Reading: Projected Principle Component Analysis In Factor Models

Meilei Jiang

Department of Statistics and Operations Research

University of North Carolina at Chapel Hill

March 3, 2016

Overview

1. Semi-parametric Factor Model

2. Projected Principal Component Analysis

Factor Model

Classical factor model:

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + u_{it}, i = 1, \cdots, p, t = 1, \cdots, T. \quad (1)$$

- Observed data $\{y_{it}\}_{i \leq p, t \leq T}$, where i indexes variable, t indexes sample.
- Unobservable common factors: $\{f_{tk}\}_{k \leq K}$, where k indexes factor.
- Corresponding factor loadings for variable i: $\{\lambda_{ik}\}_{k \leq K}$.
- Idiosyncratic component that cannot be explained by the static common factors: u_{it}.

Factor Model

Matrix form of factor model (1):

$$\mathbf{Y} = \mathbf{\Lambda}\mathbf{F}' + \mathbf{U}. \tag{2}$$

- $\mathbf{Y} \in \mathbb{R}^{p\times T}, \mathbf{\Lambda} \in \mathbb{R}^{p\times K}, \mathbf{F} \in \mathbb{R}^{T\times K}, \mathbf{U} \in \mathbb{R}^{p\times T}$.
- *Goal*: accurately estimating the loading matrices $\mathbf{\Lambda}$ and unobserved factors $\mathbf{F}$.
- High dimension low sample size settings: $p \to \infty$ and T may or may not grow.

Semi-parametric Factor Model

Semi-parametric factor model:

$$\begin{aligned} \lambda_{ik} &= g_k(\mathbf{X}_i) + \gamma_{ik}, i = 1, \cdots, p, k = 1, \cdots, K. \\ y_{it} &= \sum_{k=1}^{K} \{g_k(\mathbf{X}_i) + \gamma_{ik}\} f_{tk} + u_{it}, i = 1, \cdots, p, t = 1, \cdots, T. \end{aligned} \tag{3}$$

- Covariates associated with the ith variables: $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})'$.
  - $\mathbf{X}_i$ can be individual characteristics(e.g. age, weight, clinical and genetic information.)
- Unknown nonparametric function: $g_k(.)$.
- The component of loading coefficient that cannot be explained by the covariates $\mathbf{X}_i$: γ_{ik}.
  - γ_{ik} have mean zero.
  - γ_{ik} is independent with u_{it} and $\mathbf{X}_i$

Semi-parametric Factor Models

Unknown nonparametric function $g_k(.)$: $g_k(\mathbf{X}_i) = \sum_{l=1}^{d} g_{kl}(X_{il})$

- Not depend on t: the loadings represent the cross-sectional heterogeneity only.
- $g_{kl}(X_{il}) = \sum_{j=1}^{J} b_{j,kl}\phi_j(X_{il}) + R_{kl}(X_{il}), k \leq K, i \leq p, l \leq d.$
- Basis functions: $\{\phi_1(x), \cdots, \phi_J(x)\}$.
- The sieve coefficients for g_{kl}: $\{b_{j,kl}\}_{j \leq J}$.
- Remaining function: $R_{kl}(X_{il})$
  - $\sup_x |R_{kl}(x)| \to 0$ as $J \to \infty$.

Semi-parametric Factor Models

Matrix form of semi-parametric factor model:

$$\begin{aligned} \mathbf{\Lambda} &= \mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}, \mathbb{E}(\mathbf{\Gamma}|\mathbf{X}) = 0, \mathbb{E}(\mathbf{G}(\mathbf{X})\mathbf{\Gamma}') = 0 \\ \mathbf{Y} &= \{\mathbf{G}(\mathbf{X}) + \mathbf{\Gamma}\}\mathbf{F}' + \mathbf{U}, \\ \mathbf{G}(\mathbf{X}) &= \mathbf{\Phi}(\mathbf{X})\mathbf{B} + \mathbf{R}(\mathbf{X}). \end{aligned} \tag{4}$$

- Matrix of sieve coefficients: $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_K) \in \mathbb{R}^{(Jd)\times K}$.
  - $\mathbf{b}_k^{'} = (b_{1,k1}, \cdots, b_{J,k1}, \cdots, b_{1,kd}, \cdots, b_{J,kd}) \in \mathbb{R}^{Jd}$.
- Matrix of basis function:$\mathbf{\Phi} = (\phi(\mathbf{X}_1), \cdots, \phi(\mathbf{X}_p))' \in \mathbb{R}^{p\times(Jd)}$.
  - $\phi(\mathbf{X}_i) = (\phi_1(\mathbf{X}_{i1}), \cdots, \phi_J(\mathbf{X}_{i1}), \cdots, \phi_1(\mathbf{X}_{id}), \cdots, \phi_J(\mathbf{X}_{id}) \in \mathbb{R}^{Jd}$.
- $\mathbf{R}(\mathbf{X}) = \{\sum_{l=1}^{d} R_{kl}(X_{il})\} \in \mathbb{R}^{p\times K}$

Semi-parametric Factor Models

The model (4) can be rewritten as

$$\mathbf{Y} = \{\mathbf{\Phi}(\mathbf{X})\mathbf{B} + \mathbf{\Gamma}\}\mathbf{F}' + \mathbf{R}(\mathbf{X})\mathbf{F}' + \mathbf{U} \tag{5}$$

- The sieve approximation error: $\mathbf{R}(\mathbf{X})\mathbf{F}'$.
- The idiosyncratic: $\mathbf{U}$.

