응용통계연구 제 12권 2호, 1999년, pp. 517-525

정준판별분석의 조건부 계량화*

황선영[1] 이연숙[2]

요 약

본 논문에서는 정준판별분석에서 분석자의 사전지식을 고려한 조건부 계량화(partial quantification)를 제안하고 있다. 연관된 수식을 라그랑지 승수법으로 유도하였으며 SAS/IML을 이용한 프로그램과 예제를 간략히 설명하였다.

1. 서론

다변량 자료분석의 주요 관심사 중의 하나는 최소한의 정보 손실하에서 변수들간의 상호의존관계를 이용하여 주어진 자료를 축약, 재구성하는 것이다(cf. Dillon and Goldstein (1984)). 주성분분석, 인자분석, 판별 및 분류분석, 상관분석, 다차원척도법, 그리고 Biplot 기법이 그 대표적인 예라고 할 수 있다. 다변량 자료의 분석은 각 분야의 연구자들에 의해 오래전부터 다양한 측면에서 접근되어 왔으며, 그 과정에서 관련분야 연구자들은 자료에 관한 지식, 사전 정보, 경험을 쌓아왔다. 지금까지의 분석기법은 일단 자료와 분석기법이 정해지면 사전에 분석자의 주관이나 지식이 거의 개입되는 과정없이 기계적으로 계산하는데, 만일 분석자의 사전지식이 고려된다면 자료의 분석과 해석에 도움을 얻을 수도 있을 것이다. 자료분석에 연구자의 주관이 개입되는 경우 주어진 자료만을 가지고 분석하는 것에 비해서 객관적인 효율은 떨어질 수도 있으나, 자료분석과 사전지식의 조화를 얻을 수 있다는 점에서 심리학, 교육학, 경영학 등에서 주로 행하는 실증분석에 유용하리라 판단된다. 주성분 분석에 있어 조건부 계량화 (partial quantification)는 서혜선(1997), 서혜선과 허명회(1997)에 의해 연구되었으나, 그 이외의 분석기법에 대해서는 관련된 연구가 없는 현실이다. 본 논문에서는 정준판별분석에 있어 자료 분석전 분석자의 주관을 고려하는 조건부 계량화에 관한 연구를 목표로 한다. 2장에서는 정준판별분석에 있어 제 1 정준계수가 사전적으로(*a priori*) 주어졌을 때 나머지 정준계수를 찾는 조건부 계량화 방법을 제안하고, 3장은 연관된 알고리즘을 다루었으며 4장에서는 조건부 계량화 프로그램을 이용한 자료분석 예제를 다루고 있다.

2. 정준판별분석의 조건부 계량화

정준판별분석은 판별변수의 적절한 선형결합을 통하여 그룹을 분리하고자 하는 다변량 분석기법이다(김기영,전명식(1997), 박찬욱(1994) 참고). 자료행렬 X가 $n \times p$행렬로서 n개

* 본 연구는 숙명여대 1999년도 교내연구비 지원에 의해 수행되었음.

1) (140-742) 서울시 용산구 청파동, 숙명여대 통계학과, 부교수. shwang@sookmyung.ac.kr
2) (140-742) 서울시 용산구 청파동, 숙명여대 통계학과, 대학원.

의 개체들과 p개의 판별변수로 구성되어 있다고 하자. n개의 개체들이 g개의 그룹으로 나누어져 있고 각각의 그룹이 n_j개의 개체로 이루어져있을 때$(j = 1, 2, \cdots, g)$ 각 그룹의 평균벡터 $\bar{X}_j$와 공분산행렬 S_j는 다음과 같이 표현할 수 있다.

$$\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

$$S_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$$

전체 평균수정 제곱합과 교적합의 행렬(total mean corrected sums of squares & cross products matrix)을 T , 집단내 제곱합과 교적합의 행렬 (within-group sums of squares & cross products matrix)을 W , 집단간 제곱합과 교적합의 행렬(between-group sums of squares & cross products matrix)을 B라 하면 B와 W는 다음과 같이 정의되며 T는 $T = B + W$ 로 부터 계산할 수 있다.

