The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 387-397

An Agglomerative Hierarchical Variable-Clustering Method Based on a Correlation Matrix

Kwangjin Lee[1)]

Abstract

Generally, most of researches that need a variable-clustering process use an exploratory factor analysis technique or a divisive hierarchical variable-clustering method based on a correlation matrix. And some researchers apply a object-clustering method to a distance matrix transformed from a correlation matrix, though this approach is known to be improper. On this paper an agglomerative hierarchical variable-clustering method based on a correlation matrix itself is suggested. It is derived from a geometric concept by using variate-spaces and a characterizing variate.

Keywords : Agglomerative, Hierarchical, Variable Clustering, Correlation Matrix

1. 서론

군집분석법은 유사성 또는 비유사성에 근거하여 인접된 구성원들을 소수의 집단으로 군집화 함으로써, 자료의 구조와 군집의 형성과정을 파악하고 군집간의 관계 등을 분석하는 탐색적인 분석방법이다. 일반적으로 군집분석의 상황과 목적은 매우 다양하며 이들 각각에 대응하는 많은 분석기법이 존재한다(강현철 외, 2000).

군집분석법은 군집화의 대상에 따라 개체군집법과 변수군집법으로, 군집의 개수에 관련하여 그것이 사전에 알려진 경우와 미지인 경우로, 군집의 유형에 따라 상호배반적(최적분리) 군집, 계층적 군집, 중복 군집, 퍼지 군집 등으로 분류되어진다. 그리고 계층적 군집분석법은 다시 병합/계층적인 것과 분할/계층적인 것으로 분류된다. 군집분석법에 관한 기존의 수많은 연구에서 변수군집법에 관한 연구는 개체군집법에 관한 연구에 비해 양적으로도 미미할 뿐만 아니라 태생적으로도 많이 늦었다고 할 수 있다. 그러한 이유 때문인지는 모르겠지만 변수군집법 전용의 방법(예를 들어, SAS의 VARCLUS 절차)이 지금은 잘 개발되어 있음에도 불구하고 아직까지도 변수군집화가 필요한 상황에서 통상적으로 개체 군집분석에서 사용되는 계보적 또는 비계보적 알고리즘에 의하여 변수의 군집화를 시도하거나(강현철 외, 2000), 탐색적 인자분석이 실제로 많이 사용되고 있음(Nunnally & Berstein, 1994; Harman, 1976)은 안타까운 현실이라고 할 수 있다. 물론 이러한 사용이 적절치 못함은 '이에 따른 수리적 모형의 도입이나 회전 등의 문제가 추가적으로 발생하게

1) Associate Professor, Department of Statistics, Mokwon University, Daejeon, 302-729, Korea
E-mail : leekj@mokwon.ac.kr

되므로 변수군집분석을 직접 시도할 것'이라는 Kendall(1980)의 추천과 SAS에서 변수군집법 절차를 따로 만든 그 이유를 통해서도 알 수 있겠다. 실제로 많이 사용되어지는 통계분석 소프트웨어 중에서도 개체군집화와 변수군집화를 특별히 구분하지 않고 하나의 절차 내에서 처리되도록 한 것들도 있다.

변수군집법 중 최적분리 군집법은 통상적으로 미리 규정된 판정기준을 최적화시키도록 시도하고 있으며, 많은 경우 연구자에 의해 군집의 개수가 미리 결정되어 있다. 이에 관련된 연구로는 강현철 외(2000)의 연구에 잘 나타나 있다. 또한 변수군집법 중 분할/계층적 군집분석법에 관해서는 Harman(1976)의 방법을 보다 개선하여 만든 SAS의 VARCLUS 절차가 독보적이라고 할 수 있다. 이 절차를 이용하면 주성분군집분석법(Oblique Principal Component Variable Clustering), 중심성분군집분석법(Oblique Centroid Component Variable Clustering)이 가능하다. 물론 이 절차는 군집분할적 접근법을 취하고 있다. 즉 분석대상이 되는 p개의 모든 변수들을 포함하는 하나의 군집에서부터 시작하여 첫 단계에서는 이를 두 개의 군집으로 분할하고, 두 번째 단계에서는 첫 단계에서 생성되었던 두 군집 중 설명력이 약한 군집을 두 개의 군집으로 분할함으로써 세 개의 군집을 형성한다. 이런 과정을 반복하여 p개의 군집으로 모두 분할하게 된다.

병합/계층적 방식으로의 변수 군집화를 위해서는 사실 여태까지의 많은 응용연구에서 상관행렬을 거리행렬로 변환한 후 변수를 개체로 보고(사실은 개체와 변수를 구별하지 않고) 병합/계층적 개체군집법을 적용하는 방식을 택했던 것이 사실이다. 이 방법은 원래 비유사성 행렬에 근거하여 개체를 군집화하는 절차라는 점과 상관행렬을 거리행렬로 변환할 때 특별한 기준이 없을 뿐만 아니라 주관성이 많이 개입된다는 점 때문에 변수의 군집화 방법으로는 적절치 못한 방식으로 알려져 있다. 참고로, SAS의 CLUSTER 절차에서는 비유사성 측도를 사용하기 때문에 입력자료로 'TYPE=CORR'인 데이터를 사용할 수도 없다. 굳이 이 절차를 사용해야 한다면 Data-step 또는 IML 절차를 이용하여 $(1-r^2)$와 같은 변환을 통해 비유사성 행렬로 변환한 후 이 절차를 이용하여야 한다(SAS Institute, 1993).

