 6 is satisfied with

$$\nu_s = p^2, \quad \rho_{s,t} = \begin{cases} p^2 & \text{if } s = t, \\ p^2 & \text{otherwise,} \end{cases}$$
$$\omega_1 = 1/p, \quad \omega_2 = \omega_2 = 1.$$

We are remained to show that MCAR satisfies Assum. 6 in one-factor model with AR(1) factors. Since $|a| < 1$, for any fixed $i \in [N]$ we can write (16) as

$$\frac{1}{T}\sum_{t=1}^{T} W_{i,t}F_t^2 = \underbrace{\frac{1}{T}\sum_{t=1}^{T}(W_{i,t}-p)F_t^2}_{\triangleq A_T} + \underbrace{\frac{p}{T}\sum_{t=1}^{T} F_t^2}_{\triangleq B_T}.$$

Following from Brockwell and Davis [17, Thm. 11.2.2 and Eq. 11.2.6], $B_T \xrightarrow{\mathbb{P}} p\sigma_F^2$. Additionally, we have $\mathbb{E}[A_T] = 0$. Denote $Z_t \triangleq (W_{i,t}-p)F_t^2$. Thus, for any $s,t \in [T]$

$$\begin{aligned}\text{Cov}(Z_s, Z_t) = \mathbb{E}[Z_sZ_t] = \mathbb{E}[(W_{i,s}-p)(W_{i,t}-p)F_s^2F_t^2] =& \mathbb{E}[(W_{i,s}-p)(W_{i,t}-p)]\ \mathbb{E}[F_s^2F_t^2] \\ =& \begin{cases} p(1-p)\mathbb{E}[F_t^4] & \text{if } s = t \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Therefore under Assum. 2,

$$\text{Var}(A_T) = \frac{1}{T^2}\sum_{s,t=1}^{T}\text{Cov}(Z_s,Z_t) = \frac{1}{T^2}\sum_{t=1}^{T}\text{Var}(Z_t) = \frac{1}{T}p(1-p)\mathbb{E}[F_1^4] = \mathcal{O}(T^{-1}).$$

Using *Chebychev's inequality* for any $\delta > 0$, $\mathbb{P}(A_T > \delta) \le \text{Var}(A_T)/\delta^2 = \mathcal{O}(T^{-1}) \to 0$ as $T \to \infty$. Therefore $A_T \xrightarrow{\mathbb{P}} 0$, and we obtain $\Sigma_{F,i} = p\sigma_F^2$. Hence the asymptotic variances in (25) and (26) become

$$\xi_{i,T,h}^2 = \frac{\delta_{NT}^2 a^{2h}}{N}\left[\frac{\Lambda_i^2\sigma_\varepsilon^2}{p\sigma_\Lambda^2} + \left(\frac{1}{p}-1\right)\mathbb{E}\left\{\left(\frac{\Lambda_i^2}{\sigma_\Lambda^2}-1\right)^2\right\}\Lambda_iF_T\right] + \frac{\delta_{NT}^2 a^{2h}F_T^2\sigma_\varepsilon^2}{Tp\sigma_F^2}, \tag{31}$$

$$\tau_{i,T,h}^2 = \frac{\delta_{NT}}{T}\ h^2a^{2h-2}(1-a^2)\Lambda_i^2F_T^2. \tag{32}$$

Hence the proof is complete. □

D General assumptions

We now state the general set of assumptions on the model 4. The general assumptions impose general structure on the temporal and cross sectional dependence of the factors, loadings and the idiosyncratic errors. Under these assumptions, we prove Thm. 4.1. In all the following, M, M' denote universal positive constants that is allowed to change values from one line to another.

Assumption 7 (**Factor model with general structure**)**.** *For every i in$[N], t \in [T]$, $\mathbb{E}[\|F_t\|_2^4] \le M$ and $\mathbb{E}[\|\Lambda_i\|_2^4] \le M$. For fixed $0 \le h < T$, there exists a positive definite matrix $\Sigma_F^{(h)}$ such that*

$$\mathbb{E}\left[\left\|\sqrt{T-h}\left(\frac{1}{T-h}\sum_{t=h+1}^{T} F_tF_{t-h}^\top - \Sigma_F^{(h)}\right)\right\|^2\right] \le M. \tag{33}$$

There exists a positive definite matrix Σ_Λ such that

$$\mathbb{E}\left[\left\|\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top-\Sigma_\Lambda\right)\right\|^2\right]\le M, \tag{34}$$

and for every $\mathcal{Q}_{s,t}\subset[N]$,

$$\mathbb{E}\left[\left\|\sqrt{|\mathcal{Q}_{s,t}|}\left(\frac{1}{|\mathcal{Q}_{s,t}|}\sum_{i\in\mathcal{Q}_{s,t}}^{N}\Lambda_i\Lambda_i^\top-\Sigma_\Lambda\right)\right\|^2\right]\le M. \tag{35}$$

The eigenvalues of $\Sigma_F^{(0)}\Sigma_\Lambda$ are distinct and strictly positive.

Assumption 8 (**Idiosyncratic errors with general strucuture**). *For $i\in[N], t\in[T]$, $\mathbb{E}[\varepsilon_{i,t}]=0$, $\mathbb{E}|\varepsilon_{i,t}|^8\le M$. The autocovariance function $\gamma(s,t)\triangleq\mathbb{E}\Big[\frac{1}{|\mathcal{Q}_{s,t}|}\varepsilon_t^\top\varepsilon_s\Big]=\mathbb{E}\Big[\frac{1}{|\mathcal{Q}_{s,t}|}\sum_{i\in\mathcal{Q}_{s,t}}\varepsilon_{i,t}\varepsilon_{i,s}\Big]$, satisfies $|\gamma(s,s)|\le M,\quad\sum_{s=1}^{T}|\gamma(s,t)|\le M$. For every $s,t\in[T]$,*

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}[\varepsilon_{i,s}\varepsilon_{i,t}-\mathbb{E}(\varepsilon_{i,s}\varepsilon_{i,t})]\right|^4\right]\le M. \tag{36}$$

Assumption 9 (**Moment conditions**). *For every $s\in[T]$ and $0\le h<T$,*

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i\varepsilon_{i,t}\right\|_2^2\right]\le M, \tag{37}$$

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_{t-h}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}(\varepsilon_{i,t}\varepsilon_{i,s}-\mathbb{E}[\varepsilon_{i,t}\varepsilon_{i,s}])\right\|^2\right]\le M, \tag{38}$$

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_iF_{t-h}^\top\varepsilon_{i,t}\right\|^2\right]\le M. \tag{39}$$

Assumption 10 (**Asymptotic distributions**). *There exists a positive definite matrix $\Sigma_{F,i}$ such that (16) holds. In addition, there exists a positive definite matrix Γ_T^{obs} such that*

$$\frac{\sqrt{N}}{T}\sum_{s=1}^{T}F_sF_s^\top\frac{1}{|\mathcal{Q}_{s,T}|}\sum_{i\in\mathcal{Q}_{s,T}}\Lambda_i\varepsilon_{i,T}\xrightarrow{d}\mathcal{N}\big(0,\Gamma_T^{\text{obs}}\big). \tag{40}$$

For every $i\in[N]$, there exists a positive definite matrix Φ_i^{obs} such that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}W_{i,t}F_t\varepsilon_{i,t}\xrightarrow{d}\mathcal{N}\big(0,\Phi_i^{\text{obs}}\big). \tag{41}$$

Denote the quantities $u_i\triangleq\Sigma_F^{-1}\Sigma_\Lambda^{-1}\Lambda_i$, $J_t\triangleq\frac{1}{T}\sum_{s=1}^{T}F_sF_s^\top\Big(\frac{1}{|\mathcal{Q}_{s,t}|}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i\Lambda_i^\top-\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top\Big)$, and $R_i\triangleq\frac{1}{T}\sum_{t=1}^{T}W_{i,t}J_tF_tF_t^\top$. Then

$$\sqrt{N}\Sigma_{i,T}^{(J,R)^{-1/2}}\begin{bmatrix}J_TF_T\\R_iu_i\end{bmatrix}\xrightarrow{d}\mathcal{N}(0,I_r), \tag{42}$$

where $\Sigma_{i,T}^{(J,R)} \triangleq \begin{bmatrix} \Sigma_{J,T} & g_{i,T}^{\text{cov}}(u_i)^\top \\ g_{i,T}^{\text{cov}}(u_i) & h_i(u_i) \end{bmatrix}$ *for positive definite matrix* $\Sigma_{J,T}$*, and functions* $g_{i,T}^{\text{cov}}$ *and* h_i*.*

Under the model assumptions in the paper, the general assumptions follow. We use the general assumptions to prove Thm. 4.1. The implication of the simplified assumptions by more general set of assumptions is stated as the following lemma:

Lemma D.1 (Implication by simplified assumptions). *The general assumptions 7 to 10 are satisfied by the simplified model assumptions 1 to 6.*

D.1 Proof of Lemma D.1: Implication by simplified assumptions

We state some sub-lemmata to validate that each of the general assumptions are satisfied under the simplified assumptions.

