한국통계학회논문집 제2권 1호
1995년 4월 pp. 64-73

순차적 혼합물 실험계획을 평가하기 위한 그래픽방법

장대홍, 박상현[1)]

요 약

혼합물 실험계획의 확장시 기존의 연구들은 D-최적계획을 중심으로 전개되어 왔다. 이러한 판정기준들은 실험계획의 전 영역에 걸친 성능의 정도를 알아내는 데는 한계가 있다. 이러한 단점을 극복하기 위하여 하나의 그래픽 방법을 제안하였다. 혼합물 실험 계획에서 결측값이 발생하는 경우에도 이 그래픽 방법을 이용할 수 있다.

1. 서 론

반응표면분석에서 뿐만이 아니라 혼합물 실험계획에서도 순차적 실험계획은 중요한 연구 주제가 되고 있다. 반응표면분석에서 중심합성계획을 예로 들면, 1차모형을 가정하면 계획의 요인부분(fatorial part)을 이용하여 변수선택(variable screening)이나 영역조사(region seeking) 등의 처리를 하게 되고, 2차모형을 가정하면 계획의 요인부분에다 축부분(axial part)과 중심부분(central part)를 첨가하여 최적화(optimization)와 영역탐색(region exploration) 등의 처리를 하게 된다. 혼합물 실험계획에서의 순차적 실험은 1차모형 가정시 계획의 꼭지점들을 이용하게 되고, 2차모형 가정시 꼭지점 외에 변의 중심점이나 면의 중심점, 전체 중심점 등이 필요하게 된다.

실험계획을 확장할 때나 실험계획점 중 결측값이 발생할 때 실험계획의 성능의 정도를 알아보기 위해서 순차적 실험계획을 행하게 된다. 실험계획의 확장시 기존의 연구들은 주로 D-최적계획을 염두에 두었다. Dykstra(1966)와 Gaylor와 Merrill(1968)은 1차모형에서 다공선성을 줄이기 위한 계획의 확장에 대해 언급하였고, Covey-Cramp와 Silvey(1970), Wynn(1970), Dykstra(1971), Hebble과 Mitchell(1972), Mayer와 Hendrickson(1973)과 Evans(1979)는 실험점들의 추가시 $|X'X|$ 기준을 최대화하는 문제에 대한 연구들을 행하였다. 그러나 이러한 알파벳 최적화 기법들(예로, A-, D-, E-, G-와 V-최적화)은 하나의 수치를 이용하는 분산-최소화 기준들로서, 실험계획의 전 영역에 걸친 성능의 정도를 파악하기에는 부족한 점이 많다. 반응표면분석에서, 이러한 문제점을 지적하고, 실험계획의 전 영역에 걸친 성능을 파악할 수 있는 그래픽 방법을 Giovannitti-Jensen과 Myers(1989)가 제안하였고, Vining과 Myers(1991)는 이 그래픽 방법을 평균제곱 오차로 까지 확장하였다. Jang과 Park(1993)은 추정반응값의 기울기에 이 그래픽 방법을 적용하였다. Vining, Cornell과 Myers (1993)는 혼합물실험에서, 제한된 영역에서의 실험계획에 대하여 전 영역에 걸친 성능을 파악할 수 있는 또 다른 그래픽 방법을 제안하였다. 본 논문에서는 Vining과 Cornell이 제시한 그래픽 방법을 이용하여 실험계획의 확장시나 실험계획점들 중 결측값이 발생하였을 때 실험계획의 전 영역에 걸친 변화를 알아 보고, 실험계획들을 상호비교, 평가할 수 있는 방법을 제시하였다.