$$B = \sum_{j=1}^{g} n_j (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X})'$$

$$W = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$$

여기서 $\bar{X} = \frac{1}{n} \sum_{j=1}^{g} n_j \bar{X}_j$ 는 전체 평균벡터이다. 또한 합동집단내 공분산행렬(pooled within-group covariance matrix) S_p는 $S_p = W/(n - g)$ 를 통해 유도된다. 판별함수들의 선형결합 $y = l'X$를 고려해 보자. 분산분석에서 g개의 집단의 평균이 동일한지 검정하기 위해서

$$F = \frac{l'Bl}{(g-1)} \Big/ \frac{l'Wl}{(n-g)}$$

통계량을 이용한다. 우리의 관심은 각 집단을 잘 구분할 수 있는 계수벡터를 구하는 것이므로 F값을 최대로 하는 l을 찾는 것이다. 상수는 F값의 최대화에 영향을 미치지 않으므로 다음의 $\lambda(l)$ 을 최대로 하는 문제와 일치한다.

$$\lambda(l) = \frac{l'Bl}{l'Wl} \tag{2.1}$$

$\lambda(l)$의 값을 최대로 하는 계수벡터 l_1은 $W^{-1}B$의 최대고유값 λ_1에 해당하는 고유벡터이며, 두 번째 계수벡터 l_2는 $Cov(l_1'X, l_2'X) = 0$의 조건을 만족시키면서 $\lambda(l)$을 최대로 하는 벡터로 $W^{-1}B$의 두 번째로 큰 고유값 λ_2에 대응되는 고유벡터이다. 이러한 과정을 계속하여 계

수벡터 l_j는 $Cov(l_i'X, l_j'X) = 0$, $(i < j)$를 만족시키면서 $\lambda(l)$을 최대로 하는 벡터로 $W^{-1}B$의 j번째로 큰 고유값 λ_j에 대응되는 고유벡터이다. 이러한 과정은 $k = \min(p, g-1)$번 반복된다. 이제 제 1정준계수 a가 사전에 미리 주어졌다고 가정하면 나머지 정준계수를 구하는 과정은 다음과 같다(단, a는 $a'S_pa = 1$ 로 scale된 벡터). $a'S_pl_2 = 0$이라는 조건하에서

$$\lambda(l_2) = \frac{{l_2}'Bl_2}{{l_2}'Wl_2} \tag{2.2}$$

를 최대화하는 계수벡터 l_2를 얻은 후 다시 제약조건 $a'S_pl_3 = 0$, ${l_2}'S_pl_3 = 0$ 에서

$$\lambda(l_3) = \frac{{l_3}'Bl_3}{{l_3}'Wl_3}$$

를 최대화하는 계수벡터 l_3을 얻는다. 이러한 과정을 계속하여 정준계수 $l_2, \cdots, l_k$을 얻을 수 있다. 식 (2.2)를 임의의 벡터 l에 대해 풀기 위해 라그랑지 승수법을 이용하면 최적화 목적식은

$$\Phi(l, c) = \frac{l'Bl}{l'Wl} - 2ca'S_pl \tag{2.3}$$

이다. 이 식을 l에 대해 미분하여 다음과 같은 식을 얻는다.

$$\frac{\partial\Phi}{\partial l} = \frac{2Bl(l'Wl) - l'Bl(2Wl)}{(l'Wl)^2} - 2c{S_p}'a = 0 \tag{2.4}$$

즉,

$$Bl - Wl\frac{l'Bl}{l'Wl} - c{S_p}'a(l'Wl) = 0 \tag{2.5}$$

식 (2.3)으로부터

$$\frac{l'Bl}{l'Wl} = \Phi + 2ca'S_pl$$

를 유도할 수 있고, 이를 식 (2.5)에 대입하여 정리하면

$$Bl - \Phi Wl - 3c{S_p}'al'Wl = 0 \tag{2.6}$$

이다. 식 (2.4)를 정리하여 얻은 c 를

$$c = \frac{a'S_pW^{-1}Bl}{a'S_pW^{-1}{S_p}'a(l'Wl)}$$

식 (2.6)에 대입하면 다음과 같은 고유 방정식을 유도해 낼 수 있다.

$$QBl = \Phi l \tag{2.7}$$

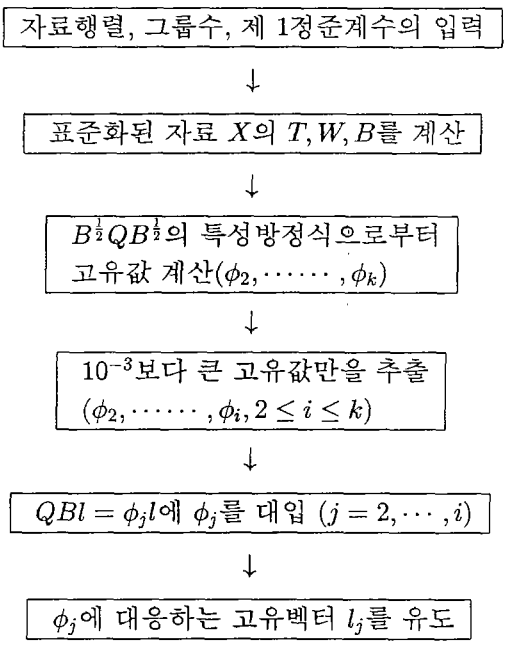
여기서 Q는 $p \times p$행렬로 다음과 같다.

