응용통계연구 제 **14**권 *1*호, *2001*년, *pp. 137-148*

Box-Cox 대비변환을 이용한 구성비율자료의 주성분분석

최병진[1] 김기영[2]

요 약

비율을 나타내는 요소들로 이루어진 구성비율자료는 각 행들의 합이 1이 되는 제약을 가지고 있어 통계적으로 다루기가 쉽지 않다. 더구나 자료의 구조가 선형적인 형태를 보이지 않는 특성을 가지기 때문에 주성분분석과 같은 선형적인 다변량기법들을 구성비율자료에 적용을 할 때 잘못된 해석과 추론이 이루어질 가능성이 있다. 본 논문에서는 구성비율자료의 주성분분석에서 기존의 방법들이 가지는 문제점을 해결하기 위해 Box-Cox 대비변환(Box-Cox contrast transformation)을 이용한 새로운 형태의 분석방법을 제시한다. 그리고 실제자료의 분석과 모의실험을 통해서 Aitchison(1983)이 제시한 방법과 수행능력을 비교하고자 한다.

주요용어: 구성비율자료, Box-Cox 대비변환, Box-Cox 대비 주성분분석.

1. 서론

지질학을 포함한 많은 응용분야에서 얻어지는 대부분의 자료들은 특별한 형태의 다변량자료인 구성비율자료와 밀접한 관련이 있다(Coakley와 Rust, 1968; Thompson, Esson과 Duncan, 1972; Carr, 1981). 구성비율자료(compositional data)는 성분들의 비율을 나타내는 p개의 양의 요소 $x_1, x_2, \cdots, x_p$들을 측정한 자료로 각 행들의 합이 1이 되는 제약을 가지게 된다. 구성비율자료의 분석에서 제약은 다양한 형태의 문제들을 야기시키고 또한 자료의 구조가 비선형적인 형태를 보이기 때문에 통상적인 다변량기법들의 적용은 적절치 않다는 것이 알려져 있다(Chayes, 1960; Chayes와 Kruscal, 1966; Gower, 1967).

고차원적인 구성비율자료를 저차원화시키기 위해 다양한 형태의 주성분분석들이 시도된 바가 있다(Butler, 1976; Le Maitre, 1968; Webb과 Briggs, 1966; Aitchison, 1983). Le Maitre(1968)와 Butler(1976)의 분석방법은 모든 구성비율들의 공분산행렬을 이용하고 있으며 특히 Le Maitre(1968)는 0이 아닌 고유값들과 이에 대응되는 고유벡터들에 기초를 둔 주성분들의 사용을 제안했다. 그러나 이들의 방법은 다음과 같은 문제점들을 내포하고 있다. 첫째는 얻어지는 주성분들은 구성비율자료가 가지는 비선형적인 변이를 제대로 포착해 주지 못하는 경향을 보인다. 둘째는 제약으로 인해 공분산들이 음의 편향을 가지기 때문에 주성분들의 상관관계 해석에 어려움이 있다. 한편 Webb과 Briggs(1966)는 다음의 변환 $x_i/x_j, i \neq j$을 통해 얻은 변환된 구성비율들의 공분산행렬에 기초한 주성분분석을 시도하

1) (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계연구소, 연구원
E-mail: choibj@mail.korea.ac.kr
2) (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 교수
E-mail: kykim@mail.korea.ac.kr

였다. 그러나 공통분모로 사용한 임의의 x_j의 선택에 따라 서로 다른 주성분들이 얻어지게 된다.

또 다른 방법으로 Aitchison(1982)는 구성비율자료분석에서 로그-비 변환(log-ratio transformation)의 사용을 제안하고 다변량 정규성 가정 하에서 개발된 기존의 방법론을 변환된 자료에 적용할 수 있음을 보여 준 바가 있다. 이를 근거로 Aitchison(1983)은 변환된 구성비율들의 공분산행렬에 기초한 비선형 형태의 새로운 주성분체계의 구축을 시도하였다. 그리고 실제자료의 분석을 통해서 기존의 방법들보다는 비선형적인 변이를 포착하는 측면에서 우수함을 보였다.

그러나 Jolliffe(1986, pp. 211-212)의 예에서 알 수 있듯이 항상 그러한 결과를 제공하지는 않는 것 같다. 더구나 로그변환된 구성비율들의 선형결합으로 표현되는 주성분들은 고려할 수 있는 많은 비선형 주성분들 중에서 특수한 형태에 지나지 않고 이를 이용하여 구성비율자료의 비선형 변이형태를 파악하는 데에는 극히 제한적일 수밖에 없다. 또한 Rayens와 Srinivasan(1991)은 로그-비 변환의 확장된 형태로 비에 대해 Box-Cox 멱변환을 적용한 다음의 변환 $\frac{(x_j/x_p)^{\lambda_j}-1}{\lambda_j}, j=1,\cdots,p-1$을 고려하여 실제 자료에 적용해 보았을 때 다변량 정규성과 관련된 변환모수들의 추정값들이 전반적으로 0에 가까운 값들을 가지지 않음을 보인 바가 있다. 이것은 결국 자료분석에서 로그-비 변환의 사용은 구성비율들의 다변량 정규성을 항상 보장해 주지 못함을 의미한다. 따라서 다변량 정규성을 충족시켜 주고 나아가 구성비율자료가 보이는 다양한 모습의 변이형태에 대처할 비선형 형태의 주성분들을 찾기 위해서는 새로운 변환을 고려해야 한다.

