응용통계연구 제 **13**권 1호, 2000년, pp. 105-113

중복분석의 확장과 이를 이용한 일반화 정준상관분석

강현철[1] 김기영[2]

요 약

Wollenberg(1977)의 중복분석(redundancy analysis)을 두 개 이상의 변수집단이 주어져 있는 경우로 확장하고, 확장된 중복분석과 일반화 정준상관분석의 관계를 논의하며, 이 관계를 이용하여 새로운 형태의 일반화 정준상관분석을 제안한다.

주요용어: 중복분석, 일반화 정준상관분석, 연관성.

1. 서론

Wollenberg(1977)에 의해서 제안된 중복분석(redundancy analysis)은 두 변수집단 간의 중복(redundancy)을 최대화하는 일단의 변량들을 유도하는 것이다. 두 변수집단 $\mathbf{x}$와 $\mathbf{y}$에 대해서 그의 구성원들이 각각 평균 0과 단위분산을 가지도록 표준화되어 있다고 할 때, 변수집단 $\mathbf{x}$의 선형결합 $z_x = \mathbf{a}'\mathbf{x}$에 대한 $\mathbf{y}$의 중복은 다음과 같다.

$$RD_{y|x} = \frac{1}{p_y}\mathbf{a}'\mathbf{R}_{xy}\mathbf{R}_{yx}\mathbf{a}, \tag{1.1}$$

여기서 p_y는 $\mathbf{y}$ 집단의 변수 개수, $\mathbf{R}_{xy}$는 $\mathbf{x}$와 $\mathbf{y}$의 집단 간 상관계수행렬, $\mathbf{a}$는 계수벡터를 나타낸다. 일반성을 잃지 않고 z_x가 단위분산을 가진다고 가정하면, 중복분석에서의 문제는 선형결합 $\hat{z}_x = \hat{\mathbf{a}}'\mathbf{x}$와 $\mathbf{y}$ 변수들과의 제곱상관의 합을 최대화로 하는 계수벡터 $\hat{\mathbf{a}}$을 찾는 것이 된다. 위 식 (1.1)에서 $\mathbf{a}'\mathbf{R}_{xy}\mathbf{R}_{yx}\mathbf{a}$는 $\mathbf{x}$의 선형결합에 의해서 설명되는 $\mathbf{y}$ 변수들의 분산을 나타내며, 이를 최대화하는 계수벡터 $\hat{\mathbf{a}}$은 다음과 같은 행렬

$$\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1/2} \tag{1.2}$$

에 대한 고유치분해(eigenvalue decomposition)에 의해서 쉽게 구할 수 있다. 이와 유사하게 $\mathbf{x}$ 변수들과의 제곱상관의 합을 최대화하는 $\mathbf{y}$의 선형결합 $\hat{z}_y = \hat{\mathbf{b}}'\mathbf{y}$는 $\mathbf{R}_{yy}^{-1/2}\mathbf{R}_{yx}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2}$에 대한 고유치분해에 의해서 구할 수 있다.

중복분석은 몇 가지 관점에서 정준상관분석과 다변량회귀분석의 중간에 위치한다고 볼 수 있다. 개념적으로 볼 때 정준상관분석은 두 변수집단 간의 연관(association)에, 그러나 다변량회귀분석은 한 변수집단에 의한 다른 변수집단의 예측(prediction)에 주 관심을 두

1) (136-701) 서울시 성북구 안암동 5가, 고려대학교 통계연구소 선임연구원
E-mail: hychkang@kuccnx.korea.ac.kr
2) (136-701) 서울시 성북구 안암동 5가, 고려대학교 통계학과 교수
E-mail: kykim@kuccnx.korea.ac.kr

고 있다. 이에 대해 중복분석은 연관과 예측 모두에 관심을 두고 있다고 볼 수 있다. 이러한 세가지 분석방법(정준상관분석, 중복분석, 다변량회귀분석)에 대한 비교는 Muller(1981)와 Lambert et. al.(1988) 등에서 자세히 다루고 있다.

