The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.4.643
29(4), 643–656

A study on bias effect of LASSO regression for model selection criteria

Donghyeon Yu[a,1]

[a]Department of Statistics, Keimyung University

(Received March 2, 2016; Revised April 26, 2016; Accepted April 28, 2016)

Abstract

High dimensional data are frequently encountered in various fields where the number of variables is greater than the number of samples. It is usually necessary to select variables to estimate regression coefficients and avoid overfitting in high dimensional data. A penalized regression model simultaneously obtains variable selection and estimation of coefficients which makes them frequently used for high dimensional data. However, the penalized regression model also needs to select the optimal model by choosing a tuning parameter based on the model selection criterion. This study deals with the bias effect of LASSO regression for model selection criteria. We numerically describes the bias effect to the model selection criteria and apply the proposed correction to the identification of biomarkers for lung cancer based on gene expression data.

Keywords: LASSO, penalized regression, bias, model selection, information criterion

1. 서론

기술의 발달에 따라 많은 양의 정보가 관측, 생성 및 축적되고 있으며, 이를 기반으로 여러 분야에서 다양한 통계적인 방법들이 개발되고 있다. 특히, 최근 관측되는 자료 중 다수는 변수의 수(p)가 표본의 수(n)보다 매우 많은 고차원 자료(high dimensional data)의 형태를 가지고 있다. 고차원 자료의 대표적인 예로는 생명 과학 분야에서 특정 질병 발생의 위험도 예측 및 질병과 연관된 바이오마커(biomarker)의 식별에 이용되는 유전자 발현량(gene expression) 자료를 들 수 있다.

이러한 고차원 자료의 형태($p > n$)에서는 회귀 모형의 최소제곱추정량(least squares estimator; LSE)이 정의되지 않기 때문에 이를 해결하기 위하여 여러 벌점화 회귀 모형(penalized regression model)이 제안되었다. 특히, 벌점화 회귀 모형 중 ℓ_1 노름(norm)을 기반으로 하는 LASSO (Tibshirani, 1996), 비볼록 벌점(nonconvex penalty)에 기반한 SCAD (Fan과 Li, 2001), MCP (Zhang, 2010) 등의 벌점화 회귀 모형은 특정 회귀 계수를 0으로 추정하여 회귀 계수의 추정과 변수의 선택을 동시에 하는 장점을 지니고 있다. 하지만 원래의 회귀 모형에서의 모형 선택과 마찬가지로 벌점화 회귀 모형을 통하여 자료를 분석하는 경우에도 조율 모수(tuning parameter)를 선택하여 최적의 모형(optimal model)을 선택해야 한다.

This research was supported by the Bisa Research Grant of Keimyung University in 2015.
[1]Department of Statistics, Keimyung University, 1095 Dalgubeol-daero, Dalseo-gu, Daegu 42601, Korea. E-mail: dyu3@kmu.ac.kr

최적의 모형을 선택하기 위하여 적용되는 기준으로 Akaike 정보 기준(Akaike information criterion; AIC), 베이지안 정보 기준(Bayesian information criterion; BIC), 일반화된 정보 기준(generalized information criterion; GIC) 및 교차검증(cross validation; CV) 등이 제안되었다 (Akaike, 1973; Nishii, 1984; Picard와 Cook, 1984; Schwarz, 1978). 이 중에서 BIC는 변수의 수가 고정된 경우(즉, 표본의 수가 증가함에 따라 변수의 수는 변하지 않는 경우)에 표본의 수가 증가함에 따라 점근적으로 참모형(true model)을 선택하는 모형 선택의 일치성(model selection consistency)을 지니며, AIC는 과적합(overfitting)으로 인하여 일치성을 지니지 않음이 알려져 있다 (Shao, 1997). 또한, AIC와 CV는 점근적으로 동일하며 두 기준을 사용할 경우 유사한 모형이 선택됨이 알려져 있다 (Yang, 2005). 최근에는 기존의 연구를 보다 일반화하여 변수의 수가 표본의 수에 따라 증가하는 경우에 대한 모형 선택의 기준들이 활발히 연구되고 있으며 (Fan과 Tang, 2013; Wang 등, 2009; Wang과 Zhu, 2011), 일반화 선형 모형(generalized linear model)에 대한 모형 선택의 일치성을 갖는 GIC 기준이 발표되었다 (Fan과 Tang, 2013).

벌점화 회귀 모형의 경우에는 벌점(penalty)을 고려하여 추정이 이루어지므로 원래의 회귀 모형의 최소제곱추정량과는 다르게 항상 편의(bias)를 지니고 있다. 이에 따라, 고차원 자료에서의 벌점화 회귀 모형의 모형 선택의 일치성은 표본의 수가 증가함에 따라 조율 모수의 값이 0으로 수렴함을 가정을 통하여 불편성을 만족하는 가정 하에서 증명이 이루어 졌다 (Wang 등, 2007). 하지만, 실제 자료를 분석하는 경우와 같이 유한 표본(finite sample)의 자료를 이용할 경우에는 추정량의 편의가 모형 선택의 기준들에 영향을 미치게 된다. 그 중에서도 LASSO 벌점은 SCAD 및 MCP 벌점들과 다르게 편의의 영향으로 일반적으로 모형 선택의 일치성이 만족하지 않는 것이 알려져 있다 (Zou, 2006).

하지만, 조건부종속성(conditional dependence)을 기반으로 변수들 사이의 의존성을 모형화하는 가우시안 그래피컬 모형(Gaussian graphical model)의 회귀 모형 기반의 접근 방법에서 이웃안정성(neighborhood stability) 조건을 만족할 경우, LASSO 벌점 기반의 추정량이 모형 선택의 일치성을 지님이 증명되었으며 (Meinshausen과 Bühlmann, 2006), 이 후, LASSO 벌점에 기반한 여러 추정방법들이 제안되었다 (Friedman 등, 2008; Khare 등, 2015; Peng 등, 2009). 특히, Peng 등 (2009)와 Khare 등 (2015) 에서는 모형 선택의 기준으로 BIC-type의 기준을 사용하였으며, Danaher 등 (2014)에서는 AIC 기준을 사용하여 조율 모수를 선택하였다. 가우시안 그래피컬 모형 중 회귀 모형 기반의 여러 추정 방법들은 LASSO 회귀 모형의 형태로 표현이 가능하며, 벌점화 모형의 추정량을 토대로 기존의 정보 기준을 적용하여 모형을 선택하고 있기 때문에 유한 표본에서는 편의의 영향으로 더 적절한 모형이 있음에도 다른 모형을 선택하는 오류가 발생할 수 있다. 따라서 본 연구에서는 벌점화 모형 중 LASSO 회귀 모형에서 편의가 모형 선택의 기준들 및 모형 선택에 미치는 영향에 대하여 모의 실험을 통하여 수치적으로 확인하고 모형 선택의 기준들을 적용 시 편의에 대한 보정이 필요함을 보이고자 한다.

