The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.199
32(2), 199–211

Joint penalization of components and predictors in mixture of regressions

Chongsun Park[a,1] · Eun Bi Mo[a]

Department of Statistics, Sungkyunkwan University

(Received November 15, 2018; Revised December 24, 2018; Accepted January 17, 2019)

Abstract

This paper is concerned with issues in the finite mixture of regression modeling as well as the simultaneous selection of the number of mixing components and relevant predictors. We propose a penalized likelihood method for both mixture components and regression coefficients that enable the simultaneous identification of significant variables and the determination of important mixture components in mixture of regression models. To avoid over-fitting and bias problems, we applied smoothly clipped absolute deviation (SCAD) penalties on the logarithm of component probabilities suggested by Huang *et al.* (*Statistical Sinica*, **27**, 147–169, 2013) as well as several well-known penalty functions for coefficients in regression models. Simulation studies reveal that our method is satisfactory with well-known penalties such as SCAD, MCP, and adaptive lasso.

Keywords: mixture regression, component, variable selection, penalty

1. 서론

Pearson (1894)에 의해 처음 제안된 유한혼합모형(finite mixture model)은 사전에 알려진 수의 모집단으로부터 얻어진 표본을 모형화하는 방법으로 생물학, 유전학, 마케팅 등 여러 분야에 사용되고 있다. 유한혼합회귀모형(finite mixture of regression; FMR)은 유한혼합모형의 관측치를 설명하는 예측변수들(predictors)을 포함하며 금융과 사회과학의 여러 분야에 적용 (Wedel과 Kamukura, 2000; Skrondal과 Rabe-Hesketh, 2004)되고 있다. 각각의 모집단에 포함될 확률을 포함하는 모수들의 추정에는 EM-알고리즘 (Dempster 등, 1977)과 Markov chain Monte Carlo (MCMC) 방법 (Metropolis 등, 1953; Hastings, 1970)이 있으며 이들을 포함하는 더 광범위한 알고리즘 (McLachlan과 Peel, 2000)이 존재한다.

Chen (1995)에 의하면 유한혼합모형에서 성분의 수를 모르는 경우 최적 수렴비가 느려지는 것으로 알려져 있으며 과도하게 많은 수의 성분을 사용하면 자료의 과적합(overfitting)에 따라 해석이 어려워질 수 있다. 반대의 경우에는 실제 내재하는 자료의 구조를 적절하게 파악하지 못하게 되며 이는 유한혼합회귀모형에서도 동일하게 적용된다. 유한혼합모형에서 성분의 수를 결정하는 다양한 검정방법 (McLachlan과 Peel, 2000)이 있었으나 유한혼합모형의 성분 수의 결정과 변수 선택의 방법으로 주로

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-Ro, Jongno-Gu, Seoul 03063, Korea. E-mail: cspark@skku.edu

전통적인 Akaike information criterion (AIC) (Akaike, 1973)와 Bayes information criterion (BIC) (Schwarz, 1978) 방법 등이 사용되었다. 그러나 이러한 방법들 또한 주어진 자료로부터 가능한 모든 후보 모형을 고려하여 최적 모형을 찾아주기 때문에 고차원(high dimension) 자료의 경우에 많은 연산 시간과 비용을 소요하는 문제가 있다. 이러한 문제를 피하는 방법으로 벌점함수를 적용하는 방법이 있다. Chen과 Khalili (2008)는 혼합확률의 과적합을 방지하기 위해 혼합확률과 위치모수의 차이값에 벌점함수를 동시에 적용하여 AIC, BIC를 사용하는 것보다 성능이 우수함을 보였다. Huang 등 (2017)은 이를 유한정규혼합모형으로 확장하여 혼합 모형의 성분에 변형된 LASSO 벌점함수를 적용하여 BIC방법보다 계산 성능이 우수하며 정규혼합모형에서 일관된 성분의 수를 선택할 수 있음을 증명하였다. 이들은 성분의 확률에 직접 벌점함수를 적용하는 경우 EM-알고리즘의 완전로그가능도함수(complete-data log-likelihood function)가 성분확률의 로그값을 포함하여 확률값이 0에 근접한 경우 이에 대한 기울기(gradient)가 빠르게 증가하는 문제점을 해결하기 위하여 성분확률의 로그값에 벌점함수를 적용하였다.

회귀모형에서의 변수선택에 벌점함수를 적용하는 대표적인 방법들로는 Tibshirani (1996)의 least absolute shrinkage and selection operator (LASSO)와 Fan과 Li (2001, 2002)가 제시한 smoothly clipped absolute deviation (SCAD) 벌점함수를 들 수 있다. 이와 더불어 계수마다 다른 가중치를 주어 LASSO 방법의 편의를 개선한 Adplasso(adaptive LASSO) 방법 (Zou, 2006)과 Zhang (2010)이 제시한 MCP (minimax concave penalty) 방법들이 널리 사용되고 있다. 유한혼합회귀모형에서 변수선택의 연구로 Khalili와 Chen (2007)은 정규혼합회귀모형의 변수 선택에 다양한 벌점 함수를 적용하여 기존의 BIC 방법보다 더 효율적임을 보였다.

혼합회귀모형에서 성분의 개수선택과 회귀계수의 선택에 동시에 벌점함수를 적용하려는 시도의 하나로 Luo 등 (2008)은 유한혼합회귀모형에서 성분의 수와 변수 선택에 벌점 함수를 적용하였다. 이 연구에서 회귀 계수 추정에 LASSO 벌점함수를 적용하고, 혼합모형의 성분 수 추정에는 추정된 회귀계수들 간의 L_2-norm 거리에 벌점함수를 적용하는 방법을 시도하였으나 성분의 수와 회귀계수가 과적합 되는 경우가 빈번하게 발생하는 것으로 나타났다.

본 연구에서는 성분과 회귀계수에 직접 벌점함수를 적용하여 적절한 성분의 수와 모형에 필요한 회귀계수들을 동시에 선택하는 방법을 제시하였다. 성분에 대한 벌점은 성분들의 로그값에 SCAD 벌점을 주는 Huang 등 (2017)의 방법을 적용하였고 회귀계수들에는 SCAD 이외에 다양한 벌점함수들을 적용하여 가상자료와 실제자료에 대하여 비교하였다. 새로운 방법은 성분의 수에 직접 벌점함수를 적용하여 Luo 등 (2008)의 방법의 문제점인 과적합을 해결할 수 있었으며 회귀계수에 적절한 벌점함수를 사용하면 계수 추정치의 편의도 크지 않은 것으로 나타났다.

