The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.5.747
26(5), 747–758

A Comparative Study of Covariance Matrix Estimators in High-Dimensional Data

DongHyuk Lee[a] · Jae Won Lee[a,1]

[a]Department of Statistics, Korea University

(Received July 29, 2013; Revised September 30, 2013; Accepted September 30, 2013)

Abstract

The covariance matrix is important in multivariate statistical analysis and a sample covariance matrix is used as an estimator of the covariance matrix. High dimensional data has a larger dimension than the sample size; therefore, the sample covariance matrix may not be suitable since it is known to perform poorly and even not invertible. A number of covariance matrix estimators have been recently proposed with three different approaches of shrinkage, thresholding, and modified Cholesky decomposition. We compare the performance of these newly proposed estimators in various situations.

Keywords: Covariance matrix estimation, shrinkage, thresholding, modified Choelsky decomposition.

1. 서론

마이크로어레이(microarray) 기술과 질량분석(mass spectrometry) 기술의 발달로 수백에서 수만에 이르는 유전자들과 단백질들을 동시에 검출할 수 있게 되면서 변수의 수가 표본의 수보다 큰 고차원 데이터를 분석하는 많은 통계적 방법들이 개발되고 발전되어 왔다. 이러한 통계적 방법들은 기본적으로 기존의 전통적인 다변량 방법들에 바탕을 두고 있다. 보통의 다변량 통계분석 방법에서 공분산행렬(covariance matrix)은 중요한 역할을 담당하고 있으며 주성분 분석, 인자분석, 군집분석, 판별분석 등에서 표본공분산행렬은 참공분산행렬의 추정량으로 이용된다. 그러나 고차원 데이터에서는 역행렬과 관련된 문제들 (Ledoit과 Wolf, 2004), 고유값 구조 등의 문제 (Schäfer와 Strimmer, 2005)로 표본공분산행렬을 그대로 사용할 수 없다. 또한 SAM (Tusher 등, 2001)과 PAM (Tibshirani 등, 2002) 등과 같은 방법들에서 변수들 사이의 독립을 가정한 분류분석 방법이 제안되어 현재까지 많이 쓰이고 있지만, 독립성 가정이 성립하지 않는 경우 정보의 손실 등으로 인해 분류오차가 늘어날 소지가 있다.

고차원 데이터에서 표본공분산행렬을 대체할 수 있는 추정량들의 연구는 Ledoit와 Wolf (2004)에 의해서 표본공분산행렬과 단위행렬과의 볼록 선형결합(convex linear combination)인 축소추정(shrinkage estimation)이 연구되었고, Schäfer와 Strimmer (2005)는 단위행렬 이외의 다른 구조를 가지는 표적행렬과의 볼록 선형결합에 대한 연구로 확장하였다. 또한 Rothman 등 (2009)은 일반화 경제추

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2008686).
[1]Corresponding author: Professor, Department of Statistics, Korea University, Anam-Dong, Seongbuk-Gu, Seoul 136-701, Korea. E-mail: jael@korea.ac.kr

정(generalized thresholding) 방법을 제안하였고, Cai와 Liu (2011)는 이를 확장한 적응적 경계추정(adaptive thresholding) 방법을 연구하였다. 한편, 수정된 콜레스키 분해법을 응용한 추정방법들도 연구되었다 (Huang 등, 2006; Levina 등, 2008; Rothman 등, 2010). 이러한 추정량들에 대한 이론적 연구로 Bickel과 Levina (2008a, 2008b)는 경계추정(thresholding estimation)과 밴딩추정(banding estimation)에 대해 그 점근적 성질들에 대하여 연구하였고, Cai 등 (2010)은 미니맥스 관점에서 최적 수렴율(rate of convergence)에 대하여 연구하였다.

본 연구에서는 널리 사용되고 있는 분석방법들의 가정들을 고려한 여러 상황하에서 추정량들의 성능 비교를 고려하였다. 본 논문의 구성은 다음과 같다. 2절에서는 고차원 데이터에서 공분산행렬을 추정할 수 있는 여러 추정량들을 소개하고, 3절에서는 여러 상황들을 고려한 모의실험 결과에 대해 설명하였다. 4절에서는 실제 자료에 적용하였으며, 마지막으로 5절에서는 본 연구에 대한 결론과 전체적인 내용에 대하여 기술하였다.

2. 공분산행렬 추정 방법

표본 크기 n을 가지는 데이터 $X_i = (X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$에서, X_{ij}를 i번째 개체의 j번째 변수의 유전자 발현량(gene expression) 또는 단백질 강도(protein intensity)라고 할 때, $\Sigma = (\sigma_{ij})_{p\times p}$를 참 공분산행렬, $S = (s_{ij})_{p\times p}$를 표본 공분산행렬이라고 하자. 여기에서 p는 변수의 수이며 $n \ll p$를 가정한다.

