응용통계연구 제 **19**권 2호, *2006*년, *pp. 331–338*

수량화 제3 방법의 축소 해

허명회[1] 이용구[2]

요 약

수량화 제3 방법은 일본의 하야시(Hayashi)에 의해 창안된 교차표 분석 기법으로 사회조사 및 마케팅 조사 자료의 분석에서 매우 유용하다. 그러나 반응빈도가 작은 일부 범주들이 특이하게 큰 수량화 값을 갖는 경우가 있어 불안정한 과잉 해석으로 이어지기도 한다. 본 교신은 이 문제를 해결하고자 한 연구로서 수량화 제3 방법을 새로 정식화하고 축소 해(shrinkage solution)를 제안할 것이다. 그리고 실제 조사 자료에 새 방법론을 적용해 보고자 한다.

주요용어: 하야시의 수량화 제3 방법, 축소 해, 상태지수, 수량화 플롯, 대응분석.

1. 연구배경

$F = \{f_{ij}\}$를 $p \times q$ 계수형(count) 자료 행렬이라고 하자. 수량화 제3 방법은 다음 정식화로부터 행 수량화 값 $p \times 1$ 벡터 x와 열 수량화 값 $q \times 1$ 벡터 y를 구하는 문제이다 (허명회, 1998, p.62).

$$\begin{aligned} &\text{maximize} \quad x^t F y/n \\ &\text{subject to} \quad x^t D_r x/n = 1, \quad y^t D_c y/n = 1. \end{aligned} \tag{1.1}$$

여기서 r과 c는 각각 행렬 F의 $p \times 1$ 행 주변, $q \times 1$ 열 주변 벡터이다 ($r = F1_q$, $c = F^t 1_p$). 표현을 단순화하기 위하여 앞으로 $F/n \to F$, $r/n \to r$, $c/n \to c$로 표기하기로 한다. 이에 따라 식 (1.1)의 정식화는

$$\begin{aligned} &\text{maximize} \quad x^t F y \\ &\text{subject to} \quad x^t D_r x = 1, \quad y^t D_c y = 1 \end{aligned} \tag{1.2}$$

이 된다. $x = 1_p$이고 $y = 1_q$이면 식 (1.2)의 제약식이 만족되면서 목적식의 값이 항상 1이 되므로 해에서 제외한다. 그 외의 경우 식 (1.2)의 목적식은 0보다 크고 1보다 작은 값을 취한다.

1) (136-701) 서울특별시 성북구 안암동 5가, 고려대학교 통계학과, 교수
E-mail: stat420@korea.ac.kr
2) (156-756) 서울시 동작구 흑석동, 중앙대학교 통계학과, 교수
E-mail: leeyg@cau.ac.kr

수량화 제3 방법은 1940년대 후반부터 1950년대 전반에 걸쳐 일본의 C. Hayashi에 의해 창안된 교차표 분석 기법으로 사회조사 및 마케팅 조사 자료의 분석에서 매우 유용하다 (岩坪, 1987; 駒澤, 1992). 그러나 반응빈도가 작은 일부 범주들이 특이하게 큰 수량화 값을 갖는 경우가 있는데 이는 식 (1.2)의 제약식 $x^tD_rx = 1$과 $y^tD_cy = 1$ 하에서 행 범주 비율 벡터 r의 i번째 요소 r_i $(i = 1, \cdots, p)$ 또는 열 범주 비율 벡터 c의 j번째 요소 c_j $(j = 1, \cdots, q)$가 작으면 해당하는 수량화값 x_i와 y_j가 비교적 자유롭게 움직일 수 있기 때문이다. 그런데 큰 수량화 값은 수량화 플롯에서 시각적 효과가 크므로 분석자는 이런 패턴을 과잉하게 해석하기 쉽다. 본 교신은 이런 문제를 해결하고자 한 시도로서, 2절에서 수량화 제3 방법을 새로 정식화하고 축소 해(shrinkage solution)를 제안할 것이다. 그리고 3절에서 실제 다중응답 조사자료에 새 방법론을 적용해보고 마지막으로 4절에서 대응분석(correspondence analysis)에의 적용을 언급하면서 소고를 맺을 것이다.

