The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 199-205

Dynamic Simple Correspondence Analysis

Yong-Seok Choi[1], Gee Hong Hyun[2], and Myung Rok Seo[3]

Abstract

In general, simple correspondence analysis has handled mainly correspondence relations between the row and column categories but can not display the trends of their change over the time. For solving this problem, we will propose DSCA(Dynamic Simple Correspondence Analysis) of transition matrix data using supplementary categories in this study. Moreover, DSCA provides its trend of the change for the future by predicting and displaying trend toward the change from a standard point of time to the next.

Keywords : Correspondence analysis, Supplementary category, Transition matrix

1. 서론

분할표 자료의 행과 열 범주들간의 대응관계를 탐구하려는 다변량 그래프적 기법으로 대응분석이 있다. 대응분석은 범주를 나타내는 변수가 둘 뿐인 이원 분할표를 위한 단순대응분석(simple correspondence analysis)과 범주를 나타내는 변수가 셋 이상인 다원 분할표를 위한 다중대응분석(multiple correspondence analysis)으로 구분하고, 대응분석의 대수적 알고리즘은 비정칙치분해(singular value decomposition)를 이용한 차원축소에 의해서 제공된다(Greenacre, 1984, 4장; 최용석, 2001, 2장). 이러한 분할표의 한 예로 전이행렬자료(transition matrix data)를 들 수 있다. 이것은 고정된 패널을 대상으로 여러 번 측정한 동일한 개수의 범주에 대하여 시간의 흐름에 따른 각 패널의 범주 사이 전이 모습을 기준시점과 기준시점의 다음 시점이라는 시차를 가지는 2차원 분할표로 정리한 것이다. 이것을 통하여 동일한 범주에 머무르는 유지정도와 다른 범주로의 전이정도를 살펴볼 수 있다.

일반적으로 단순대응분석 그림에서는 하나의 분할표 자료에 대한 행과 열의 대응관계만을 주로 다루어 왔으나 시점의 변화에 따른 행과 열 범주의 대응관계에 대한 변화의 추세를 나타내지는 못했다. 본 소고에서는 새로이 추가범주(supplementary categories)를 활용한 전이행렬자료의 동적 단순대응분석(dynamic simple correspondence analysis)을 제안하고자 한다. 이를 줄여서 동적

1) Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea
E-mail: yschoi@pusan.ac.kr
2) Ph D. Graduate Course, Department of Statistics, Pusan National University, Busan 609-735, Korea
3) Department of Statistics, Pusan National University, Busan 609-735, Korea

단순대응분석이라 부르겠다. 동적 단순대응분석은 기존의 단순대응분석에서 제공했던 행과 열의 단순한 대응관계가 아닌 시점의 변화에 따른 행과 열 범주의 변화되는 대응관계와 행 범주들의 시간적인 변화의 경향을 동적으로 보여준다. 그리고 기준시점에서 다음 시점으로의 변화의 경향도 예측하여 보여줌으로써 향후 변화의 큰 흐름을 예측할 수 있고 그 내용을 시각적으로 판단할 수 있다. 또한, 기업이미지 조사나 고객 만족도 조사 등에서 각 속성별로 어떤 회사 혹은 어떤 브랜드가 더 높은 평가를 받았는지 보려 할 때 단순대응분석을 이용할 수 있다. 이것의 장점은 수치로 구성된 분할표 대신 그림으로 결과를 보여줌으로써 해석이 쉽다는 것이다. 그러나 조사가 여러 번 시행되었을 경우 전체 조사에서의 결과를 하나의 그림으로 설명할 수 없는 단점이 있다. 이것을 해결하기 위해서 동적 단순대응분석을 사용하면 될 것이다.

