이 성 건 교수지도

석사학위 청구논문

연관성측도를 이용한 군집개수 결정에 관한 연구

2013

성신여자대학교 대학원

통 계 학 과

백 아 현

연관성측도를 이용한 군집개수 결정에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2013년 5월

성신여자대학교 대학원

통계학과

백 아 현

인 준 서

백아현의 석사학위 논문으로 인준함.

심사위원 송 일 섭 인

심사위원 이 소 정 인

심사위원 박 미 식 인

성신여자대학교 대학원

논 문 개 요

다변량 분석 중 군집분석(cluster analysis)은 주어진 관측개체를 유사한 개체끼리 몇 개의 군집으로 나눔으로써 군집을 이해하고 효율적으로 활용하기 위한 분석방법이다.

군집분석에서 군집의 적절한 개수를 추정하는 것은 중요한 문제이며 지금까지 많은 연구가 진행되어왔다(Calinski & Harabasz, 1974; Hartigan,1975; Krzanowski & Lai,1985). 군집의 개수를 결정하는 방법들의 대부분은 군집 내의 제곱거리와 군집 간의 제곱거리를 이용하는 것이다. 하지만 연구자의 주관적인 의견이 많이 반영되고 정확한 결과를 제공해주지 못한다는 문제점이 있다.

최근에는 군집 내, 군집 간의 거리개념을 넘어서 군집의 안정성에 관한 연구도 많이 진행되고 있다. 특히 군집화 불안정성을 최소화하는 군집개수를 최적의 군집개수로 결정하는 알고리즘들이 제안되었다(Wang, 2010; Fang & Wang, 2012).

본 논문에서는 군집화 거리에서 두 군집화의 일치여부가 이항자료로 나타나는 특성을 이용하여 연관성측도를 적용함으로써 군집화 불안정성을 측정하는 새로운 군집개수 결정 알고리즘을 제안하였다.

모의실험과 실제데이터를 통해 제안한 방법의 효율성을 살펴본 결과, 군집개수가 작거나 차원이 낮은 자료에서 제안한 방법의 군집개수 선택의 적중률이 기존 Wang방법보다 높았다. 결과적으로 본 연구에서 제안한 방법이 다양한 자료에서 군집개수 선택에 있어서 기존 방법보다 더 우수함을 확인할 수 있었다.

목 차

제 1 장 서 론

다변량 분석(multivariate analysis)은 서로 연관된 두 개 이상의 변수들에 대하여 구조적으로 복잡한 관계를 요약하거나 관찰개체들을 분류, 다른 개체들 간의 연관관계를 파악하는 분석이다. 다변량 분석 중에서도 군집분석(cluster analysis)은 주어진 관측개체를 몇 개의 군집으로 나눔으로써 군집을 이해하고 효율적으로 활용하기 위한 분석방법이다.

군집분석은 군집의 개수와 내용, 그리고 구조 등이 알려지지 않은 상태에서 개체들의 특성을 파악하는 것으로 군집들에 대한 사전적인 정보를 가지고 분석하지 않는다. 따라서 군집의 적절한 개수를 추정하는 것은 군집분석에서 중요한 문제이며 지금까지 많은 연구가 진행되어왔다. 군집의 개수를 결정하는 방법들의 대부분은 군집 내의 제곱거리와 군집 간의 제곱거리를 이용하는 것이었다(Calinski & Harabasz(1974); Hartigan(1975); Krzanowski & Lai(1985)). 이런 방법에 있어서는 그 기준값과 대응되는 군집의 개수를 플롯화하여 급격한 증가 혹은 감소가 일어나는 곳에서 군집개수를 결정한다. 이런 증감을 통한 군집개수의 결정은 가장 일반적으로 쓰이는 방법이기는 하지만 연구자의 주관적인 의견이 많이 반영되고 정확한 결과를 제공해주지 못한다는 문제점이 있다.

모형기반의 군집분석을 통해서도 군집개수를 결정할 수 있는데 이때, 각 군집개수 k에서의 우도(likelihood)를 이용하여 최대 우도값을 구한 후 이를 모형 선택의 기준으로 사용한다. 이 경우 AIC(Akaike information criterion)와 BIC(Bayesian information criterion)을 이용한다.

Tibshirani 등(2001)은 기존의 방법에서 더 나아가 적절한 귀무 분포 하에서 군집 내 산포의 변화를 이용한 gap 통계량을 제안하였다. 또한, Sugar & James(2003)는 jump 통계량을 통해 군집의 개수를 결정하는 알고리즘을 제안

하였다. 또한 최근에는 군집 내, 군집 간의 거리개념을 넘어서 군집화 안정성에 관한 연구도 많이 진행되고 있다. Kriegar & Green(1999)은 표본크기의 다양성을 고려하여 k-평균 군집분석방법에의 군집화 안정성 척도를 제안하였다. 또한, Steinley(2008)는 k-평균 군집분석방법을 반복적으로 시행하여 군집화(clustering) 안정성을 기반으로 군집의 개수를 결정하였다. 여기서 군집화란 군집을 결정짓는 일련의 과정으로 서로 다른 군집화 함수에 같은 자료를 적용했을 때에 유사한 군집결과가 나올 때 군집화가 안정적이라고 할 수 있다. 특히, Wang(2010)은 군집화 안정성을 최대화하는 것은 군집화 불안정성을 최소화하는 것이므로 교차타당성(cross-validation)방법을 통해 두 군집화의 거리를 계산하여 군집화 불안정성을 최소로 하는 군집개수를 최적의 군집개수로 결정하였다. Fang & Wang(2012)은 더 나아가 붓스트랩을 이용하여 군집화 불안정성을 측정하였다.

본 논문에서는 거리를 이용하여 군집화 불안정성을 측정하는 기존 Wang의 방법을 보완하여 자카드 계수(Jaccard coefficient), 카파 계수(kappa coefficient), 파이 계수(phi coefficient) 등의 연관성측도를 이용하여 군집화 불안정성을 측정하였다. 그리고 제안한 방법의 효율성을 살펴보기 위해서 모의실험을 통해 군집의 개수를 정확하게 추정하는지 비교, 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 군집분석방법과 군집화 불안정성에 기반을 둔 군집개수를 측정하는 방법에 대해 설명하고, 3장에서는 본 연구에서 제안한 연관성측도를 이용한 새로운 군집개수를 결정하는 방법을 제안한다. 4장에서는 기존의 방법과 제안한 방법을 모의실험을 통해 비교 분석하였다. 5장에서는 실제자료에 방법을 적용하여 제안한 방법의 우수성을 살펴보았다. 마지막으로 6장에서는 본 연구의 한계점 및 추후 연구방향에 대해 논의한다.

제 2 장 이론적 배경

2.1. 군집분석 방법론

군집(cluster)이란 군집 내적으로는 동질하고 다른 군집과는 이질적인 개체들의 모임을 말한다. 군집분석(cluster analysis)의 목적은 주어진 관측개체를 몇 개의 군집으로 나눔으로써 군집을 이해하고 효율적으로 활용하고자 하는 것이다. 군집분석은 군집들에 대한 사전적인 정보를 가지고 분석하지 않기 때문에 대상들의 집단구분이 이루어져 있는 상황에서 집단구분의 유의한 변수를 선정하는 판별분석(discriminant analysis)과는 다르다. 군집분석은 같은 군집에 속하는 개체들끼리는 유사성이, 다른 군집에 속하는 개체들 사이에는 비유사성이 존재하는 것을 원칙으로 한다. 따라서, 개체들 간의 유사성 또는 비유사성에 근거하여 개체를 식별함으로써 전체 다변량 자료의 구조를 파악하고 군집 간의 관계 등을 분석하는 과정의 총체가 군집분석의 목적이다.

군집분석에서는 개체들을 군집화하기 위해서 각 개체들이 얼마나 유사한가를 나타내는 척도들이 필요하다. 기본적으로 관측개체 간의 거리를 어떻게 정의하느냐에 따라 달라지는데 개체간의 거리는 통상적으로 아래와 같은 방법으로 정의된다. 먼저 개체 $\boldsymbol{x}=(x_1,...,x_p)$, $\boldsymbol{y}=(y_1,...,y_p)$ 사이의 유클리드 거리(Euclidean distance)는 아래와 같은 식(2.1)으로 표현한다.

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2 \cdots+(x_p-y_p)^2} \quad . \tag{2.1}$$

다음으로 식(2.2)는 유클리드 거리를 일반화한 민코브스키(Minkowski)거리이다(Kruskal, 1964).

$$d(\boldsymbol{x},\boldsymbol{y}) = \left[(x_1-y_1)^m+(x_2-y_2)^m \cdots+(x_p-y_p)^m\right]^{1/m} \quad . \tag{2.2}$$

민코브스키거리에서 m=2 일 때, 유클리드 거리와 같다. 유클리드거리나 민코브스키거리는 척도불변성을 가지지 않아서 척도에 따라 그 값이 크게 왜곡되므로 개체간의 거리를 측정할 시에 표준화한 개체값을 고려한다.

2.1.1. 계층적 군집화(hierarchical clustering)

계층적 군집화는 계층적으로 각 개체들을 군집화하는 방법으로 각 개체들을 병합하여 군집화하는 방법과 모든 개체를 같은 군집으로 보고 개체들을 분할하면서 군집화하는 방법이 있다. 먼저, 병합적 계층적 군집화는 각 개체가 스스로 군집인 상태에서 시작한다. 그 다음 단계에서 가장 유사한, 즉 거리가 가까운 두 개체를 군집화 함으로써 군집수를 하나 줄인다. 유사한 군집을 합하는 과정을 모든 개체가 한 군집이 될 때까지 계속한다. 분할적 계층적 군집화는 응집하는 계층적 군집화 과정과는 다르게 모든 개체가 한 군집인 상태에서 시작한다. 군집과의 거리가 가장 먼 개체들을 분할하여 군집수를 늘려간다. 최종적으로 각 개체가 군집인 상태가 된다. 계층적 군집분석 방법을 이용할 때에는 군집의 개수보다는 개체들의 계층적 구조에 관심이 있으며 개체들의 계층적 구조를 나무형 그림(dendrogram)으로 표현하여 그 구조를 쉽게 파악할 수 있다. 계층적 군집화 방법에는 최단연결법(single linkage method), 최장연결법(complete linkage method), 중심연결법(centroid linkage method), 그리고 Ward방법 등이 있다.

2.1.2. 비계층적 군집화(non-hierarchical clustering)

비계층적 군집화는 계층적으로 군집을 형성하지 않고 개체들을 몇 개의 군집으로 구분시키는 형태이다. 한번 군집이 형성되면 변경되지 않는다는 계층적 군집화 방법에서의 단점을 극복한다. 비계층적 군집화의 가장 대표적인 방법은 k-평균 군집분석이며 알고리즘은 아래의 [표 1]과 같다.

[표 1] k-평균 군집분석 알고리즘

1. k개의 각 군집에 1개의 초기값(군집의 중심값)을 설정한다. 2. 모든 개체를 단계1에서 설정한 값 중 가장 가까운 중심값을 찾아 그 군집에 할당한다. 3, 각 군집에 할당된 개체를 통하여 k개의 새로운 중심값을 계산한다. 4. 변화가 없을 때 까지 단계2-단계3을 반복한다.

비계층적 군집화는 초기값에 의해 결과가 큰 영향을 받으며 군집의 개수인 k가 미리 정해져 있어야 한다.

2.2. 군집개수의 결정

군집의 개수 k를 결정하는 문제는 인자분석에서의 공통인자의 개수를 결정하거나, 주성분분석에서 주성분의 개수를 정하는 것만큼 어려운 문제이며 계층적 군집화를 이용하여 군집개수를 정하는 방법 또한 명확하지 않다.

2.2.1. 팔꿈치 방법

군집의 개수를 결정하는 가장 기본적인 방법은 팔꿈치 방법(elbow method)으로 주성분의 개수를 구하는데 있어서도 널리 쓰이는 방법 중 하나이다. 각 군집개수에 따른 특정 통계량을 계산하여 통계량 값이 급격하게 꺾이는 지점에서 군집의 개수를 결정한다. 사용하는 통계량은 아래와 같다.

첫째로 군집 k에서 제곱근평균제곱표준편차(RMSSTD, root mean square standard deviation)는 아래의 식(2.3)과 같이 정의한다.

$$RMSSTD_k = \sqrt{\frac{W_k}{\upsilon(N_k - 1)}} \tag{2.3}$$

여기서, $W_k = \sum_{i \in C_k} \| \boldsymbol{x_i} - \overline{\boldsymbol{x_k}} \|$ 이며, N_k는 군집 k에 속해 있는 개체의 수이고 υ은 차원 수이다. 따라서 제곱근평균제곱표준편차가 가장 작은 값을 가질 때의 k를 군집개수로 정한다.

다음으로 주어진 군집 k에서 총결정계수(R^2, over-all R-squared)는 아래식(2.4)과 같다.

$$R^2 = 1 - \frac{P_G}{T} \quad , \tag{2.4}$$

여기서, $T = \sum_{i=1}^{n} \| \boldsymbol{x_i} - \overline{\boldsymbol{x}} \|$ 이고, $P_G = \sum_{j=1}^{G} W_j = \sum_{j=1}^{G} \sum_{i \in C_j} \| \boldsymbol{x_i} - \overline{\boldsymbol{x_j}} \|$ 이다. 이때, G는

주어진 계층 수준에서의 군집의 개수이며 군집의 개수가 커질수록 총결정계수는 증가한다.

다음은 부분결정계수(squared semipartial R^2)이다. 부분결정계수는 K번째 군집 C_K와 L번째 군집 C_L의 합병으로 인한 총결정계수의 감소분으로 아래 식(2.5)와 같이 나타낼 수 있다.

$$\text{부분 } R^2 = \frac{B_{KL}}{T} \quad , \tag{2.5}$$

여기서, $B_{KL} = W_M - W_K - W_L$이며 이때, $C_M = C_K \cup C_L$이다. 따라서 부분 결정계수가 급격히 증가하기 직전에 군집화를 멈추는 것이 좋다.

다음은 의사-F값(pseudo-F)으로 아래의 식(2.6)과 같이 정의한다.

$$\text{의사}-F = \frac{\frac{T-P_G}{G-1}}{\frac{P_G}{n-G}} \quad , \tag{2.6}$$

여기서, $P_G = \sum_{j=1}^{G} W_j = \sum_{j=1}^{G} \sum_{i \in C_j} \| \boldsymbol{x_i} - \overline{\boldsymbol{x_j}} \|$ 이고, G는 주어진 계층 수준에서의 군집의 개수이며. n은 개체의 수이다. 의사-F값이 가장 큰 값을 가질 때의 k를 군집개수로 정한다. 또한 군집 C_K와 C_L의 의사-t^2값(pseudo-t^2)은 아래의 식(2.7)과 같다.

$$\text{의사}-t^2 = \frac{B_{KL}}{\frac{W_K + W_L}{N_K + N_L - 2}} \quad , \tag{2.7}$$

여기서, $B_{KL} = W_M - W_K - W_L$이며 이때, $C_M = C_K \cup C_L$이다. 의사-t^2이 크다는 것은 두 군집 C_K와 C_L이 합병될 수 없다는 것을 의미한다. 따라서, 급격하게

값이 커질 때의 k를 군집개수로 정한다.

다음으로 삼차군집기준(CCC, cubic clustering criterion)은 군집의 편차를 비교 측정하는 도구로 아래의 식(2.8)과 같이 정의한다.

$$CCC = \ln\left[\frac{1-E(R^2)}{1-R^2}\right] \times K \ , \tag{2.8}$$

이때, R^2은 총결정계수를 의미하고 $E(R^2)$은 R^2의 기댓값, K는 분산 안정화 변환값이다.

삼차군집기준이 양의 큰 값을 가질수록 군집 간에 큰 차이가 있다는 것을 의미하므로 그 값이 클 때, 최적의 군집개수이다. 그러나 군집화 변수들끼리의 상관관계가 클 때에는 삼차군집기준값이 부정확할 수 있다.

2.2.2. 모형 기반 군집분석(model-based cluster analysis)

군집개수를 결정시 모형 기반 군집분석(model-based cluster analysis)을 이용할 수 있다. 이때, 군집개수는 자료에 의해 결정되는 것이 특징이다.

모형 기반 군집분석은 확률모형을 기본으로 하며 가장 기본적인 모형은 유한혼합모형(finite mixture model)이다. 군집개수가 k일 때, 개체들의 기대비율은 p_k이고, 대응되는 측정값은 확률밀도함수 $f_k(\boldsymbol{x})$에 의해 생성된다. 군집개수가 K일 때 관찰값은 아래의 식(2.9)에 의해 모형화 되며 이를 $f_1(\boldsymbol{x}), \cdots, f_K(\boldsymbol{x})$의 혼합분포(mixing distribution)라 한다.

$$f_{Mix}(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}) \ , \tag{2.9}$$

이때, $p_k \geq 0$ 이며 $\sum_{k=1}^{K} p_k = 1$이다.

