The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.201
26(1), 201–208

A Note on Model Selection in Mixture Experiments with Process Variables

Jung Il Kim[a,1]

[a]Department of Statistics, Kangwon National University

(Received January 11, 2013; Revised January 29, 2013; Accepted February 20, 2013)

Abstract

In this paper, we consider the mixture components-process variables model and propose a model selection strategy using MTS. This strategy is illustrated using an example that involves three mixture components and two process variables in a bread making experiment that was studied in several literatures.

Keywords: Mixture experiments, process variables, Mahalanobis Taguchi System.

1. 서론

화학, 식품, 약품 등 여러 분야에서 활용되는 혼합물 실험(mixture experiments)은 반응변수(y)가 혼합물의 성분(components, ingredients)인 설명변수들의 절대량이 아닌 상대적인 혼합비율에 의해 영향을 받는 성질이 있으며, Prescott (2004)에서 논의된 것처럼 공정 조건 등과 같이 혼합물의 성분과 관련이 없는 인자인 공정변수(process variables)에 의해서도 영향을 받을 수 있다.

혼합물 성분비(x_i)와 공정변수(z_k)들을 결합한 모형(combined model)으로는 Cornell (2002)에서 소개된 바와 같은 혼합물 성분 모형과 공정변수 모형의 교적모형(product model, crossed model)을 생각할 수 있다. 혼합물 성분의 수 q와 공정변수의 수 p가 큰 경우 교적모형에 포함되는 항들의 개수가 많아지므로 Prescott (2004)은 의미 있는 해석이 가능한 교호작용항, 즉 $x_i z_k^2$, $x_i z_k z_l$, $x_i x_j z_k$와 같은 꼴의 교호작용효과를 나타내는 항까지만 포함하는 부분모형을 사용할 것을 제안하였고 Lim (2012)에서는 공정변수의 곡선효과와 혼합물 성분과 공정변수간의 3차 교호작용효과까지를 반영하는 교적모형의 일부 항들로 구성된 부분모형을 실용적인 모형으로 제시하였다.

혼합물 실험에서 모형선택에 관하여 Prescott (2004)은 혼합물 성분의 공선성(collinearity)때문에 통상적인 단계별 회귀법을 사용하기 어려움을 지적하고 고차 교호작용을 배제한 실용적인 모형을 제시하여 개별 모수에 대한 추론이 아닌 모수 묶음에 대한 추론을 사용하는 전략적 접근법을 제안하였다. Lim (2012)은 혼합물 성분과 공정변수들에 대한 자료가 주어질 때 시작모형의 후보들을 교적모형의 범주와 Lim (2011)에서 제시된 실용적인 모형들 중에서 찾아 완전모형으로 간주하고 모형의 간결성 원칙에 따라 구성되는 부분모형들 중에서 축차적인 변수 선택법이나 모든 가능한 회귀법에 의한 변수선택법으로 모형 성능 통계량의 값을 비교하고 잔차 그림을 검토하여 최종적으로 적절한 모형을 추천하는 전략적인 방법을 제안하였다.

[1]Professor, Department of Statistics, Kangwon National University, 192-1 Hyoja-Dong, Chunchon, Kangwondo 200-701, Korea. E-mail: jikim@kangwon.ac.kr

이 논문에서는 Prescott (2004)과 Lim (2012)에서 제시된 교적모형을 사용하여 적절한 모형을 선택하는 실용적인 전략으로 Taguchi와 Rajesh (2000), Taguchi와 Jugulum (2002)에 소개된 마할라노비스-다구찌 시스템(Mahalanovis-Taguchi System; MTS) 절차를 활용하는 방법을 2절에서 소개한다. 3절에서는 실제적인 예로 Nas 등 (1998)에 소개된 제빵 자료에 적용하고 모형선택 결과를 논의한다.

2. 혼합물 성분과 공정변수의 모형

2.1. 혼합물 성분과 공정변수에 대한 모형

혼합물의 i번째 성분비율을 x_i로 나타내면 q개의 성분비율은 다음과 같은 제약조건을 갖는다.

$$x_1 + x_2 + \cdots + x_q = 1, \quad x_i \geq 0,\ i = 1, 2, \ldots, q. \tag{2.1}$$

따라서 이러한 혼합물 실험의 실험공간은 $(q-1)$차원 심플렉스(simplex) 공간이다. Cornell (2002)에서 제시한 바와 같이 각 혼합물 성분비율에 제약이 주어지면 실험공간은 심플렉스의 일부가 된다. 이러한 혼합물 실험공간에서 심플렉스 중심배열법 등과 같은 실험설계에 의하여 실험한 자료를 적절한 모형을 적합하여 분석하게 된다.

