응용통계연구 제10권 2호
1997년 9월 pp. 293-308

다변량 분할 역회귀모형에 관한 연구[1)]

이 용 구[2)], 이 덕 기[3)]

요 약

일변량 분할 역회귀 방법은 일반화 회귀모형에서 효과적인 차원축약방향과 공간을 추정하는 방법 이다. 본 논문에서는 두 일반화 회귀모형을 동시에 고려하여 효과적인 차원축약방향과 공간을 추정하는 방법으로 이변량 분할 역회귀를 제안한다. 이러한 이변량 분할 역회귀 방법은 모형식이 선형, 이차형, 삼차형, 비선형등의 여러 모형식에서 효과적인 차원축약방향을 추정하며, 일변량 분할 역회귀에 비하여 모형에 존재하는 오차에 크게 영향을 받지 않고 효과적인 차원축약방향을 추정한다. 특히 모형식이 대칭의 이차형인 경우에 일변량 분할 역회귀 방법이 효과적인 차원축약방향을 추정하지 못하는 문제를 해결할 수 있다.

1. 서 론

회귀분석에서 반응변수 y와 설명변수 $X_{p\times 1}=[X_1, X_2, \cdots, X_p]^T$로 이루어진 회귀모형은 선형, 비선형등 여러가지 모형이 있다. 이러한 모형을 모두 포괄하는 모형을 일반화 회귀모형이라고 하면, 일반화 회귀모형은 $y=f(X)+\varepsilon$로 나타낼 수 있다. 일반화 회귀모형에서 설명변수 $X_{p\times 1}$의 차원이 높다면 분석이 쉽지않으며, 차원을 축약하는 방법을 생각하게 된다. 만약 p차원의 일반화 회귀모형이 차수가 $k\times p$, $(k\leq p)$인 계수행렬 B에 대하여 $y=f(B^TX)+\varepsilon$와 같이 표현될 수 있다면 p차원의 일반화 회귀모형은 보다 차원이 낮은 k차원의 설명변수들의 선형결합으로 표현 가능하고 이 모형을 차원축약모형(dimension reduction model)이라고 한다. 특히 $k=1, 2$인 경우에는 반응변수와 설명변수의 관계는 2차원 또는 3차원 공간상에서 반응변수와 설명변수의 산포도를 통해서 쉽게 파악할 수 있다.

차원축약모형에서 k차원의 설명변수들의 선형결합을 효과적인 차원축약방향(effective dimension reduction direction)이라고 하며, k차원의 설명변수들로 생성된 선형부분공간 $\mathbf{B}$를 효과적인 차원축약공간(effective dimension reduction space)이라고 한다. 이러한 효과적인 차원축약방향과 공간을 추정하기 위한 방법으로는 일변량 분할 역회귀(sliced inverse regression)와 분할 평균분산 추정법(sliced average variance estimates), 주 헤이시안 방향(principal Hessian

1) 본 논문은 1996년도 중앙대학교 교내 연구비 지원에 의하여 연구하였음.
2) 서울시 동작구 흑석동 221 중앙대학교 정경대학 응용통계학과 교수
3) 서울시 동작구 흑석동 221 중앙대학교 정경대학 응용통계학과 강사

direction)이 있다.

본 연구에서는 분할 역회귀방법을 두 반응변수 모형에 이용하여 효과적인 차원축약방향과 공간을 추정하는 방법인 이변량 분할 역회귀방법에 대하여 연구한다.

2. 차원축약모형

2.1 차원축약모형

일반화 회귀모형은

$$y = f(X) + \varepsilon \tag{2.1}$$

여기서 f는 공간 R^p상에서의 미지의 함수

$\varepsilon \amalg X$, $\amalg$는 독립표시임.

으로 표현할 수 있다. 그러나 설명변수 $X_{p\times 1}$의 차원 p가 크다면 분석이 쉽지않으며, 따라서 차원을 축약하는 방법들을 생각하게 된다. 차원축약방법의 하나로 고차원의 데이타를 저차원으로 투영(projection)시키는 방법을 생각할 수 있다.

만약 모형식 (2.1)이 $k(<p)$개의 $p\times 1$벡터 $\beta_1, \beta_2, \cdots, \beta_k$에 의해서

$$y = f(\beta_1^T X, \beta_2^T X, \cdots, \beta_k^T X) + \varepsilon \tag{2.2}$$

여기서 β_i, $(i = 1, \cdots, k)$는 미지의 벡터

f는 공간 R^{k+1}상에서의 미지의 함수

$\varepsilon \amalg X$

와 같이 표현될 수 있다면, 이 모형식은 p차원의 설명변수 X를 정보의 손실없이 k차원으로 축약하는 모형이 된다. 따라서 모형식 (2.2)는 k차원의 변수 $\beta_1^T X, \beta_2^T X, \cdots, \beta_k^T X$를 통해서만 나타나는 X에 대한 y의 조건부분포라고 할 수 있으며, 따라서 조건부변수로 $\beta_i^T X$, $(i = 1, \cdots, k)$들이 주어졌을 때 y와 X는 독립이고, 이것은 축약된 k차원의 변수 $\beta_1^T X, \beta_2^T X, \cdots, \beta_k^T X$가 원래의 X가 y에 대하여 가지고 있는 정보를 손실없이 가지고 있다는 것이 된다. 특히 모형식 (2.2)에서 k가 1 또는 2라면 2차원 또는 3차원의 X와 y의 산포도를 이용하여 X와 y의 관계를 보다 쉽게 파악해 볼 수 있다.

2.2 분할 역회귀

역회귀분석은 반응변수와 설명변수의 역할을 바꾸어 분석하는 것이다. 즉, y에 대한 X의 조건부 기대값 $E(X \mid y)$를 구하는 것이라 할 수 있다. 이처럼 X와 y의 역할을 바꾸므로서 차원축

약의 효과를 얻을 수 있다. X를 평균벡터가 0이고 분산 · 공분산행렬이 단위행렬이 되도록 표준화시킨 후, y의 값을 여러개의 구간으로 분할(slice)하고, 분할된 각 구간의 y값에 대하여 그 구간에 속한 X의 변수벡터 별로 역회귀를 수행한다. 이때 $E(X \mid y)$는 y의 각 구간별로 y값의 변화에 따라 얻어지는 X의 분할평균(slice mean)이다. 이렇게 구해진 y의 각 구간별 $E(X \mid y)$의 값들에 대하여 주성분분석을 이용해서 유용한 주성분들을 구하면 이 주성분들이 효과적인 차원축약방향의 추정값이 된다. 또한 역회귀에서 y에 대한 X의 조건부 기대값 $E(X \mid y)$는 y가 변함에 따라 곡선을 그리게 되고, 이 곡선을 역회귀곡선(inverse regression curve)이라고 한다.

