Theory and Applications of High Dimensional Covariance Matrix Estimation

Yuan Liao
Princeton University

Joint work with
Jianqing Fan and Martina Mincheva

December 14, 2011

Outline

1. Applications of large covariance matrix estimation
2. Thresholding using Contaminated Data
3. Observable Factors
4. Unobservable factors
5. Simulation Studies
6. Conclusions

Needs of Large Covariance Matrix

- Portfolio Management in Finance (Markowitz 52)
- Classification (e.g. Fisher discriminant, Shao et al. 11)
- Network and graphical models
- High frequency data

Examples of high dimensional covariance matrices

Finance

Jagannathan and Ma (2003):

- p assets with returns at time t (% change values):

$$\mathbf{y}_t = (y_{1t}, y_{2t}, ..., y_{pt})'.$$

- *Portfolio* (proportions of total amount of money to invest):

$$\mathbf{w} = (w_1, ..., w_p)', \quad \sum_{i=1}^{p} w_i = 1.$$

- Return at time $t+1$: $\mathbf{w}' y_{t+1}$.
- Risk $\mathrm{var}(\mathbf{w}' y_{t+1}) = \mathbf{w}'\Sigma\mathbf{w}$, where $\Sigma = \mathrm{var}(\mathbf{y}_t)$.

Optimal Portfolio

Markowitz (1952):

- Expect to earn μ at time $t+1$,

$$\min_{\mathbf{w}} \mathbf{w}'\Sigma\mathbf{w} \quad \text{s.t. } \mathbf{w}'\mathbf{1} = 1 \quad \mathbf{w}'E\mathbf{y}_{t+1} = \mu.$$

- Solution: $\mathbf{w}^* = c_1\Sigma^{-1}E\mathbf{y}_{t+1} + c_2\Sigma^{-1}\mathbf{1}$.

- Typical data set of US stock market may contain $p = 4883$ stocks, $T = 60$ months.

Classification

Disease classification using bioinformatic data (Shao et al. 11)

- Two types of human acute leukemias
  - acute myeloid leukemia (AML)
  - acute lymphoblastic leukemia (ALL)
- Distinguishing ALL from AML is crucial for successful treatment
- Classification based solely on $p = 1,714$ genes
- A training data set
  - 47 ALL $\sim N_p(\mu_1, \Sigma)$
  - 25 AML $\sim N_p(\mu_2, \Sigma)$
- Fisher discr. $(\mu_1 - \mu_2)\Sigma^{-1}(\mathbf{x} - \bar{\mu}) \geq 0$
- p is much larger than n.

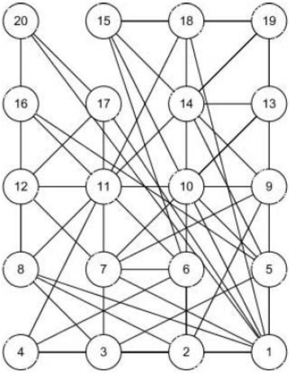
Graphical Modeling

Graphic model (Meinshausen and Bühlmann 06, Zhou et al. 11)

- Vertices: components of $\mathbf{y} = (y_1, ..., y_p)' \sim N_p(0, \Sigma)$.
- Edges: the conditional dependence

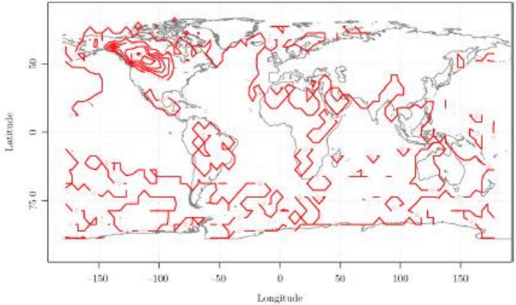
  No edge between i and $j \Longleftrightarrow y_j \perp y_j | \text{other}$

- Precision matrix: $\Sigma^{-1} = (\omega_{ij})_{p \times p}$
- $\omega_{ij} = 0$ iff y_i and y_j are cond. indep.
- A simple network graph corr. to a sparse precision matrix.