1) (608-737) 부산시 남구 대연3동 599-1, 부산수산대학교 응용수학과

2. 혼합물 계획의 순차적 실험과 평가방법

혼합물실험에서는 혼합물을 구성하는 성분의 상대적 비율에 의해 반응값이 결정된다. 혼합물의 구성성분 갯수를 q개라 하고, x_i를 i번째 구성성분의 비율이라 하면

$$0 \leq x_i \leq 1, \qquad i=1,2,\cdots,q$$

이고,

$$\sum_{i=1}^{q} x_i = 1$$

이 된다. 그리하여, 실험영역이 $(q-1)$차 단체(simplex)가 된다. i번째 구성성분의 비율에 제한이 있는

$$0 < L_i \leq x_i \leq U_i < 1, \qquad i=1,2,\cdots,q$$

일 때, 실험영역은 원 단체상의 제한된 영역이 되고, 일반적으로 불규칙한 볼록다면체가 된다. 이런 경우 볼록다면체의 꼭지점, 각 면의 중심점, 각 변의 중심점, 전체 중심점들이 실험계획을 구성할 수 있는 후보점들(candidate points)이 된다.

각 구성성분의 비율에 제한이 있는 경우, 제한된 영역의 중심이 생기는데, 이것을 전체 중심점(overall centroid), 또는 기준혼합물(reference mixture)라 한다. 제한된 영역에서 각 구성성분의 효과를 알아보기 위하여는 Cox 방향이 이용되는데, i번째 구성성분에 대한 Cox 방향이란, 전체 중심점에서 i번째 구성성분에 해당하는 꼭지점에 그은 가상선을 말한다. 전체중심점에서 각 구성성분의 비율을 $\underline{c}=(c_1, c_2, \cdots, c_q)$라 하면 $\sum_{i=1}^{q} c_i = 1$이 된다. i번째 구성성분의 비율을 Cox 방향을 따라 c_i에서 Δ_i만큼 변화시키면, i번째 구성성분의 새로운 비율은

$$x_i = c_i + \Delta_i$$

가 되고 나머지 $(q-1)$개 구성성분의 비율들은

$$x_j = c_j \frac{1-c_i-\Delta_i}{1-c_i}, \qquad j=1,2,\cdots,q, \quad j \neq i$$

이 된다. 그런데 x_j와 x_k의 비를 보면

$$\frac{x_j}{x_k} = \frac{c_j}{c_k}$$

가 되어 전체 중심점에서의 비와 같아짐을 알 수 있다. 그리하여, 제한된 영역에서 각 구성성분의 효과를 알아보기 위하여는 Cox 방향을 이용하게 되는 것이다.

혼합물 실험모형을 행렬을 이용하여 나타내면,

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \ ,$$

이 된다. 여기서, $\underline{y}$는 $n \times 1$ 반응값 벡터이고, X는 $n \times p$ 행렬로서 선택하는 회귀모형의 차수에 따라 달라지는 모형행렬이고, $\underline{\beta}$는 $p \times 1$ 모수벡터이고, $\underline{\varepsilon}$는 오차항을 나타내는 $n \times 1$ 확률오차벡터로서, 일반적으로 확률분포를 $N(0, \sigma^2)$으로 가정한다. 이 경우에 $\underline{\beta}$의 최소제곱추정량 $\underline{b}$의 분산-공분산 행렬은

$$Var(\underline{b}) = (X'X)^{-1}\sigma^2$$

이 되고, σ^2으로 나눈 추정반응값 $\hat{y}(\underline{x}) = \underline{x}'\underline{b}$의 분산은

$$V(\underline{x}) = \frac{Var(\hat{y}(\underline{x}))}{\sigma^2} = \underline{x}'(X'X)^{-1}\underline{x}$$

이 된다. 실험계획의 확장시나 실험계획점들 중 결측값이 발생하는 경우 각각은 원 실험계획에 후보점이 첨가되거나, 실험계획점이 탈락되므로 순차적 실험이 된다. 이러한 각각의 경우에 각 구성성분에 대하여 Cox 방향을 따라 양 경계영역 사이를 이동하면서 $V(\underline{x})$의 값을 구하여 그림으로 그린다. 이 그림을 추정반응값 분산그림(prediction variance trace(앞으로 PVT라 부르겠다.))이라 부른다. 이 그림을 이용하면 실험계획의 확장시나 실험계획점들 중 결측값이 있는 경우에, 제한이 있는 전 실험영역에 걸쳐 추정반응값 분산의 변화를 각 구성성분에 대해 알아 볼 수 있으므로 실험계획들을 비교, 평가하거나 확장 또는 결측의 효과를 순차적으로 알아 볼 수 있게 된다.