2. 제안 방법론

본 연구에서는 자료행렬 F가 희박(sparse)한 행 또는 열을 갖는 경우에 특히 유효한 수량화 제3 방법의 풀이에 대하여 고찰할 것이다. 이런 경우엔 불안정한 행 수량화 값 벡터 x 또는 열 수량화 값 벡터 y를 얻기 쉽다. 이런 문제를 극복하기 위하여 다음과 같이 수량화 제3 방법의 정식화를 일부 변형할 것을 제안한다.

$$\begin{aligned} &\text{maximize} \quad x^tFy - k_1x^tx - k_2y^ty \\ &\text{subject to} \quad x^tD_rx = 1, \quad y^tD_cy = 1 \end{aligned} \tag{2.1}$$

여기서 k_1과 k_2는 x^tx과 y^ty에 붙는 비음의 상수(nonnegative constant)이다. 새 정식화 식 (2.1)에서 k_1x^tx와 k_2y^ty는 각각 x의 제곱 크기(norm)와 y의 제곱 크기에 비례하는 페널티(penalty)를 의미하게 된다. 따라서 식 (2.1)의 해는 식 (1.2)의 해에 비교하여 다소 작아진다. 이와 같은 축소 해(shrinkage solution) 구하기는 제임스-스타인(James-Stein) 추정, 선형회귀에서의 능형 추정(ridge estimation), 스플라인 모델에서의 정규화(regularization) 또는 벌점화(penalization) 등의 측면에서 최근 광범위하게 받아들여지고 있다 (Casella와 Berger, 1990, p.497; Hastie, Tibshirani와 Friedman, 2001, p.59, p.144).

수량화 제3 방법의 새로운 정식화 식 (2.1)에 대한 해를 구해보자. 식 (2.1)의 목적식과 2개의 제약식을 라그랑지 승수로 묶은 함수

$$h(x, y) = x^tFy - k_1x^tx - k_2y^ty - \lambda_1(x^tD_rx - 1) - \lambda_2(y^tD_cy - 1)$$

에서

$$\partial h/\partial x = \quad Fy - 2k_1x - 2\lambda_1D_rx \tag{2.2}$$

$$\partial h/\partial y = \quad F^tx - 2k_2y - 2\lambda_2D_cy \tag{2.3}$$

이다. 식 (2.2)와 (2.3)에서 $\partial h/\partial x = 0$, $\partial h/\partial y = 0$을 풀어보자. 식 (2.2)로부터

$$x = (\lambda_1D_r + k_1I_p)^{-1}Fy/2 \tag{2.4}$$

이므로 이것을 식 (2.3)에 넣으면

$$F^t(\lambda_1 D_r + k_1 I_p)^{-1} F y = 4(\lambda_2 D_c + k_2 I_q) y$$

가 된다. 따라서 $k_1 \to \lambda_1 k_1^{'}$, $k_2 \to \lambda_2 k_2^{'}$, $4\lambda_1\lambda_2 \to \lambda$ 로 대치하면

$$F^t(D_r + k_1^{'} I_p)^{-1} F y = \lambda(D_c + k_2^{'} I_q) y \tag{2.5}$$

가 된다. 식 (2.5)의 양변 앞에 $(D_c + k_2^{'} I_q)^{-1/2}$을 곱하면

$$G(D_c + k_2^{'} I_q)^{1/2} y = \lambda(D_c + k_2^{'} I_q)^{1/2} y$$

의 형태로 표현되는데 여기서 G는 다음과 같이 정의되는 $q \times q$ 행렬이다.

$$G = (D_c + k_2^{'} I_q)^{-1/2} F^t (D_r + k_1^{'} I_p)^{-1} F (D_c + k_2^{'} I_q)^{-1/2}.$$

따라서 λ는 대칭행렬 G의 고유값이고 y는 G의 고유벡터 v에 의해

$$y \propto (D_c + k_2^{'} I_q)^{-1/2} v \tag{2.6}$$

로 표현된다. 행 수량화 값 벡터 x는 식 (2.4)에 의해

$$x \propto (D_r + k_1^{'} I_p)^{-1} F y \tag{2.7}$$

이다. 식 (2.6)과 (2.7)에서 비례 상수는 y와 x에 대한 정규화 조건 $y^t D_c y = 1$, $x^t D_r x = 1$이 충족되도록 정한다.

$k_1^{'} = k_2^{'} = 0$이면 식 (2.6)과 (2.7)에 의한 축소 해는 통상적인 수량화 제3 방법의 표준 해가 된다. 표준 해는 다음과 같은 성질들을 갖는 것으로 알려져 있다.

