The Korean Journal of Applied Statistics (2017) DOI: http://dx.doi.org/10.5351/KJAS.2017.30.6.889
30(6), 889–899

Comparison of several criteria for ordering independent components

Eunbin Choi[a] · Sulim Cho[a] · Mira Park[b,1]

[a]Department of Statistics, Korea University; [b]Department of Preventive Medicine, Eulji University

(Received September 5, 2017; Revised October 28, 2017; Accepted November 20, 2017)

Abstract

Independent component analysis is a multivariate approach to separate mixed signals into original signals. It is the most widely used method of blind source separation technique. ICA uses linear transformations such as principal component analysis and factor analysis, but differs in that ICA requires statistical independence and non-Gaussian assumptions of original signals. PCA have a natural ordering based on cumulative proportion of explained variance; howerver, ICA algorithms cannot identify the unique optimal ordering of the components. It is meaningful to set order because major components can be used for further analysis such as clustering and low-dimensional graphs. In this paper, we compare the performance of several criteria to determine the order of the components. Kurtosis, absolute value of kurtosis, negentropy, Kolmogorov-Smirnov statistic and sum of squared coefficients are considered. The criteria are evaluated by their ability to classify known groups. Two types of data are analyzed for illustration.

Keywords: independent component analysis, blind source separation, kurtosis, negentropy, Kolmogorov-Smirnov statistic, sum of squared coefficients, ordering

1. 서론

독립성분분석(independent component analysis; ICA)은 다변량 자료에서 차원을 축소시키는 통계적 방법 중 하나이며, 성분이 통계적으로 독립이고 비정규분포를 따르는 성분을 찾는 것이다. 주성분분석(principal component analysis; PCA) 및 요인분석(factor analysis; FA)과 같은 차원축소 방법과 기본적으로 다른 점은 비정규성을 가정한다는 것이다.

ICA는 주로 블라인드 음원 분리(blind source separation; BSS)를 위한 기법으로 잘 알려져 있다. BSS의 목표는 여러 사람들이 동시에 말하거나 모바일폰의 전자파의 방해와 같이 신호가 혼합된 상황에서 개별적인 신호를 구분해내는 것이다. BSS기법에서 블라인드는 원신호의 본질과 관련하여 거의 알려지지 않았다는 것을 뜻하며, BSS를 블랙박스 방법이라고도 한다 (Stone, 2004; Naik와 Kumar, 2011).

이러한 원신호를 분리하는 문제는 기존의 주성분분석이나 요인분석을 이용해서는 해결하기 어렵다. 주성분분석과 요인분석에서는 약한 가정으로 원신호들이 비상관(uncorrelatedness) 되었으며, 정규분포

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B4011504).
[1]Corresponding author: Department of Preventive Medicine, Eulji University, 77 Gyeryong-ro 771, Jung-Gu, Taejeon 34824, South Korea. E-mail: mira@eulji.ac.kr

를 따른다고 가정하는 반면, ICA는 원신호들이 독립적이고 비정규분포를 따른다고 가정한다 (Stone, 2004). 원신호들은 각각 다른 물리적 과정을 통해 생성되므로, 원신호들이 서로 독립이라는 ICA의 가정은 좀 더 현실적이라고 할 수 있다 (Naik와 Kumar, 2011; Hyvärinen, 2013).

ICA의 원리를 설명하기 위하여 자주 사용되는 예는 다음과 같다. 두 사람이 한 방의 다른 위치에 서서 각자의 마이크로 동시에 말하는 상황을 생각해보자. 이때 두 음성신호는 다른 사람에 의해 생성되므로 서로 독립이다. 그러나 스피커를 통해 나오는 혼합신호는 독립적이지 않다. 혼합신호는 공통적인 원신호들을 사용하며, 서로 다른 위치의 마이크에서는 서로 다른 비율로 독립신호가 포함되기 때문이다 (Hyvärinen와 Oja, 2000). ICA는 스피커를 통해 나오는 혼합된 음성신호로부터 각각의 독립적인 원신호를 분리하는 방법이라고 할 수 있다.

최근 ICA는 의학에서의 신호분석 (Zhu 등, 2006; Enderle와 Bronzino, 2012; Kumagai와 Utsugi, 2004; De Martino 등, 2007), 휴대폰의 신호 분석 (Cristescu 등, 2000), 이미지분석 (Cichocki와 Amari, 2002; Zhang 등, 2007) 등 매우 다양한 분야에서 활용되고 있다. 하지만 ICA에는 아직 해결해야 할 문제가 많이 남아있다. 본 연구에서는 이중에서 추출된 성분을 중요순서에 따라 구분하는 방법에 관하여 알아보고자 한다. 주성분분석의 경우 성분들은 분산이 클수록 더 중요한 성분이라고 간주한다. 하지만 ICA에서는 성분들의 분산 값을 알 수 없으며, 중요순서를 알 수 있는 통계량도 정해져 있지 않다. 따라서 본 논문에서는 중요순서와 관련된 통계량들을 정하고, 두 가지 형태의 알려진 자료를 통해 이들을 분류의 측면에서 평가하여 비교하고자 한다.