본 논문에서는 문제해결을 위해 새로운 변환기법을 도입하고 이에 기초한 분석방법을 제시한다. 그리고 자료분석을 통하여 그 유용성을 알아보고자 한다. 2절에서는 Aitchison이 고려한 변환의 확장된 형태로 볼 수 있는 Box-Cox 대비변환(Box-Cox contrast transformation)을 도입하고 이를 이용한 분석방법을 제시한다. 3절에서는 실제자료와 모의자료에 적용한 결과를 통해서 방법의 유용성을 알아보고 기존의 Aitchison의 방법과 수행능력을 비교한다. 끝으로 4절에서는 결론을 맺는다.

2. 변환과 방법론

2.1. Box-Cox 대비변환

자료분석에서 요구되는 가법성, 등분산성, 정규성 등과 같은 가정들을 충족하기 위해 종종 변수의 변환을 시도하게 된다. 이런 목적으로 제안된 변환방법들 중에서 여러 응용측면에서 폭넓게 사용되는 것이 Box-Cox 멱변환(Box-Cox power transformation)이다. Aitchison이 제안한 변환을 대치할 수 있는 새로운 변환을 찾기 위한 방법 중의 하나는 구성비율들에 대해 멱변환을 직접적으로 적용하는 것이다. 그러나 제한된 범위를 가지는 변수들에 대해서 멱변환을 적용할 경우 변환과 관련된 모수들을 제대로 추정해주지 못하는 문제가 발생한다. 이에 대안으로 Rayens와 Srinivasan(1991)의 변환을 고려할 수도 있지만 이 경우 임의로 선택한 x_p에 따라 변환된 구성비율들의 분산-공분산들은 서로 다르게 되고 이로 인해 다른 주성분들을 얻게 된다. 따라서 다른 형태의 변환을 고려해야 할 필요가 있다. 이를 위

해 비율(proportion)을 나타내는 변수에 관한 Mosteller와 Tukey(1977, pp. 92)의 접힌 멱변환(folded power transformation)을 응용하여 구성비율들에 적용할 수 있는 다변량 형태의 변환을 고려한다.

$\mathbf{x} = (x_1, x_2, \cdots, x_p)^t$를 $(p-1)$-차원 심플렉스 공간 $\mathbb{S}^{p-1}$에서 정의되는 크기 $p \times 1$인 구성비율벡터라 하면 임의로 선택한 구성비율 $x_j, j \neq p$ 와 x_p에 대한 Box-Cox 대비변환(Box-Cox contrast transformation)을 다음과 같이 정의한다.

$$T^{BCC}(x_j, x_p; \lambda_j, \lambda_p) = \begin{cases} \frac{x_j^{\lambda_j}-1}{\lambda_j} - \frac{x_p^{\lambda_p}-1}{\lambda_p}, & \text{if } \lambda_j \text{ and } \lambda_p \neq 0 \\ \frac{x_j^{\lambda_j}-1}{\lambda_j} - \log\ x_p, & \text{if } \lambda_j \neq 0 \text{ and } \lambda_p = 0 \\ \log\ x_j - \frac{x_p^{\lambda_p}-1}{\lambda_p}, & \text{if } \lambda_j = 0 \text{ and } \lambda_p \neq 0 \\ \log\ x_j - \log\ x_p, & \text{if } \lambda_j \text{ and } \lambda_p = 0 \end{cases} \tag{2.1}$$

여기서 $\lambda_1, \lambda_2, \cdots, \lambda_p$는 변환과 관련된 모수이다. (2.1)에 의해 변환된 구성비율들은 더 이상 합이 1이 되지 않기 때문에 제약으로 인한 분석의 어려움은 나타나지 않는다. 또한 모든 변환모수들이 0일 경우 (2.1)은 Aitchison의 변환이 되기 때문에 구성비율자료가 보이는 다양한 변이형태에도 더 잘 대처하고 정규성 측면에서도 로그변환에 제한적인 것보다는 더 나은 적합도의 제공을 기대할 수 있다.

