응용통계연구 제 6권 2호
1993년 9월 pp.303-309.

D-와 이분산 G-최적을 중심으로 한 오차로버스트 실험계획법

김 영 일[1)]

요 약

오차에 대한 함수관계식이 불확실한 경우 두가지 실험계획법을 제시하였다. 하나는 모든 가능한 오차함수식을 대상으로 최저의 효율성을 높이는 방안이고 다른 하나는 확률을 이용한 최저의 평균효율성을 높이는 방안이다. 이러한 두 방법을 D-와 이분산 G-최적성에 적용시켜 그 차이점을 비교연구하였다.

1. 소개 및 표현방법

특별한 경우를 명시하지 않는 한 다음의 선형모형을 간주하고자 한다.

$$y_i=f^{T}(x_i)\theta+\varepsilon_i,\ i=1,2,\cdots,N \tag{1}$$

x_i 는 예측변수 $q\times 1$ 벡터이고 $f^{T}(x_i)$는 가정된 반응함수의 형태에 의존되는 $p\times 1$ 벡터, θ 는 미지의 모수 $p\times 1$벡터이며 ε_i 는 평균 0 공분산 행렬 σ^2V 을 따르는데, V는 원소가 $v_1,\cdots,v_N$ 로 되어있는 대각행렬이다. 일반성의 손실없이 $\sigma^2=1$을 채택한다. 또한 개개의 v_i 는 x_i 의 함수로서 가정을 하고, $v_i=\omega^{-1}(x_i)$로 표기한다. (앞으로는 $\omega(x)$는 효율함수, $\omega^{-1}(x)$는 오차함수로 불리도록 한다) 실험계획문제는 실험계획영역에서 벡터 $x_i,\ i=1,2,\cdots,N$ 를 선택하는 문제인데, 어떠한 설정된 의미하에서 이러한 N벡터로 형성된 실험계획이 최적이 될 수 있게 끔 하는 것이다. 이러한 아이디어는 쉽게 확률을 이용한 실험계획으로 연장할 수 있다. 실험계획문제는 $(f,\chi\omega)$ 에 대한 확률적인 실험계획 $\xi\in E_x$ 의 선택이라 할 수 있다. 식 (1)에 관련하여 우리는 다음과 같은 정보행렬을 정의할 수 있다.

$$M(\xi)=\int_{\chi}\omega(x)f(x)f^{T}(x)d\xi(x).$$

정의 1.1 실험계획 ξ_D는 아래 조건이 만족할때 D-최적이라 한다.

$$\max_{\xi\in E_x}|M(\xi)|=|M(\xi_D)|$$

정의 1.2 실험계획 ξ_G는 아래 조건이 만족할때 G-최적이라 한다.

$$\min_{\xi\in E_x}\max_{x\in\chi}\omega(x)d(x,\xi)=\max_{x\in\chi}\omega(x)d(x,\xi_G).$$

여기서 $d(x,\xi)=f^{T}(x)M^{-1}(\xi)f(x)$는 실험계획 ξ의 주어진 x값에서의 예측치의 분산함수로 볼 수 있다.

1) (456-756) 경기도 안성군 대덕면 내리 산40-1. 중앙대학교 산업대학 산업정보학과.

최적실험계획에서는 이러한 D 및 G-기준이 많이 쓰이고 있는데, 이의 한 이유로서 $\omega(x)=1$인 경우는 이 두 기준들의 동치를 들 수 있다. 그러나 이러한 장점은 이분산 오차함수인 경우에는 그 의미를 상실한다고 할 수 있다. 이러한 맥락에서 이분산 함수하에서 $d(x,\xi)$의 최대치를 최소화하는 기준을 제시하고자 한다.