혼합물 성분에 대해서 흔히 사용되는 Scheffe의 정준 2차 다항모형은 제약조건 식 (2.1)을 고려하여 상수항과 제곱항을 포함하지 않는 다음과 같은 식으로 주어진다.

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} x_i x_j. \tag{2.2}$$

그러나 Nas 등 (1998)에서는 3개 성분간 교호작용항인 $x_i x_j x_k$까지 포함하는 다음과 같은 3차 다항모형을 사용하였다.

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{i=1}^{q-2} \sum_{i<j}^{q-1} \sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{i-j} x_i x_j (x_i - x_j). \tag{2.3}$$

그 외에도 Draper와 John (1977), Becker (1968) 등이 제안한 여러 다른 혼합물 성분 모형들이 있으나 일반적으로 많이 사용되는 Scheffe의 다항모형을 적합시키기로 한다.

한편 혼합물 성분이 q개인 혼합물 실험을 p개의 공정변수를 갖는 공정조건에서 수행하는 경우 공정변수의 실험공간은 p차원 하이퍼큐브(hypercube)가 된다. 공정변수에 대해서는 다음과 같은 2차 다항모형을 적합시킬 수 있다.

$$E(Y) = \alpha_0 + \sum_{k=1}^{p} \alpha_k z_k + \sum_{k=1}^{p} \alpha_{kk} {z_k}^2 + \sum_{k=1}^{p-1} \sum_{k<l}^{p} \alpha_{kl} z_k z_l. \tag{2.4}$$

공정변수에 대한 실험설계로는 모형의 모수에 대한 추론을 하기 위한 요인실험이나 반응표면실험 등이 사용될 수 있다.

2.2. 혼합물 성분-공정변수에 대한 결합모형

혼합물 성분 q개와 공정변수 p개를 갖는 혼합물 성분-공정변수 실험에 대한 결합모형으로는 식 (2.2) 또는 식 (2.3)과 같은 혼합물 성분 모형과 식 (2.4)와 같은 공정변수 모형의 교적모형을 생각할 수 있다.

이 교적모형의 모든 항들을 포함하는 완전모형을 시작모형으로 사용하여 적절한 모형을 선택할 수도 있겠지만 혼합물 성분과 공정변수간의 관계를 해석하기에 용이한 항들만 포함하는 모형을 시작모형으로 사용하여 적절한 모형을 선택하는 것이 실용적이다. Prescott (2004)이나 Lim (2012)에서와 같이 의미있는 해석이 용이한 교호작용항들, 즉 $x_iz_k^2$, $x_iz_kz_l$, $x_ix_jz_k$와 같은 꼴의 교호작용효과를 나타내는 항까지만 포함하는 다음과 같은 결합모형을 실용적인 시작모형으로 사용한다.

$$E(Y) = \sum_{i=1}^{q}\beta_i x_i + \sum_{i=1}^{q-1}\sum_{i<j}^{q}\beta_{ij}x_ix_j + \sum_{i=1}^{q-2}\sum_{i<j}^{q-1}\sum_{i<j<k}^{q}\beta_{ijk}x_ix_jx_k + \sum_{i=1}^{q-1}\sum_{i<j}^{q}\beta_{i-j}x_ix_j(x_i - x_j)$$
$$+ \sum_{i=1}^{q}\sum_{k=1}^{p}\gamma_{ik}x_iz_k + \sum_{i=1}^{q}\sum_{k=1}^{p}\gamma_{ikk}x_iz_k^2 + \sum_{i=1}^{q}\sum_{k=1}^{p-1}\sum_{k<l}^{p}\gamma_{ikl}x_iz_kz_l + \sum_{i=1}^{q-1}\sum_{i<j}^{q}\sum_{k=1}^{p}\gamma_{ijk}x_ix_jz_k. \quad (2.5)$$

2.3. 혼합물 성분-공정변수에 대한 모형 선택

여러 결합모형들 중 적절한 모형을 선택하는 방법은 변수선택법에 따라 다양하지만, 선택하려는 적절한 결합모형은 시작모형 (2.5)의 부분모형으로 항들의 위계 원칙(hierachy principle)이 지켜지는 모형이다. 즉 3차항 $x_ix_jz_k$가 유의하여 모형에 포함될 경우 하위 2차항인 x_ix_j, x_iz_k, x_jz_k들은 유의성에 관계 없이 모형에 포함된다.