추가적으로, 보통의 선형 회귀 모형 하에서 LASSO 벌점의 편의와 일치성을 만족하기 위한 조건에 대한 이론적인 연구는 Zang과 Huang (2008)에서 참고 할 수 있으며, Belloni와 Chernozhukov (2013)에서는 본 연구에서 적용한 것처럼 LASSO 회귀 추정량을 이용하여 모형 선택 이후, 이에 대응하는 최소제곱주청량을 이용하는 post-Lasso 추정량에 대한 이론적인 성질에 대하여 연구하였으며, 모형 선택의 일치성을 만족하지 않는 조건 하에서도 항상 LASSO 회귀 추정량 보다 이론적인 성질이 개선됨을 보였다. 본 연구는 Belloni와 Chernozhukov (2013)에서 비교, 제시한 추정량의 이론적 성질을 실제 유한 표본 하에서 여러 대표적인 정보 기준들 적용하여 편의를 보정한 추정량의 모형 선택 성능이 개선됨을 보였고, 미리 조정된 공액 경사도 알고리즘을 적용하여 효율적으로 편의를 보정하는 방법을 제안한 점에서 의미가 있다.

본 논문의 구성은 다음과 같다. 2절에서 LASSO 회귀 모형 및 모형 선택의 기준들에 대하여 소개하고 3절에서 LASSO 회귀 모형의 추정량이 지닌 편의에 대하여 설명하였다. 4절에서는 여러 정보 기준들에 대하여 편의를 지닌 추정량과 편의를 보정한 추정량을 적용하고 모형 선택의 성능을 비교하였다. 실제 자료의 분석을 통한 비교를 위하여, 5절에서는 폐암 환자들의 유전자 발현량 자료를 토대로 바이오마커를 식별하는 문제를 LASSO 회귀 모형을 적용하여 분석하였다. 마지막으로 6절에서 본 연구에 대한 결론을 정리하였다.

2. LASSO 회귀 모형 및 모형 선택의 기준들

2.1. LASSO 회귀 모형

본 연구는 회귀 계수의 추정과 변수 선택을 동시에 진행하는 벌점화 회귀 모형 중에서 대표적인 ℓ_1 노름 기반의 LASSO 회귀 모형 (Tibshirani, 1996)

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.1}$$

에 대하여 LASSO 회귀 모형의 추정량의 편의가 모형 선택의 기준들에 미치는 영향을 확인하고 이에 대한 보정이 필요함을 보이고자 한다. 여기서, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$는 n차원의 반응변수 벡터(vector), $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$는 $n \times p$차원의 설계 행렬(design matrix), $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$는 i번째 표본에 대한 설명 변수들의 관측값으로 p차원의 벡터, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$는 p차원의 회귀 계수 벡터, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$는 ℓ_1 노름, $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} |\beta_j|^2}$는 ℓ_2 노름을 나타내며 λ는 비음의 조율 모수를 나타낸다. 또한, 각 반응 변수 및 설명변수들은 $y_i - (1/n)\sum_{i=1}^{n} y_i$와 같은 연산을 통하여 간단히 0으로 중심화 가능하므로 일반성을 잃지 않고 반응변수와 개별 설명 변수들은 평균이 0으로 중심화되어 있다고 가정한다. LASSO 회귀 모형은 원래의 회귀 모형의 목적함수(objective function)인 오차의 제곱합에 벌점 함수(penalty function)로 회귀 계수의 ℓ_1 노름을 고려한 모형으로 추정된 회귀 계수 중 일부를 0으로 추정하는 성질을 지닌 벌점화 회귀 모형이다.

LASSO 회귀 모형의 추정량은 ℓ_1 노름의 미분불가능 성질로 인하여 최소제곱추정량과는 다르게 정확한 해(exact solution)의 표현이 불가능하다. 따라서 최적화(optimization) 방법들을 적용하여 식 (2.1)을 최소화하는 해를 수치적으로 찾아야 한다. 이에 대한 효율적인 알고리즘에 대한 연구가 집중적으로 이루어졌으며, 대표적인 알고리즘은 조율 모수에 따른 해의 경로(path)를 기반으로 추정하는 Efron 등 (2004)의 least angle regression(LARS) 및 주어진 조율 모수에서의 최적해를 찾는 Friedman 등 (2007)의 coordinate descent(CD) 알고리즘을 들 수 있다. 본 연구에서는 LASSO 회귀 모형의 추정량의 경로보다는 주어진 조율 모수 하에서의 편의에 대한 계산에 관심을 두고 있으므로 두 알고리즘들 중에서 CD 알고리즘을 기반으로 논의를 진행한다.

CD 알고리즘은 반복 알고리즘(iterative algorithm)으로 이전 반복의 LASSO 회귀 모형의 회귀 계수 추정값을 $\hat{\beta}^{(old)} = (\hat{\beta}_j^{(old)})$라 할 때, 각 $j = 1, 2, \ldots, p$에 대하여 β_j를 제외한 나머지 회귀 계수들을 이전 반복에서의 추정값으로 고정하여 한 번에 하나의 회귀 계수를 갱신하면서 전체 추정량이 수렴할 때까지 반복하는 알고리즘이다. 즉, 각 $j = 1, 2, \ldots, p$에 대하여 $\tilde{\beta}_k = \hat{\beta}_k^{(old)}$, $k \neq j$라 할 때, 목적함수

$$f\left(\beta_j; \tilde{\beta}_k, k \neq j\right) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k - x_{ij}\beta_j\right)^2 + \lambda\sum_{k=1}^{p}|\beta_k| \tag{2.2}$$

를 β_j에 대하여 최소화하는 해를 찾는다. 위와 같이 하나의 회귀 계수에 대한 해는 정확하게 아래와 같

이

$$\hat{\beta}_j^{(new)} = \text{sign}\left(\sum_{i=1}^n e_{i(j)}x_{ij}\right)\left(\left|\sum_{i=1}^n e_{i(j)}x_{ij}\right| - \lambda\right)_+ \Big/ \sum_{i=1}^n x_{ij}^2 \tag{2.3}$$

로 나타낼 수 있다. 여기서, $e_{i(j)} = y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k$, $(x)_+ = \max(0, x)$를 나타낸다. 본 연구에서는 위의 CD 알고리즘을 적용하여 LASSO 회귀 모형의 추정량을 계산한다.

2.2. 모형 선택의 기준들

LASSO 회귀 모형을 이용하여 모형을 선택하고자 할 경우에는 대표적으로 AIC, BIC, GIC 정보기준들 및 CV가 사용되고 있다. 앞서 언급한 정보 기준들은 일반적으로 모형 적합에 대한 측도와 모형의 복잡도에 대한 측도를 고려하여 정의된다. 먼저 관측된 자료에 대하여 회귀 모형

$$y_i = \mathbf{x}_i^T\beta + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right)$$

을 따른다고 가정할 때, 위의 정보 기준들은 아래와 같이 정의된다.

(1) AIC: $-2\log L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) + 2 \times \text{df}$.

(2) BIC: $-2\log L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) + \log n \times \text{df}$.

(3) GIC: $-2\log L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) + \log(\log n)\log p \times \text{df}$.

여기서 $\log L(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = -(n/2)\log(2\pi\sigma^2) - \sum_{i=1}^n (y_i - \mathbf{x}_i^T\beta)^2/(2\sigma^2)$로 β와 σ^2에 대한 로그가능도함수(loglikelihood function), $\text{df} = \sum_{j=1}^p I(\beta_j \neq 0)$이며 0이 아닌 회귀 계수의 수를 나타내며 $I(\cdot)$은 지시함수(indicator function)이다. 교차검증은 전체 자료를 교차적으로 훈련 자료(training data)와 검증 자료(test data)로 나누어 훈련 자료에 기반한 추정값을 검증 자료에 적용하여 예측오차를 계산한다. 일반적으로 검증 자료로 개별 표본을 이용하는 leave-one-out 교차검증(LOOCV)와 자료를 k개의 그룹으로 나누어 개별 그룹을 검증 자료로 이용하는 k-묶음 교차검증(k-fold CV) 방법이 적용된다.