역회귀곡선의 중심은 $E[E(X \mid y)] = E(X)$에 위치하고, 이 역회귀곡선에 대한 중심 역회귀곡선(centered inverse regression curve)은 $E(X \mid y) - E(X)$으로 정의하며 R^p에 존재한다. 그러나 다음의 [조건 1]과 [정리 1]에 의해서 중심역회귀곡선은 축약된 k차원의 부분공간에 존재하게 된다.

[조건 1] (Li, 1991)

R^p에서 임의의 벡터 b에 대하여 조건부 기대값 $E(b^T X \mid \beta_1^T X, \cdots, \beta_k^T X)$은 $\beta_1^T X, \cdots, \beta_k^T X$에 대하여 선형이다. 즉, 임의의 상수들 $c_0, \cdots, c_k$에 대하여

$$E(b^T X \mid \beta_1^T X, \cdots, \beta_k^T X) = c_0 + c_1 \beta_1^T X + \cdots + c_k \beta_k^T X$$

이다. 이 조건은 X의 분포가 타원으로 대칭일 때, 예를들면 정규분포에서 만족된다. ■

[정리 1] (Li, 1991)

모형식 (2.2)와 [조건 1]이 만족되면, 중심 역회귀곡선 $E(X \mid y) - E(X)$는 $\beta_k \sum_{xx}$에 의해서 생성된 k차원의 선형부분공간상에 포함된다. 여기서 $\sum_{xx}$는 X의 공분산행렬이다. ■

X를 표준화하면 $Z = \sum_{xx}^{-1/2} [X - E(X)]$이고, 따라서 모형 (2.2)는

$$y = f(\eta_1 Z, \cdots, \eta_k Z) + \varepsilon \tag{2.3}$$

여기서 $\eta_k = \beta_k \sum_{xx}^{1/2}$

으로 표현할 수 있다. X를 표준화시킨 Z에 대한 y의 역회귀곡선을 표준화된 역회귀곡선이라 하며, 이 곡선은 다음의 [따름정리 1]에 의해서 표준화된 효과적인 차원축약공간에 포함된다.

[따름정리 1] (Li, 1991)

X를 Z로 표준화시키면 표준화된 역회귀곡선 $E(Z \mid y)$는 표준화된 효과적인 차원축약방향인 $\eta_1, \eta_2, \cdots, \eta_k$들에 의해서 생성된 선형부분공간에 포함된다. ■

3. 이변량 분할 역회귀

3.1 이변량 분할 역회귀

이변량회귀의 일반적인 모형은

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = f(B^T X) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \tag{3.1}$$

$$X \amalg (\varepsilon_1, \varepsilon_2)$$

이다. 이 모형을 나누어서 표현하면,

$$y_1 = f_1(B_1^T X) + \varepsilon_1 \ , \ B_1 = [\beta_{11}, \cdots, \beta_{1k_1}]$$
$$y_2 = f_2(B_2^T X) + \varepsilon_2 \ , \ B_2 = [\beta_{21}, \cdots, \beta_{2k_2}]$$

이고, 각 식에서 B_1과 B_2가 올바르게 추정되었다면

$$y_1 \amalg X \mid B_1^T X \quad , \quad y_2 \amalg X \mid B_2^T X$$

와 같은 관계가 성립한다. 즉 p차원의 설명변수 X를 k_1과 k_2차원으로 축약하여도 정보의 손실없이 y_1과 y_2를 잘 설명하고 있음을 나타낸다. 그러나 본 연구에서는 y_1, y_2를 동시에 설명할 수 있는 효과적인 차원축약공간 $\mathcal{B}$의 기저 B를 추정하는데 목적이 있다. 즉,

$$(y_1, y_2) \amalg X \mid B^T X \tag{3. 2}$$

으로 p차원의 설명변수 X를 p보다 작은 k차원으로 축약하면서도 정보의 손실없이 y를 잘 설명하는 효과적인 차원축약공간 $\mathcal{B}$를 추정하는 것이다. $\mathcal{B}$의 추정은 Li(1991)에 의해서 제시된 SIR방법을 두 반응변수 모형에 적용할 수 있다. 즉, SIR방법의 논리적인 근거인 Duan과 Li(1991)의 정리 2.1과 보조정리 2.2는 반응변수 y가 2개인 경우에도 동일하게 적용될 수 있다.

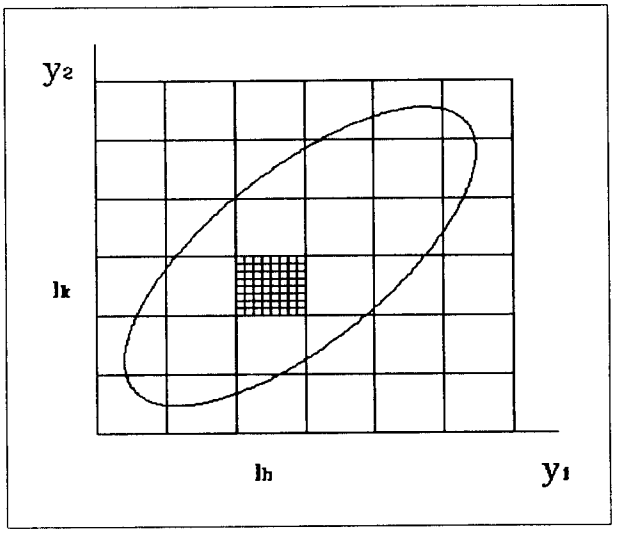
이변량 분할 역회귀를 이용하여 효과적인 차원 축약방향과 공간을 추정하는 알고리즘은 다음과 같다. 이변량 분할 역회귀는 y_1을 H개의 구간으로 분할하고, 다시 y_2를 K개의 구간으로 분할하여 형성된 격자 안에 속하는 설명변수 X의 각 변수 벡터들의 조건부평균을 이용하여 효과적인 차원축약방향과 공간을 추정하는 방법이다. 이것을 그림 으로 표현하면 [그림 3.1]과 같다.