2 동적 단순대응분석

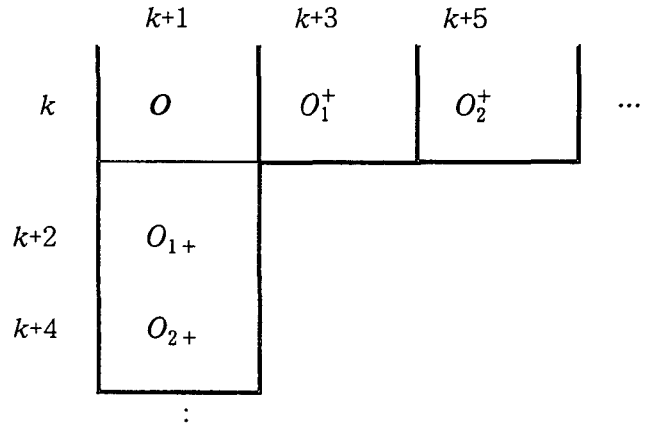
<표 1> 전이행렬자료

시점	$k+1$							합
	i \ j	1	2	⋯	j	⋯	I	
	1	$y_{11}(k+1)$	$y_{12}(k+1)$	⋯	$y_{1j}(k+1)$	⋯	$y_{1I}(k+1)$	$y_1(k)$
	2	$y_{21}(k+1)$	$y_{22}(k+1)$	⋯	$y_{2j}(k+1)$	⋯	$y_{2I}(k+1)$	$y_2(k)$
	:	:	:	:	:	:	:	:
k	:	:	:	:	:	:	:	:
	i	$y_{i1}(k+1)$	$y_{i2}(k+1)$	⋯	$y_{ij}(k+1)$	⋯	$y_{iI}(k+1)$	$y_i(k)$
	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:
	I	$y_{I1}(k+1)$	$y_{I2}(k+1)$	⋯	$y_{Ij}(k+1)$	⋯	$y_{II}(k+1)$	$y_I(k)$

<표 1>에서 k는 기준시점을 나타내며, $k+1$은 기준시점의 다음 시점을 나타내고, k=1, 2, ⋯ , T 개의 시행횟수(조사횟수)를 나타낸다. i번째 행은 기준시점 k에서의 각 개체가 속하는 상태를 j번째 열은 기준시점의 다음 시점에서 각 개체가 속하는 상태를 나타내며 i, j=1, 2, ⋯ ,I 개의 상태가 있다. $y_{ij}(k+1)$는 기준시점 k에서 i상태였다가 기준시점의 다음 시점인 $k+1$에서는 j상태가 되는 개체의 수를 나타내고, $y_i(k)$는 기준시점 k에서 i상태인 모든 개체의 수를 나타낸다.

이와 같은 전이행렬자료의 시점의 변화에 따른 행과 열 범주들 간의 대응관계를 살펴보기 위해서 크기가 $I \times I$인 전이행렬자료 O를 기준으로 시점을 달리하는 동일한 크기의 전이행렬자료들을 추가하면 <그림 1>과 같이 된다.

<그림 1>에서 k는 기준시점을 나타내며 $k+1$은 기준시점의 다음 시점을 나타낸다. 그리고 O_{1+}는 기준이 되는 전이행렬자료 O 다음에 만들어진 전이행렬자료이고, O_1^+는 O_{1+}다음에 생



<그림 1> 전이행렬자료의 추가되는 범주 표시

성된 전이행렬자료이다. O_{2+}는 O_1^+다음에 생성된 전이행렬자료이고, O_2^+는 O_{2+}다음에 생성된 전이행렬자료이다.

이러한 전이행렬자료들을 추가범주를 활용한 단순대응분석의 알고리즘(최용석, 2001, 4장)을 활용하기로 하자. 일반적으로 행과 열의 수가 각각 n과 $p(\leq n)$인 $n\times p$ 이원분할표를 행렬로 표현하면 $O=(o_{ij}), o_{ij}\geq 0,\ i=1,\cdots,n\ ;\ j=1,\cdots,p$이다. 그리고 1_n과 1_p를 크기가 $n\times 1$과 $p\times 1$으로 각각 n개와 p개가 1인 원소를 갖는 벡터라 하면 $o_{++}=1_n' O 1_p$은 O의 모든 원소들의 합이 된다. 그러면 대응행렬 $F=(f_{ij}),\ f_{ij}=o_{ij}/o_{++},\ i=1,\cdots,n\ ;\ j=1,\cdots,p$의 일반화비정칙값분해는

$$F=AD_uB' \tag{1}$$

이다. 식 (1)에서 $A'D_r^{-1}A=B'D_c^{-1}B=I$이고 $D_r=diag(f_{1+},\cdots,f_{n+})$과 $D_c=diag(f_{+1},\cdots,f_{+p})$는 크기가 $n\times n$과 $p\times p$인 대각행렬이다. $D_u=diag(1,\lambda_1,\cdots,\lambda_{p-1})$는 비정칙값(singular value)을 대각원소로 하는 대각행렬이다. $n\times p$인 자료행렬 O에 n_s개의 행과 p_s개의 열이 추가되었다고 하고 O_+는 크기가 $n_s\times p$인 추가행자료행렬이고 O^+는 크기가 $n\times p_s$인 추가열자료행렬이다. 그러면 추가행과 열프로파일행렬 R_s와 C_s도 다음과 같이 구해진다.

