*The Korean Journal of Applied Statistics* (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.1117
**29**(6), 1117–1128

# Network-based regularization for analysis of high-dimensional genomic data with group structure

Kipoong Kim[a] · Jiyun Choi[a] · Hokeun Sun[a,1]

[a]Department of Statistics, Pusan National University

(Received August 9, 2016; Revised October 5, 2016; Accepted October 5, 2016)

**Abstract**

In genetic association studies with high-dimensional genomic data, regularization procedures based on penalized likelihood are often applied to identify genes or genetic regions associated with diseases or traits. A network-based regularization procedure can utilize biological network information (such as genetic pathways and signaling pathways in genetic association studies) with an outstanding selection performance over other regularization procedures such as lasso and elastic-net. However, network-based regularization has a limitation because cannot be applied to high-dimension genomic data with a group structure. In this article, we propose to combine data dimension reduction techniques such as principal component analysis and a partial least square into network-based regularization for the analysis of high-dimensional genomic data with a group structure. The selection performance of the proposed method was evaluated by extensive simulation studies. The proposed method was also applied to real DNA methylation data generated from Illumina Infinium HumanMethylation27K BeadChip, where methylation beta values of around 20,000 CpG sites over 12,770 genes were compared between 123 ovarian cancer patients and 152 healthy controls. This analysis was also able to indicate a few cancer-related genes.

Keywords: high-dimensional genomic data, network-based regularization, genetic network, principal component analysis (PCA), partial least square (PLS)

## 1. 서론

고차원 유전체 자료(high-dimensional genomic data)를 사용하는 전장유전체 연관 분석(genome-wide association study)의 경우 일반적으로 변수의 수가 표본의 수보다 압도적으로 더 크기 때문에 회귀모형에 기반을 둔 모수 규제화(regularization) 방법이 많이 사용된다. 대표적인 예로 Lasso (Tibshirani, 1996)와 Elastic-net (Zou와 Hastie, 2005)을 들 수가 있다. 모수 규제화 방법들은 대부분 벌점함수(penalty function)의 특성에 맞추어 회귀계수의 값을 규제화 시키므로 벌점함수의 종류에 따라 다른 결과값을 갖게된다. Lasso의 경우는 $l_1$ 노름(norm)을 벌점함수로 사용하여 희소성(sparsity)의 특성을 갖고 있으며 대다수의 회귀계수 값을 정확하게 0으로 추정한다. Elastic-net의 경우 $l_1$ 노름과 $l_2$ 노름의 제곱을 함께 벌점함수로 사용하여 희소성과 유연성(smoothness)을 모두 갖고 있으며 서로 상관관계

This work was supported by a 2-Year Research Grant of Pusan National University.
[1]Corresponding author: Department of Statistics, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: hsun@pusan.ac.kr

가 높은 변수들이 함께 0으로 추정되도록 유도한다. 때문에 변수 간의 상관관계가 높을 경우 일반적으로 Lasso 보다는 Elastic-net이 변수 선택 면에서 더 우월하다.

고차원 유전체 자료의 경우 유전자들은 일정한 유전체 네트워크를 구성하여 서로서로 연결되어 있으며 이러한 유전체 네트워크 정보는 biocarta (www.genecarta.com), humancyc (humancyc.org), KEGG (www.genome.jp/kegg), nci (www.cancer.gov), panther (pantherdb.org), reactome (www.reactome.org) 등과 같은 곳에서 쉽게 찾아볼 수 있다. 한 유전체 네트워크 안에서 서로서로 연결된 유전자들은 일반적으로 유사한 연관 패턴을 가지고 있기 때문에 유전체 네트워크 정보가 연관 분석에 큰 영향을 줄 수 있다는 연구 결과가 있다 (Li와 Li, 2008; Chen 등, 2011). 하지만 Lasso와 Elstic-net의 경우 유전체 네트워크 정보를 활용할 수 없다는 단점을 가지고 있다. 이러한 이유로 네트워크 정보를 연관 분석에 활용할 수 있는 네트워크 기반의 규제화(network-based regularization) 방법이 Li와 Li (2010)에 의해서 처음으로 제안되었다. 그 이후에도 기존의 Lasso와 Elastic-net 보다 변수 선택 면에서 네트워크 기반의 규제화 방법이 훨씬 더 정확하다고 여러 연구들을 통해 입증되었으며 실제 고차원 유전체 자료를 분석하는 데에도 꾸준히 활용되고 있다 (Sun과 Wang, 2012, 2013; Sun 등, 2014).

네트워크 기반의 규제화 분석 방법은 Elastic-net의 벌점함수에 라플라스 행렬(Laplacian matrix)을 삽입하여 서로서로 연결된 유전자(변수)들끼리 유사한 회귀계수를 갖도록 유도하는 방식을 사용하고 있다. 그러므로 회귀모형에서 각 독립변수가 각각의 유전자를 나타낼 경우 네트워크 기반의 규제화 방법을 적용시켜 질병 및 표현형질에 영향을 주는 유전자를 찾아낼 수가 있다. 예를 들면, 유전자 발현량 측정 데이터(microarray gene expression data)의 경우 각 유전자의 발현량을 하나의 독립변수로 회귀모형에서 사용할 수 있으므로 네트워크 기반의 규제화 방법을 쉽게 적용시킬 수가 있다. 그렇지만 대다수의 고차원 유전체 자료는 유전자라는 하나의 그룹 안에 여러 개의 유전체 정보를 포함하는 형태로 관측된다. 단일 염기 다형성(single nucleotide polymorphism; SNP) 데이터, DNA 메틸화(methylation) 데이터 및 차세대 염기 서열(next generation sequencing) 데이터의 경우 하나의 유전자(gene) 안에 다수의 유전체 변이(genetic variants/sites)들의 정보가 있다. 이렇게 그룹으로 이루어진 고차원 유전체 자료는 하나의 독립변수가 유전자 그룹이 아니라 그룹 안에 속해 있는 하나의 유전체 변이를 나타내므로 네트워크 기반의 규제화 분석 방법을 곧바로 적용시킬 수가 없다.

