*The Korean Journal of Applied Statistics* (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.5.793
**30**(5), 793–808

# A comparative study of feature screening methods for ultrahigh dimensional multiclass classification

Kyungeun Lee[a] · Kyoung Hee Kim[b] · Seung Jun Shin[a,1]

[a]Department of Statistics, Korea University;
[b]Department of Statistics, Sungshin Women's University

(Received August 29, 2017; Revised October 12, 2017; Accepted October 13, 2017)

**Abstract**

We compare various variable screening methods on multiclass classification problems when the data is ultrahigh-dimensional. Two different approaches were considered: (1) pairwise extension from binary classification via one versus one or one versus rest comparisons and (2) direct classification of multiclass responses. We conducted extensive simulation studies under different conditions: heavy tailed explanatory variables, correlated signal and noise variables, correlated joint distributions but uncorrelated marginals, and unbalanced response variables. We then analyzed real data to examine the performance of the methods. The results showed that model-free methods perform better for multiclass classification problems as well as binary ones.

Keywords: multi-categorical classification, simulation, ultrahigh-dimensional classification

## 1. 서론

$n$개의 관측치로 이루어진 훈련자료 $\mathbb{D} = \{(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p,\ i = 1, \ldots, n\}$를 고려하자. $X$가 주어졌을 때 $Y$의 분포가 $j$번째 변수에 의존하면 $X_j$를 신호변수(signal/informative variable), 그렇지 않으면 잡음변수(noise/uninformative variable)라 정의한다. 독립변수의 차원 $p$가 관측치의 수 $n$보다 큰 고차원(high-dimensional) 자료의 분석에서는 신호변수의 수가 전체 변수의 수보다 현저히 적다고 가정한다. $\mathcal{S} \subset \{1, \ldots, p\}$를 신호변수들의 첨자 집합(index set)이라 하면, 해당 가정은 $d = |\mathcal{S}| \ll p$로 표현할 수 있다. 예를 들어, 생물정보학에서 흔히 사용되는 유전자 미세배열(microarray) 자료는 표본 크기에 비해 월등히 많은 독립변수, 즉 수많은 유전자 정보를 가지는 데 이 중에서 질병 발현에 관여하는 소수의 유전자를 찾아내는 것을 일차적 분석 목표로 한다. 이는 지도학습 모형에서 변수선택(variable selection) 문제로 볼 수 있으며, LASSO와 같은 벌점화 추정 방법 등을 통해 해결할 수 있다 (Zhang 등, 2006; Ma와 Huang, 2008; Wu 등, 2009; Gui와 Li, 2005; 외 다수).

한편, 고차원 자료에서 변수의 수를 어떤 양의 값 $\xi > 0$에 대해 $p = O(n^{\xi})$으로 표현한다면, 초고차원 자료는 $p = \exp\{O(n^{\xi})\}$인 경우로 정의한다 (Fan과 Lv, 2008). 최근 생물 정보학에서 널리 쓰이는

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government for S. Shin (NRF-2015R1C1A1A01054913).
[1]Corresponding author: Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: sjshin@korea.ac.kr

Next Generation Sequencing (NGS) (Metzker, 2010) 자료가 대표적인 초고차원 자료이다. 초고차원 자료는 신호 대비 잡음의 비중이 압도적으로 크기 때문에, 고차원 자료에서 널리 쓰이는 벌점화 모형을 바로 적용할 경우 그 성능이 안정적이지 못하고 최적화를 위한 계산도 매우 어렵거나 종종 불가능하다 (Fan과 Lv, 2008).

Fan과 Lv (2008)는 초고차원 선형회귀모형에서 주변상관(marginal correlation)만을 활용하여 대부분의 잡음변수를 걸러낸 뒤, 남은 변수들만을 이용하여 벌점화 모형을 적용하는 이단계 방법론을 제안하였다. 이 때 잡음변수를 걸러내는 첫 번째 단계를 변수선별(feature screening)이라 하며, 초고차원 자료 분석에서 가장 대표적인 방법으로 활용되고 있다. Fan과 Lv (2008)이 제안한 상관선별(correlation screening)은 알고리즘 1과 같다.

**Algorithm 1** Correlation Screening (Fan과 Lv, 2008)

1. 모든 $j = 1, \ldots, p$에 대하여, $(x_{ij}, y_i)$, $i = 1, \ldots, n$의 표본 상관계수 절대값 $\omega_j = |\hat{\rho}_j|$를 계산한다.
2. 선별변수의 첨자집합 $\hat{\mathcal{S}}_n$를 다음과 같이 계산하고

$$\hat{\mathcal{S}}_n = \left\{ j : \omega_j \geq \omega_{(d_n)} \text{ where } \omega_{(d_n)} \text{ denotes the } d_n \text{ largest value among all } \omega_j, j = 1, \ldots, p \right\}.$$

$\hat{\mathcal{S}}_n$에 해당하지 않는 변수를 모두 제거한다.

이 때, 신호변수의 수 $d$가 표본의 크기 $n$을 넘지 않는다고 가정하면, 모형 크기 $d_n = n$ 혹은 $[n/\log n]$으로 정할 수 있다. Fan과 Lv (2008)은 상관선별의 이론적 근거로 상관선별에 의해 신호변수가 제거될 확률이 근사적으로 0임을 증명하고, 이를 확실선별 성질(sure screening property)이라 정의하였다. 즉,

$$\lim_{n \to \infty} P\left( \mathcal{S} \subseteq \hat{\mathcal{S}}_n \right) = 1 \tag{1.1}$$

으로 표현할 수 있으며, 이는 모든 변수선별 방법이 반드시 만족해야 할 이론적인 성질이다.

Fan과 Lv (2008)이 제안한 상관선별(알고리즘 1)의 핵심 아이디어는 주변 연관성만을 바탕으로 대부분의 잡음변수를 걸러낼 수 있다는 것이다. 그러나 선형회귀모형 가정을 만족하지 않는 경우에는 상관계수가 아닌 다른 값을 활용하여 연관성의 측도 $\omega_j$를 정의하는 것이 더 타당할 수 있다. 예를 들어, Fan 등 (2009)은 결합회귀모형이 선형이라 하더라도 주변회귀모형은 비선형일 수 있는 문제점을 해결하기 위해 비모수적인 주변 회귀모형을 고려한 측도를 개발하였고, Li 등 (2012)은 모형무관(model-free) 방법을 제안하였는데 이는 선형과 비선형 모형 모두를 고려할 수 있어 모형 오지정에 강건한 특징을 보인다. 이외에도 다양한 형태의 변수선별 방법이 개발되었다 (Fan과 Fan, 2008; Fan과 Song, 2010; Fan 등, 2011; He 등, 2013; Mai와 Zou, 2015).

본 연구는 초고차원 다범주분류(muti-categorical classification) 모형 하에서의 변수선별 문제를 고려하였다. 기계학습 관점에서 분류 모형은 범주형 값을 가지는 반응변수 $Y \in \{1, \ldots, K\}$를 독립변수 $\mathbf{X} \in \mathbb{R}^p$에 주어진 정보를 바탕으로 예측하는 문제로 볼 수 있다. 이 때 $K = 2$이면 이항분류(binary classification), $K \geq 3$이면 다범주분류로 다시 나눌 수 있는데, 초고차원 이항분류를 위한 변수선별 방법 (Fan과 Lv, 2008; Fan과 Fan, 2008; Fan 등, 2009; Mai와 Zou, 2012)과는 달리, 다범주분류 문제는 상대적으로 덜 연구되어 왔다. 이는 일대일(one-versus-one) 혹은 일대다(one-versus-rest) 비교 등을 활용하여 이항분류 방법을 자연스럽게 확장하여 다범주분류 문제를 해결할 수 있기 때문이다. 하지만 일대일 비교의 경우 분류 정보를 학습하는데 모든 개체를 다 이용하지 못하기 때문에 추정의 효율성

이 떨어질 수 있으며, 일대다 비교의 경우 해당 개체에 대한 각 범주별 신호 크기가 비슷한 경우 왜곡된 결과를 학습할 수 있다는 단점이 있다. 따라서 다범주분류 문제에서 이항분류 기반 선별 방법들과 다범주분류 기반 선별 방법들을 함께 고려하여 성능을 비교할 필요가 있다.