Lemma D.2 (Power norm bounds under spectral radius condition). *Let* $A \in \mathbb{R}^{r\times r}$ *be a square matrix with spectral radius* $\rho(A) < 1$*. Then there exist an integer* $\underline{N} \geq 1$*, both depending on* A*, such that*

$$\|A^n\| < \left(\frac{1+\rho(A)}{2}\right)^n \quad \textit{for all } n \geq \underline{N}.$$

In particular, the series $\sum_{n=0}^{\infty} \|A^n\|$ *and* $\sum_{n=0}^{\infty} n\|A^n\|$ *converge.*

Lemma D.3. *Assum. 1, 2 and 5* $\implies$ *Assum. 7.*

Lemma D.4. *Assum. 3* $\implies$ *Assum. 8*

Lemma D.5. *Assum. 1 to 4* $\implies$ *Assum. 9*

Lemma D.6. *Assum. 1 to 6* $\implies$ *Assum. 10*

Now we provide proofs of the stated lemmata.

D.2 Proof of Lemma D.2: Power norm bounds under spectral radius condition

This lemma uses Gelfand's formula from matrix algebra [31, Cor. 5.6.14] that states the following: $\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$. Thus for any $\epsilon > 0$, there exists $N(\epsilon) \in \mathbb{N}$ such that

$$\left| \|A^n\|^{1/n} - \rho(A) \right| < \epsilon \quad \text{for } n > N(\epsilon) \qquad \implies \qquad \|A^n\| < (\rho(A)+\epsilon)^n \quad \text{for } n > N(\epsilon).$$

Choose $\epsilon = \epsilon(A) \triangleq (1-\rho(A))/2$ and define $\tilde{\rho}(A) \triangleq (1+\rho(A))/2 < 1$. We also denote $\underline{N} \triangleq N(\epsilon(A))$. Therefore,

$$\|A^n\| < \tilde{\rho}(A)^n \qquad \text{for } n > \underline{N}.$$

Hence,

$$\begin{aligned} \sum_{n=0}^{\infty} \|A^n\| = \sum_{n\leq \underline{N}} \|A^n\| + \sum_{n\leq \underline{N}} \|A^n\| &< \sum_{n\leq \underline{N}} \|A^n\| + \sum_{n>\underline{N}} \tilde{\rho}(A)^n \\ &= \sum_{n\leq \underline{N}} \|A^n\| + \frac{\tilde{\rho}(A)^{\underline{N}+1}}{1-\tilde{\rho}(A)} < \infty, \end{aligned}$$

since the first term is a finite sum, and for the the second term $\tilde{\rho}(A) < 1$. Similarly,

$$\sum_{n=\underline{N}}^{\infty} n\|A^n\| < \frac{\tilde{\rho}(A)^{\underline{N}}}{1-\tilde{\rho}(A)} (\underline{N}(1-\tilde{\rho}(A)) + \tilde{\rho}(A)) < \infty.$$

Hence the proof is done. ∎

D.3 Proof of Lem. D.3: Assum. 1, 2 and 5 $\implies$ Assum. 7

From Assum. 1, $\rho(A) < 1$. Lem. D.2 implies that F_t has the following moving average (MA) representation

$$F_t = \sum_{j=0}^{\infty} A^j \eta_{t-j}. \tag{43}$$

Additionally we can bound the fourth moment by Minkowski's inequality as follows

$$(\mathbb{E}\|F_t\|_2^4)^{1/4} = \left(\mathbb{E}\left[\left\|\sum_{j=0}^{\infty} A^j \eta_{t-j}\right\|_2^4\right]\right)^{1/4} = \sum_{j=0}^{\infty}\big(\mathbb{E}\big[\|A^j\eta_{t-j}\|_2^4\big]\big)^{1/4} \le \sum_{j=0}^{\infty}\big(\|A^j\|\mathbb{E}\big[\|\eta_{t-j}\|_2^4\big]\big)^{1/4} \le M\sum_{j=0}^{\infty}\|A^j\|.$$

Lem. D.2 implies that the infinite sum converges. Hence there exists $M > 0$ such that $\mathbb{E}[\|F_t\|_2^4] \le M$. Assum. 2 also poses $\mathbb{E}[\|\Lambda_i\|_2^4] \le M$ for every $i \in [N]$ that is satisfied by Assum. 7.

Denote the autocovariance function of the factors F_t as $\Gamma_F(h) \triangleq \mathrm{Cov}(F_t, F_{t-h})$. Then by Brockwell and Davis [17, Thm. 11.2.2 and Eq. 11.2.6], we have

$$\frac{1}{T-h}\sum_{t=h+1}^{T} F_t F_{t-h}^\top = \Gamma_F(h) + o_{\mathbb{P}}(T^{-1/2}),$$

Hence (33) holds with $\Sigma_F^{(h)} = \Gamma_F(h)$. As Λ_i are i.i.d. in Assum. 2, (34) holds. Additionally (35) holds under (13) in Assum. 5 that completes the proof. $\square$

D.4 Proof of Lem. D.4: Assum. 3 $\implies$ Assum. 8

Assum. 3 directly implies $\mathbb{E}[\varepsilon_{it}] = 0$ and $\mathbb{E}|\varepsilon_{it}|^8 \le M$. For $s \ne t$, $\gamma(s,t) = |\mathcal{Q}_{st}|^{-1}\mathbb{E}[\varepsilon_t^\top \varepsilon_s] = 0$. Furthermore, $\gamma(s,s) = |\mathcal{Q}_{ss}|^{-1}\mathbb{E}[\varepsilon_s^\top \varepsilon_s] = N^{-1}\sum_{i=1}^N \mathbb{E}[\varepsilon_{is}^2] = \sigma_\varepsilon^2 < \infty$ by Assum. 3. In addition, $\sum_{s=1}^T \gamma(s,t) = \gamma(s,s) = \sigma_\varepsilon^2$ which is finite. For fixed $s,t \in [T]$, denote $v_i^{(s,t)} \triangleq |\varepsilon_{i,s}\varepsilon_{i,t} - \mathbb{E}(\varepsilon_{i,s}\varepsilon_{i,t})|$. Therefore $\mathbb{E}[v_i^{(s,t)}] = 0$ and $\mathbb{E}[(v_i^{(s,t)})^4] < \infty$ by Assum. 3. We also have the following

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}[\varepsilon_{i,s}\varepsilon_{i,t} - \mathbb{E}(\varepsilon_{i,s}\varepsilon_{i,t})]\right|^4\right] &= \frac{1}{|\mathcal{Q}_{s,t}|^2}\mathbb{E}\left(\sum_{i\in\mathcal{Q}_{s,t}} v_i^{(s,t)}\right)^4 \\
&= \frac{1}{|\mathcal{Q}_{s,t}|^2}\left(|\mathcal{Q}_{s,t}|\mathbb{E}\left[\left(v_1^{(s,t)}\right)^4\right] + 3|\mathcal{Q}_{s,t}|(|\mathcal{Q}_{s,t}|-1)\mathbb{E}\left[\left(v_1^{(s,t)}\right)^2\right]^2\right) \\
&\le |\mathcal{Q}_{s,t}|^{-1}M + 3(1 - |\mathcal{Q}_{s,t}|^{-1})M \le 3M.
\end{aligned}$$