본 논문에서는 그룹 구조를 가지고 있는 고차원 유전체 자료를 분석하고자 네트워크 기반의 규제화 방법에 차원 축소 방법을 결합시키는 새로운 분석 방법을 제안하고자 한다. 차원 축소 방법으로는 일반적으로 통계 분석에 많이 사용하는 주성분 분석(principal component analysis; PCA)과 부분 최소 자승법(partial last square; PLS)을 사용하였다. 2장에서는 네트워크 기반의 규제화 방법에 차원 축소 방법을 결합시킨 새로운 분석 방법과 변수 선택 방법에 대해서 자세하게 소개하고 3장에서는 새롭게 제안한 분석 방법의 변수 선택의 우수성을 모의실험(simulation)을 통해서 입증하였다. 그리고 4장에서는 실제 그룹으로 이루어진 고차원 유전체 자료로서 난소암 환자들의 DNA 메틸화 자료를 가지고 새롭게 제안한 분석 방법을 통해 난소암에 영향을 주는 유전자들을 찾아보았다. 끝으로 5장에서는 본 연구에 대한 결론 및 한계점에 대해서 논의하였다.

## 2. 유전체 네트워크를 활용한 변수 선택 방법

### 2.1. 네트워크 기반 규제화(network-based regularization)의 문제점

네트워크 기반의 벌점 우도함수(network-based penalized likelihood)는 다음과 같이 정의되어 진다.

$$Q(\beta) = -l(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^{\mathrm{T}} L \beta,$$

$$= -l(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{u \sim v} \left( \frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2, \tag{2.1}$$

여기서 $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$는 $p$차원 벡터로서 추정해야 할 모수이고, $l(\cdot)$는 로그 우도함수이다. $\lambda_1$과 $\lambda_2$는 조율 모수(tuning parameter)로서 각각 $\beta$의 희소성과 유연성을 조절하는 역할을 한다. $u \sim v$는 $u$번째 유전자에 연결된 모든 유전자들의 집합 $v \in \{1, 2, \ldots, d_u\}$을 나타내고 $d_u$는 $u$번째 유전자에 연결된 총 유전자의 개수이다. 그리고 $L = \{l_{uv}\}_{p \times p}$은 $p$차원의 라플라스 행렬로서 네트워크 그래프를 상징한다.

$$l_{uv} = \begin{cases} 1, & u = v \text{ 그리고 } d_u \neq 0 \text{일 경우,} \\ -\dfrac{1}{\sqrt{d_u d_v}}, & u\text{번째 유전자와 } v\text{번째 유전자가 서로 연결되어 있을 경우,} \\ 0, & \text{그 외의 경우.} \end{cases}$$

어떤 유전자가 서로서로 연결되어 있는지 유전체 네트워크 정보를 알게 되면 위의 공식을 사용하여 손쉽게 라플라스 행렬을 구할 수 있다. 실험군(case)과 대조군(control)으로 이루어진 이진형 반응변수의 경우 로그 우도함수는

$$l(\beta_0, \beta) = \sum_{i=1}^{n} \left[ y_i \log p_i(\beta_0, \beta) + (1 - y_i) \log(1 - p_i(\beta_0, \beta)) \right],$$

여기서 $p_i(\beta_0, \beta)$는 $i$번째 표본이 실험군일 확률로서 다음과 같다.

$$p_i(\beta_0, \beta) = \frac{e^{\beta_0 + x_i^{\mathrm{T}} \beta}}{1 + e^{\beta_0 + x_i^{\mathrm{T}} \beta}}.$$

총 표본의 수를 $n$이라 가정하면 $i = 1, \ldots, n$이며 관측된 자료는 $i$번째 표본의 반응변수 $y_i$와 $i$번째 표본의 유전체 관측값인 $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$로 나타내며 실험군의 경우 $y_i = 1$이고 대조군일 경우 $y_i = 0$으로 정의한다. $\beta_0$는 $y$절편으로 추정해야 하는 모수지만 벌점함수에는 포함되지 않는다.

네트워크 기반의 규제화 분석 방법의 가장 큰 한계점은 유전체 관측값 $x_i$가 반드시 유전자 발현량 측정 데이터처럼 유전자 단위로 측정되는 자료이어야 한다는 것이다. 한 개의 표본에서 관측된 유전체 데이터가 $p$차원 벡터라면 이에 대응하는 회귀계수(regression coefficient)도 $p$차원 벡터로 구성되어야만 한다. 그렇게 되면 자연스럽게 라플라스 행렬또한 $p \times p$ 차원의 행렬로 이루어지게 된다. 결국 $p$개의 유전자가 존재할 때 어떤 유전자들끼리 서로서로 연결되어 있는지 나타내는 유전체 네트워크의 유전자 개수도 $p$개로 일치해야지만 네트워크 기반의 벌점 우도함수에 적용시킬 수 있다.

그렇지만 SNP 데이터, DNA 메틸화 데이터 및 차세대 염기 서열 데이터의 경우 각 유전자마다 여러 개의 유전체 변이를 포함하고 있기 때문에 네트워크 기반의 규제화 분석 방법을 바로 적용 시킬 수가 없다. 예를 들면, DNA 메틸화 데이터의 경우 각 유전자마다 한 개에서부터 수십 개의 CpG sites가 있으며 각 site에서 메틸화 수준(methylation level)을 측정하여 분석하게 된다. 즉 기본변수는 유전자가 아니라 CpG site가 된다. 한 개의 표본에서 관측된 DNA 메틸화 데이터가 $q$개의 CpG sites를 가지고 있다면 관측값은 $q$차원 벡터가 되고 이에 대응하는 회귀계수또한 $q$차원 벡터로 구성되어야 하지만 $p$개의 유전자에 대한 유전체 네트워크를 나타내는 라플라스 행렬은 여전히 $p \times p$차원의 행렬이며 항상 $p < q$이므로 DNA 메틸화 데이터의 경우 기존의 네트워크 기반의 규제화 분석 방법에 적용 시킬 수 없다.

### 2.2. 데이터 차원 축소 방법과의 결합

고차원 자료를 저차원 자료로 변형시키는 데이터 차원 축소 방법(data dimension reduction method)은 최근 고차원 자료를 분석하는 데 빈번하게 활용되고 있다. 유전체 데이터의 경우에도 하나의 유전자가

여러 개의 유전체 변이로 구성되어 있을 때 모든 유전체 변이의 정보를 하나의 차원으로 축소시켜 유전자 단위의 정보로 요약시킬 수 있다. 데이터 차원 축소 방법 중 대표적인 예로 PCA와 PLS를 들 수 있다 (Faraway, 2014). 두 방법 모두 기존의 독립변수들의 선형 결합(linear combination)으로 이루어진 새로운 성분(component)을 찾는 방법이지만 이러한 성분을 찾아내는 방식에서 두 방법은 차이점을 보인다. PCA의 경우 반응변수와의 연관성을 배제하여 관측된 독립변수들의 변동(variability)의 크기를 기준으로 주성분을 만들어 내지만, PLS의 경우 반응변수와의 연관성을 고려하여 주성분을 생성해낸다. 두 방법 모두 고차원 자료를 저차원 자료로 축소시켜 분석의 효율성을 높일 수 있다는 장점을 가지고 있지만 첫 번째 성분이 여러 개의 독립변수를 충분하게 설명하지 못할 경우 대표성이 떨어지게 되므로 유전체 연관 분석 결과에도 큰 영향을 미칠 수가 있다.

