응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 141–149*

다목적 다변량 자료분석을 위한 변수선택*

허명회[1] 임용빈[2] 이용구[3]

요 약

다변량 자료분석에서 최근의 추세는 관측개체의 수 n이 커지는 외에 변수의 수 p가 큰 사례들이 많아지고 있다는 것이다. n개 개체 각각에서 획득된 p개 변수들 $X_1, X_2, \ldots, X_p$ 가운데는 이름이나 개념적으로는 구분이 가능하지만 실제로 거의 중복이 되는 변수들이 있을 수 있는데, 이들 변수들이 모두 분석에 포함되면 여러 문제가 유발될 수 있다. 예컨대 주성분 분석이나 인자 분석에서는 중복 변수들이 주축(主軸, principal axis) 결정에, 관측개체 군집화에서는 개체간 거리 산출에 왜곡된 영향을 줄 수 있다. 또한 목적변수가 지정된 지도학습(supervised learning)에서 설명변수들의 중복성은 추정모형의 안정성을 해치는 결과를 초래한다. 실제 자료 분석에서는 한 자료 세트가 여러 기법으로 탐색되고 다수의 모형이 추출되므로 변수세트를 최대한 절약적(parsimonious)으로 구성할 필요가 있다.

본 연구의 목적은 $X_1, X_2, \ldots, X_p$ 중에서 필요한 변수들은 선택하고 불필요한 변수들은 제거함으로써 주어진 변수세트를 보다 적은 크기의 변수세트로 대치하는 방법을 제시하는 데 있다. 제안 방법을 몇 개의 수치적 사례에 적용해 봄으로써 선택 변수와 제거 변수간 관계의 시각화, 회귀모형에서의 유용성, 범주형 자료분석에서의 활용 등에 대해 논의하고자 한다.

주요용어: 주변수(principal variables), 변수선택, 범주형 자료.

1. 연구배경과 목적

주성분분석을 통하여 관측 자료에 내재하는 몇 개의 주요 성분을 추출할 수 있음은 잘 알려져 있다. 그러나 각 주성분이 모든 분석변수를 포함하고 있어 일부 주성분만 선택한다고 하더라도 변수세트 자체가 줄어들지는 않는다. 따라서 주성분을 결정하는 주요 변수들을 선택하는 방법들이 제안된 바 있다 (Jolliffe, 1972, 1973; Krzanowski, 1996). 그러나 주성분분석을 통해 p개 변수들 $X_1, X_2, \ldots, X_p$ 가운데 통계적으로 중복이 없는 변수들만 찾

* 이 논문은 2006년도 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-312-C00486).

1) (136-701) 교신저자. 서울특별시 성북구 안암동 5가, 고려대학교 통계학과, 교수.
E-mail: stat420@korea.ac.kr
2) (120-750) 서울특별시 서대문구 대현동 21, 이화여자대학교 통계학과, 교수.
E-mail: yblim@ewha.ac.kr
3) (156-756) 서울특별시 동작구 흑석동, 중앙대학교 통계학과, 교수.
E-mail: leeyg@cau.ac.kr

아내는 이러한 방법들에는 다소의 문제가 있다. 주성분 자체가 잉여 변수들에 의해 크게 영향을 받을 수 있기 때문이다.

McCabe (1984)는 이런 문제를 의식하고 주성분(principal components)을 대신할 '주변수(principal variables)' 개념을 제시하고 q개($\leq p$)의 주변수 선택과 나머지 변수의 폐기에 대한 이론적 성질들을 파악해냈다. 그러나 실제적으로는 p개 변수 중에서 정해진 개수의 변수를 찾아내는 것은 조합 문제(combinatoric problem)가 되므로 최적화가 쉽지 않고 선택변수 수 q의 최적 값을 정하는 것도 쉽지 않은 문제이다. 이 연구에서 우리는 McCabe (1984)를 연장하되 보다 실용적인 단계적 변수 선택/제거 방법을 제안하고자 한다.

