근사직교배열의 직교성을 평가하기 위한 측도로서의 상호정보

장대흥†

부경대학교 수리과학부 통계학전공

Mutual Information as a Criterion for Evaluating the Degree of the Orthogonality of Nearly Orthogonal Arrays

Jang, Dae-Heung†

Division of Mathematical Sciences, Pukyong National University

Key Words : Nearly Orthogonal Arrays, Evaluation of the Orthogonality, Mutual Information

Abstract

The orthogonality is an important property in the experimental designs. When we use nearly orthogonal arrays(for example, supersaturated designs), we need evaluate the degree of the orthogonality of given nearly orthogonal arrays. We can use the mutual information as a new criterion for evaluating and testing the degree of the orthogonality of given nearly orthogonal arrays.

1. 서 론

요인배치법을 완전하게 쓸 수 없을 때 요인배치법의 일부실시법(fractional factorial design), 포화계획(saturated design) 또는 초포화계획(supersaturated design)을 사용한다. Booth와 Cox(1962) 이후 많은 학자들이 이러한 (초)포화계획에 대하여 연구하여 오고 있다(예로, 2008년도에 발표되었거나 추후 발표할 초포화계획에 관한 논문으로서 Bulutoglu와 Ryan(2008), Butler(2008), Chatterjee외 2인(2008), Chen과 Liu (2008a, b), Das외 3인(2008), Fang외 2인(2008), Jones 외 2인(2008), Jones외 3인(2008), Koukouvinos와 Mylona(2008), Koukouvinos외 2인(2008a, b), Nguyen과 Cheng(2008), Siomina와 Ahlinder(2008) 등이 있다.). 통계적 품질관리나 실험계획법에서 요인의 수가 과다하게 많은 경우 주로 직교배열을 이용하여 실험을 한다. Plackett와 Burman(1946)이 포화계획(saturated design)인 Plackett-Burman design을 제안한 이후 직교배열이 실험계획분야에서 다양하게 쓰이고 있다. 직교배열은 다구찌방법의 중요한 수단으로 쓰이고 있다. 그러나, 직교계획을 쓰지 못할 때 우리는 근사직교배열(nearly orthogonal arrays)을 이용하게 되는 데 Wang과 Wu(1992)는 이러한 근사직교배열을 제안하였다. 우리가 실험계획으로서 직교배열이 아닌 근사직교배열을 쓰는 경우 이 근사직교배열의 직교성의 정도를 아는 것이 중요하다. 요즈음 실험계획에서 중요한 연구 테마 중 하나인 초포화계획도 직교배열이 되지 못하므로 이러한 초포화계획에 대하여 직교성의 정도를 평가하는 측도의 개발은 매우 중요하다. Ma외 2인(2000)는 이러한 근사직교배열의 직교성의 정도를 평가하기 위한 기준을 이용하여 근사직교배열을 구하는 방법들을 제시하였다. Jang(2002), 장대흥(2004)은 기존의 직교성 판정기준을 기반으로 하여 근사직교배열의 직교성의 정도를 평가하기 위한 측도들과 그래픽방법들을 제안하였다.

근사직교배열의 직교성의 정도를 평가할 수 있는 기

† 교신저자 doc1jjang@yahoo.co.kr

※ 본 연구는 한국학술진흥재단의 2005년도 기초과학연구(KRF-2005-015-C00082)지원으로 수행되었음.

존의 기준들 외에 또 다른 기준으로서 우리는 Shannon의 정보이론에 기초한 상호정보를 이용할 수 있다. 상호정보는 공학, 특히 화상인식 분야에서 주로 많이 이용하였으나 최근에는 통계학에서도 상호정보를 이용한 논문들이 나타나고 있다. 본 논문에서 제시하는 방법은 기존의 방법들과는 접근방법이 다르다. 이 상호정보를 이용하면 근사직교배열의 직교성의 정도를 평가할 수 있을 뿐만이 아니라 검정도 행할 수 있다. 본 논문의 내용은 상호정보를 실험계획분야에 적용하는 최초의 시도가 될 것이다. 2절에서는 기존의 직교성의 정도를 평가하는 기준들을 열거하고 3절에서 새로운 기준으로서 상호정보를 사용하는 방법에 대하여 서술하였다. 4절에 수치 예를 보이고 5절에서 결론을 내렸다.

2. 직교성의 정도를 평가하는 기준들

우리는 강도(strenth) 2인 직교배열을 직교계획(orthogonal design)이라 부른다. r 제약을 갖으며 강도가 t이고, 크기가 N인 직교배열이란, 각 열이 2개 이상의 수준을 갖는 $N\times r$ 행렬 A에서 A의 $N\times t$ 부분행렬 각각이 모든 가능한 $1\times t$ 행벡터들을 같은 빈도로 갖는 행렬을 말한다. $L_N(q_1^{m_1}\times q_2^{m_2}\times\ldots\times q_s^{m_s})$, $\sum_{i=1}^{s} m_i = r$를 q_i개의 수준을 갖는 열들이 m_i개인 직교계획이라 하자. 직교계획의 정의는 다음과 같다.