2. 혼합선형회귀모형(mixture of linear regression model)

회귀모형에서 반응변수와 이에 해당하는 p개의 예측변수들의 벡터를 y와 $\mathbf{x} = (x_1, \ldots, x_p)^T$라 두고 n개의 관측치들 $\mathbf{y} = (y_1, \ldots, y_n)^T$과 이들 각각에 해당하는 예측변수들의 짝 $(\mathbf{x}_i^T, y_i)$ $(i = 1, \ldots, n)$이 주어졌다고 가정하자. 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$이다. 관측치 y_i가 K개의 성분 중 하나인 k에 속한다면 y_i와 이에 해당하는 예측변수 $\mathbf{x}_i$는

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

인 선형모형을 따르며 $\epsilon_i \sim N(0, \sigma_k^2)$라고 가정하자. 이때 $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{pk})$는 성분 k에 속하는 i번째 관측치에 대한 p차원의 회귀계수들이며 σ_k^2은 오차항의 분산이다. K개의 성분들에 대한 혼합확률을

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$라 하고 모든 모수들을 결합하여 $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1, \ldots, \sigma_K, \alpha_1, \ldots, \alpha_K)$라고 두면 i번째 관측치가 성분 k에 속하는 경우 y_i의 주변분포는

$$f(y_i; \mathbf{x}_i, \boldsymbol{\psi}) = \sum_{k=1}^{K} \alpha_k (2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_k^2} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k \right)^2 \right\}$$

이 된다. 혼합확률들의 합은 1이며 모두 양수이다.

EM 알고리즘의 적용을 위하여 일반적인 혼합모형과 같이 i번째의 자료 $(\mathbf{x}_i, y_i)$에 아래와 같은 잠재변수인 지시변수

$$Z_{ik} = \begin{cases} 1, & \text{if } y_i|\mathbf{x}_i \text{ comes from component } k, \\ 0, & \text{otherwise} \end{cases}$$

를 고려하면 관찰된 자료들의 로그가능도함수는

$$l_o(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k (2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_k^2} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k \right)^2 \right\}$$

이 되고 로그완전가능도함수는

$$l_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \left[\alpha_k (2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_k^2} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k \right)^2 \right\} \right]$$

이 되며 이들의 최적화에는 제약식을 포함하고 있는 문제에 적합한 알고리즘이 필요하다.

3. 콤포넌트와 회귀계수에 대한 벌점함수의 적용

일반적 선형회귀모형에서 변수선택 방법에는 전진선택(forward selection), 후진제거(backward elimination), 단계적회귀(stepwise regression), 능형회귀(ridge regression) 등의 방법이 알려져 있다. 하지만 이 방법들은 자료가 고차원인 경우에 안정성 문제가 발생 (Breiman, 1996)하며 AIC와 BIC 등을 이용하는 경우 독립변수가 증가할수록 계산 비용이 많이 드는 문제점이 생긴다 (Khalili와 Chen, 2007). 이와 같은 문제를 해결하기 위해 제안된 방법으로 벌점화 회귀모형을 들 수 있다.

3.1. 벌점화 가능도함수

회귀모형에 적용할 수 있는 대표적인 벌점함수 $p_\lambda(\beta)$들은 다음과 같다.

- LASSO: $\lambda|\beta|$, $(\lambda > 0)$
- Hard: $\lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, $(\lambda > 0)$
- Adplasso: $\lambda w|\beta|$ with $w > 0$, $(\lambda > 0)$
- SCAD: $\begin{cases} \lambda|\beta|, & \text{if } |\beta| \le \lambda, \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |\beta| \le a\lambda, \ (\lambda > 0, a > 2), \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } |\beta| > a\lambda, \end{cases}$

- MCP: $\begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2a}, & \text{if } |\beta| \le a\lambda, \\ \dfrac{1}{2}a\lambda^2, & \text{if } |\beta| > a\lambda, \ (\lambda > 0, a > 1) \end{cases}$

Tibshirani (1996)에 의해 제안된 LASSO 방법은 편의의 영향으로 일치성이 만족되지 않는 것으로 알려져 있다 (Zou, 2006). 이를 보완하기 위한 방법으로 추정된 각 회귀계수에 각각 다른 가중치를 적용하는 Adplasso 방법이 있다. Fan과 Li (2001)의 SCAD 벌점함수는 가장 안정적인 방법으로 잘 알려져 있다. MCP 방법은 Zhang (2010)에 의해 제안됐으며, 고차원 자료에서 빠르고 연속적이며 비교적 편향되지 않은 정확한 변수선택 방법으로 알려져 있다. SCAD와 MCP 벌점함수에 포함되어 있는 상수 a는 모두 3.7로 고정된 값을 사용하였다.

Fan과 Li (2001)는 벌점을 부여하여 추정된 회귀계수가 좋은 추정량이 되기 위해 만족해야 하는 성질들로 불편성(unbiasedness), 희박성(sparsity), 연속성(continuity)을 제시하였다. LASSO는 불편성, HARD는 연속성을 만족하지 못하나, Adplasso, SCAD, MCP의 경우에는 희박성과 연속성을 만족하며 계수 추정치가 어느 정도 큰 경우 불편성도 만족한다. 성분에 대한 벌점함수는 Huang 등 (2017)의 방법을 적용하고 회귀계수들에는 Khalili와 Chen (2007)의 벌점함수들을 적용한 가능도함수는 다음과 같다.

$$l_P(\boldsymbol{\psi}) = l_c(\boldsymbol{\psi}) - nD_f \sum_{k=1}^{K} [\log(\epsilon + p_{\lambda_1}(\alpha_k)) - \log(\epsilon)] - \sum_{k=1}^{K} \alpha_k \sum_{j=1}^{p} p_{\lambda_2}(\sqrt{n}\beta_{jk}) \tag{3.1}$$

여기서 λ_1은 성분에 대한 벌점함수의 계수이고 λ_2는 회귀계수에 대한 벌점함수의 계수이며 D_f는 각 성분에서 자유로운 모수의 수이다. 성분에 대한 벌점함수에 포함된 ϵ값은 10^{-5}를 사용하였다.

3.2. 수정된 EM 알고리즘

이 절에서는 전통적인 EM 알고리즘을 M-step에 적용한 수정된 EM 알고리즘을 소개한다.

