Estimation of Dependence Matrices

- Covariance (correlation) matrices
- Precision matrix
- Thresholding approach
- Graphical lasso and variants
- Spectral density matrices

Dependence Matrices of Interest

- Covariance/correlation matrix

$$
\begin{array}{lll}
\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1,\dots,d} & \text{with} & \sigma_{ij} = \mathrm{Cov}(X_{i,t}, X_{j,t}) \\
R = (\rho_{ij})_{i,j=1,\dots,d} & \text{with} & \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}
\end{array}
$$

- Precision Matrix : This is just the inverse of covariance matrix

$$\Theta = (\theta_{ij})_{i,j=1,\dots,d} = \boldsymbol{\Sigma}^{-1}$$

Its main interest is due to

$$
\begin{aligned}
-\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} &= \text{partial correlation between } X_{i,t} \text{ and } X_{j,t} \\
&= \mathrm{Corr}(X_{i,t}, X_{j,t} \mid X_{-(ij),t})
\end{aligned}
$$

Gaussian Graphical Model (GGM)

- Let $\mathbf{X} = (X_1, \ldots, X_d)' \sim MVN(\mathbf{0}, \mathbf{\Sigma})$.
- Denote $V = \{1, 2, \ldots, d\}$ be the node.
- Covariance matrix, $\mathrm{Cov}(\mathbf{X}) = \Sigma$, gives marginal dependence

$$X_i \perp\!\!\!\perp X_j \iff \mathrm{Cov}(X_i, X_j) = \sigma_{ij} = 0$$

- Inverse covariance matrix (precision matrix) gives "conditional" dependence

$$X_i \perp\!\!\!\perp X_j \mid X_{-(ij)} \iff \theta_{ij} = 0$$

- This is also known as Markov Random Field (MRF).

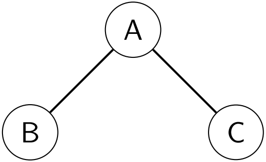
GGM

- Graph summarizes relationships between nodes $V = \{1, 2, \ldots, d\}$ and set E of edges

$$\theta_{ij} = 0 \Leftrightarrow i \nsim j$$

- For example

$$\boldsymbol{\Theta} = \begin{array}{c c} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{pmatrix} 1 & .2 & .3 \\ .2 & 2 & 0 \\ .3 & 0 & 1.2 \end{pmatrix} \end{array} \Rightarrow$$



- Hence, estimating $\boldsymbol{\Sigma}/$ (or $\boldsymbol{\Theta}$) are important in practice. Also used in PCA, MANOVA, etc.

Challenges in HD

- Estimating $\mathbf{\Sigma}$ is difficult in high dimensions.
- Natural estimator is the sample covariance matrix

$$\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}', \quad \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$$

- However, the eigenstructure of $\mathbf{S}$ tends to be systematically distorted if $\frac{d}{T} \to \lambda \in (0, \infty)$ ("Marcenko-pastur law").
- Larger eigenvalues are overestimated and small eigenvalues are underestimated.
- Shrinkage estimator is proposed by Stein (1956).

Stein's Esimator

- Spectral decomposition of $\mathbf{S}$

$$\mathbf{S} = \mathbf{Q}\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\mathbf{Q}'$$

  $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ are the eigenvalues of $\mathbf{S}$, $\mathbf{Q}$ are corresponding orthogonal eigenvectors.

- Stein (1956)

$$\hat{\mathbf{\Sigma}} = \mathbf{Q}\mathrm{diag}(\varphi_1, \ldots, \varphi_d)\mathbf{Q}'$$

$$\varphi_j = \frac{\lambda_j}{\alpha_j}, \quad \alpha_j = \frac{T - d + 1 + 2\lambda_j \sum_{i \neq j}(\lambda_j - \lambda_i)^{-1}}{T}$$

- Ledoit and Wolf(2004) also suggested a shrinkage estimator of the form

$$\hat{\mathbf{\Sigma}}^{LW} = \alpha_1 \mathbf{I} + \alpha_2 \mathbf{S}$$

  (In fact, $\varphi_j = \alpha_1 + \alpha_2$)