가장 일반적인 혼합분포로는 다변량 정규분포가 있으며 식은 아래 (2.10)과

같다.

$$f_{Mix}(\boldsymbol{x}|\boldsymbol{\mu_1},\boldsymbol{\Sigma_1},\cdots,\boldsymbol{\mu_K},\boldsymbol{\Sigma_K}) = \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma_k}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})'\boldsymbol{\Sigma_k^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_k})\right) . \quad (2.10)$$

모형 선택을 위해서 군집개수 K마다 모수의 최대우도를 계산하고 아래 식 (2.11)으로 최대우도값을 구하여 모형 선택을 위해 사용한다.

$$L_{\max} = L(\hat{p_1},\cdots,\hat{p_K},\hat{\boldsymbol{\mu_1}},\hat{\boldsymbol{\Sigma_1}},\cdots,\hat{\boldsymbol{\mu_2}},\hat{\boldsymbol{\Sigma_K}}) , \quad (2.11)$$

이때, 모형 선택은 AIC(Akalike information criterion)와 BIC(Bayesian information criterion)을 이용하며 식은 아래(2.12), (2.13)과 같다.

$$AIC = 2\ln L_{\max} - 2N\left(K\frac{1}{2}(p+1)(p+2)-1\right) . \quad (2.12)$$

$$BIC = 2\ln L_{\max} - 2\ln(N)\left(K\frac{1}{2}(p+1)(p+2)-1\right) . \quad (2.13)$$

혼합모형의 모수들을 추정할 시에는 EM(expectation-maximization) 알고리즘을 사용하며 최종적으로 가장 큰 AIC와 BIC를 갖는 군집개수가 최적의 군집개수가 된다.

2.2.3. 군집화 불안정성

군집화(clustering)는 군집을 결정짓는 일련의 과정으로 $\psi(\boldsymbol{x})$로 정의한다. 즉, $\psi(\boldsymbol{x})$은 자료를 군집 $\{1,\cdots,k\}$에 대응시켜 주는 함수이다. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$은 알려지지 않은 p차원의 분포 $F(x)$로부터의 독립표본이라 하자. 여기서, $\boldsymbol{X_n}$은 $n\times p$행렬이고, $\boldsymbol{x_1},\cdots,\boldsymbol{x_n}$은 $1\times p$벡터이다. 그리고 군집화 $\psi(\boldsymbol{x})$은 군집 $\{1,\cdots,k\}$을 출력하는 함수라고 할 때, 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$사이의 거리는 아래의

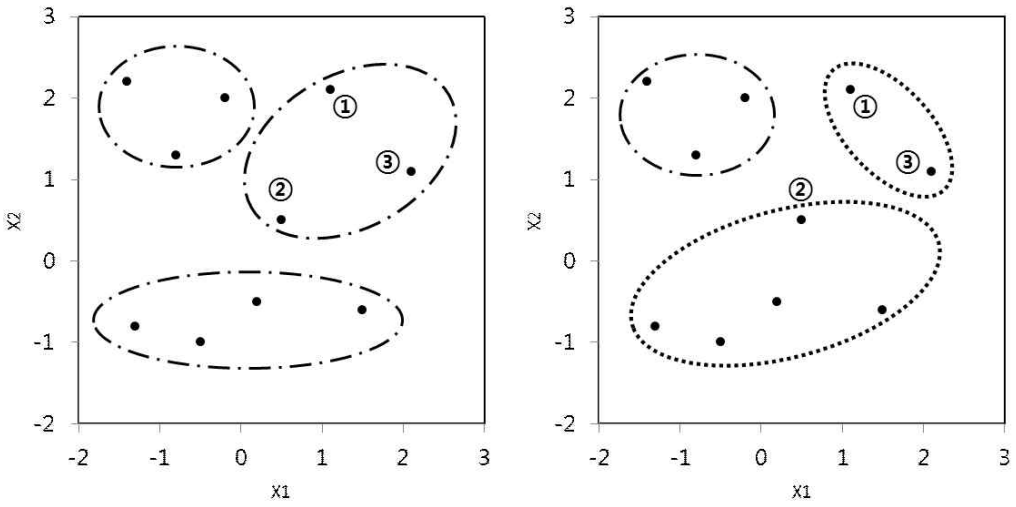
식(2.14)와 같이 정의한다(Ben-David, 2006).

$$d(\psi_1,\psi_2)=E\left[\,|\,I\{\psi_1(\boldsymbol{x}^0)=\psi_1(\boldsymbol{y}^0)\}-I\{\psi_2(\boldsymbol{x}^0)=\psi_2(\boldsymbol{y}^0)\}|\,\right]\quad , \qquad (2.14)$$

이때, $I(\cdot)$은 지시함수이고 $\boldsymbol{x}^0$와 $\boldsymbol{y}^0$은 $1\times p$벡터로서 분포 $F(x)$로부터의 독립적인 개체이다.

군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 거리는 개체 $\boldsymbol{x}^0$와 $\boldsymbol{y}^0$를 군집화 했을 때 같은 군집화 결과가 나오면 0, 아니면 1의 값으로 나타난다. 즉 각각의 개체에 대해 군집화 거리를 측정할 시에 군집화 $\psi_1(\boldsymbol{x})$에서 0 또는 1의 값을 갖고, 군집화 $\psi_2(\boldsymbol{x})$에서도 0 또는 1의 값을 갖는다. 한 쌍의 개체에 대해 두 군집화가 같은 결과를 나타내면 군집화 거리 $d(\psi_1,\psi_2)$는 0이 되고 두 군집화가 다른 결과를 나타내면 군집화 거리는 1이 된다.

아래 간단한 예시를 통해서 군집화와 군집화 거리 $d(\psi_1,\psi_2)$의 개념을 살펴볼 수 있다.



(a) 군집화 $\psi_1(\boldsymbol{x})$에 의한 결과 (b) 군집화 $\psi_2(\boldsymbol{x})$에 의한 결과

[그림 1] 군집화 $\psi_1(\boldsymbol{x})$, $\psi_2(\boldsymbol{x})$에 의한 결과

[그림 1]의 (a)은 군집화 $\psi_1(\boldsymbol{x})$에 의한 결과이며, (b)는 군집화 $\psi_2(\boldsymbol{x})$에 의한 결과이다. [그림 1]의 개체1(①), 개체2(②), 개체3(③)을 통하여 군집화 거리를 구하기 위한 식들을 구하면 아래와 같다.

$$|I\{\psi_1(①)=\psi_1(②)\}-I\{\psi_2(①)=\psi_2(②)\}|=|1-0|=1$$
$$|I\{\psi_1(①)=\psi_1(③)\}-I\{\psi_2(①)=\psi_2(③)\}|=|1-1|=0$$
$$|I\{\psi_1(②)=\psi_1(③)\}-I\{\psi_2(②)=\psi_2(③)\}|=|1-0|=1$$
$$\vdots$$

위와 같이 모든 개체끼리 마다 군집화 일치여부에 관한 식을 구하여 기댓값을 취하면 최종적인 군집화 거리가 된다.

결과적으로 모든 개체에 대해서 군집화 거리는 0과 1의 값만을 같게 되며 군집화 불안정성은 군집화 거리의 기댓값으로 나타낼 수 있다. 따라서 각 군집개수 k에 따른 군집화 불안정성은 아래 식(2.15)과 같이 나타낼 수 있다.

$$s(\Psi,k,n)=E\left[d_F(\Psi_{X_n,k},\Psi_{\widetilde{X}_n,k})\right] \ , \tag{2.15}$$

이때, X_n과 $\widetilde{X}_n$은 F로부터의 크기가 n인 독립적인 두 랜덤표본이다. 또한, $\Psi_{X_n,k}$와 $\Psi_{\widetilde{X}_n,k}$는 X_n과 $\widetilde{X}_n$으로부터 구성되어진 군집화 함수이다.

군집화 불안정성이 낮을 때의 k값이 최적의 군집개수 이므로 아래와 같이 최적의 군집개수를 정의한다.

$$k_0=k_0(n)=\operatorname{argmin} s(\Psi,k,n) \ .$$

Wang(2010)은 교차타당성(cross-validation)방법을 이용하여 군집화 불안정성을 추정하였다. 먼저 $\boldsymbol{X_n}=\{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}\}$을 3개의 데이터 세트로 분류하고 그 중 2개의 세트를 이용하여 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$을 결정하였다. 이때, 군집화

$\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$는 k-평균 군집화 알고리즘에 의해 결정된다. 그리고 나머지 세트를 이용하여 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 거리를 측정하였다. 이때, 군집화의 불안정성을 투표화와 평균을 이용하여 계산하였다. 투표를 이용한 교차타당성방법과 평균을 이용한 교차타당성 방법의 알고리즘은 아래의 [표 2], [표 3]과 같다.

[표 2] 투표화 교차타당성방법을 이용한 군집개수 결정 알고리즘

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터 n개의 개체를 비복원으로 추출하여 $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 구한다. 이때, $x\in\mathbb{R}^p$이며 $c=1,...,C$으로 C는 교차타당성방법의 반복수이다.
2. $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 세 개의 데이터 세트로 분리한다. 각각의 데이터세트는 $\boldsymbol{z_1^{*c}}=\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_m^{*c}}\}$, $\boldsymbol{z_2^{*c}}=\{\boldsymbol{x_{m+1}^{*c}},...,\boldsymbol{x_{2m}^{*c}}\}$, $\boldsymbol{z_3^{*c}}=\{\boldsymbol{x_{2m+1}^{*c}},...,\boldsymbol{x_n^{*c}}\}$ 이다. 여기서 m은 군집화를 적용하는 데이터셋의 크기(데이터의 수)이다.
3. 아래의 식을 이용하여 각 군집 k에 대해서 군집화 $\psi_1^{*c}(\boldsymbol{x})$와 $\psi_2^{*c}(\boldsymbol{x})$의 거리를 계산한다.

$$d(\psi_1^{*c},\psi_2^{*c})=\left|\left[I\{\psi_1^{*c}(\boldsymbol{x_i^{*c}})=\psi_1^{*c}(\boldsymbol{x_j^{*c}})\}-I\{\psi_2^{*c}(\boldsymbol{x_i^{*c}})=\psi_2^{*c}(\boldsymbol{x_j^{*c}})\}\right]\right|$$

이때, $I(\cdot)$는 지시함수이며 ψ_1^{*c}은 $\boldsymbol{z_1^{*c}}$에 의한 군집화 함수이고 ψ_2^{*c}은 $\boldsymbol{z_2^{*c}}$에 의한 군집화 함수이다. 군집화는 k-평균 군집분석에 의해 정의하였다. 그리고 $\boldsymbol{x_i^{*c}}$, $\boldsymbol{x_j^{*c}}$ 는 $\boldsymbol{z_3^{*c}}$으로부터의 개체이다.

4. $d(\psi_1^{*c},\psi_2^{*c})$은 0또는 1을 가지는 값이므로 $d(\psi_1^{*c},\psi_2^{*c})$의 합을 군집화 불안정성으로 정의한다. 그 식은 아래와 같다.

$$\widehat{s^{*c}}(\Psi,k,m)=\sum_{2m+1\le i<j\le n} d(\psi_1^{*c},\psi_2^{*c})$$

5. $\widehat{s^{*c}}(\Psi,k,m)$이 가장 작을 때의 k값을 $\widehat{k^{*c}}$라 하자.

$$\widehat{k^{*c}}=\mathrm{argmin}_{2\le k\le K}\,\widehat{s^{*c}}(\Psi,k,m)$$

6. $c=1,...,C$ 에 대해서 단계1~단계5를 반복하여 $\{\widehat{k^{*1}},...,\widehat{k^{*C}}\}$의 최빈값을 최적의 군집개수인 $\hat{k}$로 한다.

[표 3] 평균 교차타당성방법을 이용한 군집개수 결정 알고리즘

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터 n개의 개체를 비복원으로 추출하여 $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 구한다. 이때, $x\in\mathbb{R}^p$이며 $c=1,...,C$으로 C는 교차타당성방법의 반복수이다.
2. $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 세 개의 데이터 세트로 분리한다. 각각의 데이터세트는 $\boldsymbol{z_1^{*c}}=\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_m^{*c}}\}$, $\boldsymbol{z_2^{*c}}=\{\boldsymbol{x_{m+1}^{*c}},...,\boldsymbol{x_{2m}^{*c}}\}$, $\boldsymbol{z_3^{*c}}=\{\boldsymbol{x_{2m+1}^{*c}},...,\boldsymbol{x_n^{*c}}\}$이다. 여기서 m은 군집화를 적용하는 데이터셋의 크기(데이터의 수)이다.
3. 아래의 식을 이용하여 각 군집 k에 대해서 군집화 $\psi_1^{*c}(\boldsymbol{x})$와 $\psi_2^{*c}(\boldsymbol{x})$의 거리를 계산한다.

$$d(\psi_1^{*c},\psi_2^{*c})=\left|\left[I\{\psi_1^{*c}(\boldsymbol{x_i^{*c}})=\psi_1^{*c}(\boldsymbol{x_j^{*c}})\}-I\{\psi_2^{*c}(\boldsymbol{x_i^{*c}})=\psi_2^{*c}(\boldsymbol{x_j^{*c}})\}\right]\right|$$

이때, $I(\cdot)$는 지시함수이며 ψ_1^{*c}은 $\boldsymbol{z_1^{*c}}$에 의한 군집화 함수이고 ψ_2^{*c}은 $\boldsymbol{z_2^{*c}}$에 의한 군집화 함수이다. 군집화는 k-평균 군집분석에 의해 정의하였다. 그리고 $\boldsymbol{x_i^{*c}}$, $\boldsymbol{x_j^{*c}}$ 는 $\boldsymbol{z_3^{*c}}$으로부터의 개체이다.

4. $d(\psi_1^{*c},\psi_2^{*c})$은 0또는 1을 가지는 값이므로 $d(\psi_1^{*c},\psi_2^{*c})$의 합을 군집화 불안정성으로 정의한다. 그 식은 아래와 같다.

$$\widehat{s^{*c}}(\Psi,k,m)=\sum_{2m+1\le i<j\le n} d(\psi_1^{*c},\psi_2^{*c})$$

5. $c=1,...,C$ 에 대해서 단계1~단계4를 반복하여 $\widehat{s^{*c}}(\Psi,k,m)$의 평균을 계산한다.

$$\hat{s}(\Psi,k,m)=\frac{1}{C}\sum_{c=1}^{C}\widehat{s^{*c}}(\Psi,k,m)$$

6. $\hat{s}(\Psi,k,m)$이 가장 작을 때의 k값이 최적의 군집개수인 $\hat{k}$이다.

$$\hat{k}=\arg\min_{2\le k\le K}\hat{s}(\Psi,k,m)$$

교차타당성방법을 적용하여 군집화 불안정성을 추정하는 것에서 나아가 Fang & Wang(2012)은 붓스트랩을 이용하여 군집화 불안정성을 추정한 후 불안정성을 최소화하는 군집개수 k를 최적의 군집개수로 결정하였다. 이때, 군집화 불안정성을 측정하기위해 마찬가지로 군집화 거리를 이용하였다. 분포 F로부터 개체들을 복원 추출하여 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$를 정의하고 $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$을 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$에 적용하여 군집화간의 거리를 추정하였다. 이때, 군집화 거리를 추정하는 과정을 C번 반복하여 군집화 불안정성을 계산하였다. 붓스트랩을 이용한 군집개수 결정에 관한 알고리즘은 아래의 [표 4]와 같다.

[표 4] 붓스트랩을 이용한 군집개수 결정 알고리즘

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$은 크기가 n인 알려지지 않은 분포 $F(x)$로부터의 랜덤표본이라 하자. 이때, $x\in\mathbb{R}^p$이다. 먼저, 독립인 C개의 붓스트랩 표본 $\boldsymbol{X_n^{*c}}$와 $\widehat{\boldsymbol{X_n^{*c}}}$을 $\hat{F}$로부터 복원 추출한다. 이때, $c=1,...,C$으로 C는 붓스트랩의 반복수이다.
2. 각 군집 k에 대해서 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 경험적 거리를 계산한다.

$$d(\psi_1^{*c},\psi_2^{*c})=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|I\{\psi_1^{*c}(\boldsymbol{x_i})=\psi_1^{*c}(\boldsymbol{x_j})\}-I\{\psi_2^{*c}(\boldsymbol{x_i})=\psi_2^{*c}(\boldsymbol{x_j})\}|$$

이때, $I(\cdot)$는 지시함수이며 군집화 ψ_1^{*c}, ψ_2^{*c}은 각각 $\boldsymbol{X_n^{*c}}$와 $\widehat{\boldsymbol{X_n^{*c}}}$에 의한 함수 이다. 군집화는 k-평균 군집분석에 의해 정의하였다. 그리고 $\boldsymbol{x_i}$와 $\boldsymbol{x_j}$는 $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터의 개체이다.