- E-STEP: 반복 t $(t = 0, 1, \ldots)$에서 식 (3.1)에 대한 기댓값

$$\begin{aligned} Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}^{(t)} \log \alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}^{(t)} \log \left\{ \phi\left(y_i; \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2\right) \right\} \\ &- n\sum_{k=1}^{K} [\log(\epsilon + p_{\lambda_1}(\alpha_k)) - \log(\epsilon)] - \sum_{j=1}^{p} p_{\lambda_2}(\sqrt{n}\beta_{jk})) \end{aligned}$$

  이 되며

$$Z_{ik}^{(t)} = \frac{\alpha_k^{(t)} \phi\left(y_i; \mathbf{x}_i, \boldsymbol{\beta}_k^{(t)}, \sigma_k^{2(t)}\right)}{\sum_{k=1}^{K} \alpha_k^{(t)} \phi\left(y_i; \mathbf{x}_i, \boldsymbol{\beta}_k^{(t)}, \sigma_k^{2(t)}\right)}$$

  는 Z_{ik}의 조건부 기댓값이다.

- M-STEP: M-step에서는 반복 $t+1$에서 $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)})$를 최대화하는 $\boldsymbol{\psi}$를 구하게 되는데 우선 α_k의 추정값을 구하는 문제는 라그랑지 승수법(Lagrange multiplier method)를 이용하여 아래의 해를 구하는 것과 같다.

$$\frac{\partial}{\partial \alpha_k}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}^{(t)} \log \alpha_k - n\sum_{k=1}^{K}(\log(\epsilon + p_{\lambda_1}(\alpha_k)) - \Gamma\left(\sum_{k=1}^{K}\alpha_k - 1\right).\right] = 0$$

위에서 $\log(\epsilon+p_{\lambda_1}(\alpha_k))$를 t에서의 선형 근사인 $\log(\epsilon+p_{\lambda_1}(\alpha_k)) \approx \log(\epsilon+p_{\lambda_1}(\alpha_k^{(t)})+[p'_{\lambda_1}(\alpha_k^{(t)})/\{\epsilon+p_{\lambda_1}(\alpha_k^{(t)})\}](\alpha_k-\alpha_k^{(t)})$로 대체하고 미분하면 다음과 같은 $\alpha_k^{(t+1)}$를 구할 수 있다.

$$\alpha_k^{(t+1)} = \frac{1}{D_k}\sum_{i=1}^{n} Z_{ik}^{(t)}$$

여기서 $D_k = n + nD_f[p'_{\lambda_1}(\alpha_k^{(t)})/\{\epsilon+p_{\lambda_1}(\alpha_k^{(t)})\}] - nD_f\sum_{k=1}^{K}[p'_{\lambda_2}(\alpha_k^{(t)})\alpha_k^{(t)}/\{\epsilon+p_{\lambda_1}(\alpha_k^{(t)})\}]$이며 $p'_{\lambda_1}(\cdot)$은 벌점함수 $p_{\lambda_1}(\cdot)$의 α_k에 대한 1차 도함수이다. $\alpha_k^{(t+1)}$는 0이 될 수 없으므로 시뮬레이션과 실제자료분석에서는 추정값이 10^{-3}보다 작은 값을 갖게 되면 0으로 두고 해당 성분을 모형에서 제거하였다.

이제 남은 모수들인 $\boldsymbol{\beta}_k$ 및 σ_k들을 반복법(iterative method)인 뉴튼-라프슨 방법을 통하여 갱신하기 위하여 $\boldsymbol{\beta}_k$ 및 σ_k들의 반복 t $(t=1,\ldots)$에서의 값을 $\boldsymbol{\beta}_k^{(t)}$, $\alpha_k^{(t)}$라 두고 식 (3.1)을 Fan과 Li (2001)의 LQA 방법을 벌점함수 $p(\cdot)$에 적용하여 다시 표현하면

$$\begin{aligned} l_P(\boldsymbol{\psi}) = l_c(\boldsymbol{\psi}) &- n\sum_{k=1}^{K}\left[\log\left(\epsilon+p_{\lambda_1}(\alpha_k)\right)-\log(\epsilon)\right] \\ &- \sum_{k=1}^{K}\alpha_k\sum_{j=1}^{p}\left[p_{\lambda_2}\left(\sqrt{n}\beta_{jk}^{(t)}\right)+\frac{p'_{\lambda_2}\left(\sqrt{n}\beta_{jk}^{(t)}\right)}{2\beta_{jk}^{(t)}}\left(\beta_{jk}^2-\beta_{jk}^{2\ (t)}\right)\right] \end{aligned}$$

이 된다. 이렇게 다시 표현한 벌점화 가능도 함수를 이용하여 β_{jk}의 추정값은 순차적인 갱신을 통하여 다음을 만족하는 해가 된다.

$$\sum_{i=1}^{n} Z_{ik}^{(t)}\frac{\partial}{\partial\beta_{jk}}\left[\log\phi\left(y_i;\mathbf{x}_i,\boldsymbol{\beta}_k,\sigma_k^{2(t)}\right)\right]-\alpha_k\left(\frac{\partial}{\partial\beta_{jk}}\left[p_{\lambda_2}\left(\sqrt{n}\beta_{jk}^{(t)}\right)+\frac{p'_{\lambda_2}\left(\sqrt{n}\beta_{jk}^{(t)}\right)}{2\beta_{jk}^{(t)}}\left(\beta_{jk}^2-\beta_{jk}^{2(t)}\right)\right]\right)=0.$$

이 식은 $\beta_{jk}^{(t+1)}$이 0이 아닌 추정값에 대해 근사적으로 만족하고 추정값이 0인 경우에는 만족하지 못하므로 $\beta_{jk}^{(t+1)}$가 0인 것을 식별할 수 있다 (Khalili와 Chen, 2007). 또한, 아래 식을 만족하는 σ_k^2의 추정값은

$$\sum_{i=1}^{n} Z_{ik}^{(t)}\frac{\partial}{\partial\sigma_k^2}\left[\log\phi\left(y_i;\mathbf{x}_i,\boldsymbol{\beta}_k^{(t+1)},\sigma_k^2\right)\right]=0$$

의 해가 되며

$$\hat{\sigma}_k^{2(t+1)} = \frac{\sum_{i=1}^{n} Z_{ik}^{(t)}\left(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_k^{(t+1)}\right)^2}{\sum_{i=1}^{n} Z_{ik}^{(t)}}$$

이 된다.

