한국통계학회논문집 제4권 3호
1997년 12월 pp. 685 - 692

I_λ-최적실험계획을 위한 새로운 가중영역[1)]

김 영 일[2)]

요 약

실험자의 주관을 제일 잘 반영한다고 볼 수 있는 I_λ-최적 실험계획의 실제적 사용을 위하여 새로운 형태의 가중영역을 제시하였다. 그리고 반응표면분석에서 자주 쓰이는 이차원의 2차형식 함수를 기준으로 이에 대한 성질 및 특징을 알아보았다.

1. 소개

최적실험계획법(optimal experimental design)은 Kiefer와 Wolfowitz(1959)의 동격이론(equivalence theory)이 나온 이래 많은 이론적 발전을 하였다. 그러나 Box(1980)가 지적하였듯이 영어의 알파베트 이름을 가진 이러한 최적실험계획이론은 너무 수학적인 가정과 논리에 의거하였기 때문에 실제 실험계획을 하는 분석가들에게는 많은 도움을 주지 못한 것이 사실이다.

본 연구에서는 Fedorov(1972)가 제안한 선형(linear)-최적 실험기준의 하나인 I_λ-최적에 대해 알아보고 이를 활용할 수 있는 간단한 가중영역을 제안한다. 2절에서는 내용전개를 위한 표현방법을 알아보고, 3절에서는 I_λ-최적의 특징을 간단한 예를 통해 D-최적과 비교하여 알아본 다음, 4절에서는 분석가의 관심을 반영할 수 있는 간단한 가중영역을 제안한다. 그리고 도출된 실험계획의 특징을 기존의 통상적인 가중영역과 비교함으로써 새로운 가중영역의 유용성을 점검하여 본다.

2. 표현방법

관측값 y_{ij} 는 다음과 같은 선형모형을 통하여 이루어진다.

$$y_{ij} = \theta^T f(x_i) + \varepsilon_{ij}, \quad (1)$$
$$i = 1, \cdots, n, \ \ j = 1, \cdots, r_i, \ \ \sum r_i = N$$

여기서 $\theta \in R^m$ 는 미지의 모수이며, $f^T(x) = (f_1(x), \cdots, f_m(x))$ 는 주어진 반응함수의 형태에 의

1) 본 연구는 1996년도 중앙대학교 교내 학술연구비에 의해 지원되었음.
2) (456-756) 경기도 안성군 대덕면 내리 산 40-1 중앙대학교 산업정보학과 부교수

존하는 mx1 벡터이며, 받힘점(supporting point) x는 어떤 집합 X에서 선택될 수 있으며, 그리고 ε_{ij}는 평균 0이고 분산이 1인 서로 비상관인 확률오차이다.

미지의 모수의 최량선형불편추정량에 대한 정밀성(accuracy)은 식 (2)와 같은 정보행렬로 설명할 수 있다.

$$M(\xi)=N^{-1}\sum p_i f(x_i)f^T(x_i),\ p_i=r_i/N \tag{2}$$

이는 실험계획 $\xi=(x_i,p_i)$에 의해 완벽하게(completely) 정의된다. 연속이론으로 설명하면 식 (2)는 다음과 같이 쓰여진다.

$$M(\xi)=\int_X f(x)f^T(x)\xi(dx) \tag{3}$$

여기서 $\xi(dx)$는 받힘점 집합이 X에 속하게끔 하는 확률측도(probability measure) 혹은 연속(continuous)실험계획이다. 그리고 다음과 같은 실험계획 ξ^*를

$$\xi^*=\arg\ \min_{\xi}\Psi[M(\xi)] \tag{4}$$

Ψ-최적이라 한다. Fedorov(1972)는 $\Psi(M)=-\ln|M|$ 및 $\Psi(M)=\mathrm{tr}(M^{-1}B)$의 실험계획을 각각 D-최적, 선형-최적이라 불렀다. 여기서 B는 비음치(non-negative definte)행렬이다. Studden(1977)은 선형-최적의 하나인 I_λ-최적을 소개하였다.

$$\xi^*=\arg\ \min_{\xi}\int_\Omega f^T(x)[M(\xi)]^{-1}f(x)\lambda(dx) \tag{5}$$

여기서 B를 $\int_\Omega f(x)f^T(x)\lambda(dx)$로 치환하면 I_λ-최적은 선형-최적의 한 종류임을 쉽게 알 수 있다. 식 (5)에서 Ω는 주어진 실험영역 X과 일치할 필요는 없다. 이러한 이유로 I_λ-최적은 실험자의 관심을 끄는 가중영역(weighting space) Ω에 적절한 가중함수인 λ를 선택함으로서 외삽법(extrapolation)등 여러 방면에 활용될 수 있다. I_λ-최적은 D-최적과는 달리 문헌에서는 자주 언급되고 있지 않으나 Myers와 Montgomery(1995)가 지적하였듯이 많은 응용 가능성을 가지고 있다. I_λ-최적에 대한 알고리즘 구성은 Fedorov(1972)를 참조 바란다. 다음 절에서는 예제를 통해 이러한 I_λ-최적의 특징을 살펴본다.