3. C번의 단계1~단계3 반복 시행을 통해 계산한 $d(\psi_1^{*c},\psi_2^{*c})$을 이용하여 아래와 같이 군집화 불안정성을 계산한다.

$$\widehat{s_C}(\Psi,k,n)=\frac{1}{C}\sum_{c=1}^{C}d(\psi_1^{*c},\psi_2^{*c})$$

4. $\widehat{s_C}(\Psi,k,n)$이 가장 작을 때의 k값이 최적의 군집개수인 $\hat{k}$이다.

$$\hat{k}=\arg\min_{2\le k\le K}\widehat{s_C}(\Psi,k,n)$$

식(2.14)의 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 거리 $d(\psi_1,\psi_2)$는 각 개체를 적용할 때마다, 두 군집화 결과가 일치하면 0, 불일치하면 1의 값을 갖게 된다. 군집화 거리는 단순한 이항자료로써 각 개체에서의 군집화 거리의 평균값이 군집화 불안정성이다.

본 연구에서는 군집화의 거리에서 $I\{\psi_1(\boldsymbol{x^0})=\psi_1(\boldsymbol{y^0})\}$과 $I\{\psi_2(\boldsymbol{x^0})=\psi_2(\boldsymbol{y^0})\}$가 0 또는 1의 이항자료로 나타나는 특성을 이용하여 이항자료에서의 연관성측도를 적용함으로써 군집화 불안정성을 측정하고자한다. 이를 위해 다음 장에서 연관성의 측도를 소개하고 이를 군집화 불안정성 측정에 사용한다.

제 3 장 연관성측도를 이용한 군집개수의 결정

기존의 군집화 불안정성은 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$가 서로 다른 군집결과를 얼마나 출력했는지를 측정하여 정의하였다. 본 장에서는 범주형 자료에서의 연관성측도를 적용하여 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 불안정성을 측정하고자한다. 먼저 범주형 자료에서의 연관성측도를 살펴보자.

3.1. 연관성의 측도

여러 변수들 간의 상호연관성을 분석하는 것은 주요 관심대상이며 자료에서 연관성의 존재여부를 측정하는 여러 가지 연관성측도들이 있다. 두 범주형 변수간의 관련성을 나타내는 통계량인 연관성의 측도를 살펴보자.

1) Cohen의 카파 계수(Cohen's kappa coefficient)

카파계수는 이분 범주형 자료에서의 일치도를 측정하는 측도로 아래의 식 (3.1)과 같이 정의된다(Cohen, 1960).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad , \tag{3.1}$$

여기서 p_e은 두 변수의 범주가 일치하는 기대비율, p_o은 두 변수의 범주가 일치하는 관측비율이다.

카파계수 κ의 범위는 $0 \le \kappa \le 1$으로 $\kappa = 1$일 때 두 변수간의 연관성이 가장 높다. 이분형 자료에 대해서는 변수가 정확히 일치하거나 완전히 불일치하기 때문에 카파계수를 사용하는데 문제가 없지만, 다항변수나 연속형 자료를 임의의 수와 크기의 범주로 나누는 경우에 적용 할 시에 문제가 발생한다. 다항 범주형 자료의 경우에 한 쌍의 변수는 다른 것보다 더욱 상이한 것으로 간주

되므로 일부의 불일치의 경우에 결과가 좋지 않다.

2) 가중 카파 계수(weighted kappa coefficient)

카파계수 κ는 모든 경우의 불일치를 동일한 것으로 간주하는데 이런 약점을 처리하기 위해 가중 카파계수가 사용된다. 가중카파계수는 다음의 식(3.2)과 같이 정의한다(Cohen, 1968).

$$\kappa_w = \frac{p_{ow} - p_{ew}}{1 - p_{ew}} \quad , \tag{3.2}$$

여기서 p_{ew}은 두 변수의 범주가 일치하는 기대비율인 p_e에 가중치 w를 곱한 값이고, p_{ow}는 두 변수의 범주가 일치하는 관측비율인 p_o에 가중치 w를 곱한 값이다.

가중 카파계수 κ_w는 불일치를 보이는 관측의 가능한 조합에 고유의 가중치를 주도록 허용한다. 대각행렬에 대한 가중치가 0이고 비 대각행렬에 대한 가중치가 1일 때 카파 계수와 동일하다.

3) 자카드 유사성 계수(Jaccard similarity coefficient)

자카드 유사성 계수(Jaccard similarity coefficient)는 두 변수가 0 또는 1을 갖는 이분형 자료일 때 아래의 식(3.3)와 같이 정의된다(Jaccard, 1902, Jaccard, 1912).

$$J = \frac{a}{a + b + c} \quad , \tag{3.3}$$

여기서 a는 두 변수가 모두 1일 때의 빈도수이며 b, c는 각각 두 변수가 서로 다른 값을 가질 때의 빈도수이다.

두 변수가 모두 0을 값을 갖는 경우는 지수에서 제외되며 일치와 비일치에

동일한 가중치를 적용한다. 자카드 유사성 계수의 범위는 $0 \leq J \leq 1$으로 값이 1일 때, 두 변수가 모두 1로 일치함을 의미한다. 그리고, 자카드의 거리(Jaccard's distance)는 아래와 같이 정의할 수 있다.

$$1-J=\frac{b+c}{a+b+c} \ ,$$

자카드의 거리는 1에서 자카드 계수를 뺀 값으로 그 값이 작을수록 두 변수가 유사함을 의미한다.

4) 파이 계수(phi coefficient)

파이 계수(phi coefficient)는 이항변수를 위한 연관성측도로서 피어슨의 상관계수의 연장이다. 즉, 두 이항변수에 피어슨의 상관계수를 적용한 결과와 같은 값을 갖는다. 파이계수는 2×2 분할표에서의 카이제곱 통계량과 관련이 있으며 아래의 식(3.4)와 같다(Pearson, 1900).

$$\phi=\sqrt{\frac{\chi^2}{n}} \ , \tag{3.4}$$

여기서 n은 전체 관측값이며, 파이계수의 분자부분은 아래의 식(3.5)인 카이제곱 통계량과 일치한다.

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})}{e_{ij}} \ , \tag{3.5}$$

n_{ij}는 각 셀의 빈도이며 $e_{ij}=n_{i.}n_{.j}/n$은 기대빈도이다.

파이계수의 범위는 $-1 \leq \phi \leq 1$ 이며 절대 값이 클수록 두 범주형 변수간의 연관성이 크며 주변빈도크기에 영향을 받는다는 단점이 있다.

3.2. 새로운 군집개수 결정 알고리즘

교차타당성방법과 붓스트랩 방법을 기반으로 한 기존의 군집결정 알고리즘에서는 군집화의 거리를 식(2.14)로 정의하고 군집화 불안정성을 군집화 거리의 식(2.15)와 같이 단순평균값으로 고려하였다. 군집화 거리는 0또는 1을 가지는 이항자료로 군집화 거리의 단순 평균값을 군집화의 불안정성으로 정의할 시에 두 군집화가 일치했을 경우만 고려된다.

따라서 본 연구에서는 연관성측도를 적용하여 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 연관성이 가장 높을 때, 즉 두 군집화 불안정성이 가장 낮을 때의 k값을 최적의 군집개수로 결정한다. 여기서 군집화 $\psi(\boldsymbol{x})$은 자료를 군집 $\{1,\cdots,k\}$에 대응시켜주는 함수이고 $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$은 알려지지 않은 p차원의 분포 $F(x)$로부터의 독립표본이다. 그리고, $\boldsymbol{X_n}$은 $n\times p$행렬이고, $\boldsymbol{x_1},\cdots,\boldsymbol{x_n}$은 $1\times p$벡터이다. 군집화 불안정성 측정을 위해서 A과 B를 아래의 식(3.6)과 같이 정의한다.

$$A=I\{\psi_1(\boldsymbol{x^0})=\psi_1(\boldsymbol{y^0})\},\quad B=I\{\psi_2(\boldsymbol{x^0})=\psi_2(\boldsymbol{y^0})\}\ , \tag{3.6}$$

이때, $I(\cdot)$은 지시함수이고 $\boldsymbol{x^0}$와 $\boldsymbol{y^0}$은 $1\times p$벡터로서 분포 $F(x)$로부터의 독립적인 개체이다.

식(3.6)의 A과 B는 0또는 1값을 가지는 이항자료이다. A과 B가 같은 값을 가질 때 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$는 일치한다고 볼 수 있다. A와 B는 이항자료로 분할표를 나타내면 [표 5]와 같다.

[표 5] A와 B의 분할표

		A	
		1	0
B	1	a	b
	0	c	d

이때, a와 d는 A와 B의 값이 일치했을 때의 빈도수로 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$에 자료를 적용했을 때 같은 군집결과가 나오는 경우의 수이다. 그리고 b와 c는 불일치의 빈도수로 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$에 자료를 적용했을 때 서로 다른 군집결과가 나오는 경우의 수이다. 따라서, a와 d의 빈도수가 많다는 것은 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 결과가 유사함을 의미하며 이는 군집화가 안정적이라고 해석 할 수 있다.

본 연구에서는 연관성측도를 통하여 A와 B의 연관성을 측정하여 군집화 불안정성을 정의하고자 한다. 연관성측도 중 자카드계수, 카파계수, 파이계수를 적용하여 군집화의 연관성을 측정하고 연관성이 최대가 될 때의 군집의 개수를 최적의 군집의 개수로 정하였다.

군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 연관성이 높다는 것은 두 군집화의 안정성이 높음을 의미하며 이는 불안정성이 낮은 것이다. 따라서 연관성측도를 이용한 군집화 불안정성을 아래의 식(3.7)과 같이 정의한다.

$$\tilde{s}(\Psi, k, m) = -f(A, B) \quad , \tag{3.7}$$

이때, A와 B는 식(3.6)의 값을 의미하며 $f(\cdot)$는 자카드계수, 카파계수, 파이계수를 나타낸다.

연관성측도를 이용한 새로운 군집개수 결정 알고리즘은 투표화 교차타당성 방법과 평균 교차타당성방법, 그리고 붓스트랩방법에 적용할 수 있으며 각각 [표 6], [표 7], [표 8]으로 아래와 같다.

[표 6] 연관성측도를 이용한 군집결정 알고리즘(투표화 교차타당성 방법)

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터 n개의 개체를 비복원으로 추출하여 $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 구한다. 이때, $c=1,...,C$으로 C는 교차타당성방법의 반복수이다.
2. $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 세 개의 데이터 세트로 분리한다. 각각의 데이터세트는 $\boldsymbol{z_1^{*c}}=\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_m^{*c}}\}$, $\boldsymbol{z_2^{*c}}=\{\boldsymbol{x_{m+1}^{*c}},...,\boldsymbol{x_{2m}^{*c}}\}$, $\boldsymbol{z_3^{*c}}=\{\boldsymbol{x_{2m+1}^{*c}},...,\boldsymbol{x_n^{*c}}\}$ 이다. 여기서 m은 군집화를 적용하는 데이터 셋의 크기(데이터의 수)이다.
3. 각 군집 k에 대해서 아래의 식을 계산한다.

$$A^{*c}=I\{\psi_1^{*c}(\boldsymbol{x_i^{*c}})=\psi_1^{*c}(\boldsymbol{x_j^{*c}})\},\ \ B^{*c}=I\{\psi_2^{*c}(\boldsymbol{x_i^{*c}})=\psi_2^{*c}(\boldsymbol{x_j^{*c}})\}$$

이때, $I(\cdot)$는 지시함수이며 ψ_1^{*c}은 $\boldsymbol{z_1^{*c}}$에 의한 군집화 함수이고 ψ_2^{*c}은 $\boldsymbol{z_2^{*c}}$에 의한 군집화 함수이다. 군집화는 k-평균 군집분석에 의해 정의하였다. 그리고 $\boldsymbol{x_i^{*c}}$, $\boldsymbol{x_j^{*c}}$ 는 $\boldsymbol{z_3^{*c}}$으로부터의 개체이다.

4. 연관성측도를 통해 A와 B의 연관성을 계산하여 군집화 불안정성을 아래와 같이 정의한다.

$$\widetilde{s^{*c}}(\Psi,k,m)=-f(A^{*c},B^{*c})$$

이때, $f(\cdot)$은 연관성측도인 카파계수, 자카드계수, 파이계수를 적용한다.

5. $\widetilde{s^{*c}}(\Psi,k,m)$이 가장 작을 때의 k값을 $\widehat{k^{*c}}$라 하자.

$$\widehat{k^{*c}}=\mathrm{argmin}_{2\le k\le K}\widehat{s^{*c}}(\Psi,k,m)$$

6. $c=1,...,C$ 에 대해서 단계1~단계5를 반복하여 $\{\widehat{k^{*1}},...,\widehat{k^{*C}}\}$의 최빈값을 최적의 군집개수인 $\hat{k}$로 한다.

[표 7] 연관성측도를 이용한 군집결정 알고리즘(평균 교차타당성 방법)

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터 n개의 개체를 비복원으로 추출하여 $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 구한다. 이때, $c=1,...,C$으로 C는 교차타당성방법의 반복수이다.
2. $\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_n^{*c}}\}$을 세 개의 데이터 세트로 분리한다. 각각의 데이터세트는 $\boldsymbol{z_1^{*c}}=\{\boldsymbol{x_1^{*c}},...,\boldsymbol{x_m^{*c}}\}$, $\boldsymbol{z_2^{*c}}=\{\boldsymbol{x_{m+1}^{*c}},...,\boldsymbol{x_{2m}^{*c}}\}$, $\boldsymbol{z_3^{*c}}=\{\boldsymbol{x_{2m+1}^{*c}},...,\boldsymbol{x_n^{*c}}\}$ 이다. 여기서 m은 군집화를 적용하는 데이터셋의 크기(데이터의 수)이다.
3. 각 군집 k에 대해서 아래의 식을 계산한다.

$$A^{*c}=I\{\psi_1^{*c}(\boldsymbol{x_i^{*c}})=\psi_1^{*c}(\boldsymbol{x_j^{*c}})\},\ B^{*c}=I\{\psi_2^{*c}(\boldsymbol{x_i^{*c}})=\psi_2^{*c}(\boldsymbol{x_j^{*c}})\}$$

이때, $I(\cdot)$는 지시함수이며 ψ_1^{*c}은 $\boldsymbol{z_1^{*c}}$에 의한 군집화 함수이고 ψ_2^{*c}은 $\boldsymbol{z_2^{*c}}$에 의한 군집화 함수이다. 군집화는 k-평균 군집분석에 의해 정의하였다. 그리고 $\boldsymbol{x_i^{*c}}$, $\boldsymbol{x_j^{*c}}$는 $\boldsymbol{z_3^{*c}}$으로부터의 개체이다.

4. 연관성측도를 통해 A와 B의 연관성을 계산하여 군집화 불안정성을 아래와 같이 정의한다.

$$\widetilde{s^{*c}}(\Psi,k,m)=-f(A^{*c},B^{*c})$$

이때, $f(\cdot)$은 연관성측도인 카파계수, 자카드계수, 파이계수를 적용한다.

5. $c=1,...,C$ 에 대해서 단계1~단계4를 반복하여 $\widetilde{s^{*c}}(\Psi,k,m)$의 평균을 계산한다.

$$\tilde{s}(\Psi,k,m)=\frac{1}{C}\sum_{c=1}^{C}\widetilde{s^{*c}}(\Psi,k,m)$$

6. $\tilde{s}(\Psi,k,m)$이 가장 작을 때의 k값이 최적의 군집개수인 $\hat{k}$이다.

$$\hat{k}=\mathrm{argmin}_{2\le k\le K}\tilde{s}(\Psi,k,m)$$

[표 8] 연관성측도를 이용한 군집결정 알고리즘(붓스트랩 방법)

1. $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$은 크기가 n인 알려지지 않은 분포 $F(x)$로부터의 랜덤표본이라 하자. 이때, $x\in\mathbb{R}^p$이다. 먼저, 독립인 C개의 붓스트랩 표본 $\boldsymbol{X_n^{*c}}$와 $\widetilde{\boldsymbol{X_n^{*c}}}$을 $\hat{F}$로부터 복원 추출한다. 이때, $c=1,...,C$으로 C는 붓스트랩의 반복수이다.
2. 각 군집 k에 대해서 아래의 식을 계산한다.

$$A^{*c}=I\{\psi_1^{*c}(\boldsymbol{x_i^{*c}})=\psi_1^{*c}(\boldsymbol{x_j^{*c}})\},\ B^{*c}=I\{\psi_2^{*c}(\boldsymbol{x_i^{*c}})=\psi_2^{*c}(\boldsymbol{x_j^{*c}})\}$$

이때, $I(\cdot)$는 지시함수이며 ψ_1^{*c}은 $\boldsymbol{X_n^{*c}}$에 의한 군집화 함수이고 ψ_2^{*c}은 $\widetilde{\boldsymbol{X_n^{*c}}}$에 의한 군집화 함수이다. x_i^{*c}, x_j^{*c} 는 $\boldsymbol{X_n}=\{\boldsymbol{x_1},\cdots,\boldsymbol{x_n}\}$으로부터의 개체이다.