그룹 구조를 가지고 있는 고차원 유전체 자료의 경우 PCA 또는 PLS를 사용하여 $q$차원 그룹 자료를 각 유전자 단위별로 축소시켜서 최종적으로 $p$차원 자료로 변형시킨 후 유전체 네트워크 기반의 규제화 방법을 적용시킬 수가 있다. 예를 들면 4장에서 적용한 DNA 메틸화 데이터의 경우 총 20,461개의 CpG sites를 가지고 있는 고차원 자료이다. 이 자료는 총 12,770개의 유전자 그룹 구조로 이루어져 있으며 각 유전자의 경우 적게는 1개의 site, 많게는 22개의 sites를 포함하고 있다. 데이터 차원 축소 방법을 사용하여 20,461개의 독립변수를 갖고 있는 기존 데이터를 12,770개의 변수를 갖고 있는 새로운 데이터로 변형시킨 후 유전체 네트워크 기반의 규제화 방법을 사용할 수 있었다. PCA는 R 패키지 'prcomp'를 사용하였고 PLS는 R 패키지 'plsRglm'을 사용하였다.

### 2.3. 선택 확률(selection probability)을 이용한 결과 요약

네트워크 기반의 규제화 방법을 포함한 대다수의 모수 규제화 방법은 조율 모수의 값에 따라 최종 선택된 변수들이 결정된다. 조율 모수 값은 실질적으로 선택된 변수의 최종 개수에 영향을 주므로 조율 모수 값을 정확하게 추정하는 것이 매우 중요하다. 조율 모수 값을 결정하기 위해 가장 많이 사용하는 방법 중 하나로 교차검증(cross validation; CV)을 들 수 있다. 그러나, 교차검증은 대개 표본을 일정한 그룹의 수로 무작위로 분리하여 적용시키는 방법이기 때문에 어떻게 표본을 나누냐에 따라 그 결과값이 크게 달라질 수 있다는 단점을 가지고 있다. 이러한 문제를 해결하고자 Meinshausen과 Bühlmann (2010)은 규제화 방법에 선택 확률(selection probability; SP)을 사용할 것을 제안하였다. 고차원 유전체 데이터의 분석 결과를 선택 확률을 사용하여 요약하는 방식은 여러 문헌에서도 찾아볼 수 있다 (Alexander와 Lange, 2011; Sun과 Wang, 2012, 2013).

변수의 선택 확률을 구하는 방법은 반복적인 표본 추출을 통해 생성된 각각의 표본들에 네트워크 기반의 규제화 분석 방법을 적용시켜 변수들의 선택된 횟수의 평균을 변수별로 계산하는 방식이다. $I_k$를 크기가 $\lfloor n/2 \rfloor$인 $k$번째로 추출된 표본들의 색인(index) 집합이라고 표시하자. 조율 모수를 $\lambda = \lambda_1 + 2\lambda_2$와 $\alpha = \lambda_1/(\lambda_1 + 2\lambda_2)$로 정의한다면 $\lambda > 0$는 선택된 변수들의 개수를 조절하는 조율 모수이고 $\alpha \in [0, 1]$는 라플라스 벌점에 대한 $l_1$ 노름의 비율을 나타내는 조율 모수가 된다. $(\lambda, \alpha)$가 일정한 값들로 고정되어 있을 경우, $j$번째 변수의 선택 확률은 다음과 같다.

$$\text{SP}_j = \max_{\lambda,\alpha} \frac{1}{K} \# \left\{ k \leq K : \hat{\beta}_j^{\lambda,\alpha}(I_k) \neq 0 \right\},$$

여기서 $K$는 표본 추출의 총 반복 횟수를 나타내며 모의실험과 실제 자료 분석에서는 $K = 100$으로 설정하였다. $\hat{\beta}_j^{\lambda,\alpha}(I_k)$는 네트워크 기반의 규제화 방법에서 추정된 회귀계수로서 표본이 $i \in I_k$이고 조율 모수 $\lambda$와 $\alpha$가 고정되어 있을 경우 벌점 우도함수 (2.1)를 최소화시키는 $j$번째 회귀계수의 값을 나타낸다. 벌점 우도함수의 최소화는 볼록함수 최적화(convex optimization) 문제로서 여러 가지 알고리즘을

사용할 수 있지만, $l_1$ 노름 벌점함수에 많이 사용되는 cyclic coordinate descent 알고리즘 (Friedman 등, 2010; Simon 등, 2011)을 사용하면 빠르게 회귀계수를 구할 수 있다.

고차원 유전체 자료의 경우 각 유전자별로 선택 확률을 계산한 후에 가장 높은 선택 확률을 가진 유전자부터 가장 낮은 선택 확률을 가진 유전자 순으로 정렬시켜서 질병 또는 표현형질에 영향을 주는 유전자들의 순위를 매길 수 있다. 고차원 유전체 자료의 일변량 분석(univariate analysis)에서 개개의 유전자별로 반응변수와의 연관성을 검정하여 가장 낮은 유의확률($p$-value)을 갖는 유전자부터 높은 유의확률을 갖는 유전자 순으로 정렬시켜서 반응변수에 가장 큰 영향을 주는 일정한 수의 유전자들을 찾아 내는 것과 같은 원리라 할 수 있다. 다만 유의확률과 다른점은 선택 확률은 주어진 조율 모수 $\lambda$값의 범위에 따라 선택 확률의 크기가 변하기 때문에 선택 확률 그 자체의 값보다는 선택 확률에 의해서 매겨진 순위에 의미를 두어야 한다.

