응용통계연구 제 12권 2호, 1999년, pp. 433-447

순환모형에 대한 EM 알고리즘의 초기값 선정방법의 개선*

정미숙[1] 김성호[2]

요 약

검사관련 능력과 문항점수사이의 관계를 모형화하기 위해 사용한 순환모형에서 관측불능인 능력상태변수를 비롯한 모든 변수들이 범주형 변수라 가정하자. 이 범주형 자료를 위한 모수추정문제를 다루기 위해 EM 방법을 이용했는데, EM 방법은 사용하기에 편리하지만 순환모형에 대한 추정값이 적절하지 않는 경우가 발생한다. 그 주된 원인중의 하나로 초기값 선정의 잘못을 들 수 있는데, 본 논문에서는 이 외에 구조상의 결함도 그 원인이 됨을 경험적으로 보였다. 따라서 구조적 결함을 먼저 해결하면 보다 효과적인 초기값을 선정할 수 있으리라 기대한다.

1. 서론

검사관련 능력과 문항점수와의 관계를 구체적으로 모형화하려는 시도는 최근들어서 많이 있었는데, 특히 문항점수가 검사관련 능력들과 인과적으로 관계되어 있다는 것은 Greeno와 Simon(1988)이 실험적 연구를 통해서 잘 보여 주었다. 능력과 능력 사이의 관계는 주로 선수관계(prerequisite relation)로 되어 있고 능력과 문항사이의 관계는 인과관계로 되어있는데, 이러한 관계들을 표현하기에 가장 적합한 모형은 순환모형(recursive model(Lauritzen and Wermuth(1983))이다.

이 순환모형이 범주형 변수 $X_1, X_2, \cdots, X_K$로 구성되어 있다 하자. 여기서 $X_{L+1}, \cdots, X_K$ $(1 \leq L < K)$는 종점(terminal node)으로 관측가능한 문항점수변수이고, $X_1, \cdots, X_L$은 비종점(non-terminal node)으로 관측불능인 능력상태변수를 나타낸다 하자. 능력들사이의 화살표는 하위능력에서 상위능력으로 향하고, 능력에서 문항점수로의 화살표는 원인-결과관계(cause-effect relation)를 나타낸다. 능력들간에 화살표가 없으면 그 능력들은 서로 주변적으로 독립 (marginally independent)임을 의미하며, 모든 능력변수가 주어져 있다 할 때, 문항점수들은 조건부 독립관계에 있게 된다. 피험자가 어떤 능력을 소유하고 있으면 해당되는 능력상태 변수의 값은 1이고 그 능력을 소유하고 있지 않으면 0이며, 문항을 맞히었으면 문항점수는 1이고 못 맞히었으면 0으로 나타낸다.

순환모형에 포함된 능력과 관련된 변수들의 확률적인 추정값을 구하려 할 때, 로그선형모형(log-linear model; Fienberg, 1980 and Whittaker, 1990)을 적합시키기 위한 IPF (interative proportional fitting) 방법(Bishop, Fienberg and Holland, 1975 and Agresti, 1990)이나

* 한국학술진흥재단의 박사후 연수과정 지원에 의함.

1) (305-701) 대전시 유성구 구성동 373-1, 한국과학기술원, 박사후 연수과정
2) (305-701) 대전시 유성구 구성동 373-1, 한국과학기술원, 부교수

구조등식모형(structural equation model(SEM); Bollen 1989)을 위한 LISREL(Jöreskog and Sörbom, 1986)과 EQS(Bentler, 1985)와 같은 통계패키지가 있다. 그러나 IPF방법은 위계적 로그선형모형을 대상으로 했으므로, 본 논문의 순환모형에 적절하지 않고, SEM 모형은 연속확률변수에 대한 통계모형이기 때문에 범주형 변수만을 다루는 우리의 목적에는 적합하지가 않다.

SEM 모형의 특수한 형태로 잠재변수와 그 변수에 영향을 받는 관측 가능한 변수에 대한 통계적인 모형으로 측정모형(Measurement model)이 있는데, 본 논문에서 고려하는 범주형 능력상태변수와 문항점수변수에 대한 모수추정문제를 이 모형에 그대로 적용할 수 없다. SEM 모형은 기본적으로 연속확률변수들의 모형이며 정규분포를 가정하고 있기 때문이다.

ω를 $X_1, \cdots, X_K$의 첨자집합(index set)이라 하자. 순환모형의 모형에서 변수 X_i로 향한 화살표의 꼬리에 있는 변수들의 첨자집합을 $pa(i)$ (여기서 "pa"는 parent에서 나왔음)라 나타내고, $fa(i) = \{i\} \bigcup pa(i)$라 하고, x_θ에서의 빈도수를 $n(x_\theta)$로 나타내고, $E(n(x_\theta)) = m(x_\theta)$로 표시하자. 그러면 x_ω는 $K-$변수 분할표에서의 특정칸의 위치를 가리키고 $\phi \subset \theta \subseteq \omega$인 θ에 대해서 x_θ는 θ의 구성원소들에 해당하는 변수들에 대한 주변 분할표(marginal contingency table)에서의 칸의 위치를 가리킨다. X_i의 범주의 개수를 J_i라 할 때 집합 $\{x_\omega\}$의 크기는 $\prod_{i=1}^{K} J_i$가 된다.

