Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 3, pp. 515~521

Nonlinear Canonical Correlation Analysis for Paralysis Disease Data

Yang Kyu Shin[1)]

Abstract

Categorical data are mostly found in oriental medical research. The nonlinear canonical correlation analysis does not assume an interval level of measurement. In this paper, we apply nonlinear canonical correlation analysis to quantification and explain how similar sets of variables are to one another for paralysis disease data.

Keywords : nonlinear canonical correlation, paralysis disease, quantification

1. 서론

대부분의 한의학 분야 자료에는 여러 가지 유형의 변수들이 섞여 있다. 통계적 모형에 기초한 분석기법들은 자료 분석에 사용되는 특별한 가정들을 요구한다. 그러므로 모수의 추정과 이에 대한 검증작업을 하게 된다. 예를 들어, 종속변수가 범주형인 경우 단순선형회귀분석을 실시했다고 가정해 보자. 종속 변수가 연속형이 아니므로 관계의 선형성 및 오차에 대한 가정을 점검해 볼 수 없으므로 추정된 회귀모형의 적합성을 판단할 수가 없다. 적합한지 안 한지도 검증되지 않은 모형을 적용한다면 그 결과가 어떻게 될 것인가? 최적척도법(Optimal Scaling)은 가정들 없이 자료에 적합한 해를 구할 수 있는 통계적 분석기법으로 변수의 유형에 따라 분석기법이 적용된다(Heijden(1997), Meulman(1998)). 즉, 독립변수가 범주형이고 종속변수가 연속형이면 범주형 회귀분석(Categorical Regression)을, 모든 변수들이 범주형인 하나의 자료집합에 대하여는 다중대응분석(Multiple Correspondence Analysis)을, 일부변수들이 범주형이 아닌 하나의 자료집합에 대하여는 범주형 주성분 분석(Categorical Principal Component Analysis)을 적용한다. 비선형(혹은 범주형)정준상관분석(Nonlinear Canonical Correlation Analysis)은 모든 변수들이 명목형인 둘이상의 자료집합이나

1) 경상북도 경산시 유곡동 290번지 대구한의대학교 자산운용과학과 교수
E-mail : yks@dhu.ac.kr

일부변수들이 명목형이 아닌 여러 자료집합에 대하여 적용할 수 있는 분석기법이다. Park & Huh(1996)와 Huh(1998)는 위의 방법들과 Hayashi(1988)에 의해 제시된 수량화 방법에 대하여 연구하였다. 본 연구에서는 종속 변수인 증형이 범주형 변수인 중풍증형진단자료에 최적척도법을 적용하여 독립변수인 증후들과 종속변수인 증형간의 관계를 탐색해 보고자 한다. 증형의 척도가 명목형이고 증후들의 척도가 명목형이므로 비선형정준상관분석이 적용될 수 있다. 비선형정준상관분석에서는 변수들의 척도에 대한 조건이 필요하지 않고 또한 선형관계도 가정하지 않는다.

본 연구에서는 중풍환자의 증형진단자료에 비선형정준상관분석을 적용하여 증후들과 증형간의 관계를 탐색하고 증형과 증후들의 범주를 설명할 수 있다고 생각되는 통계량을 제시하고 이를 이용하여 증형별로 증후들과의 관계를 분석하였다.

2. 분석자료

연구에 이용된 자료는 대구한의대학교부속 한방병원에서 보건복지부 보건의료기술개발연구과제 수행을 위해 수집된 것이다. 수집에 참여한 사람은 일반수련의, 전문수련의 및 전문의로 수집된 자료의 객관성을 유지하기 위하여 자료 수집 시작 전 수집에 참여하는 전원을 대상으로 각 항목별 개념과 판단기준에 대한 설명, 질문에 대한 답변, 임상시험 등의 교육을 실시하였다. 그리고 수집된 자료에 대한 최종적인 점검은 중풍전문의가 직접 환자를 보고 수집된 자료와의 일치여부를 검증하였다. 이는 한의학에서의 자료가 대부분 정성적인 성질을 갖는 것에 인한 것으로 중국에서도 동일하게 활용되고 있다. 즉, 중국에서는 두 명의 경치의생-한명의 주치의생-과주임의 순으로 자료를 수집하고 진단의 편차가 심할 경우 과주임이 결정하는 방식을 택하고 있다. 연구과제에서는 63명의 중풍환자에 대하여 37가지의 증후를 조사한 후 중국에서 제정한 <중풍병 변증 진단표준>(양사두 (1991))에 의거하여 분류한 증형별로 환자를 진단하였다. 환자의 증형은 화영, 음허양항, 담화, 습담 그리고 기혈구허로 범주형 변수이다. 37가지의 증후들 중 증형 진단에 영향을 미치지 않는 변수들도 있으므로 범주형 주성분분석을 통하여 영향을 미치는 변수들을 가려내어 이를 다시 중풍전문의에게 확인한 결과 37가지의 증후들에서 갈음(구갈, 구건, 정상), 설태질(조태, 정상, 윤태) 그리고 맥상(삭, 정상, 지)이 증형 진단에 주 영향을 미치는 변수로 선정되었다. 본 연구에서는 이들 변수들에 대한 자료만을 이용하여 분석을 하고자 한다. 분석은 SPSS/OVERALS를 이용하였다.