2.1. Schäfer와 Strimmer 추정량(Schäfer and Strimmer Estimator; SSE)

Schäfer와 Strimmer (2005)는 표본공분산행렬과 특정 구조를 갖는 표적행렬 T와의 볼록 선형결합인 축소추정량 $\Sigma_{shrinkage}$을 추정하는 방법을 제안하였다.

$$\Sigma_{shrinkage} = (1-\alpha)S + \alpha T, \quad \alpha \in [0,1].$$

여기서 α는 축소강도(shrinkage intensity)로, $\alpha = 1$인 경우에는 축소추정량은 표적행렬 T와 같아지고, $\alpha = 0$인 경우에는 표본공분산행렬 S와 같아진다. Ledoit와 Wolf (2004)는 평균제곱오차 관점에서 단위행렬을 표적행렬로 한 경우 최적 축소강도 $\hat{\alpha}$를 해석적으로 구하였고, Schäfer와 Strimmer (2005)는 이를 확장하여 특정한 구조를 갖는 표적행렬에 대하여 최적 축소강도를 계산하였다.

$$\hat{\alpha} = \begin{cases} \dfrac{\sum_{i\neq j} \widehat{\mathrm{var}}(s_{ij}) + \sum_i \widehat{\mathrm{var}}(s_{ii})}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii}-1)^2}, & \text{if } T = I, \\ \dfrac{\sum_{i\neq j} \widehat{\mathrm{var}}(s_{ij}) + \sum_i \widehat{\mathrm{var}}(s_{ii})}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii}-\nu)^2}, & \text{if } T = D_1, \\ \dfrac{\sum_{i\neq j} \widehat{\mathrm{var}}(s_{ij})}{\sum_{i\neq j} s_{ij}^2}, & \text{if } T = D_2, \end{cases}$$

여기서 $D_1 = \begin{cases} \nu = \mathrm{avg}(s_{ii}), & \text{if } i=j, \\ 0, & \text{if } i\neq j \end{cases}$ 이고, $D_2 = \begin{cases} s_{ii}, & \text{if } i=j, \\ 0, & \text{if } i\neq j \end{cases}$ 이다. 즉, D_1은 표본 분산들의 평균을 원소로 가지는 대각행렬이 되고, D_2는 표본공분산행렬의 비대각 원소를 제거한 대각행렬이 된다.

2.2. Fisher와 Sun 추정량(Fisher and Sun Estimator; FSE)

Fisher와 Sun (2011)은 데이터가 다변량 정규분포를 따른다는 가정 하에서 선형대수와 기대값 등을 이용하여 최적 축소강도 $\hat{\alpha}$를 유도한다. 앞의 Schäfer와 Strimmer (2005) 추정량과의 차이점은, Fisher와 Sun (2011)은 데이터가 정규분포를 따른다는 가정 하에서 최적 축소강도를 추정해 내는 것이라면, Schäfer와 Strimmer (2005)는 분포에 대한 특별한 가정 없이 해석적으로 계산한다는 것이다.

$$\hat{\alpha} = \begin{cases} \dfrac{\frac{1}{n}\hat{a}_2 + \frac{p}{n}\hat{a}_1^2}{\frac{n+1}{n}\hat{a}_2 + \frac{p}{n}\hat{a}_1^2 - 2\hat{a}_1 + 1}, & \text{if } T = I, \\ \dfrac{\frac{1}{n}\hat{a}_2 + \frac{p}{n}\hat{a}_1^2}{\frac{n+1}{n}\hat{a}_2 + \frac{p-n}{n}\hat{a}_1^2}, & \text{if } T = D_1, \\ \dfrac{\frac{1}{n}\left(\hat{a}_2 + p\hat{a}_1^2\right) - \frac{2}{n}\hat{a}_2^*}{\frac{n+1}{n}\hat{a}_2 + \frac{p}{n}\hat{a}_1^2 - \frac{n+2}{n}\hat{a}_2^*}, & \text{if } T = D_2, \end{cases}$$

여기서 $\hat{a}_1 = \mathrm{tr}S/p$, $\hat{a}_2 = n^2/\{(n-1)(n+2)p\}[\mathrm{tr}S^2 - (1/n)(\mathrm{tr}S)^2]$이고, $\hat{a}_2^* = n/\{(n+2)p\}\mathrm{tr}(D_2^2)$이다.

2.3. 일반화 경계 추정량(Generalized Thresholding Estimator; GTE)

경계 추정량(thresholding estimator)은 표본공분산행렬의 원소들에 경계함수(thresholding function)를 적용하는 방법이다. 만약 표본공분산행렬의 원소가 미리 정해진 경계함수의 파라미터(parameter)보다 작게 되면 경계 추정량에서 해당하는 원소는 0이 될 것이다.

일반화 경계 추정량(Generalized Thresholding Estimator; GTE)을 구하는 절차는 hard, soft, adaptive lasso, SCAD 등의 경계함수를 표본 공분산행렬의 전체 원소에 적용시키는 것이다. Bickel와 Levina (2008a)은 hard 경계함수의 경우에 추정량의 점근적 성질을 보였고, Rothman 등 (2009)이 다른 경계함수에 대해 적용한 추정량을 제안하였으며 그 이론적 성질 등을 연구하였다.

2.4. 적응적 경계 추정량(Adaptive Thresholding Estimator; ATE)

적응적 경계 추정량(Adaptive Thresholding Estimator; ATE)은 표본 공분산행렬의 각 원소마다 다른 파라미터의 경계함수를 적용하는 방법이다. 일반화 경계 추정량이 하나의 경계함수를 표본공분산행렬의 모든 원소에 적용하는 방법이라면 적응적 경계 추정량은 원소마다 다른 경계함수를 적용시키는 방법이다.

$$\Sigma_{adap} = (s_{ij}^*)_{p\times p},$$

여기서 $s_{ij}^* = t_{\lambda_{ij}}(s_{ij})$이고, $t_\lambda(z)$는 파라미터 λ를 가지는 hard, soft, adaptive lasso, SCAD 등의 경계함수이다. Cai와 Liu (2011)는 데이터로부터 원소별 λ_{ij}를 구하는 방법을 제안하였다:

$$\lambda_{ij} := \lambda_{ij}(\delta) = \delta\sqrt{\hat{\theta}_{ij}\log\frac{p}{n}}\ ,$$

여기서 $\hat{\theta}_{ij} = 1/n \sum_{k=1}^{n} [(X_{ki} - \bar{X}^i)(X_{kj} - \bar{X}^j) - s_{ij}]^2$이고, $\bar{X}^i = 1/n \sum_{k=1}^{n} X_{ki}$는 i번째 변수의 평균이다. 따라서 $\mathrm{cov}(X_i, X_j)$에 대응하는 표본으로부터 구한 추정치가 크게 되면 경계함수의 파라미터 λ_{ij}는 커지게 되고, 결과적으로 얻어진 추정치는 0에 가까워질 것이다.