$$Q = W^{-\frac{1}{2}}[I - 3W^{-\frac{1}{2}}S_p{}'a\frac{a'S_pW^{-\frac{1}{2}}}{a'S_p{}'W^{-1}S_p{}'a}]W^{-\frac{1}{2}} \tag{2.8}$$

구하고자 하는 정준계수는 QB의 특성방정식을 이용하여 구할 수 있다. 계수벡터 l_2는 QB의 최대고유값 ϕ_1에 대응하는 고유벡터이며, l_3는 QB의 두 번째로 큰 고유값 ϕ_2에 대응하는 고유벡터이다. 이러한 과정을 계속하여 나머지$k-1$개의 고유벡터를 순차적으로 구할 수 있다. 조건부 정준계수를 구하는 과정은 다음과 같이 기하학적으로 이해할 수 도 있다. $W^{-1}B$의 양의 고유벡터를 $e_1, \cdots, e_k[k = min(p, g-1)]$라 할 때, 사전에 주어진 제 1정준계수 a와 가장 가까운 벡터e_j를 찾아서 e_j를 a로 회전시킨다. 회전에 사용된 행렬을 P로 표현하면 $a = Pe_j$가 성립한다. 동일한 회전을 $e_1, \cdots, e_{j-1}, e_j, e_{j+1}, \cdots, e_k$에 적용시킨 후, 즉 $Pe_1, \cdots, Pe_{j-1}, a, Pe_{j+1}, \cdots, Pe_k$ 으로부터 원하는 계수 $l_2 = Pe_1, \cdots, l_j = Pe_{j-1}, l_{j+1} = Pe_{j+1}, l_k = Pe_k$ 를 구할 수 있다.

3. 프로그램 흐름도

지금까지 설명한 조건부 계량화된 정준계수를 구하는 프로그램은 다음과 같은 흐름도를 가지고 있다.



SAS/IML(1990)에서 특성방정식의 해를 구하기 위해 지원되는 부프로그램(subroutine)에는 EIGEN과 GENEIG이 있다. EIGEN 부프로그램은 대칭행렬의 고유값과 고유벡터를 구하며 GENEIG 부프로그램은 특성방정식 $Ax = \lambda x$을 푸는데 있어 비대칭행렬 A가 $A = W^{-1}H$과 같이 두 대칭행렬 W, H의 곱의 형태로 표현될 수 있는 경우 편하게 이용될 수 있다. CALL GENEIG(eigenvalues, eigenvectors, symmetric matrix 1, positive definite symmetric matrix 2)의 형식으로 호출한다. 이런 부프로그램외에도 촐레스키분해를 수행하는 ROOT함수, 에르미트 정규형식(Hermite normal form)을 지원하는 HERMITE 함수등이 있으나 주어진 특성방정식 $QBl = \Phi l$을 푸는 데는 직접 이용할 수 없다. 그 이유는 행렬 Q가 양정치 행렬이 아니기 때문이다. 따라서 다음과 같은 해결방안을 모색하도록 한다. $QBl = \Phi l$의 고유치는, 다음 식에서 $l^* = B^{\frac{1}{2}}l$로 놓으면

$$B^{\frac{1}{2}}QB^{\frac{1}{2}}B^{\frac{1}{2}}l = \Phi B^{\frac{1}{2}}l$$

$B^{\frac{1}{2}}QB^{\frac{1}{2}}$의 고유치와 일치하며 $B^{\frac{1}{2}}QB^{\frac{1}{2}}$이 대칭행렬이므로 부프로그램 EIGEN 을 이용하여 구할 수 있다. 이렇게 얻은 고유값을 특성방정식에 다시 대입하여 그에 대응하는 고유벡터를 유도할 수 있다. 대입하는 과정에서 10^{-3}보다 작은 고유값은 의미가 없으므로 고려 대상에서 제외하였다. 고유벡터를 구하는 과정에는 HOMOGEN함수를 이용하였다. HOMOGEN함수는 선형방정식 $AX = 0$에서 적어도 하나의 해 X가 존재하고, A의 차수는 $m \times n(m \geq n)$이며 계수 $r(r < n)$일 때 $AX = 0, X'X = I$를 만족하는 X를 유도한다.