2절에서는 ICA의 알고리즘에 대해서 설명한다. 3절에서는 독립성분들의 순서를 결정하기 위한 몇 가지 통계량을 소개할 것이다. 4절에서 두개의 자료를 이용하여 이 통계량들의 결과를 비교할 것이다.

2. 독립성분분석의 알고리즘

ICA는 관찰된 신호인 $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$가 독립성분 $\mathbf{s} = (s_1, s_2, \ldots, s_m)'$들의 선형 결합으로 다음과 같이 나타낼 수 있다.

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.1}$$

여기서 $\mathbf{A}$는 혼합행렬(mixing matrix)이다. 식 (2.1)은 혼합행렬 $\mathbf{A}$의 행과 독립 성분들의 곱의 합으로 다음과 같이 나타낼 수 있다 (Hyvärinen와 Oja, 2000).

$$\mathbf{x}_i = \sum_{j=1}^{m} a_{ij} s_j, \quad \text{for all} \;\; i = 1, \ldots, n. \tag{2.2}$$

ICA에서는 정보가 알려지지 않은 $\mathbf{A}$와 $\mathbf{s}$를 동시에 찾아야하기 때문에, 여러 순열조합이 가능하여 결과해는 유일한 형태로 나오지 않는다.

또한 ICA에서는 독립성분 $s_1, s_2, \ldots, s_m$가 독립적이고 비정규분포를 따른다고 가정하며, 비정규성을 최대화하는 성분들을 찾도록 다음 식에서의 $\mathbf{W}$를 구하게 된다 (Hyvärinen, 2013).

$$\mathbf{s} = \mathbf{W}\mathbf{x},$$

여기서 $\mathbf{W} = \mathbf{A}^{-1}$이다. Lu와 Rajapakse (2003)는 간결성과 일반성을 잃지 않기 위해 비혼합행렬(demixing matrix) $\mathbf{W}$가 $n = m$인 경우만 고려한다. 따라서 본 논문에서도 정사각행렬(square matrix)인 경우만 고려한다. ICA의 해를 찾는 방법으로 정보이론(information theory)에 기초한 방법을 고려할 수 있다. 상호정보량(mutual information)은 확률변수간의 의존성(dependence)에 관한 자연

스런 척도로서, 이를 최소화하도록 정식화 할 수 있다. 이의 정의와 엔트로피(entropy) 및 음의 엔트로피(negentropy)와의 관계에 대해서는 다음절에서 설명하겠다.

비정규성을 측정할 수 있는 척도로는 첨도(kurtosis)와 음의 엔트로피를 사용할 수 있다. 본 논문에서 사용된 R 패키지인 fastICA는 ICA 알고리즘으로 하나씩 또는 동시에 성분들을 추정할 수 있으며, 비정규성을 최대화하는 성분들을 찾는다 (Naik와 Kumar, 2011). 비정규성을 최대화하기 전에, 두 단계의 전처리 과정이 필요하다. 첫 번째 단계는 중심화(centering)로, 관측된 벡터 $\mathbf{x}$에서 평균 벡터 $\mathbf{m} = E(\mathbf{x})$를 빼줌으로써 평균을 0으로 만드는 것이다. $\mathbf{s}$ 또한 평균이 0이 되므로, 이 단계는 ICA 알고리즘을 단순하게 만들어준다. 그 다음으로 두 번째 단계는 화이트닝(whitening)이다. 이 단계에서는 중심화된 관측 벡터 $\mathbf{x}$를 선형변환하여, 화이트닝된 새로운 벡터 $\widetilde{x}$를 얻게 해준다. 이때 $\widetilde{\mathbf{x}}$의 성분들은 비상관이고 단위 분산(unit variance)을 가진다. 즉,

$$E\left(\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}'\right) = \mathbf{I}.$$

화이트닝을 위한 방법 중 가장 알려진 방법은 고유값 분해를 이용하는 것이다. 화이트닝은 추정해야할 모수의 수를 줄여주므로, 문제의 복잡성을 감소시켜 줄 수 있다는 점에서 유용하다 (Hyvärinen와 Oja, 2000).

3. 성분의 순서 결정을 위한 통계량

독립성분분석에는 2가지 모호성이 존재한다. 첫째, 독립성분들의 분산을 알 수 없다. 이는 s_i의 크기가 $\mathbf{a}_i$의 크기에 의해 상쇄될 수 있기 때문이다. 이에 가장 자연스러운 해결 방법은 각 성분들이 단위 분산을 가진다고 가정하는 것이다 (Hyvärinen와 Oja, 2000; Hyvärinen, 2013). 둘째, 독립성분들의 중요순서를 정할 수 없다. 이는 $\mathbf{a}_i$와 s_i에 대한 정보가 없기 때문에 식 (2.2)의 항의 순서를 마음대로 변경할 수 있기 때문이다 (Hyvärinen와 Oja, 2000). 이를 해결하기 위해 ICA 알고리즘은 비정규성을 최대화하는 성분을 찾는다는 점에 착안하여 비정규성의 크기에 기반한 통계량들을 위주로 성분 순서를 결정하기 위한 후보 통계량을 정하였다.