3. 연관성측도를 통해 A와 B의 연관성을 계산하여 군집화 불안정성을 아래와 같이 정의한다.

$$\widetilde{s^{*c}}(\Psi,k,n)=-f(A,B)$$

이때, $f(\cdot)$은 연관성측도인 카파계수, 자카드계수, 파이계수를 적용한다.

4. C번의 단계1~단계3 반복 시행을 통해 계산한 $\widetilde{s^{*c}}(\Psi,k,n)$의 평균을 계산한다.

$$\tilde{s}(\Psi,k,n)=\frac{1}{C}\sum_{c=1}^{C}\widetilde{s^{*c}}(\Psi,k,n)$$

5. $\widetilde{s_C}(\Psi,k,n)$이 가장 작을 때의 k값이 최적의 군집개수인 $\hat{k}$이다.

$$\hat{k}=\mathrm{argmin}_{2\le k\le K}\tilde{s}(\Psi,k,n)$$

제 4 장 모의실험

앞 장에서는 두 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 불안정성을 측정하기 위해 연관성 측도인 자카드계수, 카파계수, 파이계수를 적용하여 군집화 불안정성을 정의하고 새로운 군집개수를 결정하기위한 알고리즘을 제안하였다.

Fang & Wang(2012)은 기존에 두 군집화의 불안정성을 단순한 거리개념으로 식(2.15)과 같이 정의하였다. 따라서 4장에서는 기존 Wang의 방법과 연관성측도인 자카드계수, 카파계수, 파이계수를 적용한 새로운 군집결정 방법의 수행능력을 비교하고자 한다.

4.1. 모의실험 설계

군집개수 결정의 효율성을 평가하기 위해 군집 간 거리, 차원, 상관관계, 군집개수, 분산 크기에 따라 4개의 시나리오를 설정하였다.

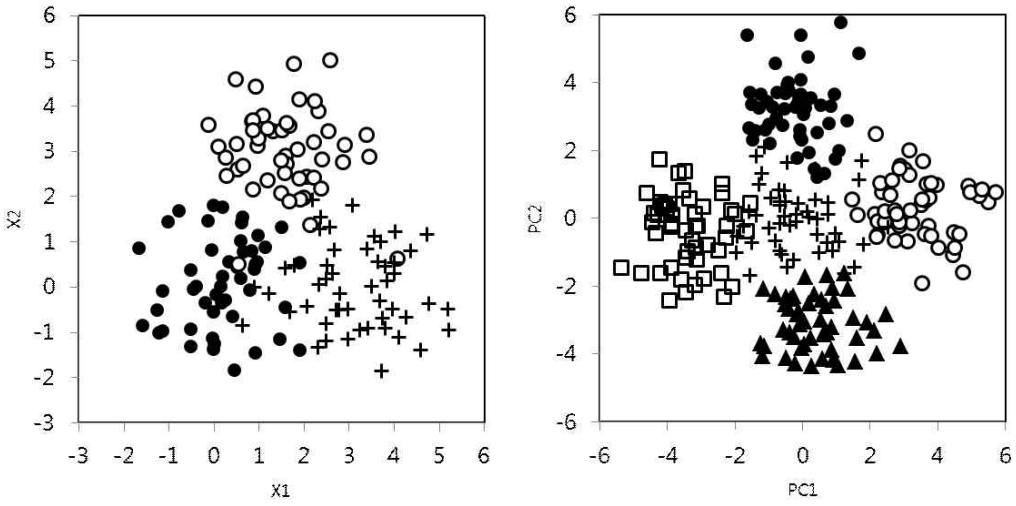
각 시나리오와 시나리오의 분산-공분산 행렬 V_k는 다음과 같다. 이때, V_k은 $p\times p$행렬이고 k는 군집을 의미하며 V_k에서 대각원소는 군집의 분산, 비 대각원소는 군집 내 상관계수이다.

시나리오 1	군집 간의 거리와 군집 내 상관계수(군집끼리는 동일)에 따라 데이터 생성. 이때, 모든 군집은 동일한 분산($\sigma^2=1$)을 갖으며, 군집 내 상관계수는 0임($\rho=0$).	$V_k=\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}.\sigma^2$

시나리오 2	군집 간의 거리와 군집 내 상관계수(군집끼리는 동일)에 따라 데이터 생성. 이때, 모든 군집은 동일한 분산($\sigma^2 = 1$)을 갖으며, 군집 내 상관계수는 0.6임.($\rho = 0.6$)	$V_k = \begin{pmatrix} 1 & 0.6 & \cdots & 0.6 \\ 0.6 & 1 & \cdots & 0.6 \\ \vdots & \vdots & \ddots & \vdots \\ 0.6 & 0.6 & \cdots & 1 \end{pmatrix} . \sigma^2$
시나리오 3	군집개수는 3개로 고정. 군집 간의 거리에 따라 데이터 생성. 이때, 군집마다 분산을 다르게 고려하며($\sigma^2 = 1.5, 0.5, 0.5$) 모든 군집 내의 상관계수는 0임($\rho = 0$).	$V_k = \begin{pmatrix} \sigma_k^2 & 0 & \cdots & 0 \\ 0 & \sigma_k^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{pmatrix}$
시나리오 4	군집개수는 3개로 고정. 군집 간의 거리에 따라 데이터 생성. 이때, 군집마다 상관계수를 다르게 고려하며($\rho = 0.6, 0, 0$) 모든 군집은 동일한 분산임($\sigma^2 = 1$).	$V_k = \begin{pmatrix} 1 & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \cdots & \rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k & \cdots & 1 \end{pmatrix} . \sigma^2$

<시나리오 1>

- 군집의 개수 k_0는 3개, 4개, 5개를 고려. ($k_0 = 3, 4, 5$)
- 각 관측치는 군집의 중심을 평균값으로 갖고 분산이 1인 이변량정규분포로부터 생성. ($n_1, \ldots, n_{k_0} = 50$)
- 상관계수는 0. ($\rho = 0$)
- 군집 간의 거리 d는 2.5에서 3.5까지 0.25씩 증가. ($d = 2.5, 2.75, 3.0, 3.25, 3.5$)
- 차원 p는 2, 5, 7, 10이며 이때, 5, 7, 10차원에서 남은 3차원, 5차원과 8차원은 군집의 정보 없이 표준정규분포로부터 생성. ($p = 2, 5, 7$)



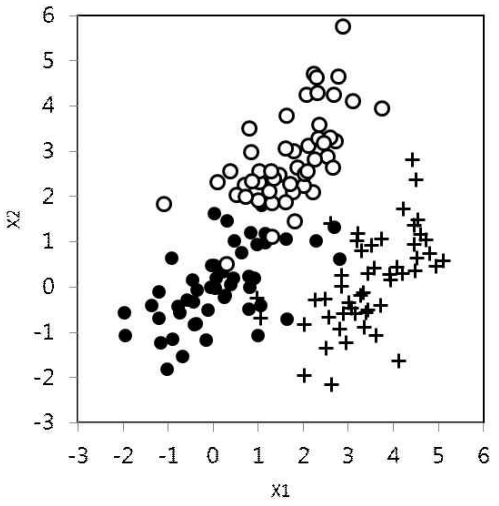
(a) k=3, p=2, ρ=0, d=3.25

(b) k=5, p=5, ρ=0, d=3.25
(주성분 분석 적용 후 산점도)

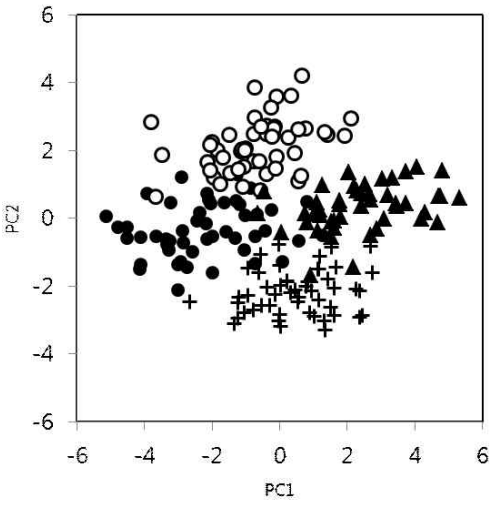
[그림 2] <시나리오 1>의 산점도

<시나리오 2>

- 군집의 개수 k_0는 3개, 4개, 5개를 고려. ($k_0 = 3, 4, 5$)
- 각 관측치는 군집의 중심을 평균값으로 갖고 분산이 1인 이변량정규분포로부터 생성. ($n_1, \ldots, n_{k_0} = 50$)
- 상관계수는 0.6. ($\rho = 0.6$)
- 군집 간의 거리 d는 2.5에서 3.5까지 0.25씩 증가. ($d = 2.5, 2.75, 3.0, 3.25, 3.5$)
- 차원 p는 2, 5, 7, 10이며 이때, 5, 7, 10차원에서 남은 3차원, 5차원과 8차원은 군집의 정보 없이 표준정규분포로부터 생성. ($p = 2, 5, 7$)



(a) k=3, p=2, ρ=0.6, d=3.25

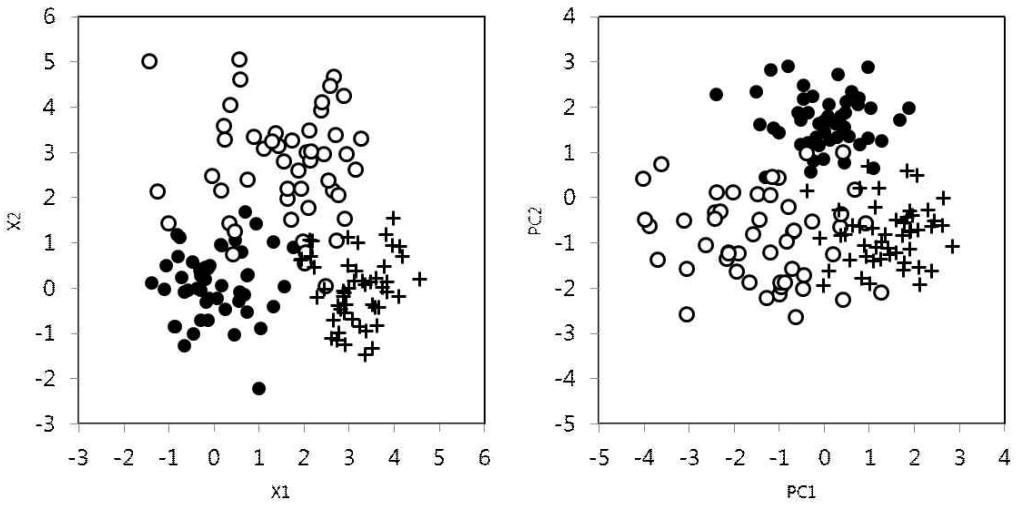


(b) k=4, p=5, ρ=0.6, d=3.25
(주성분 분석 적용 후 산점도)

[그림 3] <시나리오 2>의 산점도

<시나리오 3>

- 군집의 개수 k_0는 3개를 고려. ($k_0 = 3$)
- 각 관측치는 군집의 중심을 평균값으로 갖고 각 군집에 대응되는 분산을 갖는 이변량정규분포로부터 생성. ($n_1, ..., n_{k_0} = 50$)
- 각 군집의 상관계수는 0으로 설정. ($\rho_1, \rho_2, \rho_3 = 0$)
- 각 군집의 분산은 1.5, 0.5, 0.5로 설정. ($\sigma_1 = 1.5, \sigma_2, \sigma_3 = 0.5$)
- 군집 간의 거리 d는 2에서 3까지 0.25씩 증가. ($d = 2.0, 2.25, 2.5, 2.75, 3.0$)
- 차원 p는 2, 5이며 이때, 5차원에서 남은 3차원은 군집의 정보 없이 표준정규분포로부터 생성. ($p = 2, 5, 7$)



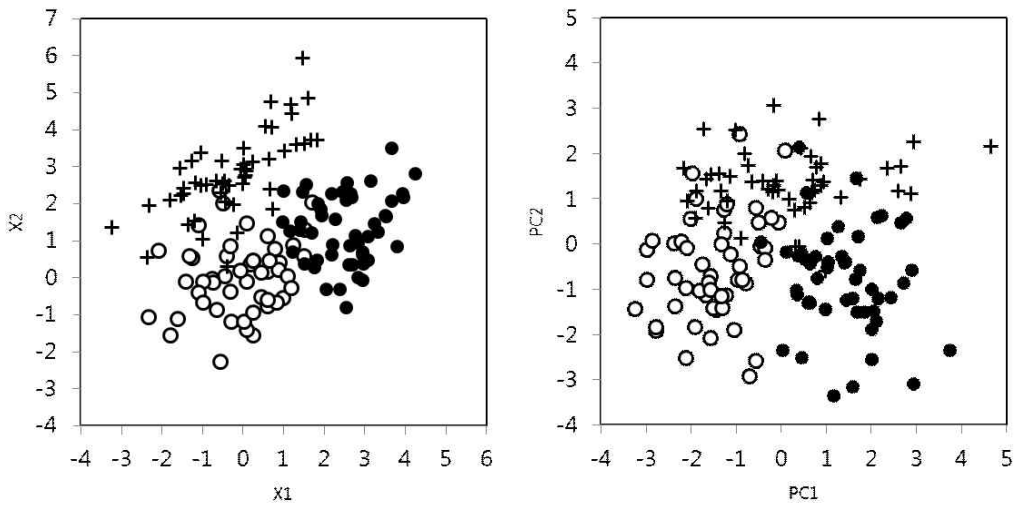
(a) k=3, p=2, d=3.00

(b) k=5, p=5, d=3.00
(주성분 분석 적용 후 산점도)

[그림 4] <시나리오 3>의 산점도

<시나리오 4>

- 군집의 개수 k_0는 3개를 고려. ($k_0 = 3$)
- 각 관측치는 군집의 중심을 평균값으로 갖고 분산이 1인 이변량정규분포로부터 생성. ($n_1, ..., n_{k_0} = 50$)
- 각 군집의 상관계수는 0.6, 0, 0으로 설정. ($\rho_1 = 0.6,\ \rho_2, \rho_3 = 0$)
- 군집 간의 거리 d는 2에서 3까지 0.25씩 증가. ($d = 2.0, 2.25, 2.5, 2.75, 3.0$)
- 차원 p는 2, 5이며 이때, 5차원에서 남은 3차원은 군집의 정보 없이 표준정규분포로부터 생성. ($p = 2, 5, 7$)



(a) k=3, p=2, d=3.00 (b) k=3, p=5, d=3.00

(주성분 분석 적용 후 산점도)

[그림 5] <시나리오 4>의 산점도

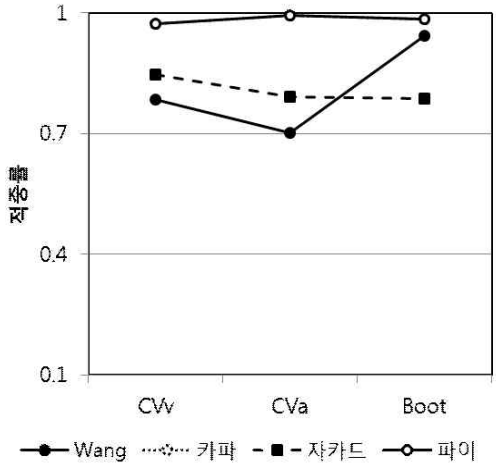
관측치는 각 군집당 50개이며 붓스트랩과 교차타당성평가를 위한 반복수 C는 50으로 하였다. 본 연구에서는 군집화 불안정성을 측정하기위해 연관성의 측도인 카파계수, 자카드계수, 파이계수를 적용하였으며 같은 자료에 대해 200

번 시행하여 그 적중률을 비교하였다.