2.5. 수정 콜레스키 분해에 기초한 추정량(Modified Cholesky Decomposition Estimator; MCDE)

수정 콜레스키 분해에 기초한 추정량(Modified Cholesky Decomposition Estimator; MCDE)은 수정 콜레스키 분해인 $\Sigma = LDL'$을 이용하여, L을 회귀 계수로, D를 분산으로 해석하는 방법이다. $\mathbf{x} = (x_1, \ldots, x_p)'$를 평균 0벡터, 분산 Σ를 갖는 확률벡터라고 하고, $1 \le i \le p$에 대하여 $\hat{x}_i$를 $x_1, \ldots, x_{i-1}$에 대한 회귀 예측량, $\epsilon_i = x_i - \hat{x}_i$는 분산 $\sigma_i^2 = \mathrm{var}(\epsilon_i)$를 가지는 예측 오차라고 하면, $1 \le i \le p$에 대하여 $x_i = \sum_{j=1}^{i-1} \phi_{ij} x_j + \epsilon_i$로 표현할 수 있다. 그리고 $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$를 오차벡터라고 하면 위의 식을 동시에 $\epsilon = T\mathbf{x}$로 나타낼 수 있다. 여기서 T는 1을 대각원소로, $-\phi_{ij}$를 하비대각원소(lower off-diagonal)로 하는 하삼각행렬이다. 또한 가정에서 $\mathrm{cov}(\epsilon) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) = D$이기 때문에 $T\Sigma T' = D$가 성립한다.

이를 이용하면, 데이터 $\mathbf{x}$에 대한 다변량 정규분포 가정 하에서 로그 가능도함수를 D와 T를 이용하여 다음과 같이 표현할 수 있다:

$$-2l(\Sigma; \mathbf{x}) = \log|\Sigma| + \mathbf{x}'\Sigma^{-1}\mathbf{x} = \log|D| + \mathbf{x}'T'\Sigma^{-1}T\mathbf{x} = \sum_{i=1}^{p} \log\sigma_i^2 + \sum_{i=1}^{p} \frac{\epsilon_i^2}{\sigma_i^2}.$$

따라서 n개의 관측치들로부터 로그 가능도 함수는

$$-2l\left(\Sigma; \mathbf{x_1}, \ldots, \mathbf{x_n}\right) = \sum_{i=1}^{p} \left(n \log\sigma_i^2 + \sum_{j=1}^{n} \frac{\epsilon_{ji}^2}{\sigma_i^2} \right)$$

이 된다. 여기서 $\epsilon_{j1} = x_{j1}$이고, $i = 2, \ldots, p$에 대하여, $\epsilon_{ij} = x_{ji} - \sum_{l=1}^{i-1} \phi_{il} x_{jl}$이다. 고차원 데이터에는 보통의 최소제곱법을 적용할 수 없기 때문에 Huang 등 (2006)은 벌점가능도(penalized likelihood)를 적용하였다. 주어진 $\lambda > 0$에 대하여, 벌점 가능도는

$$-2l(\Sigma; \mathbf{x_1}, \ldots, \mathbf{x_n}) + \lambda p(\phi_{ij})$$

와 같다. Huang 등 (2006)은 ridge와 lasso 벌점을 고려한 벌점 가능도를 고려하였고 본 연구에서도 둘을 모두 사용하였다.

3. 모의실험

고차원 데이터를 분석하는 데 있어서 희박성(sparsity)에 기초하여 sparse 주성분 분석 (Zou 등, 2006; Shen과 Huang, 2008), sparse 판별분석 (Clemmensen 등, 2011; Mai 등, 2012), sparse 편최소제곱 회귀 (Chun과 Keles, 2010a, 2010b) 등 많은 방법들이 제안되고 사용되어 왔다. 따라서 Cai와 Liu (2011)의 모형 2와 같은 희박성을 갖는 구조를 고려하는 것이 현실적인 모의실험이 될 수 있을 것이다. 한편 마이크로어레이나 질량분석 데이터를 분석하는 목적 중의 하나는 서로 다른 집단을 구별하는 것이다. 따라서 전체 표본을 서로 다른 모집단이 혼합된 혼합 모형(Mixture model)으로 간주하여 상황을 설정하는 것을 고려하였다. 혼합모형 또한 고차원 데이터를 분석하는 모형화 방법으로 많은 연구가 진행되었다 (Ghosh와 Chinnaiyan, 2002; McLachlan 등, 2002; Bouveyron 등, 2007; McNicholas와

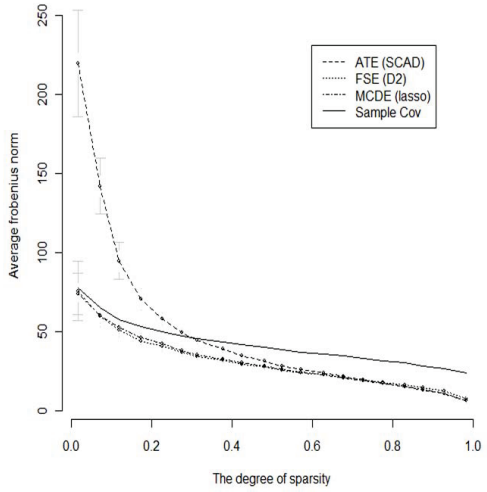

Figure 3.1. Average Frobenius norm losses by the degree of sparsity.