McCabe (1984)는 p변량 확률벡터 X를 주로 분석이 되어야 할 $q(\leq p)$변량 $X_{[1]}$과 나머지 $p-q$변량 $X_{[2]}$로 나누기 위해서는 $X=(X_{[1]}, X_{[2]})$의 공분산 행렬

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

에 대하여

i) $\min |\Sigma_{22.1}|$ (or $\max |\Sigma_{11}|$),

ii) $\min \mathrm{tr}(\Sigma_{22.1})$,

iii) $\min ||\Sigma_{22.1}||^2$,

iv) $\min \Sigma_k \rho_k^2$ (여기서 ρ_k는 선택 변수군과 비선택 변수군 사이의 정준상관)

등의 기준으로 q개 주변수를 찾을 것을 제안한 바 있다. 이 중에서 기준 ii)를 이 연구에서 차용할 것이다.

2절에서 주변수 선택 알고리즘을 제시하고 3절에서는 제안 알고리즘을 몇 개의 수치적 사례에 적용해볼 것이다. 수치적 사례에서 선택 변수와 제거 변수간 관계의 시각화, 회귀모형에서의 유용성, 범주형 자료분석에서의 활용 등을 논의하기로 한다.

2. 제안 방법론

분석 변수들 $X_1, X_2, \ldots, X_p$ 중에서 한 개의 변수 X_j를 선택하면 이 변수로 비선택 변수들의 일정 부분이 설명된다. 변수 X_l의 자료값 벡터를 $\mathbf{x}_l$로 표기할 때 ($l=1,2,\ldots,p$), 선형적인 관계로 보는 경우 변수 X_l의 총 변동 $||\mathbf{x}_l||^2$ 가운데

$$\left|\left|\mathbf{x}_l - \frac{\mathbf{x}_j^t \mathbf{x}_l}{\mathbf{x}_j^t \mathbf{x}_j}\mathbf{x}_j\right|\right|^2, \quad l=1,2,\ldots,p \tag{2.1}$$

가 남는다. 따라서 이것들을 최소화 하는 변수를 찾아야 한다. 한편으로 찾아진 변수 X_j에 의하여 (2.1)의 값이 일정 값보다 작은 변수 X_l은 더 이상 보존할 필요가 없다. 우리는 변수의 선택과 폐기를 단계별로 수행하는 다음 알고리즘을 제안한다.

0) 모든 변수들을 가지고 시작한다. 필요한 경우 변수들을 사전 표준화한다. 변수세트의 크기를 p라고 하자.

1) 크기 p의 변수세트에서

$$\sum_{l=1}^{p} \left|\left| \mathbf{x}_l - \frac{\mathbf{x}_j^t \mathbf{x}_l}{\mathbf{x}_j^t \mathbf{x}_j} \mathbf{x}_j \right|\right|^2$$

을 최소화하는 변수 X_j를 찾는다. 그 변수를 X_{j_1}이라고 하자.

2) 변수 X_l을 X_{j_1}으로 설명되지 않은 잔차인 $X_{l \cdot j_1}$으로 대치한다. 즉 자료 벡터 $\mathbf{x}_l$을

$$\mathbf{x}_{l \cdot j_1} = \mathbf{x}_l - \frac{\mathbf{x}_{j_1}^t \mathbf{x}_l}{\mathbf{x}_{j_1}^t \mathbf{x}_{j_1}} \mathbf{x}_{j_1}, \quad l = 1, 2, \ldots, p$$

로 대치한다.

3) 제곱 크기 $||\mathbf{x}_{lj_1}||^2$이 일정 값 c에 미달되면 X_l을 제거하여 변수세트를 업데이트한다. 변수

세트의 크기를 새로 구하여 p로 놓는다.

4) $p \geq 1$이면 단계 1로 돌아간다.

알고리즘 내 순환 부분이 q번 반복되면 ($q \leq p$) 그 결과로 q개의 변수들이 선택된다. 그 변수들을 $X_{j_1}, X_{j_2}, \ldots, X_{j_q}$라고 하자. 여기에 포함되지 않는 $p-q$개의 변수들은 선택된 q개의 변수들로 상당 부분이 설명되므로 잉여변수(redundant variables)로 볼 수 있다.

