*The Korean Journal of Applied Statistics* (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.6.867
**30**(6), 867–876

# Optimal number of dimensions in linear discriminant analysis for sparse data

Ga In Shin[a] · Jaejik Kim[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received August 29, 2017; Revised October 12, 2017; Accepted November 1, 2017)

**Abstract**

Datasets with small $n$ and large $p$ are often found in various fields and the analysis of the datasets is still a challenge in statistics. Discriminant analysis models for such datasets were recently developed in classification problems. One approach of those models tries to detect dimensions that distinguish between groups well and the number of the detected dimensions is typically smaller than $p$. In such models, the number of dimensions is important because the prediction and visualization of data and can be usually determined by the $K$-fold cross-validation (CV). However, in sparse data scenarios, the CV is not reliable for determining the optimal number of dimensions since there can be only a few observations for each fold. Thus, we propose a method to determine the number of dimensions using a measure based on the standardized distance between the mean values of each group in the reduced dimensions. The proposed method is verified through simulations.

Keywords: discriminant analysis, sparse data, standardized distance, dimensions

## 1. 서론

오늘날 기술과 장비의 발달로 인해 다양한 분야에서 관찰값의 개수($n$) 보다 변수의 개수($p$)가 훨씬 큰 형태를 갖는 희박 데이터(sparse data)를 쉽게 찾아볼 수 있게 되었다. 예를 들어 생물학, 화학, 유전학 분야에서는 대용량 처리기술(high throughput technology)의 발달로 인해 한 번에 수 천, 수 만 개의 화합물 또는 유전자들을 측정할 수 있게 되었다. 이러한 기술을 통해 생산된 자료는 일반적으로 표본의 수가 측정하고자 하는 변수들의 개수 보다 현저히 작게 되므로 희박 데이터에 대한 전형적인 예라고 할 수 있다. 또한, 최근들어 사물 인터넷(internet of thing; IoT) 기술의 발달로 인해 장비 또는 기계에 센서(sensor)가 장착되고 그 센서가 네트워크로 연결되어 실시간으로 데이터를 전송할 수 있게 되었다. 오늘날 자동화된 제품을 생산하는 공장에서는 중요한 기계 부품들의 선제적인 예방을 위해 또는 제품의 품질관리를 위해 기계와 장비에 센서를 장착하고 네트워크를 통해 그 센서로 부터 전송되어오는 데이터를 받아볼 수 있는 시스템을 구축하였다. 그러한 데이터는 수 많은 센서로 부터 실시간으로 전송되어져 오기 때문에 변수의 수가 생산된 제품의 개수보다 현저히 많게 되어 희박 데이터의 형태를 갖는다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. NRF-2015R1C1A1A01054808).

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

판별분석(discriminant analysis)은 분류 문제에 있어 우리가 흔히 고려하는 모형 중에 하나이고, 희박 데이터에 대한 분류 문제 역시 현실에서 종종 마주치는 문제이다. 예를 들어 우리는 마이크로어레이(microarray)나 차세대 염기서열(next generation sequencing) 기술에 의해 얻어진 유전자 발현 데이터(gene expression data)를 이용하여 유방암 환자들을 유방암의 유형별로 분류하고 예측하는 것을 목적으로 하는 분석을 원할 수 있다. 또한, 공장에서 품질관리를 위해 생산품들이 양품인지 불량품인지 기계에 부착된 센서에서 전송되는 정보를 설명변수로 고려하여 예측모형을 세우고 이를 바탕으로 실시간으로 생산되는 모든 생산품에 대해 불량품을 찾아내는 것을 원할지도 모른다. 이러한 예에서 사용될 분류모형으로 우리는 선형판별분석을 고려할 수 있지만, 기존의 선형판별분석은 이러한 희박 데이터에 적용될 수 없다. 그 이유는 기존의 선형판별분석에서는 full-rank를 갖는 그룹내 공분산 행렬(within-class covariance matrix)이 필수인데 $p \gg n$이기 때문에 그룹내 공분산 행렬은 full-rank를 만족하지 못하기 때문이다. 따라서, 희박 데이터를 위한 선형판별분석은 그룹내 공분산 행렬의 추정 문제를 해결하는 방법들에 의해 다양한 모형들이 개발되었다.

먼저 $p \gg n$인 상황에서 full-rank를 갖는 그룹내 공분산 행렬의 추정을 위해 벌점화 방법(regularization method)을 이용할 수 있다. Guo 등 (2007)은 $p \gg n$인 상황 하에서 추정된 그룹내 공분산 행렬의 대각요소에 0과 1 사이의 적당한 값을 더해줌으로써 그 문제를 해결하였고, Hastie 등 (1995)는 그룹내 공분산 행렬에 적절한 양정치 행렬(positive definite matrix)를 더하는 방법을 제안하였다. 또 다른 접근법으로 그룹들을 잘 분류할 수 있는 차원들을 찾는 방법이 있을 수 있다. 판별분석에서 그러한 새로운 차원은 $p$개의 설명변수들의 선형결합이고 그 선형결합의 계수인 판별벡터(discriminant vector)들을 구함으로써 수행된다. Witten과 Tibshirani (2011)은 Fisher (1936)의 판별분석에 $L_1$ 벌점(penalty)를 적용하여 희박 데이터에 대한 판별벡터를 구하는 방법을 소개하였고, Clemmensen 등 (2011)은 분류문제를 회귀문제로 치환한 식에 elastic-net 형태의 벌점을 적용하여 판별벡터를 구하는 방법을 제안하였다. Chung과 Keles (2010)은 Chun과 Keles (2010)가 제안한 희박부분최소제곱법(sparse partial least squares)을 적용하여 판별벡터를 구하는 방법을 소개하였다.