Sparse Estimation

- Estimating $O(d^2)$ parameters with classical estimators is not viable. Therefore, we need to reduce the number of parameters in $\mathbf{\Sigma}$.
- Sparse estimation is needed.
- Two approaches are possible:
  - Thresholding (for covariance matrix)
  - Regularized estimation (penalization) (for precision matrix)

Thresholding Estimation

- Bickel and Levina (2008)

$$\hat{\mathbf{\Sigma}} = (\hat{\sigma_{ij}}) = \begin{cases} s_{ij} & \text{, if } i = j \\ s_{ij} I(|s_{ij}| > w_T) & \text{, if } i \neq j \end{cases}$$

$$w_T = C\sqrt{\frac{\log d}{T}}, \quad \text{for some } C$$

- Hard thresholding.
- It avoids estimating small elements so that noise does not accumulate.

Thresholding Estimator I

- Cai and Lin(2011) suggested adaptive thresholding

$$\hat{\mathbf{\Sigma}} = (\sigma_{ij})_{d\times d} = \begin{cases} s_{ij} & \text{, if } i = j \\ s_{ij} I\left(\frac{|s_{ij}|}{SE(s_{ij})}\right) & \text{, if } i \neq j \end{cases}$$

where $S.E(s_{ij})$ is the estimated standard error of s_{ij}.

- It considers varying scale of the marginal standard deviation.
- Equivalently, consider

$$\begin{aligned} \hat{\mathbf{\Sigma}}^* &= \mathrm{diag}(\mathbf{S})^{1/2}\mathbf{R}\,\mathrm{diag}(\mathbf{S})^{1/2} \\ \mathbf{R} &= \mathrm{diag}(\mathbf{S})^{-1/2}\mathbf{S}\,\mathrm{diag}(\mathbf{S})^{-1/2} = (r_{ij}) \end{aligned}$$

and hard-thresholding r_{ij}, i.e.

$$r_{ij} = \begin{cases} 1 & \text{if } i = j \\ r_{ij} I(|r_{ij}| > w_T) & \text{if } i \neq j \end{cases}$$

- $\hat{\mathbf{\Sigma}}^*$ is equivalent to entry dependent thresholding, $w_{T,ij} = \sqrt{s_{ii}s_{jj}}w_T$

Thresholding Estimator II

- More generally, generalized thresholding can be applied
- shrinkage function : $h(\cdot, w_T) : \mathbb{R} \longrightarrow \mathbb{R}$
  i $|h(z, w_T)| \leq |z|$
  ii $h(z, w_T) = 0$ if $|z| \leq w_T$
  iii $|h(z, w_T) - z| \leq w_T$
- Examples include
  - Hard thresholding
  - Soft thresholding $\mathbf{h}(\mathbf{z}, w_T) = \mathrm{sign}(\mathbf{z})(|\mathbf{z}| - w_T)_+$
  - SCAD thresholding
  - MC+ hresholding
- Estimator is given by

$$\hat{\boldsymbol{\Sigma}} = \begin{cases} s_{ij} & , i = j \\ h(s_{ij}, w_T) & , i \neq j \end{cases}$$

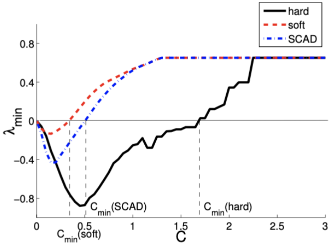
Positive Definiteness

- Thresholding estimator $\hat{\boldsymbol{\Sigma}}$ is asymptotically positive definite.
- But not guaranteed for finite sample.
- Easiest solution : choose the thresholding value to satisfy positeve definiteness such as

$$w_T = C_m \sqrt{\frac{\log d}{T}}$$

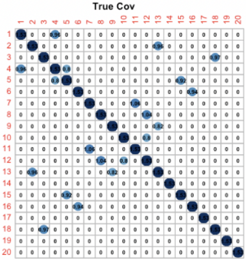
$$C_m = \inf \left\{ C > 0; \lambda_{min}(\hat{\boldsymbol{\Sigma}}) > 0 \right\}$$

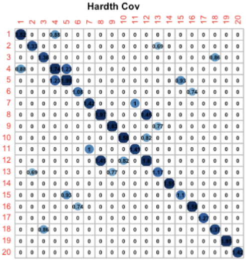
Figure 1: Minimum eigenvalue of $\widehat{\boldsymbol{\Sigma}}(C)$ as a function of C for three choices of thresholding rules. When the minimum eigenvalue reaches its maximum value, the covariance estimator becomes diagonal.