- x_1과 x_2는 r를 가중치로 직교한다: $x_1^t D_r x_2 = 0$.
- y_1과 y_2는 c를 가중치로 직교한다: $y_1^t D_c y_2 = 0$.
- x_1과 y_2는 F를 가중치로 직교한다: $x_1^t F y_2 = 0$.

그러나 축소 해는 표준 해의 섭동(攝動, perturbation)이므로 더 이상 앞의 성질들을 갖지 않는다. 다만, 0에 가까운 $k_1^{'}$과 $k_2^{'}$이 고려될 것이므로 축소 해에서도 앞의 성질들이 근사적으로 유효할 것이다. 다음 절에서 수치 예를 통하여 확인해보기로 한다.

축소 해의 방법론에서 $k_1^{'}$과 $k_2^{'}$을 어떻게 주느냐가 관건이다. 식 (2.6)을 보면 $q \times q$ 대각행렬 D_c의 모든 대각 요소에 $k_2^{'} \geq 0$을 더하여 D_c의 상태지수(condition index)가 1에 가까운 방향으로 다소 조정되는 것을 볼 수 있다. 즉, 대각행렬 $D_c + k_2^{'} I_q$의 상태지수 $\kappa(D_c + k_2^{'} I_q)$는

$$\kappa(D_c + k_2^{'} I_q) = (c_{max} + k_2^{'})/(c_{min} + k_2^{'}) \leq c_{max}/c_{min} = \kappa(D_c)$$

의 관계에 있는데 여기서 c_{max} 와 c_{min}은 열 범주 비율 c_j $(j = 1, \cdots, q)$의 최대값 및 최소값을 표기한 것이다. 따라서 열 범주 비율간 최대 비 c_{max}/c_{min} $(= \kappa(D_c))$가 열 범주 비율간 불균형도에 대해 허용가능한 상태지수 값 M_2보다 크면 비음의 상수 $k_2^{'}$을 택하여

$$\kappa(D_c + k_2^{'} I_q) = (c_{max} + k_2^{'})/(c_{min} + k_2^{'}) \leq M_2 \tag{2.8}$$

가 되게 할 필요가 있다. 식 (2.8)을 만족하는 k_2'은

$$k_2' \geq (c_{max} - M_2 c_{min})/(M_2 - 1) \tag{2.9}$$

이다. 결국 M_2의 선택에 따라 k_2'이 결정되는데 우리는 다음 절의 수치 예에서 M_2를 10으로 놓을 것이다. 수리적 논리의 대칭성에 의하여 행 대각행렬 D_r에 부여하는 비음의 상수 k_1'도 다음 식으로 정할 수 있다.

$$k_1' \geq (r_{max} - M_1 r_{min})/(M_1 - 1),$$

여기서 r_{max}와 r_{min}은 행 범주 비율 $r_i (i = 1, \cdots, p)$의 최대값 및 최소값이고 M_1은 행 범주 비율간 불균형도에 대해 허용가능한 상태지수를 나타낸다.

3. 사례 분석: 다중응답 자료의 시각화

K 리서치의 2001년 "한국인의 의식과 가치관 조사" 설문지는 2개의 다중응답(multiple response) 질문을 포함하고 있는데 다음은 그 중 한 예이다.

문 1. ○○님께서 보시기에 현재 우리 사회가 시급히 해결해야 할 문제는 무엇입니까?
다음 보기에서 두 가지만 말씀해 주십시오.

보기:

1.	빈부격차해소	2.	정치적 안정	3.	물가안정/경기활성화
4.	사회복지 향상	5.	도덕성 회복	6.	전인교육 실현
7.	민생치안 확립	8.	환경보호	9.	교통난 해결
10.	성매매 근절	11.	기초질서 확립	12.	부실공사 방지
13.	기초질서 확립	14.	지역감정 해소	15.	집단 이기주의
16.	도농간 격차 해소				

표 3.1은 268명이 응답한 자료(출처: http://www.spss.co.kr)를 남녀 4개 나이대로 분류하여 정리한 표이다. 총 빈도가 536인 이유는 각 응답자가 2개 항목을 선택하였기 때문이다. 이제 다중응답 자료에 수량화 제3 방법을 적용하여 행 범주와 열 범주 사이의 관계를 살펴보기로 한다.

