Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

격자기반(Lattice-based) 라틴 하이퍼큐브(Latin hypercube) 계획의 제안

황현식[1], 박정수[2]

요 약

라틴 하이퍼큐브 실험계획은 전산실험을 위하여 Mckay, Beckman과 Conover(1979)에 의해 고안된 방법으로 실험을 한번 시행하는데 많은 시간과 비용이 들거나 인자가 많은 실험에 효율적으로 사용할 수 있다. 하지만 이 실험계획 역시 실험영역 전체에서 골고루 배치되지 않을 가능성이 있으므로 이를 보완하려는 시도가 이루어져 왔으며, 여기서는 good lattice points(glp)와 계통추출을 응용하여 격자기반(lattice-based) Lhd의 두 가지 방법을 제안하였다. 모의실험 결과 glp 실험계획을 응용한 "방법 1"은 모형을 가정한 엔트로피에 기초한 최적 기준으로 검토한 경우 우수하였다. "방법 2"는 표본조사에 널리 쓰이는 계통추출을 응용하였으며 입력변수가 각기 다른 9개의 실험함수에 관하여 표본 평균의 추정치와 분산, MSE를 비교한 결과, 다른 실험계획들보다 우수하였다. 이 결과는 실험점이 실험영역 전체에서 골고루 퍼져서 나타난 것으로 보이며, 향후 전산실험계획에서의 응용을 기대할 수 있다.

주요용어 : Latin-hypercube design, Lattice-based Lhd, Good lattice points.

1. 서론

McKay, Beckman과 Conover(1979)에 의하여 고안된 라틴 하이퍼큐브 계획(Latin hypercube Design, 이하 Lhd라고 씀)은 Stein(1987)에 의하여 전산 실험 계획에 매우 유용하고 효율적인 것으로 밝혀졌다. 이 계획은 반복이 없이 모든 실험 영역을 고려할 수 있을 뿐만 아니라 많은 비용과 시간이 드는 실험을 비교적 적게 할 수 있는 장점이 있다. 그러나 이 실험계획 역시 여러가지 좋은 특성에도 불구하고, 실험할 수 있는 경우의 수가 많기 때문에 그 중에서도 최적인 Lhd를 찾는 작업이 계속되고 있다.

대표적인 방법으로 Sacks, Welch, Mitchell과 Wynn(1989)이 제시한 예측치의 평균 제곱 오차를 적분한 값(Integrated Mean Squared Error of Prediction, 이하 IMSE라고 씀)을 최소화하는 실험계획과 Shewry와 Wynn(1987)이 제시한 엔트로피(Entropy)를 최대화하는 계획 등이 있다.

본문에서는 편의상 표본조사에 널리 쓰이는 계통추출(systematic sampling) 방법이라고도 할 수 있으며, 또한 함수의 적분 계산에서 유사 몬테카를로(Quasi- Monte Carlo) 기법으로도 잘 알려진 격자(Lattice) 방법을 응용하여 Lhd에 적용한 격자기반(Lattice-based) Lhd를 제안하고자 한다. 격자방법 자체는 숫자이론(Number Theory)에 기초를 두었으며, 표본공간 전체에서 골고루 표본을 추출할 목적으로 실험영역 전체를 규칙적으로 배열한 것을 말한다. 격자를 Lhd에 적용한다면 Lhd의 장점을 가지면서 표본공간 전체를 커버하는 표본점들로 표본을 추출할 수 있다고 보여진다. 그 방법과 모의실험 결과를 실었다.

2. 라틴 하이퍼큐브 실험계획(Lhd)

1) 국가전문행정연수원 통계연수부 전임교수, 대전 유성구 가정동 33 hshwang@nso.go.kr
2) 전남대학교 자연과학대학 통계학과 부교수, 광주 북구 용봉동 300

Lhd는 표본공간 전체에서 표본이 추출되도록 각 확률 변수의 범위를 n 개의 범위로 나눈 다음 각 구간에서 하나씩 추출하되 중복되지 않게 n 개를 뽑는 방법이다. Lhd의 가정과 실제 모형은 다음과 같다.