3.1. 첨도

첨도는 비정규성을 측정할 수 있는 대표적인 척도로, 성분 s의 첨도는 다음과 같이 정의될 수 있다.

$$\text{kurt}(s) = E\left(s^4\right) - 3\left(E\left(s^2\right)\right)^2.$$

또한 이 식의 우변은 $E(s^2) = 1$이라는 가정을 통해 $E(s^4) - 3$으로 간단하게 나타낼 수 있으며, $\mathbf{s}$가 정규확률변수일 경우 $E(s^4) = 3(E(s^2))^2$이므로 첨도는 0이 된다. 따라서 비정규 확률변수인 성분 $\mathbf{s}$의 첨도는 0이 아니며, 첨도값은 양수와 음수 모두 가능하다.

독립성분들의 중요순서를 정하기 위해 이러한 첨도값을 사용하여, 첨도값이 클수록 더 중요한 성분이라고 주장한 연구가 있었다 (Lu와 Rajapakse, 2003). 또 다른 방법은 첨도값의 절댓값을 이용하는 것이다. 첨도의 절댓값이 클수록 첨도의 비정규성이 크다는 것을 의미하기 때문에, |첨도|가 클수록 더 중요한 성분이라고도 볼 수 있다.

3.2. 음의 엔트로피

비정규성을 측정할 수 있는 또 다른 중요한 척도로는 음의 엔트로피가 있다. 이는 정보이론에서 불확실성을 측정하기 위해 사용되는 엔트로피와 밀접하게 관련이 있다 (Shannon, 2001). 확률변수 $\mathbf{s}$의 엔트

로피 $H(\mathbf{s})$는 다음과 같이 정의되며, 정보이론의 결과를 통해 같은 분산을 가진 모든 확률변수들 중에서 정규확률변수는 가장 큰 엔트로피를 가짐을 알 수 있다.

$$H(\mathbf{s}) = -\int f(\mathbf{s}) \log f(\mathbf{s}) ds$$

또한 $\mathbf{s}_{\text{gauss}}$는 $\mathbf{s}$와 같은 공분산을 가지는 정규 확률 변수이며, 음의 엔트로피 $J(\mathbf{s})$는 다음과 같이 정의된다 (Hyvärinen와 Oja, 2000).

$$J(\mathbf{s}) = H(\mathbf{s}_{\text{gauss}}) - H(\mathbf{s}) \approx \left[E\{G(s)\} - E\{G(s_{\text{gauss}})\}\right]^2,$$
$$G(u) = -\exp\left(-\frac{u^2}{2}\right).$$

음의 엔트로피는 항상 0 이상의 값을 가지며, $\mathbf{s}$가 정규분포를 따를 때 음의 엔트로피 값은 0이 된다. 그러므로 독립성분이 비정규성을 띌수록 음의 엔트로피 값은 0에서 멀어진다고 볼 수 있으며, 음의 엔트로피 값이 클수록 더 중요한 성분이라고도 볼 수 있다. 음의 엔트로피는 계산하기 어렵기 때문에 근사하여 사용된다 (Hyvärinen와 Oja, 2000; Naik와 Kumar, 2011).

상호정보량 $I(s)$를 다음과 같이 정의할 수 있다.

$$\begin{aligned} I(s_1, s_2, \ldots, s_m) &= \sum_{j=1}^{m} H(s_j) - H(s) \\ &= \sum_{j=1}^{m} H(s_j) - H(\mathbf{x}) - \log|\det W|. \end{aligned}$$

이는 음의 엔트로피와 다음과 같은 관계를 갖는다.

$$I(s_1, s_2, \ldots, s_m) = C - \sum_{j=1}^{m} J(s_j). \tag{3.1}$$

이때, C는 W에 의존하지 않는 상수 값이다. 식 (3.1)에 의해 상호정보량을 최소화시키는 것은 음의 엔트로피가 최대화되는 방향을 찾는 것과 대략적으로 같다고 할 수 있다. 따라서 ICA를 상호정보량의 최소화는 비정규성을 최대화하는 방향으로 찾는 것과 같다고 할 수 있다 (Hyvärinen와 Oja, 2000).

