Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference

Ruoxuan Xiong and Markus Pelger

Emory University and Stanford University

Motivation

Problem: Large dimensional panel data with missing entries is prevalent:

- Macroeconomic data: Staggered releases, mixed frequencies
- Policy evaluation: Simultaneous or staggered policy rollout
- Financial data: Mergers, new firms, bankruptcy
- Recommender system: Netflix challenge

Our Goal: Impute missing values and estimate latent factor structure for panel with general observational pattern

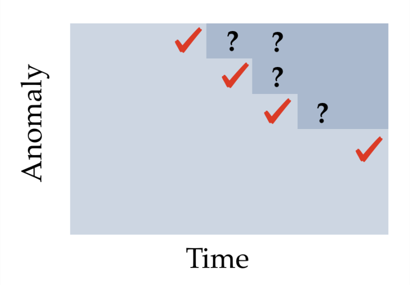
A Motivating Example: A Causal Approach to Study Publication Effect

Question: Does academic publication of a strategy affect this strategy's return?

Large-dimensional data: Many strategies and their returns over many time-periods. Strategies were published at different times

A causal inference approach: Compare the returns without and with publication. We can only observe one at one time. The other one is the counterfactual and modeled as missing obsrvation.

Impute missing observations: Use general statistical factors estimated from the partial observed large-dimensional panel data



Observation pattern for the panel of returns without publication

Contribution

Large-dimensional factor modeling:

- Simple all-purpose estimator for latent factor structure and data imputation for essentially any missing pattern
- Inferential theory for latent factor models and imputed values under general approximate factor model

Causal inference on panel data:

- Counterfactual outcomes modeled as missing values and imputed by estimated common components from latent factor
- Test for the entry-wise, time-dependent treatment effect under general treatment adoption pattern with unobserved factors

Empirical study:

- Companion paper: Study the publication effect of investment anomaly strategies

Importance

Causal inference on panel data:

Example: Publication effect on risk factors, Smoking regulation in different states
Problem: When and where is the intervention effective?
Our solution: Tests for entry-wise and weighted treatment effects
Importance: Goes beyond mean effects without assuming prespecified covariates

Large-dimensional factor modeling

Example: Panel of macroeconomic data or stock returns
Problem: How to estimate a factor model from incomplete data?
Our solution: Estimator for the factor model with confidence interval
Importance: Input for other applications, for example risk factors

Missing data imputation

Example: Financial data, mixed frequency data, users' ratings at Netflix
Problem: Whether to use imputed value?
Our solution: Estimator for each entry with confidence interval
Importance: Include observations with incomplete data instead of leaving them out for analysis which can lead to bias and efficiency loss

Related Literature (Incomplete and Partial List)

Factor modeling

- Full observations with inferential theory: Bai and Ng 2002, Bai 2003, Fan, Liao and Mincheva 2013, Pelger and Xiong 2021a+b, Lettau and Pelger 2020a+b
- Partial observations: Stock and Watson 2002, Jin, Miao and Su 2021, Bai and Ng 2021, Cahan, Bai and Ng 2021

Causal inference on panel data

- Difference in differences: Card and Krueger 1994, Bertrand, Duflo and Mullainathan 2004
- Synthetic control methods: Abadie and Gardeazabal 2003, Abadie, Diamond and Hainmueller 2010, 2015, Doudchenko and Imbens 2016
- Matrix completion: Athey, Bayati, Doudchenko, Imbens and Khosravi 2021

Matrix completion

- Independent sampling: Candes and Recht 2009, Mazumder, Hastie and Tibshirani 2010, Negahban and Wainright 2012
- Dependent sampling: Athey, Bayati, Doudchenko, Imbens and Khosravi 2021
- Independent sampling with inferential theory: Chen, Fan, Ma and Yan 2019

Theory: Model and Estimation

Model Setup: Approximate Latent Factor Model

Approximate factor model: Observe Y_{it} for N units over T time periods

$$Y_{it} = \underbrace{\Lambda_i^\top}_{1\times k} \underbrace{F_t}_{k\times 1} + e_{it}$$

In matrix notation:

$$\underbrace{Y}_{N\times T} = \underbrace{\Lambda}_{N\times k} \underbrace{F^\top}_{k\times T} + \underbrace{e}_{N\times T}$$

