High-Dimensional Data Analysis and Statistical Inference	**May 20, 2015**

Homework 2: Large covariance estimation

Student: Jun Yu *ID : 1401110054*

1. *Concentration of sample covariances:* Prove inequality (A.4) in Rothman, Levina and Zhu (2009): for uniformly sub-Gaussian X_{1j} and sufficiently small $t > 0$,

$$\mathrm{P}(\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| > t) \leq C_1 p^2 e^{-nC_2 t^2} + C_3 p e^{-nC_4 t}$$

Solution:

Without loss of generality, we assumed $\boldsymbol{EX} = \mathbf{0}$. Begin with the decomposition,

$$\hat{\Sigma} = \hat{\Sigma}^0 - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T,$$

where

$$\hat{\Sigma}^0 \equiv [\hat{\sigma}^0_{ij}] = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X_k}\boldsymbol{X_k}^T.$$

Then, we have

$$\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \leq \max_{i,j} |\hat{\sigma}^0_{ij} - \sigma_{ij}| + \max_{i,j} |\bar{\boldsymbol{X}}_{\boldsymbol{i}}\bar{\boldsymbol{X}}_{\boldsymbol{j}}^T|.$$

First, I will prove the following lemma in [1]:

Lemma (BICKEL, P. J. and LEVINA, E. (2008) Lemma A.3.)**.** *Let Z_i be i.i.d. r.v with mean $\mathbf{0}$ and variance Σ_p,and $\lambda_{max}(\Sigma_p) \leq \varepsilon_0^{-1} < \infty$. Then, if $\Sigma_p = [\sigma_{ab}]$,*

$$\mathrm{P}(|\sum_{i=1}^{n}(Z_{ij}Z_{ik} - \sigma_{jk})| \geq n\nu) \leq C_1 \exp(-C_2 n\nu^2) \textit{ for } |\nu| \leq \delta$$

where C_1, C_2 and δ depend on ε_0 only.

Proof. Write

$$\mathrm{P}(|\sum_{i=1}^{n}(Z_{ij}Z_{ik} - \sigma_{jk})| \geq n\nu) = \mathrm{P}(|\sum_{i=1}^{n}(Z^*_{ij}Z^*_{ik} - \rho_{jk})| \geq \frac{n\nu}{\sqrt{\sigma_{jj}\sigma_{kk}}}),$$

where $\rho_{jk} = \frac{\sigma_{ij}}{\sqrt{\sigma_{jj}\sigma_{kk}}}$ and (Z^*_{ij}, Z^*_{ik}) are r.v with mean 0, variance 1, covariance ρ_{jk}. Now,

$$\sum_{i=1}^{n}(Z^*_{ij}Z^*_{ik} - \rho_{jk}) = \frac{1}{4}\{\sum_{i=1}^{n}[(Z^*_{ij} + Z^*_{ik})^2 - 2(1+\rho_{jk})] + \sum_{i=1}^{n}[(Z^*_{ij} - Z^*_{ik})^2 - 2(1-\rho_{jk})]\}$$

and reduce the problem to estimating

$$2\mathrm{P}(|\sum_{i=1}^{n}(V_i^2-1)| \geq \frac{n\nu}{2(1-\rho_{jk})(\sigma_{jj}\sigma_{kk})^{1/2}}),$$

where V_i are i.i.d r.v with mean 0,variance 1.Since V_i is subguassian, it's easy to see that V_i satisfies condition (P) and (3.12) on page 45 of [2]. The lemma follows from Theorem 3.2, page 45 and (2.13) on page 19. □

By the union sum inequality, we have

$$\mathrm{P}(\max_{i,j}|\hat{\sigma}^0_{ij}-\sigma_{ij}|) \leq C_1 p^2 e^{-nC_2t^2}.$$

For the second term, we have

$$\begin{aligned}\mathrm{P}(\max_i|\bar{\boldsymbol{X}}_i|^2 \geq t) &= \mathrm{P}(\bigcup_i\{|\bar{\boldsymbol{X}}_i|^2 \geq t\}) \\ &\leq \sum_i^p \mathrm{P}(|\bar{\boldsymbol{X}}_i|^2 \geq t\}) \\ &= \sum_i^p \mathrm{P}(\exp(C_4|\bar{\boldsymbol{X}}_i|^2) \geq \exp(C_4 t)\}) \\ &\leq C_3 p e^{-nC_4t}\end{aligned}$$

the last inequality due to the fact:$Ee^{tX_{ij}^2} \leq C_3 \leq \infty$ for all i,j.

