응용통계연구 제 13권 2호, 2000년, pp. 477-488

라틴-하이퍼큐브 실험계획 간의 거리 계산과 비교*

박정수[1] 황현식[2]

요 약

전산실험계획으로 유용하게 쓰이는 라틴-하이퍼큐브 계획간의 거리를 정의하고 그 기대값을 계산하였다. 이 계산을 위해서 차원이 증가함에 따라 수리 통계학적 방법, 수치 해석적 방법 (다차원 수치 적분법), 몬테카를로 적분 방법, 극한 정규분포이론을 이용하여 거리의 기대값을 구했다. 또한 같은 구조를 가지면서 랜덤성에 차이가 있는 두 라틴-하이퍼큐브 계획 간에 반응함수의 평균에서의 차이 및 정보량의 차이를 다루었다. 본 논문에서 제시한 두 Lhd 들간의 비교 기법은 두 개의 일반 실험계획의 비교에도 유용하리라 여겨진다.

주요용어: 몬테카를로 방법, 상대효율, 수치분석, 전산실험.

1. 서론

전산 실험계획(designs for computer experiments)은 모의실험에서의 입력변수의 적절한 조합을 통하여 적은 실험으로 원하는 수준의 정보를 얻기 위한 실험계획을 말한다. 전산 모의 실험의 특징은 입력변수에 대한 반응함수가 대부분 결정적이거나 비결정적이다고 해도 그 관측오차가 물리적인 실험에 비하여 상당히 작은 값이다는 점이다. 또한 대개의 경우 입력변수 수가 매우 많으므로 특히 수행시간이 오래 걸리는 경우에 기존의 실험계획과는 다른 새로운 실험계획이 요구된다(Sacks, Welch, Mitchell, Wynn, 1989).

대표적인 전산 실험계획으로 McKay, Beckman, Conover(1979)에 의해서 제안된 라틴-하이퍼큐브 계획(Latin-hypercube design, 약해서 Lhd)이 있다. 이 실험계획은 구축하기가 매우 쉽고 모의실험을 비교적 적게 수행하면서도 실험점들이 입력변수 영역에 골고루 퍼져있게 되는(space-filling) 특성을 가진다. 이러한 장점 때문에 Lhd는 실제로 대형 전산 시뮬레이션 코드에 매우 자주 이용되고 있다 (예를들어 [5], [7], [14]).

Stein(1987)은 Lhd가 대표본에서 단순임의추출보다 분산이 감소된다는 점을 밝혔고, Owen (1992)은 Lhd 에서의 반응함수의 평균에 대한 중심 극한 정리를 다루어 이론적인 면을 추가하였다. 한편, 이러한 Lhd 역시 실험점들이 실험영역 전체에서 골고루 배치되지 않을 가능성이 있으므로 이를 보완하려는 시도가 이루어졌다. Tang(1993) 은 직교배열을 이용하여 구축한 OA-based Lhd 를 제안하였고, Morris and Mitchell(1995)은 최대최소거리(maxmin distance) Lhd 를 제안하였는데, 이들은 실험점들이 라틴-하이퍼큐브의 구조를

* 이 논문은 1999년도 전남대학교 학술연구비 지원에 의하여 연구되었음.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 정보통신연구소 및 자연과학대학 통계학과, 부교수
E-mail: jspark@chonnam.ac.kr
2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 강사

가지면서도 실험영역에 골고루 배치되도록 하고 있다. 또한 Park(1994)은 Lhd 중에서 가우시안 확률과정과 연계하여, 알고리즘을 이용한 최적 Lhd 를 연구하였다. 임미정, 권우주, 이주호(1995) 및 Lee(1999)는 충화 Lhd 에 관하여 연구하였다.

본 논문에서는 같은 구조를 가지면서 랜덤성에 차이가 있는 Lhd 간의 거리를 정의하고 그 거리의 기대값을 계산했다. 또한 두 Lhd 의 비교를 위하여 두가지 측면(반응치의 평균의 차이와 실험계획이 갖는 정보의 상대비율)에서 연구하였다. 제2절에서는 Lhd 에 관한 일반적 특징을 알아보고, 중심 Lhd와 랜덤 Lhd를 정의하였다. 제3절에서는 두 실험계획 간의(특히 Lhd 간의) 거리를 정의하며, 제4절에서는 두 Lhd 간의 거리의 기대값을 구하는 실제 계산 과정과 그 결과를 제시하였다. 제5절에서 두 Lhd의 반응치의 평균 차이를 다루었고, 제6절에서는 두 실험계획이 갖는 정보의 상대비율을 통하여 서로 비교하였다. 마지막으로 제7절에 요약 및 토의사항을 실었다.