1단계, X를 표준화하여 $\hat{X}_i = \hat{\Sigma}_{xx}^{-1/2}(X_i - \overline{X})$ $(i = 1, \cdots, n)$를 구한다.

여기서 $\hat{\Sigma}_{xx}$는 표본의 분산 · 공분산행렬이고, $\overline{X}$는 X의 표본평균이다.

2단계, y_1을 H개의 구간 $I_1, \cdots, I_H$로 분할하고, y_2를 K개의 구간 $I_1, \cdots, I_K$으로 분할한다.

y_1이 h번째 구간에 포함되고, 동시에 y_2가 k번째 구간에 포함 될 확률은 다음과 같다.



[그림 3.1] 이변량 분할 역회귀

$$\hat{p}_{hk} = \frac{1}{n} \sum_{i=1}^{n} \delta_{hk}(y_{1i}, y_{2i})$$

$$h = 1, 2, \cdots, H, \quad k = 1, 2, \cdots, K$$

여기서 $\delta_{hk}(y_{1i}y_{2i})$는 y_{1i}가 h번째 구간에 속하고, y_{2i}가 k번째 구간에 속하면 1을 취하고 아니면 0값을 취하는 지시함수이다.

3단계, 각각의 분할된 구간내에서 $\hat{X}_i$들의 표본평균을 구한다. 이 표본평균을 $\hat{m}_{hk}$라고 한다면

$$\hat{m}_{hk} = \frac{1}{n\ \hat{p}_{hk}} \sum_{y_{1i} \in I_h,\ y_{2i} \in I_k} \hat{X}_i \quad (h = 1, \cdots, H,\ k = 1, 2, \cdots, K)$$

이다.

4단계, $\hat{m}_{hk}$에 대한 가중 주성분분석을 수행한다. 우선 추정된 가중 공분산행렬

$$\hat{V} = \sum_{h=1}^{H} \sum_{k=1}^{K} \hat{P}_{hk}\ \hat{m}_{hk}\ \hat{m}_{hk}^{T}$$

를 구한다. 그리고 $\hat{V}$에 대한 고유치와 고유벡터를 구한다.

5단계, 고유치들 가운데 크기 순으로 k개의 고유치에 대응하는 고유벡터들이 표준화된 효과적인 차원축약방향 $\hat{\eta}_k$ $(k = 1, \cdots, K)$가 되고, 표준화된 효과적인 차원축약방향을 원래의 척도로 변환하여 효과적인 차원축약방향의 추정치 $\hat{\beta}_k = \hat{\eta}_k \sum_{xx}^{-1/2}$ $(k = 1, \cdots, K)$를 구한다. 그리고 추정된 효과적인 차원축약방향 $\hat{\beta}_1$, $\hat{\beta}_2, \cdots, \hat{\beta}_k$를 이용하여 효과적인

차원축약공간 $\mathbf{B}$의 추정공간인 $\hat{\mathbf{B}}$을 구한다.

이와같은 절차를 통해서 두 일반화 회귀모형을 동시에 설명할 수 있는 효과적인 차원축약방향과 공간이 추정된다. 이러한 이변량 분할 역회귀방법과 일변량 분할 역회귀방법의 차이는 분할하는데 고려하는 변수의 수에 달려 있으며, 이것은 결국 분할된 구간의 수를 변화시키는 효과를 가져온다. 그러나 일변량 분할 역회귀방법의 성질에서 분할된 구간의 수는 효과적인 차원축약방향의 추정에 큰 영향을 주지 않는 것으로 나타났다. 따라서 이변량 분할 역회귀에서 조건부변수로 어떠한 변수를 고려해도 일변량 분할 역회귀의 조건과 성질을 만족시킨다고 할 수 있다.

3.2 이변량 분할 역회귀의 응용

일반화 회귀모형 $y=f(X)+\epsilon$의 차원축약공간 추정에 이변량 분할 역회귀방법을 응용할 수 있다. 설명변수 X의 투영행렬 $H=X(X^TX)^{-1}X^T$를 이용하여 y를 X의 열 공간으로 투영한 부분과 X의 열 공간에 직교하는 공간으로 투영한 부분으로

$$\begin{aligned} y &= Hy+(I-H)y \\ &= \hat{y}+r \end{aligned} \tag{3.3}$$

와 같이 분해한 후에 $\hat{y}$과 r을 동시에 고려하여 효과적인 차원축약방향과 공간을 추정한다. y를 $\hat{y}$과 r로 나누어 고려하므로서 얻어지는 잇점은 다음과 같다.

첫째, y에 대한 정보를 잃지않는 가운데 알려져 있지 않은 y와 X의 관계를 $\hat{y}$과 r의 산포도를 통해서 잠정적으로 파악 할 수 있다.

둘째, $\hat{y}$과 r의값을 조건부 변수로 한 X에 대한 조건부 기대값 $E\left[X \mid \begin{pmatrix}\hat{y}\\ r\end{pmatrix}\right]$을 이용하므로 $\hat{y}$과 r의 산포도를 통해서 잠정적으로 파악된 y와 X의 관계를 고려하여 효과적인 차원축약방향과 공간을 추정 할 수 있다.

세째, 자료를 두번 분할하므로 자료에 존재하는 오차항에 크게 영향을 받지않고 효과적인 차원축약방향을 추정할 수 있다.

네째, 잔차 r을 고려하므로 모형식이 선형, 비선형등 어떠한 모형에서도 비교적 안정적으로 효과적인 차원축약방향을 추정할 수 있다.