본 논문에서 다루는 순환모형의 결합확률(joint probability)은

$$P(x_\omega) = \prod_{i=1}^{K} P(x_i | pa(i) = x_{pa(i)}) \tag{1.1}$$

이다. 여기서 $pa(i) = \phi$이면 $P(x_i|pa(i) = x_{pa(i)}) = p(x_i)$이다. 표본의 크기를 n이라 할 때, 모형 (1.1)은 지수족에 속하므로, 이 모형의 최대우도추정치(MLE)는

$$\hat{P}(x_\omega) = \frac{\hat{m}(x_\omega)}{n}, \quad \hat{P}(x_i|x_{pa(i)}) = \frac{\hat{m}(x_{fa(i)})}{\hat{m}(x_{pa(i)})}$$

으로 주어진다(Bishop, Fienberg and Holland, 1975). 앞에서 언급했듯이, $X_1, \cdots, X_L$은 관측불능변수이고, $X_{L+1}, \cdots, X_K$는 관측가능변수이다. 더 나아가서 모형 (1.1)에서는 $i \in \{L+1, \cdots, K\}$인 X_i에 대해서 $pa(i) \subseteq \{1, \cdots, L\}$가 성립하도록 되어 있다. 따라서 모든 i에 대해서 MLE $\hat{P}(x_i|x_{pa(i)})$를 구하기 위해서 EM 방법을 사용해야 하는데, 구체적 내용은 조금씩 다른 관점에서 Lauritzen(1995), Kim(1997), 그리고 Jeong, Kim 과 Jeong(1998)에 기술되어 있다.

EM 방법은 사용하기에 간편하고 이해하기가 쉽지만, 추정값이 적절하지 않는 경우가 발생한다. 그래서 본 논문에서는 바람직한 추정값을 얻기 위해, 충분히 큰 자료에서, 적절하지 못한 현상을 발생시키는 원인을 찾고, 복잡한 구조를 가진 순환모형의 일부 구조를 제거했을 때, 어떤 조건하에서 어떤 현상이 일어나는 지를 모의실험을 통해 알아보려 한다.

본 논문은 4절로 되어 있다. 2절은 충분히 큰 자료에서 순환모형의 모수추정에 바람직하지 못한 현상을 발생시키는 원인을 서술하였고, 3절은 순환모형의 실제 구조에서 일부

구조를 제거했을 때 일어나는 현상에 대해 모의실험을 통해 논의하였다. 4절은 관련된 논의를 하면서 결론을 맺었다.

2. 뒤틀림현상과 그 원인

EM 알고리즘에 의한 추정값이 적절하지 않는 경우로서 능력 $i(i=1,\cdots,k)$에 대해 능력의 수준이 $x_i \le x_i'$이고 적어도 하나의 능력 $j(1 \le j \le k)$에 대해 능력수준이 $x_j < x_j'$이 주어진 경우에 능력변수나 문항점수변수 X_t가 1일 확률이

$$P(X_t = 1|X_1 = x_1, \cdots, X_k = x_k) \ge P(X_t = 1|X_1 = x_1', \cdots, X_k = x_k')$$

인 경우가 있다. 즉 능력을 소유하고 있는 정도가 높을수록 어떤 능력이 있을 확률이나 어떤 문항을 맞출 확률이 높아야 하나 그 반대의 결과가 나오는 경우를 의미한다. 이런 현상을 뒤틀림현상(order-distortion)이라고 부르겠다.

충분히 큰 자료에서 일어날 수 있는 뒤틀림현상으로는
(1) EM 알고리즘 수행시 바람직하지 못한 초기값을 선정했을 때와
(2) 설정된 구조가 실제 구조와 서로 다를 때
등을 생각할 수 있는데 (1)은 이미 연구결과가 있고 (2)는 3절에서 다룰 계획이다

EM 방법은 주어진 자료가 온전히 관측되지 않았을 때, 관측되지 않은 부분을 통계적으로 처리하여, 관련된 모형에 대한 모수추정을 효과적으로 수행할 수 있도록 고안된 통계적 기법이다. 기존의 모수추정방법들을 그대로 적용할 수 있고, 방법 자체가 대체로 간단하다는 장점은 있지만, 단점으로는 수렴속도가 일반적으로 느리다는 것과 모수추정값의 불안정성이다(Wu, 1983). 여기서 수렴속도는 컴퓨터의 발달과 더불어 어느 정도 극복될 수 있는 문제이나, 모수추정값의 불안정성은 EM 방법이 안고 있는 중요한 문제이다.

이 불안정성은 모형의 복잡성에 따라서 그 정도가 달라진다. 모형의 복잡성은 우도함수에 반영되는 데, 일반적으로 모수추정값은 모형이 복잡하고 추정할 모수의 수가 많아질 수록 추정값이 초기값에 많이 의존하는 양상을 띤다. 그리고 자료의 크기가 충분히 크지 않으면 더욱 심각한 원인으로 작용할 수 있다.

Jeong, Kim과 Jeong(1998)에서는 초기값의 선택이 추정값에 얼마나 민감한 지를 모의실험을 통해 보였고, 선정된 몇몇 기본구조들에 대한 초기값 선택규준을 마련하여 보다 안정적인 모수추정값을 얻을 수 있는 방법을 제시하였다.

3. 설정된 구조가 실제 구조와 다를 때

본 절은 본 논문의 핵심 부분에 해당된다. 적절치 못한 추정결과의 대표적인 경우라 할 수 있는 뒤틀림현상 원인중의 하나로 예상되는 모형구조상의 결함에 대해서 이것이 미치는 영향이 어느 정도인지를 경험적으로 탐색하고자 한다. 구조결함의 유형으로는, 하나는 설정된 구조와 실제 구조 사이에 포함관계없이 서로 다른 경우이고 다른 하나는 설정된 구조가 실제 구조보다 화살표 하나가 부족한 경우이다. 설정된 구조가 실제 구조보다 더 큰

