The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 895-907

A Comparison of the Efficiency of Location Estimators in Bivariate t distribution[1)]

Byong Su Choi[2)] and Seung-Chun Lee[3)]

Abstract

Recent demands for representing the location of multivariate data produce various multivariate medians such as Tukey median, Oja median and spatial median. They are considered as multivariate versions of the median which is widely recognized as a robust alternative to the arithmetic mean.

Many studies show that those multivariate median preserve the robustness. However, the effectiveness of those medians is not fully identified. In this note the relative efficiencies of the multivariate medians are investigated in various configurations under the bivariate t-distribution. It is shown that Tukey median outperforms the others in most configurations.

Keywords: location, multivariate median, robustness, efficiency, bivariate t-distribution

1. 서론

자료의 탐색을 위해 많이 이용되는 통계량은 백분위수이다. 특히 중위수는 위치모수의 로버스트 추정량으로 상자그림 등 자료처리의 여러 분야에서 매우 많이 활용되고 있다. 그러나 중위수는 주로 일변량 자료의 위치를 나타내기 위해 사용되어 왔을 뿐 다차원 자료의 표현을 위한 중위수의 활용은 미비한 것이 사실이었다. 즉, 2차원 이상에서는 중위수의 정의 자체가 어려울 뿐 아니라 자료에서 다변량 중위수를 구하는 것도 쉬운 문제가 아니기 때문에 이용이 제한적일 수밖에 없었다. 그러나 최근 다차원 자료의 표현을 위한 여러 가지 통계그래픽이 연구되어 왔고, 이로 인하여 다차원 자료의 위치를 표현할 수 있는 통계량의 필요성이 인식되어 다변량 위치모수 추정량, 특히 다변량 중위수에 대한 연구가 심도있게 다루어지게 되었다.

Bickel(1964)은 다변량 위치모수 추정량으로 각 변수 별로 구한 표본평균, 중위수, Hodge-Lehmann 추정량의 효율성 비교하였다. 이러한 연구는 다변량 위치 추정량들의 로버스트성과 효율성에 대한 초기의 연구로 아직 일변량 위치모수 추정량의 일반형이라고 할 수 있는 다변량 위치모수 추정량에 대한 정의가 일반화되지 않았다는 것을 보여준다고 하겠다. 70년대 이후에는

1) This Research was financially supported by Hansung University in the year of 2003.
2) Professor, Division of Computer Engineering, Hansung University, Samsun-Dong, 2-389, seoul, Korea 137-792. E-mail : cbs@hansung.ac.kr
3) Professor, Department of Statistics, Hanshin University, Osan, Yangsan-dong, Korea, 447-791. E-mail : seung@hanshin.ac.kr

spatial 중위수 (Gower, 1974), Tukey 또는 halfspace 중위수 (Tukey, 1975), Oja 중위수 (Oja, 1983), Liu 중위수 (Liu, 1990) 등 다변량 위치 추정량들이 정의되었고, 몬테칼로 연구에 의해 주로 평균과 비교된 추정량들의 효율성 및 로버스트성에 대한 연구 결과가 발표되었다. 이러한 연구 결과는 Steiger와 Wigderson (1992), Hettmansperger와 McKean (1998), Massé와 Plante (2001) 등에서 찾아 볼 수 있다.

Massé와 Plante는 다변량 위치추정량의 효율성을 비교하기 위하여 이변량 정규분포, 이변량 t 분포 등 5 개의 이변량 분포에서 다변량 위치 추정량의 효율성을 비교하였다. 그들의 연구 결과는 다양한 이변량 분포에서 추정량들의 효율성을 몬테칼로 연구에 의해 비교하였지만, 연구에 사용된 이변량분포는 모두 독립 이변량 분포이었다. 그러나 다변량 자료들은 대개의 경우 상관관계를 갖는 것이 일반적이므로 그들의 연구 결과를 그대로 인용하기에는 무리가 있다고 하겠다. 즉, 변수들 간의 상관관계는 다변량 중위수에 영향을 미치기 때문에 상관관계의 변화에 따른 추정량들의 효율성을 비교할 필요가 있다고 판단된다.

