Maximum Likelihood Estimation for Factor Analysis

Yuan Liao

University of Maryland

Joint worth Jushan Bai

June 15, 2013

High Dim. Factor Model

$$y_{it} = \lambda_i' f_t + u_{it} \quad i \leq N, t \leq T$$

- f_t: common factor
- λ_i: loading
- u_{it}: idiosyncratic component, correlated across i
- y_{it}: is the only observable.

Approx. factor model: $\Sigma_u = \mathrm{cov}(\mathbf{u}_t)$ is high-dimensional, non-diagonal

This talk

1. **Estimate factors and loadings efficiently**

   PC= $\arg\min_{\lambda_i, f_t} \sum_{i,t}(y_{it} - \lambda_i' f_t)^2$

   ignores cross-sec. hetero, corr. of u_{it}

2. **Estimate large covariance Σ_u**

   Key assumption: Sparsity

3. **Understand the impact of large covariance estimation on statistical inference**

   Improve efficiency, but technically challenging when $N > T$

Maximum likelihood Estimation

$$\mathrm{cov}(\mathbf{y}_t) = \Lambda \mathrm{cov}(f_t)\Lambda' + \Sigma_u$$

- Suppose $\mathbf{y}_t \sim \mathcal{N}(0, \mathrm{cov}(\mathbf{y}_t))$,

  Suppose $\mathrm{cov}(\mathbf{f}_t) = I$: identification, reduce # parameters
- **log-Likelihood**: Bai and Li 2012

$$L(\Lambda, \Sigma_u) = -\frac{1}{N}\log|\det(\Lambda\Lambda' + \Sigma_u)| - \frac{1}{N}\mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \Sigma_u)^{-1})$$

- Highly non-trivial:

  (1) highly nonlinear for Λ, (2) too many para. in Σ_u

Estimating Σ_u

1. Run SVD: $\mathbf{S}_y = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T = \sum_{j=1}^{K} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \sum_{j=K+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$.
2. Compute $\sum_{j=K+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T = (\hat{r}_{ij})$.
3. Apply (adaptive) thresholding (*Cai and Liu, 11*):

$$\widehat{\Sigma}_u = (s_{ij}(\hat{r}_{ij})).$$

$s_{ij}(\cdot)$: hard, soft, SCAD, adaptive Lasso,...

Assume Σ_u to be sparse, Fan, L, Mincheva (2013) showed:

$$\|\widehat{\Sigma}_u - \Sigma_u\| = O_p(m_N \sqrt{\frac{\log N}{T}} + m_N \frac{1}{\sqrt{N}}) \approx O_p(\sqrt{\frac{\log N}{T}})$$

finite-sample positive definite

Impact of large covariance estimation on inference

- Likelihood: $\frac{1}{N}\log|\det(\Lambda\Lambda' + \widehat{\Sigma}_u)| + \frac{1}{N}\mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \widehat{\Sigma}_u)^{-1})$
- Incorporating covariance matrices often improves efficiency

$$\hat{\theta} = f(D_T, \Sigma_u^{-1}), \quad \hat{\theta}^* = f(D_T, \widehat{\Sigma}_u^{-1})$$

$$\sqrt{T}(\hat{\theta} - \theta_0) \to \textit{Gaussian}$$

- Effect of estimating Σ_u is negligible:

$$\sqrt{T}(\hat{\theta} - \hat{\theta}^*) = \sqrt{T}\mathbf{A_1}(\boldsymbol{\Sigma}_\mathbf{u}^{-1} - \widehat{\boldsymbol{\Sigma}}_\mathbf{u}^{-1})\mathbf{A_2} + o_p(1) = o_p(1)$$

- Goal: to show

$$\sqrt{T}\mathbf{A_1}(\Sigma_{\mathbf{u}}^{-1} - \widehat{\Sigma}_{\mathbf{u}}^{-1})\mathbf{A_2} = o_p(1)$$

- Classical low dim. problems, consistency suffices $\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\| = o_p(1)$, e.g., efficient GMM

- However, in high dimensions ($N > T$), Highly NON-trivial, even if $\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\|$ achieves the optimal rate, nearly root-T.

  Optimal "absolute converge" is restrictive for inference when $N > T$

- For efficient factor analysis, we need

$$\frac{\sqrt{T}}{N}\Lambda'(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1})\Lambda = o_p(1)$$

However, $\leq \frac{\sqrt{T}}{N}\|\Lambda\|^2\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\| \approx O_p(\frac{\sqrt{T}}{N} \times N \times \frac{1}{\sqrt{T}}) \neq o_p(1)$