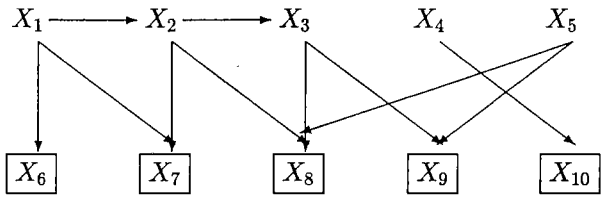


그림 3.1: 순환모형 M

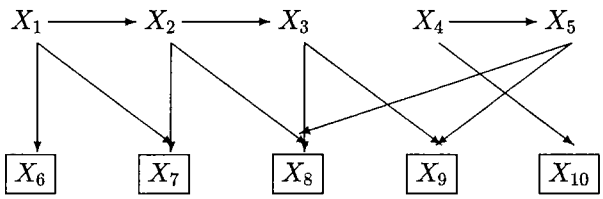


그림 3.2: 순환모형 M1

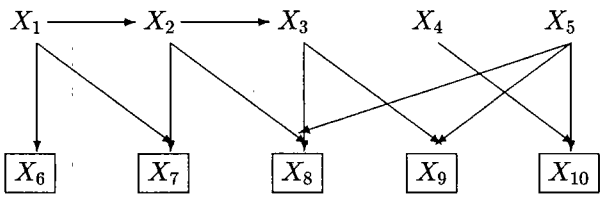


그림 3.3: 순환모형 M2

경우에는 추정 결과에 의해서 실제 구조로 접근해 갈 수 있으므로, 이 경우는 문제가 되지 않는다. 따라서 위 두 구조 결함 유형을 고려하면 되리라 보는데, 아래 두 소절에서 유형별로 다루고자 한다.

3.1. 설정된 구조가 실제 구조에 포함되지 않으면서 실제와 서로 다를 때

순환모형의 실제 구조에서 극히 일부의 구조를 변형하여 모수추정을 했다면 어떤 현상이 일어나는 지를 모의실험을 통해 알아보자. 이를 위해 그림 3.1에 있는 비교적 간단한 순환모형을 예로 들었다.

순환모형 M은 5개의 능력상태변수 $X_1, \cdots, X_5$와 문항점수변수 $X_6, \cdots, X_{10}$으로 구성되어 있다. X_1은 X_2에, X_2는 X_3에, 그리고 X_1은 X_4에 대해서 선수관계에 있고, 문항점수

변수들은 그림의 화살표대로 각각 능력변수들과 관계(인과관계)되어 있다.

순환모형 M에서 결합확률은

$$P(x_\omega) = P(x_1, \cdots, x_5)P(x_6|x_1)P(x_7|x_1, x_2)P(x_8|x_2, x_3, x_5)P(x_9|x_3, x_5)P(x_{10}|x_4) \tag{3.1}$$

이다. 단, $x_i = 0, 1, (i = 1, \cdots, 10)$.

식 (3.1)의 $P(x_1, \cdots, x_5)$는 순환모형의 능력변수들간의 구조에 따라 결정되는데 여기서는

$$P(x_1, \cdots, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_1)P(x_5)$$

이다. 그러므로 순환모형 M에서 추정해야 할 모수는 식 (3.1)의 우변에 있는 주변확률과 조건부확률들로 표 3.1의 1열에 나열된 확률들이다. 예를들어, 문항 X_7에 대한 모수는 4개로 $P(X_7 = 1|X_1 = x_1, X_2 = x_2)$, $x_1, x_2 = 0, 1$이다.

순환모형 M 대신 그림 3.2에 있는 순환모형 M1처럼 능력상태변수에 대해서 구조의 일부를 변형시킨 모형과 그림 3.3에 있는 순환모형 M2처럼 능력상태변수와 문항점수변수에 대해서 구조의 일부를 변형시킨 모형을 실제 모형인 것으로 설정해 보자. 즉, 순환모형 M1은 순환모형 M의 능력상태변수 X_4가 X_1과 서로 주변적으로 독립이고 X_5에 영향을 미치는 경우로, 순환모형 M2는 순환모형 M의 문항점수변수 X_{10}에 영향을 주는 능력상태변수를 X_4와 X_5로 한 경우이다.

표 3.1에서 4열에 있는 확률들은 순환모형 M1에서 추정해야 할 모수이고 6열에 있는 확률들은 순환모형 M2에서 추정해야 할 모수이다. 표 3.1에 있는 결과들은 실제자료의 크기를 50,000으로, EM 과정에서 초기값 선택시 사용한 자료의 크기는 100,000으로, 그리고 EM 과정에서 종결기준치(stopping criterion)는 0.001로 하였다. 각 순환모형의 모수들에 대한 초기 확률값은 실제 확률값과 같도록 했으며, 구조가 바뀐 변수 X_j들에 대한 모수의 초기 확률값은 $P(X_j = 1)$이 실제 확률값과 같도록 조정해 주었다.

표 3.1에서 3열은 순환모형 M에 대한 모수의 추정값으로 이 값을 얻는데는 몇초의 수행시간(팬티엄급 586 컴퓨터에서)만을 필요로 했지만 5열에 있는 순환모형 M1에 대한 추정값은 3시간 이상, 7열에 있는 순환모형 M2에 대한 추정값은 2시간 정도의 수행시간을 필요로 했다. 이처럼 모형의 구조 일부를 변형시키면 추정값의 수렴속도가 저하되어 최종적인 결과를 얻기까지 컴퓨터 수행시간을 많이 소요되는 경우가 많으며, 순환모형의 구조가 복잡할수록 더욱 심각하였다.

추정값에 일어난 뒤틀림현상을 살펴보자. 5열에 있는 순환모형 M1에 대한 추정값들에서 X_8, X_9에 대한 모수의 추정값에 뒤틀림현상이 일어났으며, 7열에 있는 순환모형 M2에 대한 추정값들에서도 X_8, X_9에 대한 모수의 추정값에 뒤틀림현상이 일어났다.

