Journal of the Korean Data & Information Science Society
2014, **25**(4), 881–891

http://dx.doi.org/10.7465/jkdi.2014.25.4.881
한국데이터정보과학회지

그래프 LASSO에서 모형선택기준의 비교[†]

안형석[1] · 박창이[2]

[12]서울시립대학교 통계학과

접수 2014년 6월 26일, 수정 2014년 7월 14일, 게재확정 2014년 7월 18일

요 약

그래프 모형 (graphical model)은 확률변수들간의 조건부 독립성 (conditional independence)을 시각적인 네트워크형태로 표현할 수 있기 때문에, 정보학 (bioinformatics)이나 사회관계망 (social network) 등 수많은 변수들이 서로 연결되어 있는 복잡한 확률 시스템에 대한 직관적인 도구로 활용될 수 있다. 그래프 LASSO (graphical least absolute shrinkage and selection operator)는 고차원의 자료에 대한 가우스 그래프 모형 (Gaussian graphical model)의 추정에서 과대적합 (overfitting)을 방지하는데에 효과적인 것으로 알려진 방법이다. 본 논문에서는 그래프 LASSO 추정에서 매우 중요한 문제인 모형선택에 대하여 고려한다. 특히 여러가지 모형선택기준을 모의실험을 통해 비교하며 실제 금융 자료를 분석한다.

주요용어: 가우스 그래프 모형, 고차원 자료, 조건부 독립성.

1. 서론

그래프 모형 (graphical model)은 방향 그래프 (directed graph) 혹은 무방향 그래프 (undirected graph)로 정의되는 확률모형으로서, 확률변수들간의 조건부 독립성을 시각적인 그래프로 표현할 수 있다. 따라서 그래프 모형은 생물정보학 (bioinformatics)이나 사회관계망 (social network) 등 수많은 변수들이 매우 복잡하게 연결되어 있는 복잡한 현상에 대한 확률모형을 효과적이며 직관적으로 표현할 수 있는 도구로 주목을 받고 있다 (Jordan, 2004).

특히 연속형 확률벡터에 대하여 결합분포가 다변량 정규분포를 따른다고 가정하는 경우의 그래프 모형을 가우스 그래프 모형이라고 한다. 가우스 그래프 모형에서는 로그-가능도 함수를 이용하여 공분산 행렬의 역행렬 또는 정확도 행렬 (precision matrix)에 대하여 최대 가능도 추정량 (maximum likelihood estimator)을 구할 수 있다. 그러나 $p \gg n$인 고차원 자료에서는 최대 가능도 추정은 불가능하다. 또한 최대 가능도 추정이 가능한 경우라도 n이 p에 비해 충분히 크지 않으면 추정 효율이 떨어질 수 있으므로 축소 추정 (shrinkage estimation)이 바람직하다 (Yuan과 Lin, 2007).

고차원 자료에 대한 가우스 그래프 모형 추정시 과모수화 (over-parametrization)에 따른 과대적합을 방지하기 위하여 벌점화 방법 (method of penalization)에 기반한 축소 추정법을 고려할 수 있다. 본 논문에서는 정확도 행렬에 대하여 Tibshirani (1996)의 LASSO 벌점화를 적용한 Yuan과 Lin (2007)의

[†] 이 논문은 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2012R1A1A2004901).
[1] (130-743) 서울시 동대문구 전농동 90번지, 서울시립대학교 통계학과, 석사과정.
[2] 교신저자: (130-743) 서울시 동대문구 전농동 90번지, 서울시립대학교 통계학과, 부교수.
E-mail: park463@uos.ac.kr

그래프 LASSO를 고려하기로 한다. ALASSO (adaptive LASSO)와 SCAD (smoothly clipped absolute deviation) 벌점등을 이용한 네트워크 추정은 Fan 등 (2009)를 참고하기 바란다.

사실 그래프 LASSO에서 조율모수 (tuning parameter)의 값에 따라 추정된 네트워크의 복잡도가 달라지므로 적절한 모형선택은 매우 중요한 문제이다. Yuan과 Lin (2007), Friedman 등 (2008), Guo 등 (2011) 등의 기존의 문헌에서는 AIC (Akaike information criterion), BIC (Bayesian information criterion), CV (cross-validation)와 같은 일반적인 모형선택기준을 고려하였다. 최근 Chen과 Chen (2008)에서는 $p \gg n$의 경우에 BIC는 모형선택관점에서 부적절함을 지적하며 이를 개선한 EBIC (extended BIC)를 제안하였다. 본 논문에서는 여러가지 모형선택기준을 모의실험을 통해서 비교하고 금융관련 통계지표 자료에 대하여 적용하고자 한다. 참고로 Cho와 Park (2012)에서는 사회지표조사에서의 3단계 복합 데이터마이닝의 적용 방안을 연구하였고, Choi 등 (2012)에서는 국내외 경제지표를 예측변수로 사용한 산업별 주가지수 예측 문제를 연구하였다.

본 논문의 구성은 다음과 같다. 2절에서는 그래프 LASSO 추정에 대하여 소개한다. 우선 논의에 위해 필요한 그래프 모형에 대한 기본적인 개념과 용어를 소개하며, 가우스 그래프 모형, 그리고 가우스 그래프 모형에 LASSO 벌점화를 적용한 그래프 LASSO와 모형선택기준에 대하여 설명한다. 3절에서는 앞에서 논의한 주요 모형선택기준을 모의실험을 통해 비교하며 실제 금융 자료에 적용하여 그 결과를 해석한다. 마지막으로 4절에서는 본 논문을 요약하고 향후 연구방향에 대하여 논의한다.

2. 그래프 LASSO 추정법

이 절에서는 논문의 전개에 반드시 필요한 그래프 이론의 몇 가지 개념 및 용어만을 간단히 소개하며, 가우스 그래프 모형에 대한 그래프 LASSO 추정법 및 모형선택기준에 대하여 소개하기로 한다. 그 밖의 그래프 이론과 가우스 그래프 모형에 대한 보다 자세한 사항은 Lauritzen (1996)을 참고하기 바란다.