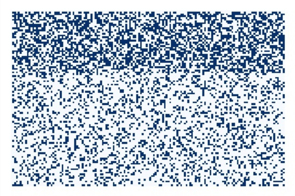
- N and T large
- Factors F_t explain common time-series movements
- Loadings Λ_i capture correlation between units
- Factors and loadings are latent and estimated from the data
- Common component $C_{it} = \Lambda_i^\top F_t$
- Idiosyncratic errors $\mathbb{E}[e_{it}] = 0$
- Number of factors k fixed

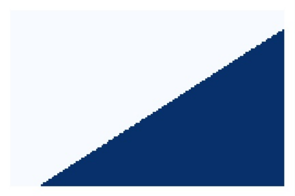
$\Rightarrow$ Estimate Λ_i, F_t, C_{it} and use estimated C_{it} to impute missing Y_{it}

General Observational Pattern

Observation matrix $W = [W_{it}] : W_{it} = \begin{cases} 1 & \text{observed} \\ 0 & \text{missing} \end{cases}$

- **W can depend on Λ, but independent of F and e**



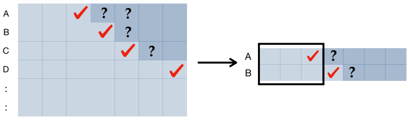


- Missing uniformly at random $P(W_{it} = 1) = p$
- Cross-section missing at random $P(W_{it} = 1) = p_t$
- Time-series missing at random $P(W_{it} = 1) = p_i$

- Staggered treatment adoption $P(W_{it} = 1) = p_{it}$ Once missing stays missing: $W_{is} = 0$ for $s \geq t$
- Mixed-frequency observations $P(W_{it} = 1) = p_{it}$ Equivalent to staggered design after reshuffling

Estimation of the Factor Model (All-Purpose Estimator)

Step 1 Estimate sample covariance matrix $\tilde{\Sigma}$ of Y using only observed entries: $\tilde{\Sigma}_{ij} = \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} Y_{it} Y_{jt}$, where $Q_{ij} = \{t : W_{it} = 1 \text{ and } W_{jt} = 1\}$ are times where both units are observed



Step 2 Estimate loadings $\tilde{\Lambda}$ (standard):

Apply principal component analysis (PCA) to $\tilde{\Sigma} = \frac{1}{N} \tilde{\Lambda} \tilde{D} \tilde{\Lambda}^\top$

Step 3 Estimate factors $\tilde{F}$ with regression on loadings for observed entries:

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

Step 4 Estimate common components/missing entries $\tilde{C}_{it} = \tilde{\Lambda}_i^\top \tilde{F}_t$

Assumptions: Approximate Factor Model

Assumption 1: Approximate Factor Model

1. Systematic factor structure: Σ_F and Σ_Λ full rank

$$\frac{1}{T}\sum_{t=1}^{T} F_t F_t^\top \xrightarrow{p} \Sigma_F \qquad \frac{1}{N}\sum_{i=1}^{N} \Lambda_i \Lambda_i^\top \xrightarrow{p} \Sigma_\Lambda$$

2. Weak dependence of errors: bounded eigenvalues of correlation and autocorrelation matrix for errors
   Simplification for presentation: $e_{it} \overset{iid}{\sim} (0, \sigma_e^2)$, $\mathbb{E}[e_{it}^8] < \infty$
3. Factors F_t and errors e_{it} independent
4. Uniqueness of factor rotation: Eigenvalues of $\Sigma_\Lambda \Sigma_F$ distinct
5. Bounded moments: $\mathbb{E}[\|F_t\|^4] < \infty$, $\mathbb{E}[\|\Lambda_i\|^4] < \infty$
   Simplification for presentation: $F_t \overset{\text{i.i.d.}}{\sim} (0, \Sigma_F)$, $\Lambda \overset{\text{i.i.d.}}{\sim} (0, \Sigma_\Lambda)$

- Standard assumptions on large dimensional approximate factor model

$\Rightarrow$ Conventional PCA consistent and asymptotically normal with full observations