잘 알려져 있는 바와 같이 정준상관분석에서 $\mathbf{x}$의 해는 행렬 $\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1/2}$에 대한 고유치분해로부터 구할 수 있는데, 중복분석에서는 식 (1.2)에서 볼 수 있듯이 고유치분해의 대상이 되는 행렬에 변수집단 $\mathbf{y}$의 상관행렬의 역행렬 $\mathbf{R}_{yy}^{-1}$을 포함하고 있지 않다. 이것은 $\mathbf{x}$에 대한 해 $\hat{\mathbf{a}}$가 중복분석에서는 $\mathbf{x}$의 집단 내 상관구조에는 영향을 받지 않으나 $\mathbf{y}$의 집단 내 상관구조에는 영향을 받는다는 것을 의미한다. 즉, 식 (1.2)에 포함되어 있는 두 개의 $\mathbf{R}_{xx}^{-1/2}$는 중복분석의 해 $\hat{\mathbf{a}}$에 대한 변수집단 $\mathbf{x}$의 집단 내 상관구조의 영향을 제거하는 역할을 수행한다. 그러나 정준상관분석에서는 두 변수집단 $\mathbf{x}$와 $\mathbf{y}$ 모두의 집단 내 상관구조가 해에 영향을 주지 않는다.

이와 같은 특성을 이용하면 정준상관분석과 중복분석의 관계를 변환불변성(transformation invariance)이라는 관점에서 보다 깊게 살펴볼 수 있다. 즉, 정준상관분석은 두 변수집단의 집단 내 상관구조를 제거하는 정칙변환(nonsingular transformation)에 대해 불변성을 가지고 있지만 중복분석은 이러한 정칙변환에 대해서 불변성을 가지고 있지 않다. 따라서 다음과 같이 변환된 상관행렬

$$\mathbf{R}_{xy}^{*} = \mathbf{T}_{x}^{-1'}\mathbf{R}_{xy}\mathbf{T}_{y}^{-1} \tag{1.3}$$

에 대해서 중복분석을 수행하여 얻은 해는 정준상관분석의 해와 동일하게 된다. 여기서 $\mathbf{T}_{x}^{-1}$와 $\mathbf{T}_{y}^{-1}$는 정칙변환으로써 다음 관계를 만족한다.

$$\mathbf{T}_{x}^{-1'}\mathbf{R}_{xx}\mathbf{T}_{x}^{-1} = \mathbf{T}_{y}^{-1'}\mathbf{R}_{yy}\mathbf{T}_{y}^{-1} = \mathbf{I}. \tag{1.4}$$

바꾸어 말하면, 집단 내 상관구조를 제거하도록 변환된 상관행렬에 중복분석을 수행하여 얻은 해는 오직 집단 간 상관구조에만 영향을 받으므로 정준상관분석의 해와 동일하게 된다.

2. 중복분석의 확장

표준화된 $m(\geq 2)$개의 변수집단 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$에 대해서, j 번째 변수집단의 선형결합 $z_j = \mathbf{a}_j'\mathbf{x}_j$에 의해서 설명되는 i 번째 변수집단의 분산($EV_{i|j}$)은 다음과 같이 표현될 수 있다.

$$EV_{i|j} = \mathbf{a}_j'\mathbf{R}_{ji}\mathbf{R}_{ij}\mathbf{a}_j, \ i \neq j. \tag{2.1}$$

이제 모든 $EV_{i|j}(i \neq j)$들의 평균(EV),

$$EV = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} EV_{i|j} \tag{2.2}$$

를 최대화하는 단위분산을 가지는 m 개의 선형결합 $\hat{z}_i = \hat{\mathbf{a}}_i'\mathbf{x}_i(i = 1, \cdots, m)$를 찾는 문제를 고려해 보자.

정리 2.1 식 (2.2)의 EV를 최대화하는 계수벡터 $\hat{\mathbf{a}}_i(i=1,\cdots,m)$는 다음과 같은 관계식을 만족한다.

$$(\mathbf{R}_{ii}^{-1/2}\sum_{j\neq i}^{m}\mathbf{R}_{ij}\mathbf{R}_{ji}\mathbf{R}_{ii}^{-1/2})\mathbf{R}_{ii}^{1/2}\hat{\mathbf{a}}_i=\lambda_i\mathbf{R}_{ii}^{1/2}\hat{\mathbf{a}}_i, \tag{2.3}$$

여기서 $\hat{\mathbf{a}}_i$는 제약조건 $\hat{\mathbf{a}}_i'\mathbf{R}_{ii}\hat{\mathbf{a}}_i=1(i=1,\cdots,m)$을 가지는 계수벡터이며, λ_i는 행렬

$$\mathbf{R}_{ii}^{-1/2}\sum_{j\neq i}^{m}\mathbf{R}_{ij}\mathbf{R}_{ji}\mathbf{R}_{ii}^{-1/2} \tag{2.4}$$

의 최대 고유값이다.