3.3 이변량 분할 역회귀의 응용사례

조건부변수로 $\hat{y}$과 r을 고려한 이변량 분할 역회귀방법이 얼마나 안정적으로 효과적인 차원축약방향을 추정하는가를 여러가지 모형을 이용하여 모의실험 하였다. 모의실험에서는 모형식이 선형, 이차형, 삼차형, 비선형등의 여러가지 모형에서 일변량 분할 역회귀와 이변량 분할 역회귀,

분할 평균분산 추정법으로 추정한 효과적인 차원축약방향과 실제 효과적인 차원축약방향과의 비교를 실시하였다. 즉, 어떤 방법이 여러 모형식에서 안정적으로 실제 효과적인 차원축약방향에 가까운 효과적인 차원축약방향을 추정하는가에 대하여 비교해 보았다. 추정된 효과적인 차원축약방향이 실제 효과적인 차원축약방향에 얼마나 가깝게 추정되었는지를 판정하는 기준으로 추정된 효과적인 차원축약방향과 실제 효과적인 차원축약방향과의 다중상관계수제곱과 코사인값 그리고 추정된 효과적인 차원축약방향과 실제 효과적인 차원축약공간이 이루는 각을 이용하였다. 모의실험은 각 모형식에서 100번씩 반복 실시하였으며, 추정된 효과적인 차원축약방향의 평균과 표준편차, 그리고 다중상관계수제곱과 코사인값, 각도의 평균과 표준편차를 결과로 제시하였다. 또한 모형식이 대칭의 이차형인 실제자료에 대하여 일변량 분할 역회귀와 이변량 분할 역회귀, 분할 평균분산 추정법을 이용한 분석을 실시하여 그 결과를 비교하였다.

모의실험에서 이용한 이변량 분할 역회귀의 프로그램은 Tierney(1990)에 의해서 만들어진 LISP-STAT을 이용하여 작성하였으며, 작성된 프로그램을 Cook과 Weisberg(1994)에 의해서 만들어진 회귀분석 프로그램인 R-code의 역회귀분석에 하나의 메뉴로 추가하였다. 일변량 분할 역회귀와 분할 평균분산 추정법은 R-code에서 제공하는 프로그램을 이용하였다.

1] 모형식이 선형인 경우

Li(1991)에서 제시한 모형식

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \tag{3.4}$$

을 이용하여 모의실험을 하였다. 여기서 x_1, x_2, x_3, x_4, x_5와 ε은 서로 독립인 표준정규난수이고 자료의 수 $n = 100$이다. 모수벡터는 $\beta^T = (0.5, 0.5, 0.5, 0.5, 0.0)$을 이용하였다. 모의실험 에서는 모형식(3.4)에서 오차항의 크기를 변화시켰을때 일변량 분할 역회귀와 이변량 분할 역회귀 방법이 얼마나 안정적으로 효과적인 차원축약방향을 추정하는지를 모의실험 하였다. 오차항의 크기 는 $k \cdot \varepsilon (k = 0, 0.5, 1, 1.5, 2)$로 변화시켜서 실험하였다. 모의실험은 각 k의 수준에서 100번씩 반복 실험하였으며, [표 3.1]에 실험의 결과를 제시하였다.

[표 3.1] 모형식 (3.4)에서 오차항의 수준변화에 따른 일변량 분할 역회귀와 이변량 분할 역회귀의 추정량 b와 실제모수 $\beta = (0.5, 0.5, 0.5, 0.5, 0)$와의 $R^2(b)$, angle, cosangle의 평균과 표준편차의 비교.

분석방법	일변량 분할 역회귀			이변량 분할 역회귀		
통계량	$R^2(b)$	Angle	Cosangle	$R^2(b)$	Angle	Cosangle
k=0.0	1.00(0.00)	1.66(0.81)	1.00(0.00)	0.93(0.07)	14.26(8.18)	0.96(0.04)
k=0.5	0.99(0.01)	6.46(2.13)	0.99(0.00)	0.98(0.01)	7.09(2.87)	0.99(0.01)
k=1.0	0.94(0.05)	13.30(5.42)	0.97(0.03)	0.96(0.03)	11.39(3.61)	0.98(0.01)
k=1.5	0.84(0.13)	22.37(9.85)	0.91(0.07)	0.92(0.05)	16.27(5.70)	0.95(0.09)
k=2.0	0.64(0.29)	36.46(19.71)	0.76(0.24)	0.89(0.07)	19.43(6.33)	0.92(0.18)

[표 3.1]에 의할 때 k=0인 경우 즉, 오차항이 존재하지 않는 경우에는 일변량 분할 역회귀로 추정된 효과적인 차원축약방향이 이변량 분할 역회귀로 추정된 효과적인 차원축약방향 보다 실제모수에 더 근사하게 추정되고 있음을 알 수 있다. 이것은 오차항이 존재하지 않는 완전한 모형의 경우에는 $\hat{y}$이 y의 정보를 대부분 확보하고 있으며, 잔차 r은 상대적으로 y에 대한 정보를 적게 가지고 있기 때문에 y대신에 $\hat{y}$과 r을 이용하는 경우 이차원 분할의 제약으로 근사정도가 낮게 나타났다. 그러나 k의 값이 0.5, 1.0, 1.5, 2.0등과 같이 커질수록 일변량 분할 역회귀방법에 의한 추정의 정확성은 급속하게 떨어지는데 비하여 이변량 분할 역회귀방법에 의한 추정의 정확성은 안정적 수준을 유지함을 알 수 있다.

2] 모형식이 이차형인 경우

Cook과 Weisberg(1991)에 의하면 모형식이 이차의 대칭형태일 때 일변량 분할 역회귀를 이용한 효과적인 차원축약방향의 추정은 정의되지 않는다. 그러나 $\hat{y}$과 r을 고려한 이변량 분할 역회귀 방법은 이차의 대칭형태인 모형식에서도 효과적인 차원축약방향을 잘 추정하며, 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 매우높다. 이것은 $\hat{y}$이 설명하지 못한 이차형의 효과를 잔차 r에서 고려했기 때문이다. 모의실험을 통해서 이와같은 사실을 알아보았으며, 모형식이 이차의 대칭형태일 때 일변량 분할 역회귀의 대안으로 Cook과 Weisberg(1991)에 의해 제시되었던 분할 평균분산 추정법과의 비교도 실시하였다. 모의실험에서 설정한 모형은

$$y = (\mu + 0.7071x_1 + 0.7071x_2)^2 \qquad (3.5)$$

으로 Cook과 Weisberg(1991)에서 분할 평균분산 추정법의 모의실험에 이용했던 모형식이다. 여기서 변수의 수 p=2 이고, 자료의 수는 n=100 이며, x_1, x_2는 표준정규난수이다.

모의실험은 μ를 0, 0.25, 0.5, 1, 2, 4, 8, 100으로 변화시켜서 완전한 이차형으로부터 선형의 효과를 증가시키면서 일변량 분할 역회귀와 이변량 분할 역회귀, 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향과 실제모수의 다중상관계수제곱, angle, cosangle을 비교하였다. 모의실험의 결과는 [표 3.2]에 제시하였다.