Assumptions: Observational Pattern

Assumption 2: Observational Pattern

1. W independent of F and e $\Rightarrow$ Important: W can depend on Λ
2. "Sufficiently many" cross-sectional observed entries
$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top W_{it} \xrightarrow{p} \Sigma_{\Lambda,t} \qquad \text{full rank for all } t$$
3. "Sufficiently many" time-series observed entries
$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_t F_t^\top \xrightarrow{p} \text{full rank matrix for all } j$$
4. "Not too many" missing entries: $q_{ij} = \lim_{T\to\infty} |Q_{ij}|/T \geq \underline{q} > 0$ and
$\omega_{jj} = \lim_{N\to\infty} \frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{q_{ij,lj}}{q_{ij}q_{lj}}$ with $q_{ij,kl} = \lim_{T\to\infty}\frac{|\mathcal{Q}_{ij}\cap\mathcal{Q}_{kl}|}{T}$;
$\omega_j = \lim_{N\to\infty}\frac{1}{N^3}\sum_{i=1}^{N}\sum_{l=1}^{N}\sum_{k=1}^{N}\frac{q_{li,kj}}{q_{li}q_{kj}}$;
$\omega = \lim_{N\to\infty}\frac{1}{N^4}\sum_{i=1}^{N}\sum_{l=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{q_{li,kj}}{q_{li}q_{kj}}$ exist.

$\Rightarrow$ Very general pattern that can depend on latent factor model

- Special case: Missing at random: $\omega_{jj} = 1/p$, $\omega_j = 1$, $\omega = 1$
- Caveat: Observed entries proportional to N and T, but we show how to relax it

Asymptotic Results

Theorem 1: Loadings

Under Assumptions 1 and 2, it holds for $N, T \to \infty$ and $\sqrt{T}/N \to 0$:

$$\sqrt{T}\Gamma_{\Lambda,j}^{-1/2}(H^{-1}\tilde{\Lambda}_j - \Lambda_j) \xrightarrow{d} \mathcal{N}(0, I_k)$$

- $\Gamma_{\Lambda,j} = \omega_{jj} \cdot \Sigma_{\Lambda}^{\text{obs}} + (\omega_{jj} - 1)\Sigma_{\Lambda,j}^{\text{miss}}$
- Convergence rate is $\sqrt{T}$
- H is a standard rotation matrix
- Missing pattern weight $\omega_{jj} = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{q_{ij,lj}}{q_{ij}q_{lj}}$, $\omega_{jj} \geq 1$
  full observations: $\omega_{jj} = 1$, missing at random $\omega_{jj} = 1/p$
- Conventional covariance matrix $\Gamma_{\Lambda}^{\text{obs}} = \Sigma_F^{-1}\sigma_e^2$
- Variance correction term $\Sigma_{\Lambda,j}^{\text{miss}}$

Theorem 2: Factors

Under Assumptions 1 and 2, it holds for $N, T \to \infty$ and $\sqrt{N}/T \to 0$:

$$\sqrt{\delta}\Gamma_{F,t}^{-1/2}(H^\top \tilde{F}_t - F_t) \xrightarrow{d} \mathcal{N}(0, I_k)$$

- $\Gamma_{F,t} = \frac{\delta}{N}\Sigma_{F,t}^{\text{obs}} + \frac{\delta}{T}(\omega - 1)\Sigma_{F,t}^{\text{miss}}$
- Convergence rate is $\delta = \min(N, T)$
- Missing pattern weight $\omega = \lim_{N\to\infty} \frac{1}{N^4} \sum_{i=1}^N \sum_{l=1}^N \sum_{j=1}^N \sum_{k=1}^N \frac{q_{li,kj}}{q_{li}q_{kj}}$
  For full observations or missing at random: $\omega = 1$
- Conventional covariance matrix $\Sigma_{F,t}^{\text{obs}} = \Sigma_{\Lambda,t}^{-1}\sigma_e^2$
- Variance correction term $\Sigma_{F,t}^{\text{miss}}$

$\Rightarrow$ Inferential theory for common components C_{it} based on

$$\sqrt{\delta}\left(\tilde{C}_{it} - C_{it}\right) = \sqrt{\delta}\left(H^{-1}\tilde{\Lambda}_i - \Lambda_i\right)^\top F_t + \sqrt{\delta}\Lambda_i^\top \left(H^\top \tilde{F}_t - F_t\right) + o_p(1),$$

convergence rate is $\min\left(\sqrt{T}, \sqrt{N}\right)$.

Propensity-Weighted Estimator

Assumption 3: Conditional Observational Pattern

Assume observations depend on observed, time-invariant covariates $S \in \mathbb{R}^{N \times K}$:

1. The probability of $W_{it} = 1$ depends on S_i and $P(W_{it} = 1 | S_i) > 0$.
2. Conditional cross-sectional independence: W independent of Λ conditional on S.
3. W_{it} is independent of W_{js} conditional on S_i, S_j.