2.1 실험계획의 확장이 있는 경우

$\underline{x}_a$를 후보점들 중 원래 실험계획에 새로이 첨가되는 벡터라 하면, 확장된 모형행렬은

$$X_a = \begin{bmatrix} X \\ \underline{x}_a \end{bmatrix}$$

이 된다. 여기서 $\underline{x}_a$는 선택된 모형의 차수에 따라 달라진다. $\hat{y}_a(\underline{x})$를 확장된 모형행렬을 이용하여 얻어지는 추정반응값이라고 하면, σ^2으로 나눈 $\hat{y}_a(\underline{x})$의 분산은

$$V_a(\underline{x}) = \frac{Var(\hat{y}_a(\underline{x}))}{\sigma^2} = \underline{x}'(X'X + \underline{x}_a\ \underline{x}_a')^{-1}\underline{x}$$

이 된다. Sherman-Morrison-Woodbury 정리(Rao(1973))를 이용하면,

$$(X'X + \underline{x}_a\ \underline{x}_a')^{-1} = (X'X)^{-1} - \frac{(X'X)^{-1}\ \underline{x}_a\ \underline{x}_a'(X'X)^{-1}}{1 + \underline{x}_a'(X'X)^{-1}\ \underline{x}_a}$$

이 되므로,

$$V_a(\underline{x}) = \underline{x}'(X'X)^{-1}\underline{x} - \frac{\underline{x}'(X'X)^{-1}\ \underline{x}_a\ \underline{x}_a'(X'X)^{-1}\underline{x}}{1 + \underline{x}_a'(X'X)^{-1}\ \underline{x}_a} \qquad (1)$$

이 된다. 실험계획의 확장에 의한 결과는 (1)식의 오른쪽 두번째 항으로 나타나, 분산이 감소하게 된다. 각 구성성분에 대하여 Cox 방향을 따라 양 경계영역 사이를 이동하면서 $V_a(\underline{x})$의 값을 구하여 PVT를 그리면, 제한이 있는 전 실험영역에 걸쳐 실험계획의 확장으로 인한 추정반응값 분산의 변화를 쉽게 알 수 있다.

2.2 실험계획점 중 결측값이 발생하는 경우

원 실험계획점 중 i번째 실험계획점이 결측값이 되는 경우의 축소된 모형행렬을 X_{-i}라 하면,

$$(X_{-i}'X_{-i})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}\ \underline{x}_i\ \underline{x}_i'(X'X)^{-1}}{1 - h_{ii}}$$

가 된다. 여기서, $\underline{x}_i$는 결측값 벡터로서, 선택된 모형의 차수에 따라 달라진다. h_{ii}는 hat matrix $H = X(X'X)^{-1}X'$의 i번째 대각선 원소이다. i번째 실험계획점이 결측값이 되는 경우에, σ^2으로

나눈 추정반응값의 분산은

$$V_{-i}(\underline{x}) = \frac{Var_{-i}(\hat{y}(\underline{x}))}{\sigma^2} = \underline{x}'(X'X)^{-1}\underline{x} + \frac{\underline{x}'(X'X)^{-1}\underline{x}_i\,\underline{x}_i'(X'X)^{-1}\underline{x}}{1-h_{ii}} \qquad (2)$$

가 된다. 여기서, i번째 실험계획점이 결측되는 경우에 대한 결과는 (2)식의 오른쪽 두번째 항으로 나타나, 분산이 증가하게 된다. 각 구성성분에 대하여 Cox 방향을 따라 양 경계영역 사이를 이동하면서 $V_{-i}(\underline{x})$의 값을 구하여 PVT를 그리면, 제한이 있는 전 실험영역에 걸쳐 실험계획의 결측으로 인한 추정반응값 분산의 변화를 쉽게 알 수 있다.