증명: EV를 최대화하는 계수벡터 $\mathbf{a}_i$를 찾는 것은 다음과 같은 목적함수

$$L=\sum_{i=1}^{m}\sum_{j\neq i}^{m}\mathbf{a}_j'\mathbf{R}_{ji}\mathbf{R}_{ij}\mathbf{a}_j-\sum_{i=1}^{m}\lambda_i(\mathbf{a}_i'\mathbf{R}_{ii}\mathbf{a}_i-1) \tag{2.5}$$

를 최대화하는 문제가 된다. 여기서 $\lambda_1,\cdots,\lambda_m$은 라그랑즈 승수이다. 목적함수 L을 계수벡터 $\mathbf{a}_i$에 대해서 편미분하고 이를 0으로 놓으면 다음과 같은 식을 얻을 수 있다.

$$\frac{\partial L}{\partial\mathbf{a}_i}=2\sum_{j\neq i}^{m}\mathbf{R}_{ij}\mathbf{R}_{ji}\mathbf{a}_i-2\lambda_i\mathbf{R}_{ii}\mathbf{a}_i=\mathbf{0}. \tag{2.6}$$

이 식 (2.6)은 다음과 같이 바꾸어 쓸 수 있다.

$$\sum_{j\neq i}^{m}\mathbf{R}_{ij}\mathbf{R}_{ji}\mathbf{a}_i=\lambda_i\mathbf{R}_{ii}\mathbf{a}_i. \tag{2.7}$$

식 (2.7)의 양변에 $\mathbf{R}_{ii}^{-1/2}$을 앞에서 곱하고 좌변의 중간에 $\mathbf{R}_{ii}^{-1/2}\mathbf{R}_{ii}^{1/2}$을 삽입하면 식 (2.3)을 얻을 수 있다. □

한편 행렬 (2.4)는 $\hat{\mathbf{a}}_j(j\neq i)$를 포함하고 있지 않기 때문에, 계수벡터 $\hat{\mathbf{a}}_i$는 다른 변수집단의 선형결합 $\hat{z}_j=\hat{\mathbf{a}}_j'\mathbf{x}_j(j\neq i)$를 고려하지 않고 구할 수 있다. 즉, $\mathbf{v}_i$를 행렬 (2.4)의 고유벡터라고 할 때 $\hat{\mathbf{a}}_i=\mathbf{R}_{ii}^{-1/2}\mathbf{v}_i$가 된다.

일반적으로 중복분석에서 주요 관심은 제 1 중복변량(redundancy variate)의 결과에 있겠지만 때로는 $2,3,\cdots,k$ 번째 단계에서의 결과를 고려할 필요가 있는데, 이 경우에 어느 단계에서 얻어진 중복변량은 그 이전의 단계에서 얻어진 동일 변수집단에 대한 중복변량들과 무상관을 가지도록 하는 것이 바람직할 것이다. 이러한 변수집단 내 중복변량들 간의 직교성(orthogonality)은 다음과 같은 추가적인 제약조건을 부여함으로써 얻어질 수 있다.

$$\hat{\mathbf{a}}_{i(k)}'\mathbf{R}_{ii}\hat{\mathbf{a}}_{i(l)}=0,\ i=1,\cdots,m;\ l=1,\cdots,k-1, \tag{2.8}$$

여기서 $\hat{\mathbf{a}}_{i(k)}$는 i 번째 변수집단에 대한 k 번째 단계에서의 계수벡터를 나타낸다. 식 (2.8)을 추가로 만족하는 $\hat{\mathbf{a}}_{i(k)}$는 행렬 (2.4)의 k 번째 고유벡터에 비례한다는 것을 쉽게 보일 수 있다. 한편 식 (2.8)을 만족하는 해로부터 얻어진 중복변량들은 두 변수집단 중복분석에서와 마찬가지로 $\hat{z}_{i(k)} = \hat{\mathbf{a}}'_{i(k)}\mathbf{x}_i$과 $\hat{z}_{j(l)} = \hat{\mathbf{a}}'_{j(l)}\mathbf{x}_j$ 간의 상관계수가 반드시 0이 될 필요는 없다($i \neq j = 1, \cdots, m$). Wollenberg (1977)는 두 집단의 중복변량들이 반드시 서로 직교할 필요는 없다고 언급한 바 있는데, 이와 같은 중복변량들의 집단 간 비직교성은 중복분석이 정준상관분석과 다른 중요한 차이점 중의 하나라고 할 수 있다.