2.1. 그래프 이론의 기본 개념 및 용어

V는 노드 (node)들의 유한집합이고 $E \subset V \times V$으로 노드간의 간선 (edge)들의 집합이라 하자. 그러면 그래프는 $\mathcal{G} = (V, E)$로 정의할 수 있다. 주어진 두 노드 $\alpha, \beta \in V$에 대하여 $(\alpha, \beta) \in E$이고 $(\beta, \alpha) \in E$이면 α와 β사이의 간선은 무방향 간선 (undirected edge)이라고 하며 $\alpha \sim \beta$로 표기한다. 만약 $(\alpha, \beta) \in E$이나 $(\beta, \alpha) \notin E$이면 두 노드 사이의 간선은 방향간선 (directed edge)이라 하고 $\alpha \to \beta$로 나타낸다. 단 동일한 한 노드로 이루어진 순서쌍인 루프 (loop)나 동일한 순서쌍에 대한 다중간선은 존재하지 않는다고 가정한다. 또한 본 논문에서는 무방향 간선으로 이루어진 무방향 그래프만을 고려하기로 한다.

$V_0 \subset V$이고 $E_0 \subset E$이면 $\mathcal{G}_0 = (V_0, E_0)$는 $\mathcal{G}$의 부분 그래프 (subgraph)라 부른다. $A \subset V$일 때 E_A를 A에서 노드를 갖는 E상의 간선들의 집합이라 정의하자. 그러면 $\mathcal{G}_A = (A, E_A)$는 A에 의해 유도되는 $\mathcal{G}$의 부분 그래프라고 한다. 모든 서로 다른 두 노드 $\alpha, \beta \in V$에 대하여 $(\alpha, \beta) \in E$이면 $\mathcal{G}$ 는 완전 그래프 (complete graph)라고 부르며, 완전 부분 그래프들중 최대의 부분집합을 클릭 (clique)이라고 부른다. V상의 서로 다른 노드들의 집합 $A = \{v_1, \ldots, v_k\}$이 주어졌을 때, 모든 $i = 1, \ldots, k-1$에 대하여 $(v_i, v_{i+1}) \in E$을 만족하는 경우에 $\mathcal{G}_A$를 경로 (path)라고 한다.

2.2. 가우스 그래프 모형

확률벡터 $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$는 평균벡터가 $\boldsymbol{\mu}$이고 공분산 행렬이 $\boldsymbol{\Sigma}$인 p차원 다변량 정규분포를 따른다고 가정하자. 정확도 행렬을 $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$로 정의하자. 그러면 정규분포의 성질에 의해 $\boldsymbol{\Theta}_{ij} = 0$은 다른 변수들이 주어졌을 때 두 확률변수 Y_i와 Y_j가 서로 조건부 독립일 필요충분 조건이다. 이때 노드집합 E를 확률변수 $Y_1, \ldots, Y_p$에 대응되는 첨자집합 $\{1, \ldots, p\}$으로 정의하고 간선집합을 $V = \{(i, j) : \boldsymbol{\Theta}_{i,j} \neq 0\}$로 정의하여 얻게 되는 그래프 $\mathcal{G} = (V, E)$를 가우스 그래프 모형이라고 한다.

자료 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$이 주어졌을 때 로그-가능도함수는

$$\frac{n}{2}\log\det(\boldsymbol{\Theta}) - \frac{n}{2}\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{S}) - \frac{n}{2}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})^T\boldsymbol{\Theta}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})$$

로 주어진다. 여기서 tr은 대각합 (trace)이고 $\boldsymbol{S}$는 표본공분산행렬을 나타낸다. $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$이므로 $\boldsymbol{\Theta}$에 대한 프로파일 (profile) 로그-가능도 함수는

$$l(\boldsymbol{\Theta}) = \log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{S})$$

로 주어진다. 최대 가능도 추정량 $\hat{\boldsymbol{\Theta}}$은 Dempster (1972)의 IPS (iterative proportional scaling) 등의 알고리즘을 이용하여 찾을 수 있다.

2.3. 그래프 LASSO 추정

가우스 그래프 모형에서 차원 p가 그리 크지 않은 경우에는 IPS 알고리즘을 이용하여 최대 가능도 추정치를 구할 수 있다. 그러나 $p \gg n$의 경우와 같은 고차원 자료에 대하여 최대 가능도 추정은 불가능하다. 사실 n이 p에 비하여 충분히 크지 않으면 최대 가능도 추정치를 구할 수 있더라도 추정 효율이 떨어지므로 축소 추정이 필요하다. 또한 모형의 해석력을 위해서도 추정된 그래프가 간결 (sparse)한 것이 바람직하다. Yuan과 Lin (2007)의 그래프 LASSO는 LASSO 벌점화를 $\boldsymbol{\Theta}$에 적용한 것으로 벌점화된 목적함수

$$\log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{S}) - \lambda\|\boldsymbol{\Theta}\|_1 \tag{2.1}$$

을 최소화한다. 여기서 $\|\boldsymbol{\Theta}\|_1$은 행렬 $\boldsymbol{\Theta}$의 원소들의 절대값의 합이며 $\lambda \geq 0$는 조율모수이다.

목적함수 (2.1)을 최소화하기 위하여 Banerjee 등 (2007)에서는 블럭별 좌표 강하 (blockwise coordinate descent) 방법을 이용하였고, Friedman 등 (2008)에서는 블럭별 좌표 강하를 기초로 `glasso` (graphical LASSO) 알고리즘을 제안하였다.

