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

Ruoxuan Xiong and Markus Pelger

Stanford University

Motivation

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

- Macroeconomic data: staggered releases, mixed frequencies
- Program evaluation: Staggered treatment design
- Financial data: Mergers, new firms, bankruptcy
- Surveys: Panel attrition
- Recommender system: Netflix challenge

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

- 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
- Key application: Casual inference
  Counterfactual outcomes modeled as missing values
  Individual treatment effects at any time with unobserved factors

Motivating Example: Publication Effect on Investment Strategies

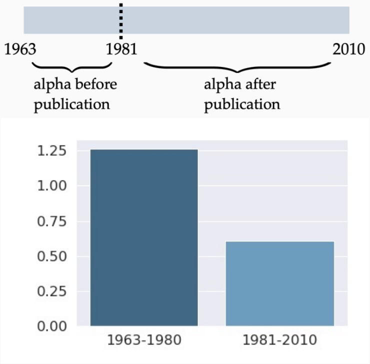
Question: Does academic publication of a strategy affect this strategy's return?

- Intuition: After publication traders exploit strategy and drive down profits
- Illustrative example (Banz 1981): Size strategy (small-minus-big portfolio) Smaller companies have higher average returns (published in 1981)
- Investment performance measure: Mean return in excess of a market index (alpha= outperformance relative to market)

Motivating Example: Publication Effect on Investment Strategies

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

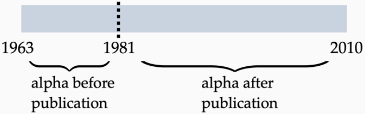
- Intuition: After publication traders exploit strategy and drive down profits
- Illustrative example (Banz 1981): Size strategy (small-minus-big portfolio)
  Smaller companies have higher average returns (published in 1981)
- Investment performance measure: Mean return in excess of a market index (alpha= outperformance relative to market)

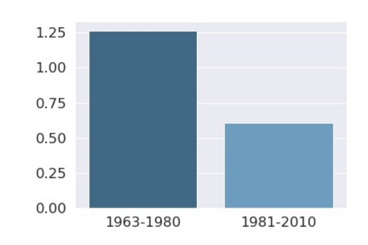


Motivating Example: Publication Effect on Investment Strategies

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

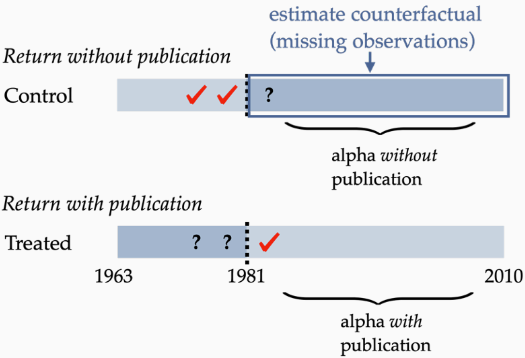
- Intuition: After publication traders exploit strategy and drive down profits
- Illustrative example (Banz 1981): Size strategy (small-minus-big portfolio)
  Smaller companies have higher average returns (published in 1981)
- Investment performance measure: Mean return in excess of a market index
  (alpha= outperformance relative to market)





Simple before-after analysis not appropriate!
It does not control for time-varying features.

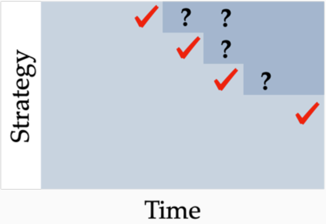
This Paper: A Causal Inference Approach



- Experiments have identical control and treatment groups
- Fundamental problem here: Only observe treated or control outcomes
- Our approach: Model counterfactual as missing observations and impute missing values
- Counterfactual = mimicking average of untreated observations

This Paper: New Methodology

- Large-dimensional panel data: Many strategies' returns over many periods.
- Complex treatment pattern: Strategies are published at different times with different probabilities



Observational pattern for the control panel

- No pre-specified model: Use general statistical factors to impute counterfactual returns without a prior what makes strategies similar
- A general causal inference approach: Model counterfactual outcomes as missing observations to obtain entry-wise control and test individual and weighted effects

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 et al. 2013, Pelger and Xiong 2020a+b, Lettau and Pelger 2020a+b
- Partial observations: Jin et al. 2020, Bai and Ng 2020, Cahan, Bai and Ng 2021, Stock and Watson 2002

Causal inference on panel data

- Difference in differences: Card 1990, Athey and Imbens 2018
- Synthetic control methods: Abadie et al. 2010, , Abadie et al. 15, Doudchenko and Imbens 2016, Li 2019
- Matrix completion: Athey et al. 2018

Matrix completion

- Independent sampling: Candes and Recht 2009, Mazumder et al 2010, Negahban and Wainright 2012
- Dependent sampling: Athey et al. 2018
- Independent sampling with inferential theory: Chen et al. 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:

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

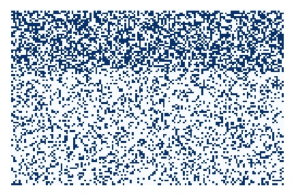
- N and T large
- Factors F_t explain common time-series movements
- Loadings Λ_i capture correlation between units
- Factors and loadings are latent and estimated from the data
- Common component $C_{it} = \Lambda_i^\top F_t$
- Idiosyncratic errors $\mathbb{E}[e_{it}] = 0$
- Number of factors k fixed

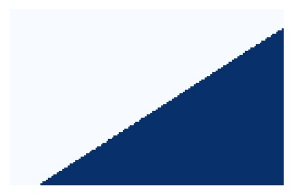
$\Rightarrow$ Estimate Λ_i, F_t, C_{it} and use estimated C_{it} to impute missing Y_{it}

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**

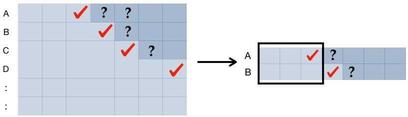




- Missing uniformly at random $P(W_{it} = 1) = p$
- Cross-section missing at random $P(W_{it} = 1) = p_t$
- Time-series missing at random $P(W_{it} = 1) = p_i$
- Staggered treatment adoption $P(W_{it} = 1) = p_{it}$ Once missing stays missing: $W_{is} = 0$ for $s \geq t$
- Mixed-frequency observations $P(W_{it} = 1) = p_{it}$ Equivalent to staggered design after reshuffling

Estimation of the Factor Model

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 $\tilde{C}_{it} = \tilde{\Lambda}_i^\top \tilde{F}_t$

Extension: A propensity weighted estimator: replace W_{it} by $\frac{W_{it}}{P(W_{it}=1|S_i)}$ in Step 3 for some observed covariates S_i

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: $e_{it} \stackrel{iid}{\sim} (0, \sigma_e^2)$, $\mathbb{E}[e_{it}^8] < \infty$
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: $F_t \stackrel{\text{i.i.d.}}{\sim} (0, \Sigma_F)$, $\Lambda \stackrel{\text{i.i.d.}}{\sim} (0, \Sigma_\Lambda)$

- Standard assumptions on large dimensional approximate factor model

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

Assumptions: Observation 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

$$\frac{1}{N}\sum_{i=1}^{N}\Lambda_i\Lambda_i^\top \frac{1}{|Q_{ij}|}\sum_{t\in Q_{ij}} F_tF_t^\top \xrightarrow{p} \text{full rank matrix for all } j$$

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 $\sqrt{T}/N \to 0$:

$$\boxed{\sqrt{T}(H^{-1}\tilde{\Lambda}_j - \Lambda_j) \xrightarrow{d} \mathcal{N}\left(0, \omega_{jj} \cdot \Sigma_{\Lambda}^{\text{obs}} + (\omega_{jj} - 1)\Sigma_{\Lambda,j}^{\text{miss}}\right)}$$

- 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 $\sqrt{N}/T \to 0$:

$$\boxed{\sqrt{\delta}(H^\top \tilde{F}_t - F_t) \xrightarrow{d} \mathcal{N}\left(0, \frac{\delta}{N}\Sigma_{F,t}^{\text{obs}} + \frac{\delta}{T}(\omega - 1)\Sigma_{F,t}^{\text{miss}}\right)}$$

- 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 $\min\left(\sqrt{T}, \sqrt{N}\right)$.

Tests for Causal Effects

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

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

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: $\tau_{\beta,i} = (Z^\top Z)^{-1} Z^\top \tau_{i,(T_{0,i}+1):T}$

Inferential theory of $\tilde{C}_{it}$ provides the test statistics for three effects.

Simulation

Simulation Design

Comparison between the four methods that provide inferential theory

1. XP_{SIM}: Our simple 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

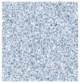
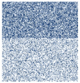
We compare the relative MSE $\sum_{i,t}(\tilde{C}_{it} - C_{it})^2 / \sum_{i,t} C_{it}^2$

- 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 $e_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$

$\Rightarrow$ Our method allows for the most general observation pattern

$\Rightarrow$ Out method provides the most efficient estimation

Simulation $N = 250$, $T = 250$

	Observation Pattern	W_{it}	XP	XP$_{\text{PROP}}$	JMS	BN
	Random	obs	**0.015**	**0.015**	0.023	348.300
		miss	**0.015**	**0.015**	0.021	363.885
		all	**0.015**	**0.015**	0.023	352.113
	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	347.082
		miss	**0.024**	**0.024**	0.067	360.409
		all	**0.021**	**0.021**	0.073	352.113
	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 the most precise

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