실험 영역 $(0,1)^d$에서 각각이 독립인 입력 변수($X_1,\cdots,X_d$)에 관하여 그에 해당하는 분포 함수 ($F_1,\cdots,F_d$)가 있고, x_{ij}는 j번째 변수(X_j)에 i번째 관측치 ($i=1,\cdots,nj=1,\cdots,d$)를 나타낸다고 하자. 여기서 $P=(P_{ij})$는 $n\times d$ 행렬이고, P의 각각의 열은 ($1,\cdots,n$)의 확률적으로 독립인 순열(permutation)을 나타낸다. r_{ij}는 $n\times d$ 행렬에서 가질 수 있는 값으로 각각이 독립적으로 균등분포[0, 1]를 따른다고 가정하면, Lhd에서 가능한 실험점 x_{ij}는 다음과 같이 정의할 수 있다.

$$x_{ij}=F_j^{-1}(\frac{1}{n}(P_{ij}-r_{ij})). \qquad \text{<식 1.1>}$$

<식 1.1>을 보면 순열 행렬 $P_{i1},\cdots,P_{id}$가 결정되고 나서 그 실험 구간에 바로 x_{ij} 가 위치하게 되고, 정확한 실험점은 $r_{i1},\cdots,r_{id}$가 결정된 후 찾을 수 있다. 이 실험점에 최종적으로 분포함수의 역을 취하면 그 분포를 따르는 실험점을 얻게 된다. 모든 알려지지 않은 분포를 균등분포(0, 1)라고 가정한다면, 특수한 경우로서 역함수를 취한 값이 실제로는 취하지 않는 것과 동일하게 된다.

Lhd 중에서 모든 i와 j에 관하여 $r_{ij}=1/2$이면 중심 Lhd(Midpoint Lhd, MLhd)라 부른다 (Park(1994)). 또한 $r_{ij}\sim U(0,1)$을 따른다고 가정하면 이를 랜덤 Lhd(Random Lhd, RLhd)라고 한다(박정수, 황현식(2000)).

3. glp(good lattice points)

glp는 Niederreiter(1992)에 의하여 체계적으로 발전하였으며, 여기서는 Fang과 Wang(1994)에 의하여 정리된 glp를 간략히 기술한다. glp는 실험공간에 실험점이 잘 퍼지도록 설계되어졌으며, 실용적이며 계산이 용이하여 자주 사용된다. 구체적으로 다음과 같은 구현 알고리즘을 갖는다.

< glp의 알고리즘>

1. 가능한 실험점의 개수 n과 입력변수의 수를 결정한다.
2. $h_1,\cdots,h_d$의 정수로 이루어진 벡터를 구성한다. 여기서 h_1=1, $0<h_i<n$.
3. 다음 생성자에 의해 새로운 숫자 q_{ki}를 구성한다.

$$q_{ki}=kh_i \ (\mathrm{mod}\ n),\ k=1,\cdots,n\ ,\ i=1,\cdots,d. \qquad \text{<식 3.1>}$$

4. 최종적으로 실험점 x_{ki}를 다음 식으로 결정한다.

$$x_{ki}=\frac{1}{n}(q_{ki}-\frac{1}{2}).$$

여기서 $h_1,\cdots,h_d$를 생성 벡터(generating vector)라고 하며, 이들이 결정되면 x_{ki} 들이 자동적으로 만들어진다. h_i에 의해 결정되어지는 실험점의 집합을 격자실험집합(lattice point set)이라고 하며, 특히 모든 가능한 생성벡터 중에서 최소 불일치(smallest discrepancy) 집합을 glp라고 한다. 따라서 glp를 적용하기 위해서는 실험점의 개수(n)과 입력변수의 수(d)에 따라 가능한 h_i를 찾아주어야 한다. h_i를 효율적으로 찾기 위하여 소수(prime number), 원근(primitive