표 3.1에 수량화 제3 방법 ($k_1' = k_2' = 0$)을 적용한 결과, F를 가중치로 한 행 수량화 값 벡터 x와 열 수량화 값 벡터 y 사이의 내적 값은 크기 순서에 따라 0.241, 0.188, 0.173, 0.156, 0.114, 0.095, 0.074 등으로 나타났다 (제곱합의 누적 점유율 = 33.2%, 53.3%, 70.4%, 84.2%, 91.7%, 96.9%, 100.0%). 이 경우 내적 값은 행 점수와 열 수량화 값간의 상관계수를 의미하기도 한다. 그림 3.1은 행 수량화 점수와 열 수량화 점수의 2개 차원 성분을 플롯한 그림이다. 열 플롯에서 V14(지역감정)와 V12(부실공사), V15(집단 이기주의), V13(기초질서 확립)이 눈에 띄게 돌출되어 나타난다. 때문에 이들 두 항목의 위치가 분석자의 자료 해석에 큰 영향을 주게 된다. 그러나 이들 범주들에 반응한 빈도는 모두 매우 작다 (구성 비율이 0.007~0.013 사이). 따라서 자료 해석은 과잉되기 쉽고 불안정성을 초래한다.

이 자료에서 행 범주비율간 최대 비는 $r_{max}/r_{min} = 1.76$으로 큰 차이가 없지만 열 범주 비율간 최대 비는 $c_{max}/c_{min} = 36.0$으로 상당히 크다 ($c_{max} = 144/536 = 0.269, c_{min} =$

표 3.1: K 리서치의 다중 응답 반응 표: 우리 사회가 시급히 해결해야 할 문제는?

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	합
M2	14	13	27	5	6	0	2	4	6	5	4	0	1	1	0	88
M3	9	13	19	3	9	1	3	3	5	5	7	0	1	0	2	80
M4	1	10	12	4	7	1	1	2	3	2	3	1	2	1	0	50
M5	3	9	14	2	9	1	1	1	1	6	5	1	0	1	0	54
F2	5	5	9	3	8	3	1	3	6	5	4	3	0	1	0	56
F3	9	8	21	3	3	3	0	3	4	4	6	0	0	0	0	64
F4	6	6	13	1	10	2	1	4	2	6	4	0	0	0	1	56
F5	14	9	29	3	7	2	5	5	2	1	7	2	1	0	1	88
합	61	73	144	24	59	13	14	25	29	34	40	7	5	4	4	536

* V1부터 V15까지는 순서대로 제시된 응답항목이고 마지막 응답항목인 도농간 격차는 반응이 없으므로 제외함. M2는 남성 20대, M3는 남성 30대, M4는 남성 40대, M5는 남성 50대 이상, F2는 여성 20대, F3는 여성 30대, F4는 여성 40대, F5는 여성 50대 이상을 나타냄.

$4/536 = 0.00746$). 행 범주 비율간 불균형도와 열 범주 비율간 불균형도에 대하여 허용가능한 상태지수 M_1과 M_2를 10으로 놓으면 $k_1^{'} = 0$, $k_2^{'} = 0.022$가 된다.

이에 따라 $k_2^{'}$을 0.022로 놓고 수량화 제3 방법의 축소 해를 구해보았다. 그림 3.2가 그 결과로 얻은 행 플롯과 열 플롯이다. 열 플롯에서 V14와 V12, V15가 중앙 부근으로 이동하여 시각적 지배력이 약화되어 있다. V13은 여전히 도드라져 보이지만 이전에 비해 원점 방향으로 많이 움직여져 있다. 이들을 대신하여 좌상단의 V1(빈부격차), 하단의 V2(정치적 안정)와 V4(사회복지 향상), 우단의 V5(도덕성 회복), 우상단의 V10(성매매 근절), 상단의 V6(전인교육) 등이 눈에 들어온다. 이들 항목에 반응한 응답자는 각각 61명, 73명, 24명, 59명, 34명, 13명이다. F2 그룹(여성 20대)과 F4 그룹(여성 40대)은 우상 4분면에 위치함으로써 V10(성매매 근절)과 V6(전인교육)에 상대적으로 많은 반응을 하였고, 반면 우하 4분면에 위치한 M4 그룹(남성 40대)은 V2(정치적 안정)와 V4(사회복지 향상)에 상대적으로 많은 반응을 하였음을 알 수 있다. 그림 3.2의 행 플롯에서 제 2축이 성 차이(gender difference)임이 드러난다 (여성이 위, 남성은 아래). F를 가중치로 한 수량화 값 벡터 x와 열 수량화 값 벡터 y사이의 내적 값은 크기 순서에 따라 0.230, 0.172, 0.160, 0.150, 0.107, 0.093, 0.073 등으로 산출되었다 (제곱합의 누적 점유율 = 33.8%, 52.7%, 69.2%, 83.7%, 91.0%, 96.6%, 100.0%). 가장 큰 2개 내적 값이 앞의 표준 해에서 0.241, 0.188이었으므로 각각 0.011, 0.016이 작아졌을 뿐이다. 제2 축까지의 제곱합의 누적 점유율, 즉 2차원 수량화 공간의 설명력은 표준해가 53.3%, 축소 해가 52.7%로 차이는 불과 0.6%P 이다.