Murphy, 2010). 마지막으로 다변량 정규성을 가정할 수 없는 상황을 고려하여 다변량 감마분포에서 생성된 표본으로 모의실험을 진행하였다.

이렇게 설정된 여러 상황들 하에서 앞서 소개한 5가지 추정방법들의 성능을 비교해 보고자 한다. 비교의 측도로는 추정량과 참공분산행렬과의 프로베니우스 거리(Frobenius norm)를 이용하였고 표본 크기 100을 가정하여 각 상황에서 100번의 반복실험을 수행하였다. 소개된 방법들 중에서 축소추정량에 대하여 위의 세 가지 표적행렬들을 모두 고려하였고, 경계추정량에 대하여 hard, soft, adaptive lasso, SCAD 등의 경계함수를 모두 고려하였으며 수정 콜레스키 방법에서는 ridge와 lasso를 고려하였다. 각각의 방법들에 대하여 가장 좋은 성능을 보여주는 하위방법만을 골라 아래 결과에 표시하였다.

3.1. 희박성의 정도에 따른 모의실험

변수의 개수가 p인 경우 전체 비대각원소의 개수는 $p(p-1)/2$가 되고 q를 0인 비대각원소의 개수라고 한다면 $2q/p(p-1)$로 정의된 희박한 정도를 구할 수 있다. 양정치행렬이면서 특별한 패턴 없이 0인 비대각원소를 가지는 참공분산행렬을 만들기 위하여 $\Sigma = LDL'$인 콜레스키 분해를 이용하였다. L은 대각원소가 1인 단위 하삼각행렬로, 이 행렬의 비대각원소들 중 일정 비율을 0으로 만들게 되면 참 공분산행렬에 원하는 만큼의 희박성을 얻을 수 있다. 본 모의실험에서 표본의 크기는 100, 변수의 수는 200을 고려하였으며 다변량 정규분포에서 데이터를 생성하였다.

희박성을 0.01에서 0.98까지 약 0.05씩 증가시켜 전체 20개의 구간에서 추정량과 참공분산행렬과 평균거리를 계산하여 Figure 3.1의 그래프로 표현하였다. SCAD 경계함수를 갖는 적응적 경계 추정량(ATE)과 표본 공분산으로 이루어진 표적행렬을 갖는 축소추정량(Fisher와 Sun 방법; FSE), lasso 벌점이 고려된 콜레스키 방법(MCDE)가 각 추정량들에서 가장 좋은 성능을 보여주었다. 우선 데이터가 정규분포에서 생성되었기 때문에 정규성을 가정하는 Fisher와 Sun 추정량(FSE)과 콜레스키 방법(MCDE)의 성능이 희박성의 정도와 관계없이 우수함을 알 수 있다. 경계추정량에 대하여 살펴보면, 희박성의 정도가 0.4 미만이 되면 표본공분산행렬보다 성능이 뒤쳐지는 것을 확인할 수 있지만 희박성의 정도가 0.8 이상인 경우에는 미세하지만 다른 추정량보다 좋은 성능을 보여주었다. 경계 추정방법의

Table 3.1. Average Frobenius norm losses by mixture model. (Id: Identity target, D1: Diagonal type 1(same diagonal elements) target, D2: Diagonal type 2(different diagonal elements) target; H: Hard Thresholding, S: Soft thresholding, SC: SCAD thresholding, AL: Adaptive lasso thresholding; R: Ridge penalty, L: Lasso penalty)

True structure	p	SSE	FSE	GTE	ATE	MCDE	SamCov
40% sparsity	30	4.068^{Id}	4.060^{Id}	3.983^{S}	$\mathbf{3.924^{S}}$	4.037^{L}	4.763
	100	15.053^{D1}	$\mathbf{15.030^{D1}}$	18.773^{SC}	17.000^{SC}	15.589^{L}	19.88
	200	32.682^{D2}	$\mathbf{32.636^{D2}}$	43.002^{SC}	37.226^{SC}	36.636^{L}	47.769
70% sparsity	30	3.585^{Id}	3.573^{Id}	3.212^{S}	$\mathbf{3.205^{S}}$	3.257^{L}	4.683
	100	11.237^{D2}	11.193^{D2}	10.995^{S}	10.626^{SC}	$\mathbf{10.411^{L}}$	17.602
	200	22.022^{D2}	21.928^{D2}	22.544^{SC}	21.298^{SC}	$\mathbf{21.198^{L}}$	39.484
Identity	30	1.435^{Id}	1.376^{Id}	1.135^{AL}	$\mathbf{0.913^{AL}}$	1.312^{R}	4.205
	100	4.731^{Id}	4.541^{Id}	2.087^{Id}	$\mathbf{1.723^{AL}}$	2.349^{R}	13.815
	200	9.283^{Id}	8.915^{Id}	2.989^{Id}	$\mathbf{2.521^{AL}}$	2.948^{R}	27.469

경우, 희박한 행렬들을 고려한 집합에서 최적 수렴율(optimal rate of convergence)을 달성한다고 한다 (Cai와 Liu, 2011). 이를 고려할 때 희박하지 않은 행렬들에서 경계 추정량을 사용하면 안 된다는 점은 확실해 보인다. Figure 3.1에서 한가지 더 알 수 있는 사실은 희박성이 낮은 경우 추정량들의 편차가 커진다는 점이다. 이와 반대로 희박한 구조를 가질수록 추정된 행렬들도 안정적인 결과를 보여주는 것을 확인할 수 있다.