2. 라틴-하이퍼큐브 계획

Lhd는 실험공간 전체에서 실험점이 골고루 추출되도록 각 입력 변수의 범위를 n개의 범위로 나눈 다음, 각 구간에서 하나씩 추출하되 중복되지 않게 n개를 뽑는 방법이다. 이를 구체적으로 기술하면 다음과 같다. 여기서 n 은 실험점의 수이고 d 는 입력변수의 수로서 전산실험의 차원을 나타낸다.

먼저 입력변수 들이 값을 가질 수 있는 범위에 바탕하여 실험영역을 각 변수에 대해 $[0,1]$로 표준화한다. 실험 영역 $[0,1\,]^d$ 에서 각각이 독립인 입력 변수 $(X_1,\cdots,X_d)$와 그에 해당하는 분포함수 $(F_1,\cdots,F_d)$ 가 있고, x_{ij} 를 j 번째 변수 (X_j)의 i 번째 실험점이라고 하자$(i=1,\cdots,n\ ,\ j=1,\cdots,d)$. 또 $P=(P_{ij})$ 는 $n\times d$ 행렬로서 P의 각 열은 $(1,\cdots,n)$의 순열이며, 이러한 각 열들은 확률적으로 독립이라고 하자. 이제 Lhd 를 이루는 실험점 x_{ij} 는 다음과 같이 정의된다.

$$x_{ij}=F_j^{-1}\left(\frac{1}{n}(P_{ij}-r_{ij})\right). \tag{2.1}$$

여기서 r_{ij} 는 각각이 독립적으로 일양분포 [0, 1]을 따르는 확률변수이다. (2.1)에서 순열 $P_{i1},\cdots,P_{id}$ 는 실험영역에서 x_{ij} 가 속하는 격자(cell)를 결정하고, 그 격자 속에서의 정확한 위치는 $r_{i1},\cdots,r_{id}$ 에 의해서 결정된다. (2.1) 에서 모든 i와 j에 관하여 $r_{ij}=1/2$ 인 Lhd 를 "중심 Lhd" (Midpoint Lhd, 약해서 MLhd) 라고 부른다(Park, 1994). 한편 r_{ij} 가 $U[0,1]$ 을 따르는 보통의 경우를 중심 Lhd 와 구분하기 위해서 "랜덤 Lhd"(Random Lhd, 약해서 RLhd)라고 하자. 또한 (2.1)에서 P_{ij} 가 같은 Lhd 들을 "구조가 같은 Lhd 들" 이라고 부른다. 따라서 같은 구조를 갖는 한 개의 MLhd 와 여러 개의 RLhd 가 있을 수 있다.

본 논문에서는 같은 구조를 갖는 MLhd 와 RLhd 를 비교하기 위하여 그들 간의 거리와 반응치의 평균에 대한 분산 및 정보량을 비교 연구하였다. 본 연구의 동기는 Park(1994) 의 최적 Lhd 를 알고리즘으로 구하는 연구에서 제기되어 시작되었다. 즉, Park(1994) 의 알고리즘은 먼저 최적 MLhd 를 구하고, 그 상태(같은 구조)에서 임의성을 고려하는 최적 RLhd 를 구하는 식으로 진행되는데, 이때 최적 RLhd 까지 가지 않고 그냥 최적 MLhd 로 만족한

다면 어떻겠는가 하는 의문에서 시작되었다. 이 의문에 대한 대답으로서 고정된 MLhd 와 같은 구조를 갖는 RLhd 들을 생각하여 그들을 비교하였다. 최적 RLhd 는 최적 MLhd 와 같은 구조를 유지하면서도 임의성을 고려해서 구해진 형태이며, 최적인 두 실험계획을 비교하는 것은 구조가 같은 일반적인 MLhd 와 RLhd 를 비교하는 것의 특수한 경우가 된다.