3. 비선형정준상관분석

비선형정준상관분석을 수행하기 위하여 먼저 각 변수들에 대한 최적척도를 선택하여야한다. 증형의 최적척도수준을 다중명목형으로 하고 증후와 관련된 변수들(갈음, 설태질, 맥상)의 최적척도수준을 순서형으로 처리하면 29번 반복에 single loss가 1.250340이고 증후와 관련된 변수들의 최적척도수준을 단일명목형으로 처리하면 26번 반복에 single loss가 1.226298이고 증후와 관련된 변수들의 최적척도수준을 다중명목형으로 처리하면 17번 반복에 single loss가 1.146794이므로 본 연구에서는 증후와

관련된 변수들의 최적척도수준을 다중명목형으로 처리하였다. 일반적으로 한의학 자료에서 범주형 변수들은 척도유형별로 분류하면 순서형보다는 명목형에 그리고 단일 명목형보다는 다중명목형의 형태라고 할 수 있다. 갈음, 설태질, 맥상 그리고 증형의 최적 척도수준을 다중명목형으로 하고 비선형정준상관분석을 수행한 결과는 <표 1>과 같다.

<표 1> 비선형정준상관분석을 수행한 결과

		차 원		합 계
		1	2	
손 실	갈 음	.426	.764	1.190
	설태질	.498	.489	.987
	맥 상	.660	.963	1.623
	증 형	.484	.303	.787
	평 균	.517	.603	1.147
고유값		.483	.370	
LDN 적합				.853

<표 1>에 의하면 제 1 고유값이 0.483이고 제 2 고유값이 0.370으로 17번의 반복을 통해 제 2축까지 고려한 고유값의 합은 0.853임을 알 수 있다. 다음 표들은 각 변수들의 범주를 2차원까지 살펴본 수량화 값이다.

<표 2> 각 변수들(갈음, 설태질, 맥상, 증형)의 수량화값

갈 음

	주변빈도	범주 수량화	
		차 원	
		1	2
구갈	7	-.333	-1.334
구건	19	1.150	.339
정상	37	-.527	.078

설태질

	주변빈도	범주 수량화	
		차 원	
		1	2
조태	22	.929	-.401
정상	15	-.820	-.848
윤태	26	-.313	.829

맥 상

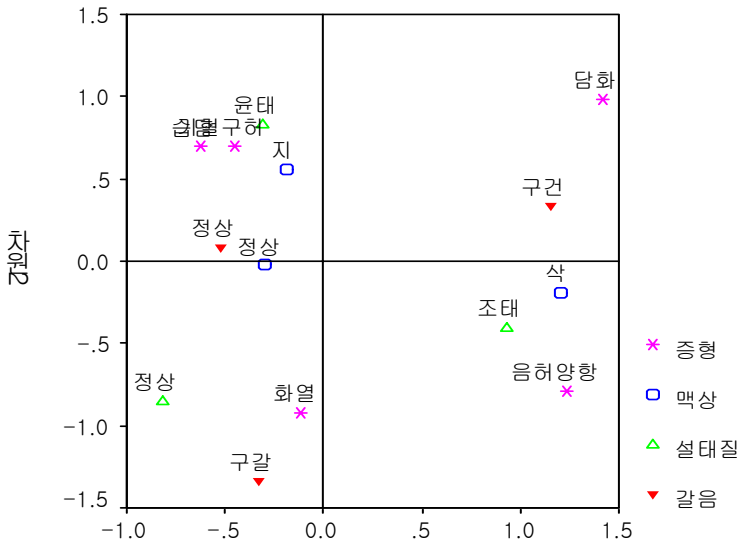
	주변 빈도	범주 수량화	
		차 원	
		1	2
삭	12	1.200	-.188
정상	45	-.295	-.025
지	6	-.189	.562

증 형

	주변 빈도	범주 수량화	
		차 원	
		1	2
화열	24	-.115	-.921
음어양항	5	1.236	-.790
담화	8	1.414	.987
습담	17	-.627	.695
기혈구허	9	-.452	.704