위 알고리즘은 cut-off 값 c에 따라 다른 결과를 내므로 c 값을 정해야 하는데 "상식적"인 수준에서 잡는 것이 좋겠다. c는 총 제곱변동 중 선택변수들로 설명되지 않은 제곱변동의 비율이므로, 모든 변수들이 사전 척도화되어 $||\mathbf{x}_l||^2$=1에서 시작한다면, $c = 0.2 \sim 0.1$로 둘 수 있을 것으로 본다. c를 크게 할수록 선택변수 세트의 크기가 작아진다.

다른 방식은 $c = 0$으로 하여 변수 1개씩 선택하는 단계를 끝까지 끌고 가되 각 단계에서

$$\max_{l=1,\ldots,p} \left|\left| \mathbf{x}_l - \frac{\mathbf{x}_{j_1}^t \mathbf{x}_l}{\mathbf{x}_{j_1}^t \mathbf{x}_{j_1}} \mathbf{x}_{j_1} \right|\right|^2 \tag{2.2}$$

을 출력·관찰하여 선택변수 세트의 크기를 사후적으로 정하는 것이다.

3. 수치적 사례

이 절에서는 3개의 수치적 사례를 통해 선택 변수와 제거 변수간 관계, 회귀모형에서의 유용성, 범주형 자료분석에서의 활용 등을 살펴보기로 한다.

신체부위 자료 (R의 gclus 라이브러리의 body 자료)

이 자료는 507명에 대한 21개 신체부위 측정값으로 구성되어 있는데, 사전 모든 변수를 척도화하고 $c = 0.2$로 놓고 제안 알고리즘을 적용한 결과 다음 15개 변수가 선택되었다. 선택된 변수들은 순서대로

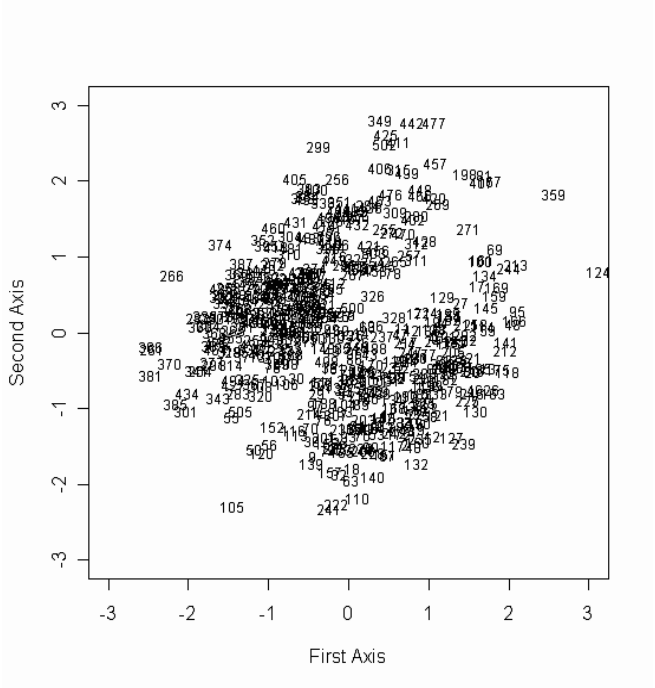

그림 3.1: 2–성분 인자분석 개체 플롯

“ForearmG” “HipG” “Biiliac” “AnkleD” “CalfG” “ChestDp”
“Biacrom” “KneeD” “AnkleG” “Bitro” “ChestD” “WristD”
“AbdG” “KneeG” “ThighG”

이고 비선택 변수 리스트는 다음과 같다.

“BicepG” “WaistG” “ShoulderG” “ChestG” “ElbowD” “WristG”

이들 6개 비선택 변수들에 대한 편공분산 행렬은 다음과 같이 나타났다.