Box-Cox 대비변환에서 변환모수들은 구성비율자료로부터 최대가능도방법에 의해 다음과 같이 추정할 수가 있다. p개의 구성비율들을 독립적으로 관찰한 벡터를 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$이라 하고 $\mathbf{y}_i^{bc} = \left(y_{i1}^{bc}, y_{i2}^{bc}, \cdots, y_{i(p-1)}^{bc}\right)^t$, $i = 1, \cdots, N$을 i번째 관찰벡터 $\mathbf{x}_i$를 (2.1)을 통해 변환한 벡터라 하자. 또한 $\mathbf{y}_i^{bc}$들은 평균이 $\boldsymbol{\mu}^{bc}$, 분산-공분산이 Σ_y^{bc}인 다변량 정규분포를 따른다고 하자. 그러면 $\mathbf{x}_i$들에 기초한 로그-가능도 함수는 다음과 같은 형태를 가지게 된다.

$$l\left(\boldsymbol{\mu}^{bc}, \Sigma_y^{bc}, \lambda_1, \lambda_2, \cdots, \lambda_p | \mathbf{x}_1, \cdots, \mathbf{x}_N\right) = k - \frac{N}{2}\ \log\ |\Sigma_y^{bc}| + \sum_{i=1}^{N}\sum_{j=1}^{p}(\lambda_j - 1)\ \log\ x_{ij} + \sum_{i=1}^{N}\ \log\ \left(\sum_{j=1}^{p} x_{ij}^{1-\lambda_j}\right) - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y}_i^{bc} - \boldsymbol{\mu}^{bc}\right)^t \Sigma_y^{bc\,-1}\left(\mathbf{y}_i^{bc} - \boldsymbol{\mu}^{bc}\right) \tag{2.2}$$

여기서 k는 $-\frac{N(p-1)}{2}\ \log 2\pi$인 상수이다. 고정된 λ_j들에 대해 $\boldsymbol{\mu}^{bc}$와 Σ_y^{bc}의 추정량으로 표본평균벡터 $\bar{\mathbf{y}}^{bc} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^{bc}$ 와 표본공분산행렬 $S_y^{bc} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i^{bc} - \bar{\mathbf{y}}^{bc}\right)\left(\mathbf{y}_i^{bc} - \bar{\mathbf{y}}^{bc}\right)^t$를 위의 식에 대입하면 다음과 같은 축소된 로그-가능도 함수를 얻게 된다.

$$l(\lambda_1, \lambda_2, \cdots, \lambda_p | \mathbf{x}_1, \cdots, \mathbf{x}_N) = k_1 - \frac{N}{2}\ \log\ |S_y^{bc}| + \sum_{i=1}^{N}\sum_{j=1}^{p}(\lambda_j - 1)\ \log\ x_{ij} + \sum_{i=1}^{N}\ \log\left(\sum_{j=1}^{p} x_{ij}^{1-\lambda_j}\right) \tag{2.3}$$

여기서 k_1은 $-\frac{N(p-1)}{2}\log 2\pi - \frac{N(p-1)}{2}$이다. 따라서 변환모수들의 최대가능도 추정량은 (2.3)을 λ_j에 대해 편미분하여 얻어지는 p개의 가능도 방정식

$$\frac{\partial l}{\partial \lambda_j} = \sum_{i=1}^{N} \log\, x_{ij} - \sum_{i=1}^{N} \left(\frac{x_{ij}^{1-\lambda_j}}{\sum_k x_{ik}^{1-\lambda_k}} \right) \log\, x_{ij} - \frac{N}{2} \frac{\partial \log\, |S_y^{bc}|}{\partial \lambda_j} = 0, \quad j = 1, \cdots, p \quad (2.4)$$

을 풀어서 얻게 된다. 그런데 (2.4)는 λ_j들에 관한 비선형 방정식이기 때문에 대수적으로 해를 구할 수는 없지만 Newton-Raphson 방법이나 Nelder-Mead 심플렉스 방법과 같은 수치적인 방법을 통해서 해를 얻을 수가 있다.

2.2. 방법론

j번째 요소로 $y_j^{bc} = \frac{x_j^{\lambda_j}-1}{\lambda_j} - \frac{x_p^{\lambda_p}-1}{\lambda_p}$를 가지는 $\mathbf{y}^{bc} = \left(y_1^{bc}, y_2^{bc}, \cdots, y_{p-1}^{bc}\right)^t$를 변환된 구성비율벡터라 하면 구성비율자료의 주성분분석은 $\mathbf{y}^{bc}$의 분산을 최대화하는 정규직교 선형결합(orthonormal linear combination)들을 찾는 것이다. 이것은 결국 $\mathbf{y}^{bc}$의 공분산행렬 $\Sigma_y^{bc} = cov\left(\mathbf{y}^{bc}\right)$의 고유해와 관련이 있다. 그런데 Σ_y^{bc}의 구조는 변환에서 공통적으로 관여하고 있는 x_p의 선택에 따라 변하기 때문에 이로 인해 서로 다른 주성분들을 얻게 된다. 따라서 Σ_y^{bc}를 이용하여 분석하는 데에는 문제가 있기 때문에 다음과 같은 접근방법을 고려한다.