정보 기준들을 적용한 모형 선택에 관한 연구는 최근 활발히 이루어 졌으며, 특히 표본의 수가 커질 때, 참모형을 선택하도록 하는 성질인 모형 선택의 일치성(consistency)에 대한 연구가 집중적으로 이루어졌다. AIC를 기준으로 선택한 모형은 예측오차(prediction error)를 최소화하며 이는 교차검증과 점근적으로 같음이 알려져 있으며 (Yang, 2005), 과적합에 의해 참모형에 대한 일치성을 갖지 않을 수 있음이 알려져 있다 (Shao, 1997). AIC와는 다르게 BIC를 기준으로 모형을 선택하는 경우에는 회귀 모형에서 모형 선택의 일치성을 지님이 알려져 있으며 (Shao, 1997), 벌점화 모형 중에서 SCAD 모형 하에서 모형 선택의 일치성이 증명되었다 (Wang 등, 2007). 앞서 모형 선택의 일치성이 변수의 수가 고정된 상태에서 표본의 수가 증가할 때 대한 점근적 성질로 증명된 것에 반하여, 최근에 Fan과 Tang (2013)에 의해 제안된 GIC는 표본의 수가 증가함에 따라 변수의 수도 같이 증가하는 조건 하에서 일반화 선형 모형에 대한 모형 선택의 일치성을 지님이 알려져 있다. 본 연구에서는 앞서 언급하였던 AIC, BIC, GIC 및 CV에 대하여 LASSO 회귀 모형의 추정량의 편의에 대한 영향을 4절의 모의실험을 통하여 수치적으로 표현하고 비교하였다.

3. Lasso 회귀 모형 추정량의 편의

2절에서 언급한 모형 선택의 일치성은 일반적으로 최소제곱추정량의 일치성(consistency)에 기반하여 이루어 졌으며 (Shao, 1997), 벌점화 모형의 경우에는 표본의 수가 증가함에 따라 조율 모수 λ_n이 0으

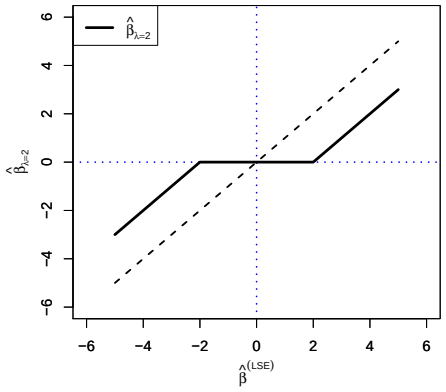


Figure 3.1. Comparison of least square estimator ($\hat{\beta}^{(\mathrm{LSE})}$) and lasso estimator ($\hat{\beta}_\lambda$) with $\lambda = 2$ when $\mathbf{X}^T\mathbf{X} = I$.

로 수렴함을 가정하고 증명이 이루어졌다 (Wang 등, 2007). 하지만, 일반적으로 실제 고차원 자료를 분석하는 경우에는 유한 표본을 이용하여 적합을 하게 되므로 벌점화 모형 적용 시 항상 추정량의 편의가 존재하게 되며 모형 선택 기준들에 영향을 미치게 된다.

본 논문에서 다루고자 하는 LASSO 추정량은 $\mathbf{X}^T\mathbf{X} = I$인 조건 하에서

$$\hat{\beta}_\lambda = \mathrm{sign}\left(\mathbf{X}^T\mathbf{y}\right)\left(\left|\mathbf{X}^T\mathbf{y}\right| - \lambda\right)_+$$

로 표현할 수 있다. 여기서, $\mathrm{sign}(x)$는 x의 부호(sign)를 나타내며 $(x)_+ = \max(0, x)$를 나타낸다. 이 경우, 0이 아닌 LASSO 회귀 계수 추정값에 대한 편의는 $-\lambda\,\mathrm{sign}(\mathbf{X}^T\mathbf{y})$를 통하여 나타낼 수 있으며 개별 LASSO 추정량의 편의 정도는 Figure 3.1에서 LASSO 회귀 계수 추정값과 LSE 회귀 계수 추정값의 비교를 통하여 나타낸 것과 같이 λ의 값에 의해 편의의 정도가 결정되어 진다.

하지만, 일반적인 $\mathbf{X}$에 대한 LASSO 회귀 모형의 추정량은 2.1절에서 설명한 바와 같이 정확한 식으로 표현할 수 없으며, 아래의 Karush-Kuhn-Tucker(KKT) 최적해의 조건

$$\mathbf{X}^T\mathbf{X}\hat{\beta}_\lambda = \mathbf{X}^T\mathbf{y} - \lambda\mathbf{s} \tag{3.1}$$

을 만족하는 해를 찾게 된다. 여기서, $\hat{\beta}_\lambda$는 주어진 λ에서의 LASSO 회귀 모형의 회귀 계수 추정량을 나타내며, $\mathbf{s} = (s_1, s_2, \ldots, s_p)^T$, $\hat{\beta}_{\lambda,i} \neq 0$이면 $s_i = \mathrm{sign}(\hat{\beta}_{\lambda,i})$, $\hat{\beta}_{\lambda,i} = 0$이면 $s_i \in [-1, 1]$을 나타낸다.

본 연구에서 관심을 두고 있는 모형 선택 기준들에 대한 편의의 영향은 $\sum_{i=1}^n (y_i - \mathbf{x}_i^T\hat{\beta}_\lambda)^2$의 계산에 대하여 LASSO 회귀 추정값 중에서 0이 아닌 값으로 추정된 회귀 계수들의 편의만 영향을 미치므로 LASSO 회귀 추정량의 편의는 추정된 LASSO 회귀 추정값 중 0이 아닌 값으로 추정된 값과 이에 대응하는 변수들로 계산된 최소제곱추정량의 비교를 통하여 측정하고자 한다. 이를 위해, 먼저 활성화 집합(active set) $I_\lambda = \{j \mid \hat{\beta}_{\lambda,j} \neq 0\}$을 정의하고 이에 대응하는 최소제곱추정량을 $\hat{\beta}_{I_\lambda}^{\mathrm{LSE}}$로 나타낸다. 대응하는 회귀 계수를 명확히 하기 위하여 활성화 집합의 원소는 크기 순서로 정렬되어 있다고 가정한다. 위의 표현을 이용하면 LASSO 회귀 추정량 중 0이 아닌 회귀 계수로 추정된 계수들의 최소제곱 추정량은

$$\hat{\beta}_{I_\lambda}^{\mathrm{LSE}} = \left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\mathbf{X}_{I_\lambda}^T\mathbf{y} \tag{3.2}$$

로 표현할 수 있다. 여기서 $\mathbf{X}_{I_\lambda}$는 설계 행렬 $\mathbf{X}$를 $\mathbf{X} = (\mathbf{X}_{I_\lambda}\ \mathbf{X}_{I_\lambda^c})$와 같이 재배열하여 분할한 행렬의 부분 행렬이다.