	BicepG	WaistG	ShoulderG	ChestG	ElbowD	WristG
BicepG	0.09	0.02	0.03	0.03	0.00	0.01
WaistG	0.02	0.14	0.02	0.03	-0.02	0.00
ShoulderG	0.03	0.02	0.11	0.04	0.01	0.00
ChestG	0.03	0.03	0.04	0.09	0.01	0.00
ElbowD	0.00	-0.02	0.01	0.01	0.15	0.00
WristG	0.01	0.00	0.00	0.00	0.00	0.11

대각선 상에는 0.2 이하의 값들만 남았고 그 외 비대각 요소들도 대부분 매우 작아졌음을 확인할 수 있다.

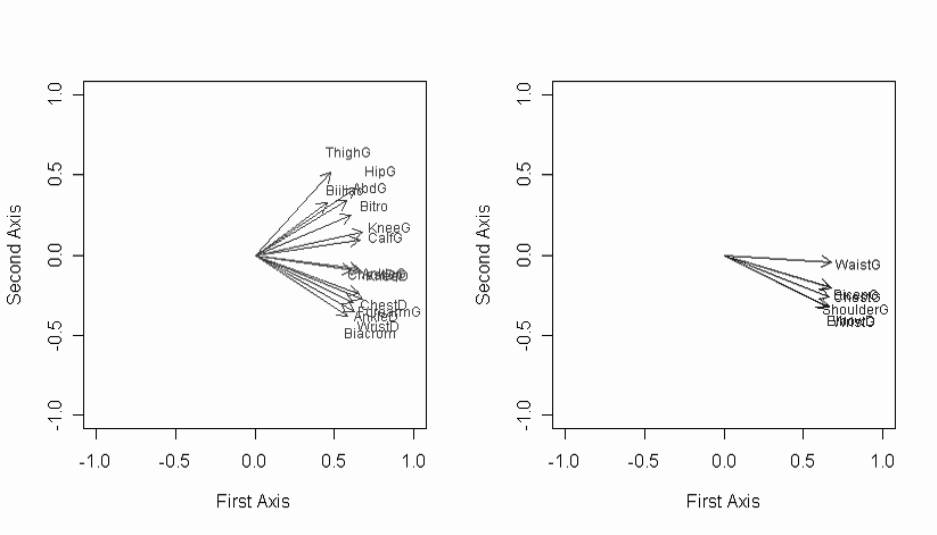

그림 3.2: 2–성분 인자분석 변수 플롯: (좌) 선택변수, (우) 제거변수

선택된 15개 변수세트에 대한 주성분의 고유값은 처음 2개만 1.0 이상이었고 (8.94와 1.98) 누적 퍼센트는 72.8%였다. 이에 따라 2-성분 인자분석을 하였고 개체 플롯으로 그림 3.1을, 변수 플롯으로 그림 3.2의 왼쪽 것을 얻었다. 그림 3.2의 오른쪽 것은 선택되지 않은 6개 변수를 추가변수(supplementary variables) 기법으로 시각화한 것이다 (허명회, 1999). 특이하게도, 비선택 6개 변수가 제2 인자의 특정 부호로 쏠려 있다. 이들 변수들은 2개의 큰 덩어리로 된 그림 3.1에서 덩어리 구분과 관련 있지만 그림 3.2의 왼쪽 그림에서 인자분석에 투입된 15개 변수 중에 그 위치에 상당수의 변수가 있음을 볼 수 있다.

NIR 자료 사례 (R의 pls 라이브러리 버전 1의 NIR 자료)

원자료는 28개 PET사(絲) 대상 268개 파장에 대한 근적외선 스펙트럼 값으로 구성되어 있다. 전처리로서 268개 변수 각각에 대하여 중심화를 하였고 척도화는 하지 않았다. 그리고 최대 표준편차를 1이라고 할 때 0.1 이상의 표준편차를 갖는 변수의 수는 107개이었지만 $c = 0.1$로 하여 2절의 알고리즘을 적용한 결과 x.11, x.112 등 2개 변수만 선택되었고 PET사의 밀도 y를 x.11과 x.112에 회귀시킨 결과 96.4%의 결정계수를 가진 다중선형회귀 모형이 도출되었다. 그림 3.3에서 y와 x.11, x.112 간 산점도 행렬을 보라.