본 연구에서는 병합/계층적 변수군집법을 새로이 개발하기 위해, 2장에서는 먼저 개체군집법과 변수군집법의 근본차이를 살펴보고, 3장에서는 변량공간과 변량공간의 특성변량이라는 개념을 도입하고 이에 관련된 기하를 설명한다. 4장에서는 병합/계층적 변수군집법을 제안하면서 그 알고리즘을 밝히고, 5장에서는 예제를 통해 그 활용가능성을 살펴보고자 한다.

2. 개체군집법과 변수군집법의 근본차이

군집분석에 관한 지금까지의 많은 논문 및 저서들 중에서 개체군집법과 변수군집법의 근본적인 차이점을 제대로 지적한 것을 저자는 아직까지 접해본 적이 없다. 군집화의 대상이 개체와 변수라는 명칭의 차이점, 변수의 군집화는 주로 상관행렬에 바탕을 두고 이루어지는데 군집화를 위해서는 상관행렬을 적절한 변환을 통해 비유사성 행렬로 바꾼 후 군집분석을 수행해야 한다는 점 정도만 지적되어 있다. 이런 정도만 지적되어 있다보니 변수를 마치 유사성 측도로 측정된 개체처럼 취급하여 별 생각없이 개체군집법을 차용하고 있는 것이 아닌가하고 저자는 여기고 있다. 물론 이러한 사용에는 개체군집법과 변수군집법을 특별히 구별하지 않는 통계패키지들의 영향도 컸을 것으로도 저자는 생각하고 있다.

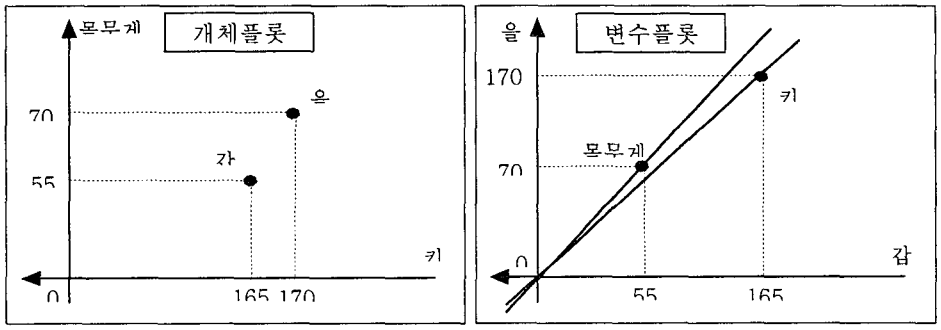
본 연구에서는 변수군집화를 위해 개체군집법을 차용하는 방법이 적절하지 않다는 점을 먼저 개체와 변수의 근본차이를 지적하는 것을 통해 보이고자 한다. 다음의 (I)자료는 가, 나 두 사람에게 키와 몸무게를 측정한 자료이다.

(I)	키	몸무게
가	165	55
나	170	70

(II)	키*	몸무게
가	1.65	55
나	1.70	70

(III)	키	몸무게
가*	330	110
나	170	70

(I)의 자료에서 키의 측정단위를 cm에서 m로 변화시켜 (II)의 자료로 바꾸더라도 정보의 손실은 없기 때문에 사실 (I)과 (II)는 같은 자료로 볼 수 있다. 그러나 (I)의 자료에서 갑의 측정치만 2배 하여 (III)의 자료로 바꾸면 (I)과 (III)은 완전히 다른 자료가 된다. 이는 변수는 변환이 허용되지만 개체는 변환이 허용되지 않음을 의미한다. 다시 말하면, 변환이란 이동이란 개념을 수반하는 것이므로 변수점은 공간상에서 어떤 형태의 룰에 따라 이동이 가능하지만 개체점은 이동이 불가능한 고정점이라는 것을 의미한다. 이를 [그림1]과 같이 좌표공간에서 표현하여 설명하면 다음과 같다. 개체플롯에서 갑과 을 두 개체점들은 조금이라도 단독적으로 이동하게 되면 자료가 완전히 변하기 때문에 움직일 수 없는 고정점이라고 할 수 있다. 하지만 변수플롯에서 키와 몸무게에 해당되는 변수점들은 원점과 그 점을 이은 직선 위를 임의로 이동하더라도 정보의 손실이 없기 때문에 개별적 이동이 가능한 유동점이라 할 수 있다. 이는 개체는 고정된 하나의 점(point)의 개념으로 보아야 하지만 변수는 1차원적인 공간(space) 즉 선(line)의 개념으로 보아야 한다는 점을 암시한다.



[그림 1] 개체플롯과 변수플롯

개체와 변수에 관한 이러한 근본차이는 개체의 군집은 단순 집합에 불과하지만 변수의 군집은 변수들의 단순 집합이 아닌 그들이 생성하는 변량공간들의 합성으로 보아야 함을 의미한다고 할 수 있다. 즉, 개체의 군집화와 달리 변수의 군집화는 그 대상이 변수 그 자체가 아니라 이들이 생성하는 공간으로 보아야 한다는 것을 의미하는 것으로 볼 수 있다. 이는 또한 개체군집법과 변수군집법을 동일시 할 수 없음을 의미하는 것으로 이런 의미에서 변수 군집화를 위해 개체군집법을 차용하는 것은 옳지 못하다고 할 수 있겠다. 또한 이러한 차이에 의해 의미상으로는 병합/계층적 변수군집법이란 용어보다는 단계적 변량공간합성법이, 분할/계층적 변수군집법이란 용어보다는 단계적 변량공간분할법이라는 용어가 더 적절할 것으로 저자는 생각한다. 하지만 본 연구에서는 기

존 연구들과의 혼동을 피하기 위해 변수군집이란 용어를 그냥 그대로 사용할 것이다.