Alternative estimator for loadings and common components:

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

- $\tilde{F}^S = \tilde{F}$ for cross-section missing at random: $P(W_{it} = 1 | S_i)$ is the same for all i

$\Rightarrow$ A larger variance in general

$\Rightarrow$ Can be robust to selection bias when we use too few latent factors

Tests for Causal Effects

Treatment effect for staggered design with $T_{0,i}$ control and $T_{1,i}$ treated

$$Y_{it}^{(\theta)} = \underbrace{\Lambda_i^{(\theta)\top} F_t^{(\theta)}}_{C_{it}^{(\theta)}} + e_{it}, \quad \theta = \begin{cases} 1 & \text{treated (missing)} \\ 0 & \text{control (observed)} \end{cases}$$

We consider three different effects:

1. Individual treatment effect: $\tau_{it} = C_{it}^{(1)} - C_{it}^{(0)}$
2. Average treatment effect: $\tau_i = \frac{1}{T_{1,i}} \sum_{t=T_{0,i}+1}^{T} \tau_{it}$
3. Weighted average treatment effect: $\tau_{\beta,i} = (Z^\top Z)^{-1} Z^\top \tau_{i,(T_{0,i}+1):T}$

The test statistic for these three effects is build on the inferential theory of $\tilde{C}_{it}$.

Simulation

Simulation Design

Comparison between the four methods that provide inferential theory

1. XP: Our all-purpose method $\tilde{C}$
2. XP_{PROP}: Our propensity-weighted method $\tilde{C}^S$
3. JMS (Jin, Miao and Su (2020)): Assuming missing at random
4. BN (Bai and Ng (2020)): Combined block PCA

We compare the relative MSE $\sum_{i,t}(\tilde{C}_{it} - C_{it})^2 / \sum_{i,t} C_{it}^2$

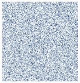
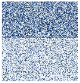
- The data generating process is $X_{it} = \Lambda_i^\top F_t + e_{it}$
- 2 factors
- $\Lambda_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2)$, $F_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_2)$ and $e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- $N = 250$, $T = 250$

All-purpose estimator: We allow for the most general observation pattern

$\Rightarrow$ Our method provides the most precise estimation in most cases

$\Rightarrow$ $\tilde{C}^S$ is very close to $\tilde{C}$, but less efficient

Simulation: Relative MSE for Different Methods

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

$\Rightarrow$ XP is precise for various observation patterns.

Simulation: Omitted Factor and Weak Factor

k	1				2			
$[\mu_{F,1}, \mu_{F,2}]$	[1,1]		[5,0.5]		[1,1]		[5, 0.5]	
$[\sigma_{F,1}, \sigma_{F,2}]$	[1,1]		[5,0.5]		[1,1]		[5, 0.5]	
Method	XP	XP_{PROP}	XP	XP_{PROP}	XP	XP_{PROP}	XP	XP_{PROP}
obs $C_{it}^{(0)}$	**0.227**	0.251	0.011	0.011	0.014	0.014	0.002	**0.003**
miss $C_{it}^{(0)}$	0.478	**0.288**	0.007	0.007	**0.044**	0.045	0.026	**0.023**
all $C_{it}^{(0)}$	0.314	**0.264**	0.009	0.009	**0.024**	0.025	0.014	**0.012**
$C_{it}^{(1)} - C_{it}^{(0)}$	0.481	**0.294**	0.008	**0.007**	0.052	0.052	0.026	**0.023**
$\beta_i^{(1)} - \beta_i^{(0)}$	0.168	**0.032**	0.002	0.002	**0.012**	0.013	0.008	**0.007**

⇒ XP_{PROP} is more precise if one factor is omitted

⇒ XP_{PROP} is more precise if the second factor is a weak factor

Conclusion

Conclusion

A new method for latent factor estimation with missing data:

- Simple all-purpose estimator for latent factor structure and data imputation
  Easy-to-adopt and applies to essentially any missing pattern
- Extension to propensity-weighted estimator:
  Less efficient but can be more robust to misspecification
- Confidence interval for each estimated entry under general and nonuniform observation patterns

Key application in causal inference:

- General tests for entry-wise and weighted treatment effects
- Generalizes conventional causal inference techniques to large panels and controls automatically for unobserved covariates

Empirical results in a companion paper:

- Weaker publication effect of investment anomaly strategies than naive before-after analysis
- Well-known strategies have no significant publication effect
  $\Rightarrow$ consistent with compensation for systematic risk
- 15% of strategies exhibit statistical significant reduction in average returns and outperformance of market