3. 두 실험계획간의 거리

d차원의 n개의 실험점을 갖는 두 개의 실험계획 D_1과 D_2에 대해서, 두 실험계획 간의 거리를 다음과 같이 정의하자.

정의 3.1 두 실험계획 $D_1 = [\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n]^t$ 와 $D_2 = \left[\underline{y}_1, \underline{y}_2, \cdots, \underline{y}_n\right]^t$ 간의 거리는

$$\delta(D_1, D_2) = \min_{\pi(i)} \sum_{i=1}^{n} ||\underline{x}_i - \underline{y}_{\pi(i)}|| \tag{3.1}$$

이다. 여기서 $\pi(i)$는 $1, \cdots, n$의 순열이다.

$\delta(D_1, D_2)$의 의미 해석은 다음과 같다. 먼저 D_1의 각 실험점들과 D_2의 각 실험점들 간의 가능한 1대1 짝을 이루었을 때의 그 거리의 합을 생각한다. 이렇게 짝을 이루는 방법은 $n!$ 만큼 존재하므로 거리의 합 역시 $n!$ 만큼 존재하게 된다. 이 중 최소의 거리의 합이 두 실험계획 간의 거리, 즉 $\delta(D_1, D_2)$이 된다. 따라서 고정된 두 실험계획 간의 거리는 유일하게 된다(비록 이 거리를 구성하는 짝들은 유일하지 않더라도). 본 논문에서는 두 실험계획 중 하나(MLhd)는 고정되어 있지만 다른 하나(RLhd)는 임의성을 포함하고 있기 때문에 평균거리, 즉 $E[\delta(D_1, D_2)]$ 를 고려하였다.

한편 위에서 고려한 모든 가능한 짝들 중에 최소의 거리의 합을 이루는 짝들을 잡는 것, 즉 최적의 순열 $\pi^*(i)$를 찾는 것을 알고리즘 이론 또는 그래프 이론에서 perfect bipartite matching 이라고 하며 보통 Hungarian algorithm 으로 구현할 수 있다 (Bondy and Murty, 1976). 특히 짝을 이루는 점들이 임의성을 가지는 경우에 이들 짝들과 거리의 특성에 관한 연구는 Talagrand (1992) 및 Friez and Janson(1995) 에서 찾아볼 수 있다. 또한 위와 같은 matching 아이디어는 통계학에서 Easton and McCulloch(1990) 에 의해서 다변량 자료에 대한 적합도 검정에 응용되었다(김남현, 1999).

위의 정의 3.1에서 $||\underline{x}_i - \underline{y}_{\pi(i)}||$ 에 사용되는 거리로서, 우리는 다음과 같은 제곱거리(L_2^2), 절대거리(L_1) 및 유클리디안 거리(L_2)를 고려하였다.

$$L_2^2 \ : \ \sum_{j=1}^{d}(x_{ij} - y_{\pi(i)j})^2 \ , \ L_1 \ : \ \sum_{j=1}^{d}|x_{ij} - y_{\pi(i)j}| \ , \ L_2 \ : \ \sqrt{\sum_{j=1}^{d}(x_{ij} - y_{\pi(i)j})^2}.$$

이제 d차원의 n개의 실험점을 갖는, 주어진 MLhd 와 구조가 같은 RLhd 와의 거리를 생각해 보자. 이를 위해서는 원래는 위에서 논의한 바대로 Hungarian algorithm 을 이용하여 최적 순열을 찾은 다음 거리를 계산해야 겠지만, Lhd 의 특별한 구조상 아래의 정리를 이용하여 쉽게 거리의 계산이 가능하다.

정리 3.1 같은 구조를 갖는 MLhd $D_M = [\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n]^t$ 과 한 고정된 RLhd $D_R = \left[\underline{y}_1, \underline{y}_2, \cdots, \underline{y}_n\right]^t$ 의 거리는 분포함수 F_j 에 상관없이 $\delta(D_M, D_R) = \sum_{i=1}^n \|\underline{x}_i - \underline{y}_i\|$ 가 된다. (여기서, $\underline{y}_i = \underline{x}_i + \underline{r}_i, i = 1, 2, \cdots, n$ 이다.)