roots), 제곱근배열(square sequence) 등을 사용하는 방법이 제안되어 왔다. 만일 소수를 쓰는 경우에 <식 3.1>은 항상 나머지가 0이 되지 않는 특성으로, 실험점은 1과 n 사이에서 중복되지 않게 나타난다. 하지만 모든 n과 d에서 존재하는 것은 아니며, Fang과 Wang(1992)은 가능한 n과 d에서 glp를 찾아서 사용하기 쉽도록 표로 정리한 바 있다. 랜덤한 실험계획이 아니고 결정적인 실험계획이라는 점에서 통계적으로 바람직하지 않다고 보여진다.

4. 격자 기반(Lattice-based) Lhd(LbLhd)

glp의 단점을 보완하는 연구로 glp의 성격을 가지면서 최종적으로 생성되는 실험점은 Lhd가 되도록 하는 격자기반(Lattice-based) Lhd(이하 LbLhd라고 씀)를 제안한다. 본 소고에서는 두 가지 방법을 제안하였는데 그 중에 "방법 1"이 glp의 성격을 가지면서 Lhd가 되는 방법이다. 이와 유사하지만, 매 k번째 표본을 조사하는 계통추출의 방법을 응용한 Lhd를 생성하는 것이 "방법 2"이다.

먼저 "방법 1"에 대한 LbLhd의 생성 방법이다. glp와 다른 점은 모든 n에서 가능하며 랜덤하다는 점이다. 또한 MLhd로 생성된 실험점을 랜덤하게 옮겨줌으로서 RLhd의 성격을 가진다는 점이 큰 차이점이다. 하지만 glp의 제약인 $n>d$인 경우에만 구성되는 것은 같다. 현실적으로 충분히 가능한 가정이라고 보여진다.

<"방법 1"의 알고리즘 : glp에 기초한 LbLhd>

(1) 실험점의 개수 n을 결정한다. 이 때 n에 대한 제약조건은 없다.

(2) $h_1, \cdots, h_d$의 정수로 이루어진 벡터를 랜덤하게 구성한다. 여기서 h_1=1로 고정하며 각 h_i 는 $0<h_i<n$로 생성하되 각 경우에 다음 규칙에 따른다.

(2.1) 만약 n이 소수(prime number)이면 h_1=1을 제외하고 h_i를 랜덤하게 생성한다.

(2.2) 만약 n이 "소수-1"이면, n+1이라고 가정하고 (2.1)과 동일하게 생성한다.

(2.3) 만약 (2.1)과 (2.2)의 조건을 만족하지 않는 n의 경우에 h_1=1 이외에 나머지 h_i 들은 n의 공약수들을 제외하고 생성한다.

(3) 다음 생성자에 의해 새로운 숫자 q_{ki}를 구성한다.

$$q_{ki} = kh_i \ (\mathrm{mod}\ n),\ \ k=1,\cdots,n\ ,\ i=1,\cdots,d$$

(4) 최종적으로 실험점 x_{ki} 는 $r_{ki} \sim U(0,1)$라고 가정하였을 때, Lhd 실험계획 <식 1.1>에 의거하여 다음 식으로 결정한다.

$$x_{ki} = \frac{1}{n}(q_{ki} - r_{ki}).$$

위 알고리즘에서 벡터인 h_i를 생성할 때 고려한 h_1=1로 고정한 이유는 입력변수의 수가 많으면 문제가 없지만 작은 경우에서는 같은 시작점이면 항상 같은 실험점들의 계통이 정해지므로 적어도 하나는 1로 고정시키려는 이유에서다. glp에서도 항상 첫 입력변수의 시작점은 항상 1로 고정되어있다. (2.3)의 공약수 제외 제약조건은 Lhd를 생성하려면 n의 모든 숫자가 출현해야 하는데, 모든 n에 대하여 현실적으로 알고리즘을 구성할 수 없었기 때문이다. 예를 들어, 이 제약 조건에 의하면 n이 21인 경우에 3의 배수, 7의 배수가 제외된다.