Thresholding Example

- Tunning parameter w_T is usually selected by using CV
- MSE : 11.81 (sample cov) vs. 2.40 (hard thresholding)





Estimating Sparse Precision Matrix

- Difference between marginal and conditional uncorrelatedness. $\mathbf{X} = (X_1, \ldots, X_5)$

$$\mathbf{\Sigma} = \begin{bmatrix} 1.05 & -.23 & .05 & -.02 & 0.05 \\ & 1.45 & -0.25 & 0.10 & -0.25 \\ & & 1.10 & -0.24 & 0.10 \\ & symm & & 1.10 & -0.24 \\ & & & & 1.10 \end{bmatrix}$$

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & .2 & 0 & 0 & 0 \\ & 1 & .2 & 0 & 0 \\ & & 1 & .2 & 0 \\ & & & 1 & .2 \\ & & & & 1 \end{bmatrix}$$

- $\mathbf{\Sigma}$: non-sparse but $\mathbf{\Theta}$ is sparse,
- $\mathbf{\Sigma}$ is dense and every pair of variable are marginally correlated.
- X_1 and X_5 are uncorrelated given the other variables, but they are margianlly correlated,

Graphical Lasso I

- Assume $\mathbf{X} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, then the log-density is given by

$$\log P_{\mathbf{\Sigma}}(\mathbf{x}) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}$$

- Rescaled log-likelihood becomes

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}\log P_{\mathbf{\Sigma}}(\mathbf{x}) &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} \\ &= -\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{\Sigma}|^{-1} - \frac{1}{2}tr(\mathbf{S}\mathbf{\Sigma}^{-1})\end{aligned}$$

where $\mathbf{S} = \frac{1}{T}\sum\limits_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i'$. Since $\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}$ is a 1×1 scalar

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}tr(\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}) = \frac{1}{T}\sum_{i=1}^{T}tr(\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}') = tr(\mathbf{\Sigma}^{-1}\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}\mathbf{x}') \\ = tr(\mathbf{\Sigma}^{-1}\mathbf{S}) = tr(\mathbf{S}\mathbf{\Sigma}^{-1}).\end{aligned}$$

Graphical Lasso II

- Therefore, log-likelihood becomes (up to constant)

$$\log|\mathbf{\Theta}| - tr(\mathbf{S\Theta})$$

-
$$\hat{\mathbf{\Theta}}^{ML} = \arg\min_{\mathbf{\Theta}\succ\mathbf{0}} \{\log|\mathbf{\Theta}| - tr(\mathbf{S\Theta})\}$$

  $\mathbf{\Theta} \succ \mathbf{0}$, means that it is positive definite.
- If $d > N$, MLE may not exist.
- Graphical lasso imposes ℓ_1-norm on the off-diagonal entries

$$\hat{\mathbf{\Theta}}^{GL} = \arg\min_{\mathbf{\Theta}>\mathbf{0}} \left\{ \log\det\mathbf{\Theta} - tr(\mathbf{S\Theta}) - \lambda \sum_{s\neq t} |\theta_{st}| \right\}$$

Graphical Lasso III

- Subgradient equation

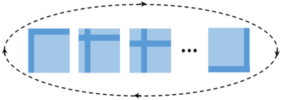
$$\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda \mathbf{\Psi} = 0$$

where

$$\mathbf{\Psi} = (\psi_{jk}) = \begin{cases} \operatorname{sign}(\theta_{jk}), & \text{if } \theta_{jk} \neq 0 \\ \text{any value in } [-1, 1], & \text{if } \theta_{jk} = 0 \end{cases}$$

Blockwise coordinate descent

Idea: repeatedly cycle through all columns/rows and, in each step, optimize only a single column/row



Notation: use $\boldsymbol{W}$ to denote working version of $\boldsymbol{\Theta}^{-1}$. Partition all matrices into 1 column/row vs. the rest

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{12}^{\top} & \theta_{22} \end{bmatrix} \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^{\top} & s_{22} \end{bmatrix} \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^{\top} & w_{22} \end{bmatrix}$$