모형의 Pearson χ^2 적합도를 비교해 보자. 문항이 5개이므로 5차원 분할표에서 2^5개의 관측도수를 관측해야 하고, 순환모형 M과 M1은 추정해야 할 모수가 28개이므로 χ^2 적합도 자유도는 $2^5 - 1 - 28 = 3$이며, 순환모형 M2는 추정해야 할 모수가 30개이므로 χ^2 적합도 자유도는 1이다. 표 3.1에서 3열에 있는 추정값 χ^2 적합도는 9.2이고, 5열은 10480.7, 7열은 6.9이다.

표 3.1: 순환모형 $M, M1, M2$에 대한 모수의 추정값

M	P	$\hat{P}_M$	M1	$\hat{P}_{M1}$	M2	$\hat{P}_{M2}$
$P(X_1=1)$	.3217	.3233	$P(X_1=1)$	.3225	$P(X_1=1)$	.3233
$P(X_2=1\|X_1)$			$P(X_2=1\|X_1)$		$P(X_2=1\|X_1)$	
0	.1500	.1452	0	.1475	0	.1418
1	.8500	.8556	1	.8102	1	.7946
$P(X_3=1\|X_2)$			$P(X_3=1\|X_2)$		$P(X_3=1\|X_2)$	
0	.3226	.3229	0	.3076	0	.2759
1	.9033	.9026	1	.7977	1	.8541
$P(X_4=1)^*$	.3558	.3571	$P(X_4=1)$	.3527	$P(X_4=1)^*$	.3515
$P(X_4=1\|X_1)$					$P(X_4=1\|X_1)$	
0	.1000	.0997			0	.0792
1	.8950	.8973			1	.9214
$P(X_5=1)$	.6500	.6506	$P(X_5=1)^{**}$	.6866	$P(X_5=1)$	.6910
			$P(X_5=1\|X_4)$			
			0	.5558		
			1	.9267		
$P(X_6=1\|X_1)$			$P(X_6=1\|X_1)$		$P(X_6=1\|X_1)$	
0	.1500	.1494	0	.1486	0	.1493
1	.9500	.9486	1	.9523	1	.9486
$P(X_7=1\|X_1,X_2)$			$P(X_7=1\|X_1,X_2)$		$P(X_7=1\|X_1,X_2)$	
0 0	.1000	.0947	0 0	.0959	0 0	.0991
0 1	.4656	.4604	0 1	.4589	0 1	.4428
1 0	.6500	.6767	1 0	.7490	1 0	.7590
1 1	.9500	.9490	1 1	.9456	1 1	.9486
$P(X_8=1\|X_2,X_3,X_5)$			$P(X_8=1\|X_2,X_3,X_5)^\sharp$		$P(X_8=1\|X_2,X_3,X_5)^\sharp$	
0 0 0	.0500	.0508	0 0 0	.1035	0 0 0	.1880
0 0 1	.1500	.1439	0 0 1	.1071	0 0 1	.0353
0 1 0	.1500	.1478	0 1 0	.1954	0 1 0	.4054
0 1 1	.4500	.4544	0 1 1	.4533	0 1 1	.4657
1 0 0	.1500	.1477	1 0 0	.0004	1 0 0	.3503
1 0 1	.4500	.4628	1 0 1	.5339	1 0 1	.1758
1 1 0	.4500	.4544	1 1 0	.6963	1 1 0	.7262
1 1 1	.8500	.8562	1 1 1	.8136	1 1 1	.8378
$P(X_9=1\|X_3,X_5)$			$P(X_9=1\|X_3,X_5)^\sharp$		$P(X_9=1\|X_3,X_5)^\sharp$	
0 0	.1000	.0980	0 0	.0747	0 0	.0081
0 1	.2200	.2173	0 1	.1828	0 1	.2655
1 0	.4800	.4768	1 0	.7805	1 0	.8723
1 1	.9300	.9297	1 1	.9079	1 1	.8185
$P(X_{10}=1\|X_4)$			$P(X_{10}=1\|X_4)$		$P(X_{10}=1\|X_4,X_5)$	
0	.1000	.0998	0	.1430	0 0	.0970
1	.7500	.7566	1	.6864	0 1	.1245
					1 0	.6708
					1 1	.7680

$*$: 추정해야 할 모수는 아니고 다른 모형과 $P(X_4=1)$를 비교하기 위함

$**$: 추정해야 할 모수는 아니고 다른 모형과 $P(X_5=1)$를 비교하기 위함

$\sharp$: 뒤틀림현상이 일어난 변수

따라서, 최종적인 추정값에 뒤틀림현상이 일어나거나, 최종결과를 얻는데 유달리 수행시간이 많이 걸리거나 또는 χ^2 적합도가 매우 클 경우에는 순환모형의 구조설정에 이상이 있을 수 있으므로 변수들간의 관계를 재 검토할 필요가 있음을 알 수 있다. 비록 한개의 순환모형만을 예로 보였지만 다른 순환모형을 통한 모의실험도 이와 유사한 결과가 나왔다.

3.2. 실제 구조에서 연결 하나를 제거했을 때

앞 소절에서는 순환모형의 구조를 일부 변형시키면 뒤틀림현상과 수렴속도 저하등의 현상이 일어나는 경우를 고찰했고, 본 절에서는 순환모형의 일부 구조를 변형시켜도 앞의 현상이 일어나지 않는 경우에 대해 알아보려 한다.