"방법 1"은 h_i를 생성한 후 시작점으로 결정하고 그로부터 h_i를 각각 더해주는 방법이다. 즉 h_i 자체가 시작점인 동시에 더해지는 수이다. 그렇기 때문에 초기점에 공약수를 제외하는 제약이 가해지는 순간에 해당되는 실험점은 항상 제외되게 되어있다. "방법 2"는 "방법 1"과 다르게 초기점 h_i 이외에 계통의 크기라고 할 수 있는 또 하나의 k_i 벡터를 랜덤하게 생성하여 LbLhd를 구현한다. 이 경우 초기치 h_i는 제약조건이 전혀 없이 n보다 작게 랜덤하게 생성되며

k_i의 경우에만 공약수 제외 제약을 두게 된다. 다음은 "방법 2"에 의한 LbLhd 생성 방법이다.

<"방법 2" : 계통추출법을 응용한 LbLhd 알고리즘>

1) 실험점의 개수 n을 결정한다. 어떤 숫자 n이라도 가능하다.
2) $h_1, \cdots, h_d$ 의 정수로 이루어진 벡터를 구성한다. 여기서 $1 \le h_i \le n$이며 h_i는 랜덤하게 생성한다.
3) $k_1, \cdots, k_d$의 정수로 이루어진 벡터를 구성한다. 여기서 $1 \le k_i < n$이며, 각 k_i는 n의 공약수들을 제외하고 랜덤하게 생성한다.
4) 다음 생성자에 의해 새로운 숫자 q_{ji}를 구성한다.

$$q_{ji} = ((j-1)k_i + h_i) \pmod{n}, \quad j = 1, \cdots, n, \; i = 1, \cdots, d$$

5) 최종적으로 실험점 x_{ji}를 다음 식으로 결정한다. $r_{ji} \sim U(0,1)$라 하면

$$x_{ji} = \frac{1}{n}(q_{ji} - r_{ji}).$$

이다.

"방법 2"에 의한 LbLhd는 모집단에서 매 k번째 표본을 조사하는 계통추출법과 유사하다. 다만 입력변수마다 k_i를 생성하는 점이 다르다. 표본조사에서 계통추출법은 모집단이 주기성이 있는 경우에는 추정치의 분산이 SRS의 분산보다 좋지 않다고 알려져 있다. 하지만 여기서의 응용은 공간상에 실험점이 넓게 퍼지도록 하는데 역점을 두었으므로, 주기성의 문제는 고려하지 않아도 적용이 가능하리라고 본다.

5. 모의실험 결과

첫 번째는 엔트로피를 이용한 최적계획 기준으로 4가지 실험계획을 비교하였다. "방법 1"에 의해 생성한 LbLhd를 LbLhd_1, "방법 2"는 LbLhd_2 그리고 MLhd, RLhd이다. n과 d는 고정하여 100개의 실험계획을 각각 생성한 다음, Mitchell, Sacks와 Ylvisaker(1994)가 제시한 방법으로 다음 상대효율을 구하였다. 예를 들어 MLhd와 RLhd의 상대효율(relative efficiency)은 다음과 같다.

$$\text{상대효율} = \frac{-n\log\sigma^2 + \log|V_{RLhd}|}{-n\log\sigma^2 + \log|V_{MLhd}|}$$

두 번째 모의실험은 임의의 함수

$$y(\underline{t}_j), \quad \underline{t}_j\text{는 } d \text{ 차원의 입력변수}, \quad j = 1, \cdots, n$$

의 평균을 잘 추정하는지를 점검하였다. 이를 목적으로 주로 다른 연구에 사용되어진 9개의 실험함수(test function)에 의한 표본평균의 MSE를 비교하였다. 입력변수의 수는 실험함수에 따라 다르며, RLhd와 MLhd 그리고 제안된 LbLhd의 두 가지 방법에 대해서 각각 100개의 실험계획을 생성하였다.