본 연구에서는 이변량 t 분포를 이용하여 상기된 Tukey. Oja, Spatial의 다변량 중위수와 표본평균의 효율성을 몬테칼로 연구로 통해 비교하기로 한다. 2 절에서는 본 연구에서 다루게 될 다변량 중위수들에 대해 기술하였고, 3 절에서는 Monte-Carlo study를 위한 실험계획 및 비교방법을 설명하였다. 또한 4 절에서는 실험의 결과 해석으로 다변량 위치 추정량들의 효율을 높이는 요인에 대한 결과를 해석하였다.

2. 다변량 위치 추정량

2.1. Tukey 중위수

일변량 중위수의 가장 직관적인 정의는 자료의 양 끝점을 계속 제거하는 정의로, 이 정의를 이용하는 *Convex hull peeling* 방법은 마지막 볼록집합(convex set)이 남을 때까지 최소볼록집합(convex hull)을 반복적으로 제거하는 방법이다. 그러나 이 방법은 극단적인 자료값에 영향을 받게 되므로 중위수로 사용하기에는 한계가 있었다.

Hotelling(1929)은 일변량 중위수를 한쪽 편에 위치한 자료의 최대수를 최소화하는 점이라고 정의하였고, 이러한 정의는 Tukey(1975)에 의해 고차원으로 일반화되었다. *Tukey* 중위수 또는 *반평면(halfspace) 중위수*는 Tukey에 의해 정의된 반평면 깊이(halfspace depth)에 의해 다변량 중위수를 구하게 되는데, 최근까지 이 반평면 깊이를 계산하기 위한 유용한 알고리즘이 없었기 때문에 그 사용은 제한적이었다. 그러나 Rousseeuw and Ruts (1996), Rousseeuw와 Ruts (1998), Rousseeuw and Struyf (1998)에 의해 고차원에서 정확한 또는 근사 알고리즘이 제시되어 최근에 가장 폭넓게 사용되고 있는 다변량 중위수이다.

Tukey에 의해 정의된 k차 닫힌 반평면(closed halfspace)은 임의 $x\in R^k$에 대해

$$H[x, u] = \{y\in R^k : u'y \geq u'x\}$$

와 같이 정의된다. 단 $u\in U$, $U=\{u\in R^k : |u|=1\}$이다. 여기서 $|\cdot|$는 유클리디안 놈을 나타낸다. 한편 다변량 분포함수 F에서 x의 반평면 깊이는

$$HD(x) = \inf_{u\in U} F(H[x, u])$$

와 같이 정의되는데 Rousseeuw와 Ruts (1999)에 의하면 HD의 상한(supremum)이 존재한다고 하였는데, Tukey는 이 값을 분포 F의 중위수로 정의하였다. 즉, Tukey의 중위수는

$$T(F) = \underset{x}{\text{argmax}}\ HD(x)$$

와 같이 정의된다.

표본 Tukey 중위수는 표본 분포함수(empirical distribution function) F_n라고 할 때, $T(F_n)$으로부터 정의된다. $T(F_n)$은 일반적으로 영역으로 표시될 수 있는데, 이 영역의 변수별 중앙값을 택하여 표본 Tukey 중위수로 정의된다.

표본에서 일변량 자료인 경우, 임의의 점 z의 Tukey의 깊이는 z의 오른쪽에 있는 자료의 수와 왼쪽에 있는 자료의 수 중 작은 값이 된다. 다변량 자료인 경우에는 자료를 모든 방향으로 투영시켜 나타나는 평면에 존재하는 자료들의 수 중 작은 값으로 정의된다. 특히 이변량인 경우에는 임의의 두 점 사이를 연결하여 나누어지는 두 평면에 존재하는 점의 개수 중 작은 값이 깊이가 되어 그림 1과 같이 나타낼 수 있다. 또, 그림에서와 같이 깊이가 k 인점을 연결하여 깊이가 k

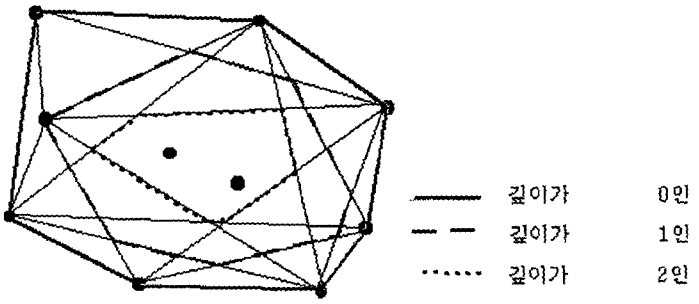

그림 1 Tukey 깊이가 같은 점을 연결한 등고선

이상이 되는 영역으로 등고선(contour)을 만들 수 있는데, 이 등고선 중 가장 안쪽에 있는 등고선은 Tukey 중위수 영역으로 정의된다.