이제 시작모형 (2.5)를 완전모형으로 적절한 부분모형을 선택하는 실용적인 전략으로 마할라노비스-다구찌 시스템(MTS) 절차를 활용하는 방법을 소개한다. MTS 절차는 Taguchi와 Rajesh (2000), Taguchi와 Jugulum (2002)이 제안한 다변량 자료를 사용하여 진단 및 예측을 하는 방법으로 다음과 같은 네 단계를 거쳐 적용한다.

단계 1: 기준(reference)이 되는 단위공간인 마할라노비스 공간을 갖는 측정척도 구성

각 관측치를 자료분석 목적에 따라 기준이 되는 그룹(대조군, 정상군)과 그렇지 않은 그룹(처리군, 비정상군)으로 분류하는데 사용할 변수들을 정한다. 정의된 변수들에 대한 대조군에서의 자료를 표준화하고 표준화된 자료에 대해 아래와 같이 정의되는 평균이 1이 되도록 단위화한 마하라노비스 거리(Mahalanobis distance; MD)를 계산하여 기준이 되는 마할라노비스 공간을 정의한다.

$$\mathrm{MD}_j = D_j^2 = \frac{Z_j^T A^{-1} Z_j}{k}, \quad (2.6)$$

여기서 k는 변수의 개수, j는 관측치 번호, Z는 표준화된 변수벡터, A는 변수들에 대한 상관계수행렬, A^{-1}는 A의 역행렬을 나타내며 공선성이 있는 경우 일반화 역행렬을 사용하기로 한다.

혼합물 성분-공정변수에 대한 모형 선택을 위해서 반응값의 속성에 따라 대조군과 처리군을 정의하고 결합모형 (2.5)를 구성하는 항들을 관측치 분류에 사용할 변수들로 정의한다.

단계 2: 측정척도의 정확성 검사

대조군의 변수들에 대한 평균과 표준편차를 사용하여 처리군의 변수들을 표준화하고 마할라노비스 거리(MD)를 계산하여 상대적으로 큰 값들이 나오면 측정척도의 정확성을 인정한다. 즉 단계 1에서 정의한 대조군과 처리군의 정당성을 확보할 수 있고 변수들을 분류에 사용할 수 있게 된다.

단계 3: 적절한 변수선택

직교배열표(Orthogonal Array; OA)와 신호잡음비(S/N ratio)를 사용하여 적절한 변수들을 찾는다. 여기서 신호잡음비는 처리군의 MD값들을 사용하여 계산한 값으로 직교배열표의 각 조합에 대한 반응

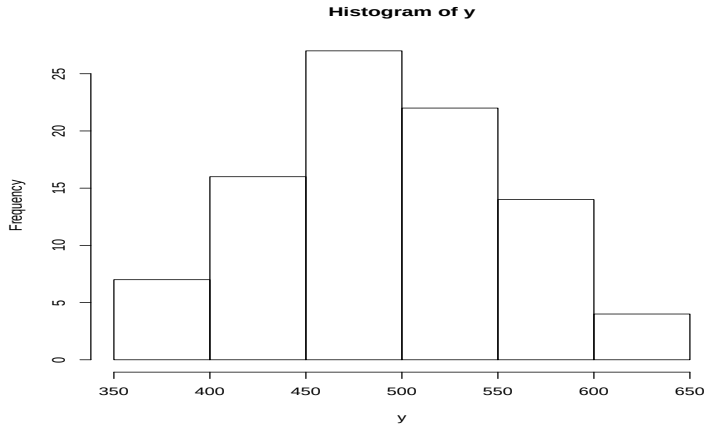

Figure 3.1. Histogram of loaf volume(y)

값으로 이용된다. 적절한 변수는 신호잡음비의 이득(gain)을 평가하여 선택한다. 이 때 항들의 위계 원칙을 지키도록 한다.

단계 4: 적절한 변수에 대한 진단

적절하다고 선택된 변수들만 사용하여 MD를 계산하여 각 관측치들에 대한 예측력을 확인하고 자료분석 목적에 따른 의사결정을 한다. 즉 선택된 변수들만에 의한 부분모형의 실용성을 확인한다.