모형의 구조를 바람직하게 설정하고 초기값을 잘 선정해도 모형의 구조가 복잡하면 최종적인 추정값을 얻는데 소요되는 컴퓨터 수행시간은 많이 걸린다. 또한 순환모형의 구조가 복잡하면 추정할 모수의 수가 많아져, 자료의 크기가 상대적으로 작으면 분할표에서 빈칸(empty cell)이 많이 발생하므로 바람직한 추정값을 얻을 수 없는 경우를 자주 본다.

이제 화살표가 하나 제거된 경우를 좀 더 구체적으로 살펴보자. 편의상 X_i에서 X_j로 향하는 화살표를 $< i, j >$로 나타내겠다. $pa(j) = \{1, 2\}$라 하고, 모형구조 설정과정에서 화살표 $< 1, j >$가 빠졌다 하자. 이것이 전체 모형의 모수추정에 어떤 영향을 미치게 될까? 실제로 추정해야 할 모수가 $P(x_j|x_{pa(j)})$인데, 설정된 모형에서는 $P(x_j|x_2)$가 추정된다. 여기서 간과해서는 안될 부분이 X_1과 X_2 사이의 관계이다. 둘 사이가 서로 독립일 때와 상호관계가 긴밀할 때, $< 1, j >$의 유무가 모수추정에 미치는 영향에 많은 차이가 있으리라 예상된다. 극단적으로 $X_1 = X_2$이면 $< 1, j >$의 유무가 아무 영향을 못 미치겠지만, 둘 사이가 서로 독립에 가까울수록, 그 영향은 커지리라 예상된다.

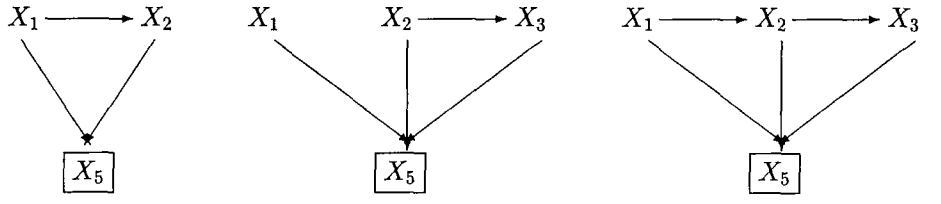
화살표 하나의 유무가 모형 전체의 모수추정에 미치는 영향을 이해하는 데 용이하도록, 비교적 간단한 모형구조를 사용하여 모의실험을 하였다. 두 범주형 변수사이의 상호관계를 표현하는 척도로 교차곱비(cross-product ratio(cpr))를 쓰는데, 0과 1을 취하는 두 이항변수들의 cpr은 $\frac{P(0,0)P(1,1)}{P(0,1)P(1,0)}$ 으로 주어진다. cpr값이 1인 것과 해당되는 두 변수의 독립성이 동치관계(Bishop, Fienberg, and Holland, 1975)인 것은 잘 알려져 있다.

한개의 문항점수변수에 영향을 주는 몇개의 능력상태변수들 중에서 화살표를 제거한 능력상태변수들을 고려하기 위해 기본구조(basic structure)들을 그림 3.4에 열거해 놓았다. 그림 3.4에 있는 모형들에서 원(circle) 안에 있는 변수는 능력상태변수이고 상자(box)안에 있는 변수는 문항점수변수이다. 예컨대, 기본구조 BS-1에서 X_1과 X_2는 능력상태변수이고 X_5는 문항점수변수이다.

본 절에서는 아래 4가지 경우의 추정값에 어떤 현상이 일어나는 지를 알아보았다:

(경우 1) 기본구조 BS-1에서 $< 1, 5 >$가 없을 때
(경우 2) 기본구조 BS-2에서 $< 2, 5 >$가 없을 때
(경우 3) 기본구조 BS-3에서 $< 1, 5 >$가 없을 때
(경우 4) 기본구조 BS-3에서 $< 2, 5 >$가 없을 때

기본구조 BS-1에 대한 분석을 위해 그림 3.5에 있는 순환모형 A를, 기본구조 BS-2에 대한 분석을 위해 그림 3.6에 있는 순환모형 B를, 기본구조 BS-3에 대한 분석을 위해 그림



기본구조 BS-1 기본구조 BS-2 기본구조 BS-3

그림 3.4: 기본구조 BS-1, BS-2, BS-3

3.7에 있는 순환모형 C를 고려했다. 여기서 세개의 순환모형으로 나눈 이유는 변수의 수가 15개 이상이면 컴퓨터 수행시간이 무척 많이 소요되기 때문이다.

경우 1 : 기본구조 BS-1에서 $< 1,5 >$가 없는 경우

순환모형 A에서 능력상태변수 X_1, X_2와 문항점수변수 X_7 사이의 인과관계는 기본구조 BS-1과 같다. 순환모형 A에서 능력상태변수 X_1과 X_2의 cpr은

$$cpr_{1,2} = \frac{P_{X_1X_2}(1,1)P_{X_1X_2}(0,0)}{P_{X_1X_2}(1,0)P_{X_1X_2}(0,1)} = \frac{P_{X_2|X_1}(1|1)P_{X_2|X_1}(0|0)}{P_{X_2|X_1}(1|0)P_{X_2|X_1}(0|1)}$$

이다.

표 3.2에서 표 3.5까지의 모의실험결과들은 실제 자료크기와 EM 과정에서 초기값 선택시 사용한 자료의 크기를 500,000으로 했으며, 각 표의 2열 또는 3열에 있는 cycle은 추정값을 얻기까지 EM 과정을 반복 수행한 횟수를 기록했다.