정의 1.3 실험계획 ξ_{G_0}는 아래 조건이 만족할때 G_0-최적이라 한다.

$$\min_{\xi\in E_\chi}\max_{x\in\chi} d(x,\xi)=\max_{x\in\chi} d(x,\xi_{G_0}).$$

정의 1.4 실험계획 ξ_1를 기준으로 실험계획 ξ의 D-효율성은 다음과 같다.

$$D(\xi,\xi_1)=\{det\ M^{-1}(\xi_1)det\ M(\xi)\}^{1/p}.$$

정의 1.5 실험계획 ξ_1를 기준으로 실험계획 ξ의 G-효율성은 다음과 같다.

$$G(\xi,\xi_1)=\max_{x\in\chi}\omega(x)d(x,\xi_1)/\max_{x\in\chi}\omega(x)d(x,\xi).$$

정의 1.6 실험계획 ξ_1를 기준으로 실험계획 ξ의 G_0-효율성은 다음과 같다.

$$G_0(\xi,\xi_1)=\max_{x\in\chi} d(x,\xi_1)/\max_{x\in\chi} d(x,\xi).$$

2절에서는 예를 들어가면서 D-최적성과 G_0-최적성의 성질을 비교하겠으며 이어서 오차구조에 대한 지식이 결여된 상태하에서의 실험계획을 위한 두가지 기준을 제시할 것이다. 각각의 기준에 따른 최적실험계획법을 G_0 및 G-(D-)최적성을 갖고 비교하겠다. 이어 결어로 이어질 것이다.

2. 오차로버스트 기준

어떠한 현상의 메카니즘을 설명하고자 할때 실험자들은 실험조건의 동질성을 보통 의심받기 쉬운 비교적 넓은 실험영역에 관심을 갖는 경우가 있는데, 이러한 경우는 등분산을 가정한 실험계획의 적용이 타당한 경우가 아니라 할 수 있다.

다음에서는 모델안에 명시되는 효율함수에 대해 "로버스트적"인 실험계획법설정을 고려하고자 한다. 참의 효율함수가 알려져 있지 않다는 가정하에서 어떠한 실험계획이 모델내에 명시되는 ω에 대해 로버스트하다함은 이러한 실험계획이 실제로 발생할 수 있는 어떠한 ω에 대해서도 효율적으로 임함을 뜻한다. 좀 더 구체적으로 ω는 효율함수들의 공간, W의 하나의 알려지지않은 원소로 간주하자. 이러한 공간들의 어떠한 원소가 효율함수로 실제 부각이 되더라도 추후 효율성에 대한 정의를 하겠지만 설정된 실험계획법이 높은 효율성을 갖게끔 실험계획을 세우고자 한다.

Atwood(1969) 에 의하면 어떠한 실험계획법의 G-효율성은 모든 $\omega\in W$에 구애받지 않고 이 실험계획법의 D-최적실험계획법에 대한 D-효율성의 아래한계를 제공하는데, 즉 실제 시행된 실험계획의 D-효율성은 산출된 G-효율성보다 더 높은 수치를 가질 것이므로 개개의 $\omega\in W$에 대한 실험계획법의 G-효율성만 간주하더라도 충분할 것이다.

정의 2.1 실험계획 $\xi^*\in E_\chi$ 는 아래 조건이 만족할때 G-오차로버스트라 한다.

$$\max_{\xi\in E_\chi}\min_{\omega\in W} G(\xi,\xi_\omega)=\min_{\omega\in W} G(\xi^*,\xi_\omega).$$

여기서 모델의 모수의 숫자, p는 $\omega(x)$에 의해 영향받지 않는 상수이므로, 정의 2.1로 형성된 실험계획은 예측값의 "평준화"된 분산 $\omega(x)d(x,\xi)$값들 중 "최악"의 값을 최소화 시킨다고

볼 수 있다. 그러나 만약 실험자가 모델의 반응면을 추정하는데 관심이 있고, 실험계획 영역하에서 예측값의 "실제" 분산의 최대치를 최소화하는데 관심이 있으면, 위의 방법보다는 아래와 같은 정의가 더 타당할 것이다.

정의 2.2 실험계획 $\xi^* \in E_\chi$ 는 아래 조건이 만족할때 G_0-오차로버스트라 한다.

$$\max_{\xi \in E_\chi} \min_{\omega \in W} G_0(\xi, \xi_\omega) = \min_{\omega \in W} G_0(\xi^*, \xi_\omega).$$

이러한 G-나 G_0-오차로버스트 실험은 Fedorov가 1972년 소개한 알고리즘의 간단한 변형을 통해 컴퓨터로 구성할 수 있다. 자세한 내용은 김(1991)을 참조 바란다.