3. 실용적 적용사례

서로 다른 세가지 밀가루 Tjalve, Folke, Hard Red Spring를 혼합한 밀가루 반죽에 의해 만들어진 빵의 품질에 대한 실험이 Nas 등 (1998)에 소개되어 있다. 제빵 공정은 제과점마다 다를 것이므로 밀가루 반죽이 제빵 공정의 조건 변동에 로버스트할 것이 요구되며, 서로 다른 공정 조건으로 혼합시간(5, 15, 25분)과 교정시간(35, 47.5, 60분)이 고려되었다. 반응변수(y)는 가공된 빵의 부피이고, 밀가루 반죽은 다음과 같은 제약을 갖는 혼합물 성분으로 생각한다.

$$0.25 \le x_1 \le 1, \quad 0 \le x_2 \le 0.75, \quad 0 \le x_3 \le 0.75, \quad x_1 + x_2 + x_3 = 1, \tag{3.1}$$

여기서 x_1, x_2, x_3는 세 종류 밀가루(Tjalve, Folke, Hard Red Spring) 각각의 혼합비율을 나타낸다. 공정변수인 혼합시간(mixing time)과 교정시간(proofing time)에 대해서는 $z_1 = (\text{mixing time} - 15)/10$, $z_2 = (\text{proofing time} - 47.5)/12.5$와 같이 변환된 값을 사용한다.

공정변수들에 대해서 3^2 요인실험이 사용되고 이 9개 공정조건 각각에 대하여 심플렉스 격자법으로 10회 실험이 반복되어 총 90회 실험의 자료가 구해졌다. 이 실험의 자세한 결과 자료 및 추가적인 사항은 Nas 등 (1998)을 참고하기로 한다. 이 실험에 적절한 모형을 선택하기 위하여 2절에서 소개한 절차를 적용한다.

단계 1: 자료분석의 목적이 가공된 빵의 부피(y)를 최대화하는 조건을 찾는데 있다고 하자. 분류에 사용할 변수들은 결합모형 (2.5)에 포함되어 있는 항들로 31개가 정의된다. Figure 3.1에 나타낸 y에 대한 히스토그램을 참고하여 처리군의 조건은 특성값이 큰 $y > 600$으로, 대조군의 조건은 최빈구간을 포함하면서 변수의 개수인 31개보다 도수가 49개로 상대적으로 더 많게 되는 $450 < y < 550$으로 정하기

Table 3.1. MD values for the treatment group

관측치번호	80	81	89	90
y	649.44	640.00	630.89	638.89
MD	6.46	25.72	17.51	70.58

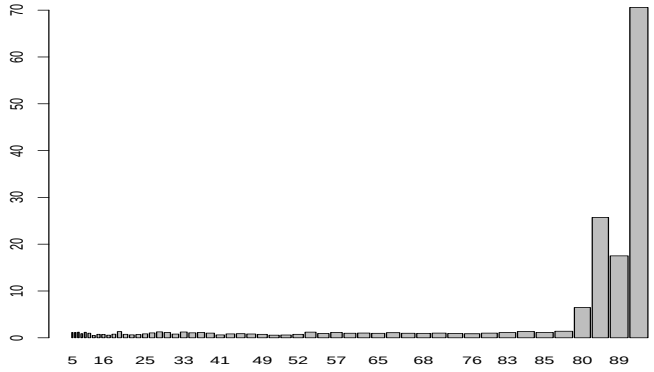

Figure 3.2. MD values for the treatment and control group

로 하였다. 처리군에 속하는 4개 관측치들의 번호는 $(80, 81, 89, 90)$인 것으로 판명되었다. 대조군에서의 변수들 값을 표준화하여 단위화된 MD를 계산하고 이를 사용하여 기준이 되는 마할라노비스 공간을 정의한다.

단계 2: 대조군의 변수들에 대한 평균과 표준편차를 사용하여 처리군의 변수들을 표준화하고 마할라노비스 거리(MD)를 계산한 결과가 Table 3.1과 Figure 3.2에 주어졌으며 처리군에 속한 관측값(번호 $80, 81, 89, 90$)들에 대한 MD가 상대적으로 큰 값들이 나왔기에 측정척도의 정확성을 인정할 수 있다. 즉 단계 1에서 정의한 대조군과 처리군의 정당성을 확보할 수 있고 변수들을 분류에 사용할 수 있게 되었다.

