Classification Using Sliced Inverse Regression and Sliced Average Variance Estimation¹⁾

Hakbae Lee²⁾

Abstract

We explore classification analysis using graphical methods such as sliced inverse regression and sliced average variance estimation based on dimension reduction. Some useful information about classification analysis are obtained by sliced inverse regression and sliced average variance estimation through dimension reduction. Two examples are illustrated, and classification rates by sliced inverse regression and sliced average variance estimation are compared with those by discriminant analysis and logistic regression.

Keywords : Dimension reduction, Central subspace, Sliced Inverse Regression, Sliced Average Variance Estimation

1. 서론

반응변수(response)가 두개의 값을 갖는 이진(binary)일 경우 로지스틱 모형(logistic model)이나 판별분석(discriminant analysis)등이 분류(classification)분석 방법으로 널리 쓰이고 있다. 본 논문은 이와 같은 분류분석에 전통적인 접근이 아닌 차원축소(dimension reduction)를 이용한 그래픽 방법론을 제시하고자 한다. Li(1991)는 역회귀(inverse regression)분석을 이용한 방법으로 Sliced Inverse Regression(SIR)을 개발하였다. Cook과 Weisberg(1991)는 Li가 개발한 SIR의 대안으로 Sliced Average Variance Estimation(SAVE)를 제시하였다. Cook과 Weisberg 그리고 Li가 제시한 SAVE와 SIR은 반응변수가 연속형(continuous)일 경우 차원축소를 이룰 수 있는 그래픽 방법(graphical method)들이다.

그래픽 방법의 기본적인 개념은 스칼라(scalar)인 반응변수 $y \in R$ 와 p 개의 설명변수(predictor)들로 이루어진 $p \times 1$ 벡터 $x \in R^p$ 를 고려하면 궁극적인 회귀분석의 목표를 다음과 같이 표현할 수 있다.

$$F(y|x) = F(y|n^T x) \quad (1.1)$$

여기서 $n = (n_1, n_2, \dots, n_d)$ 는 $p \times d$ ($d \leq p$)인 행렬로 $n^T x$ 는 $d \times 1$ 인 벡터를 나타내고 $F(u|v)$ 는

1) This work was supported (in part) by Yonsei University Research Fund of 2002.

2) Assistant Professor, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea
hblee@yonsei.ac.kr

y 가 주어졌을 때 u 의 누적 조건부 분포(cumulative conditional distribution)를 나타낸다. Cook(1991, 1994)과 Li(1991)가 주장하는 그래픽 방법은 만일 $p \times d$ 행렬 n 를 발견할 수 있다면 p 차원인 원래변수 $x \in R^p$ 를 p 보다 작은 d 차원의 새로운 변수 $n^T x = (n_1^T x, n_2^T x, \dots, n_d^T x) \in R^d$ 로 대체하면 $y|x$ 의 조건부 확률이 $y|n^T x$ 의 조건부 확률과 같아져서 x 대신 $n^T x$ 로 회귀모형을 설명할 수 있다는 것이다. 그래픽 방법의 관점에서 설명하면 개념적으로 원래 $(p+1)$ 차원 그래픽으로 설명하여야 할 회귀모형이 원래의 모든 정보를 잃어버림 없이 축소된 $(d+1)$ 차원 그래픽의 회귀모형으로 요약해도 충분하다는 것이다. 이러한 $(d+1)$ 차원 그래픽을 y 대 $n^T x$ 의 충분 요약 그림(sufficient summary plot)이라고 부른다(Cook, 1994, 1998, Cook and Weisberg, 1999). 4차원 이상의 그래픽을 인식하는 것이 현실적으로 많은 어려움이 있다. Cook(1996, 1998)은 이와 같은 4차원 이상 그래픽에 관한 논의를 하였다. 본 논문에서는 $d=0, 1, 2$ 의 경우로 3차원 이내의 그래픽으로 설명 가능한 경우만 다루고자 한다.

본 논문에서는 반응변수 y 가 오로지 두개의 값만 취하는 이진(binary)형이고 설명변수 x 는 연속형일 경우를 다루고자 한다. Cook과 Lee(1998)는 반응변수 y 가 두개의 값만 갖는 이진일 경우 SIR과 SAVE를 이용한 차원축소에 관한 논의를 하였다. 이러한 논의 과정에서 SIR과 SAVE가 분류분석의 또 다른 제안이 될 수 있다는 것이 본 논문의 관심사이다.