3. 수치 예

3.1 실험계획의 확장이 있는 경우

McLean과 Anderson(1966)은 화염실험을 통하여 4개의 성분들에 다음과 같은 제약조건을 주어 표1과 같은 27개의 후보점들을 제시하고, 이 중 15개의 실험점들로 구성된 McLean-Anderson 실험계획을 제시하고, 이 실험계획을 이용하여 2차모형을 추정하였다.

<표 1> McLean과 Anderson 화염실험에서의 후보점들

	꼭지점					변 중심점					면 중심점			
	x1	x2	x3	x4		x1	x2	x3	x4		x1	x2	x3	x4
1	0.40	0.10	0.47	0.03	9	0.40	0.100	0.445	0.055	21	0.40	0.2725	0.2725	0.055
2	0.60	0.10	0.27	0.03	10	0.40	0.445	0.100	0.055	22	0.60	0.1725	0.1725	0.055
3	0.40	0.47	0.10	0.03	11	0.40	0.285	0.285	0.030	23	0.50	0.1000	0.3450	0.055
4	0.60	0.27	0.10	0.03	12	0.40	0.260	0.260	0.080	24	0.50	0.3450	0.1000	0.055
5	0.40	0.10	0.42	0.08	13	0.60	0.100	0.245	0.055	25	0.50	0.2350	0.2350	0.030
6	0.40	0.42	0.10	0.08	14	0.60	0.245	0.100	0.055	26	0.50	0.2100	0.2100	0.080
7	0.60	0.10	0.22	0.08	15	0.60	0.185	0.185	0.030					
8	0.60	0.22	0.10	0.08	16	0.60	0.160	0.160	0.080	27*	0.50	0.2225	0.2225	0.055
					17	0.50	0.100	0.370	0.030					
					18	0.50	0.100	0.320	0.080					
					19	0.50	0.370	0.100	0.030					
					20	0.50	0.320	0.100	0.080					

* 전체 중심점

$$0.40 \le x_1 \le 0.60$$
$$0.10 \le x_2 \le 0.50$$
$$0.10 \le x_3 \le 0.50$$
$$0.03 \le x_4 \le 0.08$$

여기서, x_1은 magnesium의 비율, x_2는 sodium nitrate의 비율, x_3는 strontium nitrate의 비율, x_4는

binder의 비율이다.

다음 표2에 15개의 실험점들로 구성된 McLean-Anderson 실험계획(약자로 M-A계획이라 하겠다.)과, D-최적화 실험계획을 제시하였다. 15개로 구성된 M-A계획에다 나머지 12개의 후보점들 중 하나를 첨가해 16개로 구성된 확장 M-A계획을 만들었을 때, 추정반응값 분산의 변화가 가장 심한 후보점은 13번(또는 14번)이었다. 그림1은 13번(또는 14번) 후보점을 첨가한 확장 M-A계획과 원래의 M-A계획을 비교한 PVT이다. 15개로 구성된 D-최적화 계획에다 나머지 12개의 후보점들 중 하나를 첨가해 16개로 구성된 확장 D-최적화 계획을 만들었을 때, 추정반응값 분산의 변화가 가장 심한 후보점은 27번이었다. 그림2는 27번 후보점을 첨가한 확장 D-최적화 계획과 원래의 D-최적화 계획을 비교한 PVT이다. 확장 D-최적화 계획인 경우가 확장 M-A계획일 때보다 더 큰 변화를 나타내었다. 특히 각 구성성분의 제한영역의 가운데에서 극명한 변화가 나타났다.