이러한 깊이에 대하여 많은 연구가 있었지만 계산상의 어려움으로 인해 Tukey 중위수의 이용은 제한적이었다. 그러나 Ruts and Rousseeuw(1996)은 이변량에서 Tukey 깊이를 계산하는 알고리즘을 제시하였고, 그 후 그들의 일련의 논문에서 이를 다변량으로 확장하여 Tukey 깊이를 다양하게 이용하였다. 즉, 상자그림의 확장으로 Bag Plot (Rousseeuw, Ruts, Tukey, 1999)에도 적용시켰다.

2.2 Oja 중위수

Oja(1983)는 일변량에서 $E|X-x|$를 최소화시키는 $x \in R$ 가 중위수인 점에 착안하였다. 즉, $|X-x|$는 X와 x를 축으로 한 일차원 심플렉스의 길이이므로 이를 다변량으로 확장시키면 k차원에서의 심플렉스 $S(y_1, \ldots y_{k+1})$의 면적을 $\Delta(y_1, y_2, y_{k+1})$ 라고 할 때 $E(\Delta(X_1, \ldots, X_k, x))$를 최소화 시키는 $x \in R^k$가 Oja 중위수로 정의된다. 이를 수식으로 나타내면

$$O(F)=\underset{x}{\text{argmax}}\int \Delta(x_1,\ldots,x_k,x)dF^k(x_1,\ldots,x_k)$$

와 같이 정의된다.

이 식은 표본 분포함수 F_n에 의해

$$O(F_n)=\underset{x}{\text{argmax}}\ \frac{1}{\binom{n}{k}}\sum\Delta(X_{i_1},\ldots,X_{i_k},x) \quad \text{단 } 1\le i_1<\cdots<i_k\le n$$

와 같이 추정될 수 있다. 즉, 표본 Oja 중위수는 $O(F_n)$이 된다. Oja와 Niinimaa (1985)에 의하면 n이 짝수일 때, $O(F_n)$은 유일하다고 하다고 한다. 유일하지 않은 경우 $O(F_n)$의 집합은 볼록 다각형(convex polygon)이 된다고 하며, 이 경우 표본 Oja 중위수는 볼록 다각형의 중앙점으로 정의한다.

2.3 Spatial 중위수

L_1 중위수라고도 불리는 spatial 중위수는 고전적 정의에 의한 것으로

$$S(F)=\underset{x}{\text{argmax}}\int |x-y|dF(y)$$

으로 정의된다. 연속형 확률분포에서 spatial 중위수는 유일한 값을 갖는다. 그러나 앞에서 설명된 3 개의 다변량 중위수와는 달리 척도불변성(scale invariant)을 갖지 못한 것으로 알려져 있다. 한편 표본 spatial 중위수는

$$\frac{\sum_i |x-X_i|}{n}$$

를 최소화하는 x로 정의된다.

3. Monte Carlo 연구

본 연구는 위치모수 추정량으로서 2 절에서 열거된 여러 가지 표본 다변량 중위수들과 표본평균의 효율성을 비교하는 것을 목적으로 하고 있다. 로버스트 추정량으로서 표본 중위수들의 특성은 여러 학자들의 연구에 의해 이미 잘 알려져 있는 것으로 로버스트성에 대한 조사는 본 연구에서 제외하였다.

일변량인 경우 위치모수의 추정량으로 표본 중위수가 표본평균에 비해 효율성이 높은 경우는 분포의 꼬리가 두터운 분포라고 알려져 있다. 이 경우 분포의 모분산이 큰 값을 갖게 되어 표본평균의 효율성이 낮아지는 결과를 갖게 된다. 이에 반하여 중위수는 모분산의 크기와는 관계없이 분포 중앙의 밀도에만 의존하므로 상대적으로 표본평균에 비하여 높은 효율성을 갖을 수 있다.

t 분포는 자유도가 커짐에 따라 꼬리 부분이 아주 두터운 코쉬분포 분포로부터 그 꼬리는 점점 없어져 자유도가 무한대일 때인 꼬리가 거의 없는 정규분포로 변화하여 간다. 그러므로 본 연구에서는 t 분포를 선택하여 분포의 모양에 따라 다변량 중위수들의 효율성이 어떻게 변화하는지를 관찰하였다.