Läuter(1974)는 최적실험계획법에서의 참모델의 정확한 형태가 알려져야 한다는 제약을 둘 수 없는 경우에 대해서 기본적인 연구의 틀을 제공하였다. Läuter의 방법에 의하면 반응함수는 어떠한 모델공간의 하나의 원소로서 가정을 해야한다. 이러한 가정은 위의 두 기준에서의 설정과 비슷하다. 그리고 모델공간에 대해 사전에 알려진 확률 β를 부여할 수 있어야 한다고 설정하고 있다. 다음에서는 이러한 Läuter방법을 오차로버스트 실험계획에 응용코자 한다.

위에서 정의한 효율함수의 공간에 속해 있는 오차함수들에 대한 가중치를 고려키 위해 효율함수공간을 위한 색인집합 I에 대한 확률메져 β를 취하였다. 예로 이러한 β는 실험자의 효율함수의 공간에 속해 있는 개개의 오차함수들에 대한 타당성에 대한 사전믿음을 표시할 수도 있을 것이다.

정의 2.3 실험계획 $\xi^* \in E_\chi$ 은 아래 조건이 만족할 때 $\overline{G}$-오차로버스트라 한다.

$$\max_{\xi \in E_\chi} \min_{x \in \chi} \int_I G(\xi, \xi_\omega) d\beta(i) = \min_{x \in \chi} \int_I G(\xi^*, \xi_\omega) d\beta(i).$$

또한 이와 비슷한 정의가 따르는데

정의 2.4 실험계획 $\xi^* \in E_\chi$ 은 아래 조건이 만족할때 $\overline{G}_0$-오차로버스트라 한다.

$$\max_{\xi \in E_\chi} \min_{x \in \chi} \int_I G_0(\xi, \xi_\omega) d\beta(i) = \min_{x \in \chi} \int_I G_0(\xi^*, \xi_\omega) d\beta(i).$$

다음에서는 하나의 예를 통해 이렇게 정의된 네가지 실험기준들의 성질을 파악하고자 한다.

이차형식의 회귀모형, $f^T(x) = (1, x, x^2)$하에서 오차항에 대한 분산의 지식은 다음과 같다고 가정을 하자. 제일 작은 분산의 값과 제일 큰 값의 비는 고정된 값 $\gamma \geq 1$이며 실험영역 $\chi \in [-1, 1]$ 에서는 분산은 증가한다. 이를 수학적으로 규정하면 $\omega^{-1}(x) \propto (\gamma-1)x + (\gamma+1)$이다. 등분산일 경우 문제의 단순성을 위하여 $\omega^{-1}(x) = [(\gamma-1)x + (\gamma+1)]/2$로 한다. 표 1은 여러가지 경우의 γ에 대해 D-(G-) 최적실험계획을 보여주고 있다. γ의 값이 증가함에 따라, 이 실험계획은 중앙의 받침점을 실험영역의 좌측으로 이동시킨다. 그리하여 분산이 단순 증가한다는 사실에도 불구하고, 이러한 D-(G-)실험계획은 1/3의 디자인질량을 분산이 상대적으로 적은 지역으로 옮겨주고 있다. 이러한 사실은 얼핏 보기에 상식에 벗어나는 실험인 것 같은 착각을 일으키나, D-최적실험계획법은 표준화된 분산을 이용한 알고리즘에 의해 움직인다는 사실을 기억하면 자연스러운 현상일 수 있다. 그러나 경우에 따라서는 등분산을 가정하였을때의 D-최적실험계획법이 그대로 G_0-최적실험으로 유지될 수도 있으므로 어떠한 구조를 갖고 있는 오차함수의 경우에 혹은 어떠한 형태의 모델에 이러한 현상이 벌어지는 지는 더 연구의 대상이 되어야 하겠다.