2. *Sign consistency of covariance thresholding:* Let $\Psi = \{(i,j) : \sigma_{ij} \neq 0\}$ be the support of Σ. Correct and prove the second part of Theorem 2 in Rothman, Levina and Zhu (2009): if in addition $|\sigma_{ij}| \geq \tau$ for all $(i,j) \in \Psi$ and $\sqrt{n}(\tau-\lambda) \to \infty$, then with probability tending to 1,

$$\mathrm{sign}(s_\lambda(\hat{\sigma}_{ij})) = \mathrm{sign}(\sigma_{ij}) \text{ for all } (i,j) \in \Psi.$$

Solution:
Note the facts that

$$\{(i,j) : s_\lambda(\hat{\sigma}_{ij}) \leq 0, \sigma_{ij} > 0 \text{ or } s_\lambda(\hat{\sigma}_{ij}) \geq 0, \sigma_{ij} < 0\} \subseteq \{(i,j) : |\hat{\sigma}_{ij}-\sigma_{ij}| > \tau-\lambda\}.$$

Hence,

$$\mathrm{P}(\mathrm{sign}(s_\lambda(\hat{\sigma}_{ij})) \neq \mathrm{sign}(\sigma_{ij}) \text{ for some } (i,j) \text{ w.r.t } \sigma_{ij} \neq 0) \leq \mathrm{P}(\max_{i,j}|\hat{\sigma}_{ij}-\sigma_{ij}| > \tau-\lambda).$$

Using the result we get from 1,when $\tau-\lambda = o(1)$ we have

$$\mathrm{P}(\max_{i,j}|\hat{\sigma}_{ij}-\sigma_{ij}| > \tau-\lambda) \leq C_1p^2e^{-nC_2(\tau-\lambda)^2} + C_3pe^{-nC_4(\tau-\lambda)}$$

and right hand side tends to 0,since $\sqrt{n}(\tau-\lambda) \to \infty$

3. *Implementation of graphical Lasso:* Implement the graphical Lasso algorithm in Friedman, Hastie and Tibshirani (2008) using R, MATLAB, or any other programming language that you prefer. You may use either coordinate descent or ADMM to solve the inner Lasso problem.

   (a) Print out all source code.

   (b) Design a small simulation study to check if your program works as expected.

   (c) Conduct a timing experiment on a 1000-node problem. Describe your computing environment (CPU and memory) and report the CPU time in seconds.

   Solution:

   (a) All source code

```r
softT <- function(x, T){
  sign(x)*ifelse(abs(x)-T>0,abs(x)-T,0)
}

#The input X,y stands for X^T*X,X^T*y
lasso <- function(X, y, rho){
  b <- solve(X)%*%y
  bpre <- rep(0,length(b))
  repeat{
    for(i in 1:length(b)){
      mu <- (y[i]-X[i,-i]%*%b[-i])/X[i,i]
      b[i] <- softT(mu,rho/X[i,i])
    }
    if(sum(abs((b-bpre)))<1e-6){
      return(b)
      break
    }
    bpre <- b
  }
}

grlasso <- function(S,rho){
  W <- S + rho*diag(1,ncol(S))
  y <- W
  repeat{
```

```
      for(i in 1:ncol(S)){
        b <- lasso(W[-i,-i],S[-i,i],rho)
        W[i,-i] <- W[-i,i] <- W[-i,-i]%*%b
      }
      if(sum((y-W)^2)/sum(y^2)<1e-4){
        return(W)
        break
      }
      y <- W
    }
  }
```

(b) A small simulation study

```
set.seed(100)
x<-matrix(rnorm(50*20),ncol=20)
s<- var(x)
grlasso(s, rho=.01)
```

(c) The R code

```
sigma <- diag(1,1000)
for(i in 1:100){
  sigma[i,i+1] <- 0.5
  sigma[i+1,i] <- 0.5
}
s<-solve(sigma)
system.time(b<-grlasso(s, rho=.06))
```

The outputshows that R used 2660.09 seconds of CPU time; the operating system used 13.10 seconds of CPU time; and 2784.82 seconds elapsed during the test. Timing carried out on a Intel Core i5 2.27GHz processor with memory size 3.86GB.

4. *Unitarily invariant norms:* Let a matrix $\boldsymbol{A}$ be partitioned as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}.$$

and let $\|\cdot\|$ be a unitarily invariant norm. Show that $\|\boldsymbol{A}_{ij}\| \le \|\boldsymbol{A}\|$ for all $i, j = 1, 2$.
Solution:

Let Φ be the symmetric gauge function that generates $\|\cdot\|$ and let

$$\tilde{A}_{11} = \begin{pmatrix} \boldsymbol{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

with the same dimension as A. Then since Φ is absolute,

$$\|\tilde{A}_{11}\| = \Phi(\sigma_1, \cdots, \sigma_n) \le \Phi(\tau_1, \cdots, \tau_n) = \|\boldsymbol{A}\|$$

where $\sigma_1, \cdots, \sigma_n, \tau_1, \cdots, \tau_n$ are the eigenvalues of $\tilde{A}_{11}$ and $\boldsymbol{A}$ respectably. Using the fact $\Phi(PX) = \Phi(X)$ for any permutation matrix P, we know the result holds for all $i, j = 1, 2$.