3.1 이변량 t 분포

자유도가 1인 일변량 t 분포는 코쉬분포로 꼬리가 매우 두터운 분포이다. 코쉬분포에서는 평균과 분산이 존재하지 않는 분포로 위치모수 추정량으로 표본평균은 극단적으로 낮은 효율성을 갖는다. 그러나 일변량 표본중위수의 효율성은 모분산과는 관계없이 분포 중위수의 밀도에만 의존하기 때문에 이 경우 표본평균보다 효율성이 좋은 것으로 알려져 있다. 즉, 중위수의 근사 분산은 f와 x_m을 각각 분포의 확률밀도함수와 중위수라고 하였을 때,

$$\frac{1}{[f(x_m)]^2}\frac{1}{4n}$$

로 주어지기 때문에 표본중위수의 효율은 중위수에서 분포의 밀도(density)와 매우 밀접한 관계가 있다.

t 분포의 분산은 자유도가 3 이상에서만 정의된다. 자유도가 1 또는 2인 경우 표본평균과 중위수의 효율성 비교는 무의미하다. 한편 자유도 3 이상에서 윗 식을 이용하여 구한 표본 중위수의 근사 분산과 표본평균의 분산을 비교하여 구한 상대효율은 표 1과 같다. 표에 의하면 자유도가 5

<표 1 : 일변량 t분포에서 표본평균에 대한 표본중위수의 상대효율>

자유도 / 상대효율	3	4	5	6	7	8	9	10
$\mathrm{Var}(X_m)/\mathrm{Var}(\overline{X})$	0.617	0.889	1.041	1.138	1.205	1.254	1.291	1.321

이상에서는 분산비로 정의된 표본중위수의 상대효율이 1 보다 커져 표본평균과 비교하여 낮은 효율을 갖는 것으로 나타났다. 2 절의 다변량 중위수들은 일변량 중위수들의 일반형이므로 2변량 t 분포에서도 비슷한 특징을 갖을 것으로 예상된다.

Johnson 과 Kotz (1972)에서 정의된 두 번째 유형의 t 분포는 두 변수의 상관관계를 정규분포의 상관계수를 이용하여 정의되므로 상관관계에 대한 이해가 상대적으로 쉽다. 그런 이유로 모의실험에는 두 번째 유형의 t 분포를 사용하기로 한다. 따라서 모의실험에 필요한 이변량 t 분포의 확률난수를 만들기 위한 알고리즘은 다음과 같다.

1. 상관계수 r을 갖는 표준정규난수 Z_1과 Z_2를 만든다.
2. 자유도 f_1과 f_2를 갖는 카이제곱 확률난수 S_1와 S_2를 만든다.
3. 2변량 t 분포를 따르는 확률변수를 다음과 같이 만든다.

$$X_1 = Z_1/\sqrt{(S_1/f_1)}$$
$$X_2 = Z_2/\sqrt{(S_2/f_2)}$$

3.2 실험방법 및 효율성 측도

본 연구에서는 이변량 t 분포를 중심으로 실험을 하였고, 독립적인 형태를 포함하여 상관관계가 있는 자료에 대해서도 조사를 하였다. 즉, 모의실험의 다음과 같은 총 1296개의 조건에서 1000회의 반복을 통하여 필요한 통계값을 구하였다.

표본크기(n) : 20, 30, 50, 100
반복수 : 1000회
자유도1(df_1) : 1, 2, 3, 5, 10, 15, 20, 30, 50
자유도2(df_2) : 1, 2, 3, 5, 10, 15, 20, 30, 50
상관계수(r) : 0, 0.25, 0.5, 0.75

Rousseeuw and Ruts(1998)는 Tukey 깊이를 계산하는 알고리즘을 제안하였고, Rousseeuw and Ruts(1998)에서 Tukey 중위수를 구하는 알고리즘이 제시되었으므로 모의실험과정에서 필요한 표본 Tukey 중위수의 계산은 이들의 알고리즘을 따르기로 한다. 또 Oja 중위수는 AS 277 (Niinimaa과 Oja, 1992)을 이용하였고, Spatial 중위수는 AS 143 (Rousseeuw와 Ruts, 1996)을 이용하였다.