3. 변량공간, 특성변량, 변량공간간 거리, 그리고 관련 기하

2장에서의 서술을 근거로 본 연구에서는 변수집합과 변량공간이란 것을 다음과 같이 정의한다.

[정의 1] 변수집합과 변량공간

p개의 변수들 $x_1, x_2, \cdots, x_p$이 주어져 있을 때 이들의 단순 집합인 $x_1, x_2, \cdots, x_p$을 이 p개 변수들의 변수집합(a set of variables)이라 하고, 이들 변수들의 선형결합으로 표현되어질 수 있는 모든 변량들의 집합, 즉 $\{y | y = a_1x_1 + a_2x_2 + \cdots + a_px_p,\ a_1, a_2, \cdots, a_p$는 임의의 실수$\}$을 변수집합 $x_1, x_2, \cdots, x_p$의 변량공간(variate-space)으로 정의한다. 이를 경우에 따라서는 '변수집합 $x_1, x_2, \cdots, x_p$에 의해 생성(span/ generate)된 변량공간'으로 명명한다.

[정의 2] 변량공간의 특성변량

변량공간의 특성변량(a characterizing variate)이란 그 변량공간을 특성화할 수 있는 하나의 변량으로서 연구자의 판단에 의해 정해질 수 있는 것으로 정의하고 본 연구에서는 y로 표현한다.

특성변량을 정함에 있어, 변량공간을 단독적으로 볼 때에는 그 변량공간을 생성하는 변수집합의 제1주성분(first principal component), 중심성분(centroid component), 대표성분(representative component) 등이, 그리고 변량공간들 사이의 상대적 입장에 있을 때에는 해당 변량공간을 생성하는 변수집합들의 제1정준변량(first canonical variate)이 유력한 후보가 될 수 있을 것이다.

참고로, m개의 변수 $x_1, x_2, \cdots, x_m$이 주어져 있을 때 이들의 선형결합들 중에서 가중계수들의 제곱합이 1이면서 분산이 가장 큰 것을 제1주성분, 가중계수의 합이 1이면서 모두 동일한 것을 중심성분, 가중계수의 합이 1이면서 개별 구성변수들과의 상관계수가 모두 동일한 것을 대표성분이라 한다. 대표성분에 관해서는 이광진(1997)을 참조할 수 있겠다.

[정의 3] 변량공간간 상관계수와 거리

주어진 두 변량공간간 상관계수란 두 변량공간 모두 특성변량들이 정해진 경우라면 특성변량들 사이의 단순상관계수의 절대값으로, 그렇지 못한 경우라면 변량공간을 생성하는 변수집합들간의 제1정준상관계수로 정의한다. 그리고 이의 역코사인(arc-cosine) 값을 변량공간간 거리로 정의한다. 따라서 변량공간간 상관계수는 항상 0보다 같거나 크고 1보다 같거나 작으며, 이에 따라 변량공간간 거리는 항상 0보다 같거나 크고 $\pi/2$보다 같거나 작게 된다.

이들 정의와 그 관계에 관한 바탕이 되었던 관련 기하를 제시하면 다음과 같다. 우선 p개 변수 $x_1, x_2, \cdots, x_p$의 상관행렬을 R이라 하고, 이에 대한 아래와 같은 삼각분해(triangular decomposition, Bartlett's decomposition)를 생각해 보자.

$R = T'T$, 여기서 행렬 $T = (t_{ij}) = (\underline{t}_1\ \underline{t}_2 \cdots \underline{t}_p)$는 크기가 p인 상삼각행렬 (upper-triangular matrix)로서 대각원소들이 모두 양의 값을 가진 것이다