Climate Data

- 157January temperatures are recorded (1850-2006) by $p = 2,592$ stations over the world.
- Study the climate correlations among geographical regions in North America and Eurasia. (Bickel and Levina 08)



Statistical Inference

1. High dim. generalized least-squares
2. High dim. seemingly unrelated regression
3. Testing CAPM (mean-variance efficiency of market)

Challenge of Dimensionality

Estimating high-dim. covariance matrices is challenging.

- Suppose we have 2,000 stocks to be managed. There are 2m free parameters.
- Yet, 1-year daily returns yield only about $T = 250$. Hard to accurately estimated it.
- Risk: $\mathbf{w}'\widehat{\Sigma}\mathbf{w}$, Allocation: $\hat{c}_1\widehat{\Sigma}^{-1}\mathbf{1} + \hat{c}_2\widehat{\Sigma}^{-1}\bar{\mathbf{y}}$.
  Accumulation of millions of errors can have a huge effect.
- Sample covariance matrix is degenerate.

Approaches to Dimension Reduction

Target: $\Sigma_y = \mathrm{var}(\mathbf{y})$.

Strict Factor Model (Fan et al. 08)

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t \leq T.$$

- $\mathbf{f}_t$ =common factors $\mathbf{B}$ =factor loadings
  $\mathbf{u}_t$ =idiosyncratic component
- Fama-French-3-factor-model (Fama and French 92)
  $\mathbf{y}_t$ represents the stock returns.
  $K = 3$ known factors.

Sparsity based model (Bickel and Levina 08a,b)

thresholding penalized likelihood

Strict Factor Model

- $y_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}$. Implied covariance:

$$\Sigma_y = \mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \Sigma_u.$$

  Assume Σ_u is diagonal.

- After common factors are taken out, industry-specific factors are still correlated within the industry. (Connor and Korajczyk 93)
- Σ_u is diagonal only if K is large.
- We allow for non-diagonal Σ_u: approximate factor model (Chamberlain and Rothchild 83, Bai and Ng 02).

Sparsity Based Model

- Covariance matrix, precision matrix.
- Sparsity in Σ_y rarely occurs in many applications.
  - Returns depend on equity market risks
  - Housing prices depend on economic health
  - Gene expressions depend on cytokines

Contributions of This Talk

Model-based method

$$\Sigma_y = \mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \Sigma_u.$$

- Σ_u is sparse

  $$m_T = \max_{i \le p} \sum_{j \le p} I(\sigma_{u,ij} \neq 0)$$

  generalizable to l_q-norm.
- Investigate the estimation effect using contaminated data.
- In many cases the factors are unobservable.
- Examine impact of dependence data.

Sparse-based Matrix Estimation

Thresholding Bickel and Levina 08a, Rothman, Levina and Zhu 09, Cai and Zhou 11, etc

Adaptive thresholding Cai and Liu 11.

Banding Pourahmadi and Wu 03, Bickel and Levina 08b.

Penalization Lam and Fan 09, Bien and Tibshirani 11.

Bayesian Bhattacharya and Dunson 11.

Sparse PCA Zou, Hastie and Tibshirani 04, Jung and Marron 09, Johnstone and Lu 09.

Covariance Estimation with Contaminated Data

Suppose

$$\mathbf{u} \sim (0_p, \Sigma_u), \quad \Sigma_u \text{ sparse.}$$

- $\mathbf{u}_1, ..., \mathbf{u}_T$ are iid copies of $\mathbf{u}$.
- Instead of $\{\mathbf{u}_t\}_{t=1}^T$, we only observe contaminated $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$.
- Examples of contaminated data:
  - regression residuals
  - measurement of error
- Goal: estimate Σ_u.

1. Obtain sample covariance ($\widehat{\sigma}_{ij}$) based on $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$.
2. Apply (adaptive) thresholding (Cai and Liu 11):

$$\widehat{\Sigma}_u^{\mathcal{T}} = (\widehat{\sigma}_{ij}^{\mathcal{T}}), \quad \widehat{\sigma}_{ij}^{\mathcal{T}} = \widehat{\sigma}_{ij} I(|\widehat{\sigma}_{ij}|/\hat{\theta}_{ij} \geq \omega_T) \quad \hat{\theta}_{ij} = SD\{\hat{u}_{it}\hat{u}_{jt}\}_{t=1}^T$$

Theorem 1

Under Assumption A, with $\omega_T = (\frac{\log p}{T})^{1/2} + a_T$,

$$\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\| = O_p(\omega_T m_T) = \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|,$$

where $\max_{i \leq p} \frac{1}{T}\sum_{t=1}^T (\hat{u}_{it} - u_{it})^2 = O_p(a_T^2)$.