한 개의 실험계획에서 추정한 각 함수 $y(\underline{t}_j)$의 평균을 $\bar{y}$ 라고 하자. 즉 각 실험점에서의 결과값을 y_i, $i = 1, \cdots, n$ 할 때,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

이다. 또한 100개의 실험계획을 생성하여 추정된 $\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_{100}$의 표본평균을 $\bar{\bar{y}}$ 라 하자. 즉,

$$\bar{\bar{y}} = \frac{1}{100}\sum_{k=1}^{100} \bar{y}_k$$

이다. $\bar{y}$의 분산의 추정치는 다음과 같다.

$$\widehat{var(\bar{y})} = \frac{1}{100}\sum_{k=1}^{100}(\bar{y}_k - \bar{\bar{y}})^2.$$

그리고 각 실험계획에 의한 추정치와의 차이를 알아보기 위하여 $\bar{y}$의 MSE (Mean Squared Error)를 구하여 그 값을 비교하였다.

$$MSE(\bar{y}) = \frac{1}{100}\sum_{k=1}^{100}(\bar{y}_k - \mu)^2.$$

$MSE(\bar{y})$를 구하려면 참값(μ)이 있어야하는데, 기대함수의 적분이 가능하여 참값을 해석학적으로 구할 수 있는 경우외의 함수는 적분이 곤란하여 몬테카를로 기법에 의한 추정치를 참값으로 보고 $\widehat{MSE(\bar{y})}$를 구하였다. 이 경우 10000개의 난수를 발생하여 그 값을 다음과 같이 계산하였다.

$$\frac{1}{10000}\sum_{i=1}^{10000} y(\underline{x}_j^{(i)}),\ j=1,\cdots,n.$$

여기서 $\underline{x}_j^{(i)}$는 10000개의 입력값 $\underline{x}_j$의 i번째 값을 의미한다. μ를 몬테카를로 방법으로 추정한 측정값을 $\hat{\mu}$ 로 표기하자. $\hat{\mu}$ 에 의해 계산하는 $MSE(\bar{y})$의 추정치는 분산과 편의(bias)제곱으로 나누어진다. 그 의미는 실험계획이 함수의 평균을 얼마나 잘 추정하는지의 측도이다.

$$\widehat{MSE(\bar{y})} = \frac{1}{100}\sum_{i=1}^{100}(\bar{y}_i - \bar{\bar{y}})^2 + (\bar{\bar{y}} - \hat{\mu})^2.$$

불편추정량이면 편의제곱이 0이 되며, $\widehat{MSE(\bar{y})}$ 값이 작다면 추정 능력이 더 좋은 실험계획이다. 결국 본 장에서 보이고자 하는 것은 $\bar{y}$의 추정치를 구해서 실험계획 간에 차이를 알아보고자 하였다. 비교대상 중 추정 능력이 좋은 실험계획은 편의(bias)가 작고 $\widehat{MSE(\bar{y})}$ 값 역시 작아야 한다.

모의실험 결과를 요약하면 "방법 1"의 경우 실험점의 수가 '소수'인 경우나 '소수-1' 인 경우에는 RLhd 보다 더 작은 분산과 MSE를 보였다. 하지만 그 외의 실험점의 개수에서는 편의와 분산이 커서 RLhd보다 좋지 못하며, 특히 다른 실험계획에 비하여 많은 차이를 보임에 따라 개선 방법이 요구된다.