논의의 편의를 위하여 다음과 같이 기호를 약속하도록 한다. $U \amalg V|W$ 는 확률변수(벡터) W 가 주어진 조건일 때 또 다른 두개의 확률변수(벡터)인 U 와 V 가 독립적(independent)이라는 것을 나타내는 기호이다. 마찬가지로 $U \amalg V$ 는 두개의 확률변수(벡터)인 U 와 V 가 독립적이라는 것을 표현하는 기호이다. 부공간(subspace)은 S 로 표시하도록 하며 $S(B)$ 는 $r \times c$ 행렬 B 의 열벡터(column vector)에 의해 형성되는 부공간을 표시하도록 한다.

반응변수가 이진일 경우의 그래픽 방법론의 이론적 개념과 SIR과 SAVE에 관한 내용은 제2절에서 논의하고자 한다. 실증예제를 통하여 SIR, SAVE, 로지스틱 모형 그리고 판별분석으로부터의 정분류율(classification rate)의 비교는 제3절의 내용이다. 결론은 제4절에서 언급된다.

2. Graphical Methods

차원축소를 이용한 그래픽 방법론을 이해하기 위해 위의 식(1.1)을 다음과 같이 재구성할 수 있다.

$$y \amalg x | n^T x \quad (1.2)$$

식(1.2)는 x 의 모든 값에 대하여 $y|x$ 의 분포는 $y|n^T x$ 의 분포와 같음을 의미한다. 여기서 $p \times d$ 행렬 n 의 열벡터로 이루어진 부공간 $S(n)$ 는 반응변수를 y 로 설명변수를 x 로 하는 회귀모형의 차원축소 부공간(dimension reduction subspace)이라고 정의한다. 차원축소 부공간은 설명변수 p 차원의 x 를 p 보다 작은 d 차원의 $n^T x$ 로 설명변수의 차원을 줄일 수 있다는 이유로 유용하게 이용할 수 있다. 식(1.2)를 만족하는 많은 부공간 중에서 가장 효율적인 최소 차원축소 중심부공간

을 다음과 같이 정의한다.

<정의 1> 다음의 두 조건 (1)과 (2)를 만족하면 $S_{y|x}$ 을 중심 부공간(Central Subspace)이라고 정의한다.

(1) $S_{y|x}$ 는 차원축소 부공간이다.

(2) $S_{y|x}$ 는 모든 차원축소 부공간들의 교집합이다: $S_{y|x} = \bigcap_n S(n)$ for all n

$S_{y|x}$ 의 존재성(existence) 및 유일성(uniqueness)에 관한 내용은 Cook(1994, 1996, 1998)에 의해 논의되었다. 본 논문에서는 중심부공간 $S_{y|x}$ 이 존재한다고 가정한다. 한편 설명변수 x 는 다음과 같이 표준화된(standardized) 변수 z 를 사용하도록 한다.

$$z = \Sigma_x^{-1/2}(x - E(x))$$

여기서 Σ_x 는 x 의 공분산 행렬(covariance matrix)을 나타낸다. 이제부터 $y = \Sigma_x^{1/2} \eta$ 는 $S_{y|x}$ 의 기저(basis)가 된다고 가정한다.

그래픽 방법을 위한 차원축소를 위해서 다음의 두 가지 조건을 필요로 한다.

조건 1: 선형성 조건(linearity condition)

$$E(z|\eta^T z) = P_y z \quad (1.3)$$

조건 2: 상수 공분산 조건(constant covariance condition)

$$\text{Cov}(z|\eta^T z) = Q_y \quad (1.4)$$

여기서 $P_y = \eta(\eta^T \eta)^{-1} \eta^T$ 와 $Q_y = I_p - P_y$ 는 투사행렬(projection matrix)을 나타낸다.

<정리 1> 행렬 η 가 중심 부공간 $S_{y|x}$ 의 기저(basis)라고 하자.

만일 $E(z|\eta^T z) = P_y z$ 이고 $\text{Cov}(z|\eta^T z) = Q_y$ 가 성립하면 $E(z|y) \in S_{y|x}$ 이고 $S(I_p - \text{Cov}(z|y)) \subset S_{y|x}$ 가 성립한다.