3.3. 조율모수(tuning parameter)의 선택

위에서 제시한 모형에는 성분의 수에 대한 벌점함수와 회귀계수들에 대한 벌점함수에 두 종류의 조율모수 λ_1과 λ_2가 포함되어 있으며 이들 모수의 적절한 선택은 매우 중요하다. 본 연구에서는 과적합을 방지하기 위하여 BIC를 최적화하는 조율모수들을 선택하였다. BIC는

$$\mathrm{BIC}(\lambda_1,\lambda_2) = \sum_{i=1}^{n}\log\left\{\phi\left(y_i;\mathbf{x}_i,\hat{\boldsymbol{\beta}}_k,\hat{\sigma}_k^2\right)\right\}-\frac{r}{2}\log(n)$$

Table 4.1. Frequencies and percents of number of components for $\rho = 0.5$ and $K = 3$

Component	SCAD-SCAD		SCAD-Adplasso		SCAD-MCP	
	$n = 300$	$n = 500$	$n = 300$	$n = 500$	$n = 300$	$n = 500$
2	4(0.8%)	1(0.2%)	18(3.6%)	5(1.0%)	11(2.2%)	0(0.0%)
3	435(87.0%)	476(95.2%)	461(92.2%)	478(95.6%)	413(82.6%)	473(94.6%)
4	58(11.6%)	23(4.6%)	20(4.0%)	17(3.4%)	67(13.4%)	25(5.0%)
5	3(0.6%)	0(0.0%)	1(0.2%)	0(0.0%)	6(1.2%)	1(0.2%)
6	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	2(0.4%)	1(0.2%)
7	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	1(0.2%)	0(0.0%)
Total	500(100%)	500(100%)	500(100%)	500(100%)	500(100%)	500(100%)

이며 이를 최대화하는 조율모수들을 선택하게 된다. 여기서 r은 0이 아닌 값으로 추정한 모수들의 수가 된다.

4. 모의실험

모의실험에서는 성분이 3개이며 각 성분의 확률 $\boldsymbol{\alpha} = (0.5, 0.3, 0.2)$인 경우를 고려하였다. 모든 성분에 6개씩의 예측변수가 포함되어 있고 그 값들은 모두 표준정규분포에서 생성되었다. 성분1과 2에서는 예측변수들에 대한 회귀계수들 중 3개의 값을 0으로 두었으며 성분3은 2개의 값을 0으로 두었다. 각각의 성분에 대한 회귀계수들은 $\boldsymbol{\beta}_1 = (3, 1.5, 0, 0, 2, 0)$, $\boldsymbol{\beta}_2 = (2, -1, 1.5, 0, 0, 0)$, $\boldsymbol{\beta}_3 = (-2, 1, 0, 0, 1, 0.7)$이고 예측변수들의 상관관계는 상관계수가 0.5와 0.8인 AR(1)모형을 고려하였다. 오차는 평균이 0이고 모든 성분의 분산이 0.5인 정규분포에서 생성하였으며 관측치의 수 n은 300과 500인 두 경우를 고려하였다. 추정시 초기 성분의 수를 10으로 두었고 LQA 알고리즘에서 성분확률들의 초기값은 모두 1/10로 동일하게 주었다. 회귀계수들의 초기값은 참값을 평균으로 갖는 정규분포에서 난수를 추출하여 생성하였고 분산의 경우에는 참값을 평균으로 갖는 균일분포를 이용하였다. adplasso방법에서 w도 회귀계수에 대한 초기값을 생성하는 방법과 같은 방법을 적용하여 초기값을 부여하였다.

성분에 대한 벌점함수는 SCAD만을 고려하였으나 회귀계수들에 대한 벌점함수는 SCAD, adplasso, MCP를 적용하여 결과들을 비교하였다. 모의실험은 각각의 경우에 500번씩 반복하여 추정된 성분의 수와 회귀계수들에 대하여 참값이 0인 경우 이들을 0으로 정확하게 추정한 개수의 평균인 C(Correct)와 참값이 0이 아닌 경우 이들을 잘못 추정한 개수의 평균인 IC(InCorrect)를 살펴보았다. 추정된 성분의 확률들과 회귀계수들은 상자그림을 통하여 편의와 변동을 시각적으로 확인하였다.

Table 4.1에는 추정된 성분의 수의 빈도표와 백분율을 포함하였다. SCAD-SCAD를 제외하면 성분의 수를 정확히 추정한 비율이 모두 80%를 넘었으며 SCAD-Adplasso의 조합이 가장 높았다. 모든 경우 성분의 개수를 2로 과소 추정하는 경우보다 4로 과대추정하는 빈도가 높았으며 표본수가 300인 경우 SCAD-MCP 조합에서 상대적으로 추정된 성분의 수가 5 이상으로 나타나는 경우가 많았다. SCAD-MCP 방법에서도 표본수가 300인 경우 성분의 개수를 3으로 추정한 경우가 82.6% 나타나 다른 방법에 비해 정확도가 떨어지는 것으로 보였다.

예측변수들의 상관계수가 0.5인 경우 조건별 C, IC에 대한 결과는 Table 4.2에 포함되어 있다. 성분의 경우 C, IC는 $(7, 0)$을 회귀계수의 경우에는 각각의 성분에 대하여 $(3, 3, 2)$, $(0, 0, 0)$이 참값이다. 성분의 경우를 먼저 살펴보면 C의 경우 SCAD-Adplasso 방법에서 7에 가장 가까운 값을 갖는 것으로 나타났으나 IC값의 경우에는 SCAD나 MCP 방법에 비하여 0과의 차이가 컸다. Huang 등 (2017)이 이미 혼합모형에서 BIC, AIC를 이용하여 성분의 수를 탐색하는 경우 정확히 추정하는 비율이 벌점함수를 사

Table 4.2. C and IC of α and β ($\rho = 0.5$)

n	Mixing penalty (λ_1)	Coef. penalty (λ_2)	Avg. number of C and IC for α		Avg. number of C and IC for β					
					C			IC		
			C	IC	Comp.1	Comp.2	Comp.3	Comp.1	Comp.2	Comp.3
300	SCAD	SCAD	6.872	0.008	2.968	2.848	1.677	0.012	0.090	0.417
		Adplasso	6.956	0.036	2.993	2.928	1.911	0.017	0.704	1.496
		MCP	6.822	0.022	2.961	2.782	1.648	0.000	0.085	0.272
500	SCAD	SCAD	6.954	0.002	2.989	2.973	1.977	0.000	0.013	0.000
		Adplasso	6.966	0.010	3.000	2.969	1.922	0.004	0.449	0.960
		MCP	6.940	0.000	2.992	2.939	1.850	0.002	0.011	0.142