1. (직교조건 1) 각 열에서 각 수준은 같은 빈도수로 나타난다.
2. (직교조건 2) 임의의 두 개의 열들에서 수준의 조합이 같은 빈도수로 나타난다.

종종 (조건 1)을 만족하는 계획을 U-형 계획이라 칭한다.

우리가 실험계획으로서 직교배열이 아닌 근사직교배열을 쓰는 경우 이 근사직교배열의 직교성의 정도를 아는 것이 중요하다. 논의의 간편성을 위하여 $N\times r$ 행렬인 배열 A는 U-형 계획이고, $A=(\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_r)$라 고 하자. 그리고 i번째 열의 구성요소가 $1,2,\ldots,q$로 이루어졌다고 가정하자.

근사직교배열 A의 임의의 두 개의 열 $\mathbf{c}_i$와 $\mathbf{c}_j$의 직교성을 평가하는 기준을 $d(\mathbf{c}_i,\mathbf{c}_j)$라 할 때 다음과 같은 기준들이 있다.

$d(\mathbf{c}_i,\mathbf{c}_j)=$

1. s_{ij}^2(두 개의 열 $\mathbf{c}_i$, $\mathbf{c}_j$의 내적의 제곱) (Booth and Cox(1962))
2. $\sum_{k=1}^{q}\sum_{l=1}^{q}\left|N_{\mathbf{c}_i\mathbf{c}_j}(k,l)-\frac{N}{q^2}\right|$ (Ma외 2인(2000))
3. $\sum_{k=1}^{q}\sum_{l=1}^{q}\left(N_{\mathbf{c}_i\mathbf{c}_j}(k,l)-\frac{N}{q^2}\right)^2$ (Fang외 2인(2003))
4. $\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{\left(N_{\mathbf{c}_i\mathbf{c}_j}(k,l)-\frac{N}{q^2}\right)^2}{\frac{N}{q^2}}$ (Yamada와 Lin(1999))
5. J_{ij}(두 개의 열 $\mathbf{c}_i$, $\mathbf{c}_j$에 대한 균등지수) (Jang(2002), 장대흥(2004))

여기서, $\mathbf{c}_i$, $\mathbf{c}_j$는 A의 임의의 두 열이고, $N_{\mathbf{c}_i\mathbf{c}_j}(k,l)$은 $\mathbf{c}_i$와 $\mathbf{c}_j$의 두 개의 열들에서 (k,l)쌍의 개수이고, $\frac{N}{q^2}$은 $\mathbf{c}_i$와 $\mathbf{c}_j$의 두 개의 열들에서 수준쌍들의 평균개수이다.

기준 1은 가장 오래되었지만 지금도 2-수준계에서는 쓰이는 기준이다. 두 열 벡터의 내적(inner product)를 이용하는 기준이다. 그러나 이 기준은 2-수준에서는 사용이 가능하나 3-수준 이상이거나 혼합수준이면 사용이 불가능하다. 그래서 3-수준 이상이거나 혼합수준일 때도 사용할 수 있는 기준이 필요하다. 그래서 제시된 것들이 기준 2-5이다. 기준 2와 기준 3의 차이는 L^1 norm을 사용하느냐 아니면 L^2 norm을 사용하느냐는 차이이다. 기준 4는 카이제곱 타입의 기준이다. 기준 5는 다양성지수 중 evenness index를 사용하여 제시된 기준이다.

기준 1-4는 $d(\mathbf{c}_i,\mathbf{c}_j)$가 0에 가까울수록 직교성을 만족하게 되고 기준 5는 $d(\mathbf{c}_i,\mathbf{c}_j)$가 1에 가까울수록 직교성을 만족하게 된다.

그러면 배열 A의 직교성의 정도를 평가할 수 있는 기준들로서 다음과 같이 두 가지 기준을 제시할 수 있다. 이러한 기준들은 앞에서 제시한 $d(\mathbf{c}_i,\mathbf{c}_j)$ 기준들을 언급한 논문들에서 제시되었던 측도들이다.

1. $ave(A)=\frac{1}{\binom{r}{2}}\sum_{1\le i<j\le r} d(\mathbf{c}_i,\mathbf{c}_j)$

$(p=1,2,3,4,5)$

2. $\max(A) = \max \sum_{1 \le i < j \le r} d(\mathbf{c}_i, \mathbf{c}_j)$ $(p=1,2,3,4)$

(기준 5는 max 대신 min으로 바꿈)

$d(\mathbf{c}_i, \mathbf{c}_j)$를 계산하면 $d(\mathbf{c}_i, \mathbf{c}_j)$을 구성요소로 갖는 직교화평가행렬(장대홍(2002) 참조)인 $r \times r$ 행렬 $D_p = (d(\mathbf{c}_i, \mathbf{c}_j))$ $(p=1,2,3,4,5)$를 정의할 수 있다. 여기서, p는 p번째 기준을 가리킨다. 이 행렬 D_p의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다.