<정리 1>은 선형성 조건과 상수 공분산 조건이 만족하면 역회귀(inverse regression) 모형이 중심 부공간에 관한 정보를 제공한다는 사실을 말해 주고 있다. 이제 y 가 오직 두개의 값만 갖는 경우 ($y=0, 1$)를 고려하도록 한다. 기호의 편의상 다음과 같이 정의한다.

$$\begin{aligned}\mu_j &= E(z|y=j), \quad \Sigma_j = \text{Cov}(z|y=j), \quad j=0, 1 \\ f &= \Pr(y=1), \quad v = \mu_1 - \mu_0 \quad \Delta = \Sigma_1 - \Sigma_0\end{aligned}$$

$(y_i, x_i), i=1, \dots, n$ 를 독립적이고 같은 분포를 갖는 자료라고 한다면 v 와 Δ 의 일치 추정치(consistent estimate)는 다음과 같다.

$$\begin{aligned}\hat{v} &= \hat{\Sigma}_x^{-1/2}(\bar{x}_1 - \bar{x}_2) \\ \hat{\Delta} &= \hat{\Sigma}_x^{-1/2}(\hat{\Sigma}_{x11} - \hat{\Sigma}_{x12})\hat{\Sigma}_x^{-1/2}\end{aligned}$$

여기서 $\bar{x}_j$ 그리고 $\hat{\Sigma}_{xj}$ 는 $y=j$ 가 주어졌을 경우 원래 x 확률벡터의 표본 조건부 평균벡터와 표본 조건부 공분산 행렬이다. 만일 z 가 다변량 정규분포(multivariate normal distribution)를 따를 경우 위의 두 조건은 충족된다(Cook, 1994, 1998).

<정리 2> 행렬 y 가 중심부공간 $S_{y|z}$ 의 기저(basis)라고 하자. 반응변수 y 가 이진(binary)이고 $z|y$ 가 정규분포를 따른다고 가정하면 $E(z|y^T z) = P_y z$ 이고 $\text{Cov}(z|y^T z) = Q_y$ 가 성립한다.

<정리 2>는 조건부 확률 $z|y$ 이 정규분포를 할 경우 선형성 조건과 상수 공분산 조건을 만족한다는 것을 의미한다.

반응변수가 이진일 경우 중심부 공간을 추정하는 과정을 설명하면 다음과 같다. 추정 가능한 $p \times k$ 모집단 행렬 M 이 $S(M) \subset S_{y|z}$ 와 같은 조건을 만족한다고 하자: M 의 일치추정치(consistent estimate)는 $(\hat{\Delta}, \hat{v})$ 의 함수인 $\hat{M} = G(\hat{\Delta}, \hat{v})$ 로 표현하자. $\hat{M}$ 의 특이치 분해(singular value decomposition)의 특이치(singular value)를 $\hat{s}_1 \geq \hat{s}_2 \geq \dots \geq \hat{s}_r$ 이라하고 $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_r$ 을 특이치에 대응하는 특이벡터(singular vector)라고 표현하자. 여기서 $r = \min(p, k)$ 이다. 만일 $d = \dim(S(M))$ 을 알고 있다면

$$\hat{S}(M) = S(\hat{u}_1, \hat{u}_2, \dots, \hat{u}_d)$$

은 $S(M)$ 의 일치추정치(consistent estimate)가 된다. 그러나 현실은 d 를 모르기 때문에 $\hat{d}$ 를 추정해야 한다. 차원 추정을 하기 위해 다음의 가설을 고려한다.

$$H_0: d = m \quad \text{versus} \quad H_a: d > m$$

위 가설에 대한 검정통계량(test statistic)은 다음과 같은 형태가 된다.

$$\widehat{\Lambda}_m = cn \widehat{f}(1 - \widehat{f}) \sum_{k=m+1}^p \widehat{s}_k^2$$

여기서 $c=1$ 또는 $c=1/2$ 이고 $\widehat{f}$ 는 $y=1$ 일 경우의 표본비율을 나타낸다. 검정절차는 귀무가설을 $d=0$ 부터 시작하여 m 을 1씩 증가시키고 $d=m$ 을 기각하는데 실패할 때까지 반복한다. 만일 $H_0: d=m$ 을 기각하는데 실패한다면 $\widehat{d}=m$ 으로 추정한다. 본 논문에서 고려하는 반응변수가 이진일 때 SIR과 SAVE의 경우 검정통계량 $\widehat{\Lambda}_m$ 은 근사적으로 카이제곱 분포(chi-squared distribution)을 한다.