추정량의 효율성은 보통 평방평균오차 행렬(Mean Squared Error Matrix)에 의해 평가되어야 하므로 반복 실험을 통하여 평방평균오차 행렬을 추정하였다. 즉, 위치모수 θ의 추정량, T의 평방평균오차는 $\mathrm{MSE}(T)=\mathrm{E}(T-\theta)(T-\theta)'$ 으로 정의되는데 실험에 사용된 분포의 위치모수는 모두 $\mathbf{0}$이므로 이는 $\mathrm{MSE}(T)=\mathrm{E}(TT')$과 같이 나타낼 수 있다. T_i를 i번째 반복에서 구한 T의 값이라면 평방평균오차행렬은

$$\widehat{MSE}(T)=\frac{1}{1000}\sum T_i T_i'$$

에 의해 추정될 수 있다. 분산공분산행렬의 행렬식은 일반화 분산으로 다변량 확률변수의 분산크기를 나타내는 하나의 지표로 사용되므로, 여기에서도 추정량 T의 효율성은 $|\widehat{\mathrm{MSE}}|^{1/2}$으로 측정하였다.

4. 실험결과 및 해석

표본평균, 표본 Oja 중위수, 표본 Spatial 중위수, 표본 Tukey 중위수에 대하여 Monte-Carlo 연구를 수행한 결과의 일부는 표 2와 같다. 그러나 연구에 사용된 조건들은 총 1296개로 표를 이용할 경우, 결과 해석이 매우 어려워져 부득이 실험결과를 부록에 수록된 그림과 같이 나타내기로 하였다. Rousseeuw (1998) 등에서 Tukey 중위수는 매우 로버스트한 위치 추정량으로 효율성이 높다고 하였으므로 그림에서는 Tukey 중위수에 대한 각 추정량들의 상대효율로 표시하였다.

<표 2: $n=30, r=0.5, df_1=5$에 대한 결과표>

n	rho	df_1	df_2	Mean	Oja	Spatial	Tukey
30	0.5	5	1	70.0066	0.0686	0.8610	0.0682
30	0.5	5	2	0.1531	0.0560	0.0561	0.0570
30	0.5	5	3	0.0689	0.0504	0.0513	0.0509
30	0.5	5	5	0.0511	0.0462	0.0464	0.0468
30	0.5	5	10	0.0418	0.0422	0.0422	0.0424
30	0.5	5	15	0.0428	0.0422	0.0431	0.0418
30	0.5	5	20	0.0411	0.0423	0.0418	0.0420
30	0.5	5	30	0.0391	0.0413	0.0415	0.0414
30	0.5	5	50	0.0418	0.0442	0.0441	0.0437

각 그림은 표본크기 n, 상관계수 r, 이변량 t 분포의 첫 번째 df_1이 주어졌을 때, 두 번째 장유도 df_2가 변화할 때 4가지 추정량들의 상대효율을 표시하여 결과를 이해하기 쉽도록 하였다. 부록에 수록된 그림을 살펴보면, 그림 A.1에서는 표본크기가 20이고 상관계수의 값이 0.25일 때 Spatial 중위수의 효율성은 Tukey 중위수와 비교하여 비슷하거나 더 효율적이라는 것을 알 수 있었고, 그림 A.2에서는 표본크기가 30이고 상관계수의 값이 0.75일 때 Tukey 중위수가 안정적이고 Oja 중위수도 이와 비슷한 경향을 보이고 있었다. 또 그림 A.3에서는 표본크기가 50이고 상관계수의 값이 0.75일 때, Oja 중위수는 Tukey 중위수와 비슷하게 나타났지만 한쪽의 자유도가 1인 경우에는 극히 효율성을 잃고 있다. 한편 그림 A.4에서 표본크기가 100이고 상관계수가 0.25일 때 모든 중위수가 비슷하게 나타나지만 한쪽의 자유도가 1일 때는 Tukey 중위수가 효율적인 것으로 나타났다.