Table 4.3. Frequencies and percents of number of components for $\rho = 0.8$ and $K = 3$

Component	SCAD-SCAD		SCAD-Adplasso		SCAD-MCP	
	$n = 300$	$n = 500$	$n = 300$	$n = 500$	$n = 300$	$n = 500$
2	15(3.0%)	1(0.2%)	19(3.8%)	2(0.4%)	23(4.6%)	1(0.2%)
3	443(88.6%)	481(96.2%)	444(88.8%)	469(93.8%)	392(78.4%)	446(89.2%)
4	42(8.4%)	18(3.6%)	36(7.2%)	29(5.8%)	76(15.2%)	48(9.6%)
5	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	9(1.8%)	5(1.0%)
6	0(0.0%)	0(0.0%)	1(0.2%)	0(0.0%)	0(0.0%)	0(0.0%)
7	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)
Total	500(100%)	500(100%)	500(100%)	500(100%)	500(100%)	500(100%)

Table 4.4. C and IC of α and β ($\rho = 0.8$)

n	Mixing penalty (λ_1)	Coef. penalty (λ_2)	Avg. number of C and IC for α		Avg. number of C and IC for β					
					C			IC		
			C	IC	Comp.1	Comp.2	Comp.3	Comp.1	Comp.2	Comp.3
300	SCAD	SCAD	6.916	0.030	2.910	2.783	1.686	0.000	0.485	0.998
		Adplasso	6.922	0.038	2.977	2.862	1.849	0.054	1.287	2.321
		MCP	6.812	0.046	2.920	2.599	1.442	0.010	0.332	0.656
500	SCAD	SCAD	6.964	0.002	2.965	2.911	1.784	0.000	0.189	0.351
		Adplasso	6.942	0.004	2.994	2.931	1.910	0.017	0.705	1.498
		MCP	6.884	0.002	2.964	2.824	1.615	0.002	0.108	0.252

용하는 방법보다 많이 떨어짐을 보여 본 모의실험에서는 포함하지 않았다.

다음으로 변수의 경우 C, IC를 동시에 고려하였을 때 SCAD-MCP 방법이 가장 우수한 것으로 판단된다. SCAD-Adplasso의 경우 C값들은 $(3, 3, 2)$에 가장 가깝지만 IC값은 $(0.017, 0.704, 1.496)$으로 3번째 성분에선 0이 아닌 값을 0으로 추정하는 경우가 매우 많이 나타났다. 전체적으로 n이 증가할수록 C, IC 값이 안정적으로 추정되는 것으로 보이고, 표본의 크기가 커질수록 SCAD-SCAD 방법이 정확도가 가장 높은 것으로 보였다.

Table 4.3과 Table 4.4는 예측변수들의 자기상관계수가 0.8인 경우에 대한 결과들이다. 성분의 경우에는 SCAD-SCAD 조합이 가장 높은 정확도를 보였으며 SCAD-MCP 조합의 경우 성분의 수를 3으로 추정한 비율이 표본수 300에서는 80%에 그리고 표본수 500에서는 90%에 미치지 못하여 다른 두 방법에 비하여 결과가 좋지 않았으며 과소 및 과대추정하는 비율도 높았다. 각 성분별 변수 선택의 경우 C,

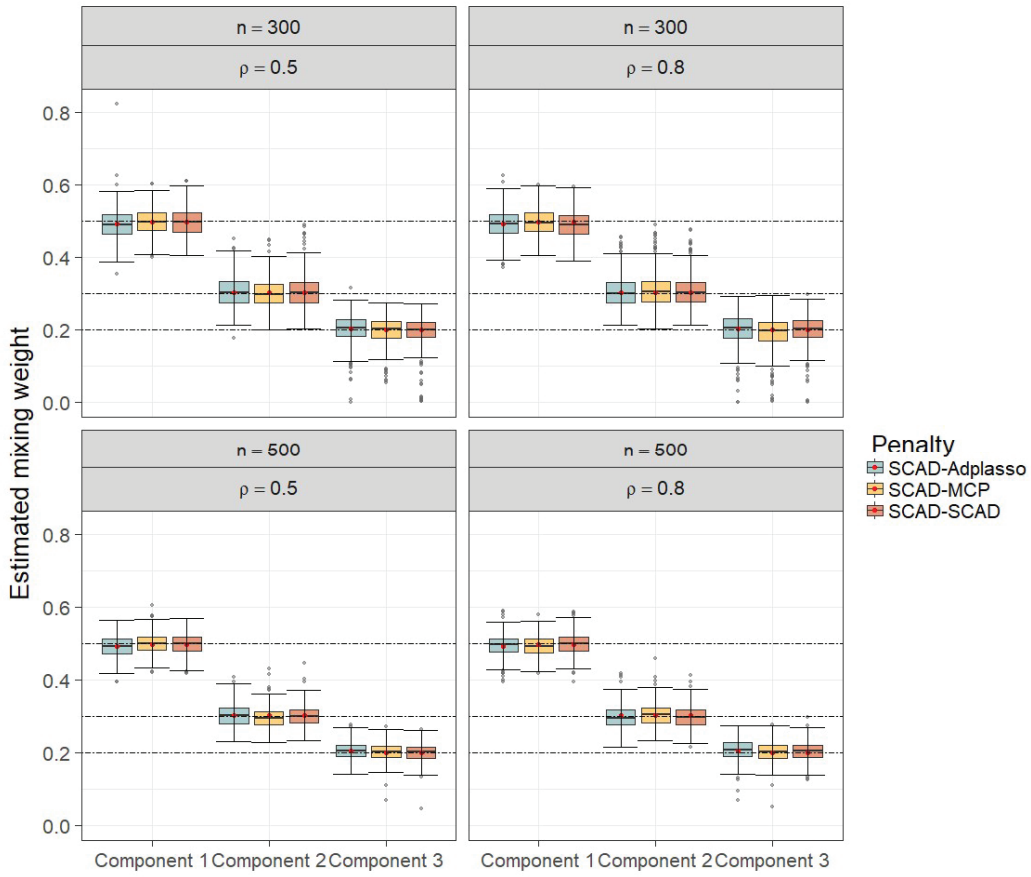
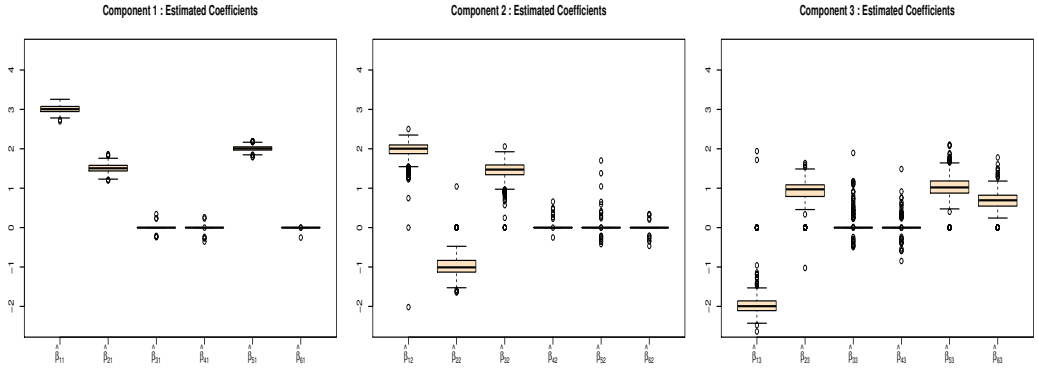