3. 확장된 중복분석을 이용한 새로운 일반화 정준상관분석

여기서는 앞에서 논의된 정준상관분석과 중복분석의 관계를 다중변수집단의 경우로 자연스럽게 확장하여 새로운 일반화 정준상관분석을 유도한다. 먼저 $\mathbf{R}^*_{ii} = \mathbf{I}$를 만족하도록 변환된 다음과 같은 상관행렬

$$\mathbf{R}^*_{ij} = \mathbf{T}_i^{-1'}\mathbf{R}_{ij}\mathbf{T}_j^{-1}, \; i = 1, \cdots, m \tag{3.1}$$

을 고려해 보자. 이 변환된 상관행렬들은 집단내 상관구조를 제거한 것이므로, 이 상관행렬들에 앞에서 전개된 확장된 중복분석의 과정을 적용하면 새로운 형태의 일반화 정준상관분석 과정을 도출할 수 있다. 즉, 식 (2.2)와 유사한 형식의 다음과 같은 통계량 EV^*

$$EV^* = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \mathbf{a}_j^{*\prime}\mathbf{R}^*_{ji}\mathbf{R}^*_{ij}\mathbf{a}_j^* \tag{3.2}$$

는 각 변수집단의 정준변량(canonical variates)이 모든 다른 변수집단의 변수들과 가지는 제곱상관들의 평균으로 해석될 수 있다.

여기서 계수벡터 $\hat{\mathbf{a}}_i^*$는 제약조건 $\hat{\mathbf{a}}_i^{*\prime}\hat{\mathbf{a}}_i^* = 1 (i = 1, \cdots, m)$ 하에서 다음과 같은 행렬

$$\sum_{j \neq i}^{m} \mathbf{T}_i^{-1'}\mathbf{R}_{ij}\mathbf{R}_{jj}^{-1}\mathbf{R}_{ji}\mathbf{T}_i^{-1} \tag{3.3}$$

의 고유치분해에 의해서 구할 수 있다. 이 때 $\hat{\mathbf{a}}_i^*$는 행렬 (3.3)의 고유벡터가 되고, 따라서 일반화 정준상관분석의 해는 $\hat{\mathbf{a}}_i = \mathbf{T}_i^{-1}\hat{\mathbf{a}}_i^*$가 된다.

한편 행렬 (3.3)의 고유벡터 $\hat{\mathbf{a}}_i^*$에 대응하는 고유값 λ_i^*는 i 번째 변수집단의 정준변량이 모든 다른 변수집단의 변수들과 가지는 제곱상관들의 합계를 의미하며, 이런 관점에서 다음과 같은 $\hat{\rho}_k$을 k 번째 단계에서의 일반화 정준상관계수의 측도로 사용할 수 있을 것이다.

$$\hat{\rho}_k^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \lambda^*_{i(k)}, \tag{3.4}$$

여기서 $\lambda^*_{i(k)}$는 행렬 (3.3)의 k 번째 고유값을 나타낸다. 두 개의 변수집단만을 고려하는 경우($m = 2$)에는 $\hat{\rho}_k^2$은 단순히 k 번째 정준상관계수의 제곱과 같음을 쉽게 보일 수 있다.

표 4.1: 세 변수집단으로 이루어진 변수들에 대한 상관행렬

Group	1			2			3		
	1	.362	.181	.017	.021	.047	.401	.632	.338
1	.362	1	-.024	.184	.327	.241	.685	.418	.174
	.181	-.024	1	.358	.384	.631	-.096	.052	-.036
	.017	.184	.358	1	.644	.438	-.100	-.082	-.073
2	.021	.327	.384	.644	1	.878	.138	.074	-.029
	.047	.241	.631	.438	.878	1	.074	.138	-.198
	.401	.685	-.096	-.100	.138	.074	1	.848	.301
3	.632	.418	.052	-.082	.074	.138	.848	1	.150
	.338	.174	-.036	-.073	-.029	-.198	.301	.150	1