증명: 같은 구조를 가지면 D_M 과 D_R 은 같은 순열행열을 가진다. 그러므로 Lhd 의 특성상 실험점이 있는 각 cell에서 중심점과 가장 가까운 실험점은 그 cell 안에 존재한다. 따라서 같은 cell 내의 점들끼리 짝을 이루어 계산되는 거리가 (3.1)에서의 최소 거리이다. □

본 논문에서 이제부터는 모든 F_j 가 일양분포 $U(0,1)$ 을 따른다고 가정한다. 그러면 MLhd와 RLhd 의 차이는 모든 i 와 j 에 대해서 $\frac{1}{n}\left(\frac{1}{2} - r_{ij}\right)$ 만큼이 된다. 이것은 $\frac{1}{2n}U(-1,1)$를 갖는 확률변수이다. 따라서 $\delta(D_M, D_R) = \sum_{i=1}^n \delta_i$ 로 표현할 때, 각 δ_i 는 i 에 무관하게, 각각 거리의 정의에 따라

$$L_2^2 \ : \ \frac{1}{2n}\sum_{j=1}^d u_j^2, \ L_1 \ : \ \frac{1}{2n}\sum_{j=1}^d |u_j|, \ L_2 \ : \ \frac{1}{2n}\sqrt{\sum_{j=1}^d u_j^2}$$

으로 된다. 이때 $\{u_j, j = 1, \cdots, d\}$ 는 서로 독립인 $U(-1,1)$ 을 따르는 확률변수들이다.

이제 구조가 같은 MLhd와 RLhd의 거리는 확률변수이므로 그 거리의 기대값 $E[\delta(S_M, S_R)]$을 구해보자. 이들은 각각 거리의 정의에 따라,

$$L_2^2 : \ \frac{1}{2}\sum_{j=1}^d E(u_j^2) = \frac{1}{2} \times \frac{1}{3}d = \frac{1}{6}d,$$

$$L_1 : \ \frac{1}{2}\sum_{j=1}^d E(|u_j|) = \frac{1}{4}d,$$

$$L_2 : \ \frac{1}{2}E\left(\sqrt{\sum_{j=1}^d u_j^2}\right)$$

이다. 이들은 세 가지 경우 모두에서 n에 무관하고 d의 함수로 표현된다는 점이 특이하다. L_2의 경우에는 곧바로 구해지지 못하는데 이의 계산은 다음 절에서 자세히 다룬다.

4. Lhd 간의 거리의 기대값의 계산 : 적분 문제

먼저 $U(-1,1)^2 = U(0,1)^2$ 이므로 L_2 경우의 거리의 기대값 계산에서 $u_j \sim U(0,1)$ 을 이용한다. 이제 $E(\sqrt{u_1^2 + u_2^2 + \cdots + u_d^2})$ 의 계산을 위하여 적분이 필요한데, $d = 2$인 경우에는 아래에서 처럼 확률밀도함수를 구하여 계산이 가능하나 d 가 3이상이면 곤란하다. 따라서 우리는 실제 계산을 위하여 다음의 근사적인 방법을 사용하였다.

(1) 확률밀도함수를 이용한 기대값의 직접 계산 ($d = 2$).

(2) 가우스 구적법(Gaussian quadrature)을 이용한 다차원 수치적분 ($d = 3$ 에서 12).

(3) 대조변수(antithetic variable)를 이용한 몬테카를로 적분 ($d = 9$ 에서 20).

(4) 중심극한정리를 적용한 점근적 정규분포의 기대값 이용 ($d = 21$ 이상).

$d = 2$ 인 경우, $U_1^2 + U_2^2$ 의 확률밀도함수를 이용하여 기대값을 계산한다. 변수변환 $Y_1 = U_1^2 + U_2^2$, $Y_2 = U_1^2 - U_2^2$ 을 통하여 Y_1과 Y_2의 결합 확률밀도함수는 다음 식으로 구해진다 (여기서 $X = U(0,1)^2$ 의 확률분포함수는 $\frac{1}{2\sqrt{x}}$ 이다는 점이 이용되었다).

$$f(y_1, y_2) = \frac{1}{4\sqrt{y_1^2 - y_2^2}} .$$

이제 Y_1의 주변 확률밀도함수를 각 영역에서 구하면,

$$f_1(y_1) = \begin{cases} \pi/4 & , \ 0 < y_1 \le 1 \\ \frac{1}{2}\sin^{-1}(\frac{2}{y_1} - 1) & , \ 1 < y_1 \le 2 \end{cases}$$