4. 예제

기업에서 발행된 채권(Bond)의 등급평가는 채권의 특수성을 고려하여 채권의 약관 내용과 운영 및 재무상태 등이 평가요소로 이용되고 있다. 채권등급은 기관 포트폴리오 관리자가 채권의 위험수준을 평가하는 지표로 자주 사용되어진다. 미국의 신용조사기관의 하나인 Moody's 등급은 신용도가 높은 순서부터 Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C의 등급으로 분류한다. Aaa, Aa, A채권등급을 우량투자성집단(1), Baa채권등급을 보통투자성집단(2), 그리고 나머지 채권등급을 투기성집단(3)이라고 한다. 1987년 Moody's International Manual에서 130개 채권을 추출하여 5개 독립변수에 대한 정보를 수집하였다.

X_1 =총자산규모(1억달러), X_2 =레버리지 척도(장기부채/총자본)
X_3 =수익성정도(순이익/총자산), X_4 =불안정척도(순이익변동계수)
X_5 =주식등급(1-6등급)

독립변수 X_5는 1986년도 자료이며, 나머지 독립변수들은 1982년부터 1986년까지의 5년간 산술평균을 사용한 자료이다(출처: 성웅현(1997)). 독립변수중 X_1, X_4가 정규성을 크게 위반하므로 자연대수변환한 자료 LX_1, LX_4를 이용하도록 하겠다. 우선 CANDISC 프로시저를 이용하여 일반적인 정준판별분석을 행한 결과는 다음과 같다. 표 4.1에서 보는 바와 같이 첫 판별함수가 총판별력의 98.48%를 점유하고 있고, 이에 비해 두 번째 판별함수는 1.52%로 상당히 미약함을 알 수 있다. 또한 집단과 정준판별함수와의 관계를 요약하는

척도인 정준상관계수를 보면 첫 판별함수는 0.8367로 매우 높은 상관을 가지고 있는 것에 비해, 두 번째 함수는 0.1867로 상당히 낮은 값을 가짐을 알 수 있다.

표 4.1: Bond자료의 고유값과 정준상관계수

정준판별함수	고유값	상대백분률	정준상관계수
1	2.3339	98.48	0.8367
2	0.0361	1.52	0.1867

표 4.2: Bond자료의 표준정준계수(standardized canonical coefficient)

판별변수	제 1정준계수(l_1)	제 2정준계수(l_2)
LX_1	-0.5328	0.8690
X_2	0.4411	0.8815
X_3	-0.1011	-0.0497
LX_4	0.4687	-0.2763
X_5	0.8212	0.1170

첫번째 판별함수에서 LX_1, X_3는 음의 값을 가지고 있으며, LX_1, X_5 은 큰 절대값을 가지고 있으므로 상대적으로 큰 설명력을 가지고 있는 것으로 생각된다. 즉 총자산규모와 수익성정도가 클수록 더 우량투자성집단의 채권으로, 레버리지척도, 불안정척도, 주식등급이 높을수록 투기성집단의 채권으로 판별된다. 두 번째 판별함수에서 X_3는 거의 설명력이 없고, LX_1, X_2가 많은 설명력을 가지는 것을 알 수 있다. 채권에 대한 신용평가기준은 평점표의 사용기관, 평가대상업체의 성격, 평점표의 사용용도에 따라 평가표를 구성하고 있는 재무비율 및 비재무항목의 종류가 다를 뿐 그 구성에서 대체로 비슷한 형태를 띠고 있다. 한국능률협회(KMA)의 우량기업분석모형을 위한 평가표는 수익성, 안정성, 규모 및 활동성, 성장성등 4개부문으로 나눈 19개 재무비율을 사용하고 있으며 수익성에 30%, 안정성 25%, 규모 및 활동성 20%, 성장성 25% 의 가중치를 부여하고 있다(출처: 한국신용평가(1995)). 이 평가표를 이용하여 X_1은 규모 및 활동성과 연관시키고 X_2 는 안정성, X_3: 수익성, X_4: 안정성 그리고 X_5: 성장성에 대응시킬 수 있으며 이에 따라 채권등급 평가자료의 각 구분의 가중치를 w =(-0.20,0.25,-0.30,0.25,0.25)로 주기로 하자. 따라서 제 1 정준계수 는 $a = w/||S_p{}^{1/2}w||$ 이며 3장에서 설명된 프로그램을 이용하여 조건부 정준판별분석을 수행하면 다음과 같은 결과를 얻을 수 있다. 두 번째 판별함수의 고유값은 0.0404이며, 정준계수는 l_2 =(0.7858, 0.8182, -0.2834, -0.1980, -0.0019) 이다. 첫 번째 정준계수 a를 주었을 때의 고유값은 1.6590이다. 원자료만을 가지고 구한 경우(표 4.1)와 첫 번째 정준변수를 사전에 정한 후 구한 경우의 고유값의 합의 비율은 71.7%로 원자료만을 이용한 경우에 비해 효율은 떨어진다. 두 번째 판별함수를 보자. 수익성(X_3)의 상대적인 공헌도는 원자료에서(표 4.2) 2.2 % : $0.0497/(0.8690 + 0.8815 + 0.0497 + 0.2763 + 0.1170)$ 인 데 반해 조건부