Assumption A:

- Σ_u is well conditioned.
- $P(|u_{it}| > s) \leq \exp(-(s/b)^r)$.
- $\max_{i,t} |\hat{u}_{it} - u_{it}| = o_p(1)$.

Observable Factors

Observable Factors

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, ..., T.$$

1. Run OLS to obtain loadings $\widehat{\mathbf{B}}$ and residuals $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$.
2. Obtain sample covariance $\widehat{\Sigma}_u$ based on $\{\widehat{\mathbf{u}}_t\}_{t=1}^T$.
3. Apply (adaptive) thresholding to get $\widehat{\Sigma}_u^{\mathcal{T}}$.
4. Compute $\widehat{\Sigma}_y = \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{f}_t)\widehat{\mathbf{B}}' + \widehat{\Sigma}_u^{\mathcal{T}}$.

Theorem 2

Under Assumptions A, B(below),

$$\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\| = O_p(m_T \omega_T) = \|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|,$$

where $\omega_T = K\left(\sqrt{\frac{\log p}{T}}\right)$ *is the threshold.*

- Minimax rate in Cai and Zhou (2010) for finite K.

Assumption B:

- $\{\mathbf{f}_t\}$ is stationary and ergodic.
- $\{\mathbf{u}_t\}$ and $\{\mathbf{f}_t\}$ are independent.
- Exponential α-mixing: $\alpha(t) \le \exp(-Ct^\gamma)$
- Exponential tail: $\forall s > 0$, $P(|f_{it}| > s) \le \exp(-(s/b)^r)$.

Accuracy of Residuals a_T

- Errors in estimating residuals:

$$\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 \leq \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{f}_t\|^2 \max_i \|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|^2.$$

- Need to bound $\max_{ij} |\frac{1}{T} \sum_{t=1}^{T} f_{it} u_{jt}|$ for dependent seq.
- Bernstein ineq. (Merlevède et al. 09)

$$P(|\frac{1}{T} \sum_{t=1}^{T} f_{it} u_{jt}| > s) \leq T \exp\left(-\frac{(Ts)^{r_3}}{C_1}\right) + \exp\left(-\frac{T^2 s^2}{C_2(1 + TC_3)}\right)$$

$$+ \text{small}.$$

- Hence, $\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u}_{it}|^2 = O_p\left(\frac{K^2 \log p}{T}\right)$.

Estimation of Σ_y

Some insights on the phenomenon: toy example.

- We know $\mathbf{b}_i = (1, 0, ..., 0)'$, $\Sigma_u = I_p$.

$$\widehat{\Sigma}_y = \mathbf{B}\widehat{\mathrm{cov}}(\mathbf{f}_t)\mathbf{B}' + I_p.$$

- The estimated errors are accumulated

$$\|\widehat{\Sigma}_y - \Sigma_y\| = O_p(\frac{p}{\sqrt{T}}).$$

Consider a different norm (entropy loss):

$$\frac{1}{p}\mathrm{tr}[(\widehat{\Sigma}_y\Sigma_y^{-1} - I_p)^2] = \frac{1}{p}\|\Sigma_y^{-1/2}(\widehat{\Sigma}_y - \Sigma_y)\Sigma_y^{-1/2}\|_F^2$$

Theorem 3

If $\lambda_{\min}(\frac{1}{p}\sum_{i=1}^{p}\mathbf{b}_i\mathbf{b}_i') > C$, *and* $\lambda_{\min}(\mathrm{cov}(\mathbf{f}_t)) > C$, *then*

$$\frac{1}{p}\mathrm{tr}(\widehat{\Sigma}_y\Sigma_y^{-1} - I_p)^2 = O_p\left(\frac{pK^2}{T^2} + \frac{m_T^2K^2\log p}{T}\right),$$

$$\|(\widehat{\Sigma}_y)^{-1} - \Sigma_y^{-1}\|^2 = O_p\left(\frac{m_T^2K^2\log p}{T}\right),$$

$$\|\widehat{\Sigma}_y - \Sigma_y\|_\infty^2 = O_p\left(\frac{K^2\log p + K^4\log T}{T}\right).$$

Recall $\frac{1}{p}\mathrm{tr}(\widehat{\Sigma}_{y,sam}\Sigma_y^{-1} - I_p)^2 = O_p(\frac{p}{T})$ (Fan et al. 08).