4. 사례분석

이 절에서는 앞에서 제안된 확장된 중복분석과 일반화 정준상관분석(기존에 제안된 일반화 정준상관분석들을 포함하여)을 비교하기 위하여 하나의 자료에 이들 분석방법들을 적용하고 결과를 비교할 것이다. 본 사례분석에서 고려하고 있는 자료는 Lambert et. al.(1988)에서 중복분석과 정준상관분석을 비교하기 위하여 사용되었는데, 원래는 각각 4개와 5개의 변수를 가진 두 개의 변수집단으로 이루어져 있었다($m = 2,\ p_1 = 4,\ p_2 = 5$). 그러나 본 절에서는 편의상 상관행렬을 재배열하고 분할하여 각 3개의 변수들을 가진 3개의 변수집단을 이루도록 상관행렬을 재구성하였다($m = 3,\ p_1 = p_2 = p_3 = 3$). 이렇게 재구성된 상관행렬이 표 4.1에 제시되어 있다.

우선 각 변수집단에 대한 선형결합 $z_i = \mathbf{a}_i'\mathbf{x}_i (i = 1, \cdots, m)$들로 이루어진 벡터 $\mathbf{z} = (z_1 \cdots z_m) = (\mathbf{a}_1'\mathbf{x}_1 \cdots \mathbf{a}_m'\mathbf{x}_m)$의 상관행렬을 $\mathbf{\Phi}$라 하고 $l_1 \geq l_2 \geq \cdots \geq l_m$를 $\mathbf{\Phi}$의 순서화된 고유값이라 하자. 이 때 기존에 제안된 여러 일반화 정준상관분석 방법들은 최적화하고자 하는 $\mathbf{\Phi}$의 함수에 의해서 특성화될 수 있다(Kettenring, 1971; Kang, 1998). 본 사례분석에서는 다음과 같은 6개의 일반화 정준상관분석 방법들: (i) SUMCOR [$\sum_{i,j}^{m} \phi_{ij}$를 최대화]; (ii) MAXVAR [l_1을 최대화]; (iii) SSQCOR [$\sum_{i,j}^{m} \phi_{ij}^2$를 최대화]; (iv) MINVAR [$l_m$을 최소화]; (v) MAXECC [$(l_1 - l_m)/(l_1 + l_m)$을 최대화]; (vi) GENVAR [det($\mathbf{\Phi}$)를 최소화], 을 고려하였다.

Kettenring(1971)과 Kang(1998) 등에서 논의된 바에 의하면 SUMCOR, MAXVAR, SSQCOR은 $\mathbf{\Phi}$의 최대 고유값에 주로(또는 전적으로) 의존하기 때문에 서로 유사한 해를 생성할 가능성이 많은데, 이들 방법들은 $\mathbf{\Phi}$의 모든 원소를 전체적으로 크게 하려는 경향이 있다. 반면에, MINVAR와 GENVAR는 $\mathbf{\Phi}$의 최소 고유값에 주로(또는 전적으로) 의존하기 때문에 이 두 방법은 유사한 해를 생성할 가능성이 많고, $\mathbf{\Phi}$를 되도록 비정칙행렬에 가깝도록 하는 경향이 있다. 한편 MAXECC의 결과는 위와 같은 두 상황의 중간에 위치하는 경향이 있다.

표 4.2: 각 단계에서의 평균설명분산

k	ERA	NEW-GCCA	SUM COR	MAX VAR	SSQ COR	MIN VAR	MAX ECC	GEN VAR
1	0.444	0.268	0.251	0.247	0.236	0.369	0.368	0.369
2	0.269	0.301	0.337	0.339	0.279	0.280	0.283	0.283
3	0.042	0.187	0.168	0.170	0.192	0.107	0.104	0.104

표 4.3: 첫 번째 단계에 대한 계수벡터와 중복/정준변량 간 상관계수

	ERA	NEW-GCCA	SUM COR	MAX VAR	SSQ COR	MIN VAR	MAX ECC	GEN VAR
	0.170	0.660	0.401	0.438	0.546	-0.701	-0.701	-0.701
$\mathbf{a}_1$	0.781	-0.853	-0.727	-0.765	-0.863	0.838	0.838	0.838
	0.502	0.373	0.637	0.591	0.438	0.649	0.649	0.649
	0.709	0.713	0.818	0.821	0.826	0.744	0.744	0.744
$\mathbf{a}_2$	-1.885	-1.998	-2.507	-2.531	-2.590	-1.803	-1.803	-1.803
	2.032	2.085	2.130	2.118	2.058	1.967	1.966	1.967
	0.854	-1.975	-1.886	-1.908	-1.955	1.898	1.898	1.898
$\mathbf{a}_3$	0.123	1.705	1.618	1.629	1.657	-1.737	-1.737	-1.737
	0.115	0.204	0.000	0.037	0.135	-0.049	-0.050	-0.050
ϕ_{12}	0.422	0.362	0.568	0.542	0.465	0.500	0.499	0.500
ϕ_{13}	0.551	0.851	0.723	0.753	0.824	0.546	0.546	0.546
ϕ_{23}	-0.190	0.436	0.576	0.570	0.550	-0.451	-0.451	-0.451
l_1	1.6111	2.1329	2.2487	2.2496	2.2413	1.5544	1.5547	1.5544
l_3	0.2062	0.1446	0.2760	0.2455	0.1682	0.0003	0.0003	0.0003