3.2. 혼합모형에서의 모의실험

앞서 기술하였듯이 실제 연구 대상이 되는 자료들은 서로 다른 모집단들에서 얻어진 자료들의 혼합인 경우가 많기 때문에, 혼합모형(mixture model)을 고려하여 모의실험을 진행하였다. 참고문헌들에서도 실제 데이터의 경우 여러 클래스를 가지는 마이크로어레이 데이터를 고려하였다. 실제 데이터의 경우 드문 상황들을 제외하고 예산 등의 문제로 많은 수의 표본을 확보하지 못하는 경우가 많다. 따라서 클래스별로 다른 공분산행렬을 가지는 경우에, 클래스별로 따로 공분산행렬을 추정하는 문제는 많은 수의 변수에 비해 소표본인 경우 추정량의 안정성이 보장되지 않기 때문에 직접적인 적용은 어려워 보인다. 따라서 얼마 되지 않는 표본들을 충분히 이용할 수 있는 경우인 같은 공분산 행렬을 가지는 경우를 고려하였다. 이때 평균벡터는 $-0.5, 0, 0.5$에서 임의추출하여 구성하였다.

전체적으로 세 형태의 공분산 행렬을 고려하였는데, 앞 절에서와 같이 희박성의 정도에 따른 모의실험의 결과를 바탕으로 단위행렬, 70%와 40%의 희박성을 가지는 행렬을 고려하였고 33개의 표본의 크기를 갖는 세 그룹을 고려하였다. 또한 변수의 수가 작은 경우에도 의미가 있을 것이라 판단되어 변수의 수는 $30, 100, 200$의 경우에 대하여 모의실험을 진행하였다. Table 3.1에 모의실험 결과가 요약되어 있다.

우선, 전체적으로 표본공분산행렬과 비교하였을 때 대체로 좋은 성능들을 보여준다. 모의실험에 사용된 세 그룹 모두 동일한 공분산 구조를 가지고 있기 때문일 수도 있겠지만, 변수의 수가 200 이 될 때 추정량과 참공분산행렬과의 차이는 표본공분산행렬과의 차이에 비해 70%의 희박성을 갖는 경우는 약 50%, 단위행렬에서는 90% 정도 줄어든다.

변수의 수가 표본의 크기보다 작은 경우, adaptive lasso 경계함수를 사용한 적응적 경계 추정량(ATE)이 전체적으로 가장 좋은 성능을 보여주었다. 하지만 변수의 수가 증가하면, 참 공분산행렬의 구조에 따라 가장 좋은 성능을 보여주는 추정량이 달라지는 것을 확인할 수 있다. 먼저 40% 희박성을 보이는 참공분산행렬에서는 Fisher와 Sun 추정량(FSE)이 좋은 성능을 보여주었고, 70% 희박성을 보이는 경우에는 lasso 벌점의 콜레스키 방법(MCDE)이 좋은 성능을 보여주었다. 그리고 마지막으로 단위행렬의

Table 3.2. Average Frobenius norm losses by multivariate gamma model. (Id: Identity target, D1: Diagonal type 1(same diagonal elements) target, D2: Diagonal type 2 (different diagonal elements) target; SC: SCAD thresholding, AL: Adaptive lasso thresholding; R: Ridge penalty, L: Lasso penalty)

True structure	dist'n	SSE	FSE	GTE	ATE	MCDE	SamCov
Block	MVN	70.063^{D1}	70.090^{D1}	$\mathbf{68.8848^{SC}}$	70.663^{SC}	205.732^{L}	71.227
	MVG	$\mathbf{85.033^{Id}}$	88.717^{Id}	87.9341^{SC}	88.744^{AL}	6802.218^{R}	89.109
Identity	MVN	3.943^{D2}	$\mathbf{3.939^{D2}}$	5.21443^{SC}	4.087^{AL}	3.956^{R}	18.361
	MVG	3.967^{D2}	$\mathbf{3.966^{D2}}$	5.34115^{SC}	4.121^{AL}	3630.541^{R}	18.382

경우, 경계 추정량들이 대체로 좋은 성능을 보여주었다.

3.3. 다변량 감마 분포에서의 모의실험

앞서 두 상황들에서 생성된 모의실험 자료는 정규성에 기초하여 생성되었다. 또한 다른 연구들의 모의실험 대부분은 다변량 정규분포에서 수행되었다. 따라서 다른 다변량 분포를 가정한 모의실험을 통해 소개된 추정량들의 로버스트성(robustness)을 확인할 수 있을 것이다. 본 연구에서는 다변량 감마분포를 고려하였는데, 다변량 감마 분포를 생성하기 위하여 Palmiesta와 Provasi(unpublished manuscript) 알고리즘을 이용하였다: 서로 독립이고 모수가 각각 $\theta_i > 0, \lambda_i > 0$인 일변량 감마 확률변수 $V_0, V_1, \ldots, V_p$를 이용하여, 다음과 같이 정의하였다.

$$Y_i = \frac{\lambda_i}{\lambda_0} V_0 + V_i, \quad i = 1, \ldots, p.$$

그러면 $Y_i (i = 1, \ldots, p)$는 모수가 $\theta_0 + \theta_i, \lambda_i$인 감마확률변수가 되고, 확률벡터 $\mathbf{Y} = (Y_1, \ \ldots, Y_p)'$의 공분산행렬의 성분은 다음과 같다.