우리는 앞에서 제시한 여러 가지 $d(\mathbf{c}_i, \mathbf{c}_j)$ 기준들을 서로 비교하여 보기 위해서는 표준화 작업이 필요하다. 가장 극단적인, 즉 가장 직교성에서 먼 경우의 $d(\mathbf{c}_i, \mathbf{c}_j)$값을 m_p라 하면 $I(i,j) = \frac{d(\mathbf{c}_i, \mathbf{c}_j)}{m_p}$를 구성요소로 갖는 직교화지수행렬(orthogonality index matrix)인 $r \times r$ 행렬 $D_p^{'} = (d(\mathbf{c}_i, \mathbf{c}_j))$를 우리는 제안할 수 있다. 이 행렬 $D_p^{'}$의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다. 기준 5의 경우는 직교하는 경우의 $d(\mathbf{c}_i, \mathbf{c}_j)$값이 m_p가 되고 이 값이 1이므로 직교화지수행렬이 직교화평가행렬과 같게 된다.

앞에서 정의한 5가지 기준들을 사용하여 구한 $I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도들로서 다음과 같은 두 가지의 측도들을 정의할 수 있다.

1. $ave(D_p^{'}) = \frac{1}{\binom{r}{2}} \sum_{1 \le i < j \le r} I(i,j)$ $(p=1,2,3,4,5)$

2. $\max(D_p^{'}) = \max \sum_{1 \le i < j \le r} I(i,j)$ $(p=1,2,3,4)$

(기준 5는 max 대신 min으로 바꿈)

기준 1-4는 $ave(D_p^{'})$나 $\max(D_p^{'})$가 0에 가까울수록 직교배열에 가깝게 되고 기준 5는 $ave(D_p^{'})$나 $\min(D_p^{'})$가 1에 가까울수록 직교배열에 가깝게 된다.

3. 직교성의 정도를 평가할 수 있는 기준으로서의 상호정보

하나의 확률변수를 X라 할 때 엔트로피 H는 X의 불확실성(uncertainty)의 정도를 나타내는 측도로 쓰인다. $H(Y|X)$를 X가 주어졌을 때 또 다른 확률변수 Y의 조건부엔트로피(conditional entropy)로 정의한다면 상호정보(mutual information) I_{XY}는 $I_{XY} = H(Y) - H(Y|X)$로 정의된다. 즉 두 개의 확률변수 X와 Y가 서로 나누어갖는 정보량을 측정하는 측도가 된다. I_{XY}는 다음과 같이 표현할 수가 있다(Brillinger와 Guha(2007) 참조).

1. (연속형)

$$I_{XY} = \iint f(x,y) \log \frac{f(x,y)}{f_1(x) f_2(y)} dxdy$$

2. (이산형)

$$I_{XY} = \sum_x \sum_y f(x,y) \log \frac{f(x,y)}{f_1(x) f_2(y)}$$

여기서, $f(x,y)$는 확률변수(또는 확률벡터) X와 Y의 결합확률(밀도)함수이고, $f_1(x)$는 확률변수(또는 확률벡터) X의 주변확률(밀도)함수이고, $f_2(y)$는 확률변수(또는 확률벡터) Y의 주변확률(밀도)함수이다.

I_{XY}를 다시 표현하면 결합확률(밀도)함수 $f(x,y)$와 주변확률(밀도)함수의 곱 $f_1(x) f_2(y)$ 사이의 Kull-back-Leibler divergence가 된다. $I_{XY} \ge 0$이 항상 성립한다(Dionisio외 2인(2004) 참조). 그리고 $I_{XY} = 0$이면 두 확률변수(확률벡터)는 서로 독립이 되고, $I_{XY} > 0$이면 두 확률변수(확률벡터)는 서로 종속이 된다. 두 변수가 범주형 변수일 때는 다음과 같이 상호정보의 추정량이 주어진다(Brillinger와 Guha(2007) 참조).

$$\widehat{I_{XY}} = \sum_k \sum_l \frac{n_{kl}}{n} log \frac{n_{kl} n}{n_{k+} n_{+l}} \tag{1}$$

여기서, n은 분할표 상의 총도수, n_{kl}는 범주형 변수 X에서 k번째 범주와 범주형 변수 Y에서 l번째 범주에 해당하는 칸(cell)의 도수, n_{k+}는 변수 X에서 k번째 범주의 도수, n_{+l}은 변수 Y에서 l번째 범주의 도수이다.

식 (1)을 이용하여 근사직교배열 A의 임의의 두 개의 열 $\mathbf{c}_i$와 $\mathbf{c}_j$의 직교성을 평가하는 기준으로서 우리는 다음과 같은 기준을 제시할 수 있다.

$$d(c_i,c_j)=\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{N_{c_i,c_j}(k,l)}{N}log_2\frac{N_{c_i,c_j}(k,l)N}{N_{c_i}(k)N_{c_j}(l)} \qquad (2)$$

여기서, N은 근사직교배열의 크기, 즉 행의 수이고, $N_{c_i,c_j}(k,l)$은 c_i와 c_j의 두 개의 열들에서 (k,l)쌍의 개수이고, $N_{c_i}(k)$는 c_i에서 k번째 구성요소의 개수, $N_{c_j}(l)$은 c_j에서 l번째 구성요소의 개수이다. 이 기준은 $d(c_i,c_j)$가 0에 가까울수록 직교성을 만족하게 된다.