먼저, $\mathbf{y}^{bc}$를 중심화시킨 벡터 $\mathbf{z}^{bc}$를 고려한다. 그러면 각 요소들은

$$z_j^{bc} = \frac{x_j^{\lambda_j}-1}{\lambda_j} - \frac{1}{p}\sum_{k=1}^{p} \frac{x_k^{\lambda_k}-1}{\lambda_k}, \qquad j = 1, \cdots, p \quad (2.5)$$

가 되므로 $\mathbf{y}^{bc}$와 $\mathbf{z}^{bc}$의 관계는 $\mathbf{y}^{bc} = F\mathbf{z}^{bc}$로 표현된다. 여기서 $F = \left(I \vdots -\mathbf{1}_{p-1}\right)$이고 I는 크기 $p-1$인 항등행렬, $\mathbf{1}_{p-1}$은 각 요소들이 1인 $(p-1)\times 1$ 벡터이다. 위의 관계를 이용하면 $\mathbf{y}^{bc}$의 주성분은 $\mathbf{z}^{bc}$의 선형결합으로 표현을 할 수가 있다. 즉 ϕ^{bc}를 $\mathbf{y}^{bc}$의 주성분이라 하면 임의의 계수벡터 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{p-1})^t \in \mathbb{R}^{p-1}$에 대하여 $\phi^{bc} = \boldsymbol{\alpha}^t\mathbf{y}^{bc} = \boldsymbol{\alpha}^t F\mathbf{z}^{bc} = \boldsymbol{\beta}^t\mathbf{z}^{bc}$가 됨을 볼 수 있다. 여기서 $\boldsymbol{\beta} = F^t\boldsymbol{\alpha}$이다. 그런데 $\boldsymbol{\beta}$는 각 요소들의 합이 0이 되기 때문에 주성분은 다음과 같이

$$\phi^{bc} = \sum_{j=1}^{p} \beta_j \left(\frac{x_j^{\lambda_j}-1}{\lambda_j} \right) \quad (2.6)$$

변환된 p개의 구성비율들의 대비형태로 표현된다. 또한 ϕ^{bc}는 x_i들의 비선형함수가 되기 때문에 구성비율자료가 보이는 비선형적인 변이에 잘 대처할 수 있을 것이고 변환에 의해 합이 1이 되는 제약이 제거되었으므로 해석의 어려움을 겪지 않아도 된다.

각각의 주성분 $\phi_i^{bc}, i = 1, \cdots, p$들은 라그랑지 승수법을 이용하여 쉽게 구할 수가 있다. Σ_z^{bc}를 $\mathbf{z}^{bc}$의 공분산행렬이라고 하면 ϕ_i^{bc}들을 결정하기 위해서는 제약조건 $\boldsymbol{\beta}_i^t\mathbf{1}_p = 0$, $\boldsymbol{\beta}_i^t\boldsymbol{\beta}_i = 1$, $\boldsymbol{\beta}_i^t\Sigma_z^{bc}\boldsymbol{\beta}_j = 0$, $i < j$하에서 분산을 최대화하는 계수벡터 $\boldsymbol{\beta}_i$들을 찾아야 하고 이

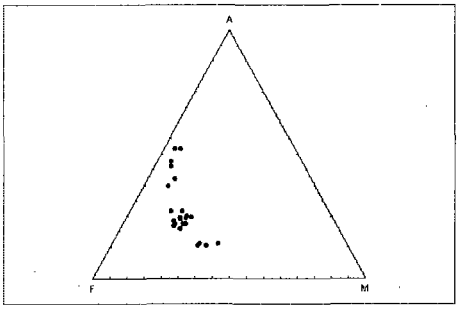


그림 3.1: AFM 자료의 삼각그림

것은 결국 Σ_z^{bc}의 고유벡터들과 관련이 있다. 따라서 Box-Cox 대비변환에 근거한 주성분분석(이하 Box-Cox 대비 주성분분석)은 특성방정식

$$\left(\Sigma_z^{bc} - \eta_i I\right)\boldsymbol{\beta}_i = \mathbf{0}, \quad i = 1, \cdots, p-1 \tag{2.7}$$

으로부터 얻어지는 $(p-1)$개의 고유해들에 기초로 하게 된다. 여기서 η_i는 Σ_z^{bc}의 i번째 고유값이고 $\boldsymbol{\beta}_i$는 이에 대응되는 고유벡터이다. 그 이유는 Σ_z^{bc}는 양반정치(positive-semidefinite) 행렬이므로 가장 작은 고유값은 0이 되고 이에 대한 고유벡터는 $\mathbf{1}_p$에 상수배가 되기 때문이다. 또한 통상적인 주성분분석에서와 마찬가지로 총변이의 측도로 $\sum_{i=1}^{p-1}\eta_i$를 사용하면 되고 첫 k개 주성분들의 설명력은 누적비율 $\sum_{i=1}^{k}\eta_i / \sum_{i=1}^{p-1}\eta_i$로 주어지게 된다.