$$\mathrm{Var}(Y_i) = (\theta_0 + \theta_i)\lambda_i^2,$$
$$\mathrm{Cov}(Y_i, Y_j) = \theta_0 \lambda_i \lambda_j, \quad i \neq j.$$

이 경우에 참공분산행렬이 분포의 모수로 들어가지 않고, 모든 원소들이 양의 값을 가지기 때문에 참공분산행렬에 특정한 구조를 부여하기 위하여 θ와 λ를 적절히 조절하였다. 위의 식에서 볼 수 있듯이 다변량 감마분포에서 참공분산행렬의 모든 원소는 양수이기 때문에 비대각 원소를 0에 아주 가깝게 만들어 대각행렬(Identity)에 가까운 참공분산행렬과, 서로 다른 구조를 갖는 행렬들을 블록으로 갖는 블록행렬(Block)을 고려하여 모의실험을 진행하였다. 그리고 비교를 위하여 같은 참 공분산 행렬을 가지는 다변량 정규분포에서 생성된 데이터들로 모의실험을 함께 진행하였다(변수의 개수 200). 우선 가장 크게 눈에 띄는 사항은 콜레스키 방법(MCDE)의 경우 정규분포 가정에 민감하게 반응한다는 점이다. Table 3.2의 콜레스키 방법(MCDE)의 수치들이 비정상적으로 보이는데 그 이유는 벌점함수의 파라미터를 교차타당성(cross validation) 분석으로 정해야 하는데, 정규분포에서 멀어지는 경우 최적의 파라미터를 잘 찾아주지 못하므로 이를 바탕으로 계산한 프로베니우스 거리가 비정상적으로 보이기 때문이다. 또한 참 공분산행렬로 대각행렬을 설정하더라도 정규분포가 아니면 추정치와 참 공분산행렬의 차이가 심하게 커진다는 점을 확인할 수 있다. 이러한 현상의 한 가지 가능한 설명은 콜레스키 방법(MCDE)이 콜레스키 분해에 기초하게 되는데, T와 D행렬의 많은 원소들을 추정해야 한다. 따라서 정규성 가정에서 약간만 벗어난다고 해도 추정된 원소들은 많이 달라질 수 있으며 따라서 추정치를 구하는 과정에서 T와 D의 곱이 들어가기 때문에 공분산행렬의 추정치는 참값에서 멀어진다고 생각된다.

하지만 같은 정규분포에서 유도된 Fisher와 Sun 추정량(FSE)의 경우, 대각행렬에서는 정규분포와 감마분포 상관없이 가장 좋은 결과를 보여주었다. 이는 콜레스키 방법(MCDE)에 비해서 Fisher와 Sun

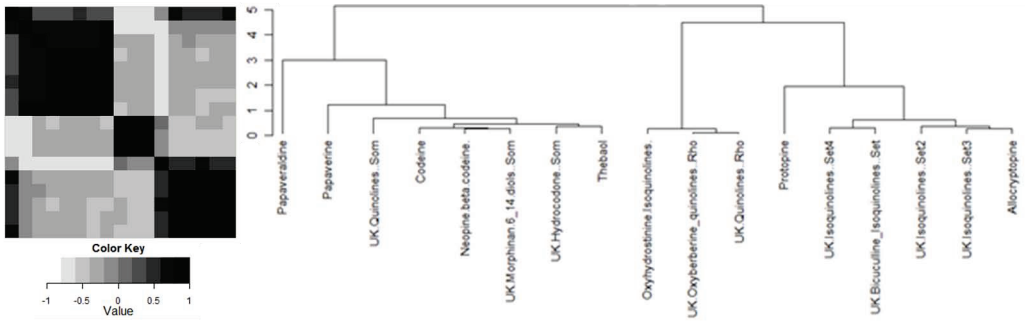

Figure 4.1. (a) Heatmap of estimated covariance matrix by SCAD thresholding(left). (b) Dendrogram of selected metabolites(right).

추정량(FSE)이 분포가정에 덜 영향받는 결과를 보여준다. 그리고 블록구조의 경우 두 분포 모두 표본 공분산행렬에 비해 개선된 성능을 보이지는 않는다는 것을 확인할 수 있다.

4. 사례연구

마이크로어레이 혹은 질량분석을 통한 실제 고차원 데이터를 분석하는 과정은 변수선택과 판별분석 등으로 많은 생물학적 표지자 후보군(biomarker candidate)을 찾고 그 중에서 생물학, 임상학적으로 실제적 유용성이 있는 최종 표지자(marker)를 추가 실험을 통해서 밝혀낸다. 하지만 이러한 과정에서 전체 변수들 또는 선택된 변수들 간의 관계에 대한 고려가 부족하다는 것은 한가지 아쉬움으로 남았다. 고차원 데이터가 가지는 근본적인 복잡성에 의하여 변수들 사이의 관계를 파악하는 것이 쉬운 문제는 아니지만, 본 논문에서 살펴본 추정 방법을 통해 변수들 사이의 관계에 대한 해답을 부분적으로라도 제공해 줄 수 있을 것이다.