식 (1)에서 식 (2)는 쉽게 유도할 수 있다. 배열 A의 두 개의 열 c_i와 c_j를 두 개의 변수로 본다면 이 두 개의 변수에 대한 분할표에 대하여 $\widehat{I_{XY}}$는 식 (1)과 같이 주어지고 $n=N$, $n_{kl}=N_{c_i,c_j}(k,l)$, $n_{k+}=N_{c_i}(k)$, $n_{+l}=N_{c_j}(l)$이 된다. 그러므로 $\widehat{I_{XY}}=d(c_i,c_j)$가 된다.

예를 들어 보자. 배열 A의 두 개의 열이 다음과 같다고 하자.

$$c_i=\begin{pmatrix}1\\1\\1\\1\\2\\2\\2\\2\\3\\3\\3\\3\\4\\4\\4\\4\end{pmatrix},\quad c_j=\begin{pmatrix}4\\2\\1\\3\\1\\3\\3\\4\\2\\4\\2\\1\\3\\1\\4\\2\end{pmatrix}$$

그러면 이 두 개의 열 c_i와 c_j를 두 개의 변수로 보고 이 두 개의 변수에 대한 분할표를 작성하면 다음 <표 1>과 같다.

<표 1> c_i와 c_j에 대한 분할표

c_i \ c_j	1	2	3	4	sum
1	1	1	1	1	4
2	1	0	2	1	4
3	1	2	0	1	4
4	1	1	1	1	4
sum	4	4	4	4	16

그러면 다음과 같은 관계가 성립한다.

$n=N=16,$
$n_{11}=N_{c_i,c_j}(1,1)=1, n_{12}=N_{c_i,c_j}(1,2)=1,$
$\ldots, n_{44}=N_{c_i,c_j}(4,4)=1,$
$n_{k+}=N_{c_i}(k)=4(k=1,2,3,4),$
$n_{+l}=N_{c_j}(l)=4(l=1,2,3,4)$

일반적으로 수준수가 q개라 하면 식 (2)의 $d(c_i,c_j)$에 대하여 다음과 같은 성질이 성립한다.

성질 1. 모든 $i,j(i,j=1,2,\ldots,r)$에 대하여

$$0\le d(c_i,c_j)\le \log_2 q$$

가 성립된다.

(증명) 가장 값이 작은 경우와 가장 값이 큰 경우를 생각하여 보자.

(1) 가장 값이 작은 경우: 두 개의 열이 직교하는 경우로서 다음 <표 2>와 같은 분할표가 만들어 진다.

<표 2> 두 열이 직교하는 경우의 $d(c_i,c_j)$ 계산을 위한 분할표(수준수: q)

c_i \ c_j	1	2	...	q	sum
1	N/q^2	N/q^2	...	N/q^2	N/q
2	N/q^2	N/q^2	...	N/q^2	N/q
...	...	...	...	...	...
q	N/q^2	N/q^2	...	N/q^2	N/q
sum	$\frac{N}{q}$	$\frac{N}{q}$	...	$\frac{N}{q}$	N

$$\therefore d(c_i,c_j)=\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{N_{c_i,c_j}(k,l)}{N}log_2\frac{N_{c_i,c_j}(k,l)N}{N_{c_i}(k)N_{c_j}(l)}$$

$$=\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{\frac{N}{q^2}}{N}log_2\frac{\frac{N}{q^2}\times N}{\frac{N}{q}\times\frac{N}{q}}=\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{1}{q^2}log_2 1=0$$

(2) 가장 값이 큰 경우: 가장 직교성에서 먼 경우로서 다음 <표 3>과 같은 분할표가 만들어 진다.

<표 3> 가장 직교성에서 먼 경우의 $d(c_i,c_j)$ 계산을 위한 분할표(수준수: q)

c_i \ c_j	1	2	...	q	sum
1	N/q	0	...	0	N/q
2	0	N/q	...	0	N/q
...	...	...	...	...	...
q	0	0	...	N/q	N/q
sum	$\frac{N}{q}$	$\frac{N}{q}$	...	$\frac{N}{q}$	N

$$\therefore d(c_i,c_j)=\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{N_{c_i,c_j}(k,l)}{N}log_2\frac{N_{c_i,c_j}(k,l)N}{N_{c_i}(k)N_{c_j}(l)}$$

$$=q\times\frac{\frac{N}{q}}{N}log_2\frac{\frac{N}{q}\times N}{\frac{N}{q}\times\frac{N}{q}}=\log_2 q$$

$\therefore$ 모든 $i,j(i,j=1,2,...,r)$에 대하여 $0\le d(c_i,c_j)\le \log_2 q$가 성립된다.