## 2. 초고차원 다범주분류를 위한 변수선별 방법

### 2.1. 이항분류 변수선별 방법

먼저, 반응변수가 이항형인 경우를 가정한 변수선별 방법들을 살펴보자. 반응변수 $y_i, \ldots, y_n \in \{-1, 1\}$는 이항형이며, 편의를 위해 독립변수 $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$는 변수별로 표준화되었다고 가정한다($\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = 1$).

Fan과 Lv (2008)과 Fan과 Fan (2008)는 반응변수 $y_i$와 $j$번째 독립변수 $x_{ij}$의 연관성을 재는 측도, 즉 주변효용(marginal utility)으로써 독립 이표본 $t$-통계량의 활용을 제안하였다:

$$\omega_j = \frac{\bar{x}_{j+} - \bar{x}_{j-}}{\sqrt{1/n_+ + 1/n_-}}, \tag{2.1}$$

여기서 $\bar{x}_{j+} = \sum_{i \in I_+} x_{ij}/n_+$와 $\bar{x}_{j-} = \sum_{i \in I_-} x_{ij}/n_-$으로 $j$번째 독립변수의 각 범주별 평균을 나타내며, $I_+ = \{i : y_i = 1\}$, $I_- = \{i : y_i = -1\}$, $n_+ = |I_+|$, $n_- = |I_-|$이다. $t$-통계량을 이용한 식 (2.1)은 $X_j|Y$는 분포의 평균을 비교하기 때문에, 특이값이 있거나, $X_j$가 신호변수임에도 $E(X_j|Y = 1) = E(X_j|Y = -1)$라면 $X_j$를 선별하지 못한다는 단점이 있다.

Fan과 Song (2010)은 일반화선형모형 하에서 주변가능도(marginal likelihood)를 활용한 변수선별 방법을 제안하였다. 이항분류의 대표적인 방법인 로지스틱 모형 하에서의 (로그)주변가능도 함수는 다음과 같다.

$$\ell_j(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n \left[ y_i \log p_j(\alpha, \beta; x_{ij}) + (1 - y_i) \log\{1 - p_j(\alpha, \beta; x_{ij})\} \right],$$

여기서 $p_j(\alpha, \beta; x_{ij}) = \exp(\alpha + \beta x_{ij})/\{1 + \exp(\alpha + \beta x_{ij})\}$이다. Fan과 Song (2010)은 변수선별을 위해 다음의 최대주변가능도(maximum marginal likelihood)를 활용하였다.

$$\omega_j = \max_{\alpha, \beta} \ell_j(\alpha, \beta). \tag{2.2}$$

의사결정론의 관점에서 보면, 가능도함수를 손실함수의 기대값인 위험함수로 이해할 수 있다. 가령, 반응변수와 $j$번째 독립변수의 관계에 대하여 $\boldsymbol{\theta}$를 모수로 가지는 모형 $f(x; \boldsymbol{\theta})$을 고려하면, 주어진 손실함수 $L$에 대하여 주변효용을 다음과 같이 정의한다:

$$\omega_j = \min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n L_j\{f(\boldsymbol{\theta}; x_{ij}), y_i\}. \tag{2.3}$$

Fan 등 (2009)은 주변효용 식 (2.3)을 이용한 변수선별 방법을 제안하였다. 대표적인 이항분류기인 서포트벡터기계(support vector machine)를 활용한 주변 효용은, 경첩손실함수(hinge loss function)를 이용하여 정의된다:

$$L(\boldsymbol{\theta}; x_{ij}, y_j) = [1 - y_i f(\boldsymbol{\theta}; x_{ij})]_+ + \frac{\lambda}{2} \|f\|_{H_K}^2, \tag{2.4}$$

여기서 $[a]_+ = \max\{a, 0\}$, $\lambda > 0$는 모형 $f$의 복잡도를 조율하는 모수이며, $H_K$는 주어진 준양정치 커널(semi-positive definite kernel) $K(x, x')$에 의해 생성된 재생 커널 힐버트 공간(reproducing kernel Hilbert space)을 의미한다. 서포트벡터기계의 결정함수 $f(x)$는 다음과 같은 $(n+1)$차원 모수 벡터 $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_n)$의 선형결합의 형태로 표현되며 (Kimeldorf와 Wahba, 1971),

$$f(x; \boldsymbol{\theta}) = \theta_0 + \sum_{i=1}^{n} \theta_i K(x, x_{ij}),$$

여기서 $\|f\|^2_{H_K} = \sum_{i=1}^{n} \sum_{i'=1}^{n} \theta_i \theta_j K(x_{ij}, x_{i'j})$이다. 커널 함수 $K$를 통해 선형은 물론 비선형 학습도 가능하다. 선형 커널 $K(x, x') = xx'$을 이용하면 선형 분류 문제를, 원형 커널(radial kernel) $K(x, x') = c(x - x')^2$을 이용하면 비선형 분류 문제를 각각 잘 해결한다. 또한 조율모수 $\lambda$를 편의상 1로 고정하여 사용하였다. 서포트벡터기계의 학습에서 조율모수 $\lambda$의 선택은 중요한 문제이지만, 변수선별 문제에서는 주변효용 식 (2.4)의 상대적인 크기만을 활용하므로 모든 독립변수가 변수별로 표준화되어있는 경우 동일한 $\lambda$를 사용하여 변수별 주변효용을 계산하여도, $\lambda$의 값이 지나치게 크거나 작지 않는 한, 선별 결과에는 큰 영향을 미치지 않기 때문이다.

한편, Mai와 Zou (2012)는 이항분류를 위한 변수선별 방법으로 콜모고로프 여과기(Kolmogorov filter)를 제안하였다. 콜모고로프 여과기는 반응변수와 독립변수 간의 연관성 측도로 $X_j|Y = 1$와 $X_j|Y = -1$의 조건부 분포의 동질성을 검정하는 콜모고로프-스미르노프(Kolmogorov-Smirov) 통계량을 사용하였다:

$$\omega_j = \sup_{-\infty < x < \infty} \left| \hat{F}_{+j}(x) - \hat{F}_{-j}(x) \right|, \tag{2.5}$$

여기서 $\hat{F}_{-j}(x) = n_-^{-1} \sum_{i \in I_-} I\{x_{ij} \le x\}$와 $\hat{F}_{+j}(x) = n_+^{-1} \sum_{i \in I_+} I\{x_{ij} \le x\}$는 각각 $\{x_{ij}; i \in I_+\}$과 $\{x_{ij}; i \in I_-\}$의 경험분포 함수이다. 콜모고로프-스미르노프 통계량이라는 비모수적 방법을 기반으로 정의되었기 때문에 콜모고로프 여과기는 자료의 분포 가정과 상관없이 작동하는 모형무관(model-free) 변수 선별법이다.

### 2.2. 이항분류 변수선별 방법에 기반한 다범주분류 변수선별 방법

다범주 반응변수 $y_i \in \{1, \ldots, K\}$가 $K$개의 범주 중 하나의 값을 취한다고 가정하자. 반응변수의 값에 따라 첨자 집합 $I_k = \{i : y_i = k\}$을 정의하고, $n_k = |I_k|$, $\mathbf{y}^{[k]} = \{y_i, i \in I_k\}$, 그리고 $\mathbf{X}^{[k]} = \{\mathbf{x}_i, i \in I_k\}$, $k = 1, \ldots, K$라 하자.