3.3. 콜모고로프-스미르노프 통계량

콜모고로프-스미르노프(Kolmogorov-Smirnov; K-S)는 비모수 검정으로 주어진 표본의 분포가 알고자 하는 분포와 동일한 지 검정할 때 쓰이며, 통계량은 다음과 같이 정의된다.

$$D_n = \max|F_n(x) - F(x)|.$$

이때, $F_n(x) = (1/n)\sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$ 표본의 누적분포 함수는 $F_n(x)$로 볼 수 있으며, 알고자하는 분포 $f(x)$의 누적분포 함수는 $F(x)$이다. $I_{[-\infty,x]}(X_i)$는 $X_i \leq x$인 경우는 1, 그 외에는 0을 가지는 지시함수를 말하며, 표본들의 분포가 $f(x)$를 따른다면 D_n의 값은 0과 가까워질 것이다. $F(x)$를 정규분포의 누적분포로 두었을 때, 이 값이 클수록 비정규성이 커지므로 콜모고로프-스미르노프 통계량이 클수록 더 중요한 성분이라고도 볼 수 있다.

3.4. 계수제곱합

ICA에서 $\mathbf{x}$는 혼합행렬 $\mathbf{A}$와 독립성분 $\mathbf{s}$의 선형 결합으로 이루어져 있다. 혼합행렬 $\mathbf{A}$는 원소인 a_{ij}들로 이루어져 있으며, 이는 각각 독립성분의 계수를 의미한다. 따라서, j번째 성분의 강도를 측정하기 위해 다음과 같이 독립성분계수 값을 이용하여 정의하였다. 계수제곱합(sum of squared coefficients; SSC)이 클수록 j번째 성분의 중요도가 크다는 것을 의미하며, 통계량은 다음과 같이 정의된다.

$$\text{SSC} = \sum_{j=1}^{m} a_{ij}^2, \quad \text{for all} \;\; i = 1, \ldots, n.$$

이 통계량은 값이 클수록 더 중요한 성분이라고도 볼 수 있다.

4. 분석결과

두 개의 자료를 이용하여 3절에서 설명된 통계량들이 독립성분들의 중요순서와 어떤 연관이 있는지 알고자 하였다. 이때 고려한 통계량은 첨도, 첨도의 절댓값(|kurtosis|), 음의 엔트로피, 콜모고로프-스미르노프 통계량, 계수제곱합이다. 이 통계량들의 값이 클수록 중요한 성분으로 볼 수 있다. 앞으로 제시할 표에서는 값이 가장 큰 성분의 성분 순서는 1로 표현하였다. R의 fastICA 패키지를 이용하여 분석을 진행하였다. ICA알고리즘은 실행시킬 때마다 추출된 성분이 다르므로, 편의상 임의의 수를 넣어 seed(1234)로 고정시켜 분석을 진행하였으며 결과가 나온 순서대로 IC1, IC2 등으로 명명하였다. 각 자료마다 추출될 성분의 수는 편의상 4까지로 하여, 총 2, 3, 4인 경우를 모두 실행시켜 한 표로 나타내었다. 또한 동시에 성분들을 추정하는 알고리즘을 사용하였고, 음의 엔트로피 근사를 위한 함수로서 지수함수를 이용하였다.

본 연구에서는 분류의 기준으로 방법을 평가하고자 하였다. 이를 위해 2차원 그래프를 통해서 어떤 성분이 알려진 그룹을 잘 구분하는지 판단하였다. 또한 보다 객관화한 평가를 위하여 BSS/WCSS를 통해 정량적인 값으로 비교하였다. 여기서 BSS(between sum of square)는 각 그룹 간 변동을 의미하고, WCSS(within clustering sum of square)는 각 그룹 내 변동을 의미한다. 먼저 군집분석을 통해 알려진 수의 그룹으로 나눈 후 BSS/WCSS 값을 계산하고, 이 값이 클수록 그룹을 잘 나누었다고 평가하였다.

4.1. 붓꽃 자료

Anderson (1935)의 붓꽃 자료(iris data)를 분석하였으며, 이 자료는 꽃잎의 길이와 너비 등 꽃의 특징과 관련된 5개의 변수와 150개의 관측치로 이루어져 있다. 꽃은 총 3가지 품종(setosa, versicolor, virginica)이며 각각 50개씩의 관측치를 가지고 있다.

추출된 성분의 수가 2개, 3개, 4개로 증가함에 따라 5개의 통계량들을 구한 결과이다 (Table 4.1). 각각의 통계량에 따른 중요성분 순서는 괄호 안에 표현하였다. 성분 수가 2개인 경우, 첨도를 제외한 나머지 4개의 통계량들은 동일하게 중요성분을 결정하였다. 또한 성분 수가 3개일 때는 |첨도|와 음의 엔트로피의 중요성분 순서가 동일하고, 콜모고로프-스미르노프 통계량과 SSC의 중요성분 순서가 동일하였다. 마지막으로 성분의 수가 4개인 경우에는 |첨도|와 음의 엔트로피의 중요성분 순서가 동일하였으며, 콜모고로프-스미르노프 통계량과 SSC의 중요성분 순서는 거의 동일하였다. 붓꽃 자료에서는 통계량 |첨도|와 음의 엔트로피의 중요성분 순서가 정확히 일치하였으며, 통계량 콜모고로프-스미르노프 통계량과 SSC의 성분 순서는 거의 비슷하였다.