## 3. 모의실험

이번 장에서는 그룹 구조를 가지고 있는 유전체 데이터를 대상으로 네트워크 기반의 규제화 방법과 네트워크 정보를 활용하지 않는 Elastic-net의 성능을 비교하기 위하여 독립변수가 DNA 메틸화 데이터와 같은 연속형 자료인 경우와 SNP 데이터와 같은 범주형 자료인 경우에 대해서 두 가지의 모의실험을 시행하였다. 모의실험을 시행하기에 앞서 모형의 성능을 비교하기 위한 척도로 true positive rate(TPR)을 사용하였는데, TPR은 민감도(sensitivity)와 동일하며 다음과 같이 정의한다.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

여기서 true postive(TP)는 실제 유의한 변수 중 선택된 변수의 수를 의미하고, false negative(FN)는 실제 유의한 변수 중 선택되지 않은 변수의 수를 의미한다. 따라서 TPR은 실제 유의한 변수들 중에서 올바르게 선택된 변수들의 비율을 의미하므로, 유의한 변수를 얼마나 잘 선택하는지 그 정확도와도 의미가 상통한다. 따라서 두 가지의 다른 방법에 대해 똑같은 개수의 변수를 선택하였을 때 TPR이 높은 방법이 변수 선택 면에서 더 우월하다고 판단할 수 있다.

유전체 네트워크 데이터는 Sun과 Wang (2013)의 방식을 사용하여 유사하게 생성하였고 총 $n = 200$개의 표본(100개의 실험군과 100개의 대조군)과 $p = 1{,}000$개의 유전자를 구성하였다. 총 1,000개의 유전자는 100개씩 동일한 하나의 유전체 네트워크로 연결되어 있다고 가정하였으며 또한 이러한 100개의 유전체는 Figure 3.1과 같은 네트워크를 형성하고 있다고 가정하였다. 즉 10개의 분리된 유전체 네트워크 그룹으로 총 1,000개의 유전자가 구성되어 있다. 여기서 우리는 10개의 유전체 네트워크 그룹 중에서 오직 하나의 네트워크 그룹만이 실험군과 대조군의 차이를 만든다고 가정하였고, 그 하나의 네트워크 안에 있는 100개의 유전자 중에서 Figure 3.1에서 색칠된 유전자 45개만이 유의한(causal) 유전자라고 가정하였다. 중앙에 위치한 1개의 유전자와 11개의 유전자가 하나의 묶음이라고 가정한다면 색칠된 유전자는 총 4개의 묶음으로 구성되었으며, 이 4개의 유전자 묶음을 각각 $g_1$, $g_2$, $g_3$, $g_4$라고 정의한다. 유의한 유전자들의 반응변수에 대한 영향력 크기를 다른 유전자와의 연결(link)된 수에 비례하도록 설정하기 위해서 평균 벡터 $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$의 $u \in \{1, 2, \ldots, 45\}$번째 평균값을 다음과 같이 정의하였다.

$$\mu_u = \begin{cases} \dfrac{\delta\sqrt{d_u}}{3}, & \text{if } u \in g_1 \text{ or } g_2, \\ -\dfrac{\delta\sqrt{d_u}}{3}, & \text{if } u \in g_3 \text{ or } g_4, \\ \delta, & \text{if centered gene.} \end{cases}$$

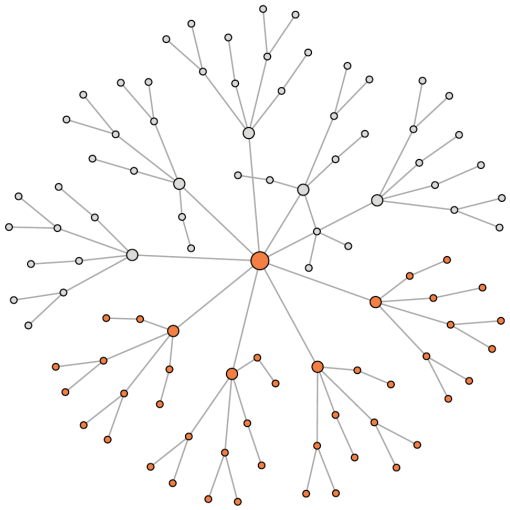

**Figure 3.1.** An example of a network graph with 100 genes used in simulation. The colored 45 genes are assumed to be causal genes and consist of one centered gene plus four different groups of genes, where each group has 11 genes.

$u > 45$의 경우 $\mu_u = 0$이다. $d_u$는 $u$번째 유전자에 연결된 다른 유전자들의 개수로서 Figure 3.1에서 중앙의 유전자는 9개이고 그 외 다른 유전자들은 1, 2, 3 또는 5개를 가지고 있다. $\delta$는 전반적인 반응변수에 대한 영향력을 결정하는 상수로 1.5로 고정시켰다. 이와 같이 주어진 유전체 네트워크에 맞추어 Gaussian graphical 모형 (Whittaker, 1990)의 성질을 이용하여 공분산 행렬 $\Sigma$를 만들었다. 공분산 행렬을 생성하는 자세한 과정은 Peng 등 (2009)을 참고하기 바란다.

첫 번째 모의실험의 경우 연속형인 독립변수를 대상으로 하였다. 위에서 구한 평균 벡터 $\mu$와 공분산 행렬 $\Sigma$가 주어지게 되면 다변량 정규분포를 이용하여 독립변수 $X$ 를 실험군($Y = 1$)과 대조군($Y = 0$)으로 나누어 다음과 같이 생성하였다.

$$X|Y = 1 \sim N(\mu, \Sigma) \quad \text{그리고} \quad X|Y = 0 \sim N(0, \Sigma).$$

마지막으로 그룹 구조를 가지고 있는 유전체 데이터를 생성하기 위해서 이미 생성된 $200 \times 1{,}000$차원의 $X$행렬을 중심으로 각 변수마다 상관관계가 있는 5개의 변수를 추가적으로 정규분포에서 생성해내었다. 즉 $j$번째 변수의 200개의 자료를 $X_j$라고 할 경우 $N(\rho X_j, \sqrt{1-\rho^2})$로부터 5개의 벡터값을 무작위로 추출하였다. 상관계수 $\rho$는 0.3, 0.5, 0.7로 다양하게 설정하여 상관관계의 크기에 따른 TPR의 변화를 확인해 보고자 하였다. 최종적으로 독립변수는 5,000개를 가지게 되며 각각의 5개 변수가 하나의 그룹을 형성하여 총 1,000개의 그룹을 갖게 된다. 실제 모의실험에서는 $200 \times 1{,}000$차원의 $X$행렬은 사용하지 않았으며, $X$행렬로 생성된 $200 \times 5{,}000$차원의 행렬을 1,000개의 그룹을 갖고 있는 고차원 유전체 데이터라고 가정하고 분석에 사용하였다.