Figure 4.1. Boxplot of estimated α.

IC의 결과를 보면 C의 경우 벌점함수의 조합들 간에 큰 차이는 없었으나 MCP 방법의 결과가 가장 좋지 않았다. IC의 경우에는 Adplasso 방법이 다른 두 벌점함수들에 비하여 상대적으로 큰 값을 보였다. 전체적으로 C, IC를 동시에 고려하였을 때 SCAD-SCAD 방법이 정확도가 가장 높은 것으로 나타났다.

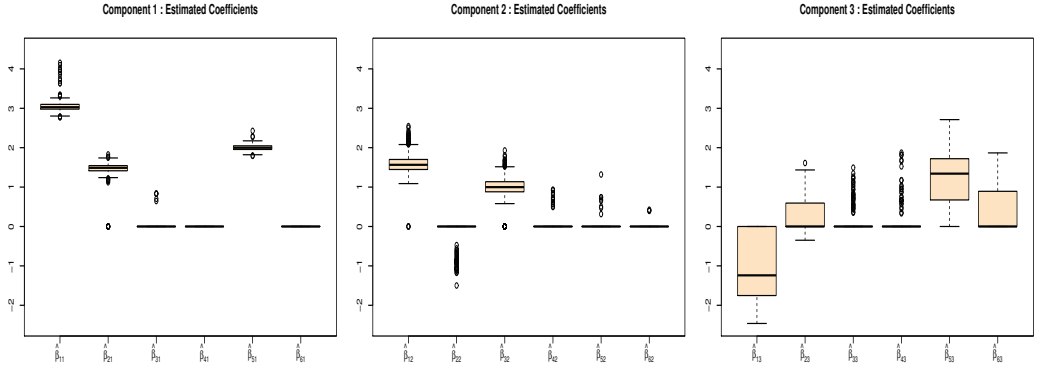
Figure 4.1에는 표본수와 자기상관계수의 조합들 각각에 대하여 500번 반복한 각 성분의 확률(참값은 (0.5, 0.3, 0.2))에 대한 추정값들의 상자그림을 포함하고 있다. 상관계수에 따른 추정값들의 평균이나 분산에는 큰 차이가 없었으나 상관계수가 큰 경우 분산이 더 컸으며 표본수가 증가하면 추정치의 변동은 줄어드는 것으로 나타났다.

Figure 4.2에는 SCAD-SCAD 조합인 경우 표본수가 500이고 상관계수가 0.8인 경우 성분별로 회귀계수들에 대하여 500번 반복 추정한 추정치들의 상자그림을 포함하였으며 Figure 4.3은 동일한 표본수와 상관계수에 대한 성분1에 SCAD-Adplasso 조합을 적용한 경우의 상자그림이다. 모든 경우 성분1에 대한 회귀계수들의 추정치는 상대적으로 편의와 변동이 작았으나 성분2와 성분3에서의 추정치들의 경우 변동이 상대적으로 크게 나타났다. 특히, 회귀계수들에 Adplasso 벌점함수를 적용한 경우 성분3에 대한 계수 추정치들은 편의와 변동이 SCAD 및 MCP 벌점함수들에 비하여 매우 크게 나타났다. 다른 경우들에 대한 상자그림들은 위의 두 경우와 큰 차이가 없어 생략하였다.



(a) Component 1 (b) Component 2 (c) Component 3

Figure 4.2. Boxplot of estimated coefficients for SCAD-SCAD case with $n = 500$, $\rho = 0.8$.

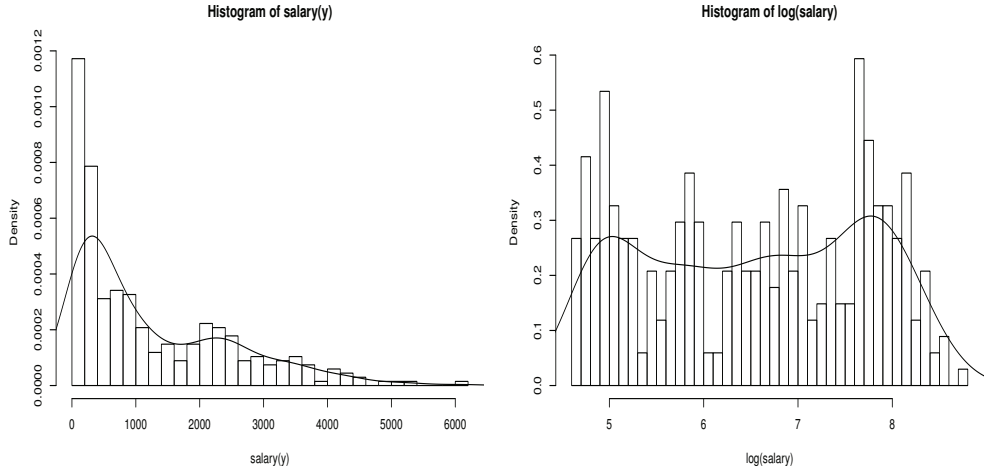


(a) Component 1 (b) Component 2 (c) Component 3

Figure 4.3. Boxplot of estimated coefficients for SCAD-Adplasso case with $n = 500$, $\rho = 0.8$.