● Sliced Inverse Regression (SIR)

SIR의 경우 $S_{SIR} = S(v)$ 이고 $d = \dim(S_{SIR}) \leq 1$ 이기 때문에 $m > 1$ 인 경우의 검정은 의미가 없다. 한편 SIR의 경우 $\widehat{M} = \widehat{v}$ 이고 검정통계량 $\widehat{\Lambda}_0^{SIR}$ 은 점근적으로(asymptotically) 다음과 같은 카이제곱 분포로 요약된다.

$$\widehat{\Lambda}_0^{SIR} = n \widehat{f}(1 - \widehat{f}) \|\widehat{v}\|^2 \rightarrow \chi^2(p)$$

SIR의 경우 $H_0: d=0$ 을 검정하는 문제는 $H_0: v=0$ 을 검정하는 문제와 일치하며 Hotelling의 통계량 T^2 가 검정을 위한 대안으로 쓰일 수 있다. 주목할 점은 반응변수가 이진일 경우 SIR은 오로지 1차 적률(first moment)인 v 에 의존하며 2차 적률(second moment)인 Δ 에 의해선 영향을 받지 않는다는 것이다(Cook and Lee, 1999).

● Sliced Average Variance Estimation (SAVE)

반응변수가 이진일 경우 SIR과 SAVE는 다음과 같이 요약된다(Cook and Lee, 1999).

<정리 2>

$$S_{SIR} = S(v) \subset S_{SAVE} = S(\Delta, v) \subset S_{yz}$$

<정리 2>는 일반적으로 반응변수가 이진일 경우 SAVE가 SIR보다 중심 부공간(central subpace)에 보다 많은 정보를 제공한다는 사실을 말해 주고 있다. 여기에 정규분포의 가정을 첨가하면 다음과 같은 사실을 얻을 수 있다(Cook and Lee, 1999)

<정리 3> 만일 $z|y$ 가 정규분포를 한다고 가정하면 다음과 같은 결과를 얻는다.

$$(1) S_{SAVE} = S(\Delta, v) = S_{yz}$$

$$(2) H_0: d=m \text{을 위한 검정통계량 } \widehat{\Lambda}_m^{SAVE} \text{은 점근적으로(asymptotically) 카이 제곱 분포를}$$

갖고 유의 수준 α 에서 다음과 같은 기각치(critical value)를 얻을 수 있다..

$$\chi^2_{1-\alpha}((p-m)(p-m+1)/2) < \hat{\Lambda}_m^{SAVE} < \chi^2_{1-\alpha}((p-m)(p-m+3)/2) \quad (1.5)$$

Cook과 Lee(1999)는 식(1.5)의 검정통계량 $\hat{\Lambda}_m^{SAVE}$ 에 관한 자세한 논의를 하였다. 한편 본 논문은 SAVE 방법론을 보다 쉽게 적용하기 위하여 다음과 같이 보수적인(conservative)한 검정을 행하기로 한다.

<정리 4> 만일 $z|y$ 가 정규분포를 한다고 가정하면 다음과 같은 결과를 얻는다.

$H_0: d=m$ 을 위한 검정통계량 $\hat{\Lambda}_m^{SAVE}$ 은 점근적으로(asymptotically) 카이제곱 분포를 갖는다.

$$\hat{\Lambda}_m^{SAVE} \rightarrow \chi^2((p-m)(p-m+3)/2) \quad (1.6)$$

현실적으로 SAVE를 적용함에 있어 $z|y$ 가 정규분포 가정을 검증하는 문제가 대두된다. 각 y 값에 대하여 z 의 분포가 다변량 정규분포를 따르게 하기 위해 Velilla(1993)의 변수변환을 이용하여 적용하기로 한다. 그렇지만 변수변환을 한 후의 분포가 정규분포 가정을 심하게 왜곡되지 않는 한 <정리 4>의 결과는 유용하다(Cook(1994, 1998, 1999), Cook and Lee(1999)).

3. 실 증 예 제

실증예제 분석을 위해 SIR과 SAVE는 lisp-stat(Tierney, 1990)로 만든 프로그램 Arc(<http://www.stat.umn.edu/arc/software.html>)를 사용하였고 로지스틱 회귀모형 및 판별분석은 SPSS를 사용하였다. 중심 부공간의 기저 γ 의 추정치 $\hat{\gamma}$ 는 SIR의 경우 Arc를 이용하였고 SAVE의 경우 Cook과 Yin(2000)이 lisp-stat을 이용하여 만든 코드를 사용하였다.