Figure 3.1에서 주어진 네트워크 정보를 정확하게 알고 있을 경우와 네트워크 정보를 모를 경우의 두 가지 방법을 비교했으며 전자는 식 (2.1)에서 네트워크 정보를 사용한 라플라스 행렬 $L$을 직접 계산해서 사용한 네트워크 기반의 규제화 방법(Network)이고 후자는 식 (2.1)에서 라플라스 행렬 $L$ 대신 항등 행렬을 사용하는 Elastic-net 방법이다. 두 방법을 사용하기 앞서 그룹 구조의 자료를 축소시키기 위해서 각 그룹마다 5개의 변수들을 PCA 또는 PLS를 사용하여 1,000개의 성분으로 변형시킨 후 Network 방

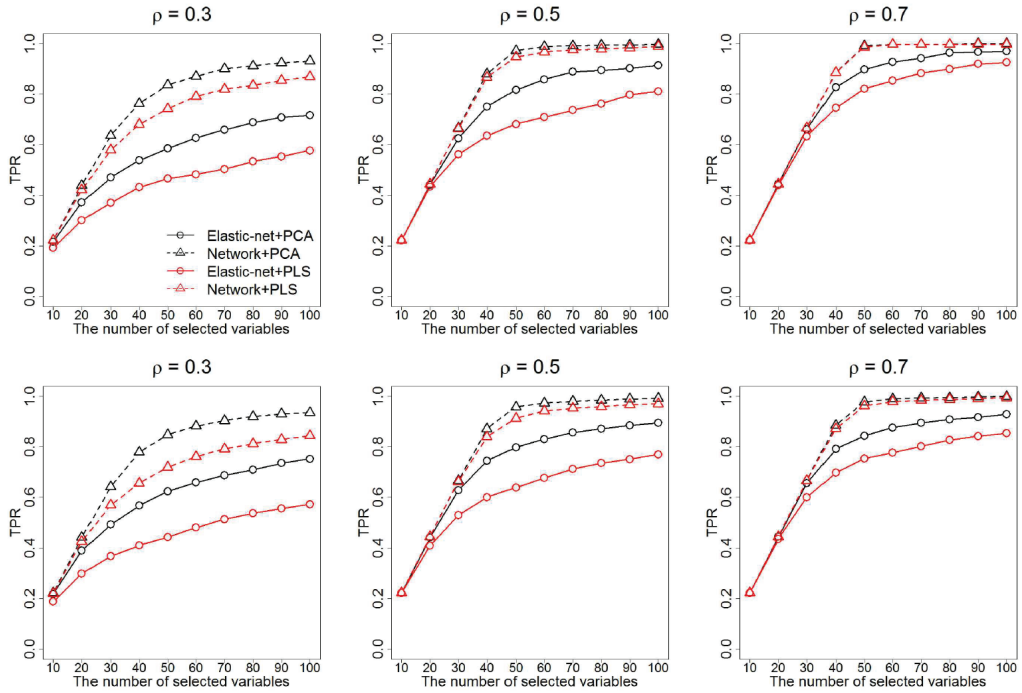


**Figure 3.2.** Averaged true positive rates (TPR) of variables selected by Elastic-net and Network-based regularization combined with PCA (principal component analysis) or PLS (partial least squares) are displayed along with different number of selected variables based on selection probability and different correlation coefficient $\rho$. In the upper three panels simulated genomic data is continuous while in the bottom three panels simulated genomic data is categorical.

법과 Elastic-net 방법을 각각 적용시켰다. 각 방법들 마다 선택 확률이 높은 변수들을 100개까지 나열하여 선택된 변수의 개수에 따른 TPR을 비교하였다. 모의실험은 총 100번을 반복하였으며 100번에 대한 TPR의 평균치를 Figure 3.2의 상위 3개의 그림에 나타내었다.

Figure 3.2의 결과를 보면, $\rho = 0.3$일 때 PCA 또는 PLS의 어떤 방법과 결합했든 간에 유전체 네트워크 정보를 사용한 Network 방법이 그렇지 않은 Elastic-net 방법 보다 대략 22% 정도 더 높은 TPR을 보여주고 있다. 상관계수 $\rho$가 높아질수록 그 차이는 줄어들지만 여전히 Network 방법이 Elastic-net 방법보다 변수 선택 면에서 우월함을 보여준다. PCA와 PLS의 비교에서는 PCA가 더 높은 성능을 보여주고 있는데, 마찬가지로 상관계수의 크기가 커짐에 따라 그 차이는 점점 감소하는 모습을 보이고 있다.

두 번째 모의실험은 독립변수가 SNP 데이터와 같은 범주형 자료라고 가정하여 실행하였다. 실제 SNP 데이터의 경우 데이터는 유전자형(genotype)으로 열성형질(minor allele)의 수 0, 1, 2 중에서 한 가지로 표현되는 범주형 자료이며 또한 여러 개의 SNP은 하나의 유전자에 속해 있기 때문에 그룹 구조를 갖고 있는 자료이다. 첫 번째 모의실험과 같은 방식으로 $X$를 생성한 후, 각 변수마다 상관관계가 있는 10개의 변수를 추가로 정규분포에서 생성해내었다. 열성형질빈도(minor allele frequency; MAF)를 10%부터 50% 사이라고 가정하고 균일분포에서 무작위로 생성해낸 후 정규분포에서 추가로 생성한 10개의 변수를 각 변수별로 백분위수를 계산하여 MAF보다 작으면 1, 크면 0으로 변환시켰다. 실제 2개의 염색체가 만나서 하나의 유전자형을 형성하듯이 0과 1로 변환된 10개의 변수를 2개씩 짝을

맞추어 그 값을 더해서 0, 1, 2 중 한 가지로 표현되도록 5개 변수로 만들어 최종적으로 $X$에서 생성된 하나의 변수가 각각 5개의 추가적인 변수를 가지도록 생성하였다. 첫 번째 모의실험과 마찬가지로 최종적으로 1,000개의 그룹으로 이루어진 $200 \times 5{,}000$차원의 범주형 자료를 가지고 같은 방법으로 TPR을 계산하였으며 그 결과는 Figure 3.2의 하위 3개의 그림에 나타내었다. 그림에서 보여주듯이 연속형 자료일 경우와 거의 유사한 결과가 나왔다. 즉 유전체 네트워크 정보를 사용할 경우 변수 선택의 정확성은 연속형 유전체 데이터와 범주형 유전체 데이터 모두에서 매우 높아진다는 것을 알 수 있었으며 기존의 그룹 유전체 데이터에 적용시킬 수 없었던 네트워크 기반의 분석 방법이 PCA 또는 PLS와 결합하여 적용시켜도 변수 선택의 성능은 여전히 뛰어나다는 것을 알 수 있었다.