5. 실제자료분석

새롭게 제안된 방법의 효과를 살펴보기 위하여 사용된 실제자료는 미국 메이저리그에 소속된 야구선수 337명의 1992년 연봉자료(반응변수 y)와 이와 연관된 1991년도의 성적과 관련이 있는 16개 예측변수들을 포함하고 있다. 자료는 Journal of Statistics Education의 웹사이트(www.amstat.org/publications/jse)에서 얻을 수 있다. 예측변수들은 타율(x_1), 출루율(x_2), 득점수(x_3), 안타수(x_4), 2루타수(x_5), 3루타수(x_6), 홈런수(x_7), 타점(x_8), 볼넷수(x_9), 삼진수(x_{10}), 도루수(x_{11}), 에러수(x_{12}), FA자격여부(x_{13}), 1991–92시즌에 FA자격을 얻었는지 여부(x_{14}), 연봉협상 자격여부(x_{15}), 1991–92시즌에 연봉협상 자격을 얻었는지 여부(x_{16})이다. Figure 5.1의 왼쪽에 포함된 반응변수(salary; y)인 연봉의 분포를 살펴보면 오른쪽으로 기울어진 형태를 나타내고 있다. 반응변수는 로그변환하여 사용하였으며 변환 후 두개 이상의 분포가 섞여있는 형태(Figure 5.1의 오른쪽)를 보여준다. 모형은 전체자료 중 임의로 선택된 90%의 자료를 사용하여 추정하였으며 나머지 10%의 자료는 검증에 사용하였다. 모형의 비교에는 검증자료에 대한 예측값의 평균제곱오차제곱근(root mean squared error of prediction; RMSEP)와 예측값의 상대오차(relative error of the prediction in percentage; REP)를 이용하였다 (Wang 등,



(a) Histogram of salary (y) (b) Histogram of log(salary)

Figure 5.1. Histogram of salary (y) (a) and log(salary) (b).

2013).

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}},$$

$$\text{REP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} y_i^2}} \times 100$$

이며 여기서 $\hat{y}_i = \sum_{k=1}^{K} \Pr(\hat{Z}_{ik} = k|\mathbf{x}_i)(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_k)$이다.

Table 5.1은 기존 모형과 제안 모형을 통해 얻은 회귀 계수를 나타낸 결과이다. 모든 분석에는 BIC를 최적화하는 모형을 선택하였다. 일반선형회귀모형은 후진제거법을 사용하여 변수선택을 하였는데, 8개의 변수(출루율, 안타수, 타점, 볼넷수, 삼진수, FA자격여부, 1991/2년도 FA자격여부, 연봉협상 자격여부)를 선택한 것으로 나타났다. 벌점화 회귀모형은 SCAD 방법으로 추정한 결과, 6개의 변수(출루율, 득점수, 타점, 삼진수, FA자격여부, 1991/2년도 FA자격여부, 연봉협상 자격여부)를 선택하였고, 초기 성분수를 2로 두고 성분과 회귀계수에 벌점함수를 적용하지 않은 혼합회귀모형에서는 혼합 확률이 (0.577, 0.423)인 두 개의 집단으로 분리하였다. SCAD-SCAD 벌점함수를 조합한 모형은 성분의 개수를 2개로 추정하였으며, 혼합 확률이 (0.567, 0.433)을 가지면서 첫 번째 집단은 3개의 변수(타점, FA자격여부, 연봉협상 자 여부), 두 번째 집단은 2개의 변수(안타수, 연봉협상 자격여부)를 최종적으로 선택하였다. 각 집단에서 선택된 변수들을 고려하면 첫 번째 집단은 베테랑선수의 집단으로 생각할 수 있고, 두 번째 집단은 신인선수 집단으로 생각할 수 있다. SCAD-MCP 방법도 성분 개수를 2개로 추정하였고, 혼합 확률이 (0.549, 0.451)이며 첫 번째 집단은 3개의 변수(타점, FA 자격여부, 연봉협상 자격여부), 두 번째 집단은 3개의 변수(안타수, 타점, FA 자격여부)를 선택한 것으로 나타났다. SCAD-MCP 방법에서는 첫 번째 집단이 신인선수 집단으로 생각이 되고, 두 번째 집단이 베테랑선수의 집단으로 생각된다. 후진선택법을 적용한 모형과 SCAD 벌점함수를 적용한 선형모형에서 선택된 변수들은 비슷하였으나 후진선택법에서 선택된 변수들의 수가 더 많았다. 혼합모형의 경우 회귀계수에 사용한 벌점함수의 종류에 따라 선택된 변수들에 차이가 있음을 알 수 있다.

Table 5.1. Coefficient estimates for various models (M: Mixture reg.)

변수	Linear reg.	Linear reg.	Mixture reg.		M + SCAD-SCAD		M + SCAD-MCP	
	+Backward	+SCAD	Comp.1	Comp.2	Comp.1	Comp.2	Comp.1	Comp.2
$\hat{\alpha_m}$			0.577	0.423	0.567	0.433	0.549	0.451
Intercept	−0.728	−0.755	−0.452	−1.082	−1.093	0.097	−1.052	−0.029
X_1			0.169	−0.042				
X_2	−0.058		−0.213	0.044				
X_3		0.173	−0.104	−0.088				
X_4	0.189		0.151	0.112		0.528	0.274	0.163
X_5			0.205	0.005				
X_6			−0.004	−0.013				
X_7			0.083	0.010				
X_8	0.283	0.276	0.271	0.176	0.261			0.333
X_9	0.130		0.112	0.037				
X_{10}	−0.156	−0.034	−0.241	−0.033				
X_{11}			0.044	0.077				
X_{12}			−0.047	0.009				
X_{13}	1.373	1.364	0.820	2.141	1.974		2.114	0.489
X_{14}	−0.278	−0.098	−0.196	−0.159				
X_{15}	1.101	1.151	0.609	1.881	0.904	0.235	1.829	
X_{16}			−0.081	0.258				

Table 5.2. RMSEP and REP for various models (L: Linear reg., M: Mixture reg.)