(2.1)의 해에 대한 서브-그래디언트 (sub-gradient)에 의한 식은

$$\boldsymbol{\Theta}^{-1} - \boldsymbol{S} - \lambda\boldsymbol{\Gamma} = \boldsymbol{0} \tag{2.2}$$

이며 $\boldsymbol{\Gamma}$는 $\boldsymbol{\Theta}$의 원소별 (elementwise) 부호로 정의되는 행렬이다. $\boldsymbol{\Theta}$의 대각원은 양수이어야 하므로 $i = 1, \ldots, p$에 대하여 $w_{ii} = s_{ii} + \lambda$이 성립한다. 여기서 $\boldsymbol{W} = \boldsymbol{\Theta}^{-1}$이다. `glasso` 알고리즘의 아이디어는 다음과 같다. $\boldsymbol{\Sigma}$의 추정량 $\boldsymbol{W}$, 표본공분산행렬 $\boldsymbol{S}$, $\boldsymbol{\Gamma}$의 서로 대응되는 분할

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^T & w_{22} \end{pmatrix}, \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^T & s_{22} \end{pmatrix}, \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \gamma_{12} \\ \boldsymbol{\gamma}_{12}^T & \gamma_{22} \end{pmatrix} \tag{2.3}$$

을 생각해보자. 이 때 $\boldsymbol{w}_{12}$는

$$\boldsymbol{w}_{12} = \arg\min_{\boldsymbol{w}}\{\boldsymbol{w}^T\boldsymbol{W}_{11}^{-1}\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{s}_{12}\|_\infty \leq \lambda\} \tag{2.4}$$

와 같이 구할 수 있다. 여기서 $\|\cdot\|_\infty$는 l_∞ 노음을 나타낸다. 식 (2.4)에 대한 볼록인 쌍대문제는

$$\min_{\boldsymbol{\beta}}\{1/2\|\boldsymbol{W}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\} \tag{2.5}$$

로 표현된다. 여기서 $\boldsymbol{b} = \boldsymbol{W}_{11}^{-1/2}\boldsymbol{s}_{12}$이고 $\|\cdot\|_1$은 l_1노음이다. $\boldsymbol{\beta}$가 (2.5)의 해이면 $\boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda\boldsymbol{\gamma}_{12} = \boldsymbol{0}$을 만족해야 한다. 따라서 (2.2)의 p열에 대응되는 식으로부터 관계식 $\boldsymbol{w}_{12} = \boldsymbol{W}_{11}\boldsymbol{\beta}$ 를 얻는다.

`glasso` 알고리즘은 다음과 같이 요약된다.

1. $\boldsymbol{W} = \boldsymbol{S} + \lambda\boldsymbol{I}$로 초기화한다.

2. 다음 과정을 수렴할 때까지 반복한다.
   (a) (2.3)과 같이 행렬 $\boldsymbol{W}$와 $\boldsymbol{S}$를 분할한다.
   (b) $\boldsymbol{W}$의 모든 행 (또는 열)에 대하여 각각 순차적으로 (2.5)를 풀어 $\hat{\boldsymbol{\beta}}$를 구한다.
   (c) $\boldsymbol{w}_{12} = \boldsymbol{W}_{11}\hat{\boldsymbol{\beta}}$으로 $\boldsymbol{W}$의 해당 행과 열을 갱신한다.
3. $\boldsymbol{W}^{-1}$를 계산한다.

2.4. 그래프 LASSO의 모형선택기준

λ는 추정된 네트워크의 복잡도 혹은 간결성을 결정하는 모수로서 그 값에 따라 서로 다른 복잡도의 네트워크를 얻게 된다. 일반적으로 사용되는 모형선택기준들은 다음과 같다.

- AIC (Akaike information criterion):

$$AIC = -\log(\det(\hat{\boldsymbol{\Theta}})) + \mathrm{tr}(\hat{\boldsymbol{\Theta}}\boldsymbol{S}) + df\frac{2}{n}.$$

  여기서 df는 추정된 그래프에서 간선의 갯수로서 무방향 그래프를 고려하므로 $\hat{\boldsymbol{\Theta}}$의 하삼각행렬에서 영이 아닌 원소의 갯수를 나타낸다.

- BIC (Bayesian information criterion):

$$BIC = -\log(\det(\hat{\boldsymbol{\Theta}})) + \mathrm{tr}(\hat{\boldsymbol{\Theta}}\boldsymbol{S}) + df\frac{log(n)}{n}.$$

- K-묶음 교차확인오차 (K-fold cross-validation):

$$CV = \sum_{k=1}^{K}\left[-\log(\det(\boldsymbol{W}^{-1(-k)})) + \mathrm{tr}(\boldsymbol{S}^{(k)}\boldsymbol{W}^{-1(-k)})\right].$$

  여기서 $\boldsymbol{W}^{-1(-k)}$는 k번째 묶음을 제외한 나머지 자료를 이용한 $\boldsymbol{\Theta}^{(-k)}$의 추정치이고 $\boldsymbol{S}^{(k)}$는 k번째 묶음을 이용한 표본공분산행렬을 나타낸다.

BIC는 AIC나 CV에 비해서 상대적으로 단순한 모형을 선택하는 특징이 있다. 또한 CV의 최소값의 표준편차 범위 내에서 가장 간결한 모형을 선택하는 Hastie 등 (2009)의 1-표준편차 규칙 (1-standard deviation rule)을 적용한 CV-SD도 고려할 수 있다. 최근 Chen과 Chen (2008)에서는 $p \gg n$의 경우에 BIC는 모형선택관점에서 부적절함을 지적하며 이를 개선한 다음과 같은 EBIC

$$EBIC = -\log(\det(\hat{\boldsymbol{\Theta}})) + \mathrm{tr}(\hat{\boldsymbol{\Theta}}\boldsymbol{S}) + df(\log(n) + 2\gamma log(p))/n$$

를 제안하였고, Chen과 Chen (2012)에서는 일반화선형모형으로 확장하였다. 여기서 γ는 p가 n에 대하여 지수적으로 증가하는 증가율과 관련된 상수로서 본 논문에서는 $\gamma = 2$를 사용하였다.