표 3.2는 순환모형 A에서 $< 1,7 >$를 제거한 변형된 모형을 가지고 $cpr_{1,2}$에 따라 모수추정을 수행한 결과로, χ^2 적합도 자유도는 28이다. 표 3.2의 5열에서 X_7과 관련된 모수의 추정값을 고려하지 않은 것은, 문항점수변수 X_7에 대한 구조를 변형한 모형에서 추정값을 구했으므로 X_7에 관련된 실제 확률값 P와 추정된 확률값 $\hat{P}$의 차이를 비교할 수 없었기 때문이다. 다른 표들도 그와 관련된 문항점수변수에 대한 구조를 변형한 모형을 가지고 수행한 결과이므로, 그 문항점수변수와 관련된 추정값의 비교는 고려하지 않았다.

표 3.2에 의하면 순환모형 A에서 X_2가 X_1에 의존하는 정도가 강하면 X_1에 대한 X_7의 관계를 무시해도 되나 X_2가 X_1에 의존하는 정도가 약하면($cpr_{1,2}$이 1에 접근하면) X_1에 대한 X_7의 관계를 무시하면 추정값에 이상이 있음을 보여 준다. 따라서, 기본구조 BS-1에서 X_1과 X_2의 관계가 강하면 $< 1,5 >$가 없어도 전체모형의 모수추정에 별 문제가 없으나, 그 관계가 약해질수록 $< 1,5 >$의 유무가 모수추정에 미치는 영향이 커짐을 알 수 있다.

경우 2 : 기본구조 BS-2에서 $< 2,5 >$가 없는 경우

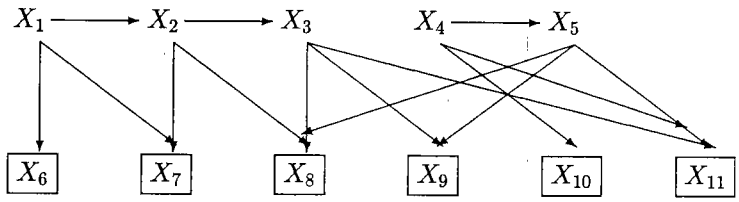


그림 3.5: 순환모형 A

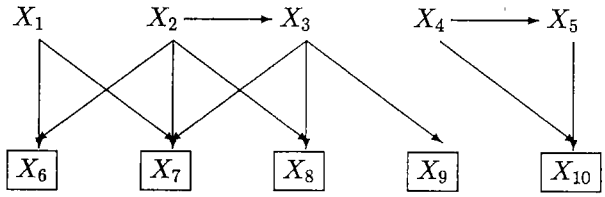


그림 3.6: 순환모형 B

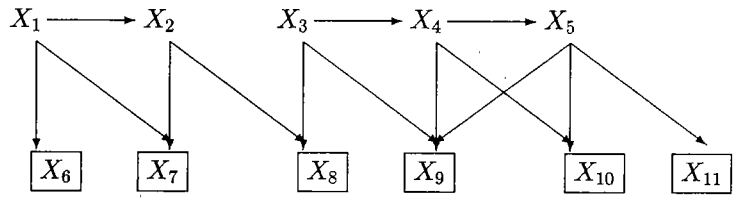


그림 3.7: 순환모형 C

순환모형 B에서 능력상태변수 X_1, X_2, X_3와 문항점수변수 X_7 사이의 인과관계는 기본구조 BS-2와 같다. 순환모형 B에서 능력상태변수 X_2와 X_3의 cpr은

$$cpr_{2,3} = \frac{P_{X_3|X_2}(1|1)P_{X_3|X_2}(0|0)}{P_{X_3|X_2}(1|0)P_{X_3|X_2}(0|1)}$$

이다.

표 3.3은 기본구조 BS-2에서 X_2와 X_5의 관계를 무시해도 좋은 경우를 알아보기 위해, 순환모형 B에서 $< 2,7 >$를 제거한 모형을 써서 EM 과정을 수행한 결과로, χ^2 적합도 자유도는 6이다. 표 3.3에 의하면 순환모형 B에서 X_3가 X_2에 의존하는 정도가 강하면 X_2에 대한 X_7의 관계를 무시해도 되나 X_3가 X_2에 의존하는 정도가 약하면($cpr_{2,3}$가 1에 접근하면) X_2에 대한 X_7의 관계를 무시하면 추정값에 이상이 있음을 보여 준다. 따라서, 기본구

표 3.2: 순환모형 A에서 X_1와 X_2의 $cpr_{1,2}$에 따른 추정결과 비교(경우 1에 해당)

$cpr_{1,2}$	cycle	χ^2	뒤틀림현상 비율	$\lvert P-\hat{P}\rvert \geq 0.2$인 비율*
614.33	23,258	27.09	0/11	0/33
171.00	24,741	23.24	0/11	0/33
32.11	21,095	26.38	0/11	0/33
5.44	23,443	22.08	0/11	2/33
1.83	24,480	116.74	0/11	3/33
1.17	28,514	313.69	1/11	5/33

* : X_7 변수에 대한 모수 4개를 제외한 33개 모수에 대해 적용함.

표 3.3: 순환모형 B에서 X_2와 X_3의 $cpr_{2,3}$에 따른 추정결과 비교(경우 2에 해당)

$cpr_{2,3}$	cycle	χ^2	뒤틀림현상 비율	$\lvert P-\hat{P}\rvert \geq 0.2$인 비율*
614.33	108,430	9.50	0/10	0/21
81.00	6,479	6.80	0/10	0/21
27.00	4,442	5.03	0/10	0/21
4.30	4,351	7.30	0/10	1/21
1.27	864	31.05	0/10	1/21
1.08	837	39.30	0/10	1/21

* : X_7 변수에 대한 모수 8개를 제외한 21개 모수에 대해 적용함.

조 BS-2에서 X_2과 X_3의 관계가 강하면 $< 2,5 >$의 유무가 전체모형의 모수추정에 별 영향을 안 미치지만, 그 관계가 약하면 많은 영향을 미침을 알 수 있다.