<표 2> 표1을 이용한 혼합물 실험계획

	M-A*	D
꼭지점 1-8	전부	전부
변 중심점 9-20		9
		11
		13
		17
		18
면 중심점 21-26	21	21
	22	24
	23	
	24	
	25	
	26	
전체 중심점 27	27	

* M - A : McLean-Anderson 실험계획, D : D-최적화 계획

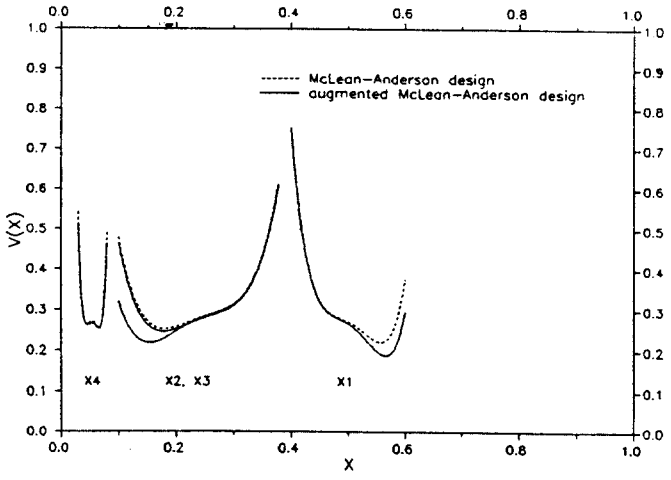

그림 1. McLean-Anderson 실험계획과 13번 후보점을 첨가한 확장 McLean-Anderson 실험계획의 추정반응값 분산

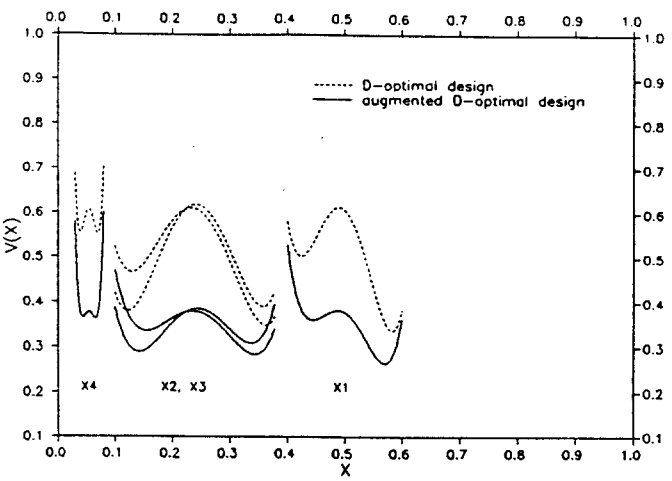

그림 2. D-최적화 실험계획과 27번 후보점을 첨가한 확장 D-최적화 실험계획의 추정반응값 분산

3.2 실험계획점 중 결측값이 발생하는 경우

3.1절의 예를 계속 이용하여 15개의 실험점들로 구성된 실험계획에서 결측값이 발생하여 14개의 실험점들로 구성된 축소 실험계획이 되었을 때의 추정반응값 분산의 변화를 PVT를 이용하여 살필 수 있다. 15개로 구성된 M-A계획에서 실험계획점을 하나 제거하여 14개로 구성된 축소 M-A계획을 만들었을 때 추정반응값 분산의 변화가 가장 심한 실험계획점은 21번이었다. 그림3은 21번