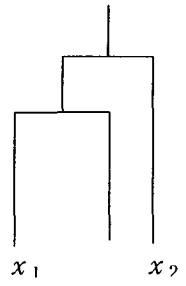
이광진(1991)에서는 $\underline{t}_i$를 '변수 x_i의 구조벡터'라 정의하고, 이에 의해 생성(span/generate)된 벡터공간(vector space)을 $V\underline{t}_i$로 표현하였다. 그리고 이를 '구조벡터 $\underline{t}_i$에 의해 생성되는 구조공간' 또는 '변수 x_i에 의해 생성되는 구조공간'이라 명명하였다. 그리고 이들 개념을 확장하여 a개의 변수들 $x_{g_1}, x_{g_2}, \cdots, x_{g_a}$이 주어진 경우 이들의 구조벡터들인 $\underline{t}_{g_1}, \underline{t}_{g_2}, \cdots, \underline{t}_{g_a}$에 의해 생성되는 벡터공간을 $V\underline{t}_{g_1}\underline{t}_{g_2}\cdots\underline{t}_{g_a}$로 표현하고, 이를 '구조벡터들 $\underline{t}_{g_1}\underline{t}_{g_2}\cdots\underline{t}_{g_a}$에 의해 생성되는 구조공간' 또는 '변수집합 $x_{g_1}x_{g_2}\cdots x_{g_a}$에 의해 생성되는 구조공간' 이라 정의하였다. 이를 이용하여 변수들 사이의 단순, 다중, 정준계수를 기하적 관점에서 구조공간들 사이의 최소 사이각으로 설명될 수 있음을 보였다. 즉, 두 변수 x_i, x_j의 단순상관계수는 이들 변수의 구조벡터들인 $\underline{t}_i$, $\underline{t}_j$의 사이각의 코사인 값과 같음을 보였다. 그리고 a개 변수들의 집합 $x_{g_1}x_{g_2}\cdots x_{g_a}$과 이에 포함되지 않은 다른 한 변수 x_i와의 다중상관계수는 이들 각각에 대응되는 구조공간 $V\underline{t}_{g_1}\underline{t}_{g_2}\cdots\underline{t}_{g_a}$와 구조벡터 $\underline{t}_i$사이의 최소 사이각에 대한 코사인 값과 같음을 보였다. 이를 확장하여 중복하지 않는 두 변수집합 $x_{g_1}x_{g_2}\cdots x_{g_a}$과 $x_{k_1}x_{k_2}\cdots x_{k_b}$의 제1정준상관계수는 이들에 대응되는 구조공간들인 $V\underline{t}_{g_1}\underline{t}_{g_2}\cdots\underline{t}_{g_a}$과 $V\underline{t}_{k_1}\underline{t}_{k_2}\cdots\underline{t}_{k_b}$ 사이의 최소 사이각의 코사인 값과 같음을 보였다.

4. 제안되는 병합/계층적 변수군집법의 알고리즘

제안되는 병합/계층적 변수군집법은 병합/계층적 개체군집법과 거의 유사한 군집화 과정을 거친다. 단지 차이가 있다면 1) 병합/계층적 개체군집법은 비유사성 행렬을 바탕으로 군집화가 이루어지지만 제안되는 병합/계층적 변수군집법은 유사성 행렬에 해당되는 상관행렬을 바탕으로 군집화가 이루어진다는 점, 2) 이에 따라 개체군집법에서 군집간의 거리 정의에서 사용되는 최단연결법, 최장연결법, 중심연결법, 중위수연결법, 평균연결법, Ward의 방법 등과 같은 방식을 사용하지 못하고 그 대신 변량공간들 사이의 상관계수를 사용한다는 점이다. 물론 변량공간의 특성변량을 무엇으로 정의하느냐에 따라 변량공간들 사이의 상관계수 값은 달라진다.

우선 군집화 대상이 되는 초기 p개 변수집합 $\{x_1, x_2, \cdots, x_p\}$의 상관행렬을 R_1이라 하고, 이들 변수들 각각이 생성하는 변량공간을 $V_1, V_2, \cdots, V_p$라 하자. 제안되는 병합/계층적 변수군집법은

p개의 변수들 각각이 자체 변량공간을 가지는 것으로 생각하고 출발하여 매 단계마다 가장 가까운(즉, 상관계수가 가장 큰) 두 변량공간을 병합하는 과정을 거쳐 최종적으로는 하나의 p차원 변량공간으로 병합하게 된다. 첫 병합 단계를 시작하기 전 사용자는 변량공간의 특성변량을 무엇으로 정의할 것인가를 미리 정해야 한다.



[그림 2]

변량공간의 특성변량을 정함에 있어서 사용자는 먼저 포괄적 입장에서 특성변량을 정할 것인가 아니면 계층적 입장에서 특성변량을 정할 것인가를 먼저 결정하여야 한다. 그 차이를 [그림2]의 간략한 덴드로그램으로 설명하면 다음과 같다. x_1의 변량공간 V_1과 x_2의 변량공간 V_2가 먼저 병합되어 변량공간 V_{12}가 생성되고, 이는 다시 x_3의 변량공간 V_3과 병합하여 변량공간 V_{123}이 생성된 경우이다. 이 때 V_{123}의 특성변량을 정할 때 포괄적 입장에서 정한다는 의미는 변량공간 V_{123}이 결국은 세 변수 x_1, x_2, x_3에 의해서 생성된 것이기 때문에 이들을 대등하게 본다는 것을 의미하며, 계층적 입장에서 정한다는 것은 V_{123}이 생성되기 전에 V_{12}가 먼저 생성되었다는 사실을 존중하여 V_{12}의 특성변량인 y_{12}와 V_3의 특성변량인 x_3만으로 V_{123}의 특성변량 y_{123}을 정해야 한다는 입장이다. 어느 입장을 취할 것인가는 사용자가 결정할 문제이다.