단계 3: 앞 단계에서 정의한 31개 변수들 중에서 적절한 선택을 하기 위하여 필요한 최소 크기인 L_{32} 직교배열표를 사용하기로 한다. 직교배열표에 의한 실험 각각에 대하여 처리군에 속한 관측치들에 대해 직교배열표의 '1'에 해당하는 변수들만 사용하여 구한 MD를 이용하여 망대특성 신호잡음비(S/N)를 계산한다. 해당 직교배열표와 신호잡음비에 대해서는 Taguchi 등 (2005)을 참고하여 계산하였고, 계산된 신호잡음비를 직교배열표의 각 실험에 대한 반응값으로 사용하여 구한 각 변수들에 대한 신호잡음비의 이득(gain)이 Figure 3.3에 나타나 있다. Figure 3.3에서 신호잡음비의 이득이 양수인 변수는 아래와 같은 20개 항목인 것으로 확인되었다.

$$x_1,\ x_3,\ x_1x_2,\ x_1x_3,\ x_1x_2(x_1-x_2),\ x_1x_3(x_1-x_3),\ x_1z_1,\ x_3z_1,\ x_2z_2,\ x_3z_2,\ x_1z_1^2,\ x_3z_1^2,$$
$$x_2z_1z_2,\ x_3z_1z_2,\ x_1x_2z_1,\ x_1x_3z_1,\ x_2x_3z_1,\ x_1x_2z_2,\ x_1x_3z_2,\ x_2x_3z_2$$

항들의 위계 원칙에 따라 $x_2, x_2x_3, x_2z_1, x_1z_2$를 모형에 추가해야 하며 따라서 선택할 변수는 아래에 주어진 24개 항목이 된다.

$$x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3,\ x_1x_2(x_1-x_2),\ x_1x_3(x_1-x_3),\ x_1z_1,\ x_2z_1,\ x_3z_1,\ x_1z_2,\ x_2z_2,$$
$$x_3z_2,\ x_1z_1^2,\ x_3z_1^2,\ x_2z_1z_2,\ x_3z_1z_2,\ x_1x_2z_1,\ x_1x_3z_1,\ x_2x_3z_1,\ x_1x_2z_2,\ x_1x_3z_2,\ x_2x_3z_2$$

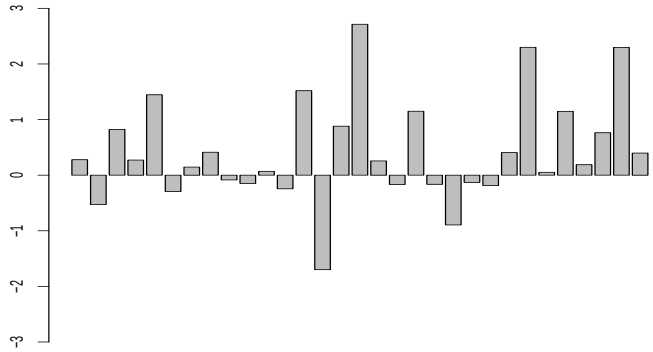

Figure 3.3. Gains of SN ratio for variables

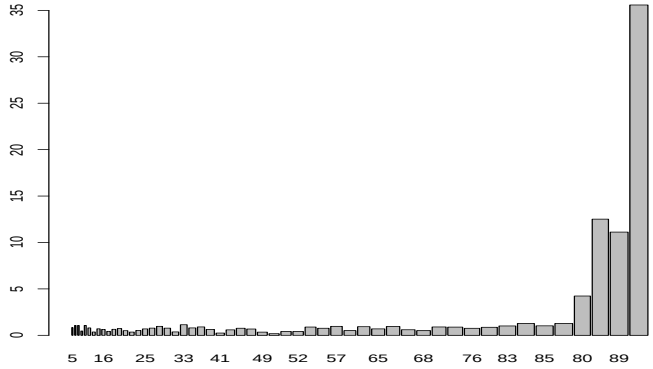

Figure 3.4. MD values for the treatment and control group

단계 4: 앞의 단계 3에서 적절하다고 선택한 24개 변수들만 사용하여 계산한 MD 값들을 Figure 3.4에 나타내었다. 처리군(번호 80, 81, 89, 90)의 최소 MD 값(4.233)이 대조군의 최대 MD 값(1.276)보다 커서 분류가 잘 되어 있음을 볼 수 있어 각 관측치에 대한 예측력이 있음을 확인할 수 있고 선택된 변수들만에 의한 부분모형이 실용성이 있다고 판단한다. 또한, 선택된 부분모형의 결정계수는 0.999, 수정결정계수는 0.998로 계산되어 Nas 등 (1998)에 제시된 28개 항목을 사용한 결합모형의 결정계수 0.934보다 크게 나와 선택된 모형이 적합함을 알 수 있다.