- For inference in high-dimensional models, need to directly evaluate "weighted convergence"

$$\sqrt{T}A_1(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1})A_2$$

- Intuition: averaged error converges faster weighted convergence $A_1(\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1})A_2$ is faster than "absolute convergence" $\|A_1\|\|A_2\|\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\|$.
- need **sparsistency**: with probability approaching one,
  If $\sigma_{u,ij} \neq 0$, $\hat{\sigma}_{ij} \neq 0$;
  If $\sigma_{u,ij} \approx 0$, $\hat{\sigma}_{ij} = 0$

$$\hat{\Lambda} = \arg\min_{\Lambda} \log|\det(\Lambda\Lambda' + \widehat{\Sigma}_u)| + \mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \widehat{\Sigma}_u)^{-1})$$

$$\hat{f}_t = \arg\min_{f_t} (\mathbf{y}_t - \hat{\Lambda} f_t)' \widehat{\Sigma}_u^{-1} (\mathbf{y}_t - \hat{\Lambda} f_t)$$

Theorem: $\forall j \leq N,\ t \leq T$,

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_t u_{jt} + o_p(1)$$

$$\sqrt{N}(\hat{f}_t - f_t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_j u_{jt} + o_p(1)$$

$$\max_{j \leq N} \|\lambda_j - \lambda_j\| = O_p(m_N \sqrt{\frac{\log N}{T}} + \frac{m_N}{\sqrt{N}})$$

$$\max_{t \leq T} \|\hat{f}_t - f_t\| = O_p(m_N \sqrt{\frac{\log N}{T}} + \frac{m_N}{\sqrt{N}})(\log \mathbf{T})^{\mathbf{r}}$$

Alternative: Penalized maximum likelihood

- Estimate Λ, Σ_u simultaneously, max:

$$-\frac{1}{N}\log|\det(\Lambda\Lambda' + \Sigma_u)| - \frac{1}{N}\mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \Sigma_u)^{-1}) - \sum_{i \neq j} w_{ij}|\sigma_{u,ij}|$$

- weight w_{ij}: Lasso, adaptive lasso, SCAD, MCP....
- Still, technically highly non-trivial

$$\frac{1}{N}\|\Lambda - \hat{\Lambda}\|^2 = o_p(1), \quad \frac{1}{N}\|\Sigma_u - \widehat{\Sigma}_u\|^2 = o_p(1)$$

Jushan and I spent months to prove this.

Numerical Examples

- 4 estimators are compared:

  PC: $\min \sum_{i,t}(y_{it} - \lambda_i' f_t)^2$

  diag ML (Bai and Li 12):

  $-\log|\det(\Lambda\Lambda' + \mathrm{diag}(\Sigma_u))| - \mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \mathrm{diag}(\Sigma_u))^{-1})$

  two-step $-\log|\det(\Lambda\Lambda' + \widehat{\Sigma}_u)| - \mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \widehat{\Sigma}_u)^{-1})$

  penalized ML

  $-\log|\det(\Lambda\Lambda' + \Sigma_u)| - \mathrm{tr}(\mathbf{S}_y(\Lambda\Lambda' + \Sigma_u)^{-1}) - \sum_{i \neq j} w_{ij}|\sigma_{u,ij}|$

- Algorithm: either **EM** (Bai and Li) or **EM+Majorize minimize** (Bien and Tibshirani 11)

				Penalized ML			
		PCA	DML	$\gamma = 1$		$\gamma = 5$	
T	N			$\mu_T = 0.08$	$\mu_T = 0.3$	$\mu_T = 0.08$	$\mu_T = 0.3$
50	50	0.205	0.199	0.212	0.222	0.230	0.234
50	100	0.429	0.558	0.591	0.613	0.627	0.631
50	150	0.328	0.470	0.494	0.495	0.515	0.507
100	50	0.496	0.519	0.560	0.537	0.558	0.537
100	100	0.394	0.574	0.621	0.648	0.648	0.658
100	150	0.774	0.819	0.837	0.829	0.840	0.836

Testing CAPM

$$y_{it} = \alpha_i + \lambda_i' f_t + u_{it}, \quad H_0 : \alpha_i = 0, \forall i$$

- Wald test $cT\hat{\alpha}'\Sigma_u^{-1}\hat{\alpha}$, under H_0:

$$\frac{cT\hat{\alpha}'\Sigma_u^{-1}\hat{\alpha} - N}{\sqrt{2N}} \to^d \mathcal{N}(0,1).$$

- Replacing with $\widehat{\Sigma}_u^{-1}$ is difficult:

$$\frac{T\hat{\alpha}'(\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1})\hat{\alpha}}{\sqrt{2N}} \leq \frac{T\|\hat{\alpha}\|^2\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\|}{\sqrt{2N}} \neq o_p(1)$$

$$T \times \frac{N}{T} \times \frac{1}{\sqrt{T}} \times \frac{1}{\sqrt{N}} = \sqrt{\frac{N}{T}}$$

Summary

Factor analysis

- Taking into account cross-sectional hetero. and corr. via Σ_u^{-1}.

Covariance estimation

- Estimating large error covariance is important for factor analysis and panel data models
- Effect of estimating covariance on inference is negligible, but tech. non-trivial