"방법 2"의 경우에는 거의 모든 실험함수에서, 실험점의 개수에 상관없이 편의와 분산이 낮아서 RLhd보다 효율적인 실험계획이다. 실험점이 골고루 퍼진 결과로, 계통추출의 성격이 그대로 반영되었다. "방법 2"는 증명하지는 않았지만 불편추정량으로 보인다. 그러나 입력변수가 아주 작은 경우에서는 RLhd나 MLhd에 비하여 "방법 2"가 분산과 MSE가 더 큰 것으로 나타났다. 하지만 이 경우에도 실험점의 개수가 증가하는 경우에는 더 낮게 나타났다. 실험함수의 평균추정 결과로 볼 때, LbLhd의 "방법 2"는 입력변수(d)가 많은 경우의 전산실험계획에도 매우 효율적으로 적용할 수 있다고 보여진다.

6. 결론 및 토의

제안된 LbLhd의 장점으로는 다음과 같은 점을 들 수 있다.

첫째로 구성하기 용이하다는 점이다. 계통추출의 방법을 응용하였으므로 매 k번째 실험점을 결정하는 방법만 있으면 가능하다. 직교기반 Lhd, 층화 Lhd, 이단계 Lhd 등 Lhd의 개선된 방법들에 비하여 비교적 구성이 편한 LbLhd가 쉽게 응용 가능하게 보인다.

둘째, 엔트로피에 기초한 상대효율 기준에서 "방법 1"은 RLhd보다 우수하다. 하지만 여기서의 최적기준은 특정한 모형을 가정하였을 때만 의미있는 기준으로 실험 함수들에 관하여서는 다른 결과를 얻었다. 결과에 의하면 "방법 2"에 의한 계통추출 응용방법이 타 실험계획에 비하

여 분산과 MSE의 추정치가 더 낮았다. 실험함수에 따라 조금씩 다르지만, RLhd나 MLhd, "방법 1"보다 현저하게 낮은 것으로 미루어 "방법 2"에 의한 LbLhd가 더 효율적인 실험계획이다.

셋째, 무엇보다 실험구간의 전 영역에서 골고루 추출된다는 점이다. 이는 glp나 계통추출의 효과를 어느 정도 반영한 결과로 보인다. 특히 전산실험계획에서 실험점의 개수인 n이 작으면서, 입력 변수의 수인 d가 큰 경우에 유용할 것으로 여겨진다. 또한 전산실험계획 이외에도 적분이나 최적화(optimization) 문제에 적용 가능하다고 본다.

LbLhd의 단점으로 "방법 1"은 $n>d$인 경우에만 구성되나, 현실적으로 가능한 가정이다. 하지만 "방법 2"는 제한이 없다. 또한 "방법 1"은 h_i 벡터 생성시에 n이 소수와 소수-1인 경우에는 문제가 없으나, 그 외의 n에서는 공약수를 제외한 상태에서 랜덤한 숫자를 요구한다. "방법 2" 역시 h_i에 더해지는 수인 k_i 벡터 생성시 이러한 제약 조건이 있다. 예를 들면 n이 20인 경우에는 2의 배수, 4의 배수, 10의 배수 등이 제외된다. Lhd를 생성하기 위한 제약 조건이지만, 앞으로 이를 보완하는 더 좋은 방법들이 나오기를 기대한다.

참고문헌

[1] 박정수, 황현식 (2000). 라틴-하이퍼큐브 실험계획 간의 거리 계산과 비교. <응용통계연구>, **13**, 477-488.

[2] Fang, K, T. and Wang, Y. (1994). *Number-theoretic Methods in Statistics.* Chapman&Hall, London.

[3] 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.

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

[5] Morris, D., Mitchell, T. and Ylvisaker, D. (1993). Bayesian design and Analysis of computer experiments : use of derivatives in surface prediction. *Technometrics,* **35,** 243-255.

[6] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods.* Soc. Industr. Appl. Math. (SIAM), Philadelphia.

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

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

[9] Shewry, M. and Wynn, H. (1987). Maximum entropy sampling. *Journal of Applied Statistics,* **14**, 409-435.

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