식 (3.1)과 위의 식 (3.2)에서 나타낸 불편 추정량인 $\hat{\beta}_{I_\lambda}^{\text{LSE}}$의 표현을 이용하면 $\hat{\beta}_{I_\lambda}^{\text{LSE}}$에 대응하는 0이 아닌 LASSO 추정량 $\hat{\beta}_{\lambda,I_\lambda}$을

$$\hat{\beta}_{\lambda,I_\lambda} = \left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\mathbf{X}_{I_\lambda}^T\mathbf{y} - \lambda\left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\text{sign}\left(\hat{\beta}_{\lambda,I_\lambda}\right) = \hat{\beta}_{I_\lambda}^{\text{LSE}} - \lambda\left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\text{sign}\left(\hat{\beta}_{\lambda,I_\lambda}\right) \quad (3.3)$$

와 같이 나타낼 수 있다. 따라서 0이 아닌 LASSO 추정량에 포함된 편의는 계산된 LASSO 추정값의 부호 정보를 이용하여 $-\lambda(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda})^{-1}\text{sign}(\hat{\beta}_{\lambda,I_\lambda})$로 계산할 수 있다. 이에 따라 모형 선택의 기준들 중에서 정보 기준들에서 요구되는 $-2\log L(\beta,\sigma^2;\mathbf{y},\mathbf{X})$의 계산의 경우, 본 연구에서 적용된 CD 알고리즘의 결과로 $\hat{\beta}_\lambda$ 및 $e_\lambda = \mathbf{y} - \mathbf{X}\hat{\beta}_\lambda$를 얻을 수 있음에 착안하여,

$$e_{I_\lambda} = \mathbf{y} - \mathbf{X}_{I_\lambda}\hat{\beta}_{I_\lambda}^{\text{LSE}} = \mathbf{y} - \mathbf{X}_{I_\lambda}\left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\mathbf{X}_{I_\lambda}^T\mathbf{y}$$

를 이용하여 계산 하지 않고 아래와 같이

$$e_{I_\lambda} = e_\lambda - \lambda\mathbf{X}_{I_\lambda}\left(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\right)^{-1}\text{sign}\left(\hat{\beta}_{\lambda,I_\lambda}\right) \quad (3.4)$$

의 관계를 이용하여 계산을 효율적으로 하고자 한다. 따라서, 회귀 모형의 가정 하에서 적합에 대한 측도인 가능도 함수의 값은 $\hat{\beta}_{I_\lambda}^{\text{LSE}}$와 $\hat{\sigma}_\lambda^2 = (1/n)\|\mathbf{y} - \mathbf{X}_{I_\lambda}\hat{\beta}_{I_\lambda}^{\text{LSE}}\|_2^2$을 이용하여 계산하게 되므로 간단히 표현하면,

$$-2\log L\left(\hat{\beta}_{I_\lambda}^{\text{LSE}},\sigma_\lambda^2;\mathbf{y},\mathbf{X}\right) = n\log\left(e_{I_\lambda}^T e_{I_\lambda}\right) + c \quad (3.5)$$

로 나타낼 수 있다. 여기서 e_{I_λ}는 식 (3.4)에 정의되어 있으며, $c = n\log(2\pi/n) + n$. 위의 결과를 토대로 본 연구에서는 모형 적합에 대한 측도를 식 (3.5)를 이용하여 계산한다.

또한, 교차검증의 경우에는 $\hat{\beta}_{I_\lambda}^{\text{LSE}}$에 대한 직접적인 계산이 필요하므로 정보 기준에서와 유사하게 편의 $\lambda(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda})^{-1}\text{sign}(\hat{\beta}_{\lambda,I})$를 먼저 계산한 뒤에 식 (3.3)을 이용하여 교차검증 오차

$$\text{CV}(\lambda) = \sum_{k=1}^{K}\left\|\mathbf{y}^{(k)} - \mathbf{X}_{I_\lambda}^{(k)}\hat{\beta}_{I_\lambda(k)}^{\text{LSE}}\right\|_2^2$$

를 계산한다. 여기서 K는 K-묶음 교차검증에서의 K개의 묶음을 나타내며, $\mathbf{y}^{(k)}$는 k번째 묶음의 검증자료의 반응변수, $\mathbf{X}_{I_\lambda}^{(k)}$는 k번째 묶음의 검증자료의 활성화 집합 I에 대응하는 설계 행렬, $\hat{\beta}_{I_\lambda(k)}^{\text{LSE}}$는 k번째 묶음의 훈련자료를 토대로 추정된 활성화 집합 I에 대응하는 최소제곱추정량을 나타낸다.

본 연구에서는 편의의 계산을 효율적으로 하기 위하여 $(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda})^{-1}$을 직접적으로 계산하지 않고

$$\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}\mathbf{a}_\lambda = \lambda\,\text{sign}\left(\hat{\beta}_{\lambda,I_\lambda}\right) \quad (3.6)$$

의 선형방정식을 Algorithm 1에 제시된 미리 조정된 공액 경사도(preconditioned conjugate gradient; PCG)방법을 이용하여 근사적으로 계산한다 (Demmel, 1997). 식 (3.6)의 해는 Algorithm 1에서 $\mathbf{A} = \mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda}$, $\mathbf{b} = \lambda\text{sign}(\hat{\beta}_{\lambda,I_\lambda})$, $\mathbf{x} = \mathbf{a}_\lambda$, $\mathbf{M} = \text{diag}(\mathbf{X}_{I_\lambda}^T\mathbf{X}_{I_\lambda})$로 적용하여 구한다. 본 논문에서 계산된 LASSO 추정량의 편의는 $(\mathbf{X}_I^T\mathbf{X}_I)^{-1}$에 의존하므로 0이 아닌 LASSO 추정량의 수가 표본의 수 (n)보다 큰 경우에는 편의를 보정할 수 없다. 본 연구에서는 Tibshirani (2013)의 결과로 부터, 설명 변수 및 종속 변수가 연속형 변수인 경우 LASSO 추정량은 n과 p에 상관없이 유일한 성질을 지니므로 0이 아닌 LASSO 추정량의 수가 n보다 큰 경우에는 편의를 보정하지 않고 모형 선택 기준을 계산하였다.

Algorithm 1 Preconditioned conjugate gradient method for solving $\mathbf{Ax} = \mathbf{b}$

Require: $\mathbf{A}, \mathbf{M}, \mathbf{b}, \delta_{tol}$

$k = 0, \mathbf{x}_0 = 0, \mathbf{r}_0 = \mathbf{b}, \mathbf{p}_1 = \mathbf{M}^{-1}\mathbf{b}, \mathbf{y}_0 = \mathbf{M}^{-1}\mathbf{r}_0$ ▷ initialization

repeat

  $k = k + 1$

  $\mathbf{z} = \mathbf{A}\mathbf{p}_k$

  $\nu_k = \dfrac{\mathbf{y}_{k-1}^T \mathbf{r}_{k-1}}{\mathbf{p}_k^T \mathbf{z}}$

  $\mathbf{x}_k = \mathbf{x}_{k-1} + \nu_k \mathbf{p}_k$

  $\mathbf{r}_k = \mathbf{r}_{k-1} - \nu_k \mathbf{z}$

  $\mathbf{y}_k = \mathbf{M}^{-1}\mathbf{r}_k$

  $\mu_{k+1} = \dfrac{\mathbf{y}_k^T \mathbf{r}_k}{\mathbf{y}_{k-1}^T \mathbf{r}_{k-1}}$