Unobservable Factors

Unobservable Factors

- In many applications, $\{\mathbf{f}_t\}_{t=1}^T$ are unobservable. (Forni et al. 00)
- PC decomposition:

$$\widehat{\Sigma}_{y,sam} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i' + \sum_{i=K+1}^{p} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i'$$

- Thresholding $\sum_{i=K+1}^{p} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i' \Rightarrow \widehat{\Sigma}_u^{\mathcal{T}}$.
- Estimator:

$$\widehat{\Sigma}_y \equiv \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i' + \widehat{\Sigma}_u^{\mathcal{T}}.$$

Least squares point of view

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + u_{it}.$$

- Need to estimate $\mathbf{B}$ and $\{\mathbf{f}_t\}_{t=1}^{T}$.
- Minimize (Bai, 03):

$$(\widehat{\mathbf{b}}_i, \widehat{\mathbf{f}}_t) = \arg\min_{\mathbf{b}_i, \mathbf{f}_t} \frac{1}{Tp} \sum_{t=1}^{T} \sum_{i=1}^{p} (y_{it} - \mathbf{b}_i' \mathbf{f}_t)^2.$$

$$\text{s.t. } \frac{1}{p} \sum_{i=1}^{p} \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i' = I_K, \quad \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t' \text{ diagonal.}$$

- Solution $\widehat{\mathbf{B}}$: K largest eigenvectors of $\widehat{\Sigma}_{y,sam}$.

- $\widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\widehat{\mathbf{f}}_t)\widehat{\mathbf{B}}' = \sum_{i=1}^K \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i'$.
- $\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t$ consistently estimates $\mathbf{b}_i'\mathbf{f}_t$ as $p \to \infty$, $T \to \infty$.
- Residual: $\hat{u}_{it} = y_{it} - \widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t$.

$$\max_i \frac{1}{T}\sum_{t=1}^T (u_{it} - \hat{u}_{it})^2 = O_p\left(\frac{K^2 \log p}{T} + \frac{K^6}{p}\right).$$

▸ Rate

- Decomposition:

$$\begin{aligned} \widehat{\Sigma}_{y,sam} &= \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\widehat{\mathbf{f}}_t)\widehat{\mathbf{B}}' + \widehat{\Sigma}_{u,sam} \\ &= \sum_{i=1}^K \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i' + \sum_{i=K+1}^p \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i'. \end{aligned}$$

Factor model v.s. PCA

- Factor model and PCA are asymptotically equivalent for high dimensional data.
- Suppose $\mathrm{cov}(\mathbf{f}_t) = I_K$, $\mathbf{B}'\mathbf{B}$ is diagonal. Chamberlain and Rothchild (1983) showed that the loadings can be obtained from eigenvalues.
- **Result**:

$$\|\xi_j - \|\widetilde{\mathbf{b}}_j\|^{-1}\widetilde{\mathbf{b}}_j\| = O_p(\frac{1}{p}\lambda_{\max}(\Sigma_u)).$$

Theorem 4

Under Assumptions A, B and C, Assumption C

$$\|\widehat{\Sigma}_u^{\mathcal{T}} - \Sigma_u\| = O_p\left(m_T K \sqrt{\frac{\log p}{T}} + \underbrace{\frac{m_T K^3}{\sqrt{p}}}_{\textit{impact of unknown factors}} \right),$$

$$\|(\widehat{\Sigma}_u^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\| = \textit{ the same order.}$$

- The impact of estimating unobservable factors vanishes when $p >> T$.
- p can be "ultra-high".