배열 A의 두 개의 열 c_i와 c_j를 두 개의 변수로 볼 때 우리는 다음과 같은 귀무가설과 대립가설을 새울 수 있다.

H_0: c_i와 c_j는 서로 독립이다[직교한다].
H_1: c_i와 c_j는 서로 독립이 아니다[직교하지 않는다].

그러면 우도비검정 통계량(likelihood ratio test statistic) G^2는 다음과 같이 표현된다.

$$G^2=2\sum_{k=1}^{q}\sum_{l=1}^{q}N_{c_i,c_j}(k,l)\log_2\frac{N_{c_i,c_j}(k,l)N}{N_{c_i}(k)N_{c_j}(l)}$$

이 G^2는 $N\to\infty$일 때 자유도가 $(q-1)^2$인 카이제곱분포를 따르게 된다. $d(c_i,c_j)=\frac{G^2}{2N}$이므로 $2N\times d(c_i,c_j)$는 $N\to\infty$일 때 자유도가 $(q-1)^2$인 카이제곱분포를 따르게 된다. 즉 우리는 $d(c_i,c_j)$의 값을 이용하여 근사직교배열 A의 임의의 두 개의 열 c_i와 c_j의 직교성을 평가할 수 있고 또한 가설검정도 행할 수 있게 된다. 이 점이 2절에서 열거하였던 기존의 직교판정 기준들과 다른 점이다. 물론 이 카이제곱검정은 배열의 크기 N이 수준수 q에 비하여 상대적으로 커야 한다는 단서가 붙는다는 단점이 있다.

$d(c_i,c_j)$를 계산하고 나면 $d(c_i,c_j)$을 구성요소로 갖는 직교화평가행렬인 $r\times r$ 행렬 $D_6=(d(c_i,c_j))$를 정의할 수 있다. 이 행렬 D_6의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다. 또한, 가장 극단적인, 즉 가장 직교성에서 먼 경우의 $d(c_i,c_j)$값을 m_6라 하면 $I(i,j)=\frac{d(c_i,c_j)}{m_6}$를 구성요소로 갖는 직교화지수행렬인 $r\times r$ 행렬 $D_6^{'}=(d(c_i,c_j))$를 우리는 제안할 수 있다. 이 행렬 $D_6^{'}$의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다.

식 (2)의 기준에 의하여 구한 $I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도들로서 다음과 같은 두 가지의 측도들을 정의할 수 있다.

1. $ave(D_6^{'})=\frac{1}{\binom{r}{2}}\sum_{1\le i<j\le r}I(i,j)$
2. $\max(D_6^{'})=\max\sum_{1\le i<j\le r}I(i,j)$

$ave(D_6^{'})$나 $\max(D_6^{'})$가 0에 가까울수록 직교배열에 가깝게 된다.

4. 수치 예

예제 1. 다음 <표 4>와 같은 배열에 대하여 고려하여 보자. 이 배열은 Fang외 3인(2000)이 제안한 균등계획(uniform design) $U_{16}(4^5)$이다.

각 열의 구성요소가 4개로 동일하고, 각 열에서 각 구성요소의 개수가 같으므로 이 배열은 U-계획이 된다. 그러나 이 배열은 직교배열이 아닌 근사직교배열이 된다.

이 배열에 대하여 빈도행렬(장대흥(2004) 참조)을 구하면 다음 <표 5>와 같다.

이 빈도행렬을 참조하여 식 (2)를 이용한 $d(c_i,c_j)$를 계산할 수 있다. 예로, <표 6>을 참조하고 식 (2)를 이용하여 $d(c_4,c_5)$을 계산할 수 있다. $d(c_4,c_5)=1$이 된다. <표 7>처럼 가장 극단적인, 즉 가장 직교성에서 먼 경우에는 $d(c_i,c_j)=2=m_6$가 된다. 그러므로 모든 $d(c_i,c_j)$에 대하여 $0\le d(c_i,c_j)\le 2$가 성립한다.

<표 4> 배열 A

$$A=\begin{pmatrix}1&1&4&3&3\\1&2&2&1&2\\1&3&1&4&4\\1&4&3&3&3\\2&1&1&2&3\\2&2&3&4&1\\2&3&3&2&4\\2&4&4&4&1\\3&1&2&4&4\\3&2&4&1&2\\3&3&2&3&1\\3&4&1&1&2\\4&1&3&1&2\\4&2&1&3&4\\4&3&4&2&1\\4&4&2&2&3\end{pmatrix}$$