성분 수가 4개일 때 가능한 총 6가지 조합에 대해 3가지 품종(setosa, versicolor, verginica)에 따라 나타내었다. IC4가 가장 세 개의 그룹을 잘 나누어 주므로, 가장 흥미롭고 중요한 성분으로 볼 수 있다

Table 4.1. Statistic of indenpent components for iris data

# of IC	IC	Kurtosis	|Kurtosis|	Negentropy	K-S	SSC
2	IC1	−1.43 (2)	1.43 (1)	0.001615 (1)	0.1978 (1)	4.08 (1)
	IC2	0.12 (1)	0.12 (2)	0.000089 (2)	0.0391 (2)	0.36 (2)
3	IC1	0.06 (2)	0.06 (3)	0.000001 (3)	0.0691 (2)	0.26 (2)
	IC2	0.11 (1)	0.11 (2)	0.000176 (2)	0.0439 (3)	0.21 (3)
	IC3	−1.46 (3)	1.46 (1)	0.001145 (1)	0.1888 (1)	4.05 (1)
4	IC1	−0.30 (3)	0.30 (4)	0.000154 (4)	0.0588 (2)	0.22 (2)
	IC2	0.69 (1)	0.69 (2)	0.000609 (2)	0.0556 (4)	0.18 (3)
	IC3	0.65 (2)	0.65 (3)	0.000226 (3)	0.0566 (3)	0.05 (4)
	IC4	−1.46 (4)	1.46 (1)	0.000849 (1)	0.1944 (1)	4.09 (1)

K-S = Kolmogorov-Smirnov; SSC = sum of squared coefficients.

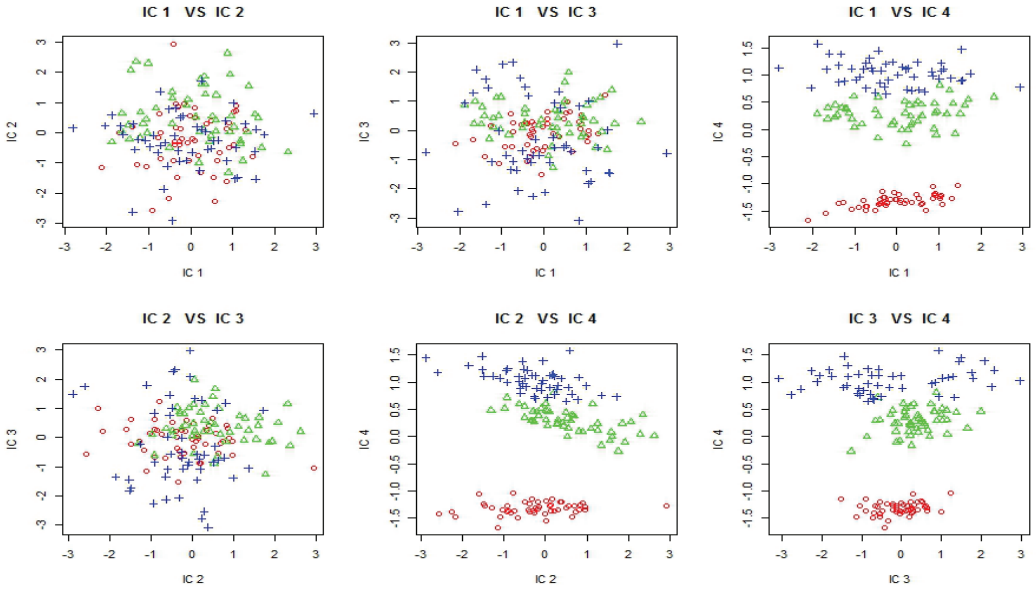

Figure 4.1. Scatter plot of independent components for iris data (○: setosa, △: versicolor, +: verginica).

(Figure 4.1). 또한 시각화를 통해서 versicolor와 virginica는 setosa에 비해 서로 더 가까운 개체들임을 알 수 있다. 그룹의 분류하는 정도를 BSS/WCSS를 통해 정량적인 값을 통해 비교한 결과, 1개의 성분을 사용하여 분류한다면 IC4가 월등하게 그룹을 잘 구분하고 있으며 (BSS/WCSS = 25.7403), 두 개의 성분으로 분류할 때에는 IC2와 IC4의 쌍이 가장 분류를 잘한다는 것을 알 수 있다(BSS/WCSS = 0.0748) (Table 4.2). 첨도를 제외한 4개의 통계량에서 IC4의 순서는 첫 번째로 나타났으며, IC2를 다음으로 뽑은 통계량은 |첨도|와 음의 엔트로피였다.