Hence the proof is done. $\square$

D.5 Proof of Lemma D.5: Assum. 1 to 4 $\implies$ Assum. 9

Next we show that (37) holds. We have the following

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i\varepsilon_{i,t}\right\|_2^2\right] &= \frac{1}{|\mathcal{Q}_{s,t}|}\mathbb{E}\left[\left(\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i^\top\varepsilon_{i,t}\right)\left(\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i\varepsilon_{i,t}\right)\right] \\
&= \frac{1}{|\mathcal{Q}_{s,t}|}\mathbb{E}\left[\sum_{i,j\in\mathcal{Q}_{s,t}}\Lambda_i^\top\Lambda_j\varepsilon_{i,t}\varepsilon_{j,t}\right] \\
&= \frac{1}{|\mathcal{Q}_{s,t}|}\cdot|\mathcal{Q}_{s,t}|\ \mathbb{E}[\|\Lambda_1\|_2^2]\mathbb{E}[\varepsilon_{1,t}^2] \ \le\ \sigma_\varepsilon^2 M,
\end{aligned}$$

where we use Assum. 2 and 3 in the last step. Next we show that (38) holds. We additionally have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}(\varepsilon_{i,s}\varepsilon_{i,t}-\mathbb{E}[\varepsilon_{i,s}\varepsilon_{i,t}])\right\|_2^2\right]\\
=&\frac{1}{T}\mathbb{E}\left[\left(\sum_{t=1}^{T}F_t^\top\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}v_i^{(s,t)}\right)\left(\sum_{t=1}^{T}F_t\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}v_i^{(s,t)}\right)\right]\\
=&\frac{1}{T}\sum_{t,t'=1}^{T}\mathbb{E}[F_t^\top F_{t'}]\cdot\frac{1}{\sqrt{|\mathcal{Q}_{s,t}||\mathcal{Q}_{s,t'}|}}\;\mathbb{E}\left[\sum_{i\in\mathcal{Q}_{s,t},\ j\in\mathcal{Q}_{s,t'}}v_i^{(s,t)}v_i^{(s,t')}\right]
\end{aligned}\tag{44}
$$

Under the general assumption 7, there exists $M>0$ such that $\mathbb{E}[F_t^\top F_{t'}]\le\mathbb{E}[\|F_t\|_2^2]\le M$. We denote $A_{s,t,t'}\triangleq\mathbb{E}[\sum_{i\in\mathcal{Q}_{s,t},j\in\mathcal{Q}_{s,t'}}v_{i,s,t}v_{i,s,t'}]$. We consider the following cases:

(Case 1.) If $t=t'=s$,

$$
\begin{aligned}
A_{s,t,t'}=&\mathbb{E}\sum_{i,j\in\mathcal{Q}_{ss}}(\varepsilon_{is}^2-\sigma_\varepsilon^2)(\varepsilon_{js}^2-\sigma_\varepsilon^2)\\
=&\mathbb{E}\left[\sum_{i=1}^{N}(\varepsilon_{is}^2-\sigma_\varepsilon^2)\right]^2=\mathbb{E}\left[\sum_{i=1}^{N}(\varepsilon_{is}^2-\sigma_\varepsilon^2)^2\right]\le NM'
\end{aligned}
$$

for some $M'>0$.

(Case 2.) If $t=t'\neq s$,

$$
A_{s,t,t'}=\sum_{i,j\in\mathcal{Q}_{s,t}}\mathbb{E}[\varepsilon_{i,s}\varepsilon_{j,s}]\mathbb{E}[\varepsilon_{i,t}\varepsilon_{j,t}]=|\mathcal{Q}_{s,t}|\sigma_\varepsilon^4.
$$

(Case 3.) If $t=s,t'\neq s$,

$$
A_{s,t,t'}=\mathbb{E}\left[\sum_{1\le i\le N,j\in\mathcal{Q}_{s,t'}}(\varepsilon_{i,s}^2-\sigma_\varepsilon^2)\varepsilon_{j,s}\varepsilon_{j,t'}\right]=0.
$$

(Case 4.) If $t\neq s,t'\neq s$,

$$
A_{s,t,t'}=\mathbb{E}\left[\sum_{i\in\mathcal{Q}_{s,t},j\in\mathcal{Q}_{s,t'}}\varepsilon_{is}\varepsilon_{i,t}\varepsilon_{j,s}\varepsilon_{j,t'}\right]=0.
$$

Combining all cases in (44), we obtain the following for some M and $M'>0$:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}(\varepsilon_{i,s}\varepsilon_{i,t}-\mathbb{E}[\varepsilon_{i,s}\varepsilon_{i,t}])\right\|_2^2\right]\le\;&\frac{1}{T}M\cdot\frac{1}{N}\cdot M'N+\frac{1}{T}M\cdot\sigma_\varepsilon^4(T-1)\\
\le\;&\frac{1}{T}M[M'+\sigma_\varepsilon^4(T-1)]\\
\le\;&M[M'+\sigma_\varepsilon^4].
\end{aligned}
$$

Finally we show that (39) holds by the mutual independence Assum. 4 as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i F_t^\top \varepsilon_{i,t}\right\|^2\right] \\
\leq\ & \mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i F_t^\top \varepsilon_{i,t}\right\|_F^2\right] \\
=\ & \mathbb{E}\left[\mathrm{Tr}\left\{\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}\Lambda_i F_t^\top \varepsilon_{i,t}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}|}}\sum_{i\in\mathcal{Q}_{s,t}}F_t \Lambda_i^\top \varepsilon_{i,t}\right)\right\}\right] \\
=\ & \mathrm{Tr}\left\{\frac{1}{T}\sum_{t,t'=1}^{T}\frac{1}{\sqrt{|\mathcal{Q}_{s,t}||\mathcal{Q}_{s,t'}|}}\sum_{i\in\mathcal{Q}_{s,t},j\in\mathcal{Q}_{s,t'}}\mathbb{E}[\|F_t\|_2^2]\Sigma_\Lambda \mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,t'})\right\} \\
\leq\ & \frac{1}{T}\times T\times \mathrm{Tr}(\Sigma_\Lambda)\mathbb{E}[\|F_t\|_2^2]\sigma_\varepsilon^2 \ \leq\ r^2\|\Sigma_\Lambda\|\ \mathbb{E}[\|F_t\|_2^2]\ \sigma_\varepsilon^2.
\end{aligned}$$

By Assum. 2 and 3, there exists $M > 0$ such that $\|\Sigma_\Lambda\| \leq M$, and $\sigma_\varepsilon^2 \leq M$. Additionally, Assum. 7 implies that $\mathbb{E}[\|F_t\|_2^2] \leq M$. Since r is fixed, we are done with the proof. □

D.6 Proof of Lemma D.6: Assum. 1 to 6 $\Longrightarrow$ Assum. 10

(16) in Assum. 6 is already implied in Assum. 10. We now show the asymptotic distributions in Assum. 10 hold under the simple conditions.

D.6.1 Proof of (40)

First we denote the Sigma-algebra $\mathcal{F}_T \triangleq \sigma\big(\{F_t\}_{t\in[T]}, \{\alpha_{t,T}\}_{t\in[T]}, \{\beta_{s,T,t,T}\}_{s,t\in[T]}\big)$ where $\alpha_{s,t}$ and $\beta_{s,t,s',t'}$ are defined for $s,t,s',t' \in [T]$ in Assum. 5. For $s \in [T]$,

$$Z_{s,T} \triangleq \frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}\sum_{i\in\mathcal{Q}_{s,T}}\Lambda_i\varepsilon_{i,T} = \frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}\sum_{i=1}^{N}W_{i,s}W_{i,T}\Lambda_i\varepsilon_{i,T}.$$

Using *Lindeberg's central limit theorem*, we show asymptotic normality of $Z_{s,T}$ stably on $\mathcal{F}_T$ as $N \to \infty$. We verify that the conditions of Lindeberg's CLT are satisfied:

1. *Zero mean.* Using independence of Λ_i's and $\varepsilon_{i,T}$'s in 4,

$$\mathbb{E}[Z_{s,T} \mid \mathcal{F}_T] = \sum_{i\in\mathcal{Q}_{s,T}}\mathbb{E}\left[\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}\Lambda_i\varepsilon_{i,T}\middle|\mathcal{F}_T\right] = \sum_{i\in\mathcal{Q}_{s,T}}\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}\mathbb{E}[\Lambda_i]\mathbb{E}[\varepsilon_{i,T}] = 0. \tag{45}$$

2. *Bounded covariance matrix.* The covariance matrix is

$$\mathrm{Cov}(Z_{s,T} \mid \mathcal{F}_T) = \sum_{i\in\mathcal{Q}_{s,T}}\mathrm{Cov}\left(\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}\Lambda_i\varepsilon_{i,T}\middle|\mathcal{F}_T\right) = \mathrm{Cov}(\Lambda_i\varepsilon_{i,T}) = \sigma_\varepsilon^2\Sigma_\Lambda,$$

which is bounded with T.