앞서 2.1절에서 소개한 이항분류 변수선별법은 일대일, 일대다 비교로 다범주 변수선별법으로 자연스럽게 확장가능하다. 이항분류 변수선별법을 기반으로 한 일대일 비교와 일대다 비교 과정은 알고리즘 2과 3에 각각 요약되어 있다.

연구 결과, 위 비교 방법간의 유의한 차이가 발견되지 않았기에, 일대일 비교 후 그 최대값을 취한 결과만 논문에 표기하였다.

### 2.3. 다범주 변수에 직접 적용 가능한 변수선별 방법

다범주 변수에 직접 적용 가능한 변수선별 방법을 설명하기 위해, $i$번째 훈련자료를 $k$번째 범주의 $l$번째 개체로 재표현하면, $\mathbb{D} = \{(y_{kl}, \mathbf{x}_{kl}) : k = 1, \ldots, K;\ l = 1, \ldots, n_k\}$가 된다.

이항분류의 변수선별에 $t$-통계량 식 (2.1)을 사용하였다는 점에 착안하여, 다범주분류의 변수선별을 위

**Algorithm 2** 일대일(one-versus-one) 비교

1. 반응변수가 가질 수 있는 임의의 두 범주쌍
$(k,l) \in \{1,\ldots,K\} \times \{1,\ldots,K\}\backslash\{k\}$에 대하여

$$\mathbb{D}_{k,l}^{\text{ovo}} = \mathbb{D}_k \cup \mathbb{D}_l$$

라 하자. 여기서 $\mathbb{D}_k$와 $\mathbb{D}_l$은 반응변수의 값이 $k$ 혹은 $l$인 자료를 바탕으로 다음과 같이 정의된다:

$$\mathbb{D}_k = \left\{ \mathbf{1}_{n_k}, \mathbf{X}^{[k]} \right\}, \qquad \mathbb{D}_l = \left\{ -\mathbf{1}_{n_l}, \mathbf{X}^{[l]} \right\},$$

여기서 $\mathbf{1}_n$은 모든 원소가 1인 $n$차원 벡터를 나타낸다.

2. 이항분류 자료 $\mathbb{D}_{k,l}^{\text{ovo}}$에 대한 $j$번째 변수의 주변연관성 측도를 계산하고 이를 $\omega_j(k,l)$라 하자. 이제, 다범주분류 변수선별을 위한 $j$번째 변수의 주변연관성 측도 $\omega_j$는

$$\omega_j = \max_{k<l} \omega_j(k,l) \quad \text{또는} \quad \omega_j = \sum\sum_{k<l} \omega_j(k,l) \Big/ \binom{K}{2}$$

로 계산된다.

**Algorithm 3** 일대다(one-versus-rest) 비교

1. 임의의 $k \in \{1,\ldots,K\}$에 대하여,

$$\mathbb{D}_k^{\text{ovr}} = \mathbb{D}_k \cup \mathbb{D}_{k^c}$$

를 정의하자. 여기서

$$\mathbb{D}_{k^c} = \left\{ -\mathbf{1}_{n-n_k}, \mathbf{X}^{[-k]} \right\}$$

이고, $\mathbf{X}^{[-k]} = \{\mathbf{x}_i, i \notin I_k\}$이다.

2. 이항분류 자료 $\mathbb{D}_k^{\text{ovr}}$에 대한 $j$번째 변수의 주변연관성 측도를 계산하고 이를 $\omega_j(k)$라 하자. 이제, 다범주분류 변수선별을 위한 $j$번째 변수의 주변연관성 측도 $\omega_j$는

$$\omega_j = \max_{k} \omega_j(k) \quad \text{또는} \quad \omega_j = \sum_{k=1}^{K} \omega_j(k)/K$$

로 계산된다.

해 다음의 $F$-통계량을 고려할 수 있다.

$$\omega_j = \frac{\sum_{k=1}^{K} n_k(\bar{x}_{jk} - \bar{x}_{j.})^2/(K-1)}{\sum_{k=1}^{K}\sum_{l=1}^{n_k}(x_{jkl} - \bar{x}_{jk})^2/(n-K)}, \tag{2.6}$$

여기서 $\bar{x}_{jk} = n_k^{-1}\sum_{l=1}^{n_k} x_{jkl}$, $\bar{x}_{j.} = n^{-1}\sum_{k=1}^{K}\sum_{l=1}^{n_k} x_{jkl}$이다.

로지스틱 모형의 자연스러운 확장으로 반응변수에 대해 다항분포를 가정하고, Fan 등 (2009)에 근거하여 다음의 다항 로지스틱 주변가능도함수를 변수선별의 측도로 사용할 수 있다.

$$\omega_j = \frac{1}{n}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\delta_{kl}\log p_{jk}\left(\hat{\alpha_k},\hat{\beta_k};x_{jkl}\right), \tag{2.7}$$

여기서 $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1,\ldots,\hat{\alpha}_K)^T$와 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_K)^T$는 최대가능도추정치이고, $\delta_{kl} = \mathbf{1}\{y_{kl} = k\}$, $p_{jk}(\hat{\alpha_k},\hat{\beta_k};x_{jkl}) = \exp(\hat{\alpha_k}+\hat{\beta_k}x_{jkl})/\{1+\sum_{k=1}^{K-1}\exp(\hat{\alpha_k}+\hat{\beta_k}x_{jkl})\}$이다.

한편, Zhu 등 (2011)는 다범주 분류모형을 위한 변수선별 기준으로 MV-SIS를 제안하였다.

$$\omega_j = n^{-1}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\frac{n_k}{n}\left\{\hat{F}_k(x_{jkl})-\hat{F}(x_{jkl})\right\}^2, \tag{2.8}$$

여기서 $\hat{F}(x) = n^{-1}\sum_{k=1}^{K}\sum_{l}^{n_k} I\{x_{jkl}\le x\}$, $\hat{F}_k(x) = n_k^{-1}\sum_{l}^{n_k} I\{x_{jkl}\le x\}$로 각각 $j$번째 독립변수의 주변 경험분포함수와, 반응변수가 $k$일 때의 조건부 경험분포함수이다. 실제 분류에 영향을 미치는 신호변수는 $\hat{F}_k$와 $\hat{F}$간에 차이가 크게 나타날 것이다. 콜모고로프 여과기와 마찬가지로, 식 (2.8)은 경험분포함수를 활용하기 때문에 모형이나 분포 가정으로부터 자유롭다는 장점이 있다. 단, 분포 함수간의 비유사성을 제곱거리를 통해 측정했다는 점에 유의하자.