4. 논의 및 결론

혼합물 성분비와 공정변수들을 결합한 모형으로 의미 있는 해석이 가능한 교호작용항만 포함하는 교적모형을 시작모형으로 하여 간결성 원칙에 따른 실용적인 부분모형을 찾는 전략적 접근법의 하나로 MTS 절차를 활용하는 방법을 제시하였다. 제시한 방법은 MTS 절차에서 사용되는 마하라노비스 거리(MD)를 공선성이 있는 경우에도 일반화 역행렬을 이용하여 안정적인 계산이 가능하므로 공선성 여부에 로버스트한 성질을 갖고 있어 혼합물 성분-공정변수에 대한 모형 선택에 활용하기에 적절함을 알 수 있다. 따라서, 혼합물 성분의 공선성때문에 추천되는 통계적 추론에 근거하는 전략적 접근법들에 더하여 자료분석적인 방법으로 보완적인 정보를 줄 수 있어 직관적이고 실용적이라 할 수 있다. 그러나 대조

군이 명확히 정의되어 있지 않을 경우 주관성을 배제할 수 없으며, 분류의 기준이나 SN비의 이득에 대한 평가 등에 관하여 엄밀한 논의가 필요하며 추후 연구 대상이다.

적용사례에서 네 단계로 이뤄진 MTS 절차를 활용하여 적절한 모형을 선택하였고 선택된 모형의 적합도가 높은 것도 확인하였다.

References

Becker, N. G. (1968). Models for the response of a mixture, *Journal of the Royal Statistical Society, Series B (Methodological)*, **30**, 349–358.

Cornell, J. A. (2002). *Experiments with Mixtures*, John Wiley & Sons, Inc.

Draper, N. R. and John, St. (1977). A mixtures model with inverse terms, *Technometrics*, **19**, 37–46.

Lim, Y. (2011). Practical designs for mixture component-process experiments, *Journal of Korean Society for Quality Management*, **39**, 400–411.

Lim, Y. (2012). Analysis of mixture experimental data with process variables, *Journal of Korean Society for Quality Management*, **40**, 347–358.

Nas, T., Fargestad, E. M. and Cornell, J. A. (1998). A comparison of methods for analyzing data from a three component mixture experiment in the presence of variation created by two process variables, *Chemometrics and Intelligent Laboratory System*, **41**, 221–235.

Prescott, P. (2004). Modelling in mixture experiments including interactions with process variables, *Quality Technology & Quantitative Management*, **1**, 87–103.

Taguchi, G., Chowdhury, S. and Wu, Y.(2005). *Taguchi's Quality Engineering Handbook*, John Wiley & Sons, Inc.

Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy*, John Wiley & Sons, Inc.

Taguchi, G. and Rajesh, J. (2000). New trends in multivariate diagnosis, *Sankhya: The Indian Journal of Statistics*, **62**, 233–248.

공정변수를 갖는 혼합물 실험에서 모형선택의 한 방법

김정일[a,1]

[a]강원대학교 통계학과

(2013년 1월 11일 접수, 2013년 1월 29일 수정, 2013년 2월 20일 채택)

요 약

이 논문에서는 공정변수를 갖는 혼합물 실험에 대하여 적절한 모형을 선택하는 한 방법으로 혼합물 성분의 공선성에 로버스트한 성질을 갖는 마할라노비스-다구찌 시스템을 활용한 전략을 소개한다. 여러 문헌에서 언급된 3개의 혼합물 성분과 2개의 공정변수를 갖는 제빵 실험 사례를 대상으로 이 전략적 방법을 적용하여 적절한 모형을 선택하였다.

주요용어: 혼합물 실험, 공정변수, 마할라노비스 다구찌 시스템.

[1](200-701) 강원도 춘천시 효자동 192-1, 강원대학교 통계학과, 교수. E-mail: jikim@kangwon.ac.kr