표 1

아래 오차분산을 갖고있는 이차형식의 회귀분석모형에 대한 G-최적실험법
$\omega^{-1}(x) \propto (\gamma-1)x + (\gamma+1)$

$\gamma = 1$	$\xi(-1)$	=	$\xi(0)$	=	$\xi(1) = 1/3$
$\gamma = 3$	$\xi(-1)$	=	$\xi(-.141191)$	=	$\xi(1) = 1/3$
$\gamma = 5$	$\xi(-1)$	=	$\xi(-.183268)$	=	$\xi(1) = 1/3$
$\gamma = 7$	$\xi(-1)$	=	$\xi(-.221089)$	=	$\xi(1) = 1/3$
$\gamma = 9$	$\xi(-1)$	=	$\xi(-.241081)$	=	$\xi(1) = 1/3$

한편으로 이 예제에 대한 G_0-최적실험계획은 표 2에 구성되어 있다. Wong과Cook (1992)은 G_0-최적실험계획의 필요충분조건을 명시하였는데, 자세한 내용은 이들의 논문을 참조 바란다. 표 2를 보면 G_0-최적실험계획은 가운데 받힘점에서의 질량 1/3은 변함이 없이 받힘점 -1에서 받힘점 1로 질량을 이동시킴을 알 수 있다. 그리고 개개의 질량은 그 점에서의 분산의 값에 비례하여 설정되어 있음을알 수 있다. 그리고 $\omega(x)$가 이 예와 같이 설정되어 있고 $\chi = [-1, 1]$, $f^T(x) = (1, x, x^2)$, $\gamma > 0$이면 $\max_{x\in\chi} d(x,\xi_{G_0}) = 3(\gamma+1)/2$ 임을 쉽게 증명할 수 있다.

표 2

아래 오차분산을 갖고있는 이차형식의 회귀분석모형에 대한 G_0-최적실험법
$\omega^{-1}(x) = [(\gamma-1)x + (\gamma+1)]/2$

max $d(x, \xi)$	γ	$\xi(-1)$	$\xi(0)$	$\xi(1)$
3.0	$\gamma = 1$	1/3	1/3	1/3
6.0	$\gamma = 3$	1/6	1/3	1/2
9.0	$\gamma = 5$	1/9	1/3	5/9
12.0	$\gamma = 7$	1/12	1/3	7/12
15.0	$\gamma = 9$	1/15	1/3	9/15

역시 같은 모형으로서 $W = \{\omega(x) | \omega^{-1}(x) \propto (\gamma-1)x + (\gamma+1),\ \gamma = 1, 3, 5, 7, 9\}$를 고려하여 보자. 즉 실험자는 이 예에서 참 $\omega(x)$은 γ에 의존하는 이 다섯가지 분산의 하나의 형태로서 적절히 대표될 수 있다고 믿고 있는 것이다.

다음의 실험계획은 G-오차로버스트라고 알고리즘에 의해 찾아졌다: $\xi(\pm1) = .325$ $\xi(.039609) = .182$ $\xi(-.260323) = .167$. γ에 대한 가정이 변할 때 실험계획의 G-효율성을 표 3에 적어 놓았다. 예를 들어 표의 첫번째행은 등분산을 가정하였을 시의 실험계획이 다양한 참의 γ의 값이 변할때 가져다 주는 G-효율성이다. 만약 γ가 추후 분석과정을 통하여 9로 판명이 날 시에는, 이러한 실험은 88.8%의 G-효율을 갖고 있다. 제일 최악의 G-효율성은 표 3의 좌측 아래에서 발생하는데 85.4%이며 여기서 γ의 가정된 값은 9이고 실제 참의 γ값은 1로 나타나는 경우이다. 이에 대비하여 G-오차로버스트 실험의 제일 나쁜 G-효율은 97.4%이다.

흥미롭게도 이러한 오차로버스트 실험은 4개의 받힘점을 선택하고 있는데 이는 다음과 같은 설명으로서 이해를 할 수 있을 것이다. 직감적으로 $x = .039609$에서의 질량은 최악의 경우가 $\gamma = 1$일 경우를 대비한 것이고, $x = -.260323$에서의 질량은 $\gamma = 9$인 경우를 대비코자 한 것으로 파악된다. 이러한 직감적인 설명은 컴퓨터 알고리즘을 시행시키는 과정에서도 $\omega(x)d(x, \xi)$의 최고값이 ±1에서만 일어나는 사실로서 뒷받힘 되어진다.