● Swiss Bank Notes Data

Flury와 Riedwyl (1988, p.5)의 자료로 스위스 은행 위조지폐 예제이다. 반응변수는 2진으로

$y=0$ = 진짜지폐

$y=1$ = 위조지폐

와 같다. 설명변수는 6군데 지폐의 길이를 측정한 것들이다.

$$x^T = (x_1, x_2, x_3, x_4, x_5, x_6)^T = (\text{bottom, diagonal, left, length, right, top})^T \in R^6$$

정규분포 가정을 만족시키기 위한 변수 변환 절차는 필요 없는 것으로 판명되어 원래 변수들을 표준화하여 SIR과 SAVE를 적용한 결과 SIR은 중심부 공간을 1차원으로 추정하였고 SAVE는 중심부 공간을 2차원으로 추정하였다. 차원축소를 이용한 그래픽 방법이 의미하는 바는 SIR은 6차원 공간에서 1차원으로 SAVE는 6차원에서 2차원으로 설명변수가 형성하는 공간을 축소하여도

원래의 자료가 내포하는 정보의 손실이 없다는 의미이다. <그림 1>은 SIR과 SAVE의 결과를 보여 주고 있다.

SIR을 이용하여 진짜지폐와 위조지폐를 분류하는 절단값을 결정하는 방법은 <그림 1> (a)에서 보는 바와 같이 두 집단을 잘 분류하는 값을 절단값으로 정하는 것이다. SIR한 중심부 공간을 1차원으로 추정된 결과를 가지고 다음과 같은 새로운 하나의 변수 SIR_1 로 지폐의 위조여부를 분류할 수 있다. 횡축(horizontal axis)에서 절단점을 0으로 취하면 SIR_1 이 양수(positive)이면 위조지폐이다. 다시 말해 SIR의 분류 기준은 다음과 같다.

$$SIR_1 = \hat{u}_1^T x = 0.613z_1 - 0.681z_2 - 0.114z_3 - 0.001z_4 + 0.13z_5 + 0.36z_6 > 0$$

이면 위조지폐로 분류하고

$$SIR_1 = \hat{u}_1^T x = 0.613z_1 - 0.681z_2 - 0.114z_3 - 0.001z_4 + 0.13z_5 + 0.36z_6 < 0$$

이면 진짜지폐로 분류한다. 한편 종축(vertical axis)의 SIR_2 는 SIR 방법이 중심 부공간을 1차원으로 추정하였으므로 분류 분석에 도움이 되지 않는다.

<그림 1> (b)는 SAVE 방법이 중심부 공간을 2차원으로 추정된 결과로 제시하는 충분 요약 그림(sufficient summary plot)이다. SAVE 방법은 차원축소를 이루는 새로운 두 변수 중 첫 번째 변수는 횡축의 $SAVE_1$ 로 지폐의 위조여부를 다음과 같이 분류 할 수 있다. <그림 1> (b)로부터 집단을 잘 분류하는 절단점을 0으로 취하면 $SAVE_1$ 이 양수(positive)이면 위조지폐이다. 다시 말해 SAVE의 분류 기준은 다음과 같다.

$$SAVE_1 = \hat{u}_1^T x = 0.768z_1 - 0.48z_2 - 0.064z_3 - 0.011z_4 + 0.09z_5 + 0.41z_6 > 0$$

이면 위조지폐로 분류하고

$$SAVE_1 = \hat{u}_1^T x = 0.768z_1 - 0.48z_2 - 0.064z_3 - 0.011z_4 + 0.09z_5 + 0.41z_6 < 0$$

이면 진짜지폐로 분류한다.

한편 종축(vertical axis)의 SIR_2 는 SIR 방법이 중심 부공간을 2차원으로 추정된 결과로 생성된 두 번째 변수이다. $SAVE_2$ 의 절단점으로 -1.7을 취하면 위조지폐를 또 다시 두 집단으로 분류하고 있다. 이것은 SIR이 제공하지 못하는 부분이다. 이것은 $SAVE_2 = -1.7$ 을 기준으로 위조지폐 집단 내에서의 또 다른 분류를 할 수 있는데 위조지폐 동질성에 대한 단서, 예를 들어 위조지폐 제조 공장이 한군데 이상이라는 단서를 제공 할 수 있는 분류분석으로 해석 할 수 있다. SIR은 각 집단내의 변동성(variability)에 별 차이를 보이고 있지 않으나 SAVE는 위조지폐 집단내의 변동성이 더 큰 것으로 보여준다. <표 1>은 판별분석(discriminant analysis), 로지스틱 회귀모형(logistic regression), SIR, SAVE간의 정분류율(classification rate)을 비교 한 것이다.