Li 등 (2012)는 상관계수 대신 거리상관계수(distance correlation)의 사용을 제안하였다. 임의의 $q$차원 확률변수 $\mathbf{U}$와 $r$차원 확률변수 $\mathbf{V}$에 대하여 거리상관계수는 다음과 같이 정의 된다.

$$\mathrm{dCor}(\mathbf{U},\mathbf{V}) = \frac{\mathrm{dCov}(\mathbf{U},\mathbf{V})}{\sqrt{\mathrm{dCov}(\mathbf{U},\mathbf{U})\mathrm{dCov}(\mathbf{V},\mathbf{V})}},$$

여기서 $\mathrm{dCov}(\mathbf{U},\mathbf{V})$는 $\mathbf{U}$와 $\mathbf{V}$의 거리공분산(distance correlation)을 나타내며

$$\{\mathrm{dCov}(\mathbf{U},\mathbf{V})\}^2 = \int_{\mathbb{R}^{r+q}}\|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})-\phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\|^2\omega(\mathbf{t},\mathbf{s})d\mathbf{t}d\mathbf{s}$$

이다. 여기서, $\phi_{\mathbf{u}}(\mathbf{t})$와 $\phi_{\mathbf{v}}(\mathbf{s})$는 각각 $\mathbf{U}$, $\mathbf{V}$의 주변 특성함수(characteristic function)를 $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})$는 $(\mathbf{U},\mathbf{V})$의 결합 특성함수를 나타낸다. $n$개의 자료 $(\mathbf{u}_i,\mathbf{v}_i)$, $i=1,\ldots,n$가 주어졌다고 하면, 거리상관함수의 표본 추정량은 다음과 같다.

$$\begin{aligned}\left\{\widehat{\mathrm{dCov}}(\mathbf{u},\mathbf{v})\right\}^2 &= \hat{S}_1+\hat{S}_2-\hat{S}_3\\ &= \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_q\|\mathbf{v}_i-\mathbf{v}_j\|_r\right)\\ &\quad+\left(\frac{1}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i-\mathbf{u}_j\|_q\right)\times\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{v}_i-\mathbf{v}_j\|_r\right)\right)\\ &\quad-\left(\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|\mathbf{u}_i-\mathbf{u}_l\|_q\|\mathbf{v}_j-\mathbf{v}_l\|_r\right).\end{aligned}$$

여기서 $\|\cdot\|_d$는 $d$차원 유클리드 거리를 나타낸다. 통상적인 상관계수와 달리, $\mathrm{dCor}(\mathbf{U},\mathbf{V})=0$은 $\mathbf{U}$와 $\mathbf{V}$가 독립임을 의미한다. 앞서 정의된 것을 통해 알 수 있듯이 두 변수의 차원이 다르거나 다차원 경우에도 잘 정의된다. Li 등 (2012)가 제안한 거리상관선별(distance correlation screening)은 다음과 같은 주변 거리상관계수를 이용하여 신호변수를 선별한다.

$$\omega_j = \left|\widehat{\mathrm{dCor}}(x_j,y)\right|.$$

반응변수가 다범주인 경우, 반응변수 $y_i$를 $(K-1)$차원 가변수 $\mathbf{z}_i = (z_{i1}, \ldots, z_{i(K-1)})^T$로 재표현하여 다음의 거리상관계수를 구함으로써 변수선별에 이용할 수 있다.

$$\omega_j = \left| \widehat{\mathrm{dCor}}(x_j, \mathbf{z}) \right|, \tag{2.9}$$

여기서 $\mathbf{z}_i$의 $k$번째 원소 $z_{ik}$는 $y_i = k$일 때, 1의 값을 가지고 아니면 0의 값을 가지는 가변수이다.

### 2.4. 반복적 변수선별

주변정보만을 이용하여 변수선별을 하는 경우, 계산상 간편함을 비롯하여 여러가지 장점이 있지만 독립변수들 간의 연관관계를 제대로 반영하지 못한다는 단점이 있다. 예를 들어 어떤 독립변수들은 반응변수와 주변적으로는 연관되어 있지 않지만 결합적으로는 연관되어 있을 수도 있다. 이러한 경우, 앞서 소개한 방법들로는 선별해내기 어렵다. 이를 보완하기 위해 Fan과 Song (2010)은 반복적 변수선별법(iterative screening)을 제안하였다. 자세한 알고리즘은 다음과 같다.

**Algorithm 4** 반복적 변수선별(iterative screening)

1. $\mathbf{y}$와 $\mathbf{X}$에 대하여 변수선별을 실행하여 $\hat{d}_{(1)}$개의 신호변수를 선별한다. 이 때 선택된 변수의 첨자집합을 $\hat{\mathcal{S}}_{(1)}$, 해당 독립변수 행렬을 $\mathbf{X}_{\hat{\mathcal{S}}_{(1)}} = \{x_{ij} : j \in \hat{\mathcal{S}}_{(1)},\ i = 1, \ldots, n\}$로 나타내자.
2. 앞의 과정에서 선택되지 않은 변수의 첨자집합을 $\hat{\mathcal{S}}_{(1)}^c$로 표현하면, 해당 독립변수 행렬은 $\mathbf{X}_{\hat{\mathcal{S}}_{(1)}^c}$로 표현할 수 있다. 이제 $\mathbf{X}_{\hat{\mathcal{S}}_{(1)}^c}$를 반응변수, $\mathbf{X}_{\hat{\mathcal{S}}_{(1)}}$를 독립변수로 하여 선형회귀식을 적합한 뒤 그 잔차행렬 $\mathbf{X}_r$를 계산한다. 즉,
$$\mathbf{X}_r = \left\{ \mathbf{I}_p - \mathbf{X}_{\hat{\mathcal{S}}_{(1)}} \left( \mathbf{X}_{\hat{\mathcal{S}}_{(1)}}^T \mathbf{X}_{\hat{\mathcal{S}}_{(1)}} \right)^{-1} \mathbf{X}_{\hat{\mathcal{S}}_{(1)}}^T \right\} \mathbf{X}_{\hat{\mathcal{S}}_{(1)}^c}.$$
3. $\mathbf{y}$와 $\mathbf{X}_r$에 대하여 변수선별을 실행하고, 선택된 $\hat{d}_{(2)}$개의 독립변수 첨자 집합을 $\hat{\mathcal{S}}_{(2)}$라 하자. 선택된 변수들의 첨자집합은 $\hat{\mathcal{S}}_{(1)} \cup \hat{\mathcal{S}}_{(2)}$가 된다.
4. 앞의 단계 2–3를 $(M-1)$번 반복하여 선택된 변수의 개수 $d_{(1)} + \cdots + d_{(M)}$가 주어진 $d_n (= [n/\log n]$ 혹은 $n)$을 넘으면 멈춘다. 최종적으로 선택된 변수들은 $\hat{\mathcal{S}}_{(1)} \cup \cdots \cup \hat{\mathcal{S}}_{(M)}$이 된다.

본 논문에서는 $d_{(1)} = d_{(2)} = d_n/2$로 설정하고, 한 번의 반복만 실시하였다.

## 3. 모의실험

본 장에서는 2장에서 살펴본 식 (2.1)부터 (2.9)까지의 방법들을 다양한 모의실험 상황에서 비교해보았다. 이항분류 기반 방법들은 Binary, 다범주분류 기반 방법들은 Multi로 구분하여 나타냈다. 이항분류 기반 방법에서 주변효용을 $t$-통계량으로 하는 방법은 t, 이항 로지스틱 주변가능도를 이용하는 방법은 Logit, 서포트벡터기계를 활용한 방법은 SVM, 그리고 콜모고르프 여과기는 KF이다. 또한 다범주분류 기반 방법에서 주변효용을 $F$-통계량으로 하는 방법은 F, 다항 로지스틱 주변가능도를 이용하는 방법은 M-Logit, MV는 경험분포함수의 차의 제곱을 이용하는 방법이며, 거리 상관계수를 이용한 방법은 DC로 표현하였다. 3.1장에서는 먼저 5가지 기본 모형을 세우고 그 결과를 살펴보았다. 그 다음 3.2장부터 3.5장에 걸쳐서는 좀 더 복잡한 데이터 상황에서의 양상을 살펴보기 위해 기본 모형에 변이

를 주었는데, 꼬리가 두꺼운 경우, 신호와 잡음이 서로 연관된 경우, 변수들이 결합 분포상으로 연관되어 있지만 주변 분포상으로는 연관되지 않은 경우, 그리고 다항 반응 범주의 분포가 불균형인 경우들을 각각 분석하였다. 모든 경우에서 $n = 200$, $p = 2{,}000$로 나타나며, 모의실험 반복 횟수는 500번이다. 분류에 실제 영향을 미치는 신호변수는 $X_j$, $j \in \mathcal{S} = \{1, \ldots, d\}$로, $Y$가 가지는 범주의 수는 $K$개 $(Y = 1, \ldots, K)$로 설정하였다. 방법 간의 성능을 비교하는 측도는 다음의 2가지를 사용하였다.