3. I_λ-최적의 특징 및 문제점

(예제 1) 모형은 $f^T(x)=(1,x,x^2)$으로 주어져 있고, $\Omega=[-g,g]$인 가중영역에서 가중함수가 균일(uniform)분포함수일 경우, 실험영역 $X=[-1,1]$에서의 I_λ-최적 실험계획의 받힘점은 -1, 0, 1 이고 이에 부여되는 질량은 각각 다음의 제곱근에 비례한다(Studden, 1971).

$$\frac{1}{4}\left(\frac{1}{5}+\frac{1}{3g^2}\right),\quad \frac{1}{5}-\frac{2}{3g^2}+\frac{1}{g^4},\quad \frac{1}{4}\left(\frac{1}{5}+\frac{1}{3g^2}\right)$$

만약 $g=1$ 이면 $\xi(\pm1)=1/4$, $\xi(0)=1/2$ 이다. 그리고 $g\to\infty$ 이면 부여되는 질량은 $g=1$ 인 경우와 같아진다. 또한 $g=1.6$ 일 때 제일 많은 질량을 양극점 (±1) 에서 부여받는다, $\xi(\pm1)=0.328$. D-최적인 경우는 $\xi(\pm1)=\xi(0)=1/3$ 임을 고려하면 I_λ-최적은 실험영역을 벗어난 $g=1.6$ 의 가중영역일 때 D-최적과 비슷한 질량을 가진다고 해석할 수 있다. 어떠한 g 값이라도 양극점에서의 질량이 D-최적의 $\xi(\pm1)=1/3$ 의 경우보다 작게 나타나 I_λ-최적은 D-최적과 달리 가운데 받힘점에 보다 많은 질량을 부여함으로서 평균분산값을 작게 하는 경향이 있다.

(예제 2) 반응표면분석에서 자주 인용되는 다차원 이차형식모형은 다음과 같다.

$$f(x)=(1;x_1,x_2,\cdots,x_k;x_1^2,x_2^2,\cdots,x_k^2;x_1x_2,\cdots,x_1x_k,x_2x_3,\cdots,x_{k-1}x_k)$$

이러한 모형 중 k가 2이고 교호작용 x_1x_2 이 없는 모형 $f^T(x)=(1,x_1,x_2,x_1^2,x_2^2)$ 인 경우, $X=[-1,1]^2$ 인 영역에서 D-최적 실험계획 ξ_D은 요인수준(factor level)이 -1, 0, 1인 3^2 요인 실험 각 점에 같은 질량 1/9 을 부여한다는 것은 잘 알려져 있는 사실이다. 실험자가 $X=[-1,1]^2$ 에 균일분포 가중함수를 도입하여 I_λ-최적을 구한다면 다음과 같다.

꼭지점: $\xi(1,1)=\xi(1,-1)=\xi(-1,1)=\xi(-1,-1)=0.073$
주변중앙점: $\xi(1,0)=\xi(0,1)=\xi(0,-1)=\xi(-1,0)=0.109$
중심점: $\xi(0,0)=0.272$

I_λ-최적은 D-최적에 비해 중심점(center point)에 대한 질량이 증가된 반면, 실험영역의 주변점의 질량이 전반적으로 감소되었고, 꼭지점(corner points)보다는 주변의 중앙점(middle point of edge)에 보다 많은 질량이 부여되는 특징을 가지고 있다. 바꾸어 말하면, D-최적은 실험영역의 가장자리에 보다 많은 질량을 부여함으로서 예측값의 최대분산값을 끌어내리는 효과를 가져다준다 할 수 있다.