앞서 언급한대로 수량화 제3 방법의 축소 해에서는 차원이 다른 행 수량화 값 벡터 x_1과 x_2가 r을 가중치로, 열 수량화 값 벡터 y_1와 y_2가 c를 가중치로 직교하지 않을 수 있는데 그 정도를 표 3.2에서 살펴보았다. 행 수량화 값 벡터들은 모두 직교하고 있고 열 수량화 벡터들 사이에서도 주요 관심이 되는 처음 몇 개의 성분들 사이에 미미한 상관만이 있는 것을

볼 수 있다 ($k_1' = 0$인 경우, 행 수량화 값 벡터 x_1과 x_2가 r을 가중치로 직교하고 서로 다른 축에 속하는 행 수량화 값 벡터 x_1과 열 수량화 값 벡터 y_2는 F를 가중치로 직교한다).

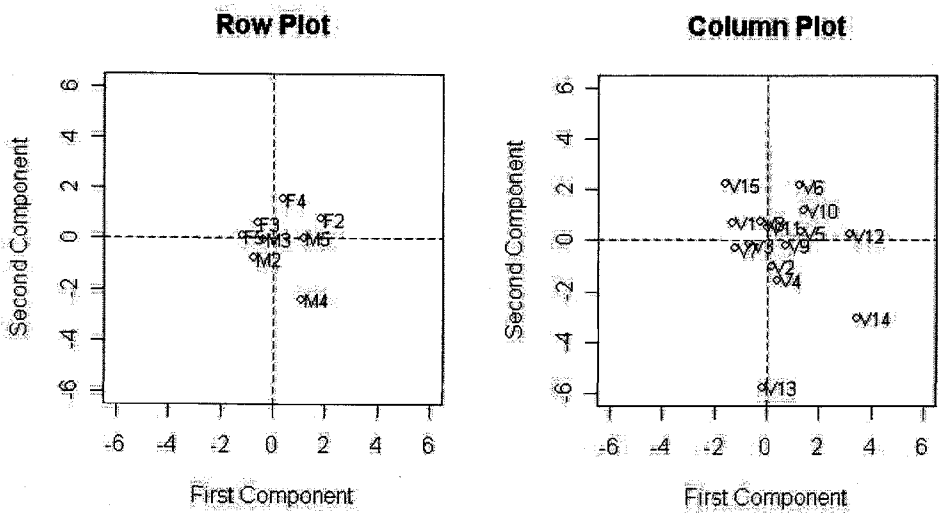

그림 3.1: 수량화 제3 방법의 표준 해 ($k_1' = k_2' = 0$): 행 플롯과 열 플롯

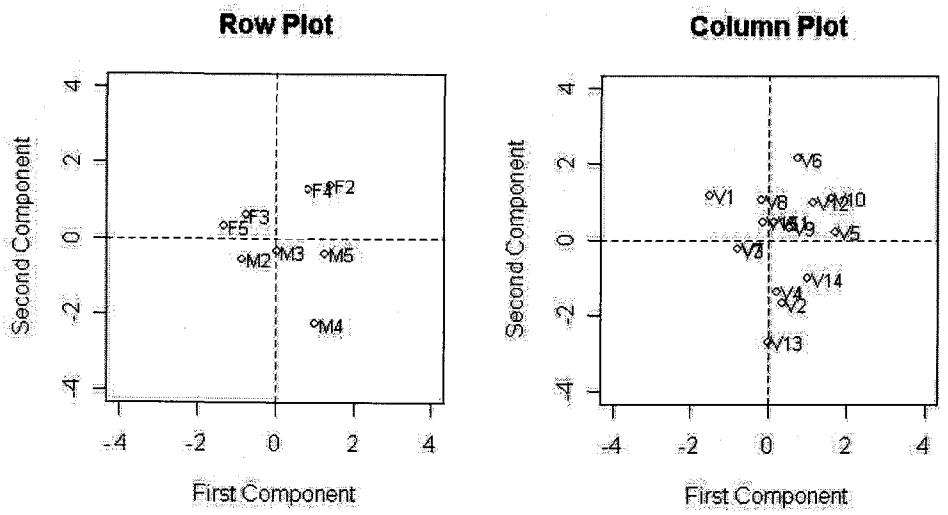

그림 3.2: 수량화 제3 방법의 축소 해 ($k_1' = 0, k_2' = 0.022$): 행 플롯과 열 플롯

표 3.2: 수량화 제3 방법의 축소 해 ($k_1' = 0, k_2' = 0.022$)에서
(a) 행 수량화 값 벡터간 내적 값과 (b) 열 수량화 값 벡터간 내적 값

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1]	1	0	0	0	0	0	0
[2]	0	1	0	0	0	0	0
[3]	0	0	1	0	0	0	0
[4]	0	0	0	1	0	0	0
[5]	0	0	0	0	1	0	0
[6]	0	0	0	0	0	1	0
[7]	0	0	0	0	0	0	1

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1]	1.00	-0.03	0.10	0.01	0.10	-0.07	0.02
[2]	-0.03	1.00	-0.01	-0.02	0.09	0.08	0.07
[3]	0.10	-0.01	1.00	-0.08	-0.16	0.14	0.02
[4]	0.01	-0.02	-0.08	1.00	0.00	0.03	-0.10
[5]	0.10	0.09	-0.16	0.00	1.00	0.15	0.00
[6]	-0.07	0.08	0.14	0.03	0.15	1.00	0.13
[7]	0.02	0.07	0.02	-0.10	0.00	0.13	1.00

4. 맺음 말

수량화 제3 방법은 대응분석(correspondence)과 외형적 정식화가 다소 다를 뿐 수리적 기초를 공유하는 것으로 알려져 있다 (허명회, 1999, pp. 35-43). 대응분석에서도 작은 빈도의 행 범주 또는 열 범주가 2-3 차원의 위치도(positioning map)에서 시각적으로 돌출되어 나타나는 경우가 드물지 않게 발생한다. 이러한 경우 안정적인 자료해석을 위해서 본 연구에서와 유사한 방식으로 축소 해를 구할 수 있을 것으로 생각한다. 또는 Choi와 Huh(2000)가 제안한 로버스트 대응분석(robust correspondence analysis) 알고리즘을 적용해볼 수 있을 것이다. 그들은 대응분석에서 찾고자 하는 저차원 주성분공간을 M-추정의 방법으로 풀어내었다. 대응분석의 축소 해와 로버스트 해에 대하여는 앞으로 추가적 연구가 필요하다.