1. $P_a$: 모형크기 $d_n$를 $[n/\log n] = 37$개로 한정하였을 때, 그 안에 신호변수가 포함된 비율
2. Minimum model size (MMS): 통계량을 큰 순서대로 나열하였을 때 신호변수들 $X_j$, $j \in \mathcal{S}$을 모두 포함하는 모형의 최소 크기를 나타낸 것

따라서 $P_a$가 1에 가까울수록 (클수록), MMS가 실제 신호변수의 개수, $d$와 가까울수록 (작을수록) 나은 성능을 보인다고 해석할 수 있다. 표의 결과는 $P_a$의 평균값이며 괄호 속의 숫자는 MMS의 평균값이다.

모의실험 결과의 표에서 t는 주변효용을 $t$-통계량으로 하는 방법, Logit은 이항 로지스틱 주변가능도를 이용하는 방법, SVM은 서포트벡터기계를 활용한 방법, KF는 콜모고르프 여과기 방법을 의미한다. 다범주분류 기반 방법 중 F는 주변효용을 $F$-통계량으로 하는 방법, M-Logit은 다항 로지스틱 주변가능도를 이용하는 방법, MV는 경험분포함수의 차의 제곱을 이용하는 방법이며 DC는 거리 상관계수를 이용한 방법을 나타낸다.

### 3.1. 기본 모형 모의실험

우선, 아래의 5개 기본 모형을 고려하였다.

(M1) $K = 4$. 신호변수의 차원은 $d = 6$이다. 분류 범주가 정해졌을 때 $X$의 조건부 분포는 $X|Y = y \sim N(\mu_y, \Sigma)$이며 범주별 평균 벡터 $\mu_y$는

$$\begin{aligned}
\mu_1 &= 0.5774 \times (1, 1, 1, 1, 1, 1, 0, \ldots, 0);\\
\mu_2 &= 0.5774 \times (1, 1, -1, -1, -1, -1, 0, \ldots, 0);\\
\mu_3 &= 0.5774 \times (-1, -1, 1, 1, -1, -1, 0, \ldots, 0);\\
\mu_4 &= 0.5774 \times (-1, -1, -1, -1, 1, 1, 0, \ldots, 0),
\end{aligned}$$

이며 $\Sigma = \mathrm{Diag}\{\mathrm{CS}_d(\rho), \mathrm{CS}_{p-d}(\rho)\}$이다. 여기서 $p$차원 행렬 $\mathrm{CS}_p(\rho) = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$이며 상관계수가 $\rho$인 복합대칭(compound symmetry) 공분산 구조를 나타낸다. $\mathbf{I}_p$와 $\rho\mathbf{J}_p$는 $p$차원이며 각각 항등행렬과 모든원소가 1인 정방행렬이다. 본 실험에서는 $\rho = 0.5$로 두었다. 즉, 신호변수들과 잡음변수들은 서로 독립이며 각 신호/잡음 그룹 내의 상관계수는 0.5이다. 베이즈 오분류율(Bayes error rate)은 22.6%이다.

(M2) $K = 3$. 신호변수의 차원은 $d = 5$이다. 다음의 다항 로지스틱 모형을 활용하여 $Y$를 생성하였다.

$$\begin{aligned}
\log \frac{\mathrm{pr}(Y = 1|X = x)}{\mathrm{pr}(Y = 3|X = x)} &= 6x_1 + 3x_2 + 4x_3 + 5x_4 + 2x_5,\\
\log \frac{\mathrm{pr}(Y = 2|X = x)}{\mathrm{pr}(Y = 3|X = x)} &= 5x_1 + 4x_2 + 7x_3 + 3x_4 + 4x_5.
\end{aligned}$$

이때 $X$들은 서로 독립적으로 $N(0, 1)$를 따른다. 베이즈 오분류율은 11.5%이다.

**Table 3.1.** Comparison under default models

| | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| M1 | 22.6% | 6 | 1.000 (6) | 1.000 (6) | 1.000 (6) | 0.998 (6) | 1.000 (6) | 1.000 (6) | 1.000 (6) | 1.000 (6) |
| M2 | 11.5% | 5 | 0.976 (8.0) | 0.972 (8) | 0.944 (11) | 0.930 (15) | 0.982 (6) | 0.882 (45) | 0.910 (22) | 0.920 (23) |
| M3 | 35.0% | 5 | 0.474 (996) | 0.436 (979) | 0.444 (1092) | 0.472 (984) | 0.492 (977) | 0.448 (928) | 0.386 (1303) | 0.418 (987) |
| M4 | 22.6% | 6 | 0.565 (336) | 0.805 (56) | 0.825 (61) | 0.998 (6.0) | 0.680 (200) | 0.837 (42) | 0.590 (182) | 0.835 (50) |
| M5 | 0.7% | 5 | 0.022 (1636) | 0.086 (1631) | 0.558 (1534) | 1.000 (5) | 0.026 (1615) | 0.014 (1636) | 0.990 (6) | 1.000 (5) |

(M3) $K = 3$. 신호변수의 차원은 $d = 5$이다. 다음의 다항 로지스틱 모형을 활용하여 $Y$를 생성하였다.

$$\log \frac{\mathrm{pr}(Y=1|X=x)}{\mathrm{pr}(Y=3|X=x)} = \frac{1}{9}\cos(x_1) + \frac{1}{9}3^{x_2} - \frac{2}{9}x_3^3 + \frac{2}{9}\cos(x_4) + \frac{1}{9}x_5^2,$$
$$\log \frac{\mathrm{pr}(Y=2|X=x)}{\mathrm{pr}(Y=3|X=x)} = -\sin(x_1) + \log(|x_2|+1) - 3{x_3}^2 + \sin(x_4) + \exp(x_5).$$

이때 $X$들은 서로 독립적으로 $N(0,1)$를 따른다. 베이즈 오분류율은 35.0%이다.

(M4) $K = 4$. 신호변수의 차원은 $d = 6$이다. 먼저 M1과 같은 방식으로 $W|Y = y \sim N(\mu_y, \Sigma)$를 생성하고, $X_j = \exp(2W_j)$와 같은 변환을 하였다. 베이즈 오분류율은 22.6%이다.

(M5) $K = 3$. 신호변수의 차원은 $d = 5$이다. 신호변수들은 독립적으로 아래와 같이 생성된다.

$$X_j|Y = 1 \sim t_4,$$
$$X_j|Y = 2 \sim 0.5N(2.5, 1) + 0.5N(-2.5, 1),$$
$$X_j|Y = 3 \sim 0.5N(5, 1) + 0.5N(-5, 1).$$

잡음변수들은 독립적으로 $N(0,1)$을 따른다. 베이즈 오분류율은 0.7%이다.