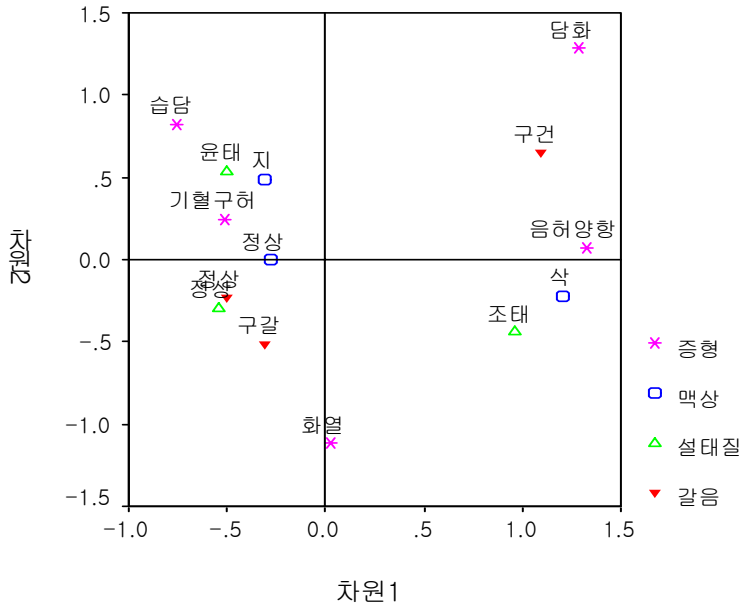
<그림 1>은 <표 2>의 수량화 값을 2차원 좌표평면위에 나타낸 것이다. 즉 갈음, 설태질 그리고 맥상의 최적 척도 수준을 다중명목형으로 하고 증형의 최적척도수준을 다중명목형으로 한 경우의 수량화값에 대한 도표이다. <그림 2>는 갈음, 설태질 그리고 맥상의 최적척도수준을 단일명목형으로 하고 증형의 최적척도수준을 다중명목형으로 한 경우의 수량화값에 대한 도표이고, <그림 3>은 갈음, 설태질 그리고 맥상의 최적척도수준을 순서형으로 하고 증형의 최적척도수준을 다중명목형으로 한 경우의 수량화값에 대한 도표이다.