3. *Lindeberg's condition.* Under Assum. 5 and for any $\delta > 0$,

$$\mathbb{I}\left[\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}}W_{i,s}W_{i,T}\|\Lambda_i\varepsilon_{i,T}\|_2 > \delta\right] \leq \mathbb{I}\Big[\|\Lambda_i\varepsilon_{i,T}\|_2 > \delta\sqrt{\underline{q}N}\Big].$$

In addition, there exists $M > 0$ such that

$$\sum_{i=1}^{N} \mathbb{P}\Big(\|\Lambda_i \varepsilon_{i,T}\|_2 > \delta\sqrt{\underline{q}N}\Big) \le \sum_{i=1}^{N} \frac{\mathbb{E}[\|\Lambda_i \varepsilon_{i,T}\|_2^2]}{\delta^2 \underline{q} N} = \frac{\mathbb{E}[\|\Lambda_i\|_2^2 \sigma_\varepsilon^2]}{\delta^2 \underline{q}^2} < \infty,$$

where the first inequality follows from *Markov inequality*. Therefore, *Borel-Cantelli Lemma* implies

$$\mathbb{I}\left[\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}} W_{i,s} W_{i,T} \|\Lambda_i \varepsilon_{i,T}\|_2 > \delta\right] \xrightarrow{\text{a.s}} 0.$$

Further using *Dominated Convergence Theorem*, we obtain the following

$$\sum_{i=1}^{N} \mathbb{E}\left[\frac{1}{|\mathcal{Q}_{s,T}|} W_{i,s} W_{i,T} \|\Lambda_i \varepsilon_{i,T}\|^2 \mathbf{1}\left\{\frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}} W_{i,s} W_{i,T} \|\Lambda_i \varepsilon_{i,T}\|_2 > \delta\right\}\right] \to 0.$$

Therefore, Lindeberg's CLT holds and using Assum. 5 and 6 we have the following for $N \to \infty$:

$$Z_{s,T} \xrightarrow{d} \mathcal{N}\big(0, \sigma_\varepsilon^2 \Sigma_\Lambda\big), \quad \sqrt{\frac{N}{|\mathcal{Q}_{s,T}|}} Z_{s,T} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\alpha_{s,T}} \sigma_\varepsilon^2 \Sigma_\Lambda\right) \quad \text{stably on } \mathcal{F}_T,$$

and the vector $\left(\frac{\sqrt{N}}{|\mathcal{Q}_{1,T}|} \sum_{i\in\mathcal{Q}_{1,T}} \Lambda_i^\top \varepsilon_{i,T}, \dots, \frac{\sqrt{N}}{|\mathcal{Q}_{T,T}|} \sum_{i\in\mathcal{Q}_{T,T}} \Lambda_i^\top \varepsilon_{i,T}\right)^\top$ is jointly asymptotically normal stably on $\mathcal{F}_T$. For each $s, s' \in [T]$, the asymptotic covariance matrix of $\frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top Z_{s,T}$ consists of the following terms

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\frac{N}{|\mathcal{Q}_{s,T}|}} Z_{s,T} \cdot \sqrt{\frac{N}{|\mathcal{Q}_{s'T}|}} Z_{s'T}^\top \middle| \mathcal{F}_T\right] =& \frac{N}{|\mathcal{Q}_{s,T}||\mathcal{Q}_{s',T}|} \sum_{i\in\mathcal{Q}_{s',T}} \sum_{j\in\mathcal{Q}_{s',T}} \mathbb{E}\big[\Lambda_i \varepsilon_{i,T} \Lambda_j^\top \varepsilon_{j,T}\big] \\
=& \frac{N|\mathcal{Q}_{s,T,s',T}|}{|\mathcal{Q}_{s,T}||\mathcal{Q}_{s',T}|} \sum_{i\in\mathcal{Q}_{s,T,s',T}} \mathbb{E}[\Lambda_i \Lambda_i^\top] \mathbb{E}[\varepsilon_{i,T}^2] \\
=& U^{(s,s')} \sigma_\varepsilon^2 \Sigma_\Lambda,
\end{aligned}$$

where $U^{(s,s')} \triangleq \frac{N|\mathcal{Q}_{s,T,s',T}|}{|\mathcal{Q}_{s,T}||\mathcal{Q}_{s',T}|}$. From Assum. 5, as $N \to \infty$

$$U^{(s,s')} \xrightarrow{\text{a.s.}} \frac{\beta_{s,T,s',T}}{\alpha_{s,T}\alpha_{s',T}} \le \frac{1}{\alpha_{s,T}} \le \frac{1}{\underline{q}} < \infty,$$

and using Assum. 6

$$\frac{1}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T} U^{(s,s')} \quad \xrightarrow[N\to\infty]{\text{a.s.}} \quad \frac{1}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T} \frac{\beta_{s,T,s',T}}{\alpha_{s,T}\alpha_{s',T}} \quad \xrightarrow[T\to\infty]{\mathbb{P}} \quad \omega_1.$$

Therefore,

$$\operatorname{Cov}\left(\frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top \sqrt{\frac{N}{|\mathcal{Q}_{s,T}|}} Z_{s,T} \middle| \mathcal{F}_T\right) = \frac{1}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T} \sigma_\varepsilon^2 U^{(s,s')} F_s F_s^\top \Sigma_\Lambda F_{s'} F_{s'}^\top \xrightarrow{\text{a.s.}} \mathcal{D}_T$$

as $N \to \infty$, where

$$\mathcal{D}_T \triangleq \frac{\sigma_\varepsilon^2}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T} \frac{\beta_{s,T,s',T}}{\alpha_{s,T}\alpha_{s',T}} F_s F_s^\top \Sigma_\Lambda F_{s'} F_{s'}^\top.$$

In order to show unconditional CLT of $\frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top Z_{s,T}$, we invoke *Chebychev's inequality* and show that $\mathcal{D}_T$ converges in probability. We denote $V^{(s,s')} \triangleq \mathrm{vec}(F_s F_s^\top \Sigma_\Lambda F_{s'} F_{s'}^\top)$. Therefore for any $m \in [r^2]$,

$$\mathbb{E}\left[\frac{1}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T} U^{(s,s')}\Big(V_m^{(s,s')} - \mathbb{E}[V_m^{(s,s')}]\Big)\right]^2 = \frac{1}{T^4}\sum_{s,s',t,t'} U^{(s,s')}U^{(t,t')}\mathrm{Cov}\Big(V_m^{(s,s')}, V_m^{(t,t')}\Big), \tag{46}$$

We handle the sum in (46) case by case. When at least one of the two equalities $s = t$ and $s' = t'$ holds, the sum has at most $\mathcal{O}(T^3)$ terms with finite second moments. When $s \neq s', t \neq t'$, there exists $(u, v) \in [r] \times [r]$ depending on m such that

$$\begin{aligned}\mathrm{Cov}\Big(V_m^{(s,s')}, V_m^{(t,t')}\Big) =& \mathrm{Cov}\big(F_{s,u}F_s^\top \Sigma_\Lambda F_t F_{t,v}, F_{s',u}F_{s'}^\top \Sigma_\Lambda F_{t'} F_{t',v}\big)\\ =& \sum_{a,b,a',b'} (\Sigma_\Lambda)_{a,b}(\Sigma_\Lambda)_{a',b'}\mathrm{Cov}(F_{s,u}F_{s,a}F_{t,b}F_{t,v}, F_{s',u}F_{s',a'}F_{t',b'}F_{t',v}).\end{aligned} \tag{47}$$