<표 5> 배열 A에 대한 빈도행렬

$c_1\begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&0&2&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&0&2&1\\0&2&0&2\\2&0&1&1\\1&2&1&0\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\2&0&1&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$
$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_2\begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\2&0&1&1\\0&2&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\1&2&0&1\\2&0&0&2\\1&1&2&0\end{pmatrix}$
$\begin{pmatrix}1&1&1&1\\1&0&2&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_3\begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&1&2\\1&1&1&1\\1&1&1&1\\2&1&1&0\end{pmatrix}$
$\begin{pmatrix}1&0&2&1\\0&2&0&2\\2&0&1&1\\1&2&1&0\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\2&0&1&1\\0&2&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_4\begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}0&4&0&0\\1&0&2&1\\1&0&2&1\\2&0&0&2\end{pmatrix}$
$\begin{pmatrix}0&1&2&1\\2&0&1&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\1&2&0&1\\2&0&0&2\\1&1&2&0\end{pmatrix}$	$\begin{pmatrix}0&1&1&2\\1&1&1&1\\1&1&1&1\\2&1&1&0\end{pmatrix}$	$\begin{pmatrix}0&4&0&0\\1&0&2&1\\1&0&2&1\\2&0&0&2\end{pmatrix}$	$c_5\begin{pmatrix}4\\4\\4\\4\end{pmatrix}$

<표 6> $d(c_4,c_5)$ 계산을 위한 분할표

c_4 \ c_5	1 2 3 4	sum
1	0 4 0 0	4
2	1 0 2 1	4
3	1 0 2 1	4
4	2 0 0 2	4
sum	4 4 4 4	16

<표 7> 가장 직교성에서 먼 경우의 $d(c_i,c_j)$ 계산을 위한 분할표

c_i \ c_j	1 2 3 4	sum
1	4 0 0 0	4
2	0 4 0 0	4
3	0 0 4 0	4
4	0 0 0 4	4
sum	4 4 4 4	16

이렇게 구한 $d(c_i,c_j)$를 이용하여 직교화평가행렬을 구하면 다음과 같다.

$$D_6=\begin{pmatrix}-&0&0.25&0.625&0.375\\0&-&0&0.25&0.625\\0.25&0&-&0&0.25\\0.625&0.25&0&-&1\\0.375&0.625&0.25&1&-\end{pmatrix}$$

이 행렬을 통하여 이 배열은 직교배열이 아님을 알 수 있다. $d(c_4,c_5)$의 값 1이 제일 커 직교성에서 제일 멀고, 1열과 2열, 2열과 3열, 3열과 4열끼리는 직교함을 알 수 있다.

참고로 기준 2에서 기준 5까지를 이용한 직교화평가행렬은 다음과 같다.

$$D_2=\begin{pmatrix}-&0&4&10&6\\0&-&0&4&10\\4&0&-&0&4\\10&4&0&-&14\\6&10&4&14&-\end{pmatrix}$$

$$D_3=\begin{pmatrix}-&0&4&10&6\\0&-&0&4&10\\4&0&-&0&4\\10&4&0&-&20\\6&10&4&20&-\end{pmatrix}=D_4$$

$$D_5=\begin{pmatrix}-&0&0.9375&0.84375&0.90625\\0&-&0&0.9375&0.84375\\0.9375&0&-&0&0.9375\\0.84375&0.9375&0&-&0.75\\0.90625&0.84375&0.9375&1&-\end{pmatrix}$$

여러 가지 $d(c_i,c_j)$ 기준들을 서로 비교하여 보기 위해서는 표준화 작업이 필요하다. $I(i,j)$를 구한 후 직교화지수행렬 D_6'를 다음과 같이 구할 수 있다.

$$D_6'=\begin{pmatrix}-&0&0.125&0.3125&0.1875\\0&-&0&0.125&0.3125\\0.125&0&-&0&0.125\\0.3125&0.125&0&-&0.5\\0.1875&0.3125&0.25&0.5&-\end{pmatrix}$$

$I(4,5)$의 값 0.5가 제일 커 직교성에서 제일 멀고, 1열과 2열, 2열과 3열, 3열과 4열끼리는 직교함을 알 수 있다.

참고로 기준 2에서 기준 5까지를 이용한 직교화지수행렬은 다음과 같다.

$$D_2' = \begin{pmatrix} - & 0 & 0.1667 & 0.4167 & 0.25 \\ 0 & - & 0 & 0.1667 & 0.4167 \\ 0.1667 & 0 & - & 0 & 0.1667 \\ 0.4167 & 0.1667 & 0 & - & 0.5833 \\ 0.25 & 0.4167 & 0.1667 & 0.5833 & - \end{pmatrix}$$

$$D_3' = \begin{pmatrix} - & 0 & 0.0833 & 0.2083 & 0.125 \\ 0 & - & 0 & 0.0833 & 0.2083 \\ 0.0833 & 0 & - & 0 & 0.0833 \\ 0.2083 & 0.0833 & 0 & - & 0.4166 \\ 0.125 & 0.2083 & 0.0833 & 0.4166 & - \end{pmatrix} = D_4'$$

$$D_5' = \begin{pmatrix} - & 0 & 0.9375 & 0.84375 & 0.90625 \\ 0 & - & 0 & 0.9375 & 0.84375 \\ 0.9375 & 0 & - & 0 & 0.9375 \\ 0.84375 & 0.9375 & 0 & - & 0.75 \\ 0.90625 & 0.84375 & 0.9375 & 1 & - \end{pmatrix}$$

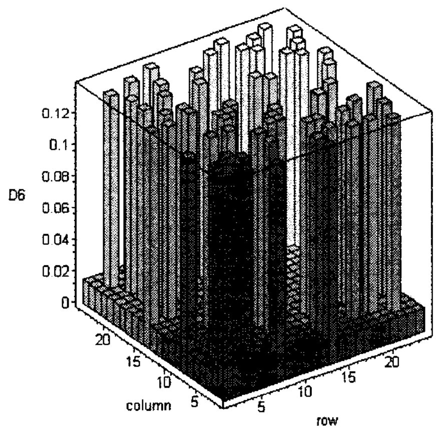
$I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도을 계산하면 $ave(D_6')=0.1688$, $\max(D_6')=0.5$가 된다. 이 두개의 측도들이 모두 0이 아니므로 이 배열은 직교배열이 아님을 알 수 있다.