표 3

표2에서의 제시된 모형의 다양한 γ에 대한 G-효율성

가정된 γ	참 γ 1	3	5	7	9
1	1.0	0.958	0.924	0.902	0.888
3	0.948	1.0	0.996	0.998	0.994
5	0.914	0.993	1.0	0.998	0.994
7	0.876	0.979	0.997	1.0	0.999
9	0.854	0.969	0.992	0.998	1.0

γ =	G-오차로버스트 실험의 G-효율성 1	3	5	7	9
	0.974	0.981	0.979	0.978	0.974

한편으로 다음의 실험계획은 G_0-오차로버스트 실험이라고 판정되었다: $\xi(-1) = 0.277$, $\xi(0) = 0.277$, $\xi(1) = 0.446$. γ에 대한 가정이 변할때, 실험계획의 G_0-효율성을 표4에 적어 놓았다. 표 4의 각각의 행은 위의 경우와 비슷한 설명을 해주고 있다. 역시 최악의 G-효율성은 표 4의 좌측아래에서 발생하는데 불과 20.0%이다. 이 경우 G-효율성보다 G_0-효율이 상대적으로 떨어지는 것은 G_0-최적실험에서는 질량을 이동시킴으로서 최적을 구하는 과정에서 기인된 현상이 아닌가 생각된다. 이에 대비하여 G_0-오차로버스트 실험의 제일 나쁜 G_0-효율은 74.3%이다.

표 4

표2에서의 제시된 모형의 다양한 γ에 대한 G_0-효율성

가정된 γ	참 γ 1	3	5	7	9
1	1.0	0.667	0.600	0.571	0.556
3	0.500	1.0	0.900	0.863	0.833
5	0.333	0.667	1.0	0.952	0.882
7	0.250	0.500	0.750	1.0	0.972
9	0.200	0.400	0.600	0.800	1.0

γ =	G_0-오차로버스트 실험의 G_0-효율성 1	3	5	7	9
	0.831	0.831	0.803	0.765	0.743

세번째 실험기준을 다양한 $i = 1, \cdots, 5$까지의 확률 $\beta(i)$를 갖고 같은 예에 적용시켜 보았다. 결과는 표 5에 요약되어 있다. γ에 대한 가중치를 크게 할수록 중앙받힘점은 점 -1로 옮겨감을 알 수 있다. 이는 세번째와 네번째 실험계획을 대비하여 보면 쉽게 파악이 된다.

네번째 실험기준에 대한 결과는 표 6에 요약이 되어있다. 이 경우는 γ에 대한 사전가중치를 증가시킴에 따라 받힘점 1에 대한 질량이 증가됨을 알 수가 있다. G_0-최적성은 어떠한 γ값에 대해서도 받힘점은 변하지 않는 상태에서 디자인질량을 이동시키는 특성을 갖고 있으므로 (표 2 참조) 이 예에서의 받힘점 역시 G_0-최적실험과 마찬가지로 점 -1, 0 및 1이 선택된다.

표 5와 6에 있어서 첫번째 실험계획법은 γ에 대한 가중치의 확률이 모두 같은 경우를 세번째와 네번째의 실험기준을 적용시킨 것이다. 이렇게 형성된 실험법의 G-효율성을 각각의 γ에 대해 계산하여 보면, γ의 크기순서로 94.9%, 99.9%, 99.5%, 98.8%, 그리고 98.1% 로 나타나는데 이의 평균은 98.24%로서 G-오차로버스터 실험을 하였을 시의 평균값, 97.72%보다는 0.52% 높다. 왜냐하면 이 경우에는 평균적인 G-효율성을 높이는데 주안점을 두었기 때문이다. 이와 유사하게 표 6에서 G_0-효율성을 계산하여 보면, γ의 크기순서로 각각 50.4%, 98.1%, 90%, 86.4%, 그리고 84.2%로 나타나는데, 이는 G_0-오차로버스트 실험에서 평균값(79.46%)을 구하는 경우보

다 2.34% 높은 것이다. 그러나 최악의 경우를 비교하여보면 네번째 기준을 적용하였을 시의 50.4%는 G_0-오차로버스트 실험시의 최악의 경우 74.3%보다 현저하게 떨어짐을 알 수 있다. 만약 실험자가 W 공간에 속해 있는 효율함수들의 가능성에 대한 사전지식이 있는 경우에는 세번째나 네번째 기준이 적합할 것이나, 그렇지 않은 경우는 제시된 첫번째와 두번째의 기준이 안정적인 G-효율성을 제공한다는 것을 이 예를 통해 부분적이나마 알 수 있다.