By Assum. 3 the errors η_t are Gaussian, and by Assum. 1 F_t is a VAR(1) process. Therefore $\|\mathbb{E}[F_s F_t^\top]\| = \mathcal{O}(\|A^{|s-t|}\|)$ and for $s \neq t$, $s' \neq t'$, we have $\mathrm{Cov}\Big(V_m^{(s,s')}, V_m^{(t,t')}\Big) = \mathcal{O}(\|A^h\|)$, where $h \triangleq \min\{|s-s'|, |t-t'|, |s-t'|, |s'-t|\}$. Fixing s, t and summing (47) we obtain

$$\sum_{s',t': s\neq t, s'\neq t'} \mathrm{Cov}\Big(V_m^{(s,s')}, V_m^{(t,t')}\Big) = \sum_{h=1}^{T} \mathcal{O}(Th)\mathcal{O}(\|A^h\|) < \mathcal{O}(T)\sum_{h=1}^{\infty} h\|A^h\| = \mathcal{O}(T),$$

where the last step follows from Lem. D.2. Hence (46) can be bounded by $T^{-4}\mathcal{O}(T^3) = \mathcal{O}(T^{-1})$. Hence we apply Chebychev's inequality to obtain $\mathcal{D}_T \xrightarrow{\mathbb{P}} \omega_1\sigma_\varepsilon^2\Sigma_F\Sigma_\Lambda\Sigma_F$ as $T \to \infty$. Using tower property of conditional expectation and (45),

$$\lim_{N,T\to\infty} \mathrm{Cov}\left(\frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top \sqrt{\frac{N}{|\mathcal{Q}_{s,T}|}} Z_{s,T}\right) = \omega_1\sigma_\varepsilon^2\Sigma_F\Sigma_\Lambda\Sigma_F.$$

Therefore,

$$\frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top \sqrt{\frac{N}{|\mathcal{Q}_{sT}|}} Z_{sT} \xrightarrow{d} \mathcal{N}\big(0, \omega_1\sigma_\varepsilon^2\Sigma_F\Sigma_\Lambda\Sigma_F\big).$$

D.6.2 Proof of (41)

Denote $Z_{i,t} \triangleq W_{i,t}F_t$. By Assum. 6, $\frac{1}{T}\sum_{t=1}^{T} W_{i,t}F_tF_t^\top = \frac{1}{T}\sum_{t=1}^{T} Z_{i,t}Z_{i,t}^\top \xrightarrow{\mathbb{P}} \Sigma_{F,i}$. Additionally we denote, $U_{i,t,T} \triangleq \frac{1}{\sqrt{T}}Z_{i,t}\varepsilon_{i,t}$, and the filtration, $\mathcal{F}_{i,t} \triangleq \sigma(\{A_{i,s}\}_{s\in[t+1]}, \{\varepsilon_{i,s}\}_{s\in[t]})$. Therefore,

$$\mathbb{E}[U_{i,t,T} \mid \mathcal{F}_{i,t-1}] = \frac{1}{\sqrt{T}} Z_{i,t}\mathbb{E}[\varepsilon_{i,t} \mid \mathcal{F}_{i,t-1}] = 0,$$

since $\varepsilon_{i,t}$ is independent of the history until time $t-1$ and $Z_{i,t}$, and it has mean zero. Therefore $(U_{i,t,T}, \mathcal{F}_{i,t,T})_{t\in[T]}$ is a Martingale difference array in $\mathbb{R}^r$. Additionally,

$$\sum_{t=1}^{T}\mathbb{E}[U_{i,t,T}\ U_{i,t,T}^\top \mid \mathcal{F}_{i,t-1}] = \sum_{t=1}^{T}\frac{1}{T}\sigma_\varepsilon^2 Z_{i,t}Z_{i,t}^\top \xrightarrow{\mathbb{P}} \sigma_\varepsilon^2\Sigma_{F,i}.$$

Hence adapted to $\mathcal{F}_{i,t-1}$, the sum of $U_{i,t,T}$'s has finite second moment. Also, $\|U_{i,t,T}\|_2 = T^{-1/2}|\varepsilon_{i,t}| \cdot \|Z_{i,t}\|_2$.

Therefore for any $\delta > 0$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\big[\|U_{i,t,T}\|_2^2 \; \mathbb{I}[\|U_{i,t,T}\|_2 > \delta] \mid \mathcal{F}_{i,t-1}\big] \\
\leq \quad & \sum_{t=1}^{T} \frac{1}{T} \mathbb{E}\Big[|\varepsilon_{i,t}|^2 \|Z_{i,t}\|_2^2 \; \mathbb{I}\Big[|\varepsilon_{i,t}| \cdot \|Z_{i,t}\|_2 \geq \sqrt{T}\delta\Big]\Big] \\
\leq \quad & \big(\mathbb{E}[|\varepsilon_{i,t}|^4] \mathbb{E}[\|Z_{i,t}\|_2^4]\big)^{1/2} \quad \Big(\mathbb{P}\Big(|\varepsilon_{i,t}| \cdot \|Z_{i,t}\|_2 \geq \sqrt{T}\delta\Big)\Big)^{1/2},
\end{aligned}$$

where the last step follows from Cauchy-Schwarz inequality. From Assum. 1, 2 and 3, $\varepsilon_{i,t}$ and $Z_{i,t}$ have finite fourth moments and are bounded in probability. Thus,

$$\mathbb{P}\Big(|\varepsilon_{i,t}| \cdot \|Z_{i,t}\|_2 \geq \sqrt{T}\delta\Big) \to 0, \quad \text{as } T \to \infty.$$

Therefore the *Lindeberg condition* is satisfied. Using *Martingale central limit theorem* [28, Thm. 3.2], 41 holds with $\Phi_i^{\text{obs}} = \sigma_\varepsilon^2 \Sigma_{F,i}$.

D.6.3 Proof of (42)

The proof proceeds analogously to step 5 in the proof of Proposition 3 in Xiong and Pelger [48]. We denote

$$\begin{aligned}
V_{s,T} \triangleq \frac{1}{|\mathcal{Q}_{s,T}|} \sum_{i \in \mathcal{Q}_{s,T}} \Lambda_i \Lambda_i^\top - \frac{1}{N} \sum_{i=1}^{N} \Lambda_i \Lambda_i^\top = & \frac{1}{\sqrt{N}} \Bigg[\Bigg(\sqrt{\frac{N}{|\mathcal{Q}_{s,T}|}} - \sqrt{\frac{|\mathcal{Q}_{s,T}|}{N}} \Bigg) \frac{1}{\sqrt{|\mathcal{Q}_{s,T}|}} \sum_{i \in \mathcal{Q}_{s,T}} \Lambda_i \Lambda_i^\top \\
& - \sqrt{\frac{N - |\mathcal{Q}_{s,T}|}{N}} \frac{1}{\sqrt{N - |\mathcal{Q}_{s,T}|}} \sum_{i \in \mathcal{Q}_{s,T}^c} \Lambda_i \Lambda_i^\top \Bigg].
\end{aligned}$$

Therefore invoking Assum. 2 and 5, $\sqrt{N}\text{vec}(V_{s,T}) \xrightarrow{d} \Big(\frac{1}{\sqrt{\gamma_s}} - \sqrt{\gamma_s}\Big) Z_1 - \sqrt{1-\gamma_s} Z_2$, where $Z_1, Z_2 \sim \mathcal{N}(0, \Theta_\Lambda)$ independent, with $\Theta_\Lambda = \mathbb{E}[\text{vec}(\Lambda_i\Lambda_i^\top - \Sigma_\Lambda)(\text{vec}(\Lambda_i\Lambda_i^\top - \Sigma_\Lambda))^\top]$. Simplifying the variance expression,

$$\sqrt{N}\text{vec}(V_{s,T}) \xrightarrow{d} \mathcal{N}\bigg(0, \Big(\frac{1}{\gamma_s} - 1\Big)\Theta_\Lambda\bigg).$$

Similar to Xiong and Pelger [48, Prop. 3], the covariance term for $s, s' \in [T]$ is

$$\lim_{T,N \to \infty} \text{Cov}(\sqrt{N}\text{vec}(V_{s,T}), \sqrt{N}\text{vec}(V_{s',T})) = \Big(\frac{\gamma_{s,s'}}{\gamma_s \gamma_{s'}} - 1\Big)\Theta_\Lambda.$$