이 사례에서는 p가 n보다 크므로 모든 설명변수에 의한 다중선형회귀는 적합이 가능하지 않다. 대안적으로 PCR 회귀(principal component regression)와 PLS 회귀(partial least squares regression)를 할 수 있는데 이 방법들은 2개 성분을 고려한다고 하더라도 y 예측에 모든 변수들을 필요로 한다. 참고로 2개 성분의 PCR 모형과 PLSR 모형의 결정계수는 각각 4.6%와 96.4%였다. 따라서 y와 관련짓지 않고 선택된 2개만의 변수를 활용하면서도 우리가 제안하는 회귀모형의 설명력이 모든 변수에 의존하는 다른 두 회귀모형에 비하여 훨

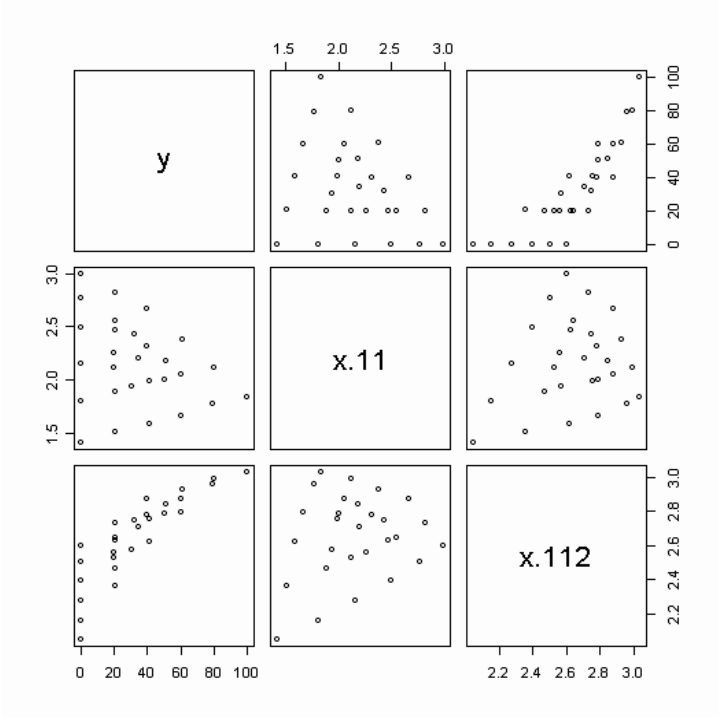

그림 3.3: NIR 자료에서 y와 2개의 선택변수 간 산점도 행렬

씬 낫거나 최소한 못지않다. 최근 허명회 등 (2007)은 이 자료에 대한 시각화 결과를 제시한 바 있다.

German Credit 자료 사례

이 사례의 자료는 신용대출자 1,000명에 대한 20개 금융 및 인구사회적 변수 X1, ..., X20과 결과변수인 신용구분 코드로 구성되어 있다 (허명회 (2003, p. 30–32) 또는 *http://mlearn.ics.uci.edu/MLSummary.html* 참조). X1부터 X20의 20개 변수 중 7개는 연속형이지만 13개는 범주형이다 (범주 수가 2개부터 11개까지). 이 사례분석에서는 범주형 변수 수를 줄이고 더 나아가 선택된 변수에 대하여는 가능한 한 범주들을 통합하는 것을 목표로 한다.