이상의 결과를 요약하면 다음과 같은 유용한 정보를 얻을 수 있었다.

1. 각각의 자유도가 모두 5이상인 경우 평균은 효율적인 추정량이지만, 어느 한쪽의 자유도라도 5이하인 경우 표본평균은 표본 다변량 중위수보다 낮은 효율성을 갖는다.

2. Tukey 중위수는 거의 모든 경우에 다른 다변량 중위수보다 안정적으로 나타났다.

3. 상관계수의 값이 클 때에는 Oja 중위수가 효율적이지만, 작은 경우에는 Spatial 중위수가 효율적이다.

4. 표본크기가 30이하이고 상관계수의 값이 크지 않을 때는 spatial 중위수는 Tukey 중위수와 비슷하거나 더 효율적이다.

5. 결론

본 연구에서는 2차원 자료에 대한 위치추정량의 효율성을 몬테칼로 연구를 통해 비교하였다. 표본평균은 대부분의 경우 효율적인 추정량이지만 어느 한쪽의 자유도라도 5이하인 경우 그 효율성을 잃는 것으로 나타났다. 이 결과는 일변량에서의 얻어진 결과와도 일치하는 것으로, 다변량에서도 분포의 꼬리가 두터우면 다변량 중위수들이 표본평균 보다 효율적이라는 것을 알 수 있었다.

다변량 중위수들 중에서 Tukey 중위수는 상대적으로 많은 계산 시간이 걸렸다. 그러나 Tukey 중위수는 거의 모든 경우에 안정적인 효율성을 보이고 있어, 계산시간이 중요한 요인이 되는 동적 그래픽을 제외한다면 다변량 자료의 위치는 Tukey 중위수를 사용하는 것이 바람직한 것으로 보인다.

본 연구의 제약점은 모의실험이 대칭분포에서만 실행되었다는 것이다. 일반적으로 중위수는 비대칭분포에서 평균보다 상대적으로 좋은 위치모수가 될 수 있으므로 중위수의 특성을 충분히 조사하려면 비대칭분포에서도 효율성 비교를 하여야 하지만, 비대칭분포에서는 다변량 분포의 중위수가 정의에 따라 다를 수 있으므로 표본 다변량 중위수는 서로 다른 위치모수를 추정하게 된다. 따라서 비대칭분포에서의 효율성 비교는 사실상 의미가 없을 수 있다.

참고문헌

[1] Bickel, P. J. (1964). On some alternative estimates for shift in the p-variate one sample problem, Annals of Mathematical Statistics, vol 35, 1079-1090.

[2] Gower, J. C. (1974). The mediancenter, Applied Statistics vol. 32, 466-470.

[3] Hettmansperger, T. P. and McKean, J. W. (1998). Robust nonparametric statistical methods, Arnold, London.

[4] Johnson, N. and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions, John Wiley & Sons, Inc. New York.

[5] Liu, R. (1990). On a notion of data depth based on random simplices, Annals of Statistics, vol. 18, 405-414.

[6] Massé, J. and Plante, J. (2001). A Monte-carlo study of the accuracy and robustness of ten bivariate location estimators., InterStat, 2001-Jan-#2.

[7] Niinimaa, G. and Oja, H.(1992). Algorithm AS 277: The Oja bivariate median, Applied Statistics, vol. 41, 611-617.

[8] Oja, H(1983). Descriptive statistics for multivariate distributions, Statistcal probability Letters vol 1. 327-332.

[9] Rousseeuw, P. J. and Ruts, I. (1996). AS 307 : Bivarite location depth, Applied Statistics. vol. 45, 516-526.

[10] Rousseeuw, P. J. and Ruts, I. (1998). Constructing the bivariate Tukey median, Statistica Sinica. vol. 8 213-244.

[11] Rousseeuw, P. J. and Ruts, I. and Tukey, J.W.(1999), The Bagplot : A Bivariate Boxplot,

The American Statistician, vol. 53, 382-387.

[12] Rousseeuw, P. J. and Struyf, A (1998). Computing location depth and regression depth in higher dimension, Statistics and Computing vol. 8, 193-203.

[13] Steiger, G. and Wigderson (1992). A. Geometric medians, Discrete medians, Discrete Mathematics vol. 108, 37-51.