I_λ-최적은 가중영역과 가중함수의 형태에 의해 지배받는 특징을 지니고 있다. 일반적으로 균일분포함수는 수학적 간편성 때문에 많은 문헌에서 가중함수로 쓰이나, 가중영역 전체에 실험자가 관심을 갖지 않는 한, 이의 무조건적인 사용은 자제해야 한다. 이러한 점을 감안한 가중함수로 제일 적합한 함수는 베타(beta)분포함수를 들 수 있다. 그러나, 설명변수의 개수가 하나인 경우는 베타분포함수를 이용하여 비교적 용이하게 다양한 가중함수의 형태를 고려할 수 있지만, 설명변수 개수가 2개 이상인 경우에는 수학적 복잡성 때문에 이의 사용은 제한될 수 밖에 없다. Kim(1993)은 단순회귀모형 $f^T(x)=(1,x)$ 과 이차형식모형 $f^T(x)=(1,x,x^2)$ 에 베타(Beta)분포함수를 가중함수로 하여 I_λ-최적의 특징을 알아보았다. 그리고 D-최적은 실험자의 주관을 반영시키지 못하는 단점이 있기 때문에, 다음 절에서는 이러한 점을 고려하여 분석가가 실제로 쓰기 쉬운 가중영역을 제시하여 본다.

4. 가중영역

Giovannitti-Jensen과 Myers(1989)가 언급하였듯이 많은 실험자는 특히나, 반응표면(response surface) 실험에서는 실험영역의 중앙보다는 가장자리(perimeter)에 보다 많은 관심을 가지게 된다. 이러한 상황에서는 실험영역 가장자리의 질량을 감소시키는 특징을 지니고 있는 (예제 1, 2)와 같은 I_λ-최적은 바람직하지 않고, 실험영역의 가장자리에 보다 많은 질량을 부여하는 특징을 가진 D-최적이 선호되는 기준이라 할 수 있다. 그러나 D-최적은 실험자의 주관을 반영시킬 수 있는 기준이 아니기 때문에 유연성을 가진 실험기준이라고는 볼 수 없다.

Giovanniti-Jensen과 Myers(1989)는 $U_r=\{x: \sum_{i=1}^{k} x_i^2=r^2\}$로 정의된 반지름이 r인 구(sphere)의 표면(surface)에 균일하게 가중치를 부여하여 계산된 예측값의 평균분산(spherical average prediction variance)이 반지름 값 r에 따라 어떻게 변하는지를 보여주는 도시적인 방법을 통하여 주어진 실험들을 비교하였다. 그러나 그들은 주어진 실험을 비교만 하였지, 이러한 가중영역을 활용한 최적실험은 생각하지 못하였다. 따라서, 실험영역의 주변에 많은 관심을 가진다면 이러한 U_r을 가중영역으로 가지는 I_λ-최적이 D-최적을 대신하는 실험기준이라 볼 수 있다. 그러면 실험자는 이러한 가중영역을 통하여 D-최적이 가지고 있지 않은 점을 보완하면서 실험자의 주관이 반영된 질량을 부여할 수 있을 것이다.

이러한 가중영역의 형태는 반드시 실험계획의 형태와 일치할 필요는 없다. 즉, U_r형태의 가중영역을 쓰더라도 실험영역은 반드시 구형태일 필요는 없다. Covey-Crump 와 Silvey (1970)가 지적하였듯이 많은 경우, 개개의 설명변수가 취할 수 있는 값의 범위가 $a\le x_i\le b$의 형태로 주어지기 때문에 오히려 실험영역은 구 형태보다는 입방형(cuboidal)형태의 영역이 자연스럽다. 따라서, 가중영역 역시 구 형태보다는 입방형으로 정의될 수 있다. 입방형의 가중영역의 겉표면 C_g은 다음과 같이 정의된다.

$$C_g=\{x: -g\le x_i\le g,\ i=1,\cdots,k;\ i\neq j;\ x_j=\pm g\}$$

가중영역의 크기는 r이나 g에 의해 결정되는데 이의 선택은 실험의 특징에 따라 실험자가 주관적으로 해야 한다.

(예제 3) 위에서 언급한 두 가지 형태의 가중영역에 균일분포함수를 부여한 I_λ-최적을 $f^T(x)=(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$인 경우, $X=[-1, 1]^2$의 실험영역에서 찾아본다. 여기서 r(혹은 g)의 값은 편의상 1 과 $\sqrt{2}$를 취하였다. 먼저 위에서 언급한 가중영역에 따른 실험을 구하기 전에 $X=[-1, 1]^2$인 실험영역에서 D-최적 실험계획 ξ_D을 알아보면 다음과 같다.

꼭지점: $\xi(1,1)=\xi(1,-1)=\xi(-1,1)=\xi(-1,-1)=0.146$
주변중앙점: $\xi(1,0)=\xi(0,1)=\xi(0,-1)=\xi(-1,0)=0.080$
중심점: $\xi(0,0)=0.096$

교호작용 변수인 x_1x_2가 모형에 들어가 있기 때문에 3절 (예제 2)에서 나온 등질량(equi-weight)의 D-최적과는 차이가 있음을 참조하기 바란다. 교호작용을 감안하였기 때문에 꼭지점에 보다 많은 질량이 부여되어 있다. $k \geq 3$ 인 경우는 Galil과 Kiefer(1977b)를 참조 바란다.