4.2. 유전자 자료

프로그램 GENOME (Liang 등, 2007)을 이용하여 유전자 자료를 모의생성 하였다. Figure 4.2에서 도식화한 것처럼 3개의 진화모델들 중에서, 두 개의 조상 모집단으로부터 도출된 세 개의 부집단을 포함하는 자료를 사용하였다 (Intarapanich 등, 2009). 자료를 생성하기 위해서는 부집단의 수와 각각 개체들

Table 4.2. BSS/WCSS for iris data

1 component		2 components	
Component	BSS/WCSS	Components	BSS/WCSS
IC1	0.0048	IC1 vs. IC2	0.0049
IC2	0.1729	IC1 vs. IC3	0.0026
IC3	0.0835	IC1 vs. IC4	0.0562
IC4	25.7403	IC2 vs. IC3	0.0075
		IC2 vs. IC4	0.0748
		IC3 vs. IC4	0.0650

BSS = between sum of square; WCSS = within clustering sum of square.

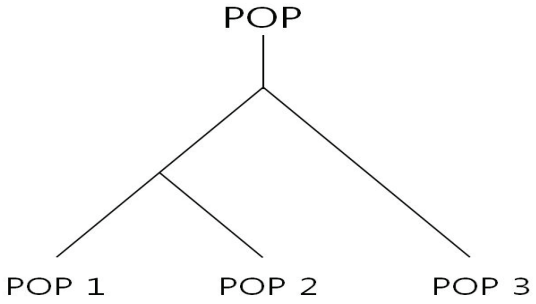

Figure 4.2. Population history trees for simulated genotype data.

Table 4.3. Statistic of indenpent components for simulated genotype data

# of IC	IC	Kurtosis	|Kurtosis|	Negentropy	K-S	SSC
2	IC1	−1.41 (2)	1.41 (1)	0.001326 (1)	0.1367 (2)	1255.19 (1)
	IC2	−1.17 (1)	1.17 (2)	0.000342 (2)	0.1791 (1)	844.50 (2)
3	IC1	−1.21 (2)	1.21 (2)	0.001046 (2)	0.1373 (3)	258.53 (2)
	IC2	0.97 (1)	0.97 (3)	0.001282 (1)	0.1508 (2)	159.14 (3)
	IC3	−1.47 (3)	1.47 (1)	0.001001 (3)	0.1736 (1)	1741.06 (1)
4	IC1	0.23 (2)	0.23 (4)	0.000007 (4)	0.0753 (4)	46.06 (4)
	IC2	1.12 (1)	1.12 (3)	0.001141 (1)	0.1128 (3)	140.61 (2)
	IC3	−1.24 (3)	1.24 (2)	0.000697 (3)	0.1270 (2)	137.41 (3)
	IC4	−1.47 (4)	1.47 (1)	0.000770 (2)	0.2278 (1)	1877.97 (1)

K-S = Kolmogorov-Smirnov; SSC = sum of squared coefficients.

의 수, 독립적인 지역들의 수, 각 독립적인 지역마다 single nucleotide polymorphisms (SNP)들의 고정된 수를 정해야 한다. 부집단의 수는 3개(POP1, POP2, POP3)이며, 각 부집단마다 25개의 개체가 존재한다. 그리고 독립적인 지역의 수는 20, 각 독립적인 지역마다 SNP들의 고정된 수는 500이다. 즉, 75개의 개체에 대해서 10,000개의 SNP를 생성하였다.

추출된 성분의 수가 2개, 3개, 4개로 증가함에 따라 5개의 통계량들을 비교하였다 (Table 4.3). 성분 수가 2개인 경우, 첨도와 콜모고로프-스미르노프 통계량을 제외한 나머지 3개의 통계량들이 동일하게 중요 성분을 결정하였다. 또한 성분 수가 3개일 때는 첨도와 음의 엔트로피의 중요성분 순서가 동일하고, |첨도|와 SSC의 중요성분 순서가 동일하였다. 마지막으로 성분의 수가 4개인 경우에는 |첨도|와 콜모고로프-스미르노프 통계량이 동일하게 중요성분 순서가 결정하였으며, SSC는 이들과 거의 비슷하게 결정하였다. 모의생성된 유전자 자료에서는 통계량 |첨도|와 SSC의 성분 순서가 거의 동일하였다.