표 5

$\gamma = 1, 3, 5, 7, 9$를 원소로 갖고있는 오차공간의 경우, 표 2에서 제시된 모델에 대한 G-오차로버스트실험법

γ 에 대한 사전확률					실험법
1	3	5	7	9	
0.2	0.2	0.2	0.2	0.2	$\xi(\pm1) = \xi(-.14) = 1/3$
0.0	0.0	0.0	0.0	1.0	$\xi(\pm1) = \xi(-.24) = 1/3$
0.6	0.1	0.1	0.1	0.1	$\xi(\pm1) = \xi(-.077) = 1/3$
0.1	0.1	0.1	0.1	0.6	$\xi(\pm1) = \xi(-.194) = 1/3$

표 6

$\gamma = 1, 3, 5, 7, 9$를 원소로 갖고있는 오차공간의 경우, 표 2에서 제시된 모델에 대한 G_0-오차로버스트실험법

γ 에 대한 사전확률					실험법		
1	3	5	7	9			
0.2	0.2	0.2	0.2	0.2	$\xi(-1) = .168$	$\xi(0) = .327$	$\xi(1) = .504$
0.0	0.0	0.0	0.0	1.0	$\xi(-1) = .067$	$\xi(0) = .333$	$\xi(1) = .600$
0.6	0.1	0.1	0.1	0.1	$\xi(-1) = .169$	$\xi(0) = .354$	$\xi(1) = .477$
0.1	0.1	0.1	0.1	0.6	$\xi(-1) = .100$	$\xi(0) = .348$	$\xi(1) = .552$

3. 결 어

본 논문에서는 오차구조의 잘못된 설정으로 야기될 수 있는 문제에 덜 민감한 실험기준을 두 가지 알아보았다. 하나는 모든 가능한 오차함수식을 대상으로 최저의 효율성을 높이는 방안이고 다른 하나는 확률을 이용한 최저의 평균효율성을 높이는 방안이다. 이러한 두 방법은 실험자의 연구의도에 따라 그 실험기준이 알고자 하는 모수의 부분집합에 대한 추정인지 아니면 예측치의 분산의 최소화인지를 구분하여 선택될 수 있을 것이다.

참 고 문 헌

[1] Atwood, C. L. (1969), "Optimal and Efficient Design of Experiments," *Annals of Mathematical Statistics*, **40**, 1570-1602.

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

[3] Kim, Y. I., and Nachtsheim, C. J. (1991), "Transformation-Robust Experimental Design with Application to Some Problems in Chemistry," *Chemometrics and Intelligent Laboratory Systems*, **10**, 261-270.

[4] Läuter, E. (1974), "Experimental Design in a Class of Models," *Mathematishe Operations Forschung Und Statistik*, **5**, 379-398.

[5] Wong, W. K., and Cook, R. D. (1992), "Heteroscedastic G-optimal Designs," Technical Manuscript, Dept.of Bio-Statistics, University of California Los Angeles.

Error-Robust Experimental Designs: D- and Heteroscedastic G-optimalities.

Young-Il Kim[1)]

Abstract

In this paper we have defined two approaches to be error-robust when the precise form of error-structure is unknown. An experiment is optimal by the first criterion if it maximizes the minimum efficiency over all candidates of error structure and is optimal by the second if it maximizes the minimum average of the efficiency over all candidates of error structure. In order to appreciate the basic implications of each design criterion, these approaches are applied to two different experimental situations, D- and heteroscedastic G-optimalities.

1) Department of Industrial Information, ChungAng University, KyungGi-Do, AhnSung-Koon, DaeDuk-Myeon, NaeRee, San 40-1, 456-756