[표 3.2]에 의할 때 완전 이차형인 $\mu = 0$인 경우와 $\mu = 0.25$인 경우에는 이변량 분할 역회귀방법과 분할 평균분산 추정법에 의한 추정결과의 정확성은 높으나 일변량 분할 역회귀에 의한 추정의 정확성은 매우 낮다. 반면에 μ의 값이 0.5, 1.0, 2.0과 같이 커질수록 모형이 선형화하며 이 경우에는 일변량 분할 역회귀와 이변량 분할 역회귀에 의한 추정결과의 정확성은 높으나 분할 평균분산 추정법에 의한 추정결과의 정확성은 상대적으로 낮음을 알 수 있다. 즉 이변량 분할 역회귀방법에 의한 추정결과는 선형과 이차형의 경우 모두 안정적인 추정을 할 수 있는데 비하여 일변량 분할 역회귀방법은 이차형에서 그리고 분할 평균분산 추정법은 일차형에서 추정의 정확성이 떨어진다.

모의실험 결과를 통해서 얻을 수 있는 결론은 일변량 분할 역회귀방법은 모형식이 이차형일 때 효과적인 차원축약방향의 추정에 있어서 불안정하다는 것이다. 그러나 $\hat{y}$과 r을 이용하는 이변량 분할 역회귀방법은 모형식이 이차형일 때도 효과적인 차원축약방향의 추정에 있어서 안

정적이며, 선형의 효과를 증가시켜도 큰 변화없이 안정적으로 효과적인 차원축약방향을 추정하고 있음을 알 수 있다.

[표 3.2] 일변량 분할 역회귀, 이변량 분할 역회귀, 분할 평균분산 추정법으로 추정된 효과적인 원축약방향과 실제모수 $\beta = (0.7071, 0.7071)$의 다중상관계수제곱과 angle, cosangle 의 비교.

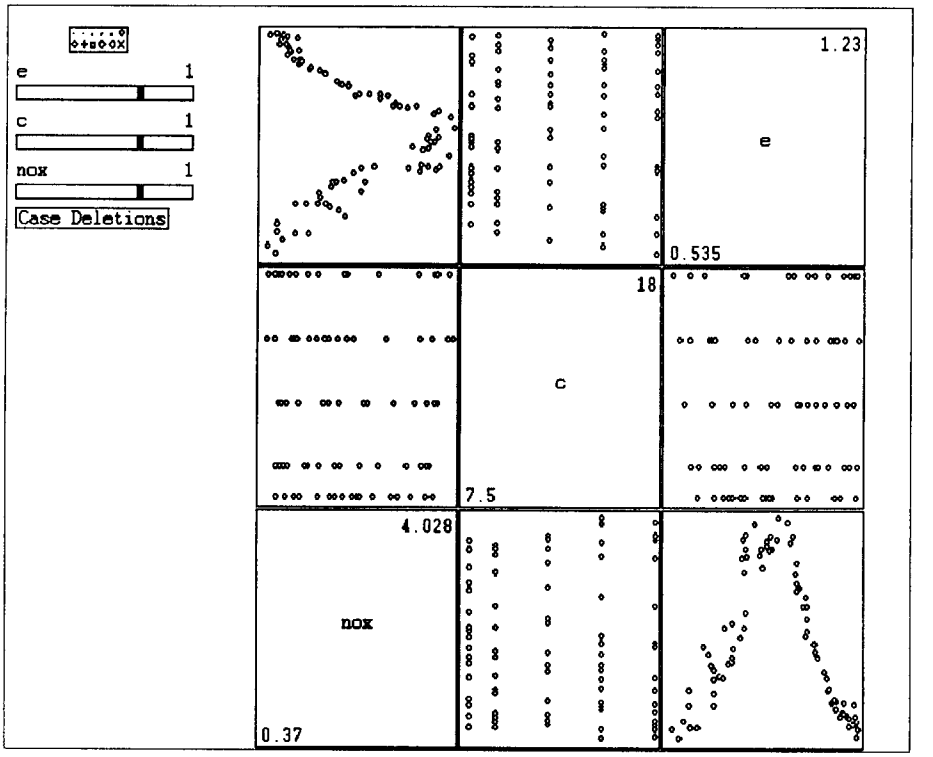
	일변량 분할 역회귀			이변량 분할 역회귀			분할 평균분산 추정법		
통계량	$R^2(b)$	Angle	Cosangle	$R^2(b)$	Angle	Cosangle	$R^2(b)$	Angle	Cosangle
μ=0.0	0.53(0.40)	43.17(31.1)	0.63(0.36)	0.95(0.08)	9.56(9.59)	0.97(0.05)	0.98(0.04)	6.41(5.89)	0.99(0.02)
μ=0.25	0.72(0.32)	29.34(24.9)	0.80(0.27)	0.98(0.05)	6.08(6.92)	0.99(0.03)	0.94(0.16)	10.35(14.11)	0.96(0.14)
μ=0.50	0.94(0.14)	10.02(13.1)	0.96(0.11)	0.99(0.03)	3.97(5.03)	0.99(0.02)	0.92(0.18)	13.0(14.5)	0.95(0.14)
μ=1.0	0.99(0.01)	2.04(2.54)	0.99(0.00)	0.99(0.01)	3.04(1.64)	0.99(0.00)	0.91(0.22)	13.0(17.78)	0.93(0.17)
μ=2.0	0.99(0.00)	2.10(1.02)	0.99(0.00)	0.99(0.00)	2.12(1.03)	0.99(0.00)	0.95(0.20)	6.11(16.61)	0.96(0.17)
μ=4.0	0.99(0.00)	0.64(0.50)	0.99(0.00)	0.99(0.00)	1.39(1.05)	0.99(0.00)	0.97(4.07)	4.07(13.53)	0.97(0.13)
μ=8.0	0.99(0.00)	0.54(0.50)	0.99(0.00)	0.99(0.00)	1.41(1.28)	0.99(0.00)	0.95(0.10)	6.45(8.65)	0.97(0.10)
μ=100	0.99(0.00)	0.64(0.52)	0.99(0.00)	0.99(0.00)	1.37(1.11)	0.99(0.00)	0.95(0.13)	6.54(10.09)	0.97(0.11)