$$R_s=diag(O_+1_p)^{-1}O_+\ ,\ C_s=diag(O^{+\prime}1_n)^{-1}O^{+\prime}\ .$$

추가행과 열프로파일 행렬의 좌표점 X_s와 Y_s는 식 (2)과 같이 된다.

$$X_s=R_sD_c^{-1}B\ ,\ Y_s=C_sD_r^{-1}A \tag{2}$$

이것을 확장해서 시점 k에서의 추가행프로파일 R_{sk}와 시점 k+1에서의 추가열프로파일 행렬 $C_{s(k+1)}$을 구하면 식 (3)과 같이 된다. 여기서 k는 k=k+2, k+3, … 이다.

$$R_{sk} = diag(O_{i+}1_n)^{-1}O_{i+}\ ,\ C_{s(k+1)} = diag(O_i^{+\prime}1_n)^{-1}O_i^{+\prime}\ ,\ i=1,2,\cdots \tag{3}$$

그리고 추가행과 열프로파일 행렬의 저차원 공간상의 좌표점 X_{sk}와 $Y_{s(k+1)}$를 구하면 식 (4)와 같이 된다.

$$X_{sk} = R_{sk}D_c^{-1}B\ ,\ Y_{s(k+1)} = C_{s(k+1)}D_r^{-1}A \tag{4}$$

식 (4)는 동적 단순대응분석 그림의 좌표를 제공하는 행렬이며 이것에서 얻어지는 좌표 그림을 이용한 분석을 동적 단순대응분석이라 한다.

3. 응용사례

현실적으로 우리 주변에서 생성되는 많은 자료들은 시점에 따라 변하는 자료들이 많고 우리의 관심 또한 시점의 변화에 따른 자료들의 변화추이를 알고 싶어 하는 경우가 많다. 다음은 그것의 한 예로서 하버드대학교 통계학과 학생 187명을 대상으로 눈을 감게 한 후 신호가 주어지면 숫자 1, 2, 3 중에서 하나만을 선택하도록 한 실험의 결과이다. 5회까지의 결과만을 사용하여 총 4개의 전이행렬자료를 <표 2>와 같이 만들었다(Bishop 외 2인, 1975, p. 264).

전이행렬자료에서 알고자 하는 것은 첫째, 전체실험에서의 선택되는 숫자의 변화모습과 둘째, 다음번에는 어떤 숫자가 많이 선택받게 될 것인가이다. 이것은 동적 단순대응분석을 통해서 알 수 있다. 우선 1회와 2회 실험 결과의 전이행렬자료를 가지고 단순대응분석을 하면 다음과 같은 <그림 2>를 얻는다.

<표 2>의 a)를 보면 1회 실험에서 1을 선택한 학생은 2회 실험에서는 3을 많이 선택하였고, 1회 실험에서 2를 선택한 학생은 2회 실험에서는 1을 많이 선택하였다. 그리고 1회 실험에서 3을

<표 2> 숫자생성실험 전이행렬자료

a) 1회와 2회

1회 \ 2회	1	2	3	합
1	10 (0.167)	16 (0.267)	34 (0.567)	60
2	42 (0.575	10 (0.137)	21 (0.288)	73
3	29 (0.537)	16 (0.296)	9 (0.167)	54
합	81 (0.433)	42 (0.225)	64 (0.342)	187

b) 2회와 3회

2회 \ 3회	1	2	3	합
1	23 (0.284)	18 (0.222)	40 (0.494)	81
2	16 (0.381)	11 (0.262)	15 (0.357)	42
3	22 (0.344)	26 (0.406)	16 (0.250)	64
합	61 (0.326)	55 (0.294)	71 (0.380)	187

c) 3회와 4회

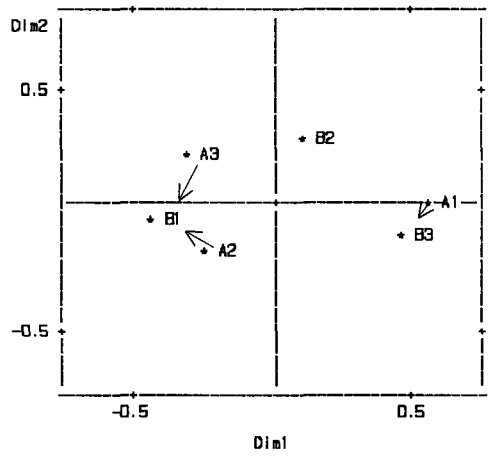
3회 \ 4회	1	2	3	합
1	11 (0.180)	33 (0.541)	17 (0.279)	61
2	22 (0.400)	10 (0.182)	23 (0.418)	55
3	34 (0.479)	27 (0.380)	10 (0.141)	71
합	67 (0.358)	70 (0.374)	50 (0.267)	187

