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 observation.

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
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
- Matrix completion: Athey, Bayati, Doudchenko, Imbens and Khosravi 2021

Matrix completion

- Independent sampling: Candes and Recht 2009, Mazumder, Hastie and Tibshirani 2010, Negahban and Wainwright 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

$$Y_{it} = \Lambda_i^T F_t + e_{it}$$

In matrix notation:

$$\mathbf{Y} = \mathbf{\Lambda} \mathbf{F}^T + \mathbf{e}$$

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

$\Rightarrow$ Estimate $\Lambda_i, F_t, C_{it}$ 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 $\Lambda$, 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:
\[
\tilde{\Sigma}_{ij} = \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} Y_{it} Y_{jt}, \text{ where } Q_{ij} = \{ t : W_{it} = 1 \text{ and } W_{jt} = 1 \} \text{ are times where both units are observed}
\]

**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: $\Sigma_F$ and $\Sigma_\Lambda$ full rank

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

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: $F_t \overset{i.i.d.}{\sim} (0, \Sigma_F)$, $\Lambda_i \overset{i.i.d.}{\sim} (0, \Sigma_\Lambda)$

- 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 \( \Lambda \)

2. “Sufficiently many” cross-sectional observed entries

\[
\frac{1}{N} \sum_{i=1}^{N} \Lambda_i \Lambda_i^T W_{it} \xrightarrow{P} \Sigma_{\Lambda,t} \quad \text{full rank for all } t
\]

3. “Sufficiently many” time-series observed entries

\[
\frac{1}{N} \sum_{i=1}^{N} \Lambda_i \Lambda_i^T \frac{1}{|Q_{ij}|} \sum_{t \in Q_{ij}} F_t F_t^T \xrightarrow{P} \text{full rank matrix for all } j
\]

4. “Not too many” missing entries: \( q_{ij} = \lim_{T \to \infty} |Q_{ij}|/T \geq q > 0 \) and

\[
\omega_{jj} = \lim_{N \to \infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{q_{ij,il}}{q_{ij}} \quad \text{with } q_{ij,kl} = \lim_{T \to \infty} \frac{|Q_{ij} \cap Q_{kl}|}{T};
\]

\[
\omega_j = \lim_{N \to \infty} \frac{1}{N^3} \sum_{i=1}^{N} \sum_{l=1}^{N} \sum_{k=1}^{N} q_{ij,kj} ;
\]

\[
\omega = \lim_{N \to \infty} \frac{1}{N^4} \sum_{i=1}^{N} \sum_{l=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} q_{lij,kj} \quad \text{exist.}
\]

\( \Rightarrow \) 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
Inferential Theory

**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 $\omega_{jj} = \lim_{N \to \infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{q_{ij,lj}}{q_{ij,q_{ij}}}$, $\omega_{jj} \geq 1$
- 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 $\Sigma_{\Lambda,j}^{\text{miss}}$
Inferential Theory

**Theorem 2: Factors**

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

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

- $\Gamma_{F,t} = \frac{\delta}{N} \Sigma_{F,t}^{\text{obs}} + \frac{\delta}{T} (\omega - 1) \Sigma_{F,t}^{\text{miss}}$
- Convergence rate is $\delta = \min(N, T)$
- Missing pattern weight $\omega = \lim_{N \to \infty} \frac{1}{N^4} \sum_{i=1}^N \sum_{l=1}^N \sum_{j=1}^N \sum_{k=1}^N \frac{q_{ii,kj}}{q_{ii}q_{kj}}$
  
  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

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

convergence rate is $\min \left( \sqrt{T}, \sqrt{N} \right)$. 
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 $\Lambda$ conditional on $S$.
3. $W_{it}$ is independent of $W_{js}$ conditional on $S_i, S_j$.

Alternative estimator for loadings and common components:

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

- $\tilde{F}^S = \tilde{F}$ 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

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

We consider three different effects:

1. Individual treatment effect: $\tau_{it} = C_{it}^{(1)} - C_{it}^{(0)}$

2. Average treatment effect: $\bar{\tau}_i = \frac{1}{T_{1,i}} \sum_{t=T_{0,i}+1}^{T} \tau_{it}$

3. Weighted average treatment effect: $\tau_{\beta,i} = (Z^\top Z)^{-1} Z^\top \bar{\tau}_{i,(T_{0,i}+1):T}$

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

14
Simulation
Simulation Design

Comparison between the four methods that provide inferential theory

1. **XP**: Our all-purpose method \( \tilde{C} \)
2. **XP\text{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 \( X_{it} = \Lambda_i^T F_t + e_{it} \)
- 2 factors
  - \( \Lambda_i \text{ i.i.d. } \mathcal{N}(0, I_2) \), \( F_t \text{ i.i.d. } \mathcal{N}(0, I_2) \) and \( e_{it} \text{ i.i.d. } \mathcal{N}(0, 1) \)
- \( N = 250, T = 250 \)

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

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

⇒ XP is precise for various observation patterns.
Simulation: Omitted Factor and Weak Factor

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[\mu_{F,1}, \mu_{F,2}]$</td>
<td>[1,1]</td>
<td>[5,0.5]</td>
</tr>
<tr>
<td>$[\sigma_{F,1}, \sigma_{F,2}]$</td>
<td>[1,1]</td>
<td>[5,0.5]</td>
</tr>
<tr>
<td>Method</td>
<td>XP</td>
<td>XP PROP</td>
</tr>
<tr>
<td>obs $C_{it}^{(0)}$</td>
<td>0.227</td>
<td>0.251</td>
</tr>
<tr>
<td>miss $C_{it}^{(0)}$</td>
<td>0.478</td>
<td>0.288</td>
</tr>
<tr>
<td>all $C_{it}^{(0)}$</td>
<td>0.314</td>
<td>0.264</td>
</tr>
<tr>
<td>$C_{it}^{(1)} - C_{it}^{(0)}$</td>
<td>0.481</td>
<td>0.294</td>
</tr>
<tr>
<td>$\beta_{i}^{(1)} - \beta_{i}^{(0)}$</td>
<td>0.168</td>
<td>0.032</td>
</tr>
</tbody>
</table>

$\Rightarrow$ **XP PROP** is more precise if one factor is omitted

$\Rightarrow$ **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
  ⇒ consistent with compensation for systematic risk
- 15% of strategies exhibit statistical significant reduction in average returns and outperformance of market