  $\mathbf{p}_{k+1} = \mathbf{y}_k + \mu_{k+1}\mathbf{p}_k$

until $\dfrac{\|\mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} < \delta_{tol}$

4. 모의 실험

본 절에서는 LASSO 회귀 추정량의 편의가 AIC, BIC, GIC 및 CV의 값에 어떠한 영향을 주는지 LASSO 회귀 추정량과 편의를 보정한 추정량의 모형 선택 기준들에 대한 비교를 통하여 확인하고자 한다. 모의 실험은 희박한(sparse) 회귀 계수를 고려하기 위하여 참회귀계수 $\beta^0 = (\beta_j^0)_{1 \le j \le p}$를

$$\beta_j^0 = \begin{cases} 2, & \text{for } j = 1, 2, \ldots, 20, \\ -1, & \text{for } j = 41, 42, \ldots, 45, \\ 3, & \text{for } j = 66, 67, \ldots, 70, \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

로 고려하여 30개의 0이 아닌 회귀계수를 포함한다. 표본의 수(n)과 변수의 수(p)에 따른 영향을 확인하기 위하여 $n = 200$으로 고정한 뒤, $p = 100, 200, 500, 1000$의 4가지 경우를 고려하였다. 본 모의 실험에서는 100개의 자료를 생성하여 각각의 모형 선택 기준들을 계산하였으며 반응 변수 $\mathbf{y} = (y_1, y_2, \ldots, y_n)$와 설계 행렬 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$는

$$y_i = \mathbf{x}_i^T \beta^0 + \epsilon_i$$

를 통하여 생성하였다. 여기서 $\mathbf{x}_i \sim N(0, I_p)$, $\epsilon_i \sim N(0, 1)$로 가정하였으며, I_p는 p차원의 단위 행렬이다. 여러 조율 모수에 대하여 AIC, BIC, GIC 및 CV를 계산하여 $p = 500$일 때의 100개의 자료에 대한 평균값을 Figure 4.1에 나타내었다. Figure 4.1에서 편의를 보정한 경우 $-2 \log L(\beta, \sigma^2; \mathbf{y}, \mathbf{X})$값의 감소를 통하여 정보기준들이 전체적으로 감소된 것을 볼 수 있으며, AIC의 경우 최소화하는 점이 크게 영향을 받지 않았으나 BIC, GIC 및 CV를 최소화하는 λ^*는 증가한 것을 확인할 수 있다. 이는 편의를 보정할 경우, 편의를 보정하지 않을 때와 비교하여 상대적으로 희박한 모형을 선택하는 것을 의미한다. 또한, Figrue 4.1의 (b) BIC와 (c) GIC의 그림을 살펴보면 작은 λ값에 대하여 BIC 및 GIC의 값이 급격하게 감소하는 패턴이 나타난다. 이는 작은 λ로 인하여 과적합(overfitting)이 발생함에 따라 모형의 자

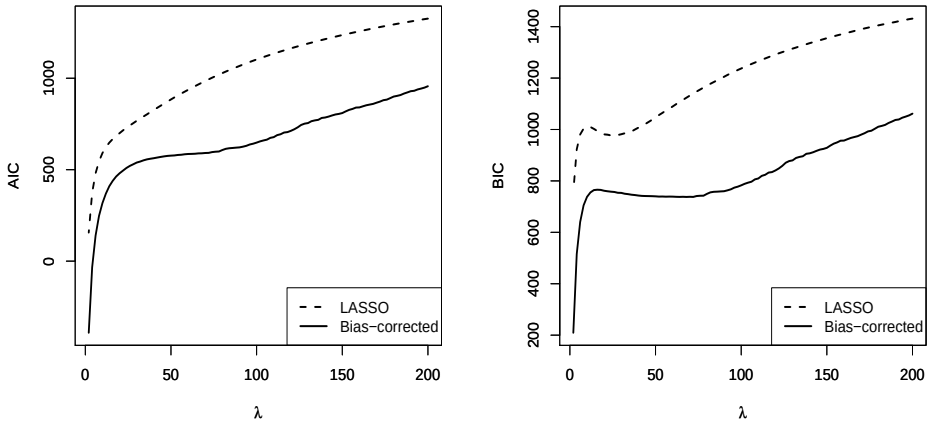
Table 4.1. Results of the estimation of β: averages of the number of nonzero estimates $\hat{\beta}_\lambda$, sensitivity (SEN), specificity (SPE), false discovery rate (FDR), ℓ_2 norm of estimation errors for "LAS" and "BC", where "LAS" and "BC" denote that the model section criteria (MSC) calculated from the naive lasso estimate and the bias-corrected estimate, respectively. For SEN, SPE, and FDR, we report the numbers multiplied by 100. Numbers in parenthesis denote standard errors.