본 사례분석에서는 확장된 중복분석(ERA)과 앞에서 제안된 일반화 정준상관분석(NEW-GCCA) 그리고 기존의 6개 일반화 정준상관분석에 대해서 해를 구하였다. 앞절에서 서술한 바와 같이, 모든 $k(=1,2,3)$ 번째 단계에 대한 ERA와 NEW-GCCA의 해는 단일 고유치 분해에 의해서 쉽게 구할 수 있다. 그러나 MAXVAR와 MINVAR의 첫 번째 단계를 제외한 기존의 6개 일반화 정준상관분석에 대해서는 Kang(1998)에서 제안된 반복적 방법에 의해서 해를 구하였다.

표 4.2는 각 k 번째 단계에서 얻어진 결과로부터 계산된 $EV_{i|j}$들의 평균, 즉 평균설명분산(mean explained variance)

$$EV_{(k)} = \frac{1}{3(3-1)} \sum_{i=1}^{3} \sum_{j \neq i}^{3} \hat{\mathbf{a}}'_{j(k)} \mathbf{R}_{ji} \mathbf{R}_{ij} \hat{\mathbf{a}}_{j(k)}, \ k = 1,2,3 \tag{4.1}$$

을 제시한 것이다. 확장된 중복분석(ERA)은 각 단계에서 평균설명분산을 최대화하는 것이기 때문에, 당연히 첫 번째 단계에서의 평균설명분산이 제일 크고 다음 단계로 갈수록 값이 작아진다. 또한 몇 개의 일반화 정준상관분석(MINVAR, MAXECC, GENVAR)은 확장된 중복분석과 비슷한 결과를 보여주고 있다. 그러나 다른 일반화 정준상관분석(NEW-GCCA, SUMCOR, MAXVAR, SSQCOR)의 결과에서는 두 번째 단계에 대한 평균설명비율이 첫 번째 단계에 대한 평균설명비율보다 크다. 즉, 일반화 정준상관분석은 변수집단들 간의 연관정도를 최대화하는 것이기 때문에, 평균설명분산이 중복분석에서와 같이 순서화될 필요는 없다는 것을 보여주고 있다.

일반적으로 연구자는 첫 번째 단계에서의 결과, 즉 제 1 중복 또는 제 1 정준변량에 주 관심을 두게 되는데, 표 4.3은 각 방법의 첫 번째 단계에서 얻어진 계수벡터 $\hat{\mathbf{a}}_{i(1)}(i=1,2,3)$과 변량 간 상관계수 $\phi_{ij}=\hat{\mathbf{a}}'_{i(1)}R_{ij}\hat{\mathbf{a}}_{j(1)}(i,j=1,2,3)$을 제시하고 있다. 이 결과를 살펴보면 일반화 정준상관분석 방법들이 두 개의 서로 다른 형태의 결과를 보이고 있음을 알 수 있다. 먼저, SUMCOR, MAXVAR, SSQCOR은 전체적으로 비슷한 결과를 보여주고 있는데, 이 방법들의 결과에서는 정준변량 간 상관계수가 전체적으로 큰 값을 나타내고 있다. 반면에, MINVAR와 GENVAR에 대해서는 두 번째 변수집단과 세 번째 변수집단의 정준변량 간 상관계수가 큰 음의 값을 가지고 있는 등 다른 패턴을 보이고 있다.