3. 응용

전반적인 수행능력을 알아보기 위해서 Box-Cox 대비 주성분분석을 실제자료와 모의자료에 적용을 해 보았다. 실제자료로는 AFM 자료와 Panoche 토양자료를 선택하였고 유도된 주성분의 설명력과 주성분점수의 플롯을 통해서 Aitchison의 방법과 비교를 하였다.

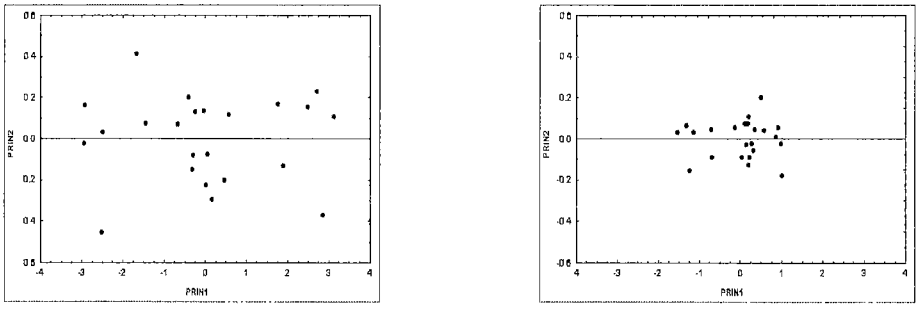
3.1. AFM 자료

AFM 자료는 스코트랜드 서부에 있는 스카이 섬의 23개 현무암 표본으로부터 산화나트륨과 산화칼륨(A), 산화철(F), 산화마그네슘(M)을 측정한 p=3인 구성비율자료로 그림 3.1에서 보듯이 곡선적인 변이형태를 가짐을 알 수 있다. 먼저, Nelder-Mead 심플렉스 방법에 의해 변환모수들의 최대가능도 추정치로 $\widehat{\lambda}_1 = -0.978$, $\widehat{\lambda}_2 = 1.736$, $\widehat{\lambda}_3 = -0.094$를 얻은 다음 이들을 사용해 자료를 변환시켰다. 표본주성분들을 얻기 위해 변환된 자료로부터 계산한 공분산행렬의 고유해들을 구하였다. 두 가지 방법을 적용한 분석결과를 표 3.1에 제시한다.

표 3.1: AFM 자료에 대한 분석결과

	Box-Cox 대비 주성분분석		Aitchison의 주성분분석	
변수	첫 번째 고유벡터	두 번째 고유벡터	첫 번째 고유벡터	두 번째 고유벡터
A	0.797	-0.178	-0.686	-0.444
F	-0.244	0.779	-0.041	0.816
M	-0.553	-0.601	0.812	-0.372
고유값	3.376	0.045	0.574	0.008
누적 비율	0.987	1.000	0.986	1.000

결과에서 보듯이 첫 Box-Cox 대비 주성분은 $\phi_1^{bc} = 0.797\left(\frac{x_1^{\hat{\lambda}_1}-1}{\hat{\lambda}_1}\right) - 0.244\left(\frac{x_2^{\hat{\lambda}_2}-1}{\hat{\lambda}_2}\right) - 0.553\left(\frac{x_3^{\hat{\lambda}_3}-1}{\hat{\lambda}_3}\right)$인 반면 Aitchison의 경우는 $\phi_1^{lc} = -0.686\log x_1 - 0.041\log x_2 + 0.812\log x_3$가 된다. 첫 주성분의 설명력을 계산해 보면 각각 98.7%, 98.6%이므로 두 방법간의 큰 차이는 없음을 알 수 있다. 보다 더 효율적인 비교를 위한 방법 중의 하나는 첫 번째 주성분점수와 두 번째 주성분점수에 대한 2차원 플롯을 동일한 척도로 그려보는 것이다. 만일 주성분점수들의 흩어짐이 비선형적인 패턴을 보이지 않는다면 전반적으로 주성분들이 전체의 변이를 잘 포착하는 것으로 생각할 수 있다. 그림 3.2에서 보는 바와 같이 두 방법 모두에서 비선형적인 패턴을 보이지 않지만 (b)보다는 (a)에서 주성분점수들이 다소 많이 흩어지는 경향을 볼 수 있다. 이상의 결과들을 종합해보면 Box-Cox 대비 주성분분석은 Aitchison의 방법만큼이나 곡선적인 변이에 대해서도 잘 대처한다는 것을 알 수 있다.