Choe 등 (2011)은 국내에서 발견되고 형태학적으로 분류가 힘든 양귀비 세 종(*P. somniferum, Papaver rhoeas, P. setigerum*)에 대하여 가스 크로마토그래피/질량 분석기(GC-MS)를 이용하여 대사체 프로파일링을 이용하여 45개의 대사체를 식별(identification)한 원자료를 얻었다. 이 데이터에 Kim 등 (2011)에서 사용한 판별분석과 변수선택 방법을 이용하여 3개의 종들을 가장 잘 분류하는 17개의 생물학적 표지자 후보군을 선택하였고, 선택된 17개의 대사체를 이용하여 이들 사이의 관계를 알아보기 위해 앞서 살펴본 공분산 행렬 추정방법을 적용하였다. 이 데이터의 표본의 크기는 30으로 엄밀한 의미에서 고차원 데이터는 아니다. 하지만 대사체들을 일일이 식별하는 과정은 실험실에서 많은 시간이 걸리는 과정으로 본 데이터를 제외하고 대사체들이 식별된 경우가 없기 때문에 부득이하게 본 데이터에 앞의 추정방법들을 적용하여 보았다.

질량분석기에 시료를 이온화시켜 분석하게 되면 시료에 들어있는 대사체들의 질량이 기록된다. 대체로 질량 순서대로 대사체들이 기록되기 때문에 비슷한 질량을 나타내는 인접한 대사체들 사이에 특정한 관계가 있을 것이라 생각할 수 있지만, 비슷한 질량을 가진 대사체라고 하더라도 화학적으로 구성이 전혀 다를 수 있기 때문에 대각원소에서 멀어질수록 값이 작아지는 등의 특수한 구조를 가지는 공분산행렬을 가정하기는 어렵다. 따라서 특정한 패턴이 없는 희박한 행렬로 가정하고, 선택된 변수가 17개이므로 136개의 상관분석을 수행하여 $0.05/136 = 0.0004$의 유의수준으로 86개의 경우에 대하여 공분산행렬의 비대각원소를 0으로 결정하였다(0.632의 희박성). 또한 다변량 정규성 검정을 위한 샤피로-윌크 검정에서 유의확률 0.001미만으로 정규분포를 가정할 수 없기 때문에 콜레스키 방법(MCDE)을 제외하고 경

계 추정방법과 축소 추정방법을 적용하였다.

Figure 4.1(a)는 SCAD 경계함수로 추정된 공분산행렬이다. 이를 이용하여 Figure 4.1(b)의 변수 군집을 얻을 수 있었다. 2절에서 언급된 다른 추정방법들도 군집 내에서 대사체들의 위치에만 변화가 있을 뿐 군집으로 묶인 대사체들은 큰 차이가 없었다. 최종적으로 선택된 17개의 변수들에 대하여 크게 3개의 군집으로 나누어볼 수 있었고, 이중에서 좌측 군집은 *P. somniferum*과 다른 종을 구별할 수 있는 대사체 집단, 중앙의 군집은 *Papaver rhoeas*을, 우측의 군집은 *P. setigerum*을 구별할 수 있는 대사체 집단이다. 선택된 생물학적 표지자 후보군을 이용하여 생물학적으로 유의미하게 나누어진 군집들을 확인할 수 있었고, 이를 통해 단일 표지자(single marker)가 아닌 표지자 집단(marker group)을 찾아낼 수 있을 것으로 기대된다.

5. 결론

본 논문에서는 공분산행렬을 추정하는 문제에 있어 추정량들의 성능을 모의실험을 통하여 비교하였다. 이를 위하여 희박성의 정도, 혼합모형, 다변량 감마분포 등을 고려하여 특정한 조건 하에서 생성된 자료들을 통해 참공분산행렬에 평균적으로 더 가까운 추정량들을 비교하여 보았다. 본 논문에서 주로 고려한 변수의 수가 더 큰 경우, 표본 공분산 행렬을 사용하는 것 보다 본 논문에서 고려한 추정량들을 사용하는 것이 평균적으로 참 공분산행렬에 더 가깝다는 것을 확인할 수 있었다. 특히, 표본의 수가 변수의 수보다 더 많은 일반적인 다변량 데이터의 경우에도 참 공분산 행렬과의 거리만을 고려한다면 본 논문에서 고려한 추정량들이 표본 공분산 행렬보다 평균적으로 참 공분산행렬에 더 가까웠다.

추정량의 성능을 결정하는 요인으로 분포 가정과 희박성의 정도가 중요하다는 것을 확인할 수 있었다. 우선, 콜레스키 방법(MCDE)의 경우 서로 다른 분산을 가지는 대각행렬에 대하여 가장 좋은 성능을 보여주었지만, 분포가정에 민감하였다. 또한 콜레스키 방법(MCDE)의 경우 벌점함수를 위한 파라미터를 결정해야 하는데, 정규분포에서는 교차타당성을 이용하여 얻은 최적의 파라미터가 참 공분산 행렬과의 거리를 최소화 시켜주는 경향이 있지만, 감마분포에서는 교차타당성을 이용하여 얻은 파라미터를 이용한 경우 참 공분산행렬과의 거리를 최소화시켜주지 못한다는 단점이 있다.

다음으로 무상관인 변수들이 많아지게 되면 경계 추정방법이 가장 좋은 성능을 보여주지만, 변수들 사이에 이러한 가정을 할 수 없다면 좋지 못한 결과를 보여주게 된다. 그리고 분포에 무관하고 희박성의 정도가 크지 않다면, 즉 많은 비대각 원소들을 0이라고 할 수 없는 경우에는 축소 추정 방법(SSE, FSE)을 사용하는 것이 좋은 결과를 보여주었다. 마지막으로 동일 공분산 행렬을 가지는 가우스 혼합모형의 경우 공통 공분산행렬을 비교적 잘 추정하는 것으로 보여지며, 변수의 수가 작은 경우 경계 추정량(GTE, ATE)이 좋은 성능을 보여주었다. 하지만 희박성을 고려하였을 때 낮은 정도(40%)에서는 축소 추정량이, 중간 정도(70%)에서는 콜레스키 방법(MCDE)이, 단위행렬의 경우 경계추정량이 좋은 성능을 보여주었다.