Estimation of Σ_y

Define

$$\|\widehat{\Sigma}_y - \Sigma_y\|_\Sigma^2 = \frac{1}{p}\mathrm{tr}(\widehat{\Sigma}_y\Sigma_y^{-1} - I_p)^2.$$

Theorem 5

When $\{\mathbf{f}_t\}$ are unobservable,

$$\|(\widehat{\Sigma}_y)^{-1} - \Sigma_y^{-1}\| = O_p\left(m_T K\sqrt{\frac{\log p}{T}} + \frac{m_T K^3}{\sqrt{p}}\right),$$

$$\|\widehat{\Sigma}_y - \Sigma_y\|_\Sigma = O_p\left(\frac{\sqrt{p}K}{T} + \frac{m_T K\sqrt{\log p} + K^2}{\sqrt{T}} + \frac{m_T K^3}{\sqrt{p}}\right),$$

$$\|\widehat{\Sigma}_y - \Sigma_y\|_\infty^2 = O_p\left(\frac{K^3\sqrt{\log K} + K\sqrt{\log p}}{\sqrt{T}} + \frac{K^3}{\sqrt{p}}\right).$$

Remarks

- Many other regularization methods can also be employed.
  - Generalized threshold (Antoniadis and Fan 01, Rothman et al. 09) $\Rightarrow$ Generalized adaptive thresholding (Cai and Liu 11)
  - Penalized likelihood (Bien and Tibshirani 11, Luo 11)
- Encompasses many estimators as special cases
  - Applied to correlation matrix of **u**
    $\lambda = 0$ $\Rightarrow$sample cov. $\lambda = 1$ $\Rightarrow$ strict factor model.
  - $K = 0$ $\Rightarrow$ sparse matrix (Bickel and Levina 08, Cai and Liu 11)

Numerical results

Simulation Designs

Model Design Fama-French 3-factor model with parameters calibrated from the market. $N_{sim} = 200$.

Calibration

- Using 30 industrial portfolios from 1/1/09 to 12/31/10 ($T = 300$), fit the Fama-French model.
- Summarize 30 factor loadings by (μ_B, Σ_B) and residuals by (μ_s, σ_s).
- Fit VAR(1) model to $\mathbf{f}_t$ and obtain model parameters.

Detailed Simulation

Generation of Factors: $\{\mathbf{f}_t\}_{t=1}^T \sim$VAR(1).

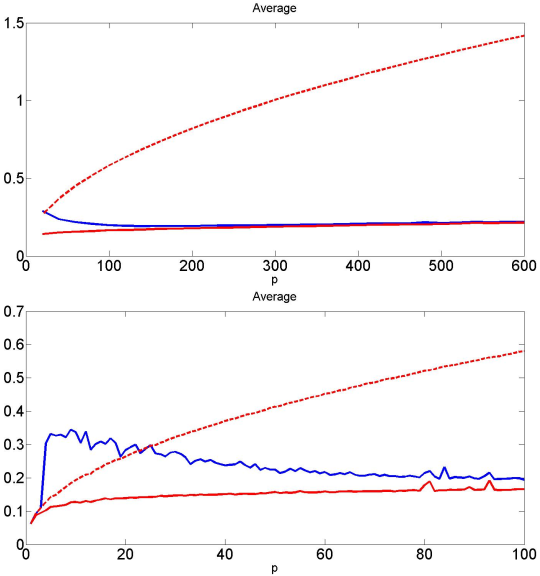
Simulation of returns: $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$:

- factor loadings: $\mathbf{b}_i \sim N_3(\mu_B, \Sigma_B)$.
- noise level: $\sigma_i \sim \Gamma(\alpha, \beta)$ with mean μ_s and SD σ_s.
- noise vector $\mathbf{u}_t \sim N_p(0, \Sigma_u)$, where

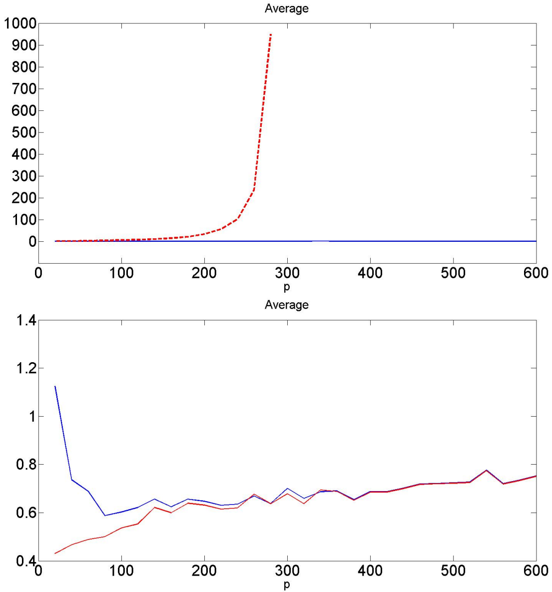
$$\Sigma_u = D\Sigma_0 D'$$

where Σ_0 is a sparse correlation matrix.