가중영역별로 구한 I_λ-최적을 r(혹은 g)의 값에 따라 <표 1>에 수록하였다. I_λ-최적을 구하기 위해 필요한 $B = \int_\Omega f(x) f^T(x)\, \lambda(dx)$ 값은 부록에 수록하였다. 참고로 통상적인 전체영역을 대상으로 균일분포의 가중함수를 도입하는 최적실험도 구인 경우는 S_r 그리고 입방인 경우는 H_g로 나누어 표기하였다.

<표 1>

r(혹은 g)의 값이 1인 경우의 질량

	U_r	S_r	C_g	H_g	D-최적
꼭지점	0.078	0.069	0.119	0.091	0.146
주변중앙점	0.156	0.099	0.122	0.091	0.080
중심점	0.064	0.328	0.036	0.272	0.096
D-효율성	89.72	82.83	97.11	90.30%	

r(혹은 g)의 값이 $\sqrt{2}$ 인 경우의 질량

	U_r	S_r	C_g	H_g
꼭지점	0.122	0.091	0.135	0.119
주변중앙점	0.125	0.126	0.097	0.102
중심점	0.012	0.132	0.072	0.116
D-효율성	96.25	93.62	99.50	98.41%

일반적으로 U_r이나 C_g인 경우 중심점에 대한 질량이 낮음을 알 수 있다. 반대로 S_r이나 H_g인 경우는 영역 전체에 관심이 있는 관계로 전체적으로 중심점에 대한 질량이 높음을 알 수 있다. 가중영역이 입방형으로 되어 있는 경우에는 구 형태의 가중영역에 비해, 꼭지점에 보다 많은 질량이 배치됨을 알 수 있다. 또한 $g=1$인 C_g인 경우 I_λ-최적을 D-최적과 비교하면 꼭지점과 비슷한 질량이 주변중앙점에 부여되는 것을 알 수 있다. 이는 D-최적은 예측값의 최대 분산값을 낮추고자 한 반면, I_λ-최적은 실험영역 $X=[-1,1]^2$의 주변전체에 가중치를 부여시켰기 때문으로 분석된다. 두 최적 모두 실험영역의 주변에 관심을 많이 부여하는 실험기준이지

만, I_λ-최적은 실험자의 주관을 반영할 수 있다는 장점을 가지고 있어 D-최적에 비해서 손색없는 실험기준을 제공한다고 본다. 참고로 <표 1> 각 하단에 특정의 실험, ξ이 D-최적인 ξ_D에 대해 가지는 D-효율성, 즉 $(|M^{-1}(\xi_D)|/|M^{-1}(\xi)|)^{1/p}$을 적어 놓았다. 여기서 모형에 들어 있는 모수의 수인 p는 6이다. r(혹은 g)의 값에 관계없이 C_g의 경우가 기대하였던 대로 D-효율성이 가장 높음을 알 수 있다. U_r인 경우는 H_g보다 효율성 면에서는 뒤떨어짐을 알 수 있다. 이는 $r=1$인 경우는 U_r의 가중영역은 꼭지점을 포함하고 있지 않고 $r=\sqrt{2}$인 경우는 가중영역이 입방형의 실험영역을 벗어나 구의 형태로 설정되었기 때문으로 분석된다. 이러한 가중영역을 이용한 I_λ-최적은 $k \geq 3$인 경우에도 쉽게 적용될 수 있다.

5. 결론

본 연구에서는 다양한 형태의 가중영역을 고려하여 I_λ-최적실험계획의 유용성을 고려하여 보았다. $k \geq 2$인 경우는 $k=1$인 경우와 달리 수학적인 어려움 때문에 균일분포함수 이외의 함수를 가중함수로 둘 수 없다. 그러나 실험기준이 가지고 있는 특성 및 관심사항을 적절한 크기의 가중영역으로 반영시킴으로서 실험의 효율성을 높일 수 있다. 실험자는 본 연구에서 제시한 가중영역의 크기를 반영하는 r이나 g의 값을 조정함으로서 여러 형태의 실험을 할 수 있는데, 이는 기존의 D-최적에 비해 큰 장점이라 하겠다.