변수의 수가 많아지는 경우 필요한 계산량도 많아지기 때문에 계산 시간 또한 추정량의 성능과 함께 고려되어야 한다. 계산시간만 고려하면 축소추정방법(SSE, FSE)이 추정량을 계산하는 데 걸리는 전체적인 시간이 가장 짧다. 적응적 경계 추정과 콜레스키 방법(MCDE) 같은 경우 단계별로 교차 타당성을 통한 최적 파라미터를 구하는 과정이 들어가기 때문에 필연적으로 많은 계산시간이 필요하게 된다. 하지만 적응적 경계추정의 경우 파라미터를 계산한 후에는 축소추정방법과 거의 비슷한 시간이 소요된다. 하지만 콜레스키 방법(MCDE)의 경우 변수의 수만큼 ridge 혹은 lasso 분석을 해야하기 때문에 많은 시간이 소요된다. 변수의 수가 1000 이상이 되면 다른 추정방법을 고려하는 것이 실용적이라고 보여진다.

실제 마이크로어레이나 질량분석 단백질 데이터를 분석하는 경우 변수들 사이의 생물학적인 상관관계를

파악하는 것이 공분산행렬을 추정하는 문제에 있어서 중요하다고 판단된다. 많은 변수들을 독립 혹은 무상관으로 간주할 수 있으면 경계추정량을 사용하는 편이 효율적이고 간주할 수 없다면 축소추정량을 사용하는 것이 좀 더 효과적이라고 보여진다. 이러한 정보는 의학연구 및 생물학 연구 등에서 고차원 데이터를 분석하는 데 도움이 될 것이다.

References

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding, *The Annals of Statistics*, **36**, 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices, *The Annals of Statistics*, **36**, 199–227.

Bouveyron, C., Girard, S. and Schmid, C. (2007). High-dimensional data clustering, *Computational Statistics & Data Analysis*, **52**, 502–519.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association*, **106**, 672–6684.

Cai, T., Zhang, C. H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation, *The Annals of Statistics*, **38**, 2118–2144.

Choe, S., Kim, S., Lee, C., Yang, W., Park, Y., Choi, H., Chung, H., Lee, D. and Hwang, B. Y. (2011). Species identification of Papaver by metabolite profiling, *Forensic Science International*.

Chun, H. and Keles, S. (2010a). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 3–25.

Chung, D. and Keles, S. (2010b). Sparse partial least squares classification for high dimensional data, *Statistical Applications in Genetics and Molecular Biology*, **9**, 17.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis, *Technometrics*, **53**, 406–413.

Fisher, T. J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix, *Computational Statistics & Data Analysis*, **55**, 1909–1918.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments, *Bioinformatics*, **18**, 275–286.

Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood, *Biometrika*, **93**, 85–98.

Kim, N., Kim, K., Choi, B. Y., Lee, D. H., Shin, Y. S., Bang, K. H., Cha, S. W., Lee, J. W., Choi, H. K., Jang, D. S. and Lee, D. (2011). Metabolomic approach for age discrimination of *Panax ginseng* using UPLC-Q-Tof MS, *Journal of Agricultural and Food Chemistry*, **59**, 10435–10441.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, **88**, 365–411.

Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty, *The Annals of Applied Statistics*, 245–263.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions, *Biometrika*, **99**, 29–42.

McLachlan, G. J., Bean, R. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, **18**, 413–422.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models, *Bioinformatics*, **26**, 2705–2712.

Palmitesta, P. and Provasi, C. (n.d.). Computer Generation of Random Vectors from Continuous Multivariate Distributions. Available from: http://www.econ-pol.unisi.it/dmq/pdf/DMQ_WP_34.pdf.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association*, **104**, 177–186.

Rothman, A. J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions, *Biometrika*, **97**, 539–550.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology*, **4**, 32.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation, *Journal of Multivariate Analysis*, **99**, 1015–1034.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences*, **99**, 6567–6572.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15**, 265–286.

고차원 데이터에서 공분산행렬의 추정에 대한 비교연구

이동혁[a] · 이재원[a,1]

[a]고려대학교 통계학과

(2013년 7월 29일 접수, 2013년 9월 30일 수정, 2013년 9월 30일 채택)

요 약

공분산 행렬은 다변량 통계분석에서 중요한 역할을 하고 있으며 전통적인 다변량 분석의 경우 표본 공분산 행렬이 참 공분산 행렬의 추정량으로 주로 사용되었다. 하지만 변수의 수가 표본의 크기보다 훨씬 큰 고차원 데이터와 같은 경우에는 표본 공분산 행렬은 비정칙행렬이 되어 기존의 다변량 기법을 사용하는 데 적절하지 않을 수가 있다. 최근 이러한 문제점을 해결하기 위해 축소추정, 경계추정, 수정 콜레스키 분해 추정 등의 새로운 공분산 행렬의 추정량들이 제안되었다. 본 논문에서는 추정량들의 성능에 영향을 미칠 수 있는 여러 현실적인 상황들을 가정하여 모의실험을 통해 참공분산 행렬의 추정량들의 성능을 비교하였다.

주요용어: 공분산 행렬 추정, 축소추정, 경계 추정, 수정 콜레스키 분해 추정.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2012R1A1A2008686).

[1]교신저자: (136-701) 서울특별시 성북구 안암동 5-1 고려대학교 통계학과, 교수. E-mail: jael@korea.ac.kr