3. 자료 분석

이 절에서는 앞에서 설명한 모형선택기준 AIC, BIC, CV, CV-SD, EBIC를 모의실험과 실제 금융 자료에 대하여 비교한다. 모든 분석은 R의 `glasso` 패키지의 `glassopath` 함수를 이용하였다.

3.1. 모의실험

모의실험에서는 그래프의 형태와 복잡도에 따른 모형선택기준들을 민감도와 특이도의 관점에서 비교하였다. 그래프의 복잡도는 전체 노드들 중에서 적어도 하나 이상의 간선들을 갖는 노드들의 비율을 고려하였으며 10%, 30%, 50%의 세 수준에서 실험하였다. 이 비율을 NSR (non-sparsity ratio)라고 부르기로 하자. 또한 완전과 경로 두 가지 형태의 그래프를 고려하였다. 그래프의 구조는 정확도 행렬 $\boldsymbol{\Theta}$을 이용하여 정할 수 있다. $\boldsymbol{\Theta}$의 대각원들은 항상 1이고, 노드 i와 j간에 간선이 존재하는 경우에는 $\theta_{ij} =$

0.5이며 간선이 존재하지 않으면 $\theta_{ij} = 0$이다. 주어진 정확도 행렬에 대하여 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$을 평균벡터가 $\mathbf{0}$이 공분산 행렬이 $\boldsymbol{\Theta}^{-1}$인 다변량 정규분포에서 생성하였다. n과 p의 상대적인 크기에 따라 모형선택 기준의 성능이 달라질 수 있으므로, 자료 갯수를 모든 가능한 간선의 수로 나눈 비율인 $\frac{2n}{p(p-1)} \approx \frac{2n}{p^2}$의 네가지 수준 $10^1, 10^0, 10^{-1}, 10^{-2}$에서 실험하였다. 모의실험에서 자료의 갯수는 $n = 1,000$으로 고정하였으므로 위의 네가지 수준의 비율에 대응되는 p는 14, 45, 141, 447이다.

완전과 경로 두 가지 그래프의 구체적인 생성법은 다음과 같다. p와 NSR의 각 수준에 대하여 정확도 행렬에서 영이 아닌 θ_{ij}들의 갯수 r을 구할 수 있다. 완전 그래프는 $\binom{s}{2} = r$을 근사적으로 만족하는 자연수 s를 구한 후에 $1 \le i \ne j \le s$에 대하여 $\theta_{ij} = 0.5$으로 설정하고 나머지는 영으로 설정하였다. 경로 그래프의 경우에는 마찬가지로 $1 \le i \le r$에 대하여 $\theta_{i,i+1} = 0.5$으로 놓고 나머지는 영으로 설정하였다.

λ의 선택을 위하여 $[0, \lambda_{\max}]$의 구간을 로그 척도 (logarithmic scale)를 이용하여 100개의 구간으로 나누고 각 모형선택기준의 값을 구하여 최소가 되는 격자점 (grid point)을 최적값으로 하였다. 여기서 $\lambda_{\max} = \|\boldsymbol{S}\|_{\max} = \max\{|s_{ij}|\}$이다. 실험의 변동성을 고려하여 자료 생성 및 모형선택기준에 의한 추정의 과정을 100회 반복하였다.

Table 3.1 Weights for *df* in simulation

$2n/p^2$	AIC	BIC	EBIC
10^1	0.0020	0.0069	0.0174
10^0	0.0020	0.0069	0.0221
10^{-1}	0.0020	0.0069	0.0267
10^{-2}	0.0020	0.0069	0.0313

Table 3.1은 $2n/p^2$의 각 수준별로 AIC, BIC, EBIC의 *df* 앞의 계수값을 보여준다. AIC와 BIC의 계수는 $2/n$과 $\log n/n$으로 p에 의존하지 않기 때문에 본 모의실험에서는 동일한 값을 갖는다. EBIC의 계수는 $(\log n + 2\gamma \log p)/n$으로 p값에 따라 증가한다. 고정된 $2n/p^2$에 대하여 *df* 앞의 계수값이 클수록 복잡한 모형에 대하여 더 많은 벌점을 주는 것을 생각할 수 있다. 결과적으로 EBIC는 BIC에 대하여 p를 고려한 추가적인 벌점을 더 주는 것으로 볼 수 있다. Table 3.2와 Table 3.3은 각각 완전 그래프와 경로 그래프의 경우에 대한 모의실험 결과를 민감도와 특이도의 평균과 표준편차로 요약한다.

Table 3.2 Results for complete graphs in simulation

		Sensitivity				
$2n/p^2$	NSR	AIC	BIC	EBIC	CV	CV-SD
10^1	10%	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	30%	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	50%	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
10^0	10%	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	30%	1.0000 (0.0000)	0.8033 (0.1037)	0.0835 (0.0396)	1.0000 (0.0000)	0.7854 (0.2212)
	50%	1.0000 (0.0000)	0.1072 (0.0442)	0.0203 (0.0127)	1.0000 (0.0000)	0.4544 (0.2959)
10^{-1}	10%	0.9575 (0.0262)	0.6430 (0.0623)	0.2817 (0.0784)	0.9436 (0.0190)	0.4003 (0.1582)
	30%	0.3070 (0.0406)	0.0624 (0.0154)	0.0142 (0.0054)	0.2890 (0.0201)	0.0305 (0.0363)
	50%	0.1678 (0.0356)	0.0187 (0.0048)	0.0041 (0.0017)	0.1557 (0.0143)	0.0059 (0.0078)
10^{-2}	10%	0.2276 (0.0320)	0.0847 (0.0173)	0.0303 (0.0077)	0.2324 (0.0327)	0.0659 (0.0334)
	30%	0.0890 (0.0130)	0.0169 (0.0034)	0.0032 (0.0010)	0.0843 (0.0103)	0.0136 (0.0090)
	50%	0.0506 (0.0087)	0.0054 (0.0014)	0.0010 (0.0003)	0.0476 (0.0028)	0.0052 (0.0046)