d) 4회와 5회

		5회			
		1	2	3	합
4회	1	20 (0.299)	21 (0.313)	26 (0.388)	67
	2	32 (0.457)	18 (0.257)	20 (0.287)	70
	3	24 (0.480)	17 (0.340)	9 (0.180)	50
	합	76 (0.406)	56 (0.299)	55 (0.294)	187

e) 예측된 4회와 5회

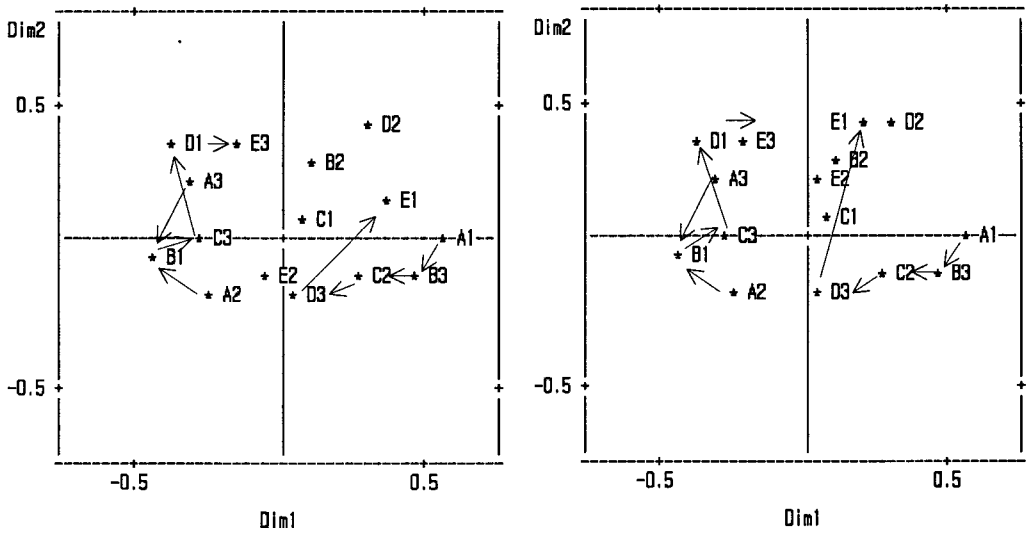
		5회			
		1	2	3	합
4회	1	15 (0.22)	22 (0.33)	30 (0.45)	67
	2	33 (0.47)	13 (0.18)	24 (0.35)	70
	3	22 (0.45)	18 (0.36)	10 (0.19)	50
	합	70	53	64	187



<그림 2> 1회와 2회의 전이행렬자료 단순대응분석 그림

선택한 학생은 2회 실험에서 1을 많이 선택하였다는 것을 알 수 있다. 이것을 단순대응분석 그림으로 확인해보면 다음과 같다. <그림 2>에서 알파벳 대문자는 실험순서를 나타내고 숫자는 1, 2, 3중에서 선택한 숫자를 뜻한다. 즉, A1는 1회 실험(A)에서 1을 선택한 것을 뜻하고 B1는 2회 실험(B)에서 1을 선택한 것을 뜻한다. 1축(Dim1)을 기준으로 보면 오른쪽의 A1→B3가 대응하는 것을 볼 수 있고, 왼쪽에서는 A2→B1과 A3→B1이 대응관계를 보이는 것을 볼 수 있다. 여기서 대응의 화살표를 1회 실험에서 선택된 숫자가 2회 실험을 했을 때 많이 선택하는 숫자의 흐름으로 보면 <표 2>의 *a)*에서의 결과와 동일한 것을 알 수 있다. 하지만 전이행렬자료의 단순대응분석은 전체실험에서의 숫자선택의 변화되는 모습을 나타내지는 못하며 전이행렬자료만큼의 단순대응분석 그림을 그려야 하는 번거로움이 있다. 따라서 전이행렬자료가 많더라도 전체내용을 하나의 2차원 그림으로 나타낼 수 있는 동적 단순대응분석이 필요하게 되었다. 따라서 전체 전이행렬자료 <표 2>의 *a)*~*d)*에 대해서 동적 단순대응분석을 실시하면 <그림 3>의 동적 단순대응분석 그림 (가)를 얻게 된다. 동적 단순대응분석 그림에서는 이제까지 단순대응분석 그림으로 표현할 수 없었던 전체 시행에서의 숫자선택의 변화되는 흐름을 한 눈에 파악할 수 있다. 먼저 <표 2>을 보고 전체적인 숫자선택의 변화 흐름을 살펴보면 다음과 같다. 1회 실험에서 1을 선택한 학생들의

변화흐름은 A1→B3→C2→D3→E1순이고, 1회 실험에서 2를 선택한 학생들의 변화흐름은 A2→B1→C3→D1→E3순이다.