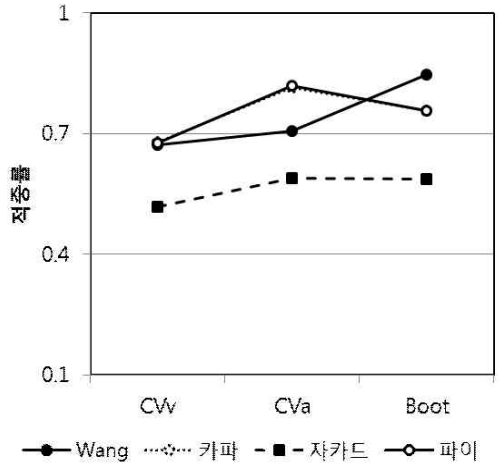
4.2. 모의실험 결과

군집개수 결정의 수행능력을 비교하기 위해 모의실험을 시행하여 기존의 Wang(2010), Fang & Wang(2012)방법과 자카드계수, 카파계수, 파이계수를 적용한 방법의 수행능력을 비교하였다. 모의실험은 각 자료마다 200번 시행하였으며 알고리즘에 의해 출력된 최적의 군집개수 결과는 [표 9]~[표 18]와 같다.

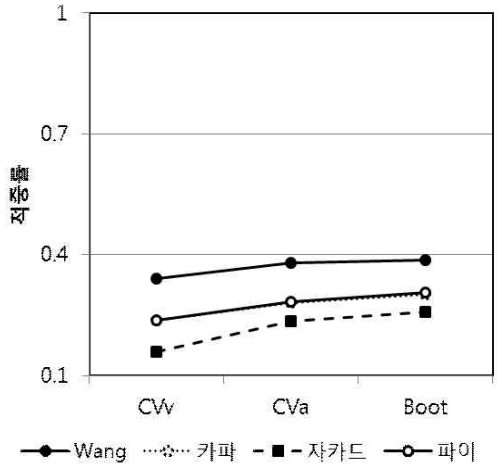
모의실험의 결과를 자료의 군집개수와 차원별로 정리하여 적중률 비교를 하였다. CV_V는 투표화 교차타당성방법, CV_a는 평균 교차타당성방법, *Boot*는 붓스트랩을 이용한 방법이다. 이때, 적중률은 총 시행중에서 실제 군집개수가 선택된 비율이다.



(a) k=3



(b) k=4



(c) k=5

[그림 6] <시나리오 1>자료의 군집개수(k=3, 4, 5)에 따른 적중률(hit ratio)

<시나리오 1>의 결과는 각 알고리즘에 따라 [표 9]~[표 12]와 같다. 먼저 자료의 실제 군집개수에 따라서 [그림 6]과 같이 각 방법의 적중률을 나타낼 수 있다. 군집개수(k)가 3인 자료에 적용한 결과는 [그림 6]의 (a)으로, 본 연구에서 제안하는 방법인 카파계수, 카자드계수, 파이계수를 이용한 방법이 군집개수를 더 정확히 선택함을 알 수 있다. 특히, 카파계수와 파이계수를 이용

한 방법은 거의 비슷한 적중률을 보이며 기존방법보다 적중률이 현저히 높음을 확인할 수 있다.

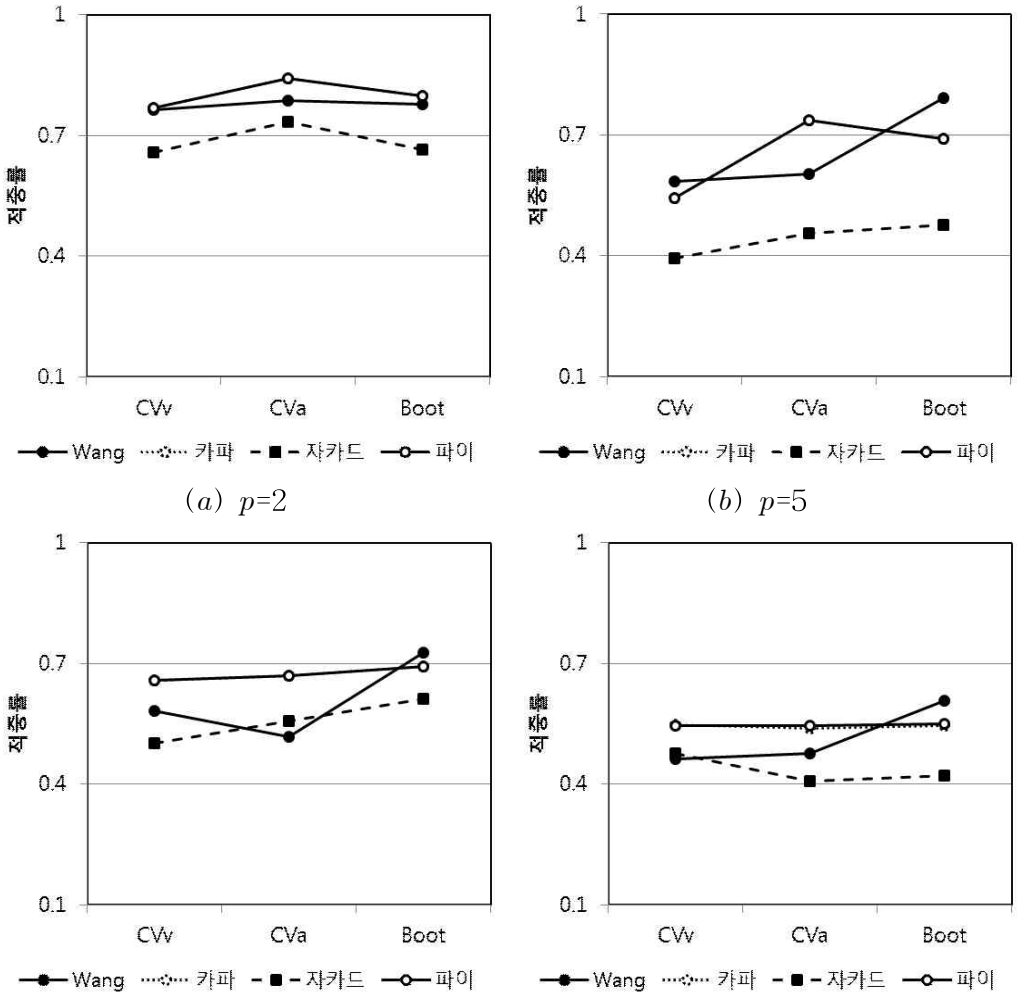
다음은 군집개수(k)가 4인 자료에 적용한 결과로 결과는 [그림 6]의 (b)와 같다. 가장 높은 적중률을 보이는 방법은 카파계수와 파이계수를 이용한 방법이고 붓스트랩을 이용한 기존방법에서도 적중률이 높았다.

군집개수(k)가 5인 자료에 적용한 결과는 [그림 6]의 (c)으로 군집개수가 높아짐에 따라 그 적중률이 낮아짐을 확인할 수 있다. 또한, 군집의 개수(k)가 3, 4일 때와는 다르게 기존 Wang방법의 적중률이 약간 더 높음을 알 수 있다.

<시나리오 1>에 대한 결과를 자료의 차원별로 살펴보면 [그림 7]와 같다. 먼저, 차원(p)이 2인 자료에 적용한 결과는 [그림 7]의 (a)로 모든 방법에서 그 적중률이 비슷하지만, 특히 카파계수와 파이계수를 적용한 방법의 적중률이 높음을 확인할 수 있다. 차원(p)이 5인 자료에 적용한 결과는 [그림 7]의 (b)와 같다. 붓스트랩을 이용한 방법에서 특히 적중률이 높으며 카파계수와 파이계수의 적중률은 비슷함을 알 수 있다. 차원(p)이 7인 자료에 적용한 경우는 [그림 7]의 (c)으로 카파계수와 파이계수를 적용한 방법이 군집개수를 가장 잘 선택함을 알 수 있고, 기존 Wang방법에서는 붓스트랩을 이용한 방법에서 적중률이 높았다. 차원(p)이 10인 자료에 적용한 결과는 [그림 7]의 (d)으로 모든 방법에서 적중률이 비슷하나 카파계수와 파이계수를 적용한 방법에서 적중률이 가장 높음을 확인할 수 있으며, 특히 붓스트랩 방법을 적용했을 시에 적중률이 높음을 알 수 있다.

결과적으로 군집개수(k)가 3, 4인 자료에서는 카파계수와 파이계수를 적용한 새로운 알고리즘의 적중률이 높았으며 군집개수(k)가 5인 자료에서는 모든 방법에서 전체적으로 적중률이 낮았으며 기존 Wang방법에서 적중률이 높았다. 또한, 차원(p)이 2인 자료에서는 카파계수와 파이계수를 적용하는 제안한 방법에서 적중률이 가장 높았으며 차원(p)이 커질수록 모든 방법에서 적중률이 낮아졌다. 차원(p)이 5이상 인 자료에서는 교차타당성방법에서 카파계수, 파이계

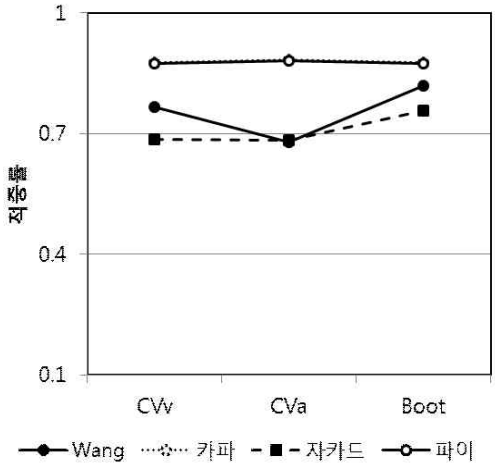
수를 적용한 방법과 기존 Wang방법 중 붓스트랩을 이용한 방법에서 적중률이 높음을 확인할 수 있다.



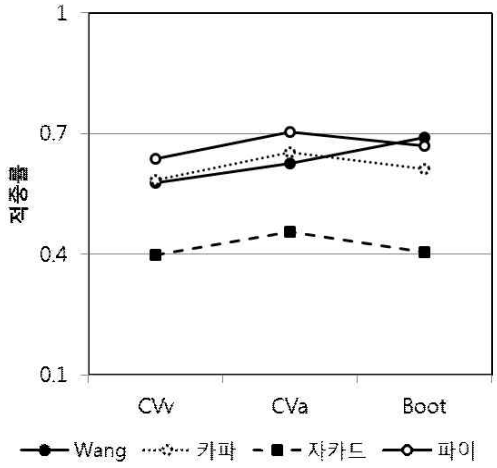
(a) p=2 (b) p=5

(c) p=7 (d) p=10

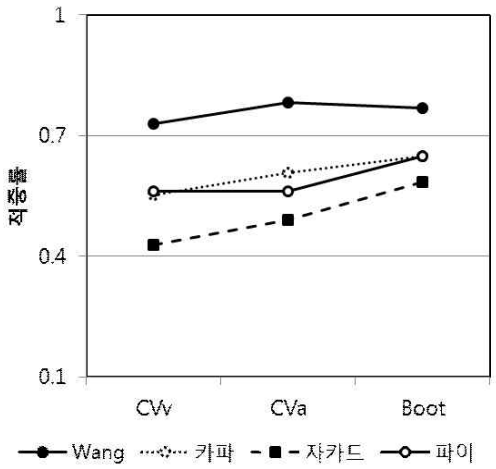
[그림 7] <시나리오 1>자료의 차원(p=2, 5, 7, 10)에 따른 적중률(hit ratio)



(a) k=3



(b) k=4



(c) k=5

[그림 8] <시나리오 2>자료의 군집개수(k=3, 4, 5)에 따른 적중률(hit ratio)

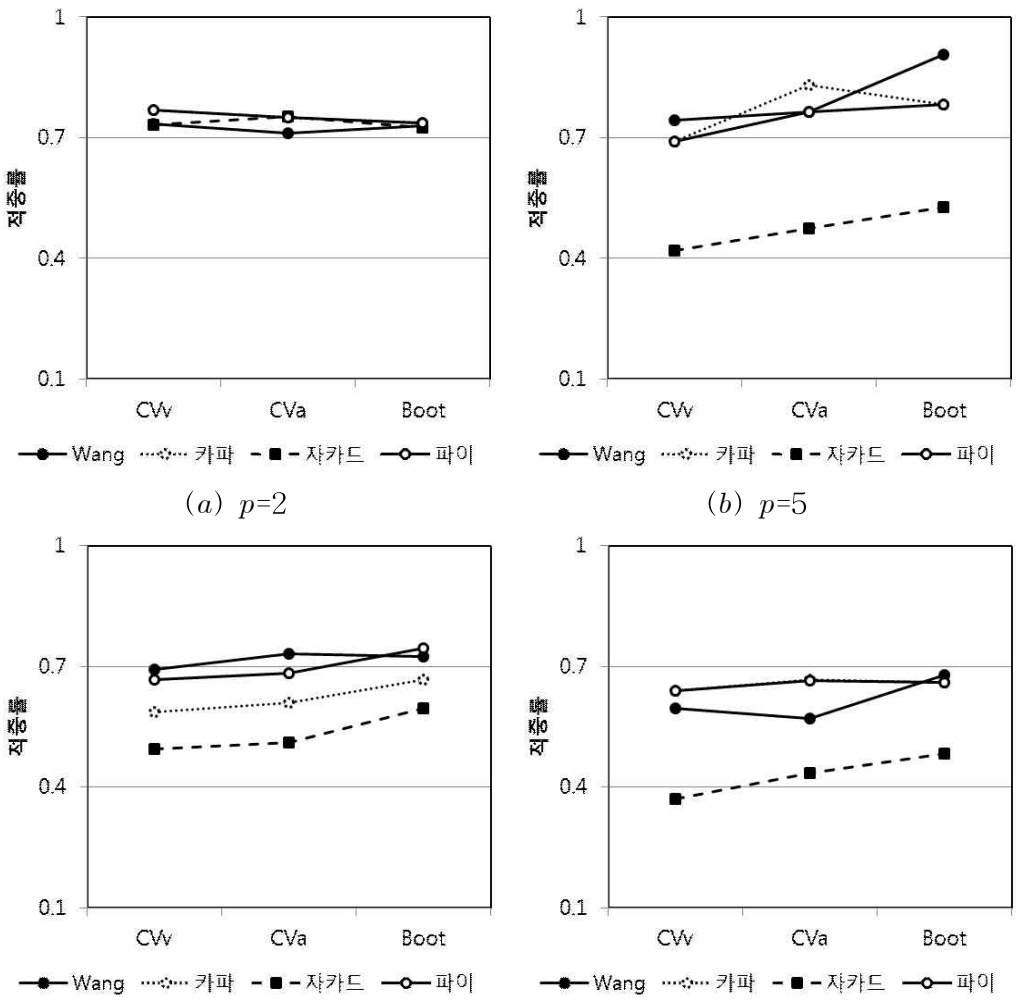
<시나리오 2>의 결과는 각 알고리즘에 따라 [표 13]~[표 16]이며 자료의 실제 군집개수에 따라 각 방법의 적중률을 보면 [그림 8]과 같다. 군집개수(k)가 3인 자료에 적용한 결과는 [그림 8]의 (a)으로, 본 연구에서 제안하는 방법인 카파계수, 카자드계수, 파이계수를 이용한 방법에서 군집개수 적중률이 현저히 높음을 할 수 있다. 다음은 군집개수(k)가 4인 자료에 적용한 결과로 [그림 8]

의 (b)와 같다. 가장 높은 적중률을 보이는 방법은 카파계수와 파이계수를 이용한 방법이며 붓스트랩을 이용한 기존방법에서도 적중률이 높았다.

군집개수(k)가 5인 자료에 적용한 결과는 [그림 8]의 (c)으로 군집의 개수가 3, 4일 때와는 다르게 기존 Wang방법의 적중률이 약간 더 높음을 알 수 있다.

<시나리오 2>에 대한 결과를 자료의 차원별로 살펴보면 [그림 9]와 같다. 먼저, 차원(p)이 2인 자료에서의 결과는 [그림 9]의 (a)로 모든 방법에서 그 적중률이 비슷하지만, 특히 카파계수와 파이계수를 적용한 방법의 적중률이 높음을 확인할 수 있다. 또한, 차원(p)이 5인 자료에 적용했을 때의 결과는 [그림 9]의 (b)와 같다. 기존 Wang방법 중 붓스트랩을 이용한 방법과 카파계수를 이용하는 제안한 방법에서 적중률이 높음을 알 수 있다. 차원(p)이 7인 자료에서의 경우는 [그림 9]의 (c)으로 파이계수를 적용한 방법과 기존 Wang 방법에서 군집개수를 가장 잘 선택함을 알 수 있다. 차원(p)이 10인 자료에 적용한 결과는 [그림 9]의 (d)으로 카파계수와 파이계수를 적용한 방법에서 적중률이 가장 높음을 확인할 수 있다.