3] 실제자료의 분석 사례

모형식이 이차형인 실제자료를 이용하여 일변량 분할 역회귀와 이변량 분할 역회귀 그리고 분할 평균분산 추정법을 이용하여 분석한 결과를 비교해 보았다. 분석에서 이용한 자료는 Cook과 Weisberg(1994) 에서 이용했던 에탄올자료이다. 에탄올자료는 연료로 에탄올을 이용하는 엔진의 연료소모에 대한 실험자료로 반응변수(NOx)는 엔진이 정상적으로 작동할때 질소 산화물의 농도와 질소 이산화물의 농도를 더한 것으로 측정 단위는 1 joule당 NOx의 micrograms이다. 설명변수의 수는 $p = 2$로 첫번째 설명변수(E)는 엔진이 작동되고 있을때 연료와 공기의 혼합농도를 측정한 것이고, 두번째 설명변수(C)는 엔진의 압축비율이다. 자료는 88개의 관측치로 이루어져 있다.

반응변수 NOx와 설명변수 E와 C에 대한 대략적인 관계를 파악하기 위해서 산포도 행렬(scatterplot matrix)을 그려보면 [그림 3.2]과 같다.

[그림 3.2]에서 반응변수(NOx)와 설명변수(C)는 특정한 형태가 없이 나타나고 있으며, 반응변수와 설명변수(E)는 이차형의 관계가 뚜렷이 나타나고 있고, 자료에 이분산이 존재하고 있음을 알 수 있다. 이러한 에탄올자료에 대하여 일변량 분할 역회귀와 $\hat{y}$과 r을 고려한 이변량 분할 역회귀, 분할 평균분산 추정법을 이용한 분석결과는 [표 3.3]과 [그림 3.3], [그림 3.4], [그림 3.5]와 같다.



[그림 3.2] 반응변수(NOx)와 설명변수(C, E)의 산포도 행렬

[표 3.3] 일변량 분할 역회귀, 이변량 분할 역회귀, 분할 평균분산 추정법으로 추정한 추정치와 R^2값.

분석방법 / 변수	일변량 분할 역회귀	이변량 분할 역회귀	분할 평균분산 추정법
C	-0.086 (0.858)	0.009 (0.174)	0.016 (0.299)
E	0.996 (0.514)	1.000 (0.985)	1.000 (0.954)
R^2	0.064	0.996	0.996

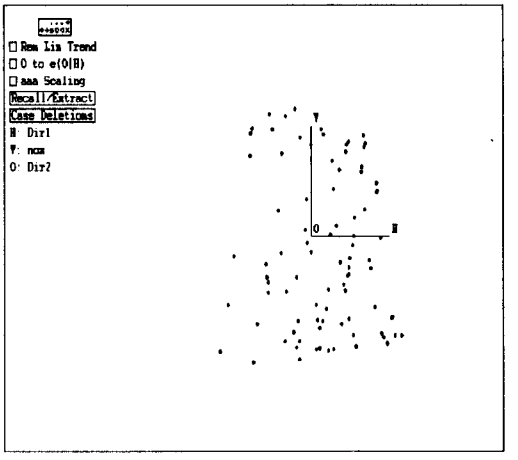
일변량 분할 역회귀방법의 분석결과를 보면, 추정된 효과적인 차원축약방향이 반응변수 y를 설명 하고 있는 설명력은 0.064로 추정된 효과적인 차원축약방향이 반응변수 y를 잘 설명하지 못

하며, 이것은 일변량 분할 역회귀에 의한 차원축약이 잘 이루어지지 않았다는 것을 의미한다. 또한 [그림 3.3]에서도 추정된 효과적인 차원축약방향이 자료에 이차형의 효과가 존재하고 있음을 나타내 주지 못하고 있다.

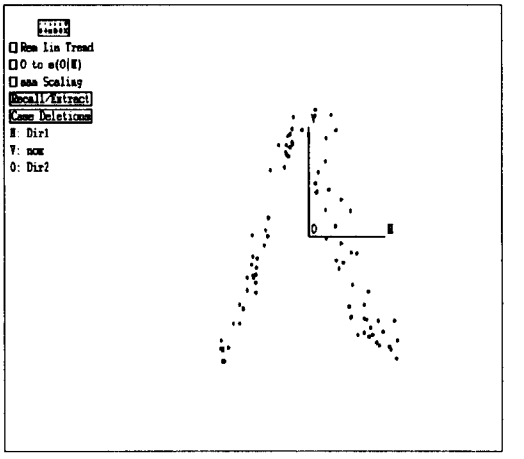
$\hat{y}$과 r을 고려하는 이변량 분할 역회귀의 분석결과를 보면, 추정된 효과적인 차원축약방향이 반응 변수 y를 설명하고 있는 설명력은 0.996이므로 추정된 효과적인 차원축약방향이 반응변수 y를 잘 설명하고 있음을 알 수 있다. 이변량 분할 역회귀로 추정된 효과적인 차원축약방향과 반응변수의 관계를 보여주는 산포도는 [그림 3.4]으로 추정된 효과적인 차원축약방향이 자료에 이차형의 효과가 존재하고 있음을 뚜렷이 나타내고 있다.

분할 평균분산 추정법의 결과를 보면, 추정된 효과적인 차원축약방향이 반응변수 y를 설명하고 있는 설명력은 0.996이므로 추정된 효과적인 차원축약방향이 반응변수 y를 잘 설명하고 있음을 알 수 있다. 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향과 반응변수의 관계를 보여주는 산포도는 [그림 3.5]와 같다. 그림을 보면 추정된 효과적인 차원축약방향은 자료에 이차형의 효과가 존재하고 있음을 뚜렷이 나타내고 있다.