## 4. 난소암 DNA 메틸화 자료 분석

이번 장에서는 난소암 환자들의 Illumina Infinium Human Methylation27 BeadChip으로 생성한 DNA 메틸화 자료 (Teschendorff 등, 2010)에 본 논문에서 제안한 방법을 적용시켜 보았다. 이 자료는 NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/)에서 이용 가능하다. 본 논문의 방법을 적용시키기 위해선 모든 CpG sites가 하나의 유전자에 속한다는 가정이 필요하기 때문에, 속한 유전자 정보가 없는 CpG sites와 결측인 $\beta$-value를 가지는 CpG sites를 제거함으로써 152명의 정상인들과 123명의 난소암 환자들의 12,770개의 유전자에 있는 20,461개의 CpG sites를 최종적으로 분석에 사용하였다. 12,770개의 유전자들 중에서 5,834개는 1개, 6,744개는 2개, 169개는 3∼9개, 23개는 10∼22개의 CpG sites를 가지고 있었다. 분석에 앞서 이 자료의 표현 형태인 0∼1의 범위를 가지는 $\beta$-value 분석법은 직관적인 생물학적 해석은 용이하지만 DNA 메틸화 수준(methylation levels)을 분석하기엔 치명적인 단점인 이분산성(heteroscedasticity)이 나타나므로, 좀 더 통계 분석을 실행하기에 알맞은 메틸화 $M$-value 분석법 (Du 등, 2010)을 사용하였다.

현재 Bioconductor의 R 패키지 'graphite'에서 제공하는 유전체 네트워크 정보를 이용할 경우 12,770개의 유전자 중 5,995개의 유전자만이 유전체 네트워크 정보에 포함되는 것을 확인하였다. 반면에 6,775개의 유전자는 현재 알려진 유전체 네트워크 정보가 없기 때문에 첫 번째 분석에서는 6,775개의 유전자를 네트워크에 연결되어 있지 않은 고립된 유전자(isolated gene)라고 가정하고 12,770개의 모든 유전자를 분석에 포함시켰으며, 두 번째 분석에서는 6,775개의 유전자를 제외시켜서 네트워크 정보를 가지고 있는 5,995개의 유전자에 대해서만 분석하였다. 앞의 모의실험에서와 동일하게 한 유전자 안에 있는 여러 개의 CpG sites를 PCA와 PLS 두 가지 방법을 이용해 하나의 차원으로 축소시킨 후 네트워크 기반의 규제화 방법을 적용시켜 가장 많이 선택된 20개의 유전자들을 선택 확률과 함께 Table 4.1에 제시하였다.

Table 4.2에서는 12,770개의 모든 유전자를 포함한 분석과 일부 5,995개의 유전자만을 포함한 분석에서 PCA와 PLS 두 가지 방법을 통해 선택된 상위 20개의 유전자들에 대한 각 분석 및 방법별 선택 여부를 확인할 수가 있다. 총 46개의 유전자가 각 분석 및 방법의 상위 20개 유전자 리스트에 포함되어 있었으며 이들 중 6개의 유전자 EGF, CFI, RHOT2, TCAP, PPBP, ADCY7는 두 가지 분석(12,770개 유전자 분석과 5,995개 유전자 분석)과 두 가지 방법(PCA와 PLS) 모두에 포함되었다. 이 중에서 유전자 EGF는 두경부암(head and neck cancer)에 영향을 미친다고 보고되었고, TCAP는 대장암, 난소암, 자궁내막암, 폐암, 요로상피암, 췌장암 등 여러 암에 영향을 미치는 것으로 조사되었다 (Marsit 등, 2009). 그 밖에도 12,770개의 유전자 분석과 5,995개의 일부 유전자를 분석한 결과를 비교해 보면 총 46개의 유전자 중 12개의 유전자가 양쪽 분석 모두에서 상위 20개의 리스트에 포함되었고, 16개의 유전자는 12,770개의 유전자 분석에는 포함되었으나 5,995개의 분석에는 유전체 네트워크 정보가 없어서 분석에서 누락되었다. 46개의 유전자 중 17개의 유전자는 5,995개의 분석에서는 상대적으로 높은 선택 확

**Table 4.1.** For analysis of ovarian cancer data, we listed top 20 genes selected by network-based regularization procedure combined with principal component analysis (PCA) and partial least square (PLS) are listed with selection probability for all 12,770 genes in the left column and only for selected 5,995 genes in the right column.

| Ranking | 12,770 genes | | | | 5,995 genes | | | |
|---|---|---|---|---|---|---|---|---|
| | PCA | | PLS | | PCA | | PLS | |
| 1 | EGF | (1.00) | LIME1 | (1.00) | TCAP | (1.00) | ADCY7 | (1.00) |
| 2 | LIME1 | (1.00) | TCAP | (1.00) | RHOT2 | (0.99) | TCAP | (1.00) |
| 3 | CFI | (0.96) | EGF | (0.98) | CFI | (0.98) | EGF | (0.99) |
| 4 | GAS2 | (0.96) | ADCY7 | (0.98) | EGF | (0.97) | TSG101 | (0.97) |
| 5 | RHOT2 | (0.95) | GAS2 | (0.96) | PPBP | (0.96) | RHOT2 | (0.96) |
| 6 | TCAP | (0.94) | RHOT2 | (0.93) | ADCY7 | (0.94) | CFI | (0.93) |
| 7 | PPBP | (0.93) | PPBP | (0.89) | HIST1H4K | (0.91) | PPBP | (0.93) |
| 8 | TNFAIP8 | (0.91) | TSG101 | (0.89) | CX3CL1 | (0.87) | SLC22A17 | (0.91) |
| 9 | SPATA12 | (0.90) | TNFAIP8 | (0.86) | SLC22A17 | (0.87) | HIST1H4K | (0.90) |
| 10 | EBI2 | (0.89) | CD300LF | (0.86) | MPO | (0.86) | NSD1 | (0.90) |
| 11 | COX7A1 | (0.88) | CFI | (0.86) | ZNF555 | (0.86) | WT1 | (0.90) |
| 12 | HIST1H4L | (0.87) | NSD1 | (0.83) | DST | (0.85) | DST | (0.89) |
| 13 | ADCY7 | (0.86) | EGFL6 | (0.83) | WT1 | (0.84) | PTPN7 | (0.89) |
| 14 | NR2F1 | (0.84) | HTR2A | (0.81) | HTR2A | (0.83) | CD81 | (0.86) |
| 15 | BRRN1 | (0.83) | COX7A1 | (0.81) | PAK7 | (0.82) | ACTN2 | (0.85) |
| 16 | CEL | (0.83) | PTPN7 | (0.81) | APBA2 | (0.81) | COX7A2L | (0.85) |
| 17 | HNRPA0 | (0.83) | C10orf27 | (0.80) | AQP8 | (0.81) | HTR2A | (0.83) |
| 18 | Bles03 | (0.82) | SLC22A17 | (0.78) | CEL | (0.81) | MAP2K7 | (0.83) |
| 19 | LIM2 | (0.82) | RBMS2 | (0.77) | MYF6 | (0.81) | PTPRO | (0.83) |
| 20 | FOXD4L4 | (0.81) | LIM2 | (0.77) | TRAF1 | (0.81) | PDCD1LG2 | (0.82) |