<그림 1>과 <표 1>에서 보는 바와 같이 네 가지 분류 방법들 판별분석, 로지스틱 회귀모형, SIR, 그리고 SAVE는 거의 정확한 분류를 하고 있다. 판별분석과 SIR, 그리고 로지스틱 회귀분석과 SAVE는 같은 결과를 보여주고 있다. SAVE는 <그림 1> (b)에서 한 개의 관찰치가 오분류되는 것처럼 보였으나 $SAVE_1$ 에서 절단점을 0으로 하여 양수와 음수로 위조지폐 여부의 기준으로 하였을 경우 정확히 분류되었다. 한편 종축의 $SAVE_2$ 가 제시하는 또 다른 기준은 판별분석, 로지스틱 회귀분석 그리고 SIR 방법이 제시하지 못하는 위조지폐 내에서의 또 다른 분류는 그 자체로서 또 다른 의미를 부여하기에 충분하다고 여겨진다. 판별분석 및 로지스틱 회귀분석을 이용한

<표 1> Swiss Banknote 자료 정분류율

분류방법	y=0	y=1	전체
판별분석	100%	99%	99.5%
로지스틱	100%	100%	100%
SIR	100%	99%	99.5%
SAVE	100%	100%	100%

<그림 1> SIR과 SAVE의 분류: (O=진짜지폐, X=위조지폐)

분류분석 결과는 <그림 1> (a)의 SIR의 결과와 거의 비슷하므로 여기서는 생략한다.

● 재 무 자 료

기업파산에 대한 통계적 예측모형을 개발하기 위해서 4가지 재무비율로 이루어진 설명변수와 46개의 관찰값으로 이루어져있다(김기영, 전명식, 1999, p.233-234).

$Y=0$: 파산기업(21개) $Y=1$: 정상기업(25개)

$$x^T = (x_1, x_2, x_3, x_4)^T \in R^4$$

x_1 = 현금흐름/총부채비율, x_2 = 순이익/총 자산비율,

x_3 = 유동자산/유동부채비율, x_4 = 유동자산/순매출액

SIR과 SAVE를 적용한 결과 SIR은 중심부 공간을 1차원으로 추정하였고 SAVE는 중심부 공간을 2차원으로 추정하였다. 그래픽 방법이 의미하는 바는 SIR은 6차원 공간에서 1차원으로 SAVE는 6차원에서 2차원으로 설명변수가 형성하는 공간을 축소하여도 원래의 자료가 내포하는 정보의

손실이 없다는 의미이다. <그림 2>는 SIR과 SAVE의 결과를 보여 주고 있다.

SIR을 이용하여 정상기업과 파산기업을 분류하는 절단값을 결정하는 방법은 <그림 2> (a)에서 보는 바와 같이 두 집단을 잘 분류하는 값을 절단값으로 정하는 것이다. SIR한 중심부 공간을 1차원으로 추정한 결과를 가지고 다음과 같은 새로운 하나의 변수 SIR_1 로 기업의 생사여부를 분류할 수 있다. 횡축에서 절단점을 0.02로 취하면 SIR의 분류 기준은 다음과 같다.

$$SIR_1 = \hat{u}^T x = -0.141z_1 - 0.938z_2 - 0.190z_3 + 0.252z_4 > 0.02$$

이면 파산기업으로 분류하고

$$SIR_1 = \hat{u}^T x = -0.141z_1 - 0.938z_2 - 0.190z_3 + 0.252z_4 < 0.02$$

이면 정상기업으로 분류한다.

<그림 2> (b)는 SAVE 방법이 제시하는 충분요약 그림이다. SAVE 방법은 2차원으로 중심 부 공간을 추정하여 새로운 설명변수 $SAVE_1$ 과 $SAVE_2$ 을 생성하여 2차원으로 축소가 가능하다고 할 수 있다.