경우 3 : 기본구조 BS-3에서 $< 1,5 >$가 없는 경우

순환모형 C에서 능력상태변수 X_3, X_4, X_5와 문항점수변수 X_9 사이의 인과관계는 기본구조 BS-3와 같다. 순환모형 C에서 능력상태변수 X_3와 X_4의 cpr은

$$cpr_{3,4} = \frac{P_{X_4|X_3}(1|1)P_{X_4|X_3}(0|0)}{P_{X_4|X_3}(1|0)P_{X_4|X_3}(0|1)}$$

이다.

표 3.4은 기본구조 BS-3에서 X_1와 X_5의 관계를 무시해도 좋은 경우를 알아보기 위해, 순환모형 C에서 $< 3,9 >$를 제거한 변형 모형을 써서 모수추정을 수행한 결과로, χ^2 적합도 자유도는 35이다. 표 3.4에 의하면 순환모형 C에서 X_4가 X_3에 의존하는 정도가 강하면 X_3에 대한 X_9의 관계를 무시해도 되나 X_4가 X_3에 의존하는 정도가 약하면($cpr_{3,4}$이 1에 접

표 3.4: 순환모형 C에서 X_3와 X_4의 $cpr_{3,4}$에 따른 추정결과 비교(경우 3에 해당)

$cpr_{3,4}$	cycle	χ^2	뒤틀림현상 비율	$\lvert P-\hat{P}\rvert \geq 0.2$인 비율*
614.33	10,370	40.17	0/11	0/24
81.00	8,890	34.92	0/11	0/24
32.11	6,655	34.65	0/11	2/24
12.00	7,475	32.11	0/11	2/24
2.25	4,187	28.78	0/11	5/24
1.49	1,701	35.31	0/11	5/24

* : X_9 변수에 대한 모수 8개를 제외한 24개 모수에 대해 적용함.

근하면) X_3에 대한 X_9의 관계를 무시하면 추정값이 실제 확률값과 차이가 남을 알 수 있다. 따라서, 기본구조 BS-3에서 X_1과 X_2의 관계가 강하면 $< 1,5 >$ 를 제거한 모형을 고려할 수 있음을 알 수 있다.

기본구조 BS-3는 기본구조 BS-1을 확장한 구조로 두 구조에 대한 결론이 유사하게 나온 것처럼, 다른 모의실험에 의하면 기본구조 BS-3를 더 확장한 구조에도 그리고 기본구조 BS-3에서 능력상태변수 X_3가 X_1과 X_2에 의존하는 구조에도 그대로 적용이 가능했다.

경우 4 : 기본구조 BS-3에서 $< 2,5 >$가 없는 경우

순환모형 C에서 능력상태변수 X_4와 X_5의 cpr은

$$cpr_{4,5} = \frac{P_{X_5|X_4}(1|1)P_{X_5|X_4}(0|0)}{P_{X_5|X_4}(1|0)P_{X_5|X_4}(0|1)}$$

이다.

표 3.5는 기본구조 BS-3에서 X_2와 X_5의 관계를 무시해도 좋은 경우를 알아보기 위해, 순환모형 C에서 $< 4,9 >$를 제거한 변형된 모형을 써서 EM 알고리즘 과정을 수행한 결과로, χ^2 적합도 자유도는 35이다. 표 3.5에 의하면 순환모형 C에서 $< 4,9 >$를 실제 모형에서 제거했을 때 $cpr_{3,4}$가 1에 가까울 때에 $\hat{P}$가 P와 0.2이상 차이있는 경우가 3개인 것으로 나타났다. $cpr_{4,5}$가 1에 가까울 때에는 이런 현상이 덜한 것은 다른 변수들의 X_3, X_4, X_5와의 관계를 살펴 보아야 한다. 예컨대, $\{4,5\} = pa(10)$인데, $\{3,4\} \subseteq pa(i)$인 i는 $i = 9$인 경우 밖에 없다. 만약에 $\{3,4\} = pa(8)$ 이라면 $cpr_{3,4}$와 $cpr_{4,5}$가 아마 유사한 정도로 모수추정에 영향을 미쳤을 것이다. 이런 현상은 경우 2의 표3.3에서 볼 수 있다. 순환모형 A에서 능력상태변수 X_2, X_3, X_5와 문항점수변수 X_8 사이의 인과관계는 경우 4와 유사한 경우로, X_3가 X_2에 의존하는 정도가 강하면 $< 2,8 >$을 무시해도 추정값에 큰 변화가 없음을 알 수 있었다.

표 3-2에서 표 3-5까지를 종합해 보면 $i \in pa(j)$일 때, x_i가 $pa(j)$안의 다른 변수들과의 상호관계가 약할 수록 $< i,j >$의 제거가 전체모형의 모수측정에 미치는 영향이 큼을 알 수

표 3.5: 순환모형 C에서 $cpr_{3,4}$와 $cpr_{4,5}$에 따른 추정결과 비교(경우 4에 해당)

$cpr_{3,4}$	$cpr_{4,5}$	cycle	χ^2	뒤틀림현상 비율	$\lvert P-\hat{P}\rvert \geq 0.2$인 비율*
81.00	614.33	9,175	43.03	0/11	0/24
81.00	34.81	3,362	30.95	0/11	0/24
81.00	1.49	2,669	31.95	0/11	0/24
34.81	614.33	6,619	42.87	0/11	0/24
34.81	34.81	2,549	32.30	0/11	0/24
34.81	1.49	2,576	35.94	0/11	0/24
1.49	614.33	2,112,571	29.72	1/11	3/24
1.49	34.81	983,271	35.34	0/11	3/24
1.49	1.49	475,196	64.31	0/11	3/24

* : X_9 변수에 대한 모수 8개를 제외한 24개 모수에 대해 적용함.