률을 가지고 있어서 상위 20개의 리스트에 포함되었으나 12,770개의 분석에서는 상대적으로 선택 확률이 높지 않아 포함되지 않았다. PCA와 PLS 방법을 비교하였을 경우, 12,770개의 유전자를 포함한 분석과 5,995개 유전자만 포함한 분석 모두에서 29개의 유전자가 상위 20개의 리스트에 포함되었고, 이 중 11개의 유전자는 PCA, PLS 두 가지 방법 모두에서 리스트에 있었으며 나머지 18개의 유전자는 둘 중 하나의 방법에서만 상위 20개 리스트에 발견되었다. 이번 분석에서 발견된 몇 가지 유전자들은 난소암뿐만 아니라 다른 종류의 암에도 영향을 미치는 것으로 알려져 있기에 이러한 유전자들을 추가로 더 면밀히 조사해 볼 필요성이 있다.

## 5. 결론

본 논문에서는 그룹 구조를 가지고 있는 SNP 데이터 및 DNA 메틸화 데이터 같은 고차원 유전체 자료를 분석할 수 있는 네트워크 기반의 규제화 방법을 소개하였다. 네트워크 기반의 규제화 방법은 이미 알려진 유전체 네트워크 정보를 활용하여 질병 및 표현형질에 영향을 주는 유전자를 보다 더 정확하게 찾아낼 수 있다는 큰 장점을 가지고 있다. 그렇지만, 기존의 네트워크 기반의 규제화 방법은 유전자 발현량 측정 데이터와 같이 각 변수가 각각의 유전자를 나타낼 경우에만 적용 가능하다는 한계점을 가지고 있었다. 그룹 구조를 가지고 있는 고차원 유전체 자료 분석에도 네트워크 기반의 규제화 방법을 적용시키고자 데이터 차원 축소 방법인 PCA와 PLS를 결합시키는 방법을 본 논문에서 제시하였고 모의실험과 실제 난소암 유전체 데이터 분석을 통해서 제안된 방법의 유용함을 입증하였다. 실제로 많은 고차원 유전체 데이터의 경우 그룹 구조를 가지고 있는 경우가 많으므로 본 논문에 제시된 방법을 사용하여 보

**Table 4.2.** For analysis of ovarian cancer data, we listed 46 distinct genes selected by network-based regularization procedure combined with principal component analysis (PCA) and partial least square (PLS) in the analysis of 12,770 genes and the analysis of the 5,995 selected genes, respectively. The distinct 46 genes are denoted by 1 or 0, representing included or not included in the top 20 list of each analysis, respectively. 'NA' means that the corresponding gene was included in the analysis of 12,770 genes but not included in the analysis of selected 5,995 genes.

| Genes | 12,770 genes | | 5,995 genes | | Genes | 12,770 genes | | 5,995 genes | |
|---|---|---|---|---|---|---|---|---|---|
| | PCA | PLS | PCA | PLS | | PCA | PLS | PCA | PLS |
| EGF | 1 | 1 | 1 | 1 | NR2F1 | 1 | 0 | NA | NA |
| CFI | 1 | 1 | 1 | 1 | BRRN1 | 1 | 0 | NA | NA |
| RHOT2 | 1 | 1 | 1 | 1 | Bles03 | 1 | 0 | NA | NA |
| TCAP | 1 | 1 | 1 | 1 | LIM2 | 1 | 0 | NA | NA |
| PPBP | 1 | 1 | 1 | 1 | FOXD4L4 | 1 | 0 | NA | NA |
| ADCY7 | 1 | 1 | 1 | 1 | CD300LF | 0 | 1 | NA | NA |
| SLC22A17 | 0 | 1 | 1 | 1 | EGFL6 | 0 | 1 | NA | NA |
| DST | 0 | 1 | 1 | 1 | RBMS2 | 0 | 1 | NA | NA |
| LIME1 | 1 | 1 | NA | NA | C9orf58 | 0 | 1 | NA | NA |
| GAS2 | 1 | 1 | NA | NA | CX3CL1 | 0 | 0 | 1 | 0 |
| TNFAIP8 | 1 | 1 | NA | NA | MPO | 0 | 0 | 1 | 0 |
| COX7A1 | 1 | 1 | NA | NA | ZNF555 | 0 | 0 | 1 | 0 |
| CEL | 1 | 0 | 1 | 0 | PAK7 | 0 | 0 | 1 | 0 |
| HNRPA0 | 1 | 1 | NA | NA | APBA2 | 0 | 0 | 1 | 0 |
| TSG101 | 0 | 1 | 0 | 1 | AQP8 | 0 | 0 | 1 | 0 |
| NSD1 | 0 | 1 | 0 | 1 | MYF6 | 0 | 0 | 1 | 0 |
| PTPN7 | 0 | 1 | 0 | 1 | TRAF1 | 0 | 0 | 1 | 0 |
| HIST1H4K | 0 | 0 | 1 | 1 | CD81 | 0 | 0 | 0 | 1 |
| WT1 | 0 | 0 | 1 | 1 | ACTN2 | 0 | 0 | 0 | 1 |
| HTR2A | 0 | 0 | 1 | 1 | COX7A2L | 0 | 0 | 0 | 1 |
| SPATA12 | 1 | 0 | NA | NA | MAP2K7 | 0 | 0 | 0 | 1 |
| EBI2 | 1 | 0 | NA | NA | PTPRO | 0 | 0 | 0 | 1 |
| HIST1H4L | 1 | 0 | 0 | 0 | PDCD1LG2 | 0 | 0 | 0 | 1 |

다 더 정확하게 관련 유전자들을 찾아낼 수 있을 것으로 기대한다.