본 연구에서는 희박 데이터에 대한 적절한 판별벡터와 같은 선형결합을 구하여 판별분석을 수행하는 방법들을 고려하고, 그러한 방법들에서 최적의 차원의 수를 구하는 방법을 제안한다. 차원의 수는 곧 그러한 방법에서 판별벡터의 개수를 의미한다. 차원의 수는 판별분석에서 분류 예측에 있어서도 중요하고 데이터에 대한 시각화에도 중요한 역할을 한다. 데이터 내의 모든 그룹들을 분류하고 구분하는데 적절한 차원의 수보다 많거나 적을 경우 과대적합(overfitting) 또는 과소적합(underfitting) 문제가 발생하여 예측오차는 커질 수 있다 (McLachlan, 2004; Hastie 등, 2009; Clemmensen 등, 2011). 또한 판별분석의 결과로써 최적의 차원을 통해 데이터를 보여주면 데이터 내 그룹들의 구조를 효율적으로 시각화할 수 있다는 장점이 있다. 판별분석은 반응변수가 존재하는 지도학습(supervised learning)에 속하므로, 일반적으로 최적의 차원 수는 Breiman 등 (1984)에 의해 제안된 $K$-묶음 교차타당성($K$-fold cross validation) 방법에 의해 구해질 수 있다. 하지만, 희박 데이터의 경우 관찰값의 개수가 작기 때문에 각 묶음에 할당되는 관찰값 개수 또한 매우 작을 수 있다. 만일 50개의 관찰값에 대해 5개의 그룹이 있고 $K = 5$라면 각 묶음에서 각 그룹에 대한 관찰값은 불과 2개 밖에 되지 않는다. Efron과 Tibshirani (1997)는 이러한 경우 교차타당성에 의한 예측오차의 추정량의 분산이 커질 수 있다고 지적했다. 이는 $K$-묶음 교차타당성 방법이 안정적으로 최적의 차원 수를 찾는데 한계가 있다는 것을 의미한다. 따라서, 이 문제를 해결하기 위해 우리는 주어진 데이터를 $K$ 묶음으로 분할하지 않고 $n$개의 모든 관찰값들을 이용한 하나의 측도로 부터 최적의 차원 수를 찾는 방법을 제안한다. 본 연구에서 제안하는 방법은 판별벡터에 의해 축소된 차원에서 각 그룹의 평균 간의 표준화된 거리에 근거하고, 이 방법은 판별벡터를 이용하는 모든 희박 데이터에 대한 선형판별분석모형에 적용 가능하다. 그러나, 본 연구에서는 그러한 모

형 중에서 Clemmensen 등 (2011)의 희박판별분석모형과 Chung과 Keles (2010)의 희박부분최소제곱 판별분석모형에 대해 제안된 차원 수 결정 방법을 적용하고 검증한다.

2절에서는 Clemmensen 등 (2011)의 희박판별분석모형과 Chung과 Keles (2010)의 희박부분최소제곱 판별분석모형을 각각 간단히 소개하고, 3절에서는 그룹들의 평균간 표준화 거리를 이용하여 판별분석의 차원 수를 결정하는 방법을 제안한다. 4절에서는 다양한 상황 하에서의 모의실험을 통해 교차타당성 방법과 본 연구에서 제안한 방법을 비교하고 그 방법의 성능을 검증한다.

## 2. 희박데이터에 대한 선형판별분석

### 2.1. 희박판별분석

본 절에서는 Clemmensen 등 (2011)이 제안한 희박판별분석모형을 간략히 소개하고자 한다. Fisher (1936)의 판별분석은 기본적으로 그룹내 분산(within-class variance)에 비해 그룹간 분산(between-class variance)을 최대로 하는 설명변수들의 선형결합을 만드는 판별벡터 $r$개를 찾는다. 여기서 $r < G$이고 $G$는 관찰값들에 대한 그룹의 개수이다. 일단 그러한 $r$개의 선형결합을 찾게되면, 일반적인 선형판별분석을 통해 관찰값들의 분류와 예측이 가능해진다. Hastie 등 (1995)은 범주형의 반응변수를 연속형인 반응변수로 변환하여 Fisher (1936)의 분류모형을 회귀모형으로 변환하는 최적 점수화(optimal scoring) 접근법을 제안하였다. $\mathbf{X}$를 $n \times p$인 각 설명변수의 평균이 0이고 분산이 같도록 변환한 표준화된 설명변수들의 행렬이라 하자. 또한, $\mathbf{Y}$를 $G$개의 그룹들에 대한 가변수(dummy variable)를 갖는 $n \times G$ 행렬이라고 하자. 즉, $\mathbf{Y}$의 요소는 $i$번째 관찰값이 $g$번째 그룹에 속한다면 $i$번째 행 $g$번째 열은 1이고 그 행의 나머지 요소들은 0의 값을 갖게 된다. 그러면, 최적 점수화 방법에 의한 판별벡터는 다음의 최적화(optimization) 문제의 해로 부터 구할 수 있다:

$$\begin{aligned} &\min_{\boldsymbol{\beta}_k, \boldsymbol{\theta}_k} \| \mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k \|^2 \\ &\text{subject to } \frac{1}{n}\boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_l = 0, \ \forall\, l < k, \end{aligned} \tag{2.1}$$

여기서 $\boldsymbol{\theta}_k$는 $G \times 1$의 $k$번째 점수벡터이고, $\boldsymbol{\beta}_k$는 $p \times 1$의 $k$번째 판별벡터이다. $\mathbf{Y}\boldsymbol{\theta}_k$가 곧 범주형 변수를 점수를 나타내는 연속형으로 변환시킨 반응변수가 되므로 이 문제는 $\boldsymbol{\theta}_k$에 대한 제약을 갖는 회귀의 문제와 일치한다. 식 (2.1)의 최소화 문제를 풀음으로써 우리는 $r\,(< G-1)$개의 판별벡터 $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r)$를 얻을 수 있고, 판별벡터들에 의해 변환된 데이터인 $n \times r$ 행렬 $(\mathbf{X}\boldsymbol{\beta}_1, \ldots, \mathbf{X}\boldsymbol{\beta}_r)$에 보통의 선형판별분석을 적용하여 분류와 예측을 할 수 있다.