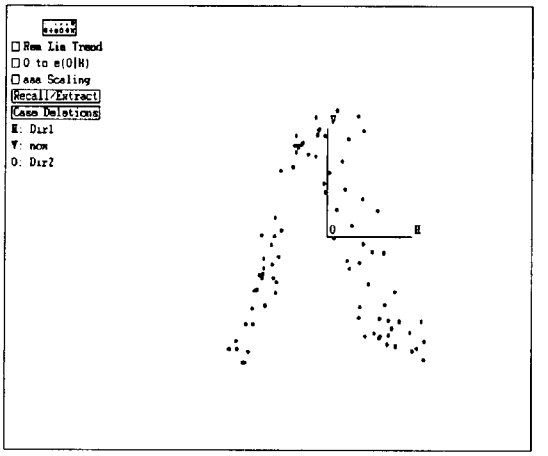
실제자료의 분석결과 모형식이 대칭의 형태를 띠는 이차형일 때 일변량 분할 역회귀에 의한 차원축약은 잘 이루어지지 않음을 알 수 있다. 그러나 $\hat{y}$과 r을 이용한 이변량 분할 역회귀에 의한 차원축약은 잘 이루어지고 있으며, 분할 평균분산 추정법에 의한 차원축약 결과와 유사함을 알 수 있다.



[그림 3.3] 일변량 분할 역회귀에 의해서 추정된 효과적인 차원축약방향과 반응변수의 산포도.



[그림 3.4] 이변량 분할 역회귀에 의해서 추정된 효과적인 차원축약방향과 반응변수의 산포도.



[그림 3.5] 분할 평균분산 추정법에 의해서 추정된 효과적인 차원축약방향과 반응변수의 산포도.

3] 모형식이 삼차형인 경우

모형식이 삼차형일 때 일변량 분할 역회귀와 이변량 분할 역회귀, 분할 평균분산 추정법에 의한 차원축약이 잘 이루어지는가를 모의실험 하였다. 모의실험에서 설정한 모형은

$$y = (0.7071x_1 + 0.7071x_2)^3 + \varepsilon \tag{3.6}$$

이고, 여기서 변수의 수 p=2 이고, 자료의 수는 n=100 이며, x_1, x_2와 ε은 서로 독립인 표준정규난수이다. 일변량 분할 역회귀와 이변량 분할 역회귀 그리고 분할 평균분산 추정법에 의해서 추정된 효과적인 차원축약방향과 실제모수의 다중상관계수제곱과 cosangle, angle을 비교한 모의실험의 결과는 [표 3.4]에 제시하였다.

[표 3.4] 일변량 분할 역회귀, 이변량 분할 역회귀, 분할 평균분산 추정법에 의한 효과적인 차원축약방향과 실제모수 $\beta = (0.7071, 0.7071)$의 다중상관계수제곱과 angle, cosangle의 비교. (괄호안의 값은 표준편차)

	일변량 분할역회귀	이변량 분할역회귀	분할 평균분산 추정법
b1	0.7062 (0.07)	0.7034 (0.05)	0.6717 (0.26)
b2	0.7021 (0.06)	0.7068 (0.05)	0.6412 (0.27)
$R^2(b)$	0.992 (0.02)	0.994 (0.01)	0.611 (0.39)
angle	4.249 (3.06)	3.247 (2.90)	35.801 (29.58)
cosangle	0.996 (0.00)	0.997 (0.00)	0.713 (0.32)

모의실험 결과를 보면, 모형식이 삼차형일 때 일변량 분할 역회귀로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 매우 높음을 알 수 있으며, $\hat{y}$과 r을 이용한 이변량 분할 역회귀로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도 또한 매우 높음을 알 수 있다. 그러나 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 매우 낮음을 알 수 있다. 이것은 분할 평균분산 추정법이 반응변수 y의 각 분할된 구간에서 조건부 분산을 이용하므로 모형식이 삼차형인 경우에 y에 대한 조건부 분산은 각 분할된 구간에서 크게 달라지지 않고 유사한 값을 갖게되고, 따라서 조건부 분산을 이용하여 구한 분할 평균분산 추정치의 고유치는 모두 유사한 값을 갖게되며, 결국 큰 고유치에 대응하는 고유벡터를 이용하여 효과적인 차원축약방향을 추정한다는 것은 어렵게 된다. 그러나 $\hat{y}$과 r을 이용한 이변량 분할 역회귀 방법은 모형식이 삼차형일 때 $\hat{y}$이 설명하지 못한 삼차형의 효과를 잔차 r을 통해서 다시 고려하므로 실제모수에 근사정도가 높은 효과적인 차원축약방향을 추정한다.

4] 모형식이 비선형인 경우

모형식을 비선형으로 설정했을 때 일변량 분할 역회귀와 이변량 분할 역회귀, 분할 평균분산 추정법에 의한 차원축약이 잘 이루어지는가를 모의실험 하였다. 모의실험에서 설정한 모형은

$$y = 0.7e^{0.7071x_1 + 0.7071x_2} + \epsilon \tag{3.7}$$

이고, 여기서 변수의 수 p=2 이고, 자료의 수는 n=100 이며, x_1, x_2와 ϵ은 서로 독립인 표준정규난수이다. 일변량 분할 역회귀와 이변량 분할 역회귀, 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향과 실제모수의 다중상관계수제곱과 cosangle, angle을 비교한 모의실험의 결과는 [표 3.5]에 제시하였다.

[표 3.5] 일변량 분할 역회귀, 이변량 분할 역회귀, 분할 평균분산 추정법으로 추정한 효과적인 차원축약방향과 실제모수 $\beta = (0.7071, 0.7071)$의 다중상관계수제곱과 angle, cosangle의 비교. (괄호안의 값은 표준편차)

	분할 역회귀	이변량 분할 역회귀	분할 평균 분산 추정법
b1	0.6923 (0.18)	0.7127 (0.12)	0.6125 (0.33)
b2	0.6766 (0.18)	0.6827 (0.11)	0.6584 (0.31)
$R^2(b)$	0.922 (0.14)	0.975 (0.03)	0.326 (0.30)
angle	12.741 (12.74)	7.485 (5.43)	56.899 (21.61)
cosangle	0.946 (0.14)	0.987 (0.02)	0.505 (0.28)

모의실험의 결과를 보면, 모형식이 비선형일 때 일변량 분할 역회귀로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 비교적 높음을 알 수 있으며, $\hat{y}$과 r을 이용한 이변량 분할 역회귀로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도 또한 매우 높음을 알

수 있다. 그러나 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 매우 낮음을 알 수 있다. 이것은 분할 평균분산 추정법이 반응변수 y의 각 분할된 구간에서 조건부 분산을 이용하므로 비선형모형 (3.7)에서 분할된 각 구간의 조건부 분산은 큰 차이없이 구해지며, 따라서 조건부 분산을 이용하여 구한 분할 평균분산 추정치의 고유치는 모두 유사한 값을 갖게되고, 결국 큰 고유치에 대응하는 고유벡터를 이용하여 효과적인 차원축약방향을 추정한다는 것은 어렵게 된다. 결과적으로 $\hat{y}$과 r을 이용한 이변량 분할 역회귀방법이 일변량 분할 역회귀와 분할 평균분산 추정법에 비하여 비선형모형에서도 실제모수에 근사정도가 높은 효과적인 차원축약방향을 추정하고 있음을 알 수 있다. 이것은 $\hat{y}$과 r을 이용한 이변량 분할 역회귀 방법이 $\hat{y}$이 설명하지 못한 비선형효과를 잔차 r에서 다시 고려하기 때문이다.