		Specificity				
$2n/p^2$	NSR	AIC	BIC	EBIC	CV	CV-SD
10^1	10%	0.7718 (0.1200)	0.9738 (0.0224)	0.9946 (0.0083)	0.7957 (0.0882)	1.0000 (0.0000)
	30%	0.3173 (0.1274)	0.8047 (0.0861)	0.9483 (0.0322)	0.3774 (0.1312)	0.9891 (0.0174)
	50%	0.2493 (0.0535)	0.5391 (0.1469)	0.8764 (0.0611)	0.2513 (0.0529)	0.8821 (0.1020)
10^0	10%	0.8232 (0.0527)	0.9835 (0.0074)	0.9964 (0.0024)	0.8539 (0.0459)	0.9992 (0.0019)
	30%	0.2567 (0.0159)	0.9829 (0.0116)	1.0000 (0.0000)	0.4522 (0.0721)	0.9559 (0.0555)
	50%	0.2418 (0.0159)	0.9996 (0.0009)	1.0000 (0.0001)	0.2434 (0.0161)	0.9105 (0.1601)
10^{-1}	10%	0.9381 (0.0225)	0.9970 (0.0012)	0.9999 (0.0001)	0.9534 (0.0084)	0.9991 (0.0020)
	30%	0.9628 (0.0128)	0.9993 (0.0004)	1.0000 (0.0001)	0.9692 (0.0047)	0.9995 (0.0015)
	50%	0.9756 (0.0114)	0.9998 (0.0002)	1.0000 (0.0000)	0.9800 (0.0040)	0.9999 (0.0002)
10^{-2}	10%	0.9901 (0.0042)	0.9993 (0.0003)	0.9999 (0.0000)	0.9894 (0.0044)	0.9995 (0.0006)
	30%	0.9931 (0.0023)	0.9997 (0.0001)	1.0000 (0.0000)	0.9939 (0.0016)	0.9998 (0.0003)
	50%	0.9963 (0.0013)	0.9999 (0.0000)	1.0000 (0.0000)	0.9968 (0.0004)	0.9999 (0.0002)

Table 3.2의 완전 그래프에 대한 결과를 살펴보면 다음과 같다. NSR이 증가할수록, 즉 그래프가 조밀할수록, 민감도는 감소하는 경향이 있다. 마찬가지로 $2n/p^2$이 감소하면, 즉 자료 갯수가 노드수에 비해 상대적으로 작을수록, 민감도는 떨어지는 것을 볼 수 있다. 비교적 더 많은 간선들을 선택하는 기준인 AIC와 CV의 민감도가 가장 높고 EBIC는 가장 낮은 값을 갖는다. BIC와 CV-SD는 그 사이의 값을 가진다. 특이도를 살펴보면 다음과 같다. $2n/p^2$이 감소하면 전반적으로 특이도가 증가하며, NSR이 증가할수록 특이도는 감소하는 경향을 보인다. 특이도 측면에서 모형선택기준을 살펴보면 EBIC가 가장 높고 AIC와 CV가 가장 낮게 나타나 민감도와 반대의 경향성을 보인다. 흥미롭게도 $2n/p^2 = 10^1$이고 NSR이 30% 또는 50%인 경우 EBIC와 CV-SD의 민감도와 특이도는 모두 0.85이상으로 다른 기준들에 비하여 높은 값을 갖는다.

Table 3.3 Results for path graphs in simulation

		Specificity				
$2n/p^2$	NSR	AIC	BIC	EBIC	CV	CV-SD
10^1	10%	0.7599 (0.1152)	0.9713 (0.0213)	0.9937 (0.0094)	0.7838 (0.0869)	1.0000 (0.0000)
	30%	0.6099 (0.1188)	0.8810 (0.0427)	0.9286 (0.0219)	0.6459 (0.0964)	0.9590 (0.0093)
	50%	0.5123 (0.1356)	0.7696 (0.0540)	0.8361 (0.0375)	0.5854 (0.1201)	0.8748 (0.0265)
10^0	10%	0.8856 (0.0459)	0.9826 (0.0049)	0.9921 (0.0027)	0.9080 (0.0133)	0.9958 (0.0013)
	30%	0.7742 (0.0254)	0.9321 (0.0073)	0.9588 (0.0073)	0.7786 (0.0345)	0.9587 (0.0119)
	50%	0.7374 (0.0918)	0.8701 (0.0113)	0.9186 (0.0086)	0.7796 (0.0115)	0.9047 (0.0229)
10^{-1}	10%	0.9440 (0.0037)	0.9869 (0.0018)	0.9939 (0.0010)	0.9440 (0.0041)	0.9902 (0.0038)
	30%	0.8835 (0.0033)	0.9611 (0.0018)	0.9777 (0.0017)	0.8835 (0.0037)	0.9509 (0.0131)
	50%	0.8632 (0.0036)	0.9211 (0.0093)	0.9590 (0.0036)	0.8634 (0.0038)	0.9022 (0.0250)
10^{-2}	10%	0.9808 (0.0006)	0.9890 (0.0009)	0.9965 (0.0002)	0.9807 (0.0007)	0.9865 (0.0037)
	30%	0.9321 (0.0012)	0.9756 (0.0005)	0.9889 (0.0006)	0.9322 (0.0014)	0.9564 (0.0125)
	50%	0.9079 (0.0013)	0.9622 (0.0006)	0.9805 (0.0008)	0.9080 (0.0014)	0.9136 (0.0134)

Table 3.3은 경로 그래프에 대한 모의실험의 결과를 보여준다. 민감도는 모형선택기준간에 차이가 없이 모두 1이므로 생략하였다. 완전 그래프의 경우와 유사하게 $2n/p^2$이 작아지면 전반적으로 특이도가 증가하는 경향을 보이며, NSR이 증가할수록 특이도는 감소한다. 간결한 그래프를 선택하는 BIC, CV-SD, EBIC가 AIC와 CV에 비하여 특이도 측면에서 더 좋다는 것을 알 수 있다. 또한 완전 그래프의 경우처럼 EBIC는 $2n/p^2 = 10^1$이고 NSR이 30% 또는 50%인 경우에 다른 방법들에 비하여 높은 특이도를 갖는다.