이러한 최적 점수화 방법을 희박 데이터에 대한 분류문제에 적용하기 위해 Clemmnesen 등 (2011)은 elastic-net 형태의 벌점을 사용하여 식 (2.1)을 다음과 같이 수정하였다:

$$\begin{aligned} &\min_{\boldsymbol{\beta}_k, \boldsymbol{\theta}_k} \left[ \| \mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k \|^2 + \delta \boldsymbol{\beta}_k^\top \boldsymbol{\Omega} \boldsymbol{\beta}_k + \lambda \| \boldsymbol{\beta}_k \|_1 \right] \\ &\text{subject to } \frac{1}{n}\boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_l = 0, \ \forall\, l < k, \end{aligned} \tag{2.2}$$

여기서 $\boldsymbol{\Omega}$은 양정치(positive definite) 행렬이고, $\delta$, $\lambda \geq 0$인 조정모수(tuning parameter)이다. 식 (2.2)의 최소화 문제를 Zou와 Hastie (2005)가 제안한 반복 알고리즘을 이용하여 풀음으로써 $k$번째 판별벡터 $\boldsymbol{\beta}_k$를 구할 수 있다. 이러한 $r$개의 판별벡터를 이용하여 우리는 $p$차원 공간의 데이터를 그룹들을 잘 구분할 수 있는 축소된 $r$차원의 공간으로 변환시킬 수 있다.

### 2.2. 희박부분최소제곱판별분석

Chung과 Keles (2010)는 Chun과 Keles (2010)가 회귀모형을 위해 개발한 희박부분최소제곱법을 판별분석으로 확장하였다. 따라서, 희박부분최소제곱판별분석을 이해하기 위해서는 먼저 부분최소제곱법(partial least squares)과 희박부분최소제곱법을 이해할 필요가 있다. 부분최소제곱법에서 $\mathbf{X}$는 각 열의 평균이 0인 $n \times p$인 설명변수들의 행렬이고, $\mathbf{Y}$ 역시 각 열의 평균이 0인 $n \times q$인 연속형 반응변수들의 행렬이다. 즉, $p$개의 설명변수와 $q$개의 연속형 반응변수들이 있는 것으로 가정한다. 부분최소제곱법은 설명변수와 종속변수 모두를 이용하여 잠재변수를 찾아내고 그 잠재변수들로 반응변수를 예측하는 방법이다. $s$개의 잠재변수들의 행렬을 $\mathbf{L}$이라 하면, $\mathbf{L} = \mathbf{XW}$에 의해 구할 수 있고, 여기서 $\mathbf{W}$는 $s$개의 방향벡터(direction vector) $(\mathbf{w}_1 \cdots \mathbf{w}_s)$를 열로 갖는 $p \times s$ 행렬이다. 즉, 잠재변수는 방향벡터를 계수로 갖는 $p$개의 설명변수들의 선형결합이고, 부분최소제곱법에서는 결국 방향벡터들인 $(\mathbf{w}_1 \cdots \mathbf{w}_s)$를 구해서 얻어진 잠재변수들을 회귀모형의 설명변수로써 사용한다. 각 방향벡터 $\mathbf{w}_k,\ k = 1, \ldots, s$는 다음의 최대화 문제를 통해 구할 수 있다:

$$\max_{\mathbf{w}_k} \left[\mathbf{w}_k^\top \mathbf{M} \mathbf{w}_k\right], \quad \text{subject to } \mathbf{w}_k^\top \mathbf{w}_k = 1,\ \mathbf{w}_k^\top \mathbf{S} \mathbf{w}_l = 0,\ l = 1, \ldots, k-1, \tag{2.3}$$

여기서 $\mathbf{M} = \mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{X}$이고 $\mathbf{S}$는 설명변수들의 표본 공분산 행렬이다.

Chun과 Keles (2010)의 희박부분최소제곱법은 식 (2.3)을 수정하여 다음의 최소화 문제를 통해 방향벡터 $\mathbf{w}_k$를 구한다:

$$\begin{aligned} &\min_{\mathbf{w}_k, \mathbf{e}} \left[-\alpha \mathbf{w}_k^\top \mathbf{M} \mathbf{w}_k + (1-\alpha)(\mathbf{e} - \mathbf{w}_k)^\top \mathbf{M} (\mathbf{e} - \mathbf{w}_k) + \lambda_1 \parallel \mathbf{e} \parallel_1 + \lambda_2 \parallel \mathbf{e} \parallel_2\right], \\ &\text{subject to } \mathbf{w}_\mathrm{k}^\top \mathbf{w}_\mathrm{k} = 1, \end{aligned} \tag{2.4}$$

여기서 $\mathbf{M} = \mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{X}$이고, $\lambda_1,\ \lambda_2 \geq 0$은 조정모수이다. 식 (2.4)에서 $L_1$은 방향벡터의 대리(surrogate) 역할을 하는 $\mathbf{e}$에 벌점을 줌으로써 변수선택을 가능하게 하고, $L_2$ 벌점은 $\mathbf{M}$의 역행렬이 존재하게 만든다. 이 최소화 과정에서 방향벡터 $\mathbf{w}_k$와 그 대리벡터인 $\mathbf{e}$는 서로 가까운 값을 갖도록 제한되고, 결국 방향벡터 $\mathbf{w}_k$를 통해 차원의 축소와 변수의 선택이 동시에 이루어지게 된다.