지금까지의 모의실험 결과 본 연구에서 제안한 $\hat{y}$과 r을 이용한 이변량 분할 역회귀방법이 선형, 이차형, 삼차형, 비선형등 여러 모형에서 일변량 분할 역회귀와 분할 평균분산 추정법에 비해 안정적으로 효과적인 차원축약방향을 추정하며, 추정된 효과적인 차원축약방향의 실제모수에 대한 근사정도는 매우 높음을 알 수 있다. 또한 모형이 대칭의 이차형태인 실제자료의 분석결과 일변량 분할 역회귀에 비하여 효과적인 차원축약방향을 잘 추정하며, 모형이 대칭의 이차형태일 때 일변량 분할 역회귀의 대안으로 제시된 분할 평균분산 추정법으로 추정한 효과적인 차원축약방향과 비교했을 때 반응변수 y에 대한 설명력에 있어서 차이없이 높은 설명력을 보여주고 있다. 그리고 $\hat{y}$과 r을 이용한 이변량 분할 역회귀 방법은 모형에 존재하는 오차항에 대하여 일변량 분할 역회귀 보다 비교적 안정적으로 효과적인 차원축약방향을 추정하고 있음을 알 수 있다.

6. 결 론

일반화 회귀모형에서 차원축약의 방법으로 Li(1991)에 의해서 제시되었던 일변량 분할 역회귀 방법은 하나의 일반화 회귀모형에서 효과적인 차원축약방향과 공간을 추정하는 방법이다. 본 논문에서는 두 일반화 회귀모형을 동시에 고려하여 효과적인 차원축약방향과 공간을 추정하는 방법으로 일변량 분할 역회귀를 확장한 이변량 분할 역회귀를 제안하였다. 이러한 이변량 분할 역회귀의 응용으로 반응변수 y를 $\hat{y}$과 r로 분해하여 동시에 고려한 이변량 분할 역회귀는 y에 대한 정보의 손실없이 효과적인 차원축약방향을 추정하며, 모형식이 선형, 이차형, 삼차형, 비선형 등의 여러 모형식에서 실제 효과적인 차원축약방향에 근사정도가 높은 효과적인 차원축약방향을 추정함을 모의실험을 통해서 확인할 수 있었다.

또한 $\hat{y}$과 r을 이용한 이변량 분할 역회귀는 일변량 분할 역회귀에 비하여 모형에 존재하는 오차에 크게 영향을 받지 않고 효과적인 차원축약방향을 추정하고 있음을 모의실험 결과를 통해서 알 수 있었다. 특히 모형식이 대칭의 이차형인 실제자료의 분석을 통해서 이미 알려져 있듯이 일변량 분할 역회귀 방법이 효과적인 차원축약방향을 잘 추정하지 못하고 있다는 것을 확인할 수 있었고, 일변량 분할 역회귀의 대안으로 제시된 분할 평균분산 추정법과 $\hat{y}$과 r을 이용한 이변량 분할 역회귀로 추정된 효과적인 차원축약방향을 비교한 결과 두 방법 모두 효과적인 차원축

약방향을 잘 추정하고 있으며, 추정된 효과적인 차원축약방향이 반응변수 y를 설명하고 있는 설명력도 높음을 알 수 있었다. 또한 이변량 분할 역회귀와 분할 평균분산 추정법으로 추정된 효과적인 차원축약방향과 반응변수 y의 산포도는 모형식이 대칭의 이차형태를 띠고 있음을 잘 보여주고 있다.

참고문헌

[1] Cook, R. D. and Weisberg, S. (1991), Discussion of "Sliced Inverse Regression" by Ker-Chau Li, *Journal of the American Statistical Association*, Vol. 86, pp 328-332.

[2] Cook, R. D. , and Weisberg, S. (1994), An Introduction to Regression Graphics, New York, JOHN WILEY & SONS.

[3] Duan, N. , and Li, K. C. (1991), "Slicing Regression : A Link-Free Regression Method", *The Annals of Statistics*, Vol. 19, No. 2, pp 505-503.

[4] Li, K. C. (1991), " Sliced Inverse Regression for Dimension Reduction", *Journal of the American Statistical Association*, Vol. 86, No. 414, pp 316-342.

[5] Tierney, L. (1990), LISP-STAT, New York , JOHN WILEY & SONS.

A Study on the Multivariate Sliced Inverse Regression

Yong Goo Lee[4], Duck-Ki Lee[5]

Abstract

Sliced inverse regression is a method for reducing the dimension of the explanatory variable X without going through any parametric or nonparametric model fitting process. This method explores the simplicity of the inverse view of regression; that is, instead of regressing the univariate output varable y against the multivariate X, we regress X against y.

In this article, we propose bivariate sliced inverse regression, whose method regress the multivariate X against the bivariate output variables y_1, y_2. Bivariate sliced inverse regression estimates the e.d.r. directions of satisfying two generalized regression model simultaneously.

For the application of bivariate sliced inverse regression, we decompose the output variable y into two variables, one variable $\hat{y}$ gained by projecting the output variable y onto the column space of X and the other variable r through projecting the output variable y onto the the space orthogonal to the column space of X, respectively and then estimate the e.d.r. directions of the generalized regression model by utilize two variables simultaneously. As a result, bivariate sliced inverse regression of considering the variable $\hat{y}$ and r simultaneously estimates the e.d.r. directions efficiently and steadily when the regression model is linear, quadratic and nonlinear, respectively.

4) Professor, Department of Applied Statistics, Chung-Ang University, Dongjak-gu, Seoul 156-756, Korea.

5) Lecture, Department of Applied Statistics, Chung-Ang University, Dongjak-gu, Seoul 156-756, Korea.