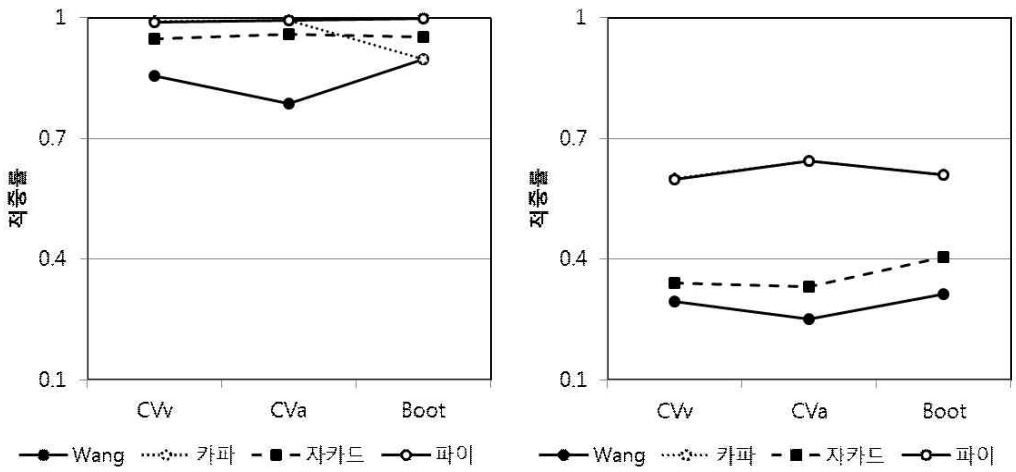
결과적으로 군집개수(k)가 3, 4인 자료에 적용했을 시에 카파계수와 파이계수를 적용한 새로운 알고리즘의 적중률이 높았으며 군집개수(k)가 5인 자료에서는 기존 Wang방법에서 적중률이 가장 높았다. 또한, 차원(p)이 2, 10인 자료에서는 카파계수와 파이계수를 적용하는 제안한 방법에서 적중률이 가장 높았으며 차원(p)이 5인 자료에 적용했을 시, 기존Wang방법과 제안한 카파계수를 적용한 방법의 적중률이 비슷함을 알 수 있다. 또한, 차원(p)이 7인 자료에서는 기존 Wang방법 중 붓스트랩을 이용한 방법에서 적중률이 높았다.



(a) p=2 (b) p=5

(c) p=7 (d) p=10

[그림 9] <시나리오 2>자료의 차원(p=2, 5, 7, 10)에 따른 적중률(hit ratio)

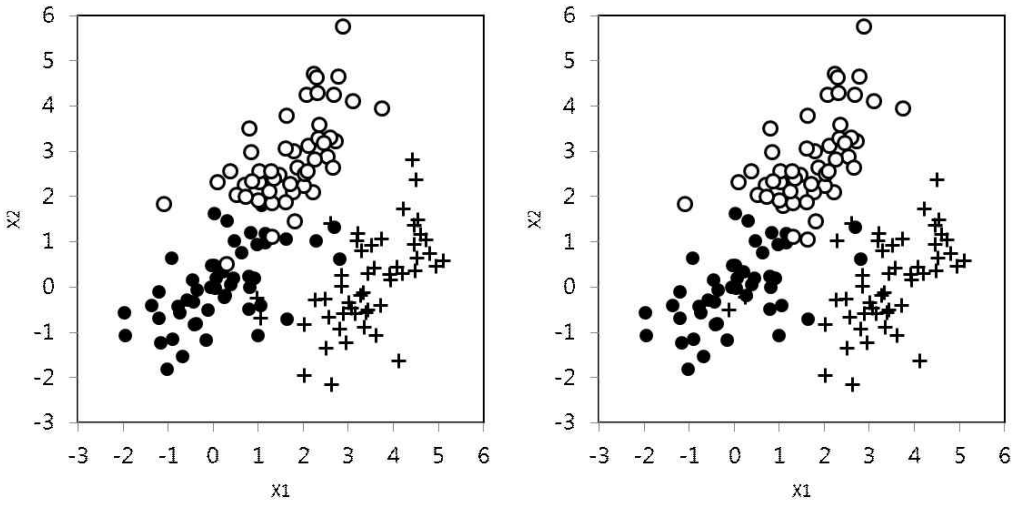


(a) <시나리오 3> (b) <시나리오 4>

[그림 10] <시나리오 3>과 <시나리오 4>에서의 적중률(hit ratio)

<시나리오 3>의 결과는 [표 17], [그림 10]의 (a)와 같다. [그림 10]의 (a)를 살펴보면 자료의 차원에 관계없이 본 연구에서 제안하는 카파계수와 파이계수, 자카드계수를 적용한 방법에서 적중률이 가장 높음을 확인할 수 있다. 다음으로 <시나리오 4>의 결과는 [표 18]과 [그림 10]의 (b)이다. [그림 10]의 (b)를 살펴보면 마찬가지로 카파계수와 파이계수를 적용한 방법에서의 적중률이 기존 Wang방법보다 현저히 높음을 확인 할 수 있다.

다음으로 연관성측도를 이용한 군집결정 방법을 이용하여 최적의 군집개수를 결정한 후 정해진 군집개수를 바탕으로 k-평균 군집분석을 실시하였다. 분석을 실시한 자료는 <시나리오 1>자료에서 군집개수(k)가 3이고 차원(p)은 2, 군집 간의 거리(d)는 3.25, 군집 내 상관계수(ρ)는 0.6인 자료이다. [그림 11]의 (a)는 자료에 관한 산점도이며 [그림 11]의 (b)는 k-평균 군집분석 결과이다. 결과를 살펴보면 k-평균 군집분석 결과와 기존 군집결과가 유사함을 확인할 수 있다.



(a) 산점도(원자료) (b) 군집분석 결과(k-평균 군집분석)

[그림 11] <시나리오 1>의 자료를 이용한 k-평균 군집분석

[표 9] <시나리오 1>에 대한 Wang방법 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0$	CVv	12	110	0	78	9	9	182	0	0	175	16	9	0	0
			CVa	0	13	0	187	0	0	200	0	0	0	127	73	0	0
			Boot	1	39	0	160	0	0	200	0	0	106	24	70	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	183	17	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	20	180	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	147	53	0	0	0
	d=3.00	$\rho=0$	CVv	3	177	0	20	0	0	200	0	0	0	0	200	0	0
			CVa	0	94	0	106	0	0	200	0	0	0	0	200	0	0
			Boot	0	193	0	7	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	180	5	15	0	0	0	0	200	0	0
			CVa	0	200	0	0	18	0	182	0	0	0	0	200	0	0
			Boot	0	200	0	0	169	0	31	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0$	CVv	42	33	2	123	0	0	53	2	145	0	200	0	0	0
			CVa	0	2	0	198	0	0	1	0	199	0	200	0	0	0
			Boot	12	187	0	1	0	0	200	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	27	5	1	5	162	15	162	23	0	0
			CVa	0	200	0	0	0	0	0	0	200	0	188	12	0	0
			Boot	0	200	0	0	0	2	169	0	29	0	119	81	0	0
	d=3.00	$\rho=0$	CVv	1	197	0	2	0	0	200	0	0	172	28	0	0	0
			CVa	0	194	0	6	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	89	111	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	87	113	0	0	0	5	195	0	0
			CVa	0	200	0	0	0	2	198	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	59	141	0	0	0	0	200	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	4	2	194	0	0	45	10	145	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	3	0	197	0	0
$p=7$	d=2.50	$\rho=0$	CVv	0	198	0	2	7	97	36	9	51	0	200	0	0	0
			CVa	0	184	0	16	0	1	11	0	188	0	200	0	0	0
			Boot	0	200	0	0	0	10	190	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	18	0	182	2	7	94	6	91	40	160	0	0	0
			CVa	0	0	0	200	0	0	32	0	168	0	199	0	1	0
			Boot	0	183	0	17	0	0	200	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	185	0	15	0	0	199	1	0	0	200	0	0	0
			CVa	0	122	0	78	0	0	198	2	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	200	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	0	187	13	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	196	4	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	193	7	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0$	CVv	0	13	0	187	0	0	0	0	200	1	194	0	2	3
			CVa	0	1	0	199	0	0	0	0	200	0	189	2	0	9
			Boot	0	200	0	0	0	0	0	0	200	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	5	0	195	199	0	0	0	1	0	200	0	0	0
			CVa	0	0	0	200	6	0	0	0	194	0	200	0	0	0
			Boot	0	174	0	26	145	0	55	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	3	167	30	0	0
			Boot	0	200	0	0	0	0	200	0	0	103	86	11	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	17	168	15	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	182	18	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	188	12	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	3	40	157	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	20	179	1	0
			Boot	0	200	0	0	0	0	200	0	0	0	31	169	0	0

[표 10] <시나리오 1>에 대한 카파(kappa)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0$	CVv	12	188	0	0	64	11	125	0	0	199	1	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	21	179	0	0	0
			Boot	13	187	0	0	1	0	199	0	0	184	15	1	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	70	130	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	198	2	0	0	0
	d=3.00	$\rho=0$	CVv	4	196	0	0	0	0	200	0	0	1	7	192	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	3	197	0	0
			Boot	1	199	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	200	0	0	0	0	0	0	200	0	0
			CVa	0	200	0	0	71	0	129	0	0	0	0	200	0	0
			Boot	0	200	0	0	195	0	5	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0$	CVv	93	107	0	0	1	0	199	0	0	0	200	0	0	0
			CVa	24	176	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	45	155	0	0	0	0	200	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	114	52	34	0	0	13	186	1	0	0
			CVa	0	200	0	0	3	41	156	0	0	0	200	0	0	0
			Boot	0	200	0	0	45	45	110	0	0	0	199	1	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	197	3	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	16	184	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	167	33	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	154	46	0	0	27	75	98	0	0
			CVa	0	200	0	0	0	44	156	0	0	0	80	120	0	0
			Boot	0	200	0	0	0	141	59	0	0	0	3	197	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	82	7	111	0	0	158	8	34	0	0
			CVa	0	200	0	0	1	0	199	0	0	0	0	200	0	0
			Boot	0	200	0	0	34	0	166	0	0	18	0	182	0	0
$p=7$	d=2.50	$\rho=0$	CVv	0	200	0	0	4	182	14	0	0	7	193	0	0	0
			CVa	0	200	0	0	0	192	8	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	116	84	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	9	16	175	0	0	50	150	0	0	0
			CVa	0	200	0	0	0	2	198	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	5	195	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	5	1	194	0	0	0	200	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	200	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	5	195	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	200	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	3	4	193	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0$	CVv	0	200	0	0	29	155	16	0	0	18	182	0	0	0
			CVa	0	200	0	0	0	192	8	0	0	2	198	0	0	0
			Boot	0	200	0	0	9	183	8	0	0	4	196	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	200	0	0	0	0	0	200	0	0	0
			CVa	0	200	0	0	182	18	0	0	0	0	200	0	0	0
			Boot	0	200	0	0	199	1	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	22	178	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	180	20	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	86	114	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	4	196	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	56	117	27	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	191	9	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	174	26	0	0

[표 11] <시나리오 1>에 대한 자카드(Jaccard)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0$	CVv	107	93	0	0	177	5	18	0	0	200	0	0	0	0
			CVa	57	143	0	0	46	0	154	0	0	194	6	0	0	0
			Boot	171	29	0	0	139	0	61	0	0	200	0	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	0	1	199	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	200	0	0	0	0
	d=3.00	$\rho=0$	CVv	90	110	0	0	0	0	200	0	0	20	23	157	0	0
			CVa	79	121	0	0	0	0	200	0	0	0	16	184	0	0
			Boot	94	106	0	0	0	0	200	0	0	0	1	199	0	0
	d=3.25	$\rho=0$	CVv	1	199	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	200	0	0	0	0	0	0	200	0	0
			CVa	0	200	0	0	199	0	1	0	0	0	0	200	0	0
			Boot	0	200	0	0	200	0	0	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0$	CVv	199	1	0	0	59	10	131	0	0	2	198	0	0	0
			CVa	200	0	0	0	100	0	100	0	0	0	200	0	0	0
			Boot	200	0	0	0	0	3	197	0	0	1	199	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	195	5	0	0	0	97	103	0	0	0
			CVa	0	200	0	0	197	3	0	0	0	7	193	0	0	0
			Boot	0	200	0	0	193	7	0	0	0	37	163	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	1	0	199	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	188	12	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	200	0	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	183	17	0	0	84	83	33	0	0
			CVa	0	200	0	0	0	160	40	0	0	0	164	36	0	0
			Boot	0	200	0	0	0	192	8	0	0	0	21	179	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	199	0	1	0	0	200	0	0	0	0
			CVa	0	200	0	0	155	0	45	0	0	58	0	142	0	0
			Boot	0	200	0	0	195	0	5	0	0	160	0	40	0	0
$p=7$	d=2.50	$\rho=0$	CVv	0	200	0	0	14	186	0	0	0	92	108	0	0	0
			CVa	0	200	0	0	13	187	0	0	0	3	197	0	0	0
			Boot	0	200	0	0	22	168	10	0	0	6	194	0	0	0
	d=2.75	$\rho=0$	CVv	82	118	0	0	106	54	40	0	0	143	57	0	0	0
			CVa	124	76	0	0	125	50	25	0	0	59	141	0	0	0
			Boot	39	161	0	0	68	64	68	0	0	86	114	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	77	16	107	0	0	13	187	0	0	0
			CVa	0	200	0	0	4	0	196	0	0	1	199	0	0	0
			Boot	0	200	0	0	4	0	196	0	0	0	200	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	7	1	192	0	0	73	127	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	1	199	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	7	193	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	136	18	46	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	27	173	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0$	CVv	2	198	0	0	186	14	0	0	0	116	84	0	0	0
			CVa	200	0	0	0	199	1	0	0	0	88	112	0	0	0
			Boot	200	0	0	0	184	16	0	0	0	64	136	0	0	0
	d=2.75	$\rho=0$	CVv	135	65	0	0	200	0	0	0	0	3	197	0	0	0
			CVa	178	22	0	0	200	0	0	0	0	0	200	0	0	0
			Boot	147	53	0	0	200	0	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	8	8	184	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	200	0	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	11	5	184	0	0	193	7	0	0	0
			CVa	0	200	0	0	3	0	197	0	0	40	160	0	0	0
			Boot	0	200	0	0	1	0	199	0	0	104	96	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	1	0	199	0	0	157	43	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	9	182	9	0	0

[표 12] <시나리오 1>에 대한 파이(phi)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0$	CVv	12	188	0	0	65	7	128	0	0	199	1	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	21	179	0	0	0
			Boot	13	187	0	0	1	0	199	0	0	184	15	1	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	70	130	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	198	2	0	0	0
	d=3.00	$\rho=0$	CVv	4	196	0	0	0	0	200	0	0	1	7	192	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	3	197	0	0
			Boot	1	199	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	200	0	0	0	0	0	0	200	0	0
			CVa	0	200	0	0	71	0	129	0	0	0	0	200	0	0
			Boot	0	200	0	0	195	0	5	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0$	CVv	93	107	0	0	1	0	199	0	0	0	200	0	0	0
			CVa	24	176	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	46	154	0	0	0	0	200	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	112	55	34	0	0	13	185	1	0	0
			CVa	0	200	0	0	3	41	156	0	0	0	200	0	0	0
			Boot	0	200	0	0	45	45	110	0	0	0	199	1	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	197	3	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	16	184	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	167	33	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	0	155	45	0	0	27	76	97	0	0
			CVa	0	200	0	0	0	44	156	0	0	0	79	121	0	0
			Boot	0	200	0	0	0	141	59	0	0	0	3	197	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	83	6	111	0	0	158	8	34	0	0
			CVa	0	200	0	0	1	0	199	0	0	0	0	200	0	0
			Boot	0	200	0	0	34	0	166	0	0	18	0	182	0	0
$p=7$	d=2.50	$\rho=0$	CVv	0	200	0	0	4	184	12	0	0	7	193	0	0	0
			CVa	0	200	0	0	0	188	12	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	115	85	0	0	0	200	0	0	0
	d=2.75	$\rho=0$	CVv	0	200	0	0	7	21	172	0	0	50	150	0	0	0
			CVa	0	200	0	0	0	3	197	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	5	195	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	5	2	193	0	0	1	199	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	200	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	1	0	199	0	0	6	194	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	200	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	1	1	198	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0$	CVv	0	200	0	0	23	169	8	0	0	9	191	0	0	0
			CVa	0	200	0	0	0	190	10	0	0	3	197	0	0	0
			Boot	0	200	0	0	8	184	8	0	0	6	194	0	0	0
	d=2.75	$\rho=0$	CVv	1	199	0	0	200	0	0	0	0	0	200	0	0	0
			CVa	0	200	0	0	188	0	12	0	0	0	200	0	0	0
			Boot	0	200	0	0	200	0	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	200	0	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	26	174	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	181	19	0	0	0
	d=3.25	$\rho=0$	CVv	0	200	0	0	1	0	199	0	0	91	109	0	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	1	199	0	0	0
	d=3.50	$\rho=0$	CVv	0	200	0	0	0	0	200	0	0	69	102	29	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	190	10	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	157	43	0	0