p	MSC	λ^*		$\{\lvert\hat{\beta}\neq 0\rvert\}$		SEN		SPE		FDR		$\lVert\hat{\beta}-\beta_0\rVert_2$	
		LAS	BC	LAS	BC	LAS	BC	LAS	BC	LAS	BC	LAS	BC
100	AIC	4.20	13.20	77.00	53.72	100.00	100.00	32.86	66.11	60.41	42.78	0.77	0.78
		(0.17)	(0.42)	(0.94)	(0.85)	(0.00)	(0.00)	(1.34)	(1.21)	(0.53)	(0.89)	(0.01)	(0.01)
	BIC	14.80	51.46	48.07	33.20	100.00	100.00	74.19	95.43	36.31	8.97	0.79	0.50
		(0.35)	(1.76)	(0.71)	(0.30)	(0.00)	(0.00)	(1.01)	(0.43)	(0.89)	(0.75)	(0.01)	(0.01)
	GIC	19.64	60.52	42.06	31.82	100.00	100.00	82.77	97.40	27.81	5.53	0.89	0.45
		(0.39)	(1.56)	(0.48)	(0.15)	(0.00)	(0.00)	(0.68)	(0.21)	(0.78)	(0.42)	(0.01)	(0.01)
	CV	6.32	55.56	70.36	33.15	100.00	100.00	42.34	95.50	43.08	91.06	0.71	0.48
		(0.15)	(1.53)	(0.70)	(0.28)	(0.00)	(0.00)	(1.00)	(0.40)	(0.45)	(0.68)	(0.01)	(0.01)
200	AIC	2.00	2.00	151.65	151.65	100.00	100.00	28.44	28.44	80.20	80.20	1.33	1.92
		(0.00)	(0.00)	(0.45)	(0.45)	(0.00)	(0.00)	(0.26)	(0.26)	(0.06)	(0.06)	(0.01)	(0.02)
	BIC	19.56	67.14	58.20	36.15	100.00	100.00	83.41	96.38	46.25	15.78	1.02	0.54
		(0.52)	(2.77)	(1.39)	(0.50)	(0.00)	(0.00)	(0.82)	(0.29)	(1.01)	(0.94)	(0.01)	(0.01)
	GIC	28.44	81.78	45.32	34.23	100.00	100.00	90.99	97.51	32.90	11.82	1.23	0.49
		(0.48)	(2.44)	(0.54)	(0.28)	(0.00)	(0.00)	(0.32)	(0.16)	(0.79)	(0.68)	(0.02)	(0.01)
	CV	8.10	58.68	98.08	36.96	100.00	100.00	59.95	95.91	31.09	82.09	0.91	0.55
		(0.19)	(1.38)	(1.27)	(0.41)	(0.00)	(0.00)	(0.75)	(0.24)	(0.40)	(0.85)	(0.01)	(0.01)
500	AIC	2.00	2.00	182.20	182.20	100.00	100.00	67.62	67.62	83.53	83.53	1.43	1.61
		(0.00)	(0.00)	(0.39)	(0.39)	(0.00)	(0.00)	(0.08)	(0.08)	(0.04)	(0.04)	(0.02)	(0.02)
	BIC	33.36	74.14	59.63	45.03	100.00	100.00	93.70	96.80	48.97	31.27	1.58	0.66
		(0.20)	(2.58)	(0.72)	(0.84)	(0.00)	(0.00)	(0.15)	(0.18)	(0.61)	(1.17)	(0.03)	(0.02)
	GIC	40.00	88.82	53.04	42.00	100.00	100.00	95.10	97.45	42.34	27.07	1.78	0.57
		(0.75)	(1.88)	(0.78)	(0.65)	(0.00)	(0.00)	(0.17)	(0.14)	(0.78)	(1.00)	(0.04)	(0.01)
	CV	7.96	50.10	140.01	52.40	100.00	100.00	76.59	95.23	21.78	60.36	1.23	0.77
		(0.27)	(1.55)	(1.83)	(1.39)	(0.00)	(0.00)	(0.39)	(0.30)	(0.27)	(1.27)	(0.02)	(0.02)
1000	AIC	10.00	10.02	150.19	150.15	100.00	100.00	87.61	87.61	79.98	79.97	1.65	1.77
		(0.00)	(0.02)	(0.72)	(0.73)	(0.00)	(0.00)	(0.07)	(0.08)	(0.10)	(0.10)	(0.04)	(0.03)
	BIC	34.78	64.06	82.22	67.72	99.97	100.00	94.62	96.11	62.76	52.41	2.03	1.06
		(0.41)	(3.01)	(1.18)	(1.83)	(0.03)	(0.00)	(0.12)	(0.19)	(0.55)	(1.29)	(0.05)	(0.04)
	GIC	80.28	98.64	56.17	53.44	94.77	98.70	97.14	97.54	47.38	42.71	3.34	0.98
		(6.12)	(2.74)	(1.36)	(1.04)	(1.26)	(0.54)	(0.11)	(0.10)	(0.94)	(1.05)	(0.21)	(0.08)
	CV	10.48	44.04	148.33	79.08	100.00	99.53	87.80	94.93	20.31	40.95	1.65	1.31
		(0.19)	(1.75)	(0.91)	(2.43)	(0.00)	(0.18)	(0.09)	(0.25)	(0.13)	(1.13)	(0.04)	(0.05)

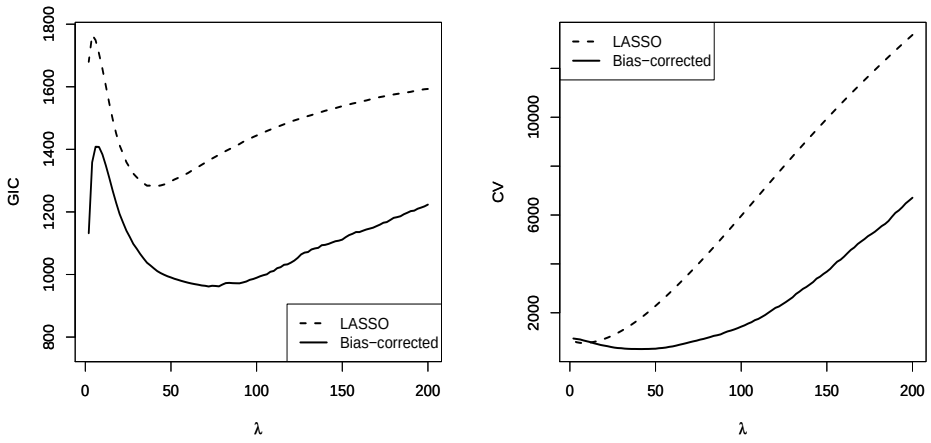
AIC = Akaike information criterion; BIC = Bayesian information criterion;
GIC = generalized information criterion; CV = cross validation.

유도를 모형 선택 기준에 반영하여도 과적합으로 인하여 모형 선택 기준의 값이 매우 작아지게 되는 것을 의미한다. 위와 같은 경우에는 과적합이 발생하기 전 까지의 모형 선택 기준을 통하여 모형을 선택하는 것이 타당하므로 본 연구에서는 $p = 500, 1000$인 경우에 대하여 BIC 기준의 경우 위와 같은 패턴이 나타나 $\lambda > 25$인 경우로 한정하여 모형 선택을 하였다.

단순히 희박한 모형을 선택하는 것이 모형 선택의 성능이 개선됨을 의미하는 것은 아니므로 선택된 모형



(a) AIC (b) BIC



(c) GIC (d) CV

Figure 4.1. Averages of Akaike information criterion (AIC), Bayesian information criterion (BIC), generalized information criterion (GIC), and cross validation (CV) for various λ with $p = 500$. "LASSO" and "Bias-corrected" denote that the model section criteria calculated from the naive lasso estimate and the bias-corrected estimate, respectively.

의 성능을 비교하기 위하여 아래에 정의한 참양성율(true positive rate; TPR)을 나타내는 민감도(sensitivity; SEN), 참음성율(true negative rate; TPR)을 나타내는 특이도(specificity; SPE), 위발견율(false discovery rate; FDR), 및 추정 오차의 ℓ_2 노름을 고려한다.

$$\text{SEN} \equiv \frac{\text{TP}}{P}, \qquad \text{SPE} = \frac{\text{TN}}{N}, \qquad \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \quad \text{and} \quad \left\| \hat{\beta}_\lambda - \beta^0 \right\|_2 \tag{4.2}$$

여기서, $\text{TP} = \sum_{j=1}^{p} I(\beta_j^0 \neq 0, \hat{\beta}_{\lambda,j} \neq 0)$, $\text{TN} = \sum_{j=1}^{p} I(\beta_j^0 = 0, \hat{\beta}_{\lambda,j} = 0)$, $P = \sum_{j=1}^{p} I(\beta_j^0 \neq 0)$, $N = \sum_{j=1}^{p} I(\beta_j^0 \neq 0)$, $\text{FP} = N - \text{TN}$이며, $\|x\|_2 = \sqrt{\sum_{j=1}^{p} |x_j|^2}$로 ℓ_2 노름을 나타낸다.