Graphical Lasso IV

- Denote W be the working version of $\mathbf{\Theta}^{-1}$, then w_{12} satisfy

$$W_{11}\beta - s_{12} + \lambda \text{sign}(\beta) = 0, \quad \beta = -\frac{\theta_{12}}{\theta_{22}} \tag{1}$$

$$\because \begin{bmatrix} W_{11} & w_{12} \\ w_{12}' & w_{22} \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}_{11} & \theta_{12} \\ \theta_{12}' & \theta_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 1 \end{bmatrix}$$
$$\Rightarrow w_{12} = -W_{11}\frac{\theta_{12}}{\theta_{22}} = W_{11}\beta$$

- This can be viewed as a modification of lasso. In the regression form, lasso is

$$\frac{1}{2N}\|\mathbf{y} - \mathbf{z}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

  The subgradient equations are

$$\frac{1}{N}\mathbf{z}'\mathbf{z}\boldsymbol{\beta} - \frac{1}{N}\mathbf{z}'\mathbf{y} + \lambda \text{sign}(\boldsymbol{\beta}) = 0 \tag{2}$$

Graphical Lasso V

- Hence, replacing

$$\begin{cases} W_{11} & \longrightarrow \frac{1}{N}\mathbf{z}'\mathbf{z} \\ s_{12} & \longrightarrow \frac{1}{N}\mathbf{z}'\mathbf{y} \end{cases}$$

  gives a solution.

Algorithm 9.1 GRAPHICAL LASSO.

---

1. Initialize $\mathbf{W} = \mathbf{S}$. Note that the diagonal of $\mathbf{W}$ is unchanged in what follows.
2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the j^{th} row and column, and part 2: the j^{th} row and column.
   (b) Solve the estimating equations $\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \lambda \cdot \mathrm{sign}(\boldsymbol{\beta}) = 0$ using a cyclical coordinate-descent algorithm for the modified lasso.
   (c) Update $\mathbf{w}_{12} = \mathbf{W}_{11}\hat{\boldsymbol{\beta}}$
3. In the final cycle (for each j) solve for $\hat{\boldsymbol{\theta}}_{12} = -\hat{\boldsymbol{\beta}} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - \mathbf{w}_{12}^T\hat{\boldsymbol{\beta}}$.

Graphical Lasso VI

- Tuning parameters can be selelcted by CV or BIC. Theory suggests $\lambda_T = 2\frac{\log d}{T}$.
- Debiasing can be done by solving exact soltion. That is, apply glasso to find sparsity pattern and re-estimate parameters with constraints.
- This is the same as put $\lambda = 0$ with constraints in β

  $$W_{11}^* \beta^* - s_{12}^* = 0 \iff \beta^* = (W_{11}^*)^{-1} s_{12}^*$$

  Therefore,

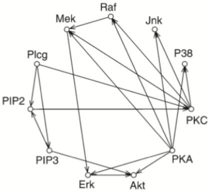
  $$\hat{\beta}_j = \begin{cases} \beta_j^* & \text{if } j\text{-th variable is non-zero} \\ 0 & \text{constrained to be zero} \end{cases}$$

- In `R`, use `glasso` package

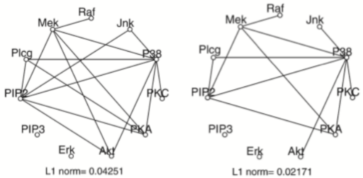
Glasso: Example

Example from Friedman et al. (2007), cell-signaling network:

Believed network

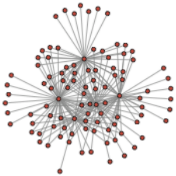


Graphical lasso estimates

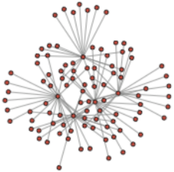


Example from Liu et al. (2010), hub graph simulation:

True graph



Graphical lasso estimate



Application: Portfolio Optimization

- Mean-variance portfolio (MVP) theory uses covariance matrix to hedge risk.
- Minimum variance portfolio (given $\mathbf{\Sigma}$) is defined as to minimize $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ subject to $\sum w_i = 1$.