Chung과 Keles (2010)는 Chun과 Keles (2010)의 방법을 $G$개의 그룹에 대한 분류 문제에 사용하기 위해 $(G-1)$개의 가변수를 갖는 $n \times (G-1)$인 반응변수 행렬 $\mathbf{Y}$를 설정한다. 즉, $G$개의 그룹 중 임의의 그룹을 기준 그룹(baseline group)으로 정해 $\mathbf{Y}$의 모든 가변수에 대해 0을 코딩(coding)하고, $g$ 그룹에 속하는 $i$번째 관찰값에 대해서는 $\mathbf{Y}$의 $i$번째 행 $g$번째 열만 1이고 그 행의 나머지 열에 대해서는 0을 갖도록 코딩한다. 결국, Chung과 Keles (2010)의 희박부분최소제곱판별분석은 연속형이 아닌 이러한 가변수를 갖는 $\mathbf{Y}$에 식 (2.4)의 희박부분최소제곱법을 적용하여 얻어진 잠재변수들을 설명변수로 사용한 일반적인 선형판별분석을 통해 이루어진다. 즉, 이 과정을 통해 구한 잠재변수의 수 $s$는 $p$와 $n$ 보다 작기 때문에 차원의 축소를 통한 보통의 선형판별분석의 사용이 가능해진다.

## 3. 최대표준화거리방법에 의한 최적의 차원 수 결정

데이터에 대한 분석시 최적의 차원 수보다 더 많거나 더 적은 수의 차원을 사용하는 것은 여러 가지 문제를 초래할 수 있다. 먼저 예측의 측면에서 살펴보면 차원의 수가 더 많은 경우 굳이 필요하지 않은 축까지 분석에 사용되었기 때문에 과대적합현상이 일어나게 되어 모형의 분산이 커져 예측력이 떨어지고, 차원의 수가 적으면 과소적합이 일어나 분류에 중요한 구조를 놓쳐 역시 예측력에 문제가 생긴다. 뿐만 아니라 자료의 시각화, 해석의 측면 등 여러 가지 측면에서 문제가 있을 수 있다. 따라서 고차원 자료의 차원 축소시 적절한 차원 수의 결정은 예측을 비롯한 여러가지 측면에서 매우 중요하다.

앞서 소개한 희박판별분석이나 희박부분최소제곱판별분석에서도 분류를 위한 최적의 차원 수의 결정은 그러한 이유로 역시 중요하다. 일반적으로 희박판별분석에서는 최대 $(G-1)$개의 차원이면 분류에 있어 충분하지만 때때로 그 보다 더 적은 차원에서 분류가 충분히 잘 이루어지는 경우가 존재한다. 또한, 희박부분최소제곱판별분석에서는 차원 수에 대한 적절한 상한이 존재하지 않는다. 하지만, 판별분석은 반응변수가 존재하므로 예측의 측면에서 최적의 차원 수가 결정될 수 있고, 이 경우 가장 일반적인 차원의 수를 결정하는 방법은 $K$-묶음 교차타당성 방법이다. 기본적으로 이 방법은 데이터셋 전체를 각 묶음이 동일한 관찰값의 개수를 갖도록 $K$개의 묶음으로 임의(random)로 분할한 후 $(K-1)$개의 묶음은 모형을 세우기 위한 트레이닝(training) 데이터셋으로 사용하고 나머지 하나의 묶음은 모형을 평가하기 위한 테스트(test) 데이터셋으로 사용한다. 각 묶음이 돌아가면서 한 번씩 테스트 데이터셋이 되고 각 테스트 데이터셋에 대응하는 모형에 대한 평가값을 가지게 된다. 이 $K$개의 모형 평가값의 평균이 최종적인 그 모형에 대한 평가값으로 고려된다. 경험적인 연구로 부터 보통 5-묶음 또는 10-묶음의 교차타당성 방법이 사용된다. 결국 판별분석에서는 판별벡터의 각각의 개수를 적용한 각 모형에 대한 예측오차의 추정값을 교차타당성 방법으로 구하고 그 중에서 가장 작은 추정값에 대응하는 차원의 수를 결정하게 된다.

하지만, $p \gg n$인 희박 데이터의 경우 관찰값의 개수가 적은 경우가 흔하고, 이러한 경우 교차타당성 방법 적용시 1 묶음 당 관찰값의 수는 훨씬 더 작아지는 문제를 갖는다. 게다가 만일 그룹의 수가 다수일 경우 한 묶음에 포함되는 그룹당 관찰값의 개수는 훨씬 더 적어지고, 이는 교차타당성에 의해 구해진 예측오차의 추정값의 분산이 더 커질 수 있음을 의미한다. 따라서, 이 경우 교차타당성 방법에 의해 구해진 차원의 수는 자료의 참된 구조를 반영하지 못하게 될 가능성이 크고 이는 모형을 과대적합 또는 과소적합으로 이끌어 실질적인 예측오차가 커질 수 있다. 또한, 교차타당성 방법은 모형을 세우고 예측오차를 추정하는 과정을 반복적으로 요구하기 때문에 계산적으로 높은 시간과 비용을 요구한다. 이에 본 절에서는 보다 안정적이고 계산적으로 간단한 거리측도를 이용하여 희박 데이터에 대한 선형판별분석에서 최적의 차원 수를 구하는 방법을 제안한다.