이 된다. 이를 이용하여 결국 구하는 기대값은 다음과 같다.

$$\begin{aligned} E(\sqrt{y_1}) &= \int_0^1 \frac{\pi}{4}\sqrt{y_1}dy_1 + \int_1^2 \frac{\sqrt{y_1}}{2}\sin^{-1}(\frac{2}{y_1} - 1)dy_2 \\ &\doteq (\pi/6 + 0.2416)/2 \\ &= 0.3826. \end{aligned}$$

이상에서와 같이 $d = 2$인 경우에는 기대값을 정확히 구할 수 있으나, d 가 3 이상인 경우에는 위의 방법으로 구하기 어렵다. 따라서 우리는 수치적분 및 몬테카를로 적분을 이용하여 기대값을 근사적으로 구한다. 이러한 계산을 하기 전에 먼저 $E(\sqrt{u_1^2 + u_2^2 + \cdots + u_d^2})$의 상한과 하한을 구하여 보자.

상한은 Jensen의 부등식을 적용하여, 서로 독립인 $U(0,1)$ 분포이므로

$$\begin{aligned} E\left(\sqrt{u_1^2 + u_2^2 + \cdots + u_d^2}\right) &\le \sqrt{E(u_1^2 + u_2^2 + \cdots + u_d^2)} \\ &= \sqrt{dE(u_1^2)} = \sqrt{d\frac{1}{3}} . \end{aligned}$$

하한은

$$\sqrt{d\left(\frac{u_1 + u_2 + \cdots + u_d}{d}\right)^2} \le \sqrt{\sum_{i=1}^{d} u_i^2}$$

을 이용하여

$$\frac{\sqrt{d}}{2} \le E\left(\sqrt{u_1^2 + u_2^2 + \cdots + u_d^2}\right)$$

로 구한다. 따라서 거리 기대값의 하한과 상한은 각각 $\sqrt{d}/2$ 와 $\sqrt{d/3}$ 이다.

표 4.1: Lhd 간 거리의 최종 기대값과 상하한

d	하한/2	기대값/2	상한/2	비고
1	0.2500	0.2500	0.2886	확률밀도함수를
2	0.3535	0.3826	0.4082	이용
3	0.4330	0.4802	0.5000	
4	0.5000	0.5609	0.5773	
5	0.5590	0.6312	0.6455	가우스 적분을 이용한
6	0.6123	0.6942	0.7071	다차원 수치 적분
7	0.6614	0.7520	0.7637	
8	0.7071	0.8056	0.8164	
9	0.7500	0.8558(0.8549)	0.8660	수치적분(몬테카를로)
10	0.7905	0.9032(0.9066)	0.9128	
11	0.8291	0.9481(0.9480)	0.9574	몬테카를로(수치적분)
12	0.8660	0.9915(0.9913)	1.0000	
13	0.9013	1.0329	1.0408	
14	0.9354	1.0724	1.0801	대조변수를 이용한
15	0.9682	1.1100	1.1180	몬테카를로 적분
20	1.1180	1.2845	1.2909	
25	1.2500	1.4433	1.4433	
30	1.3693	1.5811	1.5811	
35	1.4790	1.7078	1.7078	점근적 정규분포 이용
40	1.5811	1.8257	1.8257	
50	1.7677	2.0412	2.0412	

이제 수치적분 및 몬테카를로 적분을 위하여 목적하는 기대값을 적분 형태로 표현하면 다음과 같다.

$$\int_0^1 \int_0^1 \cdots \int_0^1 \sqrt{u_1^2 + u_2^2 + \cdots + u_d^2}\, du_1 du_2 \cdots du_d.$$

$d = 3 \sim 12$ 에서는 가우스 구적법(Gaussian quadrature)을 적용한 다차원 수치 적분법으로 거리의 기대값을 구했다 [표 4.1] (IMSL의 DQAND를 이용하여 계산). 비교적 입력변수의 수가 적은 저차원인 경우에는 수치적분을 이용하였지만, $d > 8$ 인 경우에서는 몬테카를로 적분으로 계산하였다. 고차원에서는 수치적분의 적용보다는 몬테카를로 방법이 더 정확할 것으로 알려졌기 때문이다(Ripley, 1987). 이 몬테카를로 적분을 적용시, 거리의 기대값은 u_j 들의 단조 증가함수이므로, 분산 감소기법인 대조변수(antithetic variable)를 이용하였다(Ripley, 1987). 즉 500개의 일양난수(u)를 먼저 생성하고, u^2 과 $(1-u)^2$ 으로 1000개를 최종적으로 다음과 같이 사용하였다[표 4.1].