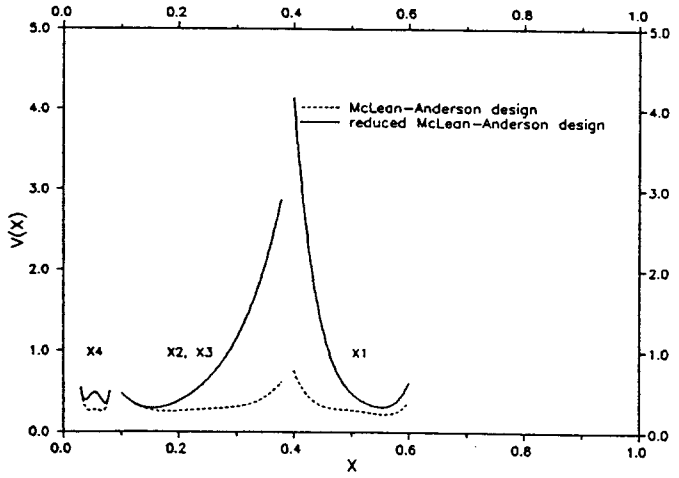

그림 3. McLean-Anderson 실험계획과 21번 실험점이 결측된 축소McLean-Anderson 실험계획의 추정반응값 분산

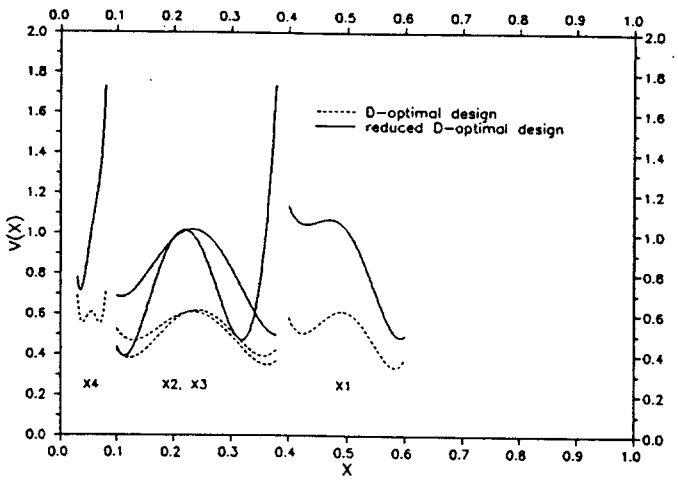

그림 4. D-최적화 실험계획과 3번 실험점이 결측된 축소 D-최적화 실험계획의 추정반응값 분산

을 제거한 축소 M-A계획과 원래의 M-A계획을 비교한 PVT이다. 15개로 구성된 D-최적화 계획에서 실험계획점을 하나 제거하여 14개로 구성된 축소 D-최적화 계획을 만들었을 때, 추정반응값 분산의 변화가 가장 심한 실험계획은 3번과 24번이었다. 그림4와 그림5는 3번과 24번을 각각 제거한 축소 D-최적화 계획과 원래의 D-최적화 계획을 비교한 PVT들이다. 두 그림에서 알 수 있듯이 3번을 제거하였을 때의 변화는 24번을 제거하였을 때의 변화하고는 상당히 다른 변화를 나타내고 있다. 그림6에서 그림8까지는 M-A계획에서 21,22,23,24번의 실험계획점들을 순차적으로 제거시켰을 때의 각 구성성분에 대한 추정반응값 분산의 변화를 나타낸 그림들이다. 여기서 알 수 있는 것은 변화가 가장 심한 구성성분은 x_1이고, 변화가 가장 적은 구성성분은 x_4이고, 각 구성성분에 따라 변화의 모습이 다르다는 것이다.