전처리로서 범주형 변수가 명목형인 경우는 더미 변수로 바꾸었다. 예컨대 X20은 foreign worker로 'yes' 또는 'no'의 두 값을 갖는 이항형 변수인데 이 변수를

$$\text{A201} = 1 \quad \text{if} \quad \text{'yes'}, \quad = 0 \text{ otherwise},$$
$$\text{A202} = 1 \quad \text{if} \quad \text{'no'}, \quad = 0 \text{ otherwise}$$

로 표현하였다. 따라서 두 변수 중 하나는 redundant하므로 하나가 선택되면 자동적으로 다른 하나는 폐기된다. 이렇게 처리된 변수는 X20 (foreign worker, 2개 범주) 외에 X4 (purpose,

11개 범주), X9 (personal status and sex), X14 (other installment plans, 3개 범주), X15 (housing, 3개 범주), X19 (telephone, 2개 범주) 등이다. 이 때 k개의 범주를 갖는 변수를 같은 수의 더미 변수로 표현하였는데 물론 그 중 하나는 redundant하다. 그러나 더미 변수들의 어느 조합이라도 선택될 수 있도록 하기 위하여 이런 표현 방법이 바람직하다. 물론 k개의 더미 변수가 모두 선택되지는 않는다.

두 번째 유형은 순서 범주형 변수의 경우인데 누적 범주 방식으로 이항형 변수를 생성시켰다. 예를 들어 X3는 credit history로 all credits paid back duly(= 0), all credits at this bank paid back duly(= 1), existing credits paid back duly(= 2), delay in paying off in the past(= 3), critical account(= 4)의 5개 범주를 취하는데 다음과 같이 4개 이항형 코드로 변환하였다.

$$\begin{aligned}
&\text{A31} = 1 \quad \text{if} \quad \text{X3} \geq 1, \quad = 0 \text{ otherwise,}\\
&\text{A32} = 1 \quad \text{if} \quad \text{X3} \geq 2, \quad = 0 \text{ otherwise,}\\
&\text{A33} = 1 \quad \text{if} \quad \text{X3} \geq 3, \quad = 0 \text{ otherwise,}\\
&\text{A34} = 1 \quad \text{if} \quad \text{X3} \geq 4, \quad = 0 \text{ otherwise.}
\end{aligned}$$

이렇게 처리된 변수는 X3 (credit history, 5개 범주) 외에 X7 (present employment since, 5개 범주), X10 (other debtors, 3개 범주), X12 (property, 4개 범주), X17 (job, 4개 범주) 등이다.

세 번째 유형은 순서형과 명목형이 혼합된 범주형 변수인데 이 경우 순서형 부분은 누적 범주 방식으로, 명목형 범주는 구분되도록 이항형 변수를 생성시켰다. 예를 들어 X1은 checking account인데 0 or less(= 1), between 0 and 200(= 2), 200 or more(= 3), no account(= 4)의 4개 범주를 다음과 같이 코딩하였다.

$$\begin{aligned}
&\text{A11} = 1 \quad \text{if} \quad 1 \leq \text{X1} \leq 3, \quad = 0 \text{ otherwise,}\\
&\text{A12} = 1 \quad \text{if} \quad 2 \leq \text{X1} \leq 3, \quad = 0 \text{ otherwise,}\\
&\text{A13} = 1 \quad \text{if} \quad 3 \leq \text{X1} \leq 3, \quad = 0 \text{ otherwise,}\\
&\text{A14} = 1 \quad \text{if} \quad \text{X1} = 4, \quad = 0 \text{ otherwise.}
\end{aligned}$$

따라서 이들 중 하나는 redundant하다. 이렇게 처리된 변수는 X1 (checking account, 4개 범주) 외에 X6 (savings account, 5개 범주)가 있다.