희박판별분석의 최적의 판별벡터의 수와 희박부분최소제곱판별분석의 최적의 방향벡터의 수를 결정하는 문제를 고려해보자. 두 방법 모두 판별벡터 또는 방향벡터를 사용하여 축소된 차원에서 관찰값들의 판별분석을 진행한다. $\mathbf{Z}$를 희박판별분석에서 판별벡터 또는 희박부분최소제곱판별분석에서 방향벡터에 의해 축소된 새로운 차원에서의 설명변수들의 관찰값이라 하면

$$\mathbf{Z} = (\mathbf{X}\boldsymbol{\beta}_1 \cdots \mathbf{X}\boldsymbol{\beta}_r) \quad \text{or} \quad (\mathbf{X}\mathbf{w}_1 \cdots \mathbf{X}\mathbf{w}_s), \tag{3.1}$$

여기서 $\boldsymbol{\beta}_k$와 $\mathbf{w}_k$는 각각 $k$번째 판별벡터와 방향벡터이다. 판별벡터들과 방향벡터들은 모두 그룹들을 잘 구별되게 해주는 서로 직교하는 차원을 제공한다. 결국 최적의 차원들은 첫 번째 차원부터 각 차원별로 차례대로 탐색했을 때 모든 그룹들이 명확히 구분될 수 있는 최소 개수의 차원이 된다. 따라서, 각 차원에서는 적어도 하나의 그룹이 다른 그룹들과 명확히 구분이 되어야 한다. 결국 그룹들이 구분이 잘 된다는 의미는 그룹 내의 변동이 각 차원 별로 일정한 상태 하에서 그룹 간의 거리가 커야 한다는 것을 뜻한다.

두 그룹 간의 거리는 두 그룹의 평균벡터 사이의 거리로써 고려될 수 있다. 하지만, 단순히 이 거리가 멀다고 해서 두 그룹이 잘 구별되는 것은 아니다. 즉, 두 그룹의 중심이 멀리 떨어져 있어도 두 그룹의 분산이 크다면 두 집단은 잘 구분되지 않을 것이다. 따라서, 우리는 그룹들의 분산을 이용해 표준화된 차원에서 두 집단의 중심 간의 거리를 최대화하는 차원들을 식별하여 최적의 차원 수를 결정하는 방법을 제안한다. 먼저 희박판별분석의 판별벡터 또는 희박부분최소제곱판별분석의 방향벡터에 의해 축소된 차원에서의 관찰값들인 식 (3.1)의 $n \times t$인 $\mathbf{Z}$를 고려한다. 여기서 $\mathbf{Z}$의 열의 수인 $t$는 희박판별분석의

경우 $t = r$이고, 희박부분최소제곱판별분석의 경우 $t = s$이다. 총 $G$개의 그룹이 있고 $g$번째 그룹에 속하는 변환된 관찰값을 $\mathbf{z}_{ig} = (z_{i1g}, \ldots, z_{itg})^\top,\ i = 1, \ldots, n_g$라고 하자. 여기서 $\sum_{g=1}^{G} n_g = n$이다. 그러면, $j$번째 축소된 차원에 대한 각 그룹의 평균은 $\bar{z}_{jg} = (\sum_{i=1}^{n_g} z_{ijg})/n_g,\ j = 1, \ldots, t,\ g = 1, \ldots, G$이다. 이를 이용하여 $j$번째 축소된 차원에서 $g$와 $g'$ 그룹 간의 표준화된 거리 $d_j(g, g')$는 다음과 같이 정의된다:

$$d_j(g, g') = \frac{|\bar{z}_{jg} - \bar{z}_{jg'}|}{\sqrt{\sum_{g=1}^{G} v_{jg}}}, \quad j = 1, \ldots, t,\ g, g' = 1, \ldots, G, \tag{3.2}$$

여기서 $v_{jg} = \sum_{i=1}^{n_g} (z_{ijg} - \bar{z}_{jg})^2 / (n_g - 1)$이다. 즉, 두 그룹 간의 평균 간의 거리는 $G$개의 그룹들의 분산의 합에 의해 표준화 된다.

$j$번째 축소된 차원에 대해 식 (3.2)를 이용하여 우리는 $g$번째 행과 $g'$번째 열이 $d_j(g, g')$인 $G \times G$의 표준화 거리행렬을 구성할 수 있고 이 행렬을 $\mathbf{D}_j,\ j = 1, \ldots, t$라고 표시하자. 표준화 거리행렬 $\mathbf{D}_j$는 대각요소가 모두 0이고 대칭인 행렬이므로 우리는 $\mathbf{D}_j$의 상삼각 또는 하삼각만을 고려할 수 있다. 이제 최적의 차원 수를 결정하기 위해 $t$개의 표준화 행렬들의 상삼각의 같은 위치(같은 행과 열)의 거리들을 비교하여 최대거리를 구하고, 그 최대거리에 대응하는 축소된 차원을 찾는다. 즉,

$$h_{gg'} = \max_j \left\{ d_j(g, g'), j = 1, \ldots, t \right\}, \quad g, g' = 1, \ldots, G,\ g < g'. \tag{3.3}$$

최종적으로 식 (3.3)으로 부터 찾아진 $G(G-1)/2$개의 축소된 차원들은 같은 차원들이 여러 개 중복되어 있을 것이다. 이때 이러한 중복을 제거하여 유일하게 한 차원의 수가 곧 최적의 수로 결정되어질 수 있다. 예를 들어 4개의 그룹 ($G = 4$)이 있으면 우리는 6개의 식 (3.3)의 $h_{gg'}$값들을 갖는다. 이때 $(h_{12} = 1, h_{13} = 2, h_{14} = 1, h_{23} = 3, h_{24} = 1, h_{34} = 2)$라고 가정하자. 그러면 $h_{12} = 1$은 첫 번째 그룹과 두 번째 그룹이 첫 번째 축소된 차원에서 구분될 수 있다는 것을 의미한다. 마찬가지로 $h_{13} = 2$는 첫 번째 그룹과 세 번째 그룹이 두 번째 축소된 차원에서 구별될 수 있다고 해석될 수 있다. 이런 의미에서 보면 이 경우 모든 그룹들을 구별하는데는 6개의 $h_{gg'}$값 중에서 중복을 제거하여 유일하게 한 첫 번째, 두 번째, 세 번째, 세 개의 차원이면 충분하다고 판단할 수 있다. 따라서, 예측을 위해 필요한 최적의 차원 수는 3으로 결정할 수 있다.