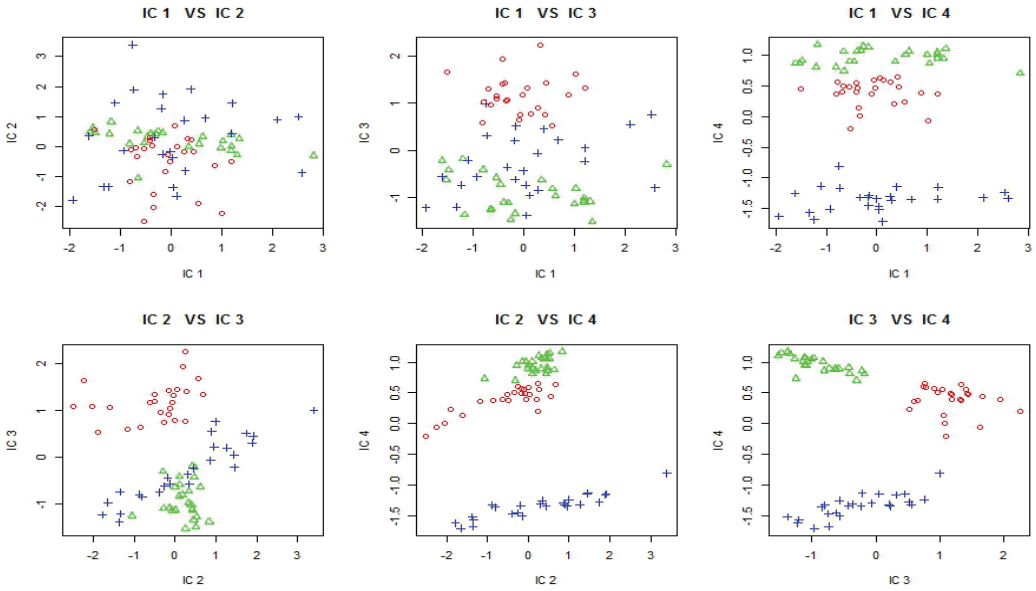

Figure 4.3. Scatter plot of independent components for simulated genotype data (○: POP1, △: POP2, +: POP3).

Table 4.4. BSS/WCSS for genotype data

1 component		2 components	
Component	BSS/WCSS	Components	BSS/WCSS
IC1	1.5351	IC1 vs. IC2	0.0090
IC2	1.1597	IC1 vs. IC3	0.0738
IC3	4.3308	IC1 vs. IC4	0.1132
IC4	86.7180	IC2 vs. IC3	0.0967
		IC2 vs. IC4	0.1471
		IC3 vs. IC4	0.7501

BSS = between sum of square; WCSS = within clustering sum of square.

성분 수가 4개일 때 가능한 총 6가지 조합에 대해 부집단(POP1, POP2, POP3)별로 나타내었다. IC4가 세 개의 그룹을 가장 잘 나누어 주므로, 가장 흥미롭고 중요한 성분으로 볼 수 있다 (Figure 4.3). 다음으로 IC3이 IC2에 비해 POP1과 POP3 그룹을 잘 구분하여 그 다음으로 중요한 성분으로 볼 수 있다. 또한 POP1과 POP2는 POP3에 비해 서로 더 가까운 개체들임을 알 수 있다. 마찬가지로 BSS/WCSS 값을 통해서 IC4(BSS/WCSS = 86.7180)가 세 개의 그룹을 가장 잘 나누어준다는 것을 파악할 수 있었다 (Table 4.4). 2개의 성분을 사용하는 경우 IC3과 IC4가 그룹을 잘 구분하였다는 것을 정량적인 값을 통해서도 확인할 수 있었다. |첨도|, 콜모고로프-스미르노프 통계량, SSC에서 IC4의 순서는 첫 번째로 나타났다. IC3를 두 번째로 뽑은 것은 |첨도|와 콜모고로프-스미르노프 통계량이었다.

5. 결론 및 토의

본 연구에서는 다양한 통계량들을 이용하여 독립성분분석에서 구해지는 독립성분들의 중요순서를 파악하고자 하였다. 본 연구에서는 분류의 측면에서 그룹을 잘 분리해주는 정도를 이용하여 성분의 중요도

를 평가하였다. ICA의 기본 목적은 혼합된 신호로부터 원신호를 분리하는 것이지만, 고차원 자료의 경우에는 중요한 소수의 신호를 선택하는 것이 필요하다. 이러한 차원축소의 측면에서 볼 때에도 뚜렷한 패턴을 가진 성분의 선택이 모든 표본에서 비슷한 신호를 보이는 성분보다 더 의미가 있다고 생각되므로, 연구의 목적이 분류가 아닌 경우라도 이들 통계량의 활용이 타당하다고 생각된다.

시각화 결과와 함께 BSS/WCSS의 값을 통하여 통계량을 비교하였으며, 성분 수가 4개인 경우에 대하여 첫 번째 자료에서는 |첨도|와 음의 엔트로피, 두 번째 자료에서는 |첨도|와 콜모고로프-스미르노프 통계량이 중요성분을 선택하였다. 첨도를 이용한 독립성분들의 중요순서의 결과는 좋지 않았다. Lu와 Rajapakse (2003)에서는 첨도를 이용한 결과가 우수했는데, 이들이 모의생성한 데이터에서 첨도의 크기와 |첨도|의 크기가 비례하여 나타난 결과로 생각된다.

본 연구에서는 비정규성을 근거한 통계량들을 위주로 제시하고 이들을 비교하였다. 기존 연구로서 왜도나 첨도를 사용한 방법 (Lu와 Rajaparkse, 2003; Koch, 2014)이나, IC의 분산에 따라 순서를 매긴 연구 (Lu와 Rajaparkse, 2003)가 있으며, 데이터의 파워를 이용한 순서화 (Hendrikse 등, 2007)가 제시된 바 있다. 그러나 아직 이를 결정하는 전형적인 방법으로 정착된 것은 없다. 자료에 따라 다른 양상이 나타날 수도 있으므로, 본 연구에서 제시된 바와 같이 |첨도|, 콜모고로프-스미르노프 통계량 등의 기준을 종합적으로 고려하고, 그래프와 함께 탐색적인 분석을 할 필요가 있다.