(a) (b)

그림 3.2: 주성분점수 플롯: (a) Box-Cox 대비 주성분분석 (b) Aitchison의 주성분분석

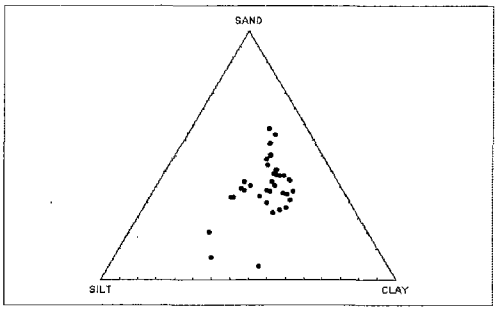


그림 3.3: Panoche 토양자료의 삼각그림

3.2. Panoche 토양자료

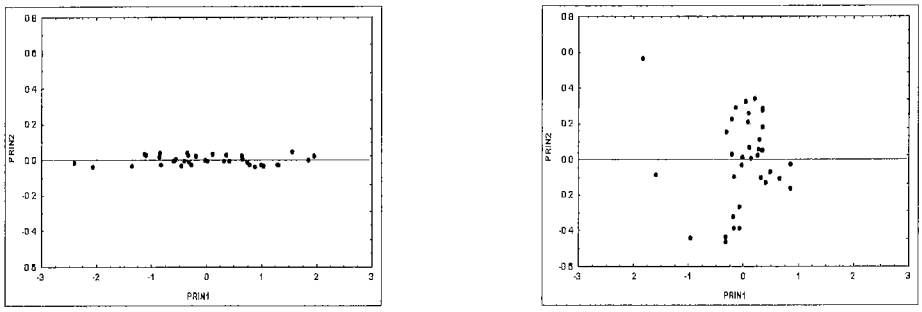
Panoche 토양자료는 Nielsen 등(1973)이 Panoche지역의 토양성분조사를 위해 20개 구획지의 서로 다른 깊이에서 얻은 토양표본들로부터 모래(sand), 침니(silt), 점토(clay)성분들을 측정한 원래의 자료에서 분석목적으로 일부분을 선택한 36개의 표본에 관한 자료이다. 그림 3.3은 자료를 삼각좌표에 플롯한 것으로 드문드문 쪼개진 형태의 모습을 보이고 있다. 변환모수들의 최대가능도 추정치로 각각 $\widehat{\lambda}_1 = 1.557$, $\widehat{\lambda}_2 = -0.847$, $\widehat{\lambda}_3 = 4.730$를 얻었고 이를 통해 자료를 변환시킨 다음 분석을 하였다.

표 3.2의 분석결과를 보면 두 방법에 의해 얻어진 첫 주성분의 설명력은 각각 99.9%, 82.9%로 나타났고 Box-Cox 대비 주성분이 Aitchison의 것보다는 설명력이 상대적으로 약 17% 정도 더 높다는 것을 알 수 있다. 그림 3.4는 두 방법간의 효율적인 비교를 위해서 주성분점수들을 플롯한 것으로 (b)에서는 뚜렷한 비선형적인 패턴을 보이는 것과는 대조적으로 (a)에서는 타원적인 패턴을 보인다. 따라서 이 경우에는 Aitchison의 방법이 변이의 포착에 효과적으로 대처하지 못하는 것 같다.

Aitchison의 방법론에서 제기될 수 있는 문제 중의 하나는 변환된 자료에 대해 다변량 정규성을 가정한 다음 표본주성분들에 기초한 통계적 추론의 가능성을 제시한 부분이다.

표 3.2: Panoche 토양자료에 대한 분석결과

	Box-Cox 대비 주성분분석		Aitchison의 주성분분석	
변수	첫 번째 고유벡터	두 번째 고유벡터	첫 번째 고유벡터	두 번째 고유벡터
Sand	-0.433	-0.692	0.748	-0.327
Silt	0.816	-0.029	-0.657	-0.485
Clay	-0.383	0.721	0.091	0.811
고유값	1.047	0.001	0.304	0.063
누적비율	0.999	1.000	0.829	1.000



(a) (b)