## 4. 모의실험

본 연구에서 제안된 최대표준화거리방법의 희박데이터에 대한 선형판별분석에서의 성능을 평가하기 위해 본 절에서는 모의실험을 통해 $K$-묶음 교차타당성 방법의 성능과 비교를 시도한다. 본 모의실험에서는 모든 그룹들을 분류하는데 필요한 최적의 차원의 수가 2차원, 3차원, 5차원인 경우들을 고려한다. 각각의 경우에 대해 관찰값의 수는 100개, 독립변수의 수는 1,000개로 ($n = 100$, $p = 1000$) 희박 데이터를 고려하고, 종속변수는 모두 동일한 크기를 갖는 10개의 집단으로 설정한다 (즉, $G = 10$이고 $n_1 = \cdots = n_{10} = 10$).

각 최적의 차원 수에 대한 설명변수에 대한 관찰값은 다음의 다변량 정규분포로 부터 생성된다:

$$\begin{aligned}
&2D: (X_1, \ldots, X_{10}) \sim \mathrm{MVN}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma} + 2\mathbf{I}), \quad (X_{11}, \ldots, X_{1000}) \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}), \quad g = 1, \ldots, 10, \\
&3D: (X_1, \ldots, X_{15}) \sim \mathrm{MVN}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma} + 3\mathbf{I}), \quad (X_{16}, \ldots, X_{1000}) \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}), \quad g = 1, \ldots, 10, \\
&5D: (X_1, \ldots, X_{25}) \sim \mathrm{MVN}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma} + 4\mathbf{I}), \quad (X_{26}, \ldots, X_{1000}) \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}), \quad g = 1, \ldots, 10,
\end{aligned}$$

**Table 4.1.** Mean vector for each group and dimension

| 차원 | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ | $X_{18}$ | $X_{19}$ | $X_{20}$ | $X_{21}$ | $X_{22}$ | $X_{23}$ | $X_{24}$ | $X_{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2D | $\boldsymbol{\mu}_1$ | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_2$ | 8 | 8 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_3$ | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_4$ | 16 | 16 | 16 | 16 | 16 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_5$ | 20 | 20 | 20 | 20 | 20 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_6$ | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 4 | 4 | 4 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_7$ | 0 | 0 | 0 | 0 | 0 | 8 | 8 | 8 | 8 | 8 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_8$ | 0 | 0 | 0 | 0 | 0 | 12 | 12 | 12 | 12 | 12 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_9$ | 0 | 0 | 0 | 0 | 0 | 16 | 16 | 16 | 16 | 16 | | | | | | | | | | | | | | | |
| | $\boldsymbol{\mu}_{10}$ | 0 | 0 | 0 | 0 | 0 | 20 | 20 | 20 | 20 | 20 | | | | | | | | | | | | | | | |
| 3D | $\boldsymbol{\mu}_1$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |
| | $\boldsymbol{\mu}_2$ | 8 | 8 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |
| | $\boldsymbol{\mu}_3$ | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 0 | 20 | 20 | 20 | 20 | 20 | | | | | | | | | | |
| | $\boldsymbol{\mu}_4$ | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 8 | 8 | 8 | 8 | 8 | | | | | | | | | | |
| | $\boldsymbol{\mu}_5$ | 20 | 20 | 20 | 20 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |
| | $\boldsymbol{\mu}_6$ | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |
| | $\boldsymbol{\mu}_7$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 16 | 16 | 16 | 16 | 16 | | | | | | | | | | |
| | $\boldsymbol{\mu}_8$ | 0 | 0 | 0 | 0 | 0 | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |
| | $\boldsymbol{\mu}_9$ | 0 | 0 | 0 | 0 | 0 | 16 | 16 | 16 | 16 | 16 | 20 | 20 | 20 | 20 | 20 | | | | | | | | | | |
| | $\boldsymbol{\mu}_{10}$ | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | | | | | | | | | | |
| 5D | $\boldsymbol{\mu}_1$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_2$ | 8 | 8 | 8 | 8 | 8 | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_3$ | 12 | 12 | 12 | 12 | 12 | 8 | 8 | 8 | 8 | 8 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 20 | 20 | 20 | 20 | 20 |
| | $\boldsymbol{\mu}_4$ | 16 | 16 | 16 | 16 | 16 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 4 | 4 | 4 | 8 | 8 | 8 | 8 | 8 | 4 | 4 | 4 | 4 | 4 |
| | $\boldsymbol{\mu}_5$ | 20 | 20 | 20 | 20 | 20 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_6$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 12 | 12 | 12 | 12 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_7$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 16 | 16 | 16 | 16 | 16 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_8$ | 8 | 8 | 8 | 8 | 8 | 12 | 12 | 12 | 12 | 12 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 |
| | $\boldsymbol{\mu}_9$ | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 4 | 4 | 4 | 4 | 4 | 8 | 8 | 8 | 8 | 8 | 4 | 4 | 4 | 4 | 4 |
| | $\boldsymbol{\mu}_{10}$ | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 4 | 4 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

여기서 $\boldsymbol{\mu}_g$는 $g$번째 집단의 평균벡터이다. 또한, $\boldsymbol{\Sigma}$는 대각요소는 0이고 비대각요소들은 모두 0.5인 행렬이고 $\mathbf{I}$는 단위행렬이다. 각 차원 수에 대한 각 그룹의 평균벡터 $\boldsymbol{\mu}_g$의 상세한 값들은 Table 4.1에 나타난다.