3.2. 실제자료

한국은행 경제통계시스템에서는 한국은행 및 타기관 작성 통계 중 정책수립 및 동향분석에 유용한 통계지표를 선정하여 100대 통계지표로 제공하고 있다. 본 논문에서는 한국은행의 100대 통계지표와 국가통계포털 (KOSIS)에서 제공하는 코스피 (KOSPI)와 코스닥 (KOSDAQ)의 산업별 시가총액 자료를 통합하여 국내 산업구조에 대한 그래프를 추정하고자 한다.

100대 통계지표에서는 2005년 4월부터 2013년 9월까지의 자료를 사용하였고, 총 103개의 지표들 중 그 기간의 값이 없는 6개와 년도별로 제공되는 7개의 지표들을 제외한 나머지 90개의 지표들을 분석할 변수로 포함시켰다. 90개의 지표들 중 일별 혹은 분기별로 제공되는 것들은 분석을 위해 월별로 변환하였다. 100대 통계지표와 동일한 기간의 어업, 광업, 음식료품 등 24개의 코스피 산업분류와 제조, 섬유·의류, 음식료·담배 등 30개의 코스닥 산업분류의 총 54개를 분석할 변수로 포함시켰다. 전월대비 증감을 보기 위하여 증감을 나타내는 변수들을 제외한 나머지는 변수별로 차분한 후 표준화 하였다. 따라서 분석에 사용될 전체 변수와 자료의 갯수는 각각 $p = 144$이고 $n = 102$이다.

2000년대 말 미국의 금융 시장에서 시작되어 전 세계로 파급된 대규모의 금융위기가 있었으며, 당시 미국 뿐만 아니라 우리나라의 주가도 큰 폭으로 하락했다. 본 논문에서는 우선 전체 자료에 대한 그래프 추정결과를 살펴보고, 자료를 금융위기 이전 (2005년 9월 ∼ 2007년 10월), 금융위기 기간 (2007년 11월 ∼ 2008년 12월), 금융위기 이후 (2009년 1월 ∼ 2013년 9월)로 나누어 그래프 추정 결과를 비교하기로 한다.

전체 기간의 자료에 대한 분석

우선 2005년 4월에서 2013년 9월까지 전체 기간의 자료를 분석하여 다음과 같은 결과를 얻었다. 모형선택기준 AIC, BIC, EBIC, CV, CV-SD에 의해 최적으로 선택된 λ값은 각각 0.0069, 0.1398, 0.8590, 0.0741, 0.2003이다. λ값이 작을수록 추정된 그래프가 조밀하고 그 값이 클수록 추정된 그래프는 간결하다. 편의상 가장 간결한 그래프 추정치를 주는 EBIC를 중심으로 해석하기로 한다. Figure 3.1 (a)는 EBIC에 추정된 그래프를 보여주며 간선이 존재하는 노드들에 대응되는 변수를 정리하면 다음과 같다. 단 각 항목 내의 변수들은 동일 항목의 변수들로만 간선이 연결되어 있다.

- 제조업생산지수, 제조업출하지수, 제조업가동률지수 [산업활동·소비·투자]
- 소매판매액지수, 도소매업지수 [산업활동·소비·투자]
- 경제활동인구, 취업자수, 고용률 [고용·임금·가계]
- KORIBOR, CD수익률, 통안증권수익률, 예금은행 수신금리, 예금은행 대출금리 [금리]
- 국고채수익률(3년), 국고채수익률(5년) [금리]
- 소비자물가지수, 생활물가지수 [물가]
- 수출, 수출금액지수 [국제수지·대외거래]
- 수입, 수입금액지수 [국제수지·대외거래]
- 원/달러 환율(기준), 원/달러 환율(종가) [환율·외환보유]

결과를 살펴보면 100대 통계지표의 변수들에 대해서만 간선이 존재하며 코스피와 코스닥의 변수들 사이에는 간선이 존재하지 않는다. 사각괄호안은 100대 통계지표의 대분류 항목을 나타낸다. 100대 통계지표의 동일 대분류에 속하는 변수들끼리 관련된 것을 알 수 있다. 그 이유는 산업활동, 물가, 고용, 금리 등 다양한 경제현상을 설명하기 위해서 대분류 항목 간에는 연관성이 적고 대분류 항목 내의 소분류 항목 간에는 연관성이 크도록 지표를 구성하였기 때문이다. 예를 들어, 첫 번째 항목인 산업활동·소비·투자 내의 변수인 제조업 생산지수, 제조업출하지수, 제조업가동률지수는 서로 연관이 있는데 Figure 3.1 (a)에서 확인할 수 있다.

AIC등의 다른 기준에 의해 추정된 그래프에서는 EBIC에 비해 좀더 많은 간선을 추정하지만 대분류 항목 내부의 간선들이 대부분이다. 이는 앞서 설명한 것과 같이 대분류 내부에서는 연관성이 있지만 대분류들간에는 연관성이 적도록 지표를 구성하였기 때문인 것으로 생각된다. 한 가지 언급할 만한 것은, 가구당월평균소득 [고용·임금· 가계]과 교육비지출률 [경제관련 사회통계]은 서로 다른 대분류에 속하지만 EBIC이외의 다른 모든 기준에 의해서 간선을 갖는 것으로 추정되었다. 따라서 자료의 전체 기간에서 가구당월평균소득이 증가하면 교육비지출은 증가하고 가구당월평균소득이 감소하면 교육비지출은 감소하는 것으로 해석할 수 있다.