그림 3.4: 주성분점수 플롯: (a) Box-Cox 대비 주성분분석 (b) Aitchison의 주성분분석

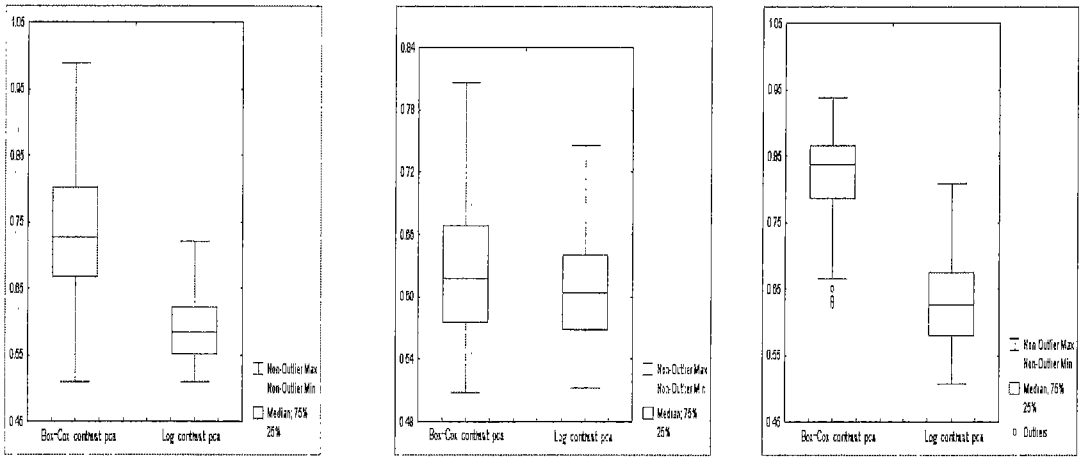
실제로 주성분분석이 자료의 요약이라는 측면에 더 큰 비중을 두고 있다고 하더라도 추론을 통한 주성분들의 모형화에 관심이 있는 경우 정규성 가정의 검토는 중요한 측면을 차지한다. 이런 측면에서 Anderson-Darling 통계량(T_{AD}), Cramer-von Mises 통계량(T_{CM}), Watson 통계량(T_W), Shapiro-Wilk 통계량(T_{SW}), 다변량 왜도와 첨도에 기초한 Mardia 통계량들(T_{MS}, T_{MK})의 계산을 통해 변환된 Panoche 토양자료에 관한 일련의 정규성 검정을 해 보았다.

표 3.3: 변환된 자료에 대한 정규성 검정결과

		Box-Cox 대비변환					
		검정통계량					
검정	변환된 변수**	T_{AD}	T_{CM}	T_W	T_{SW}	T_{MS}	T_{MK}
일변량	y_1^{bc}	0.653	0.102	0.010	0.962	.	.
	y_2^{bc}	0.232	0.035	0.035	0.980	.	.
이변량		0.289	0.020	0.026	.	.	.
다변량		0.450	0.063	0.068	.	1.119	-0.850
		Aitchison의 변환					
		검정통계량					
검정	변환된 변수**	T_{AD}	T_{CM}	T_W	T_{SW}	T_{MS}	T_{MK}
일변량	y_1^{lr}	1.639*	0.240*	0.208*	0.841*	.	.
	y_2^{lr}	1.432*	0.242*	0.208*	0.895*	.	.
이변량		1.389	0.264	0.102	.	.	.
다변량		1.275	0.194	0.047	.	32.066*	3.901*

*: $\alpha = 0.05$에서 유의

**: $y_j^{bc} = \frac{x_j^{\hat{\lambda}_j}-1}{\hat{\lambda}_j} - \frac{x_3^{\hat{\lambda}_3}-1}{\hat{\lambda}_3}$, $y_j^{lr} = \log x_j - \log x_3$, $j = 1, 2$



(a) (b) (c)

그림 3.5: 모의실험을 통해 얻은 첫 주성분의 설명력에 관한 상자그림

표 3.3의 검정결과를 보면 Box-Cox 대비변환인 경우 모든 결과들이 유의하지 않는 반면 Aitchison의 변환은 16개의 검정결과 중 10개가 유의성을 보이므로 전반적으로 다변량 정규성을 보장해 주지 못한다. 따라서 정규성 하에서 개발된 통계적인 기법들을 적용하여 추론을 할 경우 잘못된 해석을 할 가능성이 있음을 의미한다.

3.3. 모의자료분석

Box-Cox 대비 주성분분석의 전반적인 수행능력을 주성분들이 보유하게 되는 설명력 관점에서 Aitchison의 방법과 비교하기 위해 모의실험을 하였다. 표본의 크기는 50, 변수의 수는 3, 반복은 200으로 하였고 다음과 같은 일련의 과정을 통해서 구성비율자료를 생성하였다. 먼저, 양의 값을 갖는 기저자료(basis data)를 각각 대수정규, 감마, 역가우스분포에서 독립적으로 발생시킨 다음 적절한 변환을 통해 종속적인 구조를 가지는 자료로 만들었다. 그런 다음 각 행의 요소들을 행의 합으로 나누어서 구성비율자료를 얻었다. 생성시킨 구성비율자료에 두 가지 방법을 적용하여 첫 주성분의 설명력을 계산하였다. 각 방법에 의해 얻어진 결과들의 비교를 위해서 상자그림을 그려보았다.