Table 5.1. list of genes and their estimated regression coefficients from ordinary least squares identified by the generalized information criterion with the lasso estimate and the bias-corrected estimate

Bias-corrected		LASSO	
Gene	Coefficient	Gene	Coefficient
DNAJC13	0.2746	DNAJC13	0.2667
DCUN1D1	0.2730	DCUN1D1	0.1689
NMD3	0.2109	NMD3	0.0543
ORMDL3	−0.2241	ORMDL3	−0.0895
		ZNF639	0.2251
		ZNF267	0.1558
		DNAJB4	0.1497

Table 4.1에 각 모형 선택 기준에 따라서 선택된 λ^*, 0이 아닌 회귀 계수 추정값의 수, SEN, SPE, FDR 및 추정 오차의 ℓ_2 노름을 100개의 자료에 대한 평균(average)과 표준오차(standard error)를 요약하였다. 요약된 자료를 보면, 먼저 AIC 기준으로는 $n > p$인 경우에 대하여 편의를 보정할 때, 더 희박한 모형을 선택하며 SPE가 증가, FDR이 감소하는 패턴이 나타났으나 $p \geq n$인 다른 모든 경우에는 모의실험에서 고려한 λ의 최솟값으로만 선택되었다. 이에 반하여 BIC, GIC 및 CV의 모형 선택 기준들은 편의를 보정한 경우, 보정하지 않았을 때 보다 λ^*는 대부분 증가하여 추정량을 상대적으로 희박하게 추정하였으며, SEN, SPE, FDR 및 추정 오차의 ℓ_2 노름 측도들에 대하여 대부분 개선되는 경향을 나타내었다. 여러 모형 선택 기준 중에서 GIC가 차원이 커지는 경우에도 전반적으로 다른 모형 선택 기준들보다 추정 성능이 좋은 모형을 선택하는 경향이 나타났다. 이러한 경향은 이미 Fan과 Tang (2013)에서 설명되었고 기존 논문의 모의 실험에서도 드러난 경향으로 p가 n보다 큰 경우에 대하여 GIC가 다른 모형 선택 기준들 보다 추정 성능이 좋은 모형을 선택하는 경향이 있다.

본 모의실험을 요약하면, LASSO 회귀 모형을 이용한 모형 선택에 있어서 모형 선택 기준들을 적용할 때에 LASSO 회귀 모형의 편의를 보정하여 적용하는 것이 편의를 지닌 기존의 LASSO 추정량을 적용하는 것보다 추정 성능이 더 나은 모형의 선택이 이루어짐을 보일 수 있었다. 특히, 적합에 대한 잔차의 보정으로 인하여 상대적으로 더 희박한 모형을 선택하는 경향을 지님을 Figure 4.1과 Table 4.1을 통하여 확인하였다.

5. 폐암 환자의 유전체 자료를 통한 바이오마커 식별에의 응용

폐암(lung cancer)은 5년 생존율이 약 15%정도로 낮은 생존율을 갖는 심각한 질병이다 (Jemal 등, 2000). 폐암에 대한 치료 및 예후는 암의 진행 단계에 따라 크게 영향을 받기 때문에 조기 진단 및 치료가 매우 중요하다. 최근 암에 대한 유전적 연구가 활발히 이루어지면서 유전적인 바이오마커(biomarker)를 식별하여 질병의 예후 및 예측에 이용하고 있다 (Tang 등, 2013). 본 논문에서는 LASSO 회귀 모형을 Tomida 등 (2009)의 폐암 환자에 대한 유전자 발현량 자료에 적용하여 모형 식별의 편의에 대한 영향을 비교하며 기존에 알려진 바이오마커 PIK3CA (El-Telbany와 Ma, 2012)와의 연관성을 토대로 잠재 바이오마커(potential biomarker)를 식별하고자 한다. 본 연구에서 다루는 Tomida 등 (2009) 자료는 Agilent-014850 Whole Human Genome Microarray 4x44K G4112F 플랫폼을 이용한 117명의 폐암 환자의 유전자 발현량(gene expression level) 자료이다. 표준화를 위한 전처리(preprocessing)로, RMA(robust multiarray average) 알고리즘과 분위수 표준화(quantile normalization)을 적용하였다 (Irizarry 등, 2003).

잠재 바이오 마커의 식별에 앞서, Yu 등 (2015)에서 유전자 조절 네트워크(gene regulatory network)

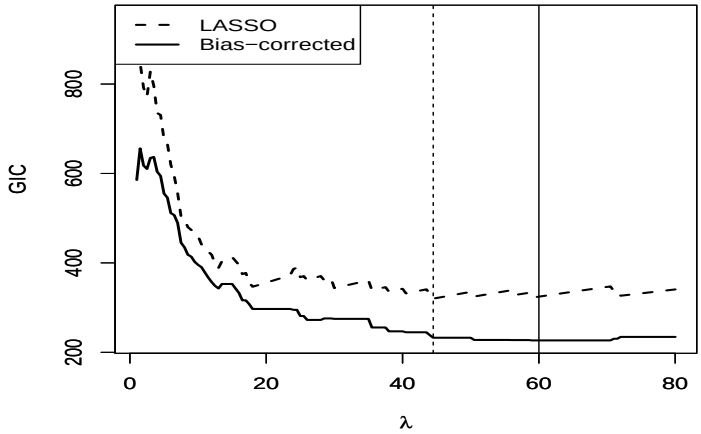


Figure 5.1. Plot of generalized information criterion (GIC) values for the lung cancer data. "LASSO" and "Bias-corrected" denote that the GIC calculated from the naive lasso estimate and the bias-corrected estimate, respectively. Vertical lines at $\lambda = 44.5$ and 60 denote the mininum points of GIC for "LASSO" and "Bias-corrected".

추정을 위해 환자의 생존과 관련된 유전자를 선택하였던 절차를 적용하여 환자의 생존 시간과 연관성이 높은 유전자 군을 식별하여 연구를 진행한다. 이에 따라, 유전자 발현량 자료와 함께 제공된 임상(clinical) 및 생존(survival) 자료를 기반으로 Cox 회귀 모형 (Anderson과 Gill, 1982)과 베타-균일 모형(Beta-Uniform model)을 적용하여 Benjamini와 Hochberg (1995)에 의해 제안 된 위발견율(false discovery rate)을 0.05이하로 조절하였다 (Pounds와 Morris, 2003). 위의 절차를 통하여 1975개의 유전자 군이 식별 되었으며 PIK3CA의 유전자 발현량을 반응 변수로 나머지 1974개의 변수는 설명변수로 LASSO 회귀 모형을 적용하였다. 또한, 각 유전자 발현량은 평균 0, 분산 1을 갖도록 표준화되었다.

모형 선택 기준으로 GIC를 이용한 결과, Figure 5.1에 나타낸 바와 같이 편의를 보정하지 않은 LASSO 추정량을 적용한 경우는 $\lambda = 44.5$에서 GIC값이 최소가 되었으며, 편의를 보정한 경우 $\lambda = 60$에서 GIC가 최소가 되었다. 따라서, 편의를 보정한 경우, 보다 희박한(sparse) 결과를 나타냄을 알 수 있으며, 이는 모의 실험 결과와 일치한다. 각 기준을 적용한 결과에 따라 PIK3CA의 유전자 발현량과 연관성이 높은 유전자 목록은 Table 5.1에 정리하였다. LASSO 추정량을 직접 적용한 GIC 기준으로는 6개의 유전자(DNAJC13, DCUN1D1, NMD3, ORMDL3, ZNF639, ZNF267)가 식별 되었으며, 편의가 보정된 추정량을 적용한 경우는 6개의 유전자 목록과 일치하는 4개의 유전자(DNAJC13, DCUN1D1, NMD3, ORMDL3)가 식별되었다. 이 중에서 DNAJC13, DCUN1D1은 유전자 발현이 암과 관련이 있음이 문헌에 보고 되었으며 (Sterrenberg 등, 2011; Yoo 등, 2012), ORMDL3의 발현은 천식(asthma)과 관련있음이 보고되었다 (Cantero-Recasens 등, 2010). 따라서, 본 연구에 의해 식별된 6개의 유전자들은 기존에 알려진 PIK3CA와 유전자 발현에 대한 연관성이 존재하며, 각각의 유전자들의 발현량을 이용하여 폐암의 예후 및 예측 연구에 이용 될 수 있을 것이라 기대한다. 또한, 모의 실험 결과를 통하여 편의가 보정된 모형 선택의 기준이 더 낮은 위발견율을 지니므로 편의를 보정하여 식별된 4개의 유전자를 폐암에 대한 잠재적인 바이오마커로 제안한다.