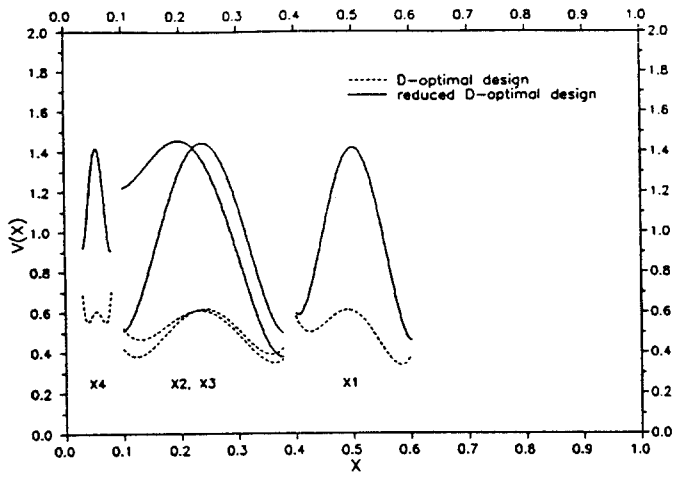

그림 5. D-최적화 실험계획과 24번 실험점이 결측된 축소 D-최적화 실험계획의 추정반응값 분산

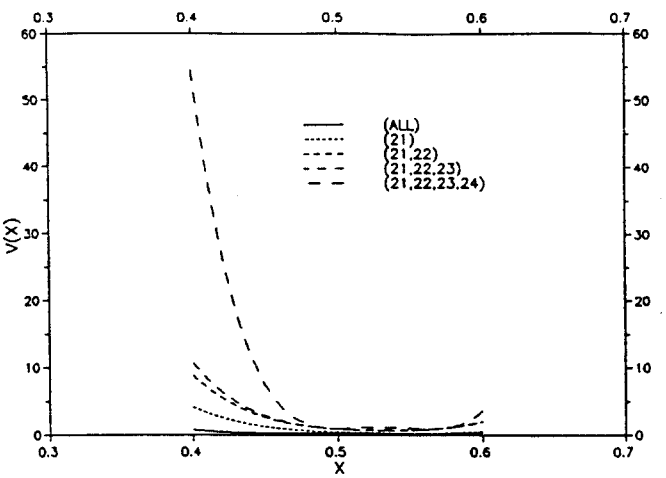

그림 6. McLean-Anderson 실험계획에서 순차적으로 21,22,23,24실험계획점들을 제거시켰을 때의 x_1 성분에 대한 추정값 반응값 분산

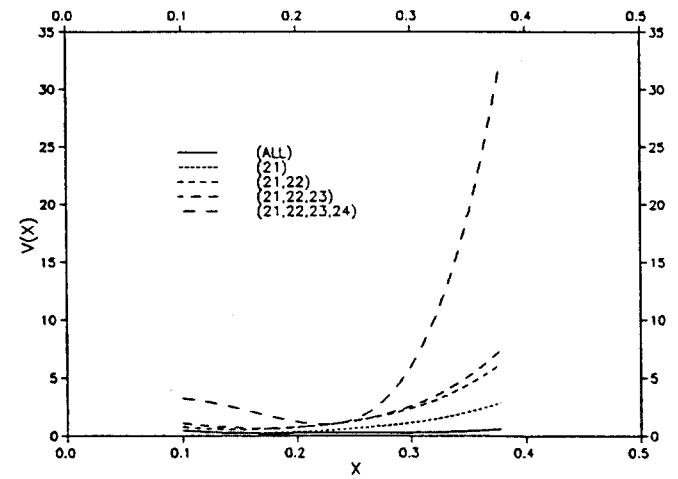

그림 7. McLean-Anderson 실험계획에서 순차적으로 21,22,23,24 실험계획점들을 제거시켰을 때의 x_2 성분 (또는 x_3 성분)에 대한 추정반응값 분산

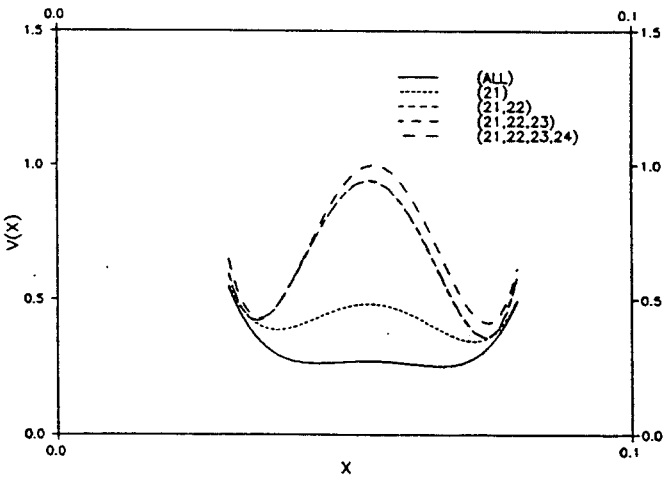

그림 8. McLean-Anderson 실험계획에서 순차적으로 21,22,23,24실험계획점들을 제거시켰을 때의 x_4 성분에 대한 추정반응값 분산