(42) consists of $J_T F_T$ where $J_T = \frac{1}{T}\sum_{s=1}^{T} F_s F_s^\top V_{s,T}$. Following the route of Xiong and Pelger [48, Prop. 3.1(b), step 5.2], and invoking Assum. 6 the sum has a limiting covariance term

$$\lim_{T,N \to \infty} \text{Cov}(\sqrt{N}\text{vec}(J_T)) = (\omega_1 - 1)(I_r \otimes \Sigma_F)\Theta_\Lambda(I_r \otimes \Sigma_F).$$

Using the stable convergence in Law [30, Thm. 6.1], $\sqrt{N}\Sigma_{J,T}^{-1/2} J_T F_T \xrightarrow{d} \mathcal{N}(0, I_r)$ where $\Sigma_{J,T} = (\omega_1 - 1)(F_T^\top \otimes \Sigma_F)\Theta_\Lambda(F_T \otimes \Sigma_F)$. Similarly the proof for asymptotic normality of R_i is similar to Xiong and Pelger [48, Prop. 3.1(b), step 5.4] with the asymptotic covariance of $R_i u_i$ being $h_i(u_i) = (\omega_3 - 1)(\Sigma_{F,i} \otimes \Sigma_F)\Theta_\Lambda(\Sigma_{F,i} \otimes \Sigma_F)$. Additionally the asymptotic covariance term between $J_T F_T$ and $R_i u_i$ being $g_{i,T}^{\text{cov}}(u_i)^\top = (\omega_2 - 1)(\Sigma_{F,i} \otimes \Sigma_F)\Theta_\Lambda(I_r \otimes \Sigma_F)$. Thus the proof is complete. □

E Identifiability of $\theta_{i,T:T+h}$ under rotation

In (4), factors and loadings are identifiable only up to a non-singular rotation. For any invertible H, the rotated representation yields the factors $G_t = HF_t$, loadings $(H^\top)^{-1}\Lambda_i$, and coefficient matrix HAH^{-1}. The forecast target remains invariant, since $\theta_{i,T}^{(h,H)} = \big((H^\top)^{-1}\Lambda_i\big)^\top (HAH^{-1})^h G_T = \theta_{i,T:T+h}$. Hence identification assumption determines H but not affecting the forecast target.

F Assum. 5 under MCAR: High-Probability Guarantee

In this section we show that the condition (13) in Assum. 5 regarding the observation W holds with high probability when the observations are missing completely at random.

Lemma F.1 (MCAR)**.** *Suppose the observation matrix W satisfies Def. 4.1. Then for $T = \mathcal{O}(e^N)$ and $N \to \infty$,* (13) *holds almost surely.*

Proof. We have $|\mathcal{Q}_{s,t}|/N = \frac{1}{N}\sum_{i=1}^N W_{i,s}W_{i,t}$ that is a sum of independent Bernoulli(p^2) random variables. Hence we can apply Hoeffding's inequality [46, Thm. 2.2.5] for $\underline{q} \in (0, p^2)$ and $s, t \in [T]$ as follows

$$\mathbb{P}\big(|\mathcal{Q}_{s,t}| < N\underline{q}\big) = \mathbb{P}\bigg(\frac{|\mathcal{Q}_{s,t}|}{N} < \underline{q}\bigg) \le \mathbb{P}\bigg(\bigg|\frac{|\mathcal{Q}_{s,t}|}{N} - p^2\bigg| > p^2 - \underline{q}\bigg) \le 2e^{-2N(p^2-\underline{q})^2}.$$

Union bound over $s, t \in [T]$ implies

$$\mathbb{P}\big(|\mathcal{Q}_{s,t}| < N\underline{q} \text{ for some } s,t \in [T]\big) \le 2T^2 e^{-2N(p^2-\underline{q})^2} = 2e^{-2[N(p^2-\underline{q})^2 - \log T]}. \tag{48}$$

Hence in the regime $T = \mathcal{O}(e^N)$, we have $\log T \le CN$ for some $C > 0$. Consequently,

$$\sum_{N=1}^{\infty} \mathbb{P}\big(|\mathcal{Q}_{s,t}| < N\underline{q} \text{ for some } s,t \in [T]\big) \le \sum_{N=1}^{N} e^{-2N[(p^2-\underline{q})^2 - C]} < \infty.$$

Hence by first Borel-Cantelli lemma, $\mathbb{P}(|\mathcal{Q}_{s,t}| < N\underline{q}$ infinitely often for some $s,t) = 0$. Equivalently, $|\mathcal{Q}_{s,t}| \ge N\underline{q}$ almost surely as $N \to \infty$. Hence the proof is complete. □

G Additional details of simulation studies in Sec. 5.1

In this section, we defer the experimental details of Sec. 5.1. The benchmarks mSSA and SyNBEATS implemented from the publicly available repositories `https://github.com/AbdullahO/mSSA` and `https://github.com/Crabtain959/SyNBEATS` respectively.

Generative models. We describe the three DGPs as follows–

1. **DGP-1 (Purely autoregressive process).** The factors are generated from an AR(1) process as $F_t = 0.5F_{t-1} + \eta_t$ with $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, (0.5)^2\big)$.
2. **DGP-2 (Autoregressive process with quadratic component).** The factors are generated as a sum of quadratic function $F_t^{(1)} = 2t^2/T^2$ and the AR(1) component $F_t^{(2)}$ as DGP-1.
3. **DGP-3 (ARMA process and quadratic component).** The quadratic component $F_t^{(1)}$ of the factors are generated similar to DGP-2. The autpregressive moving average component i.e. ARMA(1,1) has the following generative model:

   $$F_t^{(2)} = 0.5F_{t-1}^{(2)} - 0.4F_{t-2}^{(2)} + 0.2F_{t-2}^{(2)} + \eta_t + 0.5\eta_{t-1},$$

   with $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, (0.7)^2\big)$.

Simultaneous adoption observation pattern. We implement simultaneous adoption pattern in DGP-2 that is generated as follows. Unit-specific characteristics are generated as $X_i = \mathbb{I}[\Lambda_i \ge 0]$. For the units with $X_i = 1$, 25% randomly selected units have missing entries onward $\lceil 0.75T \rceil$, and the remaining 75% have all the entries observed. For the units with $X_i = 0$, 62.5% randomly selected units have missing entries onward $\lceil 0.375T \rceil$, and the remaining 37.5% have all the entries observed.

Tuning the penalized smoothing splines. The penalized smoothing splines is a method for estimating functions with nonparametric regression [27, 29]. We implement the penalized spline with penalty tuned with 10-fold block-cross validation and block size $\lceil \sqrt{T} \rceil$, with previous blocks training the temporally next blocks accordingly.