Table 3.1은 각 모형별로 선별방법들이 얼마나 변수를 잘 추려냈는 지를 $P_a$로 나타낸 것이다. 괄호 안은 MMS를 나타낸다. M3을 제외하면, 4개 모형에서 가장 좋은 성능을 보인 선별 방법은 콜모고로프 여과기였다. 모형별 결과를 상세히 살펴보면, M1, M2는 모수적 가정에 부합하는 모형들로 정규 분포를 따르거나 로짓 모형에 의해 생성된 자료들이기 때문에 모든 선별 방법들이 잘 작동하는 것을 볼 수 있다. 그러나 비선형 로지스틱 모형인 M3에서는 모든 방법들의 $P_a$가 0.5 미만으로 떨어졌다. 정규 분포에 지수 변환을 한 M4는 콜모고로프 여과기의 탁월한 성능을 볼 수 있다. 콜모고로프 여과기와 같이 경험분포 함수를 이용하는 MV는 그 성능이 눈에 띄게 떨어졌는데, 이는 MV가 경험분포 사이의 차에 절댓값 대신 제곱을 취하기 때문에 비대칭의 꼬리가 긴 자료에 취약한 것으로 이해할 수 있다. M5에서는 모형 무관 방법들(콜모고르프 여과기, MV, 거리상관선별)은 거의 완벽하게 선별을 한 반면, 여타 모수적 방법들은 선별이 제대로 이루어지지 않았다.

### 3.2. 꼬리가 두꺼운 경우

이번에는 자료가 정규 분포가 아닌 이상치를 가지는 다양한 경우에서 각 선별 방법의 성능을 비교하기

**Table 3.2.** Comparison under heavy-tailed models

| | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| M1 | 30.9% | 6 | 0.767 (94) | 0.778 (54) | 0.823 (36) | 1.000 (6) | 0.787 (80) | 0.727 (67) | 1.000 (6) | 0.970 (9) |
| M2 | 7.4% | 5 | 0.960 (9) | 0.888 (40) | 0.940 (14) | 0.922 (24) | 0.972 (6) | 0.866 (44) | 0.924 (17) | 0.960 (12) |
| M3 | 25.9% | 5 | 0.416 (1302) | 0.422 (1220) | 0.454 (1641) | 0.444 (1066) | 0.454 (1340) | 0.394 (1316) | 0.454 (1232) | 0.494 (1250) |
| M5(1) | 0.4% | 5 | 0.030 (1447) | 0.108 (1212) | 0.694 (6) | 1.000 (5) | 0.040 (1394) | 0.032 (1466) | 0.992 (6) | 1.000 (5) |
| M5(2) | 8.5% | 5 | 0.042 (1650) | 0.122 (1419) | 0.380 (1318) | 1.000 (5) | 0.050 (1632) | 0.028 (1645) | 0.728 (54) | 1.000 (5) |

위해 모의실험을 실시하였다. M1, M2, M3은 분포 가정을 $t_{\mathrm{df}=2}$로 변경하였으며, M5는 2가지 변형을 만들었다.

(M1) $X|Y=y \sim t_{\mathrm{df}=2}(\mu_y, \Sigma)$이고, 이때 $\mu_y, \Sigma$는 기존 M1과 같다.

(M2) $X$들은 서로 독립적으로 $t_{\mathrm{df}=2}$를 따른다. 기존 M2의 로지스틱 모형에 의해 $Y$가 생성된다.

(M3) $X$들은 서로 독립적으로 $t_{\mathrm{df}=2}$를 따우른다. 기존 M3의 로지스틱 모형에 의해 $Y$가 생성된다.

(M5) 신호변수들은 독립적으로 아래와 같이 생성되고, 잡음변수들은 두 경우 모두에서 독립적으로 $N(0,1)$을 따른다.

(1) $X_j|Y=1 \sim N(0,1)$,
$X_j|Y=2 \sim 0.5N(2.5,1)+0.5N(-2.5,1)$,
$X_j|Y=3 \sim 0.5N(5,1)+0.5N(-5,1)$.

(2) $X_j|Y=1 \sim t_4$,
$X_j|Y=2 \sim 0.5t_4(2.5)+0.5t_4(-2.5)$,
$X_j|Y=3 \sim 0.5t_4(5)+0.5t_4(-5)$.

Table 3.2는 위 모형에 대한 모의실험 결과이다. 자료가 꼬리가 두꺼운 $t$분포를 따르게 되면 신호보다 잡음이 커지므로 평균을 기반으로 변수를 선별하는 $t$-통계량과 $F$-통계량은 쉽게 망가질 것을 예상할 수 있다. 가장 간단한 모형인 M1의 결과를 보면 $t$분포 변형으로 인한 효과를 가장 잘 확인할 수 있는데, 모형무관 방법인 콜모고로프 여과기, MV 그리고 거리상관선별이 이러한 변이에 관계없이 선별을 잘했다. 각각 선형과 비선형 로지스틱 모형에서 자료를 생성하는 M2, M3에서는 모든 방법에서 성능 차이가 거의 보이지 않았다. 한편 M5의 변형에서는 (1)과 (2) 모두에서 대체로 모형무관 방법들이 잘 작동했다. MV만 (2) 변형에서 다소 떨어지는 성능을 보였는데, 이는 MV가 이상치에 영향을 크게 받는 것으로 생각할 수 있다.

### 3.3. 신호변수와 잡음변수가 서로 연관된 경우

변수선별은 결합정보가 아닌 주변정보만을 활용하기 때문에 신호변수들과 잡음변수들 간에 상관관계가 강할 경우 적절히 작동하지 않을 수 있다. 신호변수가 $X$행렬의 앞에 몰려있는 우리 모의실험의 설정상

**Table 3.3.** Comparison under models with correlated signal and noise variables.

| | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| M1 | 20.5% | 6 | 1.00 (6) | 1.000 (6) | 1.000 (6) | 0.998 (6) | 1.000 (6) | 1.000 (6) | 1.000 (6) | 0.992 (6) |
| M2 | 10.3% | 5 | 0.820 (72) | 0.806 (75) | 0.662 (201) | 0.736 (149) | 0.786 (101) | 0.601 (616) | 0.616 (216) | 0.653 (323) |
| M3 | 36.7% | 5 | 0.291 (1858) | 0.246 (1857) | 0.244 (1829) | 0.332 (1742) | 0.290 (1862) | 0.292 (1865) | 0.174 (1645) | 0.290 (1702) |

상관 구조가 자기 회귀(autoregressive)하는 경우에는 신호변수군끼리의 정보는 증폭되고, 잡음변수와의 연관은 더 작아지므로 복합 대칭성(compound symmetry)의 경우보다 더 좋은 결과를 나타낸다. 따라서 우리는 신호변수들과 잡음변수들 간에 상관구조가 복합 대칭성을 가지는 경우만을 살펴보았다. 이를 위해 변형된 기본 모형 M1, M2, M3을 이용하여 모의실험을 수행하였다.

(M1) $X|Y = y \sim N(\mu_y, \mathrm{CS})$이고, 이때 $\mu_y$는 M1과 같다.

(M2) $X \sim N(0, \mathrm{CS})$. 기존 M2의 로지스틱 모형에 의해 $Y$가 생성된다.

(M3) $X \sim N(0, \mathrm{CS})$. 기존 M3의 로지스틱 모형에 의해 $Y$가 생성된다.

CS는 신호변수군과 잡음변수군 전체에 걸쳐 상관관계가 0.5로 존재하는 것을 의미한다. 변수들간의 연관성으로 인해 $Y$ 분류 정보가 잡음변수로 퍼져 선별의 난이도가 올라간다.

Table 3.3을 보면 전반적으로 Table 3.1에 비해 선별이 잘 이루어지지 않았음을 알 수 있다. 가장 간단한 정규 분포 모형인 M1에서는 변수들의 연관이 큰 영향을 미치지 않았으나, M2와 M3에서는 모두 눈에 띄게 선별 성능이 떨어졌다. t는 주변효용을 $t$-통계량으로 하는 방법, Logit은 이항 로지스틱 주변가능도를 이용하는 방법, SVM은 서포트벡터기계를 활용한 방법, KF는 콜모고르프 여과기 방법을 의미한다. 다범주분류 기반 방법 중 F는 주변효용을 $F$-통계량으로 하는 방법, M-Logit은 다항 로지스틱 주변가능도를 이용하는 방법, MV는 경험분포함수의 차의 제곱을 이용하는 방법이며 DC는 거리 상관계수를 이용한 방법이다.