(가) (나)

<그림 3> <표 2>의 *a)~d)*와 <표 2>의 *a)~c)*와 *e)*에 대한 각각의 동적 단순대응분석 그림 (가)와 (나)

마지막으로 1회 실험에서 3을 선택한 학생들의 변화흐름은 A3→B1→C3→D1→E3순이다. 이 결과를 <그림 3>의 (가)에서 확인하여보면 제1축(Dim1)을 기준으로 오른쪽에서는 A1→B3→C2→D3→E1로 대응되면서 변화되는 것을 볼 수 있다. 왼쪽에서는 A2→B1→C3→D1→E3와 A3→B1→C3→D1→E3가 서로 대응되면서 변화되는 것을 볼 수 있다. 단순대응분석에서는 두 실험간의 숫자의 변화흐름을 살펴보았지만 동적 단순대응분석에서는 전체 실험에서의 숫자의 변화흐름을 2차원상의 그림에서 동적으로 살펴볼 수 있었다. 특히 전이행렬자료의 단순대응분석 그림과 전이행렬자료를 하나하나 대조하면서 숫자선택의 변화흐름을 추적할 필요 없이 한눈에 파악할 수 있다는 것이 가장 큰 장점이라 할 수 있다. 이제 추정된 전이확률 값을 이용한 동적 단순대응분석의 예측에서 예측의 정도가 어느 정도인지 그림으로 알아보려고 한다.

<표 2>의 *d)*와 *e)*를 비교하면 4회 실험에서 1을 선택한 학생들은 5회 실험에서 3을 많이 선택하게 되며, 4회 실험에서 2를 선택한 학생들은 5회 실험에서 1을 많이 선택하며, 4회 실험에서 3을 선택한 학생들은 5회 실험에서는 1을 많이 선택하는 것을 알 수 있다. 즉, 셀 안의 수치는 정확하게 맞지는 않지만 변화의 경향은 같음을 알 수 있다. <그림 3>의 (나)는 실제 4회와 5회의 전이행렬자료인 <표 2>의 *d)* 대신에 *e)*의 전이행렬자료를 전체 실험 자료에 넣고 동적 단순대응분석을 한 것으로 <그림 3>의 (가)와 비교해 보면 E1과 E2의 위치는 틀리지만 대응되는 그룹 내에 위치하고 있음을 알 수 있다. 따라서 동적 단순대응분석 그림은 예측의 정도가 높다는 것을 알 수 있다.

4. 결론

단순대응분석 그림에서는 하나의 분할표 자료에 대한 행과 열의 대응관계만을 주로 다루어 왔으나 시점의 변화에 따른 행과 열 범주의 대응관계에 대한 변화의 추세를 나타내지는 못했다. 본 소고에서는 새로이 동적 단순대응분석을 제안해서 기존의 단순대응분석에서 제공했던 행과 열의 단순한 대응관계가 아닌 시점의 변화에 따른 행과 열 범주의 변화되는 대응관계와 행 범주들의 시간적인 변화의 경향을 동적으로 볼 수 있었다. 그리고 무엇보다 기준시점에서 다음 시점으로의 변화의 경향도 예측하여 보여줌으로써 향후 변화의 큰 흐름을 예측할 수 있고 그 내용을 시각적으로 판단할 수 있었다.

이와 같은 기법을 Tracking 조사에서나 패널조사에 사용한다면 전체 조사에서의 각 범주별 변화의 모습과 앞으로의 변화되는 모습을 예측하는데 유용하게 사용할 수 있을 것이다.

참고문헌

[1] 최용석 (2001), 「SAS 대응분석의 이해와 응용」, 자유아카데미, 서울.

[2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975), *Discrete Multivariate Analysis : Theory and Practice*, The MIT Press, London.

[3] Greenacre, M. J.(1984), *Theory and Application of Correspondence Analysis*, Academic Press, London.

[2004년 10월 접수, 2005년 2월 채택]