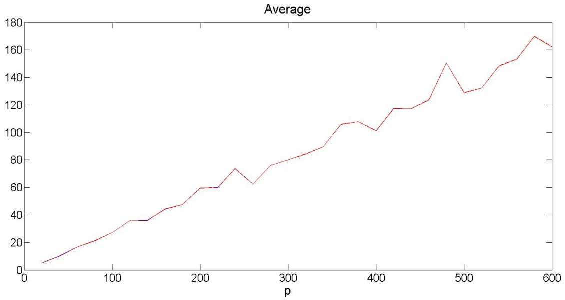
Simulation Results: $\mathrm{tr}^{1/2}(\widehat{\Sigma}_y \Sigma_y^{-1} - I_p)^2$



Simulation Results: $\|(\widehat{\Sigma}_y)^{-1} - \Sigma_y^{-1}\|$



Simulation Results: $\|\widehat{\Sigma}_y - \Sigma_y\|$



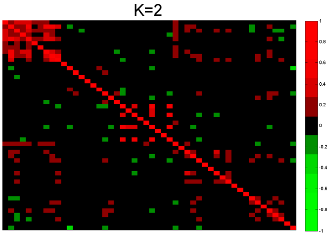
Empirical Example

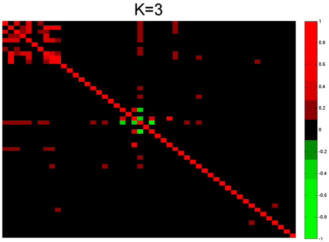
- $p = 50$ stocks from CRSP database. 5 industries, 10 companies each
  1. consumer goods & apparel clothing
  2. financial-credit services
  3. health care
  4. services-restaurants
  5. utilities-water
- $T = 252$ daily returns, Jan 2010-Dec 2010
- eigenvalues of sample covariance:

$$\lambda_1 = 0.010, \quad \lambda_2 = 0.004, \quad \lambda_3 = 0.004, \quad \lambda_{i \geq 4} < 0.002$$

- threshold has been chosen by leave-one-out CV.

Thresholded error correlation matrix





Conclusions

Conditional Sparsity widens scope of applicability

- direct sparsity rarely occurs in Econ and Fin, and biology.
- strict factor model is also very restrictive.

Method:

- easy to compute: keep first K PCs, threshold remaining
- avoid numerical minimization w/ pd. constraints.

Results:

- convergence rates for weighted l_2 loss, spectral norm, l_∞
- when estimating Σ_u, Σ^{-1}: $\log p \ll T^a$
- PCA and factor model are asym. equiv. for high dim. data

Impacts:

- impact of unob. factor vanishes for high dim.
- cov. estimation using contaminated data
- weakly dependent processes with mixing conditions

Assumption C

Assumption 1

1. $\{\mathbf{u}_t\}_{t=1}^T$ *is stationary and ergodic*
2. $E[p^{-1/2}(\mathbf{u}_s'\mathbf{u}_t - E\mathbf{u}_s'\mathbf{u}_t)]^4 < M$,
3. $E\|(pK)^{-1/2}\sum_{i=1}^p \mathbf{b}_i u_{it}\|^4 < M$.
4. $p^{-1}\mathbf{B}'\mathbf{B}$ *is well conditioned for all large* p.

jumpback

Some references

- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*. **71** 135-171.
- CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation.
  To appear in *J. Amer. Statist. Assoc.*.
- CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure and mean-variance analyssi in large asset markets. *Econometrica*. **51** 1305-1324.
- FAMA, E. and FRENCH, K. (1992). The cross-section of expected stock returns. *Journal of Finance*. **47** 427-465.
- FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*. **147** 186-197