5. *RSC condition:* For an approximately low-rank matrix Θ^* with $\sum_{j=r+1}^{m} \sigma_j(\Theta^*) > 0$, prove the following claims in Negahban and Wainwright (2011):

   (a) The set of all Δ satisfying the constraint

   $$\|\Delta_{\mathcal{B}^r}\|_* \le 3\|\Delta_{\mathcal{A}^r}\|_* + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*)$$

   includes an open ball around the origin.

   (b) If the operator $\mathfrak{X}$ fails to satisfy the RSC condition

   $$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \ge \kappa(\mathfrak{X})\|\Delta\|_F^2$$

   for all Δ, then it will also fail to satisfy it under the constraint in part (a).

Solution:

(a) Using Lemma 1 in together with triangle inequality, we have

$$\|\Delta\|_* \le 4\|\Delta_{\mathcal{A}^r}\|_* + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*).$$

The result is obvious.

(b) Note that we have the decomposition $\Theta^* = \Pi_{\mathcal{A}^r}(\Theta^*) + \Pi_{\mathcal{B}^r}(\Theta^*)$ where $\Pi_{\mathcal{A}^r}$ denote the projection operator onto the subspace $\mathcal{A}^r$.Using this decomposition together with the triangle inequality , we have

$$\begin{aligned}
\|\hat{\Theta}\|_* &= \|\Pi_{\mathcal{A}^r}(\Theta^*) + \Delta_{\mathcal{B}^r} + \Pi_{\mathcal{B}^r}(\Theta^*) + \Delta_{\mathcal{A}^r}\|_* \\
&\ge \|\Pi_{\mathcal{A}^r}(\Theta^*) + \Delta_{\mathcal{B}^r}\|_* - \|\Pi_{\mathcal{B}^r}(\Theta^*) + \Delta_{\mathcal{A}^r}\|_* \\
&\ge \|\Pi_{\mathcal{A}^r}(\Theta^*)\|_* + \|\Delta_{\mathcal{B}^r}\|_* - (\|\Pi_{\mathcal{B}^r}(\Theta^*)\|_* + \|\Delta_{\mathcal{A}^r}\|_*)
\end{aligned}$$

the last inequality due to the construction of $\Delta_{\mathcal{B}^r}$.Consequently, we have

$$\|\Theta^*\|_* - \|\hat{\Theta}\|_* \le 2\|\Pi_{\mathcal{B}^r}(\Theta^*)\|_* + \|\Delta_{\mathcal{A}^r}\|_* - \|\Delta_{\mathcal{B}^r}\|_*$$

Recall that the error $\Delta = \hat{\Theta} - \Theta^*$ associated with any optimal solution must satisfy the inequality (30) in [3] , which implies that

$$0 \leq \|\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})\|_{op}\|\Delta\|_* + \lambda_N(\|\Theta^*\|_* - \|\hat{\Theta}\|_*),$$

Substituting this inequality into above inequality, we obtain

$$0 \leq \|\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})\|_{op}\|\Delta\|_* + \lambda_N(2\|\Pi_{\mathcal{B}^r}(\Theta^*)\|_* + \|\Delta_{\mathcal{A}^r}\|_* - \|\Delta_{\mathcal{B}^r}\|_*)$$

Since $\|\Pi_{\mathcal{B}^r}(\Theta^*)\|_* = \sum_{j=r+1}^m \sigma_j(\Theta^*)$, we obtain

$$\|\Delta_{\mathcal{B}^r}\|_* \leq \|\Delta_{\mathcal{A}^r}\|_* + 2\sum_{j=r+1}^m \sigma_j(\Theta^*) + \frac{1}{\lambda_N}\|\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})\|_{op}\|\Delta\|_*$$

Since each inequality is not strick, there is Δ makes the equation come into existence. When $\|\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})\|_{op} > \lambda_N/2$, the operator $\mathfrak{X}$ fails to satisfy the RSC condition for all Δ, then it will also fail to satisfy it under the constraint in part (a).

References

[1] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

[2] L. Saulis and V. Statulevicius, *Limit theorems for large deviations*, vol. 73. Springer Science & Business Media, 1991.

[3] S. Negahban, M. J. Wainwright, *et al.*, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.