데이터가 생성되면 희박판별분석과 희박부분최소자승판별분석을 각각 적용하여 충분한 차원 수의 판별벡터와 방향벡터들을 찾는다. 찾아진 판별벡터와 방향벡터를 이용하여 생성된 데이터를 새로운 차원의 값들의 행렬인 식 (3.1)의 $\mathbf{Z}$로 변환시킨다. $\mathbf{Z}$에 5-묶음 교차타당성 방법과 본 연구에서 제안하는 최대 표준화거리 방법을 모두 적용하여 설정한 참인 차원 수를 찾는지 조사하고, 이러한 과정을 모두 1,000번 반복하여 각 방법이 참인 차원 수를 1,000번 중 몇 번 찾아내는지 비교한다. 또한, 참인 차원 수에서 실제로 예측오차가 가장 낮은지 알아보기 위해 5,000개의 관찰값을 갖는 테스트 데이터셋을 만들고 각 반복에서 만들어진 희박판별분석과 희박부분최소제곱판별분석 모형을 테스트 데이터셋에 적용하여 각 차원 수에서 오분류율(misclassification rate)을 구한다. 최종적으로 각 차원 수에서 1,000개의 오분류율 값에 대한 평균을 구해 제시한다.

**Table 4.2.** Test error rate and proportion correctly detected by the maximum standardized distance and CV method

| 모형 | 참인 차원수 | 방법 | 판별벡터 또는 방향벡터의 개수 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 희박 판별 분석 | 2D | 최대표준화거리 | 0.127 | **0.609** | 0.229 | 0.030 | 0.003 | 0.001 | 0.001 | 0 | 0 | |
| | | 교차타당성 | 0.005 | 0.151 | 0.166 | **0.178** | 0.173 | 0.131 | 0.116 | 0.052 | 0.028 | |
| | | 테스트 오분류율 | 0.196 | **0.128** | 0.157 | 0.178 | 0.191 | 0.205 | 0.216 | 0.216 | 0.218 | |
| | 3D | 최대표준화거리 | 0 | 0 | **0.909** | 0.090 | 0.001 | 0 | 0 | 0 | 0 | |
| | | 교차타당성 | 0 | 0.023 | **0.386** | 0.209 | 0.145 | 0.101 | 0.067 | 0.046 | 0.023 | |
| | | 테스트 오분류율 | 0.501 | 0.129 | **0.026** | 0.044 | 0.050 | 0.051 | 0.052 | 0.053 | 0.054 | |
| | 5D | 최대표준화거리 | 0 | 0 | 0 | 0.062 | **0.879** | 0.059 | 0 | 0 | 0 | |
| | | 교차타당성 | 0 | 0.032 | 0.263 | **0.380** | 0.288 | 0.016 | 0.014 | 0.004 | 0.003 | |
| | | 테스트 오분류율 | 0.592 | 0.299 | 0.154 | 0.071 | **0.018** | 0.023 | 0.024 | 0.023 | 0.025 | |
| 희박 부분 최소 제곱 판별 분석 | 2D | 최대표준화거리 | **0.681** | 0.294 | 0.025 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | 교차타당성 | 0.071 | 0.123 | **0.244** | 0.154 | 0.096 | 0.093 | 0.072 | 0.051 | 0.055 | 0 |
| | | 테스트 오분류율 | 0.016 | **0.010** | 0.019 | 0.021 | 0.022 | 0.021 | 0.021 | 0.020 | 0.020 | 0.020 |
| | 3D | 최대표준화거리 | 0 | 0.001 | **0.938** | 0.061 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | 교차타당성 | 0.001 | 0.070 | **0.357** | 0.226 | 0.104 | 0.074 | 0.063 | 0.047 | 0.037 | 0.021 |
| | | 테스트 오분류율 | 0.368 | 0.011 | **0.005** | 0.013 | 0.013 | 0.013 | 0.013 | 0.012 | 0.013 | 0.013 |
| | 5D | 최대표준화거리 | 0 | 0 | 0 | 0.014 | **0.983** | 0.003 | 0 | 0 | 0 | 0 |
| | | 교차타당성 | 0.004 | 0.004 | 0.112 | 0.107 | **0.680** | 0.045 | 0.013 | 0.013 | 0.009 | 0.013 |
| | | 테스트 오분류율 | 0.461 | 0.103 | 0.024 | 0.011 | **0.003** | 0.005 | 0.005 | 0.006 | 0.005 | 0.005 |

Table 4.2는 희박판별분석과 희박부분최소제곱판별분석의 두 모형에 대해 각 참인 차원 수에서 최대표준화거리방법과 5-묶음 교차타당성 방법이 찾은 차원 수를 보여준다. 또한, 두 모형을 각 차원 수에 대해 테스트 데이터셋에 적용하여 구한 오분류율을 보여준다. Table 4.2에서 보듯이 $K$-묶음 교차타당성 방법은 희박판별분석의 2차원과 5차원인 경우 그리고 희박부분최소제곱판별분석의 2차원의 경우 1,000번의 반복 중에서 참인 차원 수보다는 다른 차원 수들을 더 많이 찾았다. 또한, 참인 차원 수를 최다 횟수로 찾은 경우에도 그 비율이 상대적으로 낮은 편이었다. 이에 반해, 최대표준화거리 방법은 희박부분최소제곱판별분석의 2차원인 경우를 제외하고는 1,000번의 반복 중에서 참인 차원 수를 교차타당성 방법에 비해 매우 높은 비율로 찾아냈음을 알 수 있다. 또한, $K$-묶음 교차타당성 방법은 잘못 찾은 경우에 대한 차원 수가 넓게 분포되어 있는데 반해, 최대표준화거리 방법은 잘못 찾은 경우에도 그 차원 수가 참인 차원 수 근처임을 알 수 있다. 즉, 이 결과는 최대표준화거리 방법이 $K$-묶음 교차타당성 방법에 비해 희박데이터에 대해서 안정적으로 참인 차원 수를 찾음을 보여준다. 또한, Table 4.2로 부터 두 모형에서 모두 참인 차원 수에서 가장 낮은 테스트 오분류율을 갖는다는 것을 알 수 있고, 이는 참인 차원 수가 최적의 차원임을 증명한다. 모의실험의 결과로 부터 우리는 본 연구에서 제안한 최대표준화거리 방법이 $K$-묶음 교차타당성 방법에 비해 희박 데이터에 대한 선형판별분석에서 최적의 차원 수를 더 정확하고 안정적이게 찾아준다는 것을 알 수 있다.