참고로 기준 2에서 기준 5까지를 이용한 $ave(D_p')$와 $\max(D_p')$(기준 5는 $\min(D_p')$)는 다음과 같다.

$$ave(D_2') = 0.2167,\ \max(D_2') = 0.5833$$
$$ave(D_3') = 0.1208,\ \max(D_3') = 0.4166$$
$$ave(D_4') = 0.1208,\ \max(D_4') = 0.4166$$
$$ave(D_5') = 0.6156,\ \min(D_5') = 0.75$$

기준 5는 나머지 기준들과는 성격이 다르므로 제외하면 이 논문에서 제시하고 있는 상호정보를 이용한 기준 6이 나머지 기준들과 비슷한 결과를 산출한다. 그러므로 새로 제시되는 상호정보를 이용한 기준 6이 직교성의 정도를 평가하는 기준으로서 의미가 있다고 할 수 있다. 만일 상호정보를 이용한 기준 6이 나머지 기준들과 아주 다른 결과를 산출한다면 이 기준 6을 우리는 사용하기 어렵게 된다.

예제 2. Zhang외 2인(2006)의 논문에 나오는 초포화계획 $SSD(14,2^{23})$에 대하여 생각하여 보자. 이 계획은 수준수 q가 2, 배열의 크기 N이 14, 열의 개수 r이 23인 근사직교배열이다. 각 열의 구성요소가 2개(-1과 1)로 동일하고, 각 열에서 각 구성요소의 개수가 같으므로 이 배열은 U-계획이 된다. 이 배열에 대하여 직교화평가행렬 D_6을 구하면 23×23 행렬이 되므로 숫자로 나타내기에는 너무 커서 다음 <그림 1>과 같은 직교화평가행렬그림으로 나타내었다. 이 그림에서 기둥의 높이가 $d(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 값이다. 대각선 원소는 모두 0임을 통하여 이 계획이 U-계획이 임을 알 수 있다. 또한 $d(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 값이 0.0148(444개)과 0.1369(62개) 두 가지로 나뉨을 알 수 있다. $I(i,j)$를 구한 후 직교화지수행렬 D_6'를 구하면 $D_6'=D_6$임을 알 수 있다. 이 계획의 직교성의 정도를 평가할 수 있는 측도을 계산하면 $ave(D_6')=0.0297$, $\max(D_6')=0.1369$가 된다. 이 두 개의 측도들이 모두 0이 아니므로 이 배열은 직교배열이 아니나 예제 1에서 살펴보았던 균등계획보다는 직교배열에 더 가까움을 알 수 있다.



<그림 1> 직교화평가행렬그림

이 계획의 임의의 두 개의 열 $\boldsymbol{c}_i$와 $\boldsymbol{c}_j$에 대하여 우도비검정을 행하면 p-값이 계산된다. 총 253개의 p-값들이 나오는 데 이 253개의 p-값들 중 222개는 0.4098, 31개는 0.0121이었다. 그러므로 유의수준을 5%라 할 때 총 253개의 검정 중 222개는 귀무가설(H_0: $\boldsymbol{c}_i$와 $\boldsymbol{c}_j$는 서로 직교한다.)을 기각할 수 없고, 31개만이 귀무가설을 기각하게 된다. 상호정보의 과추정(over-estimation) 문제를 고려하면 전체적으로 이 계획은 직교배열에 가깝다고 결론을 내릴 수 있다.

5. 결 론

산업현장에서는 실험계획 시 인자의 수는 매우 많고

실험횟수는 적을 수밖에 없는 경우가 자주 일어난다. 그래서 초포화계획이 매우 중요한 수단이 된다. 이 때 직교성의 정도를 평가하는 측도의 개발은 매우 중요하다. 직교성의 정도를 평가하는 측도로서 상호정보를 이용하는 방법은 정보이론을 바탕으로 하는 방법으로서 기존의 방법과는 다른 접근방법이다. 상호정보를 이용하면 근사직교배열의 직교성의 정도를 평가할 수 있고, 제한적이기는 하지만 가설검정도 행할 수 있게 된다.

참고문헌

[1] 장대홍(2004), "근사직교배열의 직교성의 정도를 평가하기 위한 그래픽방법", *한국품질경영학회지*, 32권 4호, 220-228.