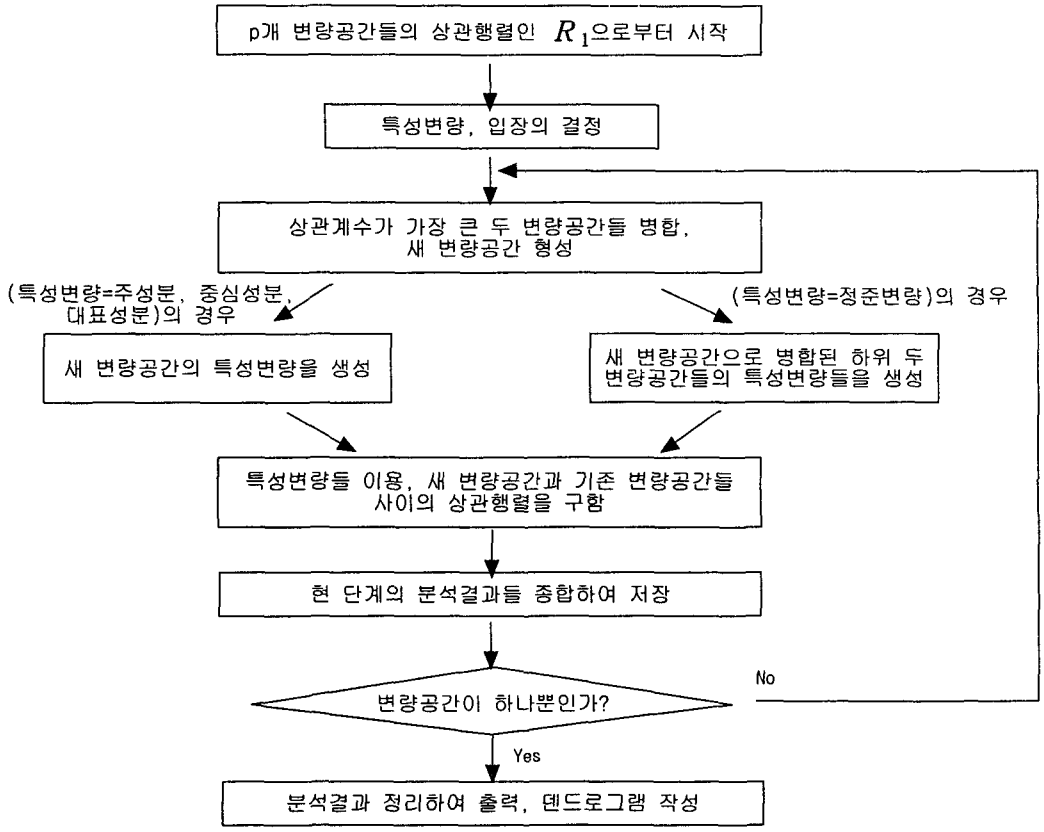
본 연구에서 제안되는 병합/계층적 변수군집법은 특성변량을 정할 때 취하는 입장과 사용되는 성분에 따라 다음의 [표1]과 같이 모두 6가지의 방법이 존재하게 된다. 참고로 계층적 입장에서는 주성분, 중심성분, 대표성분 어느 것을 특성변량으로 사용하든 동일한 결과를 준다는 것은 쉽게 이해될 수 있다. 왜냐하면 두 변량만 주어진 경우 이들의 주성분, 중심성분, 대표성분은 모두 동일한 1차원 변량공간을 형성하게 되기 때문이다.

[표 1] 특성변량의 설정에 따른 병합/계층적 변수군집법의 6가지 방법 구분

특성변량 / 입장	주성분	중심성분	대표성분	정준변량
포괄적 입장	방법1	방법2	방법3	방법4
계층적 입장	방법5			방법6

실제 군집화 과정에서는 각 단계마다 크게 두 가지의 절차는 반드시 거치게 된다. 첫 번째 절차는 새로운 변량공간을 형성할 두 개의 변량공간을 찾아 병합하는 것이고, 두 번째 절차는 기존 변량공간과 새로이 형성된 변량공간 사이의 상관계수를 계산하는 것이다. 이러한 단계와 절차들은 병합/계층적 개체군집법에서도 거의 동일하게 나타난다고 볼 수 있다.

제안되는 병합/계층적 변수군집법의 알고리즘을 그림으로 정리하면 [그림3]과 같다.



[그림 3] 제안되는 병합/계층적 변수군집법 알고리즘

5. 실제 예의 분석결과를 통한 분석과정의 자세한 설명

본 연구에서 제안하는 병합/계층적 변수군집법의 알고리즘을 실제 예의 분석결과를 통해 좀 더 자세히 설명하기 위해 Harman(1976)에 주어진 8개 신체변수들에 대한 상관행렬인 다음의 자료를 이용하였다. 이 상관행렬은 SAS의 분할/계층적 변수군집법 처리절차인 VARCLUS 절차의 사용설명서에서 예제로 사용된 것으로 비교를 위해 본 연구에서도 그대로 이용하고자 한다.

1.0	0.846	0.805	0.859	0.473	0.398	0.301	0.382	x_1='Height'
0.846	1.0	0.881	0.826	0.376	0.326	0.277	0.415	x_2='Arm Span'
0.805	0.881	1.0	0.801	0.380	0.319	0.237	0.345	x_3='Length of Forearm'
0.859	0.826	0.801	1.0	0.436	0.329	0.327	0.365	x_4='Length of Lower Leg'
0.473	0.376	0.380	0.436	1.0	0.762	0.730	0.629	x_5='Weight'
0.398	0.326	0.319	0.329	0.762	1.0	0.583	0.577	x_6='Bitrochanteric Diameter'
0.301	0.277	0.237	0.327	0.730	0.583	1.0	0.539	x_7='Chest Girth'
0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.0	x_8='Chest Width'