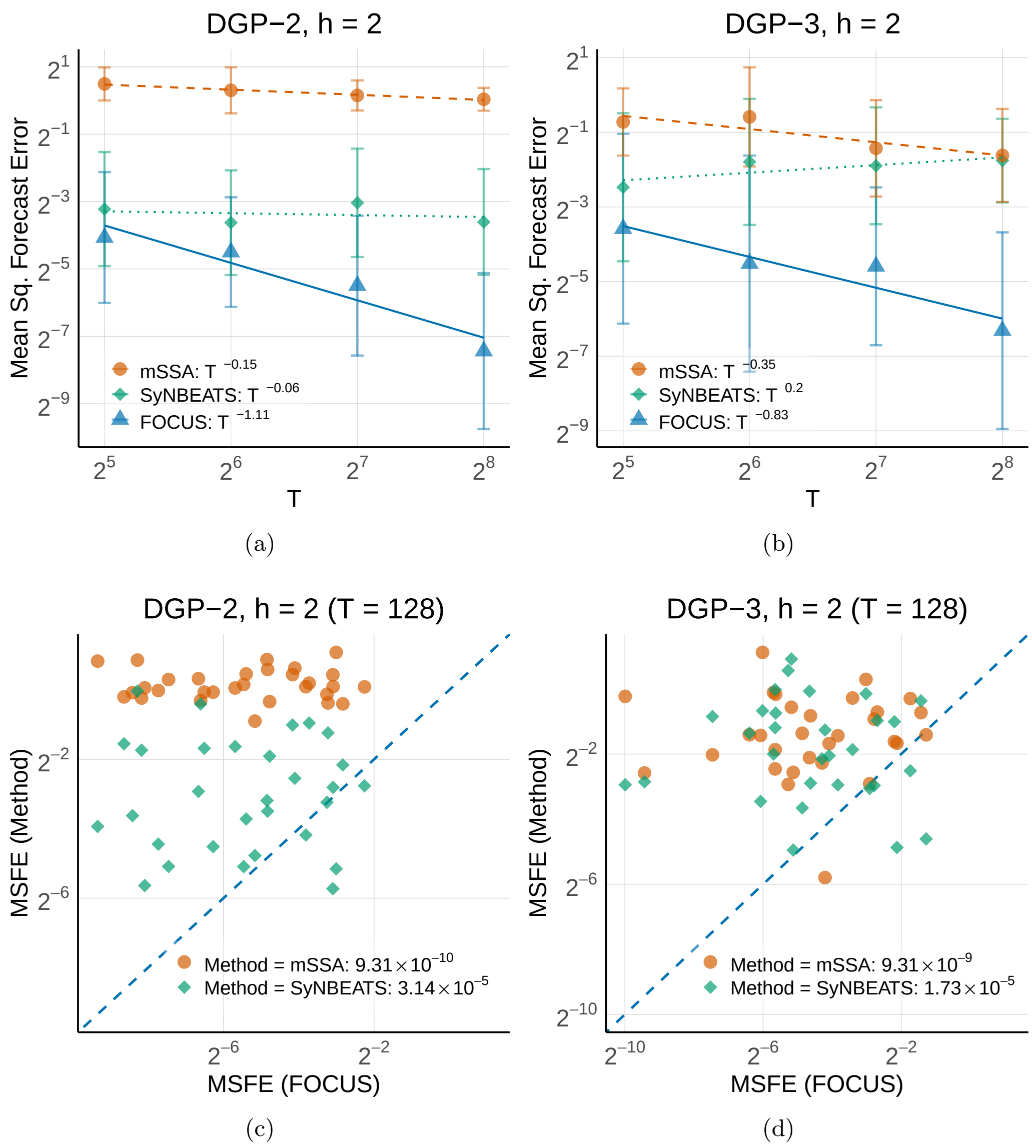


Figure 3: **Additional figures for MSFE (averaged over 30 trials) for $h = 2$ across the benchmarks for $N = 64$ and three generative models.** Panels (a) and (b) present the average MSFE of Focus (blue triangle), mSSA (orange circle) and SyNBEATS (green diamond) across $T \in \{2^5, 2^6, 2^7, 2^8\}$, and the vertical lines mark the one standard deviation error bars. As comapared to SyNBEATS and mSSA, Focus has lower average MSFE that decreases faster with T (empirical rates in the legends). Panels (c) and (d) present scatter plots of difference of MSFE (Focus Benchmark method) for $T = 128$ and $h = 2$. The errors of Focus are significantly lower (p-values of Wilcoxon's one-sided pairwise test in legends < 0.01), resulting the scatter plots concentrated in $y > x$ region.

H Additional details of the HeartSteps case study in Sec. 5.2

The HeartSteps data has 37 users and the maximum number of decision points across users is 315. Users were considered unavailable (i.e., not nudged) when driving, offline and in similar circumstances. The readers are referred to `https://github.com/klasnja/HeartStepsV1` for the data source, and Klasnja et al. [35], Liao et al. [36] for more details.

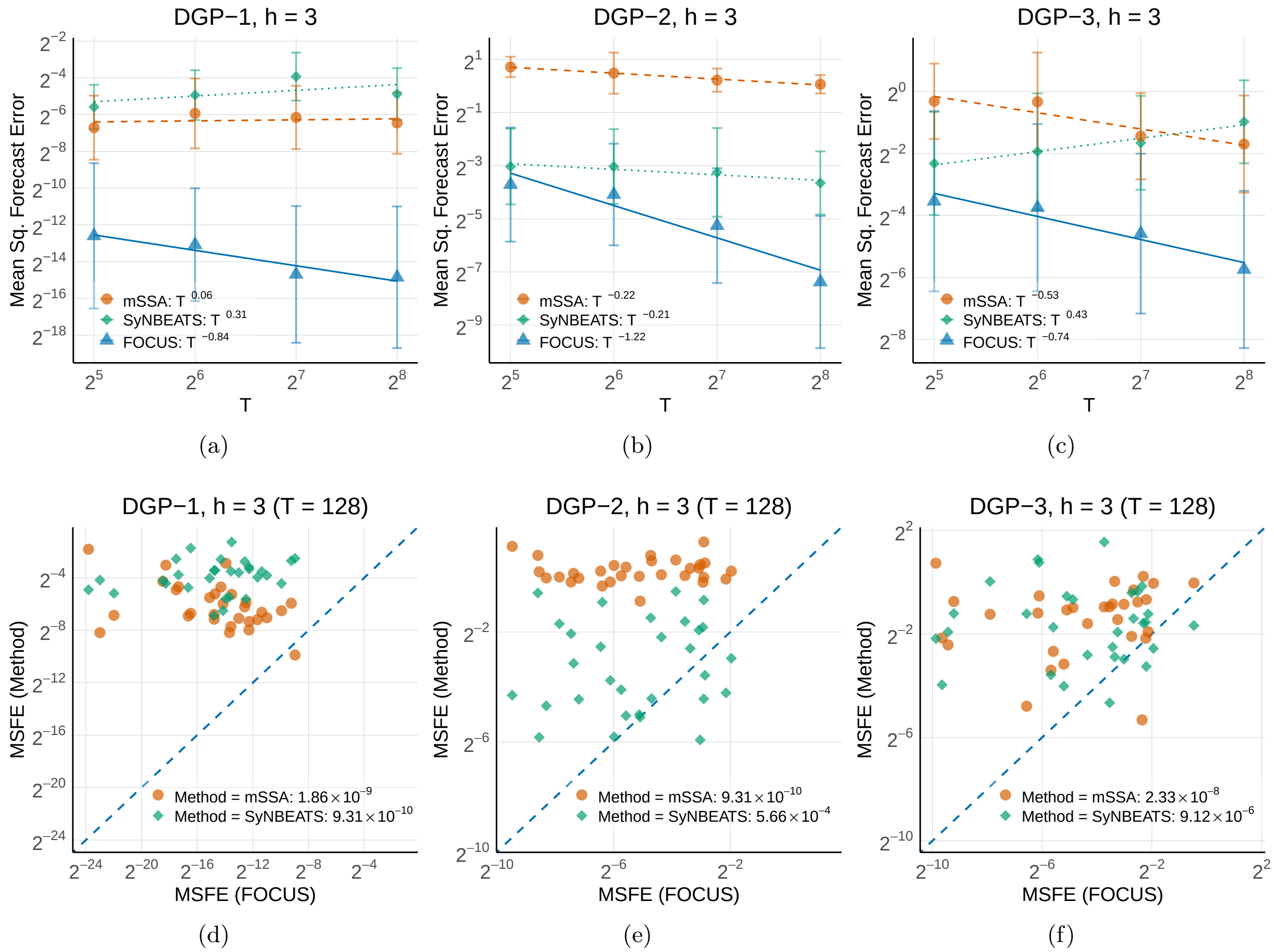


Figure 4: **Additional figures for MSFE (averaged over 30 trials) for $h = 3$ across the benchmarks for $N = 64$ and three generative models.** Panels (a) and (b) present the average MSFE of Focus (blue triangle), mSSA (orange circle) and SyNBEATS (green diamond) across $T \in \{2^5, 2^6, 2^7, 2^8\}$, and the vertical lines mark the one standard deviation error bars. As comapared to SyNBEATS and mSSA, Focus has lower average MSFE that decreases faster with T (empirical rates in the legends). Panels (c) and (d) present scatter plots of difference of MSFE (Focus Benchmark method) for $T = 128$ and $h = 3$. The errors of Focus are significantly lower (p-values of Wilcoxon's one-sided pairwise test in legends < 0.01), resulting the scatter plots concentrated in $y > x$ region.

H.1 Preprocessing the data

We consider three variables in the data– the binary variable `available` that indicates user availability at the underlying decision slot, the binary variable `send` that tracks whether a user was nudged with a notification, and `jbsteps30` that measures the number of steps accomplished by the user 30 minutes after sending the activity prompt. The outcome variable is $\log(1 + \texttt{jbsteps30})$, and the treatment variable is `available * nudge`.