	L + Backward	L + SCAD	Mixture reg.	M + SCAD-SCAD	M + SCAD-MCP
RMSEP	0.373	0.352	0.282	0.213	0.220
REP	40.515	38.479	29.302	24.700	23.847

Table 5.2는 분석 방법별 최종 모형의 RMSEP와 REP의 값을 비교한 결과이다. 제안 모형인 SCAD-SCAD 방법과 SCAD-MCP 방법에서 RMSEP와 REP가 기존 분석 방법에 비해 작은 값을 갖는 것으로 나타나 예측력이 더 높음을 확인할 수 있었다.

6. 결론 및 향후 과제

반응변수와 예측변수들을 포함하는 관측치들이 이질적인 모집단에 속하여 있는 자료에 대한 분석방법으로 유한혼합회귀모형을 들 수 있으며 이 경우 적절한 하위 모집단의 수와 각 모집단에서 의미있는 예측변수들을 선택하는 것은 매우 중요한 문제이다. 본 연구에서는 유한혼합회귀모형의 성분의 수와 각 성분에 포함된 선형회귀모형의 예측변수들을 동시에 선택할 수 있는 벌점화 모형을 제시하였다. 성분과 예측변수 모두에 벌점함수를 적용하는 기존의 방법 (Luo 등, 2008)을 해결하기 위하여 예측변수들에 대한 벌점함수를 적용하여 성분의 수를 선택하는 대신 각 성분에 대한 확률과 회귀계수들에 직접 벌점함수를 적용하여 과적합 등의 문제를 해결하였다.

성분에 대한 벌점함수는 Huang 등 (2017)이 제시한 각각의 성분에 대한 확률의 로그변환에 SCAD 벌점함수를 적용하였으며 예측변수들에는 SCAD, Adplasso, 및 MCP 벌점함수들을 모의자료와 실제자료에 적용하고 그 결과를 비교한 결과 SCAD-SCAD 조합과 SCAD-MCP 조합에서 적절한 성분의 수와 의미있는 예측변수들을 효과적으로 선택하는 것을 확인할 수 있었다. 예측변수에 Adplasso를 사용하는 경우에는 표본의 수가 작은 경우 편의와 변동이 다른 벌점함수들에 비하여 크게 나타나 LASSO방

법의 단점을 완벽하게 극복하지 못하는 것으로 판단된다. 모의실험 결과에 포함하지는 않았지만 성분에 LASSO 벌점함수를 적용하는 경우 알고리즘이 수렴하지 않는 경우가 빈번하였으며 MCP의 경우에는 SCAD와 큰 차이가 없어 생략하였다. 그리고 성분의 확률에 SCAD 벌점함수를 적용하는 경우 Huang 등의 지적과 달리 로그 벌점함수를 사용하지 않고 직접 벌점함수를 사용하여도 수렴 등에 문제가 없었으며 이에 대한 이론적인 부분은 추가적인 연구가 필요하다.

추후 더욱 다양한 모의자료들에 적용한 결과의 비교가 필요한 것으로 보이며 선형모형 뿐만 아니라 일반화선형모형 등 일반적인 모형들에 확장하여 적용하는 것도 가능할 것이다. 본 연구에서 제시한 방법에서 얻어지는 추정치들의 일치성 및 점근적 성질들에 대한 이론적인 연구는 현재 진행중에 있다.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds. Petrox, B. N., and Caski, F., Budapest: Akademiai Kiado, pp. 267.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**, 2350–2383.

Chen, J. (1995). Optimal rate of convergence for finite mixture models, *The Annals of Statistics*, **23**, 221–233.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty, *Journal of the American Statistical Association*, **104**, 187–196.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, **30**, 74–99.

Hastings. W. (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

Huang, T., Peng, H. and Zhang, K. (2017). Model Selection for Gaussian mixture models, *Statistica Sinica*, **27**, 147–169.

Khalili, A. and Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models, *Journal of the American Statistical Association*, **102**, 1025–1038.

Luo, R., Wang, H., and Tsai, C. (2008), On mixture regression shrinkage and selection via the MR-lasso, *International Journal of Pure and Applied Mathematics*, **46**, 403–414.

McLachlan, G. and D. Peel (2000). *Finite Mixture Model*, John Wiley & Sons, Inc.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machine, *Journal of Chemical Physics*, **21**, 1087–1091.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society of London. A*, **185**, 71–110.

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

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC, FL.

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

Wedel, M. and Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed), Kluwer Academic Publishers, Boston.

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

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

혼합회귀모형에서 콤포넌트 및 설명변수에 대한 벌점함수의 적용

박종선[a,1] · 모은비[a]

[a]성균관대학교 통계학과

(2018년 11월 15일 접수, 2018년 12월 24일 수정, 2019년 1월 17일 채택)

요 약

주어진 회귀자료에 유한혼합회귀모형을 적합하는 경우 적절한 성분의 수를 선택하고 선택된 각각의 회귀모형에서 의미있는 예측변수들의 집합을 선택하며 동시에 편의와 변동이 작은 회귀계수 추정치들을 얻는 것은 매우 중요하다. 본 연구에서는 혼합선형회귀모형에서 성분의 개수와 회귀계수에 벌점함수를 적용하여 적절한 성분의 수와 각 성분의 회귀모형에 필요한 설명변수들을 동시에 선택하는 방법을 제시하였다. 성분에 대한 벌점은 성분들의 로그값에 SCAD 벌점함수를 적용하였고 회귀계수들에는 SCAD와 더불어 MCP 및 Adplasso 벌점함수들을 사용하여 가상자료와 실제자료들에 대한 결과를 비교하였다. SCAD-SCAD 벌점함수 조합과 SCAD-MCP 조합의 경우 기존의 Luo 등 (2008)의 방법에서 문제가 되었던 과적합 문제를 해결함과 동시에 선택된 성분의 수와 회귀계수들을 효과적으로 선택하였으며 회귀계수들의 추정치에 대한 편의도 크지 않았다. 본 연구는 성분의 수가 알려져 있지 않은 회귀자료에서 적절한 성분의 수와 더불어 각 성분에 대한 회귀모형에서 모형에 필요한 예측변수들을 동시에 선택하는 방법을 제시하였다는데 의미가 있다고 하겠다.

주요용어: 혼합회귀모형, 성분선택, 변수선택, 벌점함수

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cspark@skku.edu