$$E\left(\sqrt{u_1^2+u_2^2+\cdots+u_d^2}\right) \approx \frac{1}{1000}\left(\sum_{k=1}^{500}\sqrt{\sum_{i=1}^{d}u_i^2}+\sum_{k=1}^{500}\sqrt{\sum_{i=1}^{d}(1-u_i)^2}\right).$$

특히 d 가 9에서 12까지는 두 가지 방법(가우스 구적법과 몬테카를로 방법)을 모두 적용하였다.

$d>20$ 인 경우에는 중심극한정리를 이용하여 다음과 같이 근사적 정규분포를 구하였다.

$$\begin{aligned} u_1^2+u_2^2+\cdots+u_d^2 &\sim \mathrm{AN}(d\,E(u^2),\ d\,Var(u^2)) \\ &= \mathrm{AN}(d/3,\ 4d/45). \end{aligned}$$

여기서 확률변수 $Y \sim \mathrm{AN}[d/3, 4d/45]$ 이면, d 가 20 이상만 되어도 $P(Y\le 0)\to 0$ 이 되므로 델타법(delta method)인 $g(Y)\sim \mathrm{AN}(g(\mu),\ g'(\mu)^2\sigma^2)$ 을 이용하여

$$\sqrt{Y} \sim \mathrm{AN}(\sqrt{d/3},\ \sqrt{3d}/45)$$

을 구하게 된다. 따라서 이 경우 $E(\sqrt{u_1^2+u_2^2+\cdots+u_d^2}\,)\approx\sqrt{d/3}$ 로 근사된다[표 4.1]. 주시할 점은 이 값이 앞에서 구한 기대값의 상한과 같다는 점이다.

5. Lhd에 대한 반응치의 평균 차이

여기서는 두 실험계획의 반응치의 평균, 즉 MLhd 에서의 $\bar{Y}$ 의 기대값 $E(\bar{Y}_{ML})$과 RLhd 에서의 $\bar{Y}$ 의 기대값 $E(\bar{Y}_{RL})$의 차이를 구한다. McKay, Beckman, Conover(1979) 에 의해서 이미 $\bar{Y}_{RL}$ 은 $E(Y)$ 의 불편추정치임이 알려져 있다. 먼저 $Y_i=h(\underline{x}_i)=h(\underline{m}_i+\underline{u}_i)$ 라 하고 또한 다음과 같이 표시하자.

$$\underline{m}_i = i\text{ 번째 중심점의 좌표},\quad \underline{u}_i = \text{서로 독립인 } U(-1,1)/2n\text{인 } d\times 1 \text{ 벡터}.$$

테일러 급수전개에 의해서

$$h(\underline{x}_i) \approx h(\underline{m}_i)+G_i^t\underline{u}_i+0.5\underline{u}_i{}^tH_i\underline{u}_i.$$

여기서 G_i 는 $\underline{m}_i$ 에서의 h 의 미분벡터(gradient vector) 이고, H_i 는 $\underline{m}_i$ 에서의 h 의 헤시안 행렬(Hessian matrix) 이다. 그런데 $\bar{Y}_{ML}=\frac{1}{n}\sum_{i=1}^{n}h(\underline{m}_i)$ 이므로,

$$\begin{aligned} E(\bar{Y}_{RL}-\bar{Y}_{ML}) &\approx \frac{1}{n}\sum_{i=1}^{n}E(G_i^t\underline{u}_i+.5\underline{u}_i{}^tH_i\underline{u}_i) \\ &= \frac{1}{8n^2}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}h_j''(\underline{m}_i) = O(\frac{d}{n^2}). \end{aligned}$$

따라서 $\bar{Y}_{ML}$ 은 편의되어(biased) 있지만 n이 d에 비하여 충분히 클 때, 또한 특히 함수 h가 부드러울수록 편의의 양은 줄어든다.