<그림 1> 수량화값 도표



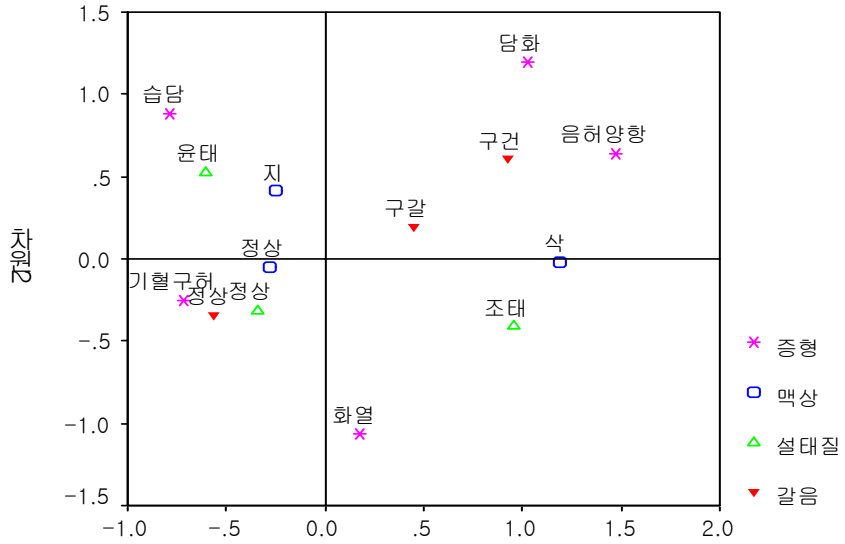
주: 갈음, 설태질, 맥상, 증형의 최적척도수준이 다중명목형인 경우

<그림 2> 수량화값 도표



주: 갈음, 설태질, 맥상의 최적척도수준이 단일명목형이고 증형의 최적척도수준은 다중명목형인 경우

<그림 3> 수량화값 도표



주: 갈음, 설태질, 맥상의 최적척도수준이 순서형이고 증형의 최적척도수준은 다중명목형인 경우

위의 세 개의 그림에서 설명변수의 하나인 갈음에 대하여 살펴보면 갈음에는 구건, 정상, 구갈이 있는데 구건은 (그림1), (그림2), (그림3)에서 모두 1사분면에 위치하나 수량화 값에 있어서 차이가 있음을 볼 수 있다. 정상은 (그림2)에서는 2사분면에 있는데 (그림2)에서는 3사분면에 있고 구갈도 (그림1)과 (그림2)에서는 3사분면에 있는 반면 (그림3)에서는 4사분면에 있다. 그리고 설명변수인 맥상, 설태질, 갈음의 수량화값을 증형의 수량화 값들과 연관시켜보면 (그림1)에서는 구건이 담화와 관련이 있고 구갈이 화열과 관련이 있는 것으로 나타나는데 (그림2)에서는 구건이 담화와 음허양항의 중간에 위치하고 구갈도 화열과 기혈구허간에 같은 위치에 있음을 알 수 있다. (그림3)에서는 구갈이 모든 증형의 유형들과 같은 거리에 있음을 볼 수 있다. 그러므로 위의 (그림1), (그림2) 그리고 (그림3)에 의하면 척도유형에 따라 분석결과가 다름을 알 수 있다. 각각에 대한 적합값은 0.853, 0.774, 0.750이므로 앞에서 언급한 바와 같이 증형과 증후들의 척도수준을 다중명목형으로 처리한 (그림1)이 가장 적합하다고 할 수 있다. 따라서 (표2)와 (그림1)로부터 증형이 화열인 환자는 갈음이 구갈이고 맥상은 정상이라고 할 수 있다. 그리고 증형이 담화로 진단되는 환자는 갈음이 구건인 증후를 나타낸다고 할 수 있다. 또, 증형이 습담과 기혈구허로 진단된 환자는 설태질이 윤태이고 맥상이 지인 증후를 나타냄을 알 수 있다.

4. 결론

비선형정준상관분석은 종속변수가 범주형인 경우 적용될 수 있는 독립변수와 종속변수간의 관계를 분석하는 기법으로 범주형인 종속변수와 역시 범주형인 독립변수를 모두 가변수를 이용하여 표현한 후 종속변수 가변수들의 선형결합과 독립변수 가변수들의 선형결합간의 상관계수를 최대화함으로써 모두 범주에 수량화 값을 부여하는 방법이다. 수량화 값은 증형과 각 증후들 간의 상관관계를 나타내는 척도로 (표2)에 각 범주에 대한 수량화 값이 계산되어 있다.

그러므로 (표2)와 같은 수량화 결과를 잘 해석하는 것이 자료 분석에서 가장 중요하다. 이 때 각 변수들의 수량화 값들을 그래프로 표현한 (그림1)과 같은 도표을 활용하는 것도 큰 도움이 된다. 도표들에서 볼 수 있듯이 변수들에 대한 척도선정에 따라 분석결과가 다르므로 적합한 척도유형을 선정하도록 주의하여야한다. (표2)와 (그림1)로부터 증형이 음허양항인 환자는 증후가 삭과 조태이고, 화열 환자는 맥상은 정상이나 구갈이 있는 상태, 담화환자는 증후가 구건이고, 습담이나 기혈구허환자는 윤태와 지의 증후와 관계가 있다고 할 수 있다. 특히 비선형정준상관분석은 독립변수 군과 종속변수 군과 같이 자료가 2개 군으로 나뉘어 지는 경우뿐만 아니라 3개 군 이상으로 나뉘어 지는 경우에도 적용 가능하므로 특성상 여러 개념이 복합적으로 합쳐져 있는 한의 진단과정을 객관화하는데 유용하게 이용될 것으로 기대된다.

참고문헌

1. 강효신, 신양규, 권영규, 박창국, 김상철 (1998). 전문가시스템을 이용한

한의진단의 객관화에 관한 연구, 연구과제최종보고서, 보건복지부.
2. 양사두, 진귀연 (1991). 실용중서의 결합진단 치료학, 중국의약과기출판사, 북경.
3. 허명회 (1998). 수량화 방법 I, II, III, IV, 자유아카데미, 서울.
4. 허명회, 양경숙 (2001). 다변량자료분석, SPSS 아카데미, 서울.
5. Hayashi, C. (1988). New Developments in multidimensional data analysis, In: Recent Developments in Clustering and Data Analysis. (Edited by Diday,E., Hayashi,C., Jambu,M. and Dhsumi,N.), 3-16. Academic Press, London.
6. Heijden, P. G. M. (1997). Multiple correspondence analysis as a tool for quantification or classification of career data, Journal of Educational and Behavioral Statistics, Vol. 22, 447-477.
7. Meulman, J. J. (1998). Optimal scaling methods for graphical display of multivariate data, Proceedings in Computational Statistics, 13th Symposium, 65-76.
8. Park, M. and Huh, M. H. (1996). Quantification Plots for Several Sets of Variables, Journal of the Korean Statistical Society, Vol. 25, No. 4, 589-601.
9. SPSS Inc. (1999). SPSS Categories 10.0, SPSS Inc., Chicago.

[2004년 4월 접수, 2004년 7월 채택]