이와 같이 사전처리로 모두 1/0 코드화한 다음 중심화하였으나 척도화는 하지 않았다. 따라서 m개의 1과 $n-m$개의 0으로 코딩된 열의 중심화된 벡터를 $\mathbf{x}_j$로 표기할 때 $||\mathbf{x}_j||^2/n$은 $r \cdot (1-r)$가 된다 (여기서 $r = m/n$). 알고리즘 적용시 cut-off 값을 0.09 ($= 0.1 \cdot 0.9$)로 한 결과, A11, A12, A33, A40, A42, A41, A61, A62, A73, A74, A75, A92, A93, A122, A123, A143, A152, A173, A174, A191 등이 선택되었다 (여기서 마지막 숫자는 범주를, 그 앞 숫자는 변수를 나타낸다). 즉, X10 (other debtors, 3개 범주)과 X20 (foreign worker, 2개 범주)에 해당하는 1/0 코드 변수가 모두 선택되지 않았다. 따라서 두 변수 X10과 X20은 실제로 redundant하다고 할 수 있다. 또한 다수의 범주구분이 불필요함을 알 수 있는데 예컨대 X3에 대한 코드 중에서 A33만 선택되었으므로 X3 (credit history $= 0, 1, 2, 3, 4$)의 5개 범주

중 0,1,2를 한 범주로 묶고 3,4를 다른 한 범주로 병합할 수 있다.

4. 맺음말

이상으로 주어진 p개 변수 $X_1, X_2, \ldots, X_p$ 중에서 필요한 변수들은 선택하는 단계적 방법과 사례를 제시하였다. cut-off 값 c를 0으로 놓는 경우 2절의 알고리즘을 Gram-Schmidt 직교화 과정으로도 볼 수 있겠는데 통상적인 Gram-Schmidt 과정과는 열(변수) 선택 기능에서 차이가 있다. 이 연구에서 제안한 "변수선택 Gram-Schmidt 직교화 알고리즘"은 다변량 자료의 다목적 분석에 유용한 축소된 변수세트를 제공한다.

참고문헌

허명회 (1999). <다변량 수량화>, 자유아카데미, 서울.

허명회 (2003). <데이터마이닝 모델링과 사례>, SPSS 아카데미, 서울.

허명회, 이용구, 이성근 (2007). PLS 기법에 의한 (X, Y) 자료의 시각화, <응용통계연구>, **20**, 345–355.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data, *Applied Statistics, Journal of the Royal Statistical Society*, Ser. C., **21**, 160–173.

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis. II: Real data, *Applied Statistics, Journal of the Royal Statistical Society*, Ser. C., **22**, 21–31.

Krzanowski, W. J. (1996). A stopping rule for structure-preserving variable selection, *Statistics and Computing*, **6**, 51–56.

McCabe, G. P. (1984). Principal variables, *Technometrics*, **26**, 137–144.

[2007년 8월 접수, 2007년 9월 채택]

Variable Selection for Multi-Purpose Multivariate Data Analysis*

Myung-Hoe Huh[1)] Yong Bin Lim[2)] Yonggoo Lee[3)]

Abstract

Recently we frequently analyze multivariate data with quite large number of variables. In such data sets, virtually duplicated variables may exist simultaneously even though they are conceptually distinguishable. Duplicate variables may cause problems such as the distortion of principal axes in principal component analysis and factor analysis and the distortion of the distances between observations, *i.e.* the input for cluster analysis. Also in supervised learning or regression analysis, duplicated explanatory variables often cause the instability of fitted models. Since real data analyses are aimed often at multiple purposes, it is necessary to reduce the number of variables to a parsimonious level.

The aim of this paper is to propose a practical algorithm for selection of a subset of variables from a given set of p input variables, by the criterion of minimum trace of partial variances of unselected variables unexplained by selected variables. The usefulness of proposed method is demonstrated in visualizing the relationship between selected and unselected variables, in building a predictive model with very large number of independent variables, and in reducing the number of variables and purging/merging categories in categorical data.

Keywords: Principal variables, variable selection, categorical data.

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2006-312-C00486).

1) Corresponding author. Professor, Dept. of Statistics, Korea University, Anam-dong 5, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

2) Professor, Dept. of Statistics, Ewha Women's University, Seoul 120-750, Korea.
E-mail: yblim@ewha.ac.kr

3) Professor, Dept. of Statistics, Chungang University, Seoul 156-756, Korea.
E-mail: leeyg@cau.ac.kr