6. 정보량 비교

Mitchell, Sacks and Ylvisaker(1994)는 두 실험계획의 비교를 위하여 다음과 같이 엔트로피(entropy)의 비율로 표현되는 상대효율(relative efficiency) 을 사용하였다.

$$\text{상대효율} = \frac{-n\log\sigma^2 + \log|V_R|}{-n\log\sigma^2 + \log|V_M|}. \tag{6.1}$$

여기서 한 실험계획에 대해 계산되는 공분산 행렬 V 는 다음과 같은 원소로 구성된다고 가정된다. 두 개의 실험점

표 6.1: MLhd 와 RLhd 의 상대효율

d = 2			d = 4		
n	σ^2	상대효율	n	σ^2	상대효율
5	0.1	0.9565	5	0.1	0.9991
	1	1.0313		1	1.0047
	10	1.0115		10	1.0006
12	0.1	1.0047	12	0.1	1.0005
	1	1.0020		1	0.9994
	10	1.0013		10	0.9998
25	0.1	1.0016	25	0.1	1.0239
	1	1.0011		1	0.9994
	10	1.0009		10	0.9996
d = 9			d = 15		
n	σ^2	상대효율	n	σ^2	상대효율
15	0.1	1.0005	15	0.1	1.0002
	1	0.9959		1	0.9921
	10	0.9996		10	0.9998
25	0.1	1.0004	25	0.1	1.0001
	1	0.9982		1	0.9962
	10	0.9997		10	0.9998
40	0.1	1.0002	40	0.1	1.0001
	1	0.9994		1	0.9987
	10	0.9999		10	0.9999
50	0.1	1.0000	50	0.1	1.0000
	1	1.0000		1	1.0000
	10	1.0000		10	1.0000

$\underline{t}=(t_1, t_2, \cdots, t_d)$ 와 $\underline{u}=(u_1, u_2, \cdots, u_d)$ 사이의 공분산은

$$v(\underline{t}, \underline{u}) = \sigma^2 \; exp\{-\theta \sum_{j=1}^{d} |t_j - u_j|^p\}, \; 0 < \theta, \; 0 < p \leq 2 \tag{6.2}$$

이다.

여기서 σ^2은 크기를 결정하는 분산 모수이고 θ와 p는 공분산을 특성짓는 모수인데, 큰 θ는 가까운 반응치들간에 작은 상관관계를 반영하고 작은 θ는 큰 상관관계를 반영한다. 만일 실험점 간에 거리가 가까우면 V/σ^2의 각 원소는 1에 가까운 값을 갖게되고 멀리 떨어져 있으면 0에 가까운 값을 가지게 되어 보통의 상관행렬과 동일한 성격을 지니게 된다. 이 공분산과 전산실험의 계획과 분석에 관한 자세한 논의는 Sacks, Welch, Mitchell and Wynn(1989)를 참조하기 바란다.

(6.1)에서 $log|V_M|$은 한 MLhd에서 계산된 것이며, $log|V_R|$은 100개의 RLhd에서 계산되어, 이 100개의 상대효율의 평균 값이 실제로 비교에 이용됐다[표6.1]. σ^2=1인 경우에는 상대효율은 $\log|V_R|/\log|V_M|$ 이 되어 행열식의 상대적인 크기가 된다. [표 6.1]은 상관행렬의 $\theta = 1$, $p = 2$ 인 경우, 즉 (6.2)가 $exp(-\sum_{j=1}^{d} |t_j - u_j|^2)$ 라고 가정하고 모의실험을 해 본 결과이다.

결과를 요약하면 다음과 같다. 상대효율은 분산 모수인 σ^2에 따라 차이를 보이지만 결국 n과 d가 증가하면 두 실험계획이 거의 같아짐을 알 수 있다. $p = 1$ 인 경우에도 역시 n과 d가 증가하면 같은 결과를 보이므로 그 결과를 생략하였다. 그러나 $p = 1$ 인 경우가 $p = 2$ 인 경우보다 효율이 약간씩 감소한 결과를 주었음을 밝혀둔다.