Intuition

Classical Principal Component Analysis: running PCA on the original $\mathbf{Y}$ to estimate $\mathbf{F}$ and $\mathbf{\Lambda}$.

Projected Principal Component Analysis: running PCA on the projected data $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$ to estimate $\mathbf{F}$ and $\mathbf{P}\mathbf{\Lambda}$.

- $\mathcal{X}$ is a space spanned by $\mathbf{X} = \{\mathbf{X}_i\}_{i \leq p}$, which is orthogonal to $\mathbf{U}$.
- $\mathbf{P}$ is the projection matrix onto $\mathcal{X}$ and $\mathbf{P}\mathbf{U} \approx \mathbf{0}$.
- Analyzing the projected data $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$ is an approximately noiseless problem

Key Assumptions

Identification assumptions:

- $\frac{1}{T}\mathbf{F}'\mathbf{F} = \mathbf{I}_K$
- $\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}$ is a diagonal matrix with distinct entries.

The two assumptions mean that the columns of factors and loadings can be orthogonalized.

Key Assumptions

Genuine projection assumptions: There are positive constants $c_{\min}$ and $c_{\max}$ such that, with probability approaching one (as $p \to \infty$),

$$c_{\min} < \lambda_{\min}(p^{-1}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}) < \lambda_{\max}(p^{-1}\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}) < c_{\max}.$$

- Require the covariates $\mathbf{X}$ have nonvanishing explaining power on the loading matrix, so that the projection matrix $\mathbf{\Lambda}'\mathbf{P}\mathbf{\Lambda}$ has spiked eigenvalues.
- Rule out the case when $\mathbf{X}$ is completely unassociated with the loading matrix $\mathbf{\Lambda}$.

Projected Principal Component Analysis

Estimation of $\mathbf{F}$

- $\frac{1}{T}\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \frac{1}{T}\mathbf{Y}'\mathbf{P}\mathbf{Y} \approx \frac{1}{T}\mathbf{F}\boldsymbol{\Lambda}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{F}'$.
- $\frac{1}{T}\mathbf{Y}'\mathbf{P}\mathbf{Y}\mathbf{F} \approx \frac{1}{T}\mathbf{F}\boldsymbol{\Lambda}'\mathbf{P}\boldsymbol{\Lambda}$
- The columns of $\mathbf{F}/\sqrt{T}$ are approximately the first K PCs of $\frac{1}{T}\mathbf{Y}'\mathbf{P}\mathbf{Y}$.

Two estimations of $\mathbf{P}\boldsymbol{\Lambda}$

1. $\frac{1}{T}\mathbf{P}\mathbf{Y}\mathbf{F} = \mathbf{P}\boldsymbol{\Lambda} + \frac{1}{T}\mathbf{P}\mathbf{U}\mathbf{F} \approx \mathbf{P}\boldsymbol{\Lambda}$
2. The columns of $\mathbf{P}\boldsymbol{\Lambda}$ are approximately the first K PCs of $\frac{1}{T}\hat{\mathbf{Y}}\hat{\mathbf{Y}}'$.
   - $\frac{1}{T}\hat{\mathbf{Y}}\hat{\mathbf{Y}}' = \frac{1}{T}\mathbf{P}\mathbf{Y}\mathbf{Y}'\mathbf{P} = \mathbf{P}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{P} + \tilde{\boldsymbol{\Delta}} \approx \mathbf{P}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{P}$.
   - $(\frac{1}{T}\mathbf{P}\mathbf{Y}\mathbf{Y}'\mathbf{P})\mathbf{P}\boldsymbol{\Lambda} = \mathbf{P}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\mathbf{P}\boldsymbol{\Lambda})$

Projected Principal Component Analysis For Semi-parametric Factor Model

- $\mathbf{P} = \mathbf{\Phi}(\mathbf{X})(\mathbf{\Phi}(\mathbf{X})'\mathbf{\Phi}(\mathbf{X}))\mathbf{\Phi}(\mathbf{X})'$.
- $\hat{\mathbf{F}}/\sqrt{T}$ are the first K PCs of $\frac{1}{T}\mathbf{Y}'\mathbf{P}\mathbf{Y}$.
- $\hat{\mathbf{G}}(\mathbf{X}) = \frac{1}{T}\mathbf{P}\mathbf{Y}\hat{\mathbf{F}}$.
- $\hat{\mathbf{\Lambda}} = \mathbf{Y}\hat{\mathbf{F}}/T$.
- $\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Lambda}} - \hat{\mathbf{G}}(\mathbf{X}) = \frac{1}{T}(\mathbf{I} - \mathbf{P})\mathbf{Y}\hat{\mathbf{F}}$.

References

Jianqing Fan, Yuan Liao, And Weichen Wang (2016)
Projected principal component analysis in factor models.
Annals of Statistics 2016, Vol. 44, No. 1, 219254.

The End