실질적으로 I_λ-최적이 실험자에게 도움이 되기 위해서는 앞으로도 다양한 가중함수의 개발이 있어야 한다. 또한, 본 연구에서는 연속 최적화이론이 입각하여 논리를 전개하였으나 실제 실험자가 도움을 받기 위해서는 이산형태의 알고리즘에 대한 설명도 필요하다. 그리고 구의 중심을 이동시켜 원하는 가중영역을 설정할 수 있다. 특히 반응표면분석과 같이 실험이 일회로 끝나지 않고 탐색적으로 이루어지는 경우는 이러한 가중함수의 개발도 고려해 볼 만하다.

참고문헌

[1] 김영일(1993). 단순선형회귀모형과 이차형식회귀모형을 중심으로 D-와 이분산 G-최적에 비교한 I_λ-최적실험기준의 특성연구, 품질관리학회지, 제 21권 제 2호, 140-155.

[2] Box, G. E. P. (1980). Choice of response surface design and alphabetical optimality, *Utilas Mathematica,* 21B, 11-55.

[3] Covey-Crump, P. A. K. and Silvey, S. D. (1970). Optimal regression with previous observations, *Biometrika,* 57, 551-566.

[4] Fedorov, V. (1972). *Theory of Optimal Experiments*, New York: Academic Press.

[5] Galil, Z. and Kiefer, J. (1977a). Comparison of rotatable designs for regression on balls, I (quadratic), *Journal of Statistical Planning and Inference,* 1, 27-40.

[6] Galil, Z. and Kiefer, J. (1977b). Comparison of design for quadratic regression on cubes, *Journal of Statistical Planning and Inference,* 1, 121-132.

[7] Giovannitti-Jensen, A. and Myers, R. H. (1989). Graphical assessment of response surface designs, *Technometrics*, 31, 158-171.

[8] Kiefer, J. and Wolfowitz, J. (1959). Optimum Designs in Regression Problems, *Annals of Mathematics*, 30, 271-294.

[9] Myers, R. H. and Montgomery, D.C. (1995). *Response Surface Methodology*, New York: John Wiley.

[10] Rozum, M. A. and Myers, R.H. (1991). Variance Dispersion Graphs for Cuboidal Regions, paper presented at ASA Meetings, Atlanta, GA.

[11] Studden, W. J. (1971). Optimal designs and spline regression, *In optimizing Methods in Statistics*, 63-76, New York: Academic Press.

[12] Studden, W. J. (1977). Optimal designs for integrated variance for polynomial regression, *In Statistical Decision Theory and Related Topics II*, New York: Academic Press.

부 록

2절에서 언급한 $B=\int_{\Omega} f(x)f^{T}(x)\,\lambda(dx)$ 의 요소값을 구성하고 있는 차수(order)가 $\delta=\sum\delta_i$ 인 식 (a.1)의 모멘트값 중 δ_i 값이 홀수인 경우에 해당되는 모멘트값은 가중영역의 형태가 대칭(symmetrical region)이므로 0이다.

$$\sigma_{\delta_1\delta_2\cdots\delta_k}=\Psi\int_{\Omega}x_1{}^{\delta_1}x_2{}^{\delta_2}\cdots x_k{}^{\delta_k}d\boldsymbol{x} \tag{a.1}$$

여기서 $\Psi^{-1}=\int_{\Omega}d\boldsymbol{x}$이고, Ω는 가중영역을 의미한다. 각각의 가중영역에 해당되는 0이 아닌 모멘트값을 기술한다.

(a-1) 구의 표면: $U_r=\{\boldsymbol{x}:\sum_{i=1}^{k}x_i{}^2=r^2\}$: Giovannitti-Jensen과 Myers(1989) 및 Galil과 Kiefer(1977a)를 참조바람.

$$\sigma_2=\Psi\int_{U_r}x_i^2\,d\boldsymbol{x}=\frac{r^2}{k}$$

$$\sigma_4=\Psi\int_{U_r}x_i^4\,d\boldsymbol{x}=\frac{3r^4}{k(k+2)}$$

$$\sigma_{22}=\Psi\int_{U_r}x_i^2x_j^2\,d\boldsymbol{x}=\frac{r^4}{k(k+2)}$$

(a-2) 입방면체의 표면: $C_g=\{\boldsymbol{x}:-g\le x_i\le g,\ i=1,\cdots,k;\ i\neq j;\ x_j=\pm g\}$: Rozum과 Myers(1991)를 참조바람.

$$\sigma_2=\Psi\int_{C_g}x_i^2\,d\boldsymbol{x}=\frac{(k+2)g^2}{3k}$$

$$\sigma_4=\Psi\int_{C_g}x_i^4\,d\boldsymbol{x}=\frac{(k+4)g^4}{5k}$$

$$\sigma_{22}=\Psi\int_{C_g}x_i^2x_j^2\,d\boldsymbol{x}=\frac{g^4}{3}$$