* 출처 : SAS Institute, 1993

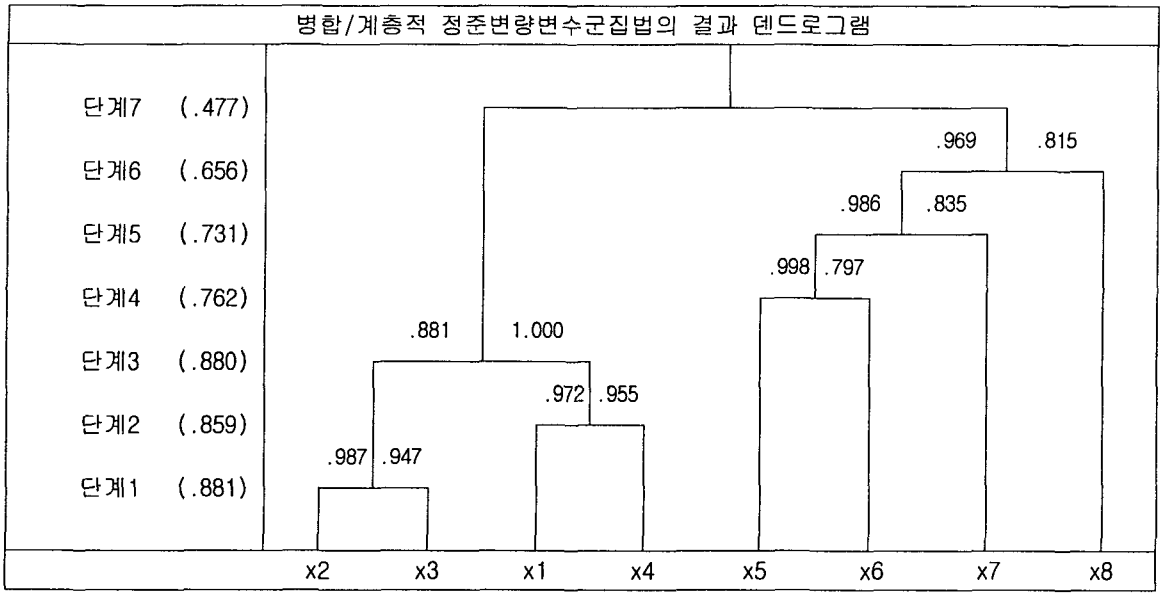
본 자료에 대한 분석에서는 변량공간의 특성변량을 계층적 입장에서 정준변량으로 정한 경우인 <방법6>을 사용하였으며, 분석결과의 일부만을 제시하면 [표1]과 [그림4]와 같다. 참고로, 본 방법에 의한 분석결과 덴드로그램([그림4])과 SAS의 VARCLUS 절차에 의한 분석결과 덴드로그램(SAS Institute, 1993 참조)의 구조는 완전히 일치함을 미리 밝힌다. 아래의 분석결과들을 보기 전에 혼동을 피하기 위해 다음을 인식해 둘 필요가 있다. 제안되는 병합/계층적 변수군집법은 변수의 병합이 아니라 변량공간의 병합을 전제하고 있기 때문에 변수를 표시하는 x라는 기호, 변량공간을 표시하는 V, 특성변량을 표시하는 y라는 기호를 잘 구별할 필요가 있다.

[표 2] 예제 자료에 대한 병합/계층적 정준변량변수군집법 분석결과

Eigenvalues of the Correlation Matrix

	1	2	3	4	5	6	7	8
Eigenvalue	4.673	1.771	0.481	0.421	0.233	0.187	0.137	0.096
Difference	2.902	1.290	0.060	0.188	0.047	0.049	0.041	
Proportion	0.584	0.221	0.060	0.053	0.029	0.023	0.017	0.012
Cumulative	0.584	0.806	0.866	0.918	0.947	0.971	0.988	1.000

Stage	Spaces Joined		New Space	Dimension of New Space	Correlation between Spaces Joined	Standardized Canonical Coefficients		Number of Spaces
1	V2	V3	V23	2	0.881	0.682	0.346	7
2	V1	V4	V14	2	0.859	0.579	0.458	6
3	V23	V14	V2314	4	0.880	0.005	0.995	5
4	V5	V6	V56	2	0.762	0.932	0.087	4
5	V2314	V7	V567	3	0.731	0.807	0.245	3
6	V567	V8	V5678	4	0.656	0.759	0.325	2
7	V2314	V5678	V23145678	8	0.477			1



[그림 4] 예제 자료에 대한 분석결과 덴드로그램

상기 [표2]와 [그림4]의 분석결과는 저자가 SAS를 이용하여 직접 작성한 프로그램으로 분석한 결과의 일부를 재정리한 것인데, 이러한 결과가 얻어지게 되는 과정을 단계적으로 설명하면 다음과 같다. 단계1에서는 원 상관행렬 R_1 중에서 가장 큰 상관계수값 0.881을 가지는 변량공간 V_2와 V_3이 먼저 병합되고 병합된 새로운 변량공간을 V_{23}이라 한다. 단계1을 마친 후 $V_1, V_4, V_5, V_6, V_7, V_8, V_{23}$들의 상관행렬 R_2를 계산한다. 이 때 V_5와 V_6의 상관계수 같은 것은 R_1의 것을 그대로 사용하면 되지만, V_7과 V_{23}과 같은 상관계수는 R_1을 이용하여 구한 $\{x_2, x_3\}$과 $\{x_7\}$의 정준상관계수 값을 사용한다. 이 때 유의할 점은 $\{x_2, x_3\}$의 정준변량 y_{23}은 다른 변수와의 정준상관계수를 구할 때마다 달라진다는 점이다. 이러한 이유 때문에 정준변량을 특성변량으로 사용하는 정준변량 변수군집법(방법4와 방법6)에서는 한 변량공간의 특성변량을 그 변량공간이 다른 변량공간과 처음으로 병합될 때의 것으로 정의하게 된다. 따라서 단계1에서는 V_{23}이 아직 다른 변량공간과 병합하지 않았기 때문에 V_{23}의 특성변량 y_{23}은 아직 결정되어지지 않는 것이다.