References

Anderson, E. (1935). The irises of the gaspe peninsula, *Bulletin of American Iris Society*, **59**, 2–5.

Cichocki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing: Learning Al- gorithms and Applications*, Number V. 1 in Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley.

Cristescu, R., Ristaniemi, T., Joutsensalo, J., and Karhunen, J. (2000). Delay estimation in cdma communications using a fast ica algorithm. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*. In press.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., and Formisano, E. (2007). Classification of fmri independent components using ic-fingerprints and support vector machine classifiers, *Neuroimage*, **34**, 177–194.

Enderle, J. D. and Bronzino, J. D. (2012). *Introduction to Biomedical Engineering*, Academic Press, Seoul.

Hendrikse, A., Veldhuis, R., and Spreeuwers, L. (2007). Component ordering in independent component analysis based on data power, *28th Symposium on Information Theory in the Benelux*, 211–218.

Hyvärinen, A. (2013). Independent component analysis: recent advances, *Philosophical Transactions of the Royal Society A*, **371**, 20110534.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, **13**, 411–430.

Intarapanich, A., Shaw, P. J., Assawamakin, A., Wangkumhang, P., Ngamphiw, C., Chaichoompu, K., Piriyapongsa, J., and Tongsima, S. (2009). Iterative pruning pca improves resolution of highly structured populations, *BMC Bioinformatics*, **10**, 382.

Koch, I. (2014). *Analysis of Multivariate and High-Dimensional Data*, Cambridge University Press, Cambridge.

Kumagai, T. and Utsugi, A. (2004). Removal of artifacts and fluctuations from MEG data by clustering methods, *Neurocomputing*, **62**, 153–160.

Liang, L., Zöllner, S., and Abecasis, G. R. (2007). Genome: a rapid coalescent-based whole genome simulator, *Bioinformatics*, **23**, 1565–1567.

Lu, W. and Rajapakse, J. C. (2003). Eliminating indeterminacy in ICA, *Neurocomputing*, **50**, 271–290.

Naik, G. R. and Kumar, D. K. (2011). An overview of independent component analysis and its applications, *Informatica*, **35**, 63–81.

Shannon, C. E. (2001). A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**, 3–55.

Stone, J. V. (2004). *Independent Component Analysis: A Tutorial Introduction*, A Bradford Book.

Zhang, Q., Sun, J., Liu, J., and Sun, X. (2007). A novel ica-based image/video processing method, *Independent Component Analysis and Signal Separation*, 836–842.

Zhu, Y., Chen, T. L., Zhang, W., Jung, T.-P., Duann, J.-R., Makeig, S., and Cheng, C.-K. (2006). Noninvasive study of the human heart using independent component analysis. In *BioInformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on*, 340–347.

독립성분의 순서화 방법 비교

최은빈[a] · 조수림[a] · 박미라[b,1]

[a]고려대학교 통계학과, [b]을지대학교 예방의학교실

(2017년 9월 5일 접수, 2017년 10월 28일 수정, 2017년 11월 20일 채택)

요 약

독립성분분석은 혼합된 신호에서 원신호들을 분리하기 위해서 사용되는 다변량 분석방법으로서, 블라인드 음원 분리 중 가장 널리 사용되는 방법이다. 독립성분분석은 주성분분석이나 요인분석과 같이 선형변환을 사용하지만, 원신호들의 통계적 독립과 비정규성 가정을 필요로 한다는 점에서 다르다. 설명되는 분산의 누적비율이 클수록 더 중요한 성분을 의미하게 되는 주성분분석과 달리, 독립성분분석에서는 독립성분들의 중요순서를 결정하는데 적절한 유일한 기준이 정해지지 않는다. 군집분석이나 차원축소된 그래프 작성 등과 같은 후속 연구를 진행하기 위해서는 일부의 주요 독립성분을 사용하게 되므로, 성분의 순서를 정하는 것은 의미가 있다. 본 연구에서는 성분의 순서를 결정하기 위한 몇 가지 기준의 성능을 비교하였다. 첨도와 첨도의 절댓값, 음의 엔트로피, 콜모고로프-스미르노프 통계량, 계수제곱합을 이용한 방법이 고려되었다. 이들은 알려진 그룹을 분류하는 능력을 기준으로 평가되었다. 두 가지 형태의 자료를 이용한 분석결과를 제시하였다.

주요용어: 독립성분분석, 블라인드 음원 분리, 첨도, 음의 엔트로피, 콜모고로프-스미르노프 통계량, 계수제곱합, 순서화

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2017R1A2B4011504).

[1]교신저자: (34824) 대전 중구 계룡로1번길 77, 을지대학교 예방의학교실. E-mail: mira@eulji.ac.kr