7. 요약 및 토의

지금까지 MLhd 와 RLhd 의 단순거리계산 및 $\bar{Y}$ 의 기대값 차이, 정보량 비교를 이용한 차이 결과를 제시하였다. 특히 거리의 기대값 계산을 위해 사용된 네 가지 다른 방법은 대학원 교육에서 유용한 예제로서 활용될 수 있다고 본다. 본 논문에서 사용한 실험계획의 여러 가지 비교 기법은 구조가 같지만 랜덤성에서 다른 두 개의 실험계획의 비교에 유용하다. 그리고 엔트로피의 비율인 상대효율은 실험계획의 일반적 비교에 사용 가능하다.

참고문헌

[1] 김남현 (1999). The limits of bivariate Q-Q plots based on matching that minimizes a distance. <한국통계학회논문집>, **6**, 645-658.

[2] 임미정, 권우주, 이주호 (1995). 이단계 Latin Hypercube 추출법과 그 응용. <응용통계연구>, **8**, 99-108.

[3] Bondy, J.A. and Murty, U.S.R. (1976). *Graph Theory with Applications*, American Elsevier Pub. Co., New York.

[4] Easton, G.S. and McCulloch, R.E. (1990). A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, **85**, 376-386.

[5] Ekberg, C., Borjesson, S., Emren. A.T. and Samuelsson, A. (2000). MINVAR and UNCCON, computer programs for uncertainty analysis of solubility calculations in geological systems. *Computers and Geosciences*, **26**, 219-226.

[6] Frieze, A. and Janson, S. (1995). Perfect matchings in random s-uniform hypercube. *Random Structure and Algorithms*, **7**, 41-57.

[7] Lahkim, M.B., Garcia, L.A. and Nuckols, J.R. (1999). Stochastic modeling of exposure and risk in a contaminated heterogeneous aquifer. 2: Application of Latin hypercube sampling. *Environmental Engineering Science*, **16**, 329-343.

[8] Lee, J. (1999). Asymptotic comparison of Latin hypercube sampling and its stratified version. *Journal of the Korean Statistical Society*, **28**, 135-150.

[9] McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239-245.

[10] Mitchell, T., Sacks, J. and Ylvisaker, D. (1994). Asymptotic Bayes criteria for nonparametric response surface design. *Ann. Statist.*, **22**, 634-651.

[11] Morris, M. and Mitchell, T. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381-402.

[12] Owen, A.B. (1992). A central limit theorem for Latin hypercube sampling, *Journal of the Royal Statistical Society, B*, **54**, 541-551.

[13] Park, J.S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, **39**, 95-111.

[14] Portielje, R., Hvitved-Jacobsen, T. and Schaarup-Jensen, K. (2000). Risk analysis using stochastic reliability methods applied to two cases of deterministic water quality models. *Water Research*, **34**, 153-170.

[15] Ripley, B.D. (1987). *Stochastic Simulation.*, John Wiley & Sons, New york.

[16] Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989). Designs and analysis of computer experiments (with discussion), *Statistical Science*, **4**, 409-435.

[17] Stein, M. (1987). Large sample properties of simulation using Latin hypercube sampling. *Technometrics*, **29**, 143-151.

[18] Talagrand, M. (1992). Matching random samples in many dimensions. *Annals of Applied Probability*, **2**, 846-856.

[19] Tang, B. (1993). OA-Based Latin hypercubes. *Journal of the American Statistical Association*, **88**, 1392-1397.

[1999년 6월 접수, 2000년 7월 채택]

Comparison and Distance Calculation Between Two Latin-hypercube Designs*

Jeong-Soo Park [1] Hyun-Sik Hwang [2]

ABSTRACT

A distance measure between two Latin-hypercube designs is defined and its expected value is computed. It was computed by using mathematical statistics, numerical analysis (multidimensional numerical integration), Monte-carlo method, and the theory of asymptotic normal distribution. For the comparison of two Latin-hypercube designs with same structure but different randomness, the difference of expected values of response function and information mass of experimental designs are considered. These methods may be useful in comparison between two general experimental designs.

Keywords: Computer experiments; Distance between designs; Monte-carlo method; Numerical integration; Relative efficiency.

* This study was financially supported by Chonnam National University in the program, 1999.

1) Associate Professor, Information and Telecommunication Research Institute, and Department of Statistics, Chonnam National University. E-mail: jspark@chonnam.ac.kr

2) Lecturer, Department of Statistics, Chonnam National University.