계량화된 경우에는 13.6 % : 0.2834/(0.7858 + 0.8182 + 0.2834 + 0.1980 + 0.0019) 로 대폭 증가하게 된다.

표 4.3: 한국능률협회(KMA)의 우량기업분석모형을 위한 평가표

구분	평가항목	가중치(%)
수익성	총자본영업이익률(당해년도)	12
	총자본영업이익률(전년도)	4
	자기자본경상이익률	4
	납입자본이익률	4
	매출액순이익률	6
안정성	자기자본비율	12
	고정장기적합	5
	유동비율	5
	자본장기율	3
규모 및 활동성	총자산	3
	매출액	6
	임원, 종업원 총급여액	3
	총자본회전율	4
	매출채권회전율	4
성장성	매출신장율(당해년도)	9
	매출신장율(전년도)	3
	총자산신장율	5
	자기자본신장율	5
	고정자산신장율	3
	합계	100

즉, 조건부 계량화된 판별분석결과 수익성(X_3)의 상대적인 중요성이 6배정도 증가하였으며 반대로 주식등급(X_5)의 공헌도는 거의 없는 것으로 해석할 수 있다. 이는 회사의 신용도를 외형적인 면 보다는 수익성 기준의 내실있는 평가를 하는 요즈음의 조류와 일치한다고 할 수 있겠다. 결론적으로 조건부 계량화된 판별분석은 주어진 판별함수를 강조하여 자료를 해석하고 동시에 의미있는 나머지 판별함수를 유도해 낼 수 있는 상대적인 장점을 가질 수도 있음을 알 수 있다.

감사의 글

본 논문을 심사해주신 두분의 심사위원님께 감사를 드립니다.

참고문헌

[1] 김기영, 전명식(1997). 〈SAS 판별 및 분류분석〉, 자유아카데미.

[2] 박찬욱(1994). 판별분석. *Marketing Forum*, Vol. 5, No.2, 36-52.

[3] 서혜선(1997). 사회연구를 위한 세가지 측면에서의 통계적 방법의 개발과 응용. 고려대학교 대학원 통계학과 박사학위논문.

[4] 성내경(1994). 〈SAS/IML 행렬연산〉, 자유아카데미.

[5] 성웅현(1997). 〈응용다변량분석:이론과 SAS 활용〉, 자유아카데미.

[6] 한국신용평가(1995). 〈부실예측에 관한 계량적 기업신용분석〉.

[7] 허명회(1994). 〈SAS 최적척도법〉, 자유아카데미.

[8] 허명회(1998). 〈1998년 춘계 한국통계학회 다변량 특강〉.

[9] H.S. Suh and M.H. Huh(1997). Partial quantification in principal component analysis. 〈한국통계학회 논문집〉, Vol. 4, 637-644.

[10] SAS Institute Inc.(1990). *SAS/IML Software : Usage and Reference.* Version 6. SAS Institute Inc.

[11] W.R. Dillon and M. Goldstein(1984). *Multivariate Analysis : methods and Applications.* Wiley, New York.

[1998년 11월 접수, 1999년 7월 최종수정]

Partial Quantification in Canonical Discriminant Analysis*

Sun Y. Hwang[1] Youn-suk Lee[2]

ABSTRACT

This article discusses the partial quantification method in canonical discriminant analysis. When a canonical discrimination function(CDF) comes logically from the researcher's opinion, it may be incorporated into the analysis by deriving the other orthogonal CDFs. An algorithm to figure out the orthogonal CDFs conditional on the given, *a priori*, CDF is developed and an example is presented illustrating and implementing the algorithm.

* This research was supported by the Sookmyung Women's Univ. Research Grants.

1) Associate Professor, Department of Statistics, Sookmyung Women's Univ.

2) Graduate student, Department of Statistics, Sookmyung Women's Univ.