For $i = 1, \ldots, 37$, the i^{th} row of the outcome matrix Y and the observation matrix W are constructed by stacking $\log(1+\texttt{jbsteps30})$ and `available * sent` indicators for user i at each quintuplets for a single decision day. Next, we discard user 31 from the analysis due to consistently low nudges across time points, resulting to $N = 36$. Out of the 315 total time points, the horizon T for model training is considered in the range $[100, 200]$ in multiples of 10 for obtaining enough time points to extract the factor dynamics, as well as to rule out the last few intervention periods when many users discontinued the experiment.

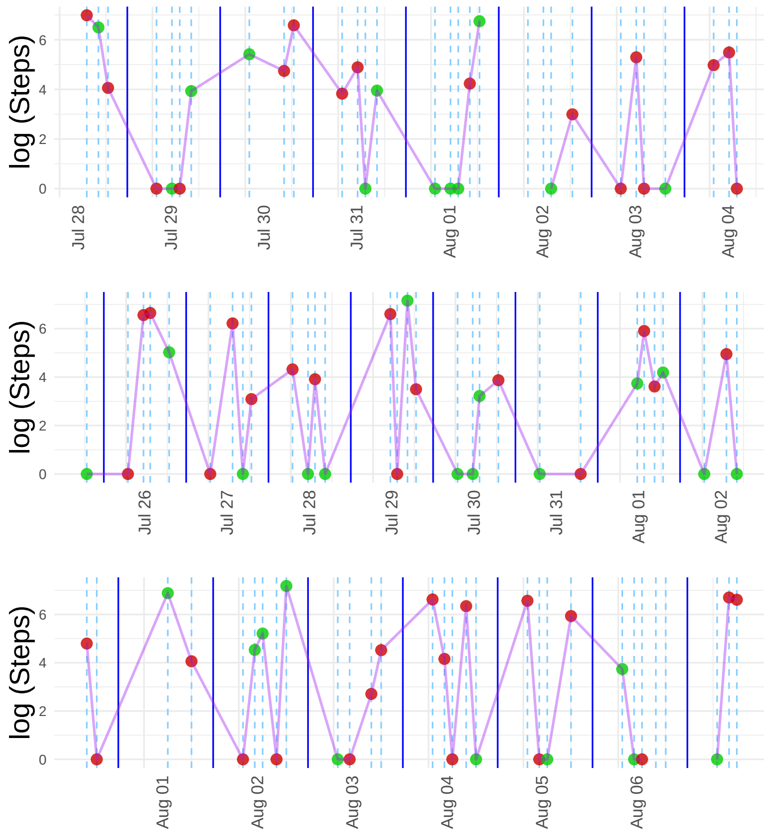


Figure 5: $\log(1 + \texttt{jbsteps30})$ **vs time for three users in HeartSteps data.** The 5 decision slots each day are marked by the dashed blue vertical lines. The green (and red) dots represent that the user was nudged (not nudged). Between consecutive slots, the steps exhibit a negative correlation shared across users.

H.2 Focus on slot pair (4,5)

The factors are estimated with the PCA method of Xiong and Pelger [48]. Since $T > N$ in most cases and interchanging the roles of $\hat{\Lambda}$ and $\hat{F}$ in Step 1 of Focus lowers the chance of $(|\mathcal{Q}_{i,j}|)_{i,j\in[N]}$ being small or zero. Upon observing the scree plots of the PCA for most slices of the data, we choose the dimension $r = 7$ to explain at east 80% of the data for most choices of T.

H.3 Expression of Focus estimator for HeartSteps

For forecast horizon at $T = 5K,\ K \geq 2$, the estimated factors are $\{\hat{F}_t, t = 1, \ldots, 5K\}$. We select the estimated factors at slot $s \in \{4, 5\}$ as $\hat{F}^{(s)}$, where

$$\left(\hat{F}^{(s)}\right)^\top := \left\{\hat{F}_{s+5j} : j = 0, \ldots, K-1\right\}.$$

Next, we regress the estimated factors at slot 5 on that of slot 4 to obtain the estimated coefficient matrix as

$$\hat{A}_{4\to 5} = \left(\hat{F}^{(5)}\right)^\top \hat{F}^{(4)} \left[\left(\hat{F}^{(4)}\right)^\top \hat{F}^{(4)}\right]^{-1}.$$

This approach is similar to the VAR(1) coefficient matrix estimation in (11). The 5-step forecast estimator for a user i for $T = 5K$ is $\hat{\theta}_{i,T+5} = \hat{\Lambda}_i^\top \hat{A}_{4\to 5} \hat{F}_{5K}$.

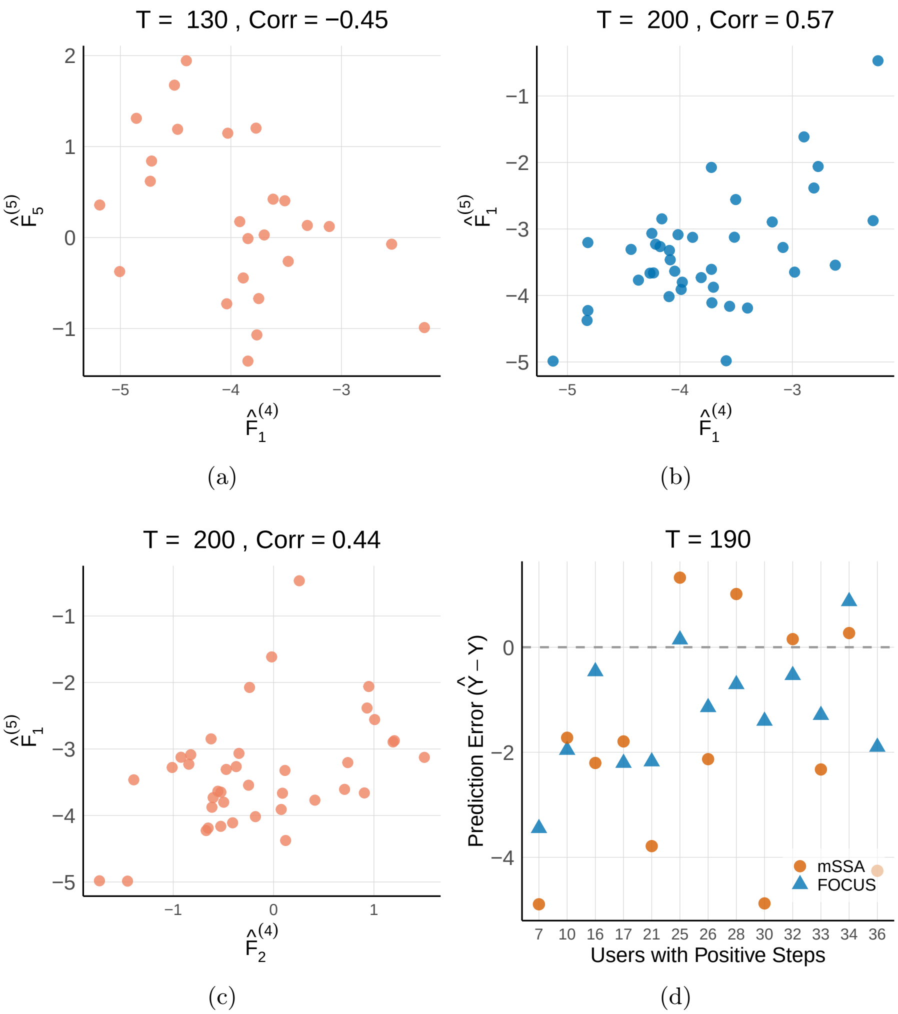


Figure 6: **Additional results for Focus and mSSA on HeartSteps data.** Panels (a), (b) and (c) present scatter plots of the estimated factors at slot 5 i.e. $\hat{F}^{(5)}$ vs the same at slot 4 denoted by $\hat{F}^{(4)}$. In all three panels, $\hat{F}_i^{(4)}$ and $\hat{F}_j^{(5)}$ are strongly correlated for several pairs of (i, j). Panel (d) shows a comparison of the prediction errors between Focus and mSSA at $T = 190$ and $h = 5$. The prediction error is measured with Mean squared prediction error for the users with positive steps at $T + h$. The blue triangles (Focus) are closer to the origin axis than the orange circles (mSSA)– indicating that Focus exhibits better forecasting performance than mSSA at $T = 190$ and $h = 5$.