단계2에서는 R_2로부터 가장 큰 상관계수값 0.859를 가지는 변량공간 V_1과 V_4를 병합한다. 단계1을 마친 후와 마찬가지로 단계2를 마친 후에도 $V_5, V_6, V_7, V_8, V_{23}, V_{14}$들의 상관행렬 R_3을 계산한다. 이 때에도 V_5와 V_6, V_7과 V_{23}의 상관계수 같은 것은 단계1에서처럼 구하면 되지만, V_{23}과 V_{14}의 상관계수는 $\{x_2, x_3\}$과 $\{x_1, x_4\}$의 정준상관계수 값을 사용한다. 이 때에도 V_{23}과 V_{14}가 아직 다른 변량공간과 병합하지 않았기 때문에 V_{23}, V_{14}의 특성변량들 y_{23}, y_{14}가 아직도 결정되어지지 못하게 된다.

단계3에서는 단계2에서 얻어진 R_3으로부터 가장 큰 상관계수값 0.880을 가지는 변량공간 V_{23}과 V_{14}가 병합되어 V_{2314}가 된다. 이 때 V_{23}과 V_{14}가 처음으로 다른 변량공간과 병합되기 때문에 이제는 V_{23}과 V_{14}의 특성변량 y_{23}과 y_{14}가 결정되게 된다. 이는 $\{x_2, x_3\}$과 $\{x_1, x_4\}$의 제1정준변량들로 $y_{23}=0.682x_2+0.346x_3$, $y_{14}=0.579x_1+0.458x_4$로 표현되어진다. 단계3을 마치면서 $V_5, V_6, V_7, V_8, V_{2314}$의 상관행렬 R_4를 얻기 위해서는 $\{y_{23}, y_{14}\}$와 $\{x_5\}$ $\{y_{23}, y_{14}\}$와 $\{x_6\}$, $\{y_{23}, y_{14}\}$와 $\{x_7\}$, $\{y_{23}, y_{14}\}$와 $\{x_8\}$의 정준상관계수를 각각 구할 필요가 있을 것이다.

단계4에서는 R_4로부터 가장 큰 상관계수값 0.762를 가지는 V_5와 V_6이 병합되어 변량공간 V_{56}이 생성되지만, 이 때에도 V_{56}이 다른 변량공간과 병합된 것이 아니기 때문에 V_{56}의 특성변량 y_{56}은 아직 결정되지는 못한다. R_5는 $V_7, V_8, V_{2314}, V_{56}$의 상관계수들이 필요한데 이를 위해 $\{x_5, x_6\}$과 $\{x_7\}$, $\{x_5, x_6\}$과 $\{x_8\}$, $\{x_5, x_6\}$과 $\{y_{23}, y_{14}\}$의 정준상관계수를 각각 구해야 한다.

단계5에서는 R_5로부터 가장 큰 상관계수값 0.731을 가지는 V_{56}과 V_7이 병합되어 V_{567}이 생성된다. 이 때 V_{56}이 다른 변량공간과 처음으로 병합하기 때문에 V_{56}의 특성변량 y_{56}이 결정되게 되는데 이는 $\{x_5, x_6\}$과 $\{x_7\}$의 정준상관분석 과정에서 쉽게 얻어진다. 이는 $y_{56}=0.932x_5+0.087x_6$으로 표현된다. 단계5를 마치면서 V_8, V_{2314}, V_{567}의 상관행렬 R_6을 얻기 위해서는 $\{y_{56}, x_7\}$와 $\{x_8\}$, $\{y_{56}, x_7\}$와 $\{y_{23}, y_{14}\}$의 정준상관계수를 구해야 할 것이다.

단계6에서는 R_6으로부터 가장 큰 상관계수값 0.656을 가지는 V_{567}과 V_8이 병합하여 V_{5678}이 생성되면서, V_{567}이 다른 변량공간과 처음으로 병합하기 때문에 V_{567}의 특성변량 y_{567}이 구해지게 된다. 이는 $y_{567}=0.830\,y_{56}+0.217\,x_7$이 된다. V_{2314}, V_{5678}의 상관행렬 R_7를 계산할 때에는 $\{y_{23}, y_{14}\}$와 $\{y_{567}, x_8\}$의 정준상관계수가 필요하게 되는데, 분석결과에 의하면 그 값은 0.477이 된다.

단계7에서는 V_{2314}와 V_{5678}이 병합하여 하나의 변수군집이 생성되고, 이 때 V_{2314}와 V_{5678}이 처음으로 타 변수군집과 병합하기 때문에 V_{2314}와 V_{5678}의 특성변량 y_{2314}와 y_{5678}이 결정된다. 이는 $\{y_{23}, y_{14}\}$와 $\{y_{567}, x_8\}$의 정준상관분석을 통해 얻어지게 되는데, 이는 $y_{2314}=0.005\,y_{23}+0.995\,y_{14}$, $y_{5678}=0.759\,y_{567}+0.325\,x_8$로 표현되어진다.