표 4.3에는 각 방법에 의해서 얻어진 행렬 $\mathbf{\Phi}$의 최대 고유값 l_1과 최소 고유값 l_3들도 제시되어 있다. 특히, MAXVAR에 의해서 얻어진 $l_1^{Max}=2.2496$과 MINVAR에 의해서 얻어진 $l_3^{Min}=0.0003$은 모든 가능한 $\mathbf{\Phi}$에 대해서 얻을 수 있는 최대 고유값과 최소 고유값으로써 각각 행렬 $\mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2}$의 최대 고유값 그리고 최소 고유값과 같다(Kettenring, 1971; Kang, 1998). 여기서 $\mathbf{D}$는 i 번째 블록을 $\mathbf{R}_{ii}$로 가지는 블록대각행렬이다. 한편 MAXECC는 $\mathbf{\Phi}$의 두 극단 고유값으로부터 큰 영향을 받는데, 본 사례에서와 같이 $\mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2}$의 최소 고유값이 상대적으로 0에 가까운 경우에는 $\mathbf{\Phi}$의 최소 고유값에 영향을 더 많이 받게 되므로 당연히 MINVAR와 GENVAR에 가까운 결과를 보이고 있다. 일반적으로 $\mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2}$의 최대 고유값과 최소 고유값이 상대적으로 큰 경우에는, 모든 방법들이 $\mathbf{\Phi}$의 최소 고유값에 덜 영향을 받게 되기 때문에 서로 유사한 결과를 제공하게 될 것이다.

한편 NEW-GCCA는 각 정준변량과 다른 변수집단에 속하는 변수들과의 모든 상관계수의 제곱합을 최대화하는 것이기 때문에 간접적으로 $\mathbf{\Phi}$의 모든 원소를 전체적으로 크게하는 경향이 있고, 따라서 본 사례분석의 결과에서와 같이 SUMCOR, MAXVAR, SSQCOR와 유사한 결과를 제공하는 경향이 있다. 그러나 다른 방법들과는 달리 제약조건 (2.8)을 가지는 모든 해를 단일 고유치분해에 의해서 쉽게 구할 수 있다는 것이 이 방법의 장점이라 할 수 있을 것이다.

표 4.3으로부터 두 번째와 세 번째 변수집단에 대응하는 ERA의 중복변량 간 상관계수가 $\phi_{23}=-0.190$으로써 다소 작은 음의 값을 가지고 있다는 것을 알 수 있다. 그러나 일반화 정준상관분석은 최적화 함수를 통하여 집단 간 연관정도를 최대화하고자 하는 것임을 주지할 때, 정준변량 간 상관계수의 절대값이 상대적으로 큰 경향을 나타낸다. 표 4.1의 원래 상관행렬을 자세히 살펴보면 특히 두 번째와 세 번째 변수집단에 속하는 변수들 간의 집단 간 상관계수의 일부가 다소 작지만 음의 상관계수들을 가지고 있다는 것을 알 수 있는

데, 따라서 이 자료에 대해서는 ERA의 결과가 표 4.1에 제시된 상관행렬 **R**의 구조를 적절히 표현하고 있음을 보여주고 있다.

참고문헌

[1] Kang, H. (1998). A study of Multivariate Structural Relationships for Groups of Variables, *Ph.D. Dissertation*, Korea University, Seoul, Korea.

[2] Kettenring, J.R. (1971). Canonical analysis of several sets of variables, *Biometrika*, vol. 58, 433-451.

[3] Lambert, Z.V., Wildt, A.R. and Durand, R.M. (1988). Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interset associations, *Psychological Bulletin*, vol. 104, 282-289.

[4] Muller, K.E. (1981). Relationships between redundancy analysis, canonical correlation, and multivariate regression, *Psychometrika*, vol. 46, 139-142.

[5] Van den Wollenberg, A.L. (1977). Redundancy analysis: An alternative for canonical analysis, *Psychometrika*, vol. 42, 207-219.

[1999년 8월 접수, 1999년 11월 채택]

A Note on Generalized Canonical Correlation Analysis Via an Extended Redundancy Analysis

Hyuncheol Kang[1] Keeyoung Kim[2]

ABSTRACT

An extension of Wollenberg's (1977) redundancy analysis is made to the case of more than two groups of variables, and relationships between the extended redundancy analysis and the generalized canonical correlation analysis are discussed. This paper suggests an alternate form of the generalized canonical correlation analysis from the relationships.

Keywords: Redundancy Analysis; Generalized Canonical Correlation Analysis; Association.

1) Senior Researcher, Institute of Statistics, Korea University. E-mail: hychkang@kuccnx.korea.ac.kr
2) Professor, Department of Statistics, Korea University. E-mail: kykim@kuccnx.korea.ac.kr