[표 13] <시나리오 2>에 대한 Wang방법 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0.6$	CVv	1	149	0	50	0	0	0	0	200	0	21	178	0	1
			CVa	0	64	0	136	0	0	0	0	200	0	2	198	0	0
			Boot	0	198	0	2	0	0	0	0	200	0	13	187	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	0	7	6	61	126	5	159	35	1	0
			CVa	0	200	0	0	0	0	0	28	172	0	122	72	4	2
			Boot	0	200	0	0	0	0	0	93	107	0	193	0	7	0
	d=3.00	$\rho=0.6$	CVv	0	50	9	141	0	0	200	0	0	0	0	194	3	3
			CVa	0	9	0	191	0	0	200	0	0	0	0	194	0	6
			Boot	0	3	0	197	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	0	192	0	8	0	0	200	0	0
			CVa	0	200	0	0	0	0	198	0	2	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0.6$	CVv	7	85	1	107	3	2	5	4	186	0	48	151	1	0
			CVa	0	2	0	198	0	0	0	0	200	0	12	187	1	0
			Boot	0	198	0	2	0	1	189	0	10	0	3	197	0	0
	d=2.75	$\rho=0.6$	CVv	1	199	0	0	0	152	45	0	3	1	83	116	0	0
			CVa	0	200	0	0	0	2	130	1	67	0	72	128	0	0
			Boot	1	199	0	0	0	59	141	0	0	0	97	103	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	0	0	198	0	2	5	3	192	0	0
			CVa	0	200	0	0	0	0	194	0	6	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	1	0	199	0	0	0	161	36	2	1
			CVa	0	200	0	0	0	0	200	0	0	0	147	52	0	1
			Boot	0	200	0	0	0	0	200	0	0	0	109	91	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=7$	d=2.50	$\rho=0.6$	CVv	92	0	0	108	0	0	0	0	200	2	103	84	5	6
			CVa	15	0	0	185	0	0	0	0	200	0	67	121	5	7
			Boot	198	0	0	2	0	0	0	0	200	0	101	99	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	0	114	86	0	0	0	2	197	1	0
			CVa	0	199	0	1	0	9	186	4	1	0	1	198	1	0
			Boot	0	200	0	0	0	2	198	0	0	0	0	200	0	0
	d=3.00	$\rho=0.6$	CVv	0	189	0	11	0	6	194	0	0	2	55	142	1	0
			CVa	0	118	0	82	0	0	200	0	0	0	18	182	0	0
			Boot	0	200	0	0	0	2	198	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	200	0	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	199	0	0	1	0	0	200	0	0
			Boot	0	200	0	0	0	200	0	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	187	13	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	189	9	2	0	0	200	0	0	0	0	200	0	0
			Boot	0	75	125	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0.6$	CVv	17	0	0	183	1	109	1	14	75	1	192	0	2	5
			CVa	1	0	0	199	0	1	0	0	199	0	194	0	0	6
			Boot	197	0	0	3	0	164	36	0	0	0	200	0	0	0
	d=2.75	$\rho=0.6$	CVv	0	25	0	175	52	0	0	9	139	0	200	0	0	0
			CVa	0	0	0	200	1	0	0	0	199	0	200	0	0	0
			Boot	0	200	0	0	193	0	3	0	4	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	13	179	0	8	0	0	200	0	0	1	0	193	2	4
			CVa	0	129	1	70	0	0	200	0	0	0	0	194	6	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	11	186	2	1	0	0	200	0	0
			CVa	0	200	0	0	0	0	191	0	9	0	0	200	0	0
			Boot	0	200	0	0	0	2	198	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0

[표 14] <시나리오 2>에 대한 카파(kappa)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0.6$	CVv	0	200	0	0	192	7	0	1	0	1	109	90	0	0
			CVa	0	200	0	0	198	0	0	0	2	0	131	69	0	0
			Boot	0	200	0	0	195	0	0	0	5	0	124	76	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	1	59	53	70	17	4	188	8	0	0
			CVa	0	200	0	0	0	0	32	152	16	0	197	3	0	0
			Boot	0	200	0	0	0	0	6	173	21	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	0	160	8	32	0	0	200	0	0	0	7	193	0	0
			CVa	0	151	0	49	0	0	200	0	0	0	4	196	0	0
			Boot	0	133	0	67	0	0	200	0	0	0	7	193	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0.6$	CVv	38	162	0	0	41	58	101	0	0	1	180	19	0	0
			CVa	17	183	0	0	14	8	178	0	0	0	196	4	0	0
			Boot	0	200	0	0	13	36	151	0	0	0	75	125	0	0
	d=2.75	$\rho=0.6$	CVv	3	197	0	0	0	155	45	0	0	17	164	19	0	0
			CVa	3	197	0	0	0	78	122	0	0	0	196	4	0	0
			Boot	3	197	0	0	0	143	57	0	0	2	189	9	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	45	11	144	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	15	5	180	0	0	1	196	3	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	195	5	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=7$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	119	74	7	0	0	5	195	0	0	0
			CVa	200	0	0	0	48	123	29	0	0	0	200	0	0	0
			Boot	200	0	0	0	107	76	17	0	0	4	196	0	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	0	186	14	0	0	0	42	158	0	0
			CVa	0	200	0	0	0	150	50	0	0	0	49	151	0	0
			Boot	0	200	0	0	0	129	71	0	0	0	2	198	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	0	20	180	0	0	11	186	3	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	45	155	0	0	0	15	185	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	200	0	0	0	0	1	199	0	0
			CVa	0	200	0	0	0	200	0	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	200	0	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	178	22	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	27	172	1	0	0	0	200	0	0	0
			CVa	200	0	0	0	15	184	1	0	0	0	200	0	0	0
			Boot	200	0	0	0	0	200	0	0	0	0	200	0	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	200	0	0	0	0	0	200	0	0	0
			CVa	0	200	0	0	200	0	0	0	0	0	200	0	0	0
			Boot	0	200	0	0	200	0	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	15	185	0	0	7	2	191	0	0	11	13	176	0	0
			CVa	1	199	0	0	0	0	200	0	0	0	2	198	0	0
			Boot	3	197	0	0	6	0	194	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	33	167	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	8	192	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0

[표 15] <시나리오 2>에 대한 자카드(Jaccard)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0.6$	CVv	27	173	0	0	200	0	0	0	0	10	154	36	0	0
			CVa	10	190	0	0	200	0	0	0	0	0	190	10	0	0
			Boot	43	157	0	0	200	0	0	0	0	0	163	37	0	0
	d=2.75	$\rho=0.6$	CVv	4	196	0	0	20	112	44	23	1	37	162	1	0	0
			CVa	0	200	0	0	1	22	88	89	0	0	200	0	0	0
			Boot	5	195	0	0	0	11	14	170	5	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	0	195	1	4	0	0	200	0	0	18	32	150	0	0
			CVa	0	198	0	2	0	0	200	0	0	0	31	169	0	0
			Boot	0	195	0	5	0	0	200	0	0	0	17	183	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	2	1	197	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	1	199	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	5	195	0	0
$p=5$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	200	0	0	0	0	7	191	2	0	0
			CVa	200	0	0	0	200	0	0	0	0	0	200	0	0	0
			Boot	138	62	0	0	192	8	0	0	0	0	141	59	0	0
	d=2.75	$\rho=0.6$	CVv	158	42	0	0	32	168	0	0	0	141	59	0	0	0
			CVa	182	18	0	0	39	160	1	0	0	8	192	0	0	0
			Boot	144	56	0	0	6	191	3	0	0	71	127	2	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	52	8	140	0	0	147	12	41	0	0
			CVa	0	200	0	0	40	0	160	0	0	18	0	182	0	0
			Boot	0	200	0	0	8	0	192	0	0	59	0	141	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	160	7	33	0	0	14	186	0	0	0
			CVa	0	200	0	0	140	0	60	0	0	0	200	0	0	0
			Boot	0	200	0	0	138	0	62	0	0	0	199	1	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=7$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	200	0	0	0	0	48	152	0	0	0
			CVa	200	0	0	0	200	0	0	0	0	7	193	0	0	0
			Boot	200	0	0	0	200	0	0	0	0	53	147	0	0	0
	d=2.75	$\rho=0.6$	CVv	60	140	0	0	1	199	0	0	0	8	86	106	0	0
			CVa	60	140	0	0	1	198	1	0	0	0	131	69	0	0
			Boot	19	181	0	0	0	190	10	0	0	0	5	195	0	0
	d=3.00	$\rho=0.6$	CVv	33	167	0	0	0	51	149	0	0	72	128	0	0	0
			CVa	43	157	0	0	0	12	188	0	0	3	197	0	0	0
			Boot	0	200	0	0	0	136	64	0	0	1	62	137	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	200	0	0	0	28	30	142	0	0
			CVa	0	200	0	0	1	199	0	0	0	0	22	178	0	0
			Boot	0	200	0	0	0	200	0	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	11	9	180	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	1	199	0	0
			Boot	0	197	3	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	198	2	0	0	0	44	156	0	0	0
			CVa	200	0	0	0	200	0	0	0	0	36	164	0	0	0
			Boot	200	0	0	0	98	102	0	0	0	43	157	0	0	0
	d=2.75	$\rho=0.6$	CVv	165	35	0	0	200	0	0	0	0	7	193	0	0	0
			CVa	171	29	0	0	200	0	0	0	0	1	199	0	0	0
			Boot	0	200	0	0	200	0	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	199	1	0	0	185	1	14	0	0	84	25	91	0	0
			CVa	199	1	0	0	190	0	10	0	0	10	39	151	0	0
			Boot	196	4	0	0	193	0	7	0	0	3	10	187	0	0
	d=3.25	$\rho=0.6$	CVv	3	197	0	0	54	84	62	0	0	0	0	200	0	0
			CVa	0	200	0	0	43	40	117	0	0	0	0	200	0	0
			Boot	1	199	0	0	7	114	79	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	1	199	0	0	39	13	148	0	0	28	8	164	0	0
			CVa	0	200	0	0	7	0	193	0	0	0	0	200	0	0
			Boot	19	181	0	0	11	0	189	0	0	0	0	200	0	0

[표 16] <시나리오 2>에 대한 파이(phi)계수 적용 결과

(k_0=군집개수, d=군집 간거리, ρ=상관계수, p=차원)

	추정된 군집 개수			k_0 = 3				k_0 = 4					k_0 = 5				
				2	3	4	≥5	2	3	4	5	≥6	≤3	4	5	6	≥7
$p=2$	d=2.50	$\rho=0.6$	CVv	0	200	0	0	192	7	0	1	0	1	111	88	0	0
			CVa	0	200	0	0	198	0	0	0	2	0	131	69	0	0
			Boot	0	200	0	0	195	0	0	0	5	0	125	75	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	1	59	54	69	17	4	188	8	0	0
			CVa	0	200	0	0	0	0	30	154	16	0	197	3	0	0
			Boot	0	200	0	0	0	0	6	173	21	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	0	160	8	32	0	0	200	0	0	0	6	194	0	0
			CVa	0	151	0	49	0	0	200	0	0	0	4	196	0	0
			Boot	0	131	0	69	0	0	200	0	0	0	7	193	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=5$	d=2.50	$\rho=0.6$	CVv	36	164	0	0	42	59	99	0	0	1	180	19	0	0
			CVa	17	183	0	0	14	8	178	0	0	0	196	4	0	0
			Boot	0	200	0	0	13	36	151	0	0	0	75	125	0	0
	d=2.75	$\rho=0.6$	CVv	4	196	0	0	0	155	45	0	0	17	163	20	0	0
			CVa	3	197	0	0	0	78	122	0	0	0	196	4	0	0
			Boot	3	197	0	0	0	143	57	0	0	2	189	9	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	45	11	144	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	15	5	180	0	0	1	195	4	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	200	0	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	195	5	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
$p=7$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	116	70	14	0	0	6	194	0	0	0
			CVa	200	0	0	0	45	126	29	0	0	0	200	0	0	0
			Boot	200	0	0	0	119	63	18	0	0	4	196	0	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	0	182	18	0	0	0	42	158	0	0
			CVa	0	200	0	0	0	146	54	0	0	0	49	151	0	0
			Boot	0	200	0	0	0	112	88	0	0	0	2	198	0	0
	d=3.00	$\rho=0.6$	CVv	0	200	0	0	0	19	181	0	0	12	156	32	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	183	17	0	0
			Boot	0	200	0	0	0	33	167	0	0	0	14	186	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	178	22	0	0	0	200	0	0	0	0	200	0	0
$p=10$	d=2.50	$\rho=0.6$	CVv	200	0	0	0	42	158	0	0	0	0	200	0	0	0
			CVa	200	0	0	0	17	183	0	0	0	0	200	0	0	0
			Boot	200	0	0	0	0	200	0	0	0	0	200	0	0	0
	d=2.75	$\rho=0.6$	CVv	0	200	0	0	200	0	0	0	0	0	200	0	0	0
			CVa	0	200	0	0	200	0	0	0	0	0	200	0	0	0
			Boot	0	200	0	0	200	0	0	0	0	0	200	0	0	0
	d=3.00	$\rho=0.6$	CVv	20	180	0	0	8	5	187	0	0	11	9	180	0	0
			CVa	6	194	0	0	0	0	200	0	0	0	1	199	0	0
			Boot	6	194	0	0	5	0	195	0	0	0	0	200	0	0
	d=3.25	$\rho=0.6$	CVv	0	200	0	0	2	28	170	0	0	0	0	200	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	9	191	0	0	0	0	200	0	0
	d=3.50	$\rho=0.6$	CVv	0	200	0	0	1	0	199	0	0	1	0	199	0	0
			CVa	0	200	0	0	0	0	200	0	0	0	0	200	0	0
			Boot	0	200	0	0	0	0	200	0	0	0	0	200	0	0

[표 17] <시나리오 3>에 대한 결과

(k_0=군집개수, d=군집 간거리, p=차원)

			k_0 = 3															
	추정된 군집개수		Wang				카파계수				자카드계수				파이계수			
			2	3	4	≥5	2	3	4	≥5	2	3	4	≥5	2	3	4	≥5
$p=2$	d=2.00	CVv	0	191	2	7	0	198	2	0	0	200	0	0	0	196	4	0
		CVa	0	161	19	20	0	198	2	0	0	200	0	0	0	200	0	0
		Boot	0	196	4	0	0	0	200	0	0	200	0	0	0	200	0	0
	d=2.25	CVv	7	193	0	0	8	192	0	0	80	120	0	0	13	187	0	0
		CVa	0	199	0	1	0	200	0	0	69	131	0	0	2	198	0	0
		Boot	3	197	0	0	8	192	0	0	96	104	0	0	5	195	0	0
	d=2.50	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
	d=2.75	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
	d=3.00	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
$p=5$	d=2.00	CVv	0	129	0	71	0	200	0	0	1	199	0	0	0	200	0	0
		CVa	0	15	0	185	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
	d=2.25	CVv	0	0	0	200	0	194	5	1	23	177	0	0	0	196	3	1
		CVa	0	0	0	200	0	187	9	4	12	188	0	0	0	190	4	6
		Boot	0	0	0	200	0	200	0	0	0	200	0	0	0	200	0	0
	d=2.50	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
	d=2.75	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
	d=3.00	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0

[표 18] <시나리오 4>에 대한 결과

(k_0=군집개수, d=군집 간거리, p=차원)

			k_0 = 3															
	추정된 군집개수		Wang				카파계수				자카드계수				파이계수			
			2	3	4	≥5	2	3	4	≥5	2	3	4	≥5	2	3	4	≥5
$p=2$	d=2.00	CVv	3	0	0	197	195	3	1	1	200	0	0	0	199	1	0	0
		CVa	0	0	0	200	161	0	7	32	200	0	0	0	146	0	4	50
		Boot	1	0	0	199	199	0	0	1	200	0	0	0	199	0	0	1
	d=2.25	CVv	2	0	0	198	198	0	0	2	200	0	0	0	199	0	1	0
		CVa	0	0	0	200	169	0	2	29	200	0	0	0	167	0	3	30
		Boot	0	0	0	200	198	0	1	1	200	0	0	0	198	0	1	1
	d=2.50	CVv	19	26	0	155	63	137	0	0	192	8	0	0	65	135	0	0
		CVa	0	0	0	200	0	200	0	0	197	3	0	0	3	197	0	0
		Boot	2	34	0	164	62	138	0	0	196	4	0	0	65	135	0	0
	d=2.75	CVv	0	6	1	193	0	171	9	20	1	198	0	1	0	163	5	32
		CVa	0	0	0	200	0	169	0	31	0	200	0	0	0	175	0	25
		Boot	0	0	0	200	0	121	0	79	0	193	0	7	0	120	0	80
	d=3.00	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
$p=5$	d=2.00	CVv	0	0	0	200	13	140	23	24	200	0	0	0	20	134	18	28
		CVa	0	0	0	200	0	108	34	58	200	0	0	0	1	111	20	68
		Boot	0	0	0	200	1	197	2	0	195	5	0	0	1	199	0	0
	d=2.25	CVv	0	0	0	200	0	200	0	0	147	53	0	0	0	200	0	0
		CVa	0	0	0	200	0	200	0	0	149	51	0	0	0	200	0	0
		Boot	0	4	0	196	0	200	0	0	4	196	0	0	0	200	0	0
	d=2.50	CVv	38	155	1	6	51	149	0	0	178	22	0	0	41	159	0	0
		CVa	0	102	0	98	5	195	0	0	192	8	0	0	5	195	0	0
		Boot	13	187	0	0	38	162	0	0	187	13	0	0	37	163	0	0
	d=2.75	CVv	66	0	0	134	198	2	0	0	200	0	0	0	196	4	0	0
		CVa	0	0	0	200	185	15	0	0	200	0	0	0	188	11	1	0
		Boot	108	0	0	92	198	2	0	0	200	0	0	0	200	0	0	0
	d=3.00	CVv	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		CVa	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0
		Boot	0	200	0	0	0	200	0	0	0	200	0	0	0	200	0	0

제 5 장 실제 데이터의 적용

다음으로 군집개수 결정 방법을 실제 자료에 적용해봄으로써 기존의 방법과 제안한 방법의 수행능력을 평가하고자 한다.