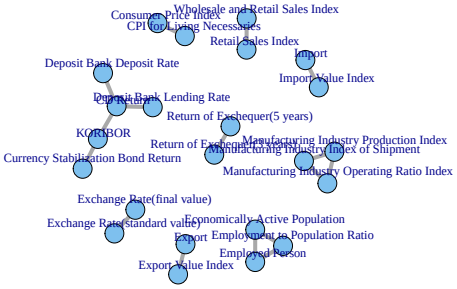
금융위기를 고려한 기간별 자료에 대한 분석

이제 자료를 금융위기 이전, 금융위기, 금융위기 이후의 세 기간으로 나누어 그래프를 추정한 후 그 결과를 비교해보자. $p = 144$로 고정되어 있지만 기간별로 자료를 분할함으로써 자료의 갯수가 각각 14, 32, 56개로 줄어들게 되어 고차원성이 심화된다. 따라서 `glasso` 알고리즘의 계산시간이 매우 늘어나게 되어 자료를 분할하는 방법인 CV나 CV-SD는 계산하는데 시간이 오래 걸리기 때문에 생략한다. 알고리즘의 계산 시간에 대하여 관심있는 독자는 Friedman 등 (2008)을 참고하기 바란다.

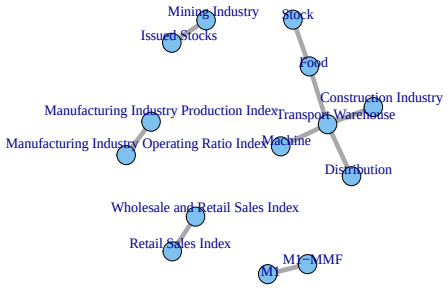
Table 3.4 λ selected by AIC, BIC, and EBIC

	AIC	BIC	EBIC
Before financial crisis	0.0111	0.2479	0.7984
During financial crisis	0.0185	0.0185	3.4143
After financial crisis	0.0096	0.2341	0.8797

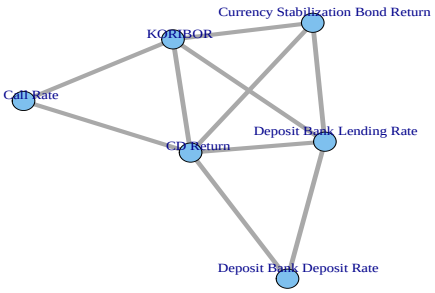
Table 3.4는 기간별 자료에 대하여 모형선택기준 AIC, BIC, EBIC에 의해 선택된 최적의 λ값을 보여준다. AIC, BIC, EBIC 순으로 그래프를 조밀하게 추정함을 알 수 있다. Figure 3.1 (b)-(d)는 기간별로 EBIC를 이용하여 추정한 그래프를 보여준다. 기간별로 그래프 구조가 차이가 남을 쉽게 알 수 있다.



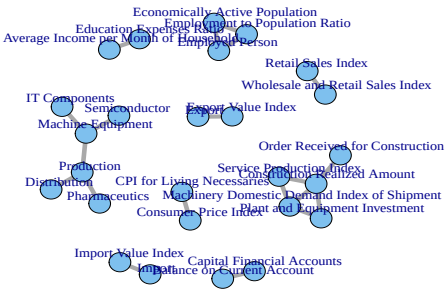
(a) All period



(b) Before financial crisis



(c) All period



(d) Before financial crisis

Figure 3.1 Estimated graph using EBIC

금융위기 이전에 대한 추정 그래프 Figure 3.1 (b)에서 간선이 존재하는 노드들에 해당되는 변수를 정리하면 다음과 같다.

- 제조업생산지수, 제조업가동률지수 [산업활동·소비·투자]
- 소매판매액지수, 도소매업지수 [산업활동·소비·투자]
- M1(협의통화), M1-MMF [통화·금융]
- 주식발행액 [증권], 광업 [코스피]
- 음식료품, 기계, 유통업, 건설업, 운수창고업, 증권 [코스피]

예를 들어, 네번째 항목을 보면 100대 통계지표의 주식발행액과 코스피의 광업이 연결되어 있는데, 이는 발행되는 주식 수의 변화와 광업의 시가총액 변화에 연관성이 있음을 의미한다.

금융위기 기간에 대한 추정 그래프 Figure 3.1 (c)에서 간선이 존재하는 노드들에 해당되는 변수들은 다음과 같다.

- 콜금리, KORIBOR, CD수익률, 통안증권수익률, 예금은행 수신금리, 예금은행 대출금리 [금리]

금융위기 기간에는 100대 통계지표의 금리에 해당하는 부분에서만 간선이 존재하는데, 금융위기 하에서 금리와 관련된 변수들끼리 긴밀하게 연결되어 있다. 사실 앞서 살펴본 전체 기간의 결과에서도 금리 항목에 속한 일부 변수들이 서로 연결되어 있는데 금융위기 기간과의 차이점은 콜금리 변수의 유무이다. 금융위기 기간에서는 콜금리가 CD 수익률과 KORIBOR와 연결되어 있다. 콜금리는 초단기금융시장을, CD 수익률과 KORIBOR는 단기금융시장의 자금 상황을 반영한다. 따라서 금융위기 기간에는 초단기 금융시장의 변화가 단기 금융시장의 변화와 연관성이 있음을 알 수 있다.

마지막으로 금융위기 이후에 대한 추정 그래프 Figure 3.1 (d)에서 간선이 존재하는 노드들에 해당되는 변수들은 다음과 같이 요약된다.