있다.

4. 결론

실제 검사자료를 가지고 능력상태변수와 문항점수변수 사이에 존재하는 모수의 추정값을 구하려고 할 때, EM 방법의 장점에도 불구하고 최종적인 추정값이 왜곡되거나 터무니없는 컴퓨터 수행시간을 많이 소요하게 되는 경우가 종종 발생하였다. 그런데, 본 논문에서 고려한 뒤틀림현상 원인들을 먼저 신중하게 고려한 뒤, 분석에 임하면 수행시간의 단축뿐만아니라 바람직한 결론으로 유도되는 경우가 많았다.

본 논문에서는, 충분히 큰 자료에서, 일어날 수 있는 뒤틀림현상의 원인을 분석했는데 그 결과로는 EM 알고리즘 수행시 바람직하지 못한 초기값의 선정과 순환모형의 잘못된 구조 설정 등을 들 수 있었다.

서로 독립관계가 강한 능력상태변수들과 인과관계를 이루고 있는 문항점수변수에서, 능력상태변수와 문항점수변수 사이의 화살표를 제거하면 수행시간, χ^2 적합도, 추정의 정도 등이 매우 떨어지나 서로 종속관계가 강한 능력상태변수들과 인과관계를 이루고 있는 문항점수변수에서, 능력상태변수와 문항점수변수 사이의 화살표를 제거하면 실제 구조를 가지고 수행한 결과와 모수추정에 있어서 별로 차이가 없음을 알 수 있었다.

따라서 순환모형의 모수추정을 EM 알고리즘을 사용해서 할 때, 추정값에 뒤틀림현상 등 적절치 못한 현상이 발생하면, 초기값을 점검하거나 구조상의 결함이 없는 지를 점검할 것을 권한다. 물론 초기값 잘못과 구조상의 결함이 동시에 있을 수도 있다. 이 때에는 초기값 선정을 Jeong, Kim과 Jeong(1998)의 방법에 따라서 한 다음에 그래도 추정결과가 이상하면 구조설정을 다시 할 필요가 있다고 본다. 실제 구조보다 더 크게 설정했을 때에는 별 문제가 없겠으나 실제 구조보다 작거나 포함 관계없이 서로 다르게 설정했을 때에는, 3절

에서 본 바와 같이 추정결과가 적절치 못하게 나올 가능성이 높다. 구조적 결함을 해결한 뒤에 초기값을 선정하면 보다 효과적인 초기값 선정이 되리라 기대한다.

참고문헌

[1] Agresti, A.(1990). *Categorical Data Analysis.* New York: John Wiley & Sons.

[2] Bentler, P.M.(1985). *Theory and Implementation of EQS: A Structural Equations Program.* Los Angeles: BMDP Statistical Software.

[3] Bishop, Y.m., Fienberg, S.E., and Holland, P.W.(1975). *Discrete Multivariate Analysis: Theory Practice.* Sixth printing. Cambridge, MA: MIT Press.

[4] Bollen, K.A.(1989). *Structural Equations with Latent Variables.* NY: John Wiley & Sons.

[5] Fienberg, S. E.(1980). *The Analysis of Cross-Classified Categorical Data.* 2nd ed. Cambridge, MA : MIT Press.

[6] Greeno, J.G. and Simon, H.A.(1988). Problem solving and reasoning. In R.C. Atkinson, R.J. Hernstein, G. Lindzey, and R.D. Luce(Eds.), *Stevens' handbook of experimental psychology(2nd ed.)*, Vol II. New York: John Wiley & Sons, 1988, 589-672.

[7] Jeong, M. S., Kim, S.-H., and Jeong, K. M.(1998). Initial value selection in applying an EM algorithm for recursive models of categorical variables. *Journal of The Korean Statistical Society*, 27, 1, 25-55.

[8] Jöreskog, K.G. and Sörbom, D.(1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Method.* Mooresville, IN: Scientific Software, Inc.

[9] Kim, S.-H.(1997). Iterative proportional fitting for nonhierarchical log-linear models, *Comm. Stat.-Theory and methods*, 26, 6, 1443-1460.

[10] Lauritzen, S. L.(1995). The EM algorithm for graphical association model with missing data, *Computational statistical & Data Analysis*, 19, 191-201.

[11] Lauritzen, S. L. and Wermuth, N.(1983). Graphical and recursive model for contingency tables. *Biometrika*, 70, 3, 537-552.

[12] Whittaker, J.(1990), *Graphical Models in Applied Multivariate Statistics.* Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons.

[13] Wu, C. F.(1983). On the convergence properties of the EM algorithm. *The Annals of*

Statistics, 11, 1, 95-103.

[1998년 7월 접수, 1999년 7월 최종수정]

An improvement on initial value selection in applying an EM algorithm for recursive models*

Mi-Sook Jeong[1] Sung-Ho Kim[2]

ABSTRACT

We assume that all the variables that are involved in a recursive model of item-score variables and ability-state variables are categorical. An EM algorithm is applied for this model. The EM is relatively easy to use though, we often see unreasonable-looking estimates from the algorithm. One of the reasons for that is an ill-selection of initial values for the EM. And it is shown in this paper that an ill-selection of the model structure is also a reason for the undesirable estimates. Therefore, a better selection of the initial values for the EM is anticipated when any defect in the selected model-structure is cleared off before hand.

* This research was supported by a Post-Doctor program of the Korea Research Foundation, 1997.

1) In Post-Doctor program, Korea Advanced Institute of Science and Technology, Daejon, 305-701, Korea.

2) Associate Professor, Korea Advanced Institute of Science and Technology, Daejon, 305-701, Korea.