### 3.4. 결합분포상으로는 연관되어 있지만 주변분포상으로는 연관되지 않은 경우

선별이 주변정보만을 활용하기 때문에 내재된 취약점을 좀 더 심도 깊게 살펴보자. $Y$ 분류 정보가 가려지게 되는 경우, 즉 결합분포 상으로는 $X$가 $Y$에 대한 정보를 가지지만 주변분포 상으로는 정보가 없는 경우에는 선별이 제대로 작동하기 어렵다. 이러한 상황을 만들기 위해 아래와 같은 모형을 만들었다.

(M0) $\tilde{X_1}, \ldots, \tilde{X_4}$은 균일 분포 $[-\sqrt{3}, \sqrt{3}]$, $\tilde{X_5}, \ldots, \tilde{X_p}$는 $N(0,1)$에서 생성되었을 때,

(1) $j = 1, \ldots, p$에 대해, $X_j = \tilde{X_j}$와 $\mathrm{sd}(X_j) = 1$를 만족한다.

(2) $X_1 = \tilde{X_1} - \sqrt{2}\tilde{X_5}$, $X_2 = \tilde{X_2} + \sqrt{2}\tilde{X_5}$, $X_3 = \tilde{X_1} - \sqrt{2}\tilde{X_5}$, $X_4 = \tilde{X_4} + \sqrt{2}\tilde{X_5}$이고 $j = 5, \ldots, p$에서는 $X_j = \sqrt{3}\tilde{X_j}$이다. 이때 $j = 1, \ldots, p$에서 $\mathrm{sd}(X_j) = \sqrt{3}$를 만족한다.

$X = x$가 주어졌을 때, $Y$는 $k = 1, \ldots, 4$에서 $P(Y = k|\tilde{X} = \tilde{x}) \propto \exp\{f_k(\tilde{x})\}$의 확률로 생성된다. 여기서 $a = 5/\sqrt{3}$라 할 때, $f_1(\tilde{x}) = -a\tilde{x_1} + a\tilde{x_4}$, $f_2(\tilde{x}) = -a\tilde{x_1} - a\tilde{x_2}$, $f_3(\tilde{x}) = -a\tilde{x_2} - a\tilde{x_3}$, 그리고 $f_4(\tilde{x}) = -a\tilde{x_3} - a\tilde{x_4}$를 만족한다.

**Table 3.4.** Comparison under models with predictors not marginally but jointly correlated with response

| | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| M0(1) | 14.1% | 4 | 1.000 (4) | 1.000 (4) | 1.000 (4) | 1.000 (4) | 1.000 (4) | 1.000 (4) | 1.000 (4) | 1.000 (4) |
| M0(2) | 13.8% | 5 | 0.794 (996) | 0.788 (984) | 0.764 (1014) | 0.768 (914) | 0.796 (1039) | 0.722 (948) | 0.716 (1050) | 0.638 (945) |

**Table 3.5.** Comparison results for iterative versions

| | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| M0(2) | 13.8% | 5 | 1.000 (20) | 1.000 (20) | 1.000 (20) | 1.000 (20) | 1.000 (20) | 1.000 (20) | 1.000 (20) | 1.000 (23) |

**Table 3.6.** Comparison under unbalanced cases

| | | Bayes error | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| Case1 | M1 | 18% | 6 | 0.883 (18) | 0.855 (25) | 0.835 (134) | 0.775 (92) | 1.00 (6) | 0.997 (6) | 0.987 (6) | 0.988 (6) |
| Case2 | M1 | 36% | 6 | 0.725 (120) | 0.412 (255) | 0.454 (815) | 0.523 (247) | 0.988 (6) | 0.987 (6) | 0.599 (149) | 0.845 (19) |

여기서 (1)은 비교를 위한 간단한 모형으로 신호변수에 $Y$ 분류 정보가 그대로 드러나 있다. (2)는 트릭을 통해 $X_5$가 실제로는 $Y$를 생성하는 데 영향을 미치나, 주변분포 상으로는 아무 정보를 가지지 않은 것처럼 나타나게 하였다.

Table 3.4를 보면 (1)에 비해 (2)의 변형에 모든 선별 방법들이 얼마나 취약한지 알 수 있다. $X_1$부터 $X_4$까지는 쉽게 선별되지만, 분류 정보가 가려진 $X_5$는 모형크기를 최소 875개로 설정해야 선별되었다. 즉 $X_5$에 한해서는 선별이 적절히 작동하지 않았다. 이런 자료에서는 변수선별의 취약점을 극복하기 위해 제안된 반복적 변수선별을 고려할 수 있다. 2.4장에서 설명한 것처럼 반복을 한번만 실시하였으며, $d_{(1)} = 19,\ d_{(2)} = 18$로 설정하였다. Table 3.5는 반복적 변수선별을 적용한 결과이다. 모든 방법에서 $P_a$는 1.000, MMS는 20개 가량으로 나타나서 반복적 변수선별을 적용하기 전에 비해 선별이 훨씬 더 잘 작동하는 것을 알 수 있다. 즉 변수 간 연관관계가 존재하거나, $Y$의 분류 정보가 가려진 복잡한 상황에서는 반복적 변수선별을 활용함으로써 선별의 성능을 보완할 수 있다.

### 3.5. 다범주 반응변수의 분포가 불균형인 경우

마지막으로 고려한 변형은 $Y$의 범주가 불균형인 경우이다. $Y$가 불균형인 경우에는 분류 정보의 불균형으로 인하여 변수 선별이 어려워질 수 있다. 이 경우 선별 방법에 따른 성능을 비교하기 위해 아래와 같은 두 가지 경우에 대해 모의실험을 하였다.

- Case 1 : 범주별 생성 확률의 차이가 서로 큰 경우 – (M1)에서 $Y$를 뽑을 확률이 $(P(Y=1), P(Y=2), P(Y=3), P(Y=4)) = (0.05, 0.10, 0.30, 0.55)$로 변경된다.
- Case 2 : 어느 한 쪽의 비율이 매우 높고, 나머지의 비율은 작은 경우 – (M1)에서 $Y$를 뽑을 확률이 $(P(Y=1), P(Y=2), P(Y=3), P(Y=4)) = (0.85, 0.05, 0.05, 0.05)$로 변경된다.

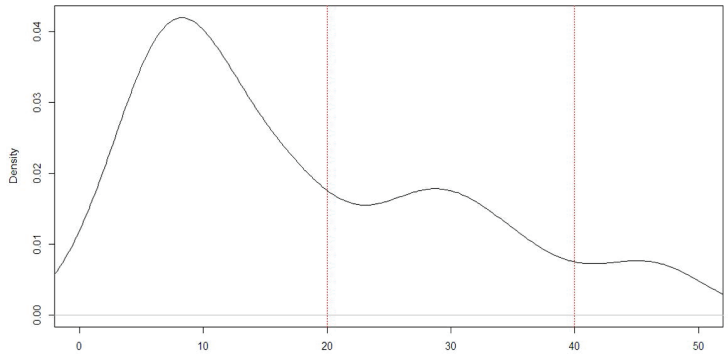


**Figure 4.1.** Density plot of response for tecator and its categorization.

Table 3.6에서 $Y$가 불균형인 경우의 결과를 볼 수 있다. Case 1과 Case 2 모두에서 이항 분류를 기반으로 한 방법들이 다범주 분류를 기반으로 한 방법들보다 선별 성능이 떨어짐을 확인할 수 있었다. 이는 앞서 언급한 것처럼 $Y$ 분류 정보가 불균형인 경우에 이항 분류 기반 방법이 적절하게 작동하지 않기 때문이다. 어느 한 쪽의 비율이 매우 높은 Case 2에서는 다범주 분류 기반 방법 중 MV만 유독 성능이 떨어졌다. MV는 경험 분포 함수를 기반으로 하기 때문에 다수 범주의 비율이 적은 경우 정확도가 떨어지는 것으로 생각할 수 있다. 따라서 $Y$가 불균형인 경우에는 다범주 분류 기반 방법 중 MV를 제외한 선별 방법을 적용하는 것이 추천된다.