[14] Tukey, J. W. (1975). Mathematics and picturing data, In Proceedings of 1975 International Congress of Mathematics vol. 2, 523-531.

[2003년 9월 접수, 2003년 11월 채택]

부록

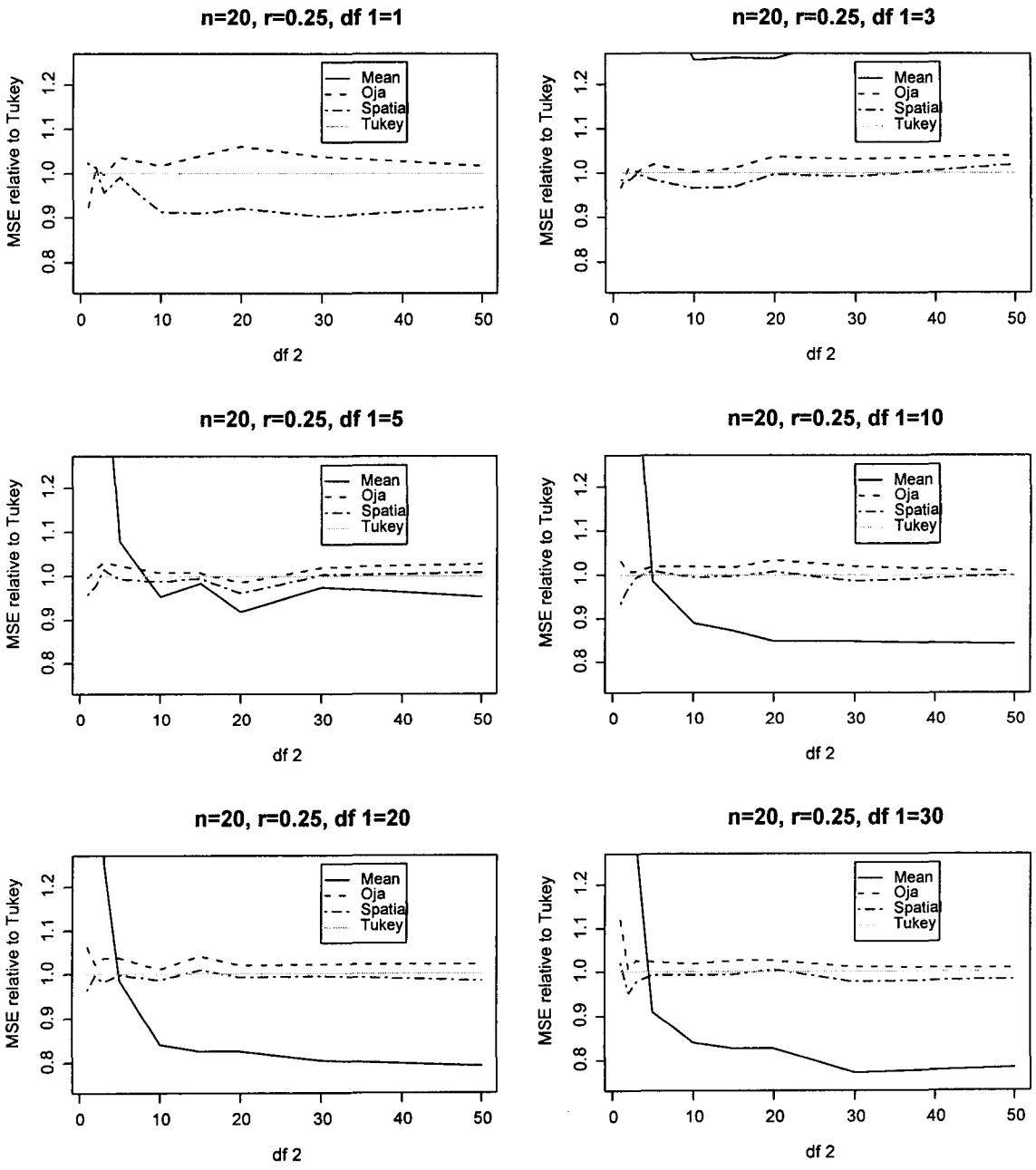

그림 A.1: n=20, r=0.25에 대한 결과