- 서비스업생산지수, 설비투자지수, 기계류내수출하지수, 건설수주액, 건설기성액 [산업활동·소비·투자]
- 소매액판매지수, 도소매업지수 [산업활동·소비·투자]
- 경제활동인구, 취업자수, 고용률 [고용·임금·가계]
- 가구당월평균소득 [고용·임금·가계], 교육비지출률 [경제관련 사회통계]
- 소비자물가지수, 생활물가지수 [물가]
- 경상수지, 자본·금융계정 [국제수지·대외거래]
- 수출, 수출금액지수 [국제수지·대외거래]
- 수입, 수입금액지수 [국제수지·대외거래]
- 제조, 제약, 기계·장비, 유통, 반도체, IT부품 [코스닥]

금융위기 이후에는 100대 통계지표 외에도 코스닥의 변수들간에 간선이 존재한다. 코스닥의 변수중에 기계·장비와 제조가 연결되어 있고, 기계·장비는 반도체와 IT부품과, 제조는 제약과 유통과 연결되어 있다. 이들 산업군이 기계·장비와 제조를 중심으로 연관관계가 있음을 알 수 있다.

4. 결론

본 논문에서는 가우스 그래프 모형에서 축소 추정을 하는 그래프 LASSO에서 여러가지 모형선택기준을 모의실험을 통하여 비교하였다. AIC와 CV는 조밀하게 연결된 그래프를 추정할 때 민감도 측면에서 좋게 나타났고, BIC, EBIC, CV-SD는 간결한 그래프를 추정할 때 특이도 측면에서 좋음을 알 수 있었다. 각 기준마다 조금씩 차이는 있지만 전반적으로 노드의 개수가 증가하면 민감도는 감소하고 특이도는 증가하며, 노드들이 더 조밀하게 연결될수록 민감도와 특이도는 감소하는 것으로 나타났다. 흥미롭게도 p가 n에 대하여 지수적으로 증가하는 고차원을 상정한 기준인 EBIC가 차원에 비하여 자료의 갯수가 많고 조밀한 그래프일수록 다른 기준들보다 더 좋게 나타났다. 또한 한국은행에서 제공하는 100대 통계지표와 코스피, 코스닥의 산업별 시가총액을 통합한 자료에 대하여 전체 기간의 자료 및 금융위기를 고려한 기간별 자료로 나누어 추정하여 경제 상황을 나타내는 지표들간의 관계를 시각적으로 살펴보았다.

본 논문에서 모의실험과 자료분석을 통해 살펴본 결과에 기반하여 다음과 같은 제언을 할 수 있다. 사실 EBIC의 모형선택에서의 일치성은 일반화선형모형에 대한 것으로 그래프 LASSO에서 연구된 바는

없다. 일반화선형모형의 경우 γ는 n과 p의 차수가 $p = O(n^{\kappa})$인 경우 $\gamma > 1 - \frac{1}{2\kappa}$을 만족하도록 선택해야 모형선택의 일치성이 성립한다. 유의한 간선의 대략적인 비율을 아는 경우에는 조건을 만족하는 대략적인 γ값의 범위를 계산할 수도 있지만 유의한 간선의 비율을 모르는 경우에는 계산이 불가능하다. 따라서 실제 자료분석에서는 몇 가지 수준의 γ값을 시험해보고 모형의 간결성을 살펴본 후 적절한 γ값을 선택하는 수 밖에 없을 것이다. 또한 본 논문에서처럼 EBIC이외에도 AIC 와 BIC 등 몇 가지 선택기준을 적용한 후에 공통적으로 선택되는 간선들과 다르게 선택되는 간선들을 비교해 보는 것이 바람직할 것으로 생각된다.

그래프 LASSO는 기본적으로 연속형 자료에 대한 가우스 그래프 모형에 기반하는데, 현실의 많은 자료는 연속형뿐만 아니라 이산형 변수들을 포함하는 혼합자료 형태인 경우가 일반적이다. 따라서 Tibshirani 등 (2005)의 규합벌점 (fused penalty) 등을 이용한 방법을 고려해 볼 수 있다. 규합벌점화의 경우 연속형과 이산형 변수가 혼재하기 때문에 모형선택이 그래프 LASSO보다 더 어렵기 때문에 적절한 모형선택기준에 대한 연구가 필요하다. 또한 유전자, 사회관계망, 교통 등 보다 다양한 자료에 대하여 여러가지 모형선택기준의 특성을 비교해 볼 수도 있다.

References

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, **9**, 485-516.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759-771.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, **22**, 555.

Cho, K.-H. and Park, H.-C. (2012). A study on 3-step complex data mining in society indicator survey. *Journal of the Korean Data & Information Science Society*, **23**, 983-935.

Choi, I., Kang, D., Lee, J., Kang, M., Song, D., Shin, S. and Son, Y. S. (2012). Prediction of the industrial stock price index using domestic and foreign economic indices. *Journal of the Korean Data & Information Science Society*, **23**, 271-283.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157-175.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Statistics*, **3**, 521-541.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, **98**, 1-15.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of statistical learning*, Springer, New York.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, **19**, 140-155.

Lauritzen, S. L. (1996). *Graphical models*, Clarendon Press, Oxford.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91-108.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19-35.

Journal of the Korean Data & Information Science Society
2014, **25**(4), 881–891

http://dx.doi.org/10.7465/jkdi.2014.25.4.881
한국데이터정보과학회지

Comparison of model selection criteria in graphical LASSO[†]

Hyeongseok Ahn[1] · Changyi Park[2]

[12]Department of Statistics, University of Seoul

Received 26 June 2014, revised 14 July 2014, accepted 18 July 2014

Abstract

Graphical models can be used as an intuitive tool for modeling a complex stochastic system with a large number of variables related each other because the conditional independence between random variables can be visualized as a network. Graphical least absolute shrinkage and selection operator (LASSO) is considered to be effective in avoiding overfitting in the estimation of Gaussian graphical models for high dimensional data. In this paper, we consider the model selection problem in graphical LASSO. Particularly, we compare various model selection criteria via simulations and analyze a real financial data set.

Keywords: Conditional independence, Gaussian graphical model, high-dimensional data.

[†] This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1A2004901).

[1] Graduate student, Department of Statistics, University of Seoul, Seoul 130-743, Korea.
[2] Corresponding author: Associate professor, Department of Statistics, University of Seoul, Seoul 130-743, Korea. E-mail: park463@uos.ac.kr