6. 결론

본 연구에서는 LASSO 회귀 모형에서 모형 선택 기준들을 적용할 때, 0이 아닌 추정량을 기반으로 계산

이 이루어짐을 이용하여 LASSO 추정량의 편의가 모형 선택 기준들에 어떠한 영향을 미치는치 모의실험을 통하여 살펴보았다. 또한, 0이 아닌 추정량의 수가 표본의 수보다 작은 경우에는 LASSO 추정량의 부호와 0이 아닌 성분에 대응하는 설명 변수를 이용하여 PCG 알고리즘을 활용하여 효율적으로 편의를 보정하는 방법을 제안하였다. 모의실험의 결과를 토대로 편의를 보정하여 모형 선택 기준들을 적용하는 것이 보다 나은 모형 선택의 결과를 주는 것을 알 수 있었으며, 이를 토대로 폐암 환자의 유전자 발현량 자료에 적용하여 잠재적인 바이오마커를 식별하였다.

본 연구에서 0이 아닌 LASSO 추정량에 대응하는 $\mathbf{X}_I^T\mathbf{X}_I$의 역행렬을 기반으로 편의를 보정하여 0이 아닌 추정량의 수가 표본의 수보다 큰 경우에는 편의를 보정하는데 어려움이 있다. 후속 연구로 $\mathbf{X}_I^T\mathbf{X}_I$의 일반화 역행렬(generalized inverse matrix) 또는 Paige와 Saunders (1975)의 최소잔차법(minimal residual method; MINRES) 방법 등을 활용하는 편의를 보정하는 방법을 연구하고자 한다.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceeding of the Second International Symposium on Information Theory*, Budapest: Akademiai Kiado.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes, a large sample study, *Annals of Statistics*, **10**, 1100–1120.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models, *Bernoulli*, **19**, 521–547.

Cantero-Recasens, G., Fandos, C., Rubio-Moscardo, F., Valverde, M. A., and Vicente, R. (2010). The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress, *Human Molecular Genetics*, **19**, 111–121.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes, *Journal of the Royal Statistical Society. Series B (Methodological)*, **76**, 373–397.

Demmel, J. W. (1997). *Applied Numerical Linear Algebra*, SIAM, Philadelphia.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics*, **32**, 407–499.

El-Telbany, A. and Ma, P. C. (2012). Cancer genes in lung cancer: racial disparities: are there any?, *Genes & Cancer*, **3**, 467–480.

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.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society. Series B (Methodological)*, **75**, 531–552.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization, *Annals of Applied Statistics*, **1**, 302–332.

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

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, **31**, e15.

Jemal, A., Siegel, R., Xu, J., and Ward, E. (2010). Cancer statistics, *CA: A Cancer Journal for Clinicians*, **60**, 277–300.

Khare, K., Oh, S.-Y., and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees, *Journal of the Royal Statistical Society. Series B (Methodological)*, **77**, 803–825.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso,

Annals of Statistics, **34**, 1436–1462.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Annals of Statistics*, **12**, 758–765.

Paige, C. C. and Saunders, M. A. (1975). Solution of sparse indefinite systems of linear equations, *SIAM Journal on Numerical Analysis*, **12**, 617–629.

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.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models, *Journal of the American Statistical Association*, **79**, 575–583.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values, *Bioinformatics*, **19**, 1236–1242.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221–264.

Sterrenberg, J. N., Blatch, G. L., and Edkins, A. L. (2011). Human DNAJ in cancer and stem cells, *Cancer Letters*, **312**, 129–142.

Tang, H., Xiao, G., Behrens, C., Schiller, J., Allen, J., Chow, C. W., Suraokar, M., Corvalan, A., Mao, J., White, M. A., Wistuba, I., Minna, J. D., and Xie, Y. (2013). A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients, *Clinical Cancer Research*, **19**, 1577–1586.

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness, *Electronic Journal of Statistics*, **7**, 1456–1490.

Tomida, S., Takeuchi, T., Shimada, Y., Arima, C., Matsuo, K., Mitsudomi, T., Yatabe, Y., and Takahashi, T. (2009). Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis, *Journal of Clinical Oncology*, **27**, 2793–2799.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society. Series B (Methodological)*, **71**, 671–683.

Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553–568.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression, *Journal of Multivariate Analysis*, **102**, 1141–1151.

Yang, Y. (2005). Can the strengths of aic and bic be shared?: a conflict between model identification and regression estimation, *Biometrika*, **92**, 937–950.

Yoo, J., Lee, S.-H., Lym, K., Park, S. Y., Yang, S.-H., Yoo, C.-Y., Jung, J.-H., Kang, S.-J., and Kang, C.-S. (2012). Immunohistochemical expression of DCUN1D1 in non-small cell lung carcinoma: its relation to brain metastasis, *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, **44**, 57–62.

Yu, D., Son, W., Lim, J., and Xiao, G. (2015). Statistical completion of partially identified graph with application to estimation of gene regulatory network, *Biostatistics*, **16**, 670–685.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, **38**, 894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression, *Annals of Statistics*, **36**, 1567–1594.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

모형 선택 기준들에 대한 LASSO 회귀 모형 편의의 영향 연구

유동현[a,1]

[a]계명대학교 통계학과

(2016년 3월 2일 접수, 2016년 4월 26일 수정, 2016년 4월 28일 채택)

요 약

고차원 자료(high dimensional data)는 변수의 수가 표본의 수보다 많은 자료로 다양한 분야에서 관측 또는 생성되고 있다. 일반적으로, 고차원 자료에 대한 회귀 모형에서는 모수의 추정과 과적합을 피하기 위하여 변수 선택이 이루어진다. 벌점화 회귀 모형(penalized regression model)은 변수 선택과 회귀 계수의 추정을 동시에 수행하는 장점으로 인하여 고차원 자료에 빈번하게 적용되고 있다. 하지만, 벌점화 회귀 모형에서도 여전히 조율 모수 선택(tuning parameter selection)을 통한 최적의 모형 선택이 요구된다. 본 논문에서는 벌점화 회귀 모형 중에서 대표적인 LASSO 회귀 모형을 기반으로 모형 선택의 기준들에 대한 LASSO 회귀 추정량의 편의가 어떠한 영향을 미치는지 모의실험을 통하여 수치적으로 연구하였고 편의의 보정의 필요성에 대하여 나타내었다. 실제 자료 분석에서의 영향을 나타내기 위하여, 폐암 환자의 유전자 발현량(gene expression) 자료를 기반으로 바이오마커 식별(biomarker identification) 문제에 적용하였다.

주요용0어: LASSO, 벌점화 회귀 모형, 편의, 모형 선택, 정보 기준

이 연구는 2015년도 계명대학교 비사연구기금으로 이루어졌음.

[1](42601) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과. E-mail: dyu3@kmu.ac.kr