분석에 사용된 데이터는 1975년 미국의 22개의 공공기업에 대한 8개의 경제지표 자료이다. 자료는 [표 19]와 같다.

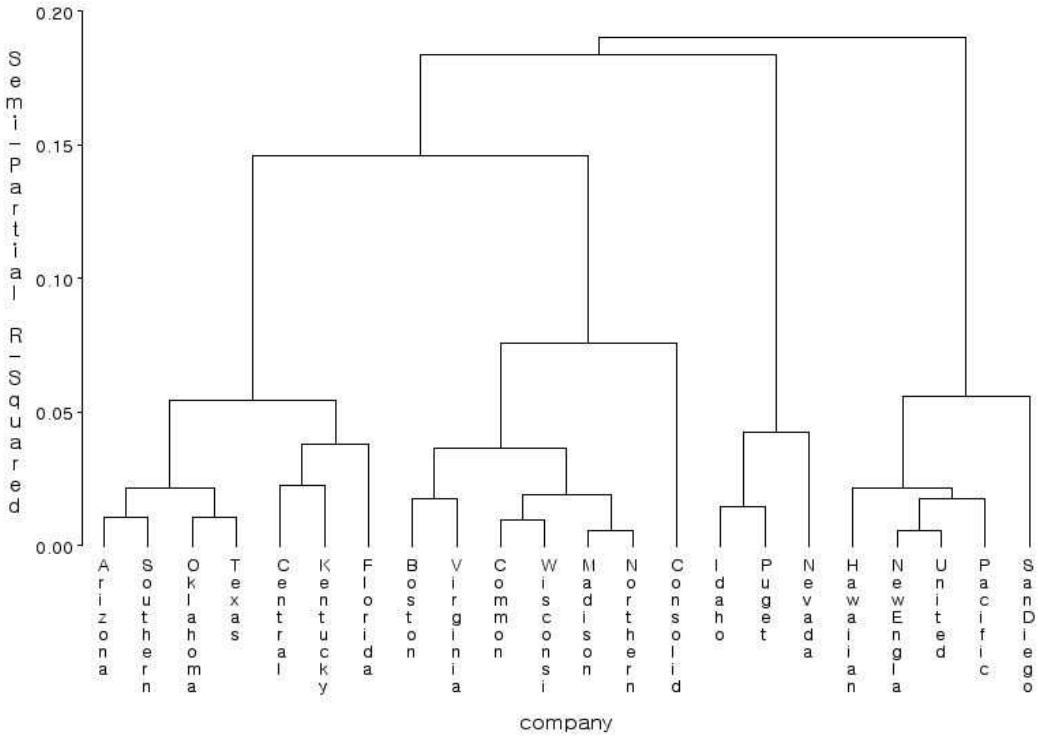
[표 19] 미국의 22개의 공공기업의 경제지표 자료(1975).

obs	Company	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	Arizona	1.06	9.2	151	54.4	1.6	9077	0	0.628
2	Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central	1.43	15.4	113	53	3.4	9212	0	1.058
4	Commonwealth	1.02	11.2	168	56	0.3	6423	34.3	0.7
5	Consolidated	1.49	8.8	192	51.2	1	3300	15.6	2.044
6	Florida	1.32	13.5	111	60	-2.2	11127	22.5	1.241
7	Hawaiian	1.22	12.2	175	67.6	2.2	7642	0	1.652
8	Idaho	1.1	9.2	245	57	3.3	13082	0	0.309
9	Kentucky	1.34	13	168	60.4	7.2	8406	0	0.862
10	Madison	1.12	12.4	197	53	2.7	6455	39.2	0.623
11	Nevada	0.75	7.5	173	51.5	6.5	17441	0	0.768
12	New England	1.13	10.9	178	62	3.7	6154	0	1.897
13	Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma	1.09	12	96	49.8	1.4	9673	0	0.588
15	Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
16	Puget	1.16	9.9	252	56	9.2	15991	0	0.62
17	SanDiego	0.76	6.4	136	61.9	9	5714	8.3	1.92
18	Southern	1.05	12.6	150	56.7	2.7	10140	0	1.108
19	Texas	1.16	11.7	104	54	-2.1	13507	0	0.636
20	Wisconsi	1.2	11.8	148	59.9	3.5	7287	41.1	0.702
21	United	1.04	8.6	204	61	3.5	6650	0	2.116
22	Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

X_1 : Fixed-charge coverage ratio(income/debt).
X_2 : Rated of return on capital.
X_3 : Cost per KW capacity in place.
X_4 : Annual load factor.
X_5 : Peak kWh demand growth from 1974 to 1975.
X_6 : Sales (kWh use per year).
X_7 : Percent nuclear.
X_8 : Total fuel costs(cents per kWh).

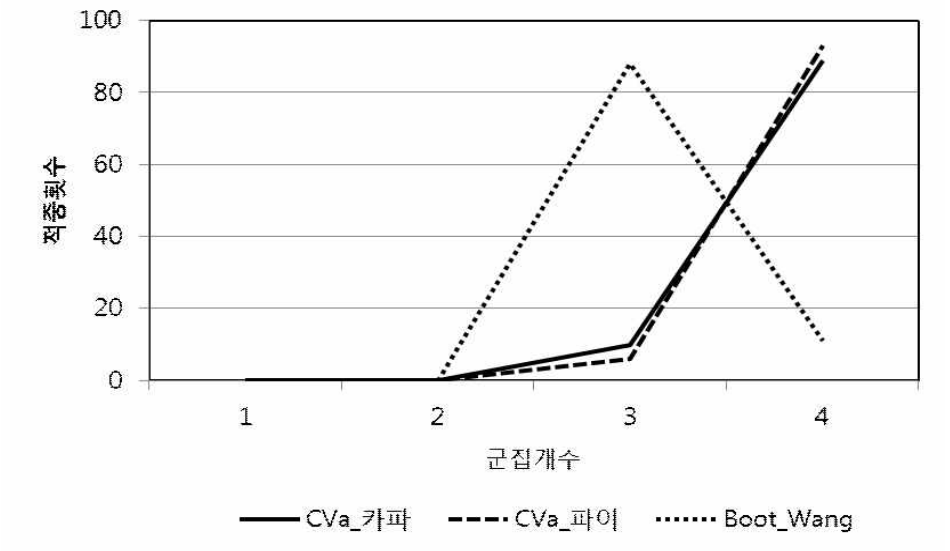
먼저, 계층적 군집화 결과인 덴드로그림은 [그림 12]이다. 또한, 통상적으로 군집개수를 결정하는데 가장 널리 사용되는 삼차군집기준값(CCC, cubic clustering criterion)을 살펴보면 [그림 13]의 (a)와 같다. 삼차군집기준값이 군집개수가 4일 때에 가장 높으므로 최적의 군집개수는 4이다. 자료에 제안한 방법을 적용한 결과는 [그림 13]의 (b)으로 99번의 시행중에서 최적의 군집개수를 출력한 경우의 수를 나타내었다.

기존 Wang방법 중 붓스트랩 방법에서는 최적의 군집개수가 3으로 나타났다. 반면에 평균을 이용한 교차타당성방법에 파이계수와 카파계수를 적용한 알고리즘에서는 최적의 군집개수가 4로 나타났다. 결과적으로 연관성측도를 이용한 군집개수 결정방법의 결과와 CCC에 의한 결과가 일치함을 확인 할 수 있다.



[그림 12] 계층적 군집분석 방법에 의한 덴드로그램

(a) 실제자료의 삼차군집기준값



(b) 군집개수 결정 방법 적용 결과

[그림 13] 첫 번째 실제자료의 CCC와 군집개수 결정 방법 적용 결과

최적의 군집개수는 4로 결정되었으므로 정해진 군집개수를 바탕으로 계층적 군집분석과 k-평균 군집분석을 실시하였다. 결과를 [표 20]과 같으며 Boston, Consolidated의 군집을 제외하고는 군집화 결과가 일치한다.

[표 20] 계층적 군집화 방법과 k-평균 군집분석 결과

	계층적 군집화 방법	k-평균 군집분석
군집1	Commonwealth, Madison, Northern, Virginia Wisconsi, Boston, Consolidated	Commonwealth, Madison, Northern, Virginia Wisconsi,
군집2	Arizona, Central, Florida, Kentucky, Oklahoma, Southern, Texas	Arizona, Central, Florida, Kentucky, Oklahoma, Southern, Texas
군집3	Idaho, Nevada, Puget,	Idaho, Nevada, Puget
군집4	Hawaiian, New England, Pacific, SanDiego United,	Hawaiian, New England, Pacific, SanDiego United, Boston, Consolidated,

제 6 장 결 론

본 논문에서는 군집화 불안정성을 이용한 기존의 군집개수를 결정하는 알고리즘을 보완하여 이항자료에서의 연관성측도를 적용한 새로운 군집개수 결정 알고리즘을 제안하였다. 기존의 군집개수 결정 알고리즘은 군집화 $\psi_1(\boldsymbol{x})$와 $\psi_2(\boldsymbol{x})$의 단순 거리를 이용하여 군집화 불안정성을 측정하였으나 본 연구에서는 군집화 거리에서 $I\{\psi_1(\boldsymbol{x}^0)=\psi_1(\boldsymbol{y}^0)\}$과 $I\{\psi_2(\boldsymbol{x}^0)=\psi_2(\boldsymbol{y}^0)\}$가 이항자료로 나타나는 특성을 이용하여 이항자료에서의 연관성측도인 카파계수, 자카드계수, 파이계수를 적용하여 군집화 불안정성을 측정하였다.

기존의 군집개수 결정 방법과 본 연구에서 제안한 새로운 방법을 모의실험과 실제자료에 적용시킨 결과 연관성측도를 적용한 알고리즘에서 그 적중률이 높았다. 특히, 모의실험 시 카파계수와 파이계수를 적용한 알고리즘에서 적중률이 현저히 높아 제안한 방법의 우수성을 확인할 수 있었다. 모의실험의 주요 결과를 살펴보면 아래와 같다.

첫째, 군집개수가 3, 4인 자료일 때 카파계수와 파이계수를 적용한 방법이 더 좋은 성능을 가지는 것을 확인 할 수 있다.

둘째, 자료의 차원이 2일 때에는 카파계수와 파이계수를 적용했을 때의 적중률이 높으며, 차원이 3이상 일 때에는 기존 Wang방법 중 붓스트랩 방법과 제안한 방법에서 큰 차이가 없었다.

셋째, 서로 다른 분산을 가지는 군집에 대한 자료에서는 제안한 카파계수, 파이계수, 자카드계수를 적용한 방법의 적중률이 현저히 높았다.

넷째, 서로 다른 상관계수를 가지는 군집에 대한 자료에서는 카파계수와 파이계수를 적용할 때의 방법이 더 좋은 성능을 보였다.

다섯째, 자료의 군집개수가 5일 때에는 제안한 방법의 적중률이 약간 낮음을 확인할 수 있었다.

또한, 모의실험에서 군집개수가 5이상 일 때에는 기존 Wang방법의 적중률이 높았으며 차원이 10이상 일 때는 Wang방법과 제안한 방법 모두에서 적중률이 전체적으로 낮아짐을 확인할 수 있었다. 그리고, 실제자료를 통해서 제안한 방법의 결과와 삼차군집기준값을 이용할 때의 결과가 일치함을 확인할 수 있었다.

결과적으로 자료의 차원과 군집개수가 증가할수록 기존방법과 제안한 방법의 적중률의 차이는 줄어들지만 본 연구에서 제안한 방법이 다양한 자료에서 기존 Wang방법보다 군집개수 선택에 있어서 우수하다.

추후에는 군집화를 정의함에 있어서 k-평균 군집분석방법 이외에 계층적 군집분석방법, 잠재군집분석(latent class analysis)방법 등을 활용하여 군집개수 결정 알고리즘을 제안하고자 한다.

참 고 문 헌

[1] Ben-David, S., von Luxburg, U., Pal, D. (2006). A sober look at stability of clustering. *19th Annual Conference on Learning Theory*.

[2] Caliniski, R. B., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*. 3, 1-27.

[3] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20 (1), 37-46.

[4] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*. 70 (4), 213-220.

[5] Fang, J., Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computation Statistics and Data Analysis*. 56, 468-477.

[6] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

[7] Jaccard P. (1902). Lois de distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*. 38, 67-130.

[8] Jaccard P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*. 11(2), 37-50.

[9] Krieger, A.W., Green, P.E., (1999). A cautionary note on using internal crossvalidation to select the number of clusters. *Psychometrika*. 64, 341 - 353.

[10] Krzanowski, W. J., Lai, Y. T. (1985). A criterion for determining the number of clusters in a data set. *Biometrics*. 44, 23 - 34.

[11] Pearson, K. (1900). Mathematical contributions to the theory of

evolution. VII. *On the correlation of characters not quantitatively measurable*. Philosophical Transactions of the Royal Society of London, Series A, 195, 1 – 47.

[12] Steinley, D. (2008). Stability analysis in K-means clustering. *British Journal of Mathematical and Statistical Psychology*. 61, 255 – 273.

[13] Sugar, C., James, G. (2003). Finding the number of clusters in a data set: an information theoretic approach. *Journal of American Statistical Association*. 98, 750 – 763.

[14] Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society*. Series B 63, 511 – 528.

[15] Wang, J. (2010). Consistent selection of the number of clusters via cross-validation. *Biometrika*. 97, 893-904.

[16] Kruskal J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis. *Psychometrika*. 29(1), 1-27.

ABSTRACT

A Study on the Number of Clusters Using Measures of Association

AHHYEON BAEK
Department of Statistics
The Graduate School
Sungshin Women's University

In cluster analysis, it is important to estimate the number of clusters. Many ways to determine the number of cluster have been proposed sush as Calinski & Harabasz(1974), Hartigan(1975), Krzanowski & Lai(1985). Most of them are based on the between cluster and/or within-cluster sum of squared distances.

Recently, researches on the stability of the clustering have been studied. It has been proposed to select the number of clusters as the one minimizing the clustering instability(Wang, 2010; Fang & Wang, 2012). Also, Wang(2010) and Fang & Wang(2012) developed an estimate scheme for clustering instability based on bootstrap and cross-validation.

In this paper, we define the clustering instability by using measure of

association such as kappa coefficient, Jaccard coefficient, phi coefficient.

The proposed methods are demonstrated on a variety of numerical experiments using the simulation and real data application. The simulation study and real application showed that the hit ratio of the proposed method is higher than the previous methods. As a result, we know that the methods using measures of association are competitive.