<그림 2> 재무자료의 SIR과 SAVE의 분류: (X: 정상기업, O: 파산기업)

$$SAVE_1 = 0.153z_1 - 0.986z_2 - 0.019z_3 + 0.063z_4$$

$$SAVE_2 = -0.272z_1 - 0.955z_2 - 0.113z_3 + 0.045z_4$$

SAVE 방법의 판별은 <그림 2> (b)에서 보는 바와 같이 다음과 같이 2차원 평면에서 $SAVE_1$ 과 $SAVE_2$ 의 선형식으로 절단식을 구성하고 다음과 같이 분류한다.

$$SAVE = 0.83 - 2.77SAVE_1 - 3.01SAVE_2 > 0$$

이면 파산기업으로 분류하고

$$SAVE = 0.83 - 2.77SAVE_1 - 3.01SAVE_2 < 0$$

이면 정상기업으로 분류한다.

<표 2> 재무자료 정분류율

분류방법	파산기업	정상기업	전체
판별분석	81%	96%	89%
로지스틱	86%	96%	91%
SIR	86%	96%	91%
SAVE	90%	96%	93%

<표 2>에서 보는 바와 같이 2차원의 중심 부공간으로 추정된 SAVE 방법의 정분류율이 판별 분석, 로지스틱 회귀모형이나 SIR보다 높게 나온 것을 확인할 수 있다.

4. 결 론

그래픽 방법은 여러 설명변수들이 본질적으로 내포하고 있는 정보를 잃어버림 없이 고차원 (high dimension)의 문제를 저차원(low dimension)으로 축소시킨 공간에서 모형화를 시키려는 시도이다. 그래픽 방법을 이용한 분류분석을 이용할 때 자료가 분류되어 있는 형태를 눈으로 쉽게 확인할 수 있는 장점이 있다. 그래픽 방법이 요구하는 두 조건인 선형성 조건(linearity condition)과 상수 공분산 조건(constant variance condition)이 현실적으로 실현되지 않는 경우 SIR이나 SAVE 방법들의 문제점 또한 내포하고 있다. 실제 적용에 있어 차원 추정을 위한 검정 통계량이 정규분포에 대한 가정을 요구하므로 이에 대한 논란 또한 있을 수 있으나 정규분포의 가정이 심하게 왜곡되어 있지 않는 한 사용 가능해 보인다(Cook 1994, 1998). 많은 경우 변수변환의 과정을 통하여 선형성 조건과 상수 공분산 조건을 극복할 여지가 있다.

예제에서 본 바와 같이 SIR이나 SAVE와 같은 차원축소에 기초한 방법들은 변수 선택의 관점에서 보면 결과의 해석에 어려움이 있다. 이것은 차원축소를 이용한 기법들이 원래 변수들(original variables)의 선형 결합으로 표시되기 때문이다.

모든 통계 방법들과 마찬가지로 그래픽을 이용한 방법인 SIR이나 SAVE 방법들이 기존의 분류 방법보다 우월하다고 단정할 수는 없다. 그러나 기존의 분류방법으로 널리 사용되는 판별분석이나 로지스틱 회귀모형 등에 대한 새로운 대안 또는 보충적 역할이라고 하는 관점으로 보면 의미있는 시도라고 생각된다. 현재도 논의가 진행되고 있는 차원 축소를 이용한 그래픽 방법론은 분류 방법에 새로운 제시를 할 것으로 여겨진다.

참고문헌

- [1] 김기영, 전명식 (1999). 다변량 통계자료분석, 자유아카데미, 서울
- [2] Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, Vol. 89, 177-189.
- [3] Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, Vol. 91, 983-992.
- [4] Cook, R. D. (1998). *Regression Graphics, Ideas for studying regressions through graphics*. New York: Wiley.
- [5] Cook, R. D. and Lee, H. (1998). Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association*, Vol. 94, 1187-1200.
- [6] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association*, Vol. 86, 328-332.
- [7] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association*, Vol. 86, 328-332.
- [8] Cook, R. D. and Yin, X. (2001). Special Invited Paper: Dimension Reduction and Visualization in Discriminant Analysis. *Australian and New Zealand Journal of Statistics*, Vol. 43, 147-199
- [9] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.
- [10] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, Vol 86, 316-342.
- [11] Tierney, L. (1990). *Lisp-Stat*, New York: Wiley.
- [12] Velilla, S. (1993). A note on multivariate Box-Cox transformations to normality. *Statistics and Probability Letters*, Vol. 17, 315-322.

[2004년 4월 접수, 2004년 5월 채택]