덴드로그램에서 단계별 수평선은 변량공간으로 이해하면 된다. 예를 들어, 단계1에서의 수평선은 V_{23}을, 단계5에서의 수평선은 V_{567}을 나타내는 것이다. 그리고 수직선들은 모두 수직선 아래의 변량공간에 대한 특성변량들에 해당되는데, 예를 들어 x_2에서 위로 뻗은 수직선은 x_2의 변량공간 V_2의 특성변량인 x_2를 나타내며, V_2와 V_3이 병합한 부분에서 위로 뻗은 수직선은 V_{23}의 특성변량 y_{23}이 된다.

본 정준변량군집법에서 얻어지는 특성변량은 전통적인 정준변량과 약간의 차이가 있다. 이는 계층적인 입장에서의 특성변량을 정의하였기 때문으로, 예를 들어, V_{2314}와 V_{5678}의 특성변량 y_{2314}와 y_{5678}은 $\{x_1, x_2, x_3, x_4\}$와 $\{x_5, x_6, x_7, x_8\}$의 제1정준변량들이 아니라 $\{y_{23}, y_{14}\}$와 $\{y_{567}, x_8\}$의 제1정준변량들이라는 점이다. 물론 y_{567}은 $\{y_{56}, x_7\}$과 $\{x_8\}$의 제1정준변량이 된다. 즉, 본 연구에서의 정준변량은 계층구조를 고려한 축차적 의미를 지닌 정준변량라고 할 수 있다.

덴드로그램에서 각 단계별 수평선 위에 적혀 있는 수치들은 병합되는 변량공간의 특성변량과 병합 후 생성된 변량공간의 특성변량과의 상관계수로서, 예를 들어 단계5의 0.986, 0.835라는 값들은 각각 V_{567}의 특성변량 y_{567}과 V_{56}의 특성변량 y_{56}의 상관계수, V_{567}의 특성변량 y_{567}과 x_7의 상관계수가 된다.

6. 결론

본 논문에서는 변량공간과 변량공간의 특성변량의 개념을 도입하여 상관행렬을 이용한 병합/계층적 변수군집법을 제안하고, 예제를 통해 그 방법의 사용법과 알고리즘을 상술하였다. 이는 SAS의 VARCLUS 절차인 분할/계층적 변수군집법과는 병합적 입장이냐 분할적 입장이냐의 차이가 있는 것으로 서로가 장단점을 가진다고 할 수 있겠다. 저자의 자료분석 경험에 의하면 변수군집의 수에 관한 정보가 없는 상태에서 초기 분석변수의 수가 많은 경우는 분할/계층적 변수군집법이, 수가 적은 경우에서는 제안되는 병합/계층적 변수군집법이 좀 더 유효한 것으로 판단되지만, 변수들 사이의 군집현상과 군집의 속성을 조금이라도 더 자세히 알고 싶다면 두 방법을 모두 사용해 보는 것이 좋겠다.

제안된 병합/계층적 변수군집법은 특성변량의 정의에 따라 6가지의 방법이 있음을 제시하였지만, 주성분, 중심성분, 대표성분, 정준변량 이외의 것으로 좋은 성질을 가지는 성분이 있어 이를 특성변량으로 정의할 수 있다면 더 많은 방법들이 존재할 수도 있다. 이 부분에 관한 추가연구가 필요하리라고 본다.

그리고 지면의 제약에 의해 6가지 방법 중 알고리즘이 가장 까다롭다고 볼 수 있는 계층적 입장에서 특성변량을 정준변량으로 정하는 방법(방법6)의 사용 예와 이의 자세한 알고리즘만을 제시하였지만, 나머지 방법들의 알고리즘도 이와 거의 비슷하기 때문에 이에 관해서는 독자의 몫으로 남긴다. 그리고 분석결과로서 본문에서 제시된 정보 이외의 다른 어떤 유용한 정보들이 있을 수 있는지에 관한 연구도 후속 연구의 과제가 될 수 있겠다.

본문에서는 지적하지 않은 문제로서 병합의 매 단계마다 생성되는 변량공간들의 상관행렬에는 단순, 다중, 정준 상관계수들이 혼재할 수 있는데 이들의 직접 비교로 변량공간간의 가까움과 멀음을 논할 수 있는지와 이에 대한 대안도 연구 중에 있지만 후속 연구의 몫으로 남긴다.

참고문헌

[1] 강현철, 김윤규, 김기영 (2000). 군집성분을 이용한 변수군집분석, 「한국분류학회지」, 제4권, 25-33.

[2] 이광진 (1991). 상관구조의 기하적 고찰과 고유변환에 관한 연구, 박사학위논문, 고려대학교.

[3] 이광진 (1997). 대표성분점수화법의 제안과 이의 타당성 및 활용성에 관한 연구, 「응용통계연구」, 제10권 제2호, 275-291.

[4] Harman, H. H. (1976). *Modern Factor Analysis*, The University of Chicago Press.

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

[6] SAS Insititute (1993). *SAS/STAT User's Guide* (Vol. 1), Version 6 Fourth Edition. SAS Institute, NC:Cary.

[2003년 3월 접수, 2003년 7월 채택]