4. 결론

혼합물 실험계획의 확장시 기존의 연구들은 D-최적계획을 중심으로 전개되어 왔다. 이 논문에서 이러한 알파벳 최적화 기법들의 문제점을 해결할 수 있는 그래픽 방법을 제안하였다. 이 그래픽 방법인 추정반응값 분산 그림을 이용하여 실험계획의 확장시나 실험계획점들 중 결측값이 발생하는 경우, 제한이 있는 전 실험계획 영역에 걸쳐 추정반응값 분산의 변화를 순차적으로 볼 수 있고, 실험계획들을 서로 비교, 평가할 수 있다. 특히, 서로 경쟁적인 혼합물 실험계획에 대하여 제한이 있는 전 실험계획 영역에 걸쳐 추정반응값 분산의 변화를 한 눈에 확인할 수 있으므로 실험계획들의 장,단점을 쉽게 파악할 수 있다. 또한, 한 실험계획에 대하여 실험계획의 확장 또는 결측의 결과를 추정반응값 분산의 변화를 보면서 확인할 수 있다. 이 과정을 통하여 각 구성성분들에 대하여 추정반응값 분산의 크기 변화를 알아내어 구성성분들의 중요도에 서열을 매길 수 있다. 이 그래픽 방법은 혼합물 실험계획 영역에 제한이 없는 경우에도 활용할 수 있다. 이 논문의 확장으로서, 분산 뿐만이 아니라 편의도 고려하여 평균제곱오차로 까지 확장하여 그림을 작성할 수 있다.

참 고 문 헌

[1] Covey-Cramp, P.A.K. and Silvey, S.D.(1970). Optimal Regression Designs with Previous Observations, *Biometrica*, 57, 551-566.

[2] Dykstra, O. Jr.(1966). The Orthogonalization of Undesigned Experiments, *Technometrics*, 8, 279-290.

[3] _____ (1971). The Augmentation of Experimental Data to Maximize $|X'X|$, *Technometrics*, 13, 682-688.

[4] Evans, J. W. (1979). Computer Augmentation of Experimental Designs to Maximize $|X'X|$, *Technometrics*, 21, 321-330.

[5] Gaylor, D. W. and Merrill, J. A. (1968). Augmenting Existing Data in Multiple Regression, *Technometrics*, 10, 73-81.

[6] Giovannitti-Jensen, A. and Myers, R. H. (1989). Graphical Assessment of the Prediction Capability of Response Surface Designs, *Technometrics*, 31, 159-172.

[7] Hebble, T. L. and Mitchell, T. J. (1972). Repairing Response Surface Designs, *Technometrics*, 14, 767-779.

[8] Jang, D. H. and Park, S. H. (1993). A Measure and A Graphical Method for Evaluating Slope Rotatability in Response Surface Designs, *Communications in Statistics-Theory and Methods*, 22, 1849-1863.

[9] Mayer, L. S. and Hendrickson, A. D. (1973). A Method for Constructing on Optimal Regression Design After an Initial Set of Input Values Has Been Selected, *Communications in Statistics- Theory and Methods*, 2, 465-477.

[10] McLean, R. A. and Anderson, V. L. (1966). Extreme Vertices Design of Mixture

Experiments, *Technometrics*, 8, 447-454.

[11] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., John Wiley and Sons, New York.

[12] Vining, G. G., Cornell, I. A. and Myers, R. H. (1993). A Graphical Approach for Evaluating Mixture Designs, *Applied Statistics*, 42, 127-138.

[13] Vining, G. G. and Myers, R. H. (1991). A Graphical Approach for Evaluating Response Surface Designs in terms of The Mean Squared Error of Prediction, *Technometrics*, 33,315-326.

[14] Wynn, H. P. (1970). The Sequential Generation of D-optimum Experimental Designs, *The Annals of Mathematical Statistics*, 41, 1655-1664.