## 4. 실제자료 적용

이 장에서는 우리의 방법들이 실제 자료에서는 어떠한 양상을 보이는 지 살펴보기 위해 Tecator 자료를 활용하였다. Tecator 자료는 예측 변수인 100가지 흡수율과 반응 변수인 지방률로 이루어져있다. 상세한 자료 설명은 http://lib.stat.cmu.edu/datasets/tecator에서 볼 수 있다. 우리는 전체 자료 중 이상치로 보이는 103번 자료와 105번 자료를 제외한 처음 213개를 활용하였다. 반응 변수인 지방률이 연속형으로 이루어져 있으므로 다음과 같이 구간을 나누어 범주화하였다:

| 범주 | 반응변수 $y$ 값의 범위 | 관측치의 개수 |
|---|---|---|
| 1 | $y_i \leq 20$ | 138 |
| 2 | $20 < y_i \leq 40]$ | 57 |
| 3 | $40 \leq y_i$ | 18 |

Figure 4.1은 반응 변수의 분포를 나타낸 그림이다. 또한 전체 자료의 수($n = 213$)보다 예측변수의 수가 훨씬 더 큰 상황을 만들기 위해 $t_{\mathrm{df}=7}$에서 독립적으로 생성된 4,900개 잡음변수를 추가하였다. 전체 모의 실험의 상황을 간략하게 정리하면 $n = 213$, $d = 100$, $p = 5000$이며, 잡음변수 생성을 랜덤하게 하여 100번 반복 시행하였다. Table 4.1은 모형크기를 100개로 한정하였을 때 신호변수가 포함된 비율을 나타낸 것이다. 괄호 안은 MMS를 나타낸다.

Table 4.1의 결과를 보면 전반적으로 이항분류를 기반한 방법들보다 다범주분류 방법들이 더 나은 성능을 보였다. 이는 반응변수의 범주 분포가 불균형이기 때문인데 3.5장에서 살펴보았던 모의실험과 상응하는 결과임을 알 수 있다. 다범주분류 방법에 비해서는 다소 떨어지지만 모형무관 방법인 콜모고로프 여과기도 선별을 우수하게 하였다. 이처럼 실제 자료에서는 자료가 분포 가정을 만족하기 어려우므로

**Table 4.1.** Comparison for the real data set

| | $d$ | Binary | | | | Multi | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | t | Logit | SVM | KF | F | M-Logit | MV | DC |
| Tecator | 100 | 0.77 (169) | 0.57 (320.5) | 0.21 (1668) | 0.94 (106) | 1.00 (100) | 1.00 (100) | 1.00 (100) | 1.00 (100) |

모형무관 방법을 적용하는 것이 추천된다. 이때 $Y$가 불균형인지 아닌지에 따라 다범주분류 방법만 적용할지 여부가 결정된다.

## 5. 결론

본 논문에서는 초고차원 자료의 다범주분류를 위한 변수선별에 있어 비교 연구를 수행하였다. 보다 다각적인 관점에서 보기 위해 실제 자료에서 흔히 예상되는 상황들을 모의실험을 통해 재연해 살펴보았고, 실제 자료에도 적용해보았다. 그 결과, 전반적으로 모형 가정으로부터 자유로운 방법들인 콜모고르프 여과기, 거리상관선별, 그리고 MV가 다른 방법들에 비해 잘 작동하는 것을 확인할 수 있었다. 신호변수와 잡음변수가 서로 연관되거나 $Y$와 $X$가 실제로는 서로 연관이 있지만 주변 분포상으로는 연관이 없게 나타나는 경우에서는 선별 방법들이 모두 어려움을 겪었지만, 반복적 변수선별을 적용한 후에는 적절하게 선별이 이루어졌다. 변수선별 방법을 적용할 때는 대부분의 경우에서 모형무관 방법을 1번만 적용하여도 충분하지만, 보수적으로 접근할 때는 반복적 변수선별을 적용함으로써 더 좋은 결과를 얻을 수 있을 것으로 예상할 수 있다. $Y$ 분류 범주가 불균형일 때는 조금 다른 양상을 확인하였는데, 다범주분류 방법이 이항분류 방법에 비해 월등히 좋은 성능을 보였다. 따라서 실제 자료에 변수선별을 적용할 때는 $Y$가 불균형인지 먼저 확인한 후 모형무관 선별 방법 중 어떤 선별 방법을 적용할 지 결정할 수 있다.

## References


Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules, *The Annals of Statistics*, **36**, 2605.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models, *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 849–911.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model, *The Journal of Machine Learning Research*, **10**, 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, *The Annals of Statistics*, **38**, 3567–3604.

Gui, J. and Li, H. (2005), Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, *Bioinformatics*, **21**, 3001–3008.

He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *The Annals of Statistics*, **41**, 342–369.

Kimeldorf, G. and Wahba, G. (1971). Some Results on Tchebycheffian Spline Functions, *Journal of Mathematical Analysis and Applications*, **33**, 82–95.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning, *Journal of the American Statistical Association*, **107**, 1129–1139.

Ma, S. and Huang, J. (2008). Penalized feature selection and classification in Bioinformatics, *Briefings in Bioinformatics*, **9**, 392–403.

Mai, Q. and Zou, H. (2012). The Kolmogorov filter for variable screening in high-dimensional binary classification, *Biometrika*, **100**, 229–234.

Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: a nonparametric model-free screening method, *The Annals of Statistics*, **43**, 1471–1497.

Metzker, M. L. (2010). Sequencing technologies-the next generation, *Nature Reviews Genetics*, **11**, 31–46.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, **25**, 714–721.

Zhang, H. H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machines with non-convex penalty, *Bioinformatics*, **22**, 88–95.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data, *Journal of the American Statistical Association*, **106**, 1464–1475,

# 초고차원 다범주분류를 위한 변수선별 방법 비교 연구

이경은[a] · 김경희[b] · 신승준[a,1]

[a]고려대학교 통계학과, [b]성신여자대학교 통계학과

(2017년 8월 29일 접수, 2017년 10월 12일 수정, 2017년 10월 13일 채택)

**요 약**

본 논문에서는 초고차원 자료의 다항분류를 위한 변수선별 방법에 대해 비교 연구를 진행하였다. 다항분류를 위한 변수선별 방법에는 일대일 혹은 일대다 비교를 통해 이항분류를 위한 방법을 확장시켜 적용하는 방법과 다항 반응 변수에 직접 적용할 수 있는 방법이 있다. 다항분류를 위한 변수선별 성능을 확인하기 위하여 여러가지 상황—설명변수의 꼬리가 두꺼운 경우, 신호변수와 잡음변수가 서로 연관된 경우, 결합분포상으로 연관되어 있지만 주변분포 상으로는 연관되어 있지 않은 경우, 다범주 반응변수의 분포가 불균형인 경우—을 가정하고 모의실험을 진행하였고, 실제 자료에도 적용해 보았다. 그 결과, 모형 가정을 필요로 하지 않는 방법들이 안정적인 성능을 보이는 것을 확인하였다.

주요용어: 다범주 분류, 모의실험, 초고차원 분류

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015R1C1A1A01054913).

[1]교신저자: (02841) 서울시 성북구 안암로 145, 고려대학교 통계학과. E-mail: sjshin@korea.ac.kr