[2] Booth, K. H. V. and Cox, D. R.(1962), "Some Systematic Supersaturated Designs", *Technometrics*, Vol. 4, 489-495.

[3] Brillinger, D. R. and Guha, A.(2007), "Mutual Information in the Frequency Domain", *Journal of Statistical Planning and Inference*, Vol. 137, 1076-1084.

[4] Bulutoglu, D. A. and Ryan, K. J.(2008), "$E(s^2)$-optimal Supersaturated Designs with Good Minimax Properties when N is Odd", *Journal of Statistical Planning and Inference*, Vol. 138, 1754-1762.

[5] Butler, N. A.(2008), "Two-Level Supersaturated Designs for 2^k Runs and Other Cases", *Journal of Statistical Planning and Inference*, to appear.

[6] Chatterjee, K., Sarkar, A. and Lin, D. K. J.(2008), "Supersaturated Designs with High Searching Probability", *Journal of Statistical Planning and Inference*, Vol. 138, 272-277.

[7] Chen, J. and Liu, M. Q.(2008), "Optimal Mixed-level Supersaturated Design with General Number of Runs", *Statistics and Probability Letters*, to appear.

[8] ________(2008), "Optimal Mixed-level k-circulant Supersaturated Designs", *Journal of Statistical Planning and Inference*, Vol. 138, 4151-4157.

[9] Das, A., Dey, A., Chan, L. Y. and Chatterjee, K.(2008), "On $E(s^2)$-optimal Supersaturated Designs", *Journal of Statistical Planning and Inference*, Vol. 138, 3749-3757.

[10] Dionisio, A., Menezes, R. and Mendes, D. A.(2004), "Mutual Information: A Measure of Dependency for Nonlinear Time Series", *Physica A*, Vol. 344, 326-329.

[11] Fang, K. T., Lin, D. K. J. and Liu, M.(2003), "Optimal Mixed-level Supersaturated Design", *Metrika*, Vol. 58, 279-291.

[12] Fang, K. T., Lin, D. K. J., Winker, P., and Zhang, Y.(2000), "Uniform Design: Theory and Application", *Technometrics*, Vol. 42, 237-248.

[13] Fang, K. T., Tang, Y. and Yin, J.(2008), "Lower Bounds of Various Criteria in Experimental Designs", *Journal of Statistical Planning and Inference*, Vol. 138, 184-195.

[14] Jang, D. H.(2002), "Measures for Evaluating Non-orthogonality of Experimental Designs", *Communications in Statistics-Theory and Methods*, Vol. 31, 249-260.

[15] Jones, B. A., Lin, D. K. J. and Nachtsheim, C. J.(2008), "Bayesian D-optimal Supersaturated Designs", *Journal of Statistical Planning and Inference*, Vol. 138, 86-92.

[16] Jones, B. A., Li, W., Nachtsheim, C. J. and Ye, K. Q.(2008), "Model-Robust Supersaturated and Partially-Supersaturated Designs", *Journal of Statistical Planning and Inference*, to appear.

[17] Koukouvinos, C., Mylona, K.(2008), "A Method for Analyzing Supersaturated Designs with a Block Orthogonal Structure", *Communications in Statistics-Simulation and Computation*, Vol. 37, 290-300.

[18] Koukouvinos, C., Mylona, K. and Simos, D. E.(2008), "A Hybrid SAGA Algorithm for the Construction of $E(s^2)$-optimal Cyclic Supersaturated Designs", *Journal of Statistical Planning and Inference*, to appear.

[19] ________(2008), "$E(s^2)$-optimal and Minimax-optimal Cyclic Supersaturated Designs via Multi-objective Simulated Annealing", *Journal of Statistical Planning and Inference*, Vol. 138, 1639-1646.

[20] Ma, C., Fang, K. T., and Liski, E.(2000), "A New Approach in Constructing Orthogonal and Nearly Orthogonal Arrays", *Metrika*, Vol. 50, 255-268.

[21] Nguyen, N. K. and Cheng, C. S.(2008), "New $E(s^2)$-optimal Supersaturated Designs Constructed from Incomplete Block Designs", *Technometrics*, Vol. 50, 26-31.

[22] Plackett, R. L. and Burman, J. P.(1946), "The Design of Optimum Multifactorial Experiments", *Biometrika*, Vol. 33, 305-325.

[23] Siomina, I. and Ahlinder, S.(2008), "Lean Optimization Using Supersaturated Experimental Design", *Applied Numerical Mathematics*, Vol. 58, 1-15.
[24] Wang, J. C. and Wu, C. F. J.(1992), "Nearly Orthogonal Arrays with Mixed Levels and Small Runs", *Technometrics*, Vol. 34, 409-422.
[25] Yamada, S. and Lin, D. K. J.(1999), "Three-level Supersaturated Designs", *Statistics and Probability Letters*, Vol. 45, 31-39.
[26] Zhang, Q. Z., Zhang, R. C. and Liu, M. Q.(2006), "A Method for Screening Active Effects in Supersaturated Designs", *Journal of Statistical Planning and Inference*, Vol. 137, 2068-2079.