참고문헌

허명회 (1998). 「수량화 방법 Ⅰ, Ⅱ, Ⅲ, Ⅳ」 서울: 자유아카데미.

허명회 (1999). 「다변량 수량화」 서울: 자유아카데미.

岩坪秀一 (1987). 「數量化法の 基礎」 東京: 朝倉書店.

駒澤 勉 (1992). 「數量化理論」 東京: 放送大學教育振興會.

Casella, G., and Berger, R. L. (1990). *Statistical Inference.* Duxbury, CA: Belmont.

Choi, Y. S., and Huh, M. H. (1999). "Robust simple correspondence analysis". *Journal of the Korean Statistical Society.* **28**, 337-346.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

[2005년 8월 접수, 2005년 10월 채택]

Shrinkage Solution of Quantification Method Ⅲ

Myung-Hoe Huh[1] Yonggoo Lee[2]

ABSTRACT

Quantification method Ⅲ is designed by C. Hayashi as visualizing technique for two-way cross-classified tables. Specially in Japan, its usefulness is timely proven in social and marketing surveys. In several instances, relatively large quantification scores are assigned to low-frequency categories. Thus, they lead to unreliable data interpretation. The aim of this study is to develop stable solution to overcome such traits of quantification method Ⅲ. The solution is of shrinkage type induced by small perturbations and is applied to a multiple response data obtained in a Korean social survey.

Keywords: Hayashi's Quantification method Ⅲ, Shrinkage solution, Condition index, Quantification plot, Correspondence analysis.

1) Professor, Dept. of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul, 136-701 Korea.
E-mail: stat420@korea.ac.kr

2) Professor, Dept. of Statistics, Chung-Ang University, Heuksuk Dong 221, Dongjak-Gu, Seoul, 156-756 Korea.
E-mail: leeyg@cau.ac.kr