  (e.g) Your portfolio has Samsung, Apple, Google.
  We want to allocate your total budget into

  $100w_1$ % of Samsung
  $100w_2$ % of Apple
  $100w_3$ % of Google

  to minimize "variance" to hedge risk.
- Analytic solution exists $\mathbf{w}^* = (\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{\Sigma}^{-1}\mathbf{1}$

Application: MVP

- Due to non-stationarity, use rebalancing strategy on every 4 weeks.
- Use past N_{est} days to estimate $\mathbf{\Sigma}^{-1}$

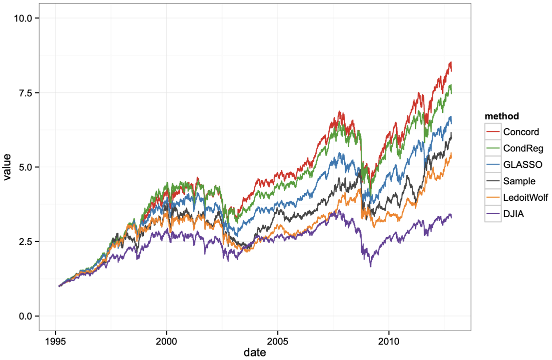


Figure: $N_{\text{est}} = 75$ days, rebalance every 4 weeks

Results from Khare, Oh and Rajarathan (2014).

Extension to Time Series Data

- In HDTS context, we are interested in the estimation of spectral density matrix.
- The spectral density $f_X(\omega)$ is estimated by periodogram

$$\hat{f}_X(\omega_k) = \frac{1}{2m+1} \sum_{|j| \leq m} I_X(\omega_{k+j})$$

where $I_X(\omega) = \sum_{|\ell|<n} \hat{\Gamma}(\ell) e^{-i\omega\ell}$

- Thresholding estimators are defined as

$$S_\tau\left(\hat{f}_X(\omega_k)\right)$$

for threshold τ.

- Key reference is Sun et al. (2018).

Spectral Density Matrices

- Note that spectral density / periodogram are definded in the complex field.
- Some forms of thresholding functions
  - Hard thresholding; $S_\tau(z) = \begin{cases} z & \text{if } |z| > \tau \\ 0 & \text{o.w.} \end{cases}$
  - Soft thresholding; $S_\tau(z) = \frac{z}{|z|}(|z| - \lambda)_+,\ z \in \mathbb{C}$
- Special case when $\omega = 0$:

$$2\pi f_X(0) = \sum_{h=-\infty}^{\infty} \Gamma_X(h) = \sum_{h=-\infty}^{\infty} E(\mathbf{X}_{t+h} - \boldsymbol{\mu})(\mathbf{X}_t - \boldsymbol{\mu})'$$

is the long-run variance

Sparse Long-run Variance

- fMRI series to study brain connectivity (Data dimension $= 86 \times 210$)
- Tuning parameters are selected from CV in the frequency domain

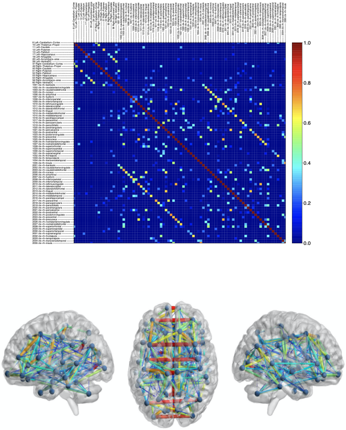


Figure: Brain connectivity

References

- P. J. Bickel and E. Levina. Covariance regularization by thresholding. The Annals of Statistics, pages 2577–2604, 2008.
- T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association, 106(494):672–684, 2011.
- Fan, Jianqing, Yuan Liao, and Han Liu. "An overview of the estimation of large covariance and precision matrices." The Econometrics Journal 19.1 (2016): C1-C32.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96 1348–1360.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." Biostatistics 9, no. 3 (2008): 432-441.
- O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2):365–411, 2004.
- Sun, Y., Li, Y., Kuceyeski, A. and Basu, S. (2018). Large Spectral Density Matrix Estimation by Thresholding. arXiv preprint arXiv:1812.00532.
- MEINSHAUSEN, N. AND BUHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. Annals of Statistics 34, 1436–1462.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 894–942.