그림 3.5의 (a)는 대수정규분포에서 생성한 구성비율자료를 분석한 결과로 Box-Cox 대비 주성분의 설명력(왼쪽 상자그림)이 Aitchison의 경우보다 높게 나타남을 볼 수 있는데 각각의 평균적인 설명력은 73.4%, 58.7%로 약 15%정도의 차이를 보였다. 역가우스분포에서 생성한 구성비율자료인 경우에도 (c)에서 보는 바와 같이 Box-Cox 대비 주성분이 상대적으로 높은 설명력을 가짐을 알 수 있다. 각각의 설명력은 82.4%, 62.7%로 Box-Cox 대비 주성분의 설명력이 약 20%정도 높게 나타났다. 반면에 감마분포에서 얻은 경우((b))에는 별 차이를 보이지 않았다. 전반적으로 Box-Cox 대비변환을 이용한 분석방법이 Aitchison이 제안한 방법보다 일정하게 좋은 성능을 보여 줌을 알 수 있다.

4. 결론

구성비율자료의 주성분분석에서 기존의 방법들이 가지는 문제들을 해결하기 위해 본 논문에서는 Box-Cox 대비변환을 이용한 새로운 형태의 방법을 제시했다. 실제자료에 적용했을 때 전반적으로 Aitchison의 방법보다 더 좋은 결과를 제공했고 모의실험에서도 비슷한 결과를 얻을 수 있었다. 또한 Box-Cox 대비 주성분은 구성비율자료가 보이는 비선형적인 변이를 비교적 잘 포착한다는 것도 주성분점수의 플롯을 통해서 확인할 수 있었다. 그리고 검정결과에서 보여주듯이 변환된 자료는 다변량 정규성 가정을 전반적으로 위배하지 않기 때문에 주성분들에 관한 통계적 추론시 잘못된 해석을 할 가능성이 Aitchison의 경우보다 상대적으로 낮음을 알 수 있다. 그러나 방법론을 적용할 때 야기되는 문제는 자료의 변환과 관련된 변환모수들을 수치적으로 구하기 위해서는 복잡한 계산을 해야 하기 때문에 변수가 많은 경우에는 많은 시간이 요구된다는 것이다.

참고문헌

[1] Aitchison, J. (1982). The statistical analysis of compositional data(with discussion), *Journal of Royal Statistical Society*, Series B, vol. 44, 139-177.

[2] Aitchison, J. (1983). Principal component analysis of compositional data, *Biometrika*, vol. 70, 57-65.

[3] Butler, J.C. (1976). Principal component analysis using the hypothetical closed array, *Journal of Mathematical Geology*, vol. 8, 25-36.

[4] Carr, D.F. (1981). Distinction between Permian and Post-Permian igneous rocks in the Southern Sydney Basin, New-South Wales, on the basis of major-element geochemistry, *Journal of Mathematical Geology*, vol. 13, 193-199.

[5] Chayes, F. (1960). On correlations between variables of constant sum, *Journal of Geophysical Responses*, vol. 60, 4185-4193.

[6] Chayes, F. and Kruscal, W. (1966). An approximate statistical test for correlations between proportions, *Journal of Geology*, vol. 74, 692-702.

[7] Coakley, J.P. and Rust, B.R. (1968). Sedimentation in an Artic lake, *Journal of Sedimentary Petrology*, vol. 38, 1290-1300.

[8] Gower, J.C. (1967). Multivariate Analysis and Multidimensional Geometry, *Statistician*, vol. 17, 13-28.

[9] Jolliffe, I.T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

[10] Le Maitre, R.W. (1968). Chemical variation within and between volcanic rock series - a

statistical approach, *Journal of Petrology*, vol. 9, 220-252.

[11] Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Mass.

[12] Nielsen, D.R., Biggar, J.W. and Erh, K.T. (1973). Spatial variability of field-measured soil-water properties, *Hilgardia*, vol. 42, 215-259.

[13] Rayens, W.S. and Srinivasan, C. (1991). Box-Cox transformations in the analysis of compositional data, *Journal of Chemometrics*, vol. 5, 227-239.

[14] Thompson, R.N., Esson, J. and Duncan, A.C. (1972). Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland, *Journal of Petrology*, vol. 13, 219-253.

[15] Webb, W.N. and Briggs, L.I. (1966). The use of principal component analysis to screen mineralogical data, *Journal of Geology*, vol. 74, 716-720.

[2000년 10월 접수, 2001년 2월 채택]

Principal Component Analysis of Compositional Data using Box-Cox Contrast Transformation

Byungjin Choi[1] Keeyoung Kim[2]

ABSTRACT

Compositional data found in many practical applications consist of non-negative vectors of proportions with the constraint which the sum of the elements of each vector is unity. It is well-known that the statistical analysis of compositional data suffers from the unit-sum constraint. Moreover, the non-linear pattern frequently displayed by the data does not facilitate the application of the linear multivariate techniques such as principal component analysis. In this paper we develop new type of principal component analysis for compositional data using Box-Cox contrast transformation. Numerical illustrations are provided for comparative purpose.

Keywords: Compositional data; Box-Cox contrast transformation; Box-Cox contrast principal component analysis.

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