## 5. 결론

본 논문에서는 희박 데이터에 대한 선형판별분석에서 최적의 차원 수를 결정하는데 있어 계산적으로 간단하고 안정적인 방법을 제안하였다. 적절한 차원의 수를 선택하는 문제는 예측의 정확도를 결정할 수 있는 하나의 요인이 될 수 있고, 자료의 구조를 파악하고 그 시각화 측면에서도 중요한 역할을 한다. 판

별분석 모형은 반응변수가 있는 지도학습의 한 종류이므로 일반적으로 모형에서 결정되어야 할 조정모수와 차원의 수는 $K$-묶음 교차타당성 방법에 의해 정해진다. 그러나, $K$-묶음 교차타당성 방법은 주어진 데이터에 의존하는 비모수적인 방법이고 희박데이터의 경우 각 묶음이 갖는 데이터의 개수가 매우 적을 수 있으므로 정확한 차원의 수를 선택하는데 한계가 있을 수 있다. 이 문제를 보완하고 해결하기 위해 판별벡터 또는 방향벡터에 의해 축소된 차원에서 각 그룹 간의 거리에 기반한 최대표준화거리 방법이 본 연구에서 소개 되었다.

판별벡터 또는 방향벡터들은 그룹들을 더 잘 구분할 수 있는 차원을 제공한다. 따라서, 그룹들의 모든 쌍들이 모두 구분될 수 있는 최소 개수의 차원이 최적의 차원이라 고려될 수 있고, 축소된 차원에서 그룹 간의 표준화된 거리가 최대가 되는 차원들 모두의 집합이 곧 관찰값들을 분류하는 최적의 차원들이 될 수 있다. 또한, 계산적인 측면에서 $K$-묶음 교차타당성 방법은 각 차원 수에 대해 반복적인 계산을 요구하지만 최대표준화거리 방법은 한 번의 계산으로 차원 수를 결정할 수 있다는 이점이 있다.

그러나, 최대표준화거리 방법은 각 축소된 차원에서 그룹 내의 분산들이 그룹들 간의 거리에 비해 상대적으로 크다면 최적의 차원 수를 찾지 못할 수 있다. 기본적으로 선형판별분석 모형이 그룹 내의 변동과 비교해서 그룹 간의 변동을 최대로 하는 차원들을 찾기 때문에 그룹들을 잘 구분해주는 판별벡터 또는 방향벡터들이 찾아졌다면 최대표준화거리 방법이 잘 작동할 것이다. 하지만, 그룹들이 잘 분리되지 않고 많이 겹치는 경우, 그룹 내 변동이 그룹 간의 거리에 비해 크게 되므로 최대표준화거리 방법의 성능은 좋지 않을 것이다.

결론적으로 본 연구에서는 모의실험을 통해 희박데이터의 선형판별분석에서 $K$-묶음 교차타당성 방법이 최적의 차원 수를 결정하는데 있어 한계가 있다는 것을 보였고, 이에 대한 대안으로써 최대표준화거리 방법이 보다 나은 결과를 제공할 수 있다는 것을 보였다.

## References


Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth International Group.

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

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data, *Statistical Applications in Genetics and Molecular Biology*, **9**, 1544–6115.

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

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method, *Journal of the American Statistical Association*, **92**, 548–560.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its applications in microarrays, *Biostatistics*, **8**, 86–100.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics*, **23**, 73–102.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Element of Statistical Learning*, Springer, New York.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New Jersey.

Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant, *Journal of Royal Statistical Society, Series B*, **73**, 753–772.

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

# 희박한 데이터에 대한 선형판별분석에서 최적의 차원 수 결정

신가인[a] · 김재직[a,1]

[a]성균관대학교 통계학과

(2017년 8월 29일 접수, 2017년 10월 12일 수정, 2017년 11월 1일 채택)

**요 약**

오늘날 관찰값의 개수에 비해 변수의 개수가 큰 희박한 데이터셋은 다양한 분야에서 쉽게 찾아볼 수 있고, 통계학에서 그러한 데이터셋에 대한 분석은 하나의 도전이 되어 왔다. 그러한 희박한 데이터에 대한 분류를 위해 판별분석모형들이 최근에 개발되었다. 그러한 판별분석모형들 중 하나의 접근법은 그룹들을 잘 구분해주는 차원들을 찾기를 시도하는데, 그러한 차원들은 데이터의 변수의 개수보다 훨씬 적다. 그러한 모형에서 차원의 수는 예측과 자료의 시각화를 위해 중요한 역할을 하고 일반적으로 $K$-묶음 교차타당성 방법에 의해 결정된다. 하지만, 희박한 데이터의 경우 $K$-묶음 교차타당성 방법 적용시 각 묶음에 대한 관찰값의 개수가 매우 적을 수 있기 때문에 교차타당성에 의한 차원 수 결정은 신뢰성이 떨어질 수 있다. 따라서, 본 연구에서는 그러한 희박판별분석모형에 의해 찾아진 차원들에서 각 그룹들의 평균 간의 표준화된 거리에 근거한 측도를 사용하여 최적의 차원 수를 결정하는 방법을 제안하고, 제안된 방법은 모의실험을 통해 검증된다.

주요용어: 판별분석, 희박데이터, 표준화 거리, 차원

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2015R1C1A1A01054808).

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대 통계학과. E-mail: jaejik@skku.edu