PCA와 PLS의 경우 하나의 유전자에 속해 있는 여러 가지 유전체 변이들의 정보를 하나의 성분으로 요약하는 것이므로 그 하나의 주성분이 가지고 있는 유전체 변이들의 설명력이 상대적으로 낮을 경우 유전자의 요약된 정보의 대표성은 떨어지게 되며 이 경우 질병 및 표현형질에 영향을 주는 유전자를 찾아내는 연관 분석 결과에도 악영향을 미칠 수가 있다. 실제로 모의실험에서 하나의 유전자에서 일정한 상관관계를 가지고 변수들을 생성할 때, 상관계수가 낮은 경우에는 유의한 변수 선택의 비율이 전반적으로 많이 떨어진다는 것을 목격하였다. 따라서 같은 그룹 안에 있는 변수들이 서로서로 높은 상관관계를 가지고 있을 경우 이 논문에서 제안된 방법으로 손쉽게 유의한 변수를 찾아낼 수 있지만 같은 그룹 안의 변수들 간 상관관계가 낮을 경우에는 대체할 수 있는 새로운 방법을 개발하는 것이 다음 과제로서 의미가 있을 것이라 생각한다.

## References

Alexander, D. and Lange, K. (2011). Stability selection for genome-wide association, *Genetic Epidemiology*,

**35**, 722–728.

Chen, M., Cho, J., and Zhao, H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies, *PLoS Genetics*, **7**, e1001353.

Du, P., Zhang, X., Huang, C., Jafari, N., Kibbe, W., Hou, L., and Lin, S. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, *BMC Bioinformatics*, **11**, 587.

Faraway, J. (2014). *Linear Models with R* (2nd ed.), Chapman and Hall/CRC.

Friedman J., Hastie T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**, 1–22.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data, *Bioinformatics*, **24**, 1175–1182.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics, *Annals of Applied Statistics*, **4**, 1498–1516.

Marsit, C., Christensen, B., Houseman, E., Karagas, M., Wrensch, M., Yeh, R., Nelson, H., Wiemels, J., Zheng, S., Posner, M., McClean, M., Wiencke, J., and Kelsey, K. (2009). Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma, *Carcinogenesis*, **30**, 416–422.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society, Series B*, **72**, 417–473.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association*, **104**, 735–746.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent, *Journal of Statistical Software*, **39**, 1–13.

Sun, H. and Wang, S. (2012). Penalized logistic regression for high-dimensional DNA methylation data with case-control studies, *Bioinformatics*, **28**, 1368–1375.

Sun, H. and Wang, S. (2013). Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data, *Statistics in Medicine*, **32**, 2127–2139.

Sun, H., Lin, W., Feng, R., and Li, H. (2014). Network-regularized high-dimensional Cox regression for analysis of genomic data, *Statistca Sinica*, **24**, 1433–1459.

Teschendorff, A., Menon, U., Gentry-Maharaj, A., Ramus, S., Weisenberger, D., Shen, H., Campan, M., Noushmehr, H., Bell, C., Maxwell, A., Savage, D., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S., Jones, A., Beck, S., Wagner, W., Laird, P., Jacobs, I., and Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is hallmark of cancer, *Genome Research*, **20**, 440–446.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Whittaker, J. (1990). *Graphical Models in Applied Mathematical Multivariate Statistics*, Wiley, New York.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

# 그룹 구조를 갖는 고차원 유전체 자료 분석을 위한 네트워크 기반의 규제화 방법

김기풍[a] · 최지윤[a] · 선호근[a,1]

[a]부산대학교 통계학과

(2016년 8월 9일 접수, 2016년 10월 5일 수정, 2016년 10월 5일 채택)

**요 약**

고차원 유전체 자료를 사용하는 유전체 연관 분석에서는 벌점 우도함수 기반의 회귀계수 규제화 방법이 질병 및 표현형질에 영향을 주는 유전자를 발견하는데 많이 이용된다. 특히, 네트워크 기반의 규제화 방법은 유전체 연관성 연구에서의 유전체 경로나 신호 전달 경로와 같은 생물학적 네트워크 정보를 사용할 수 있으므로, Lasso나 Elastic-net과 같은 다른 규제화 방법들과 비교했을 경우 네트워크 기반의 규제화 방법이 보다 더 정확하게 관련 유전자들을 찾아낼 수 있다는 장점을 가지고 있다. 그러나 네트워크 기반의 규제화 방법은 그룹 구조를 갖고 있는 고차원 유전체 자료에는 적용시킬 수 없다는 문제점을 가지고 있다. 실제 SNP 데이터와 DNA 메틸화 데이터처럼 대다수의 고차원 유전체 자료는 그룹 구조를 가지고 있으므로 본 논문에서는 이러한 그룹 구조를 가지고 있는 고차원 유전체 자료를 분석하고자 네트워크 기반의 규제화 방법에 주성분 분석(principal component analysis; PCA)과 부분 최소 자승법(partial least square; PLS)과 같은 차원 축소 방법을 결합시키는 새로운 분석 방법을 제안하고자 한다. 새롭게 제안한 분석 방법은 몇 가지의 모의실험을 통해 변수 선택의 우수성을 입증하였으며, 또한 152명의 정상인들과 123명의 난소암 환자들로 구성된 고차원 DNA 메틸화 자료 분석에도 사용하였다. DNA 메틸화 자료는 대략 20,000여개의 CpG sites가 12,770개의 유전자에 포함되어 있는 그룹 구조를 가지고 있으며 Illumina Infinium Human Methylation27 BeadChip으로부터 생성되었다. 분석 결과 우리는 실제로 암에 연관된 몇 가지의 유전자를 발견할 수 있었다.

주요용어: 고차원 유전체 자료, 네트워크 기반 규제화, 유전체 네트워크, 주성분 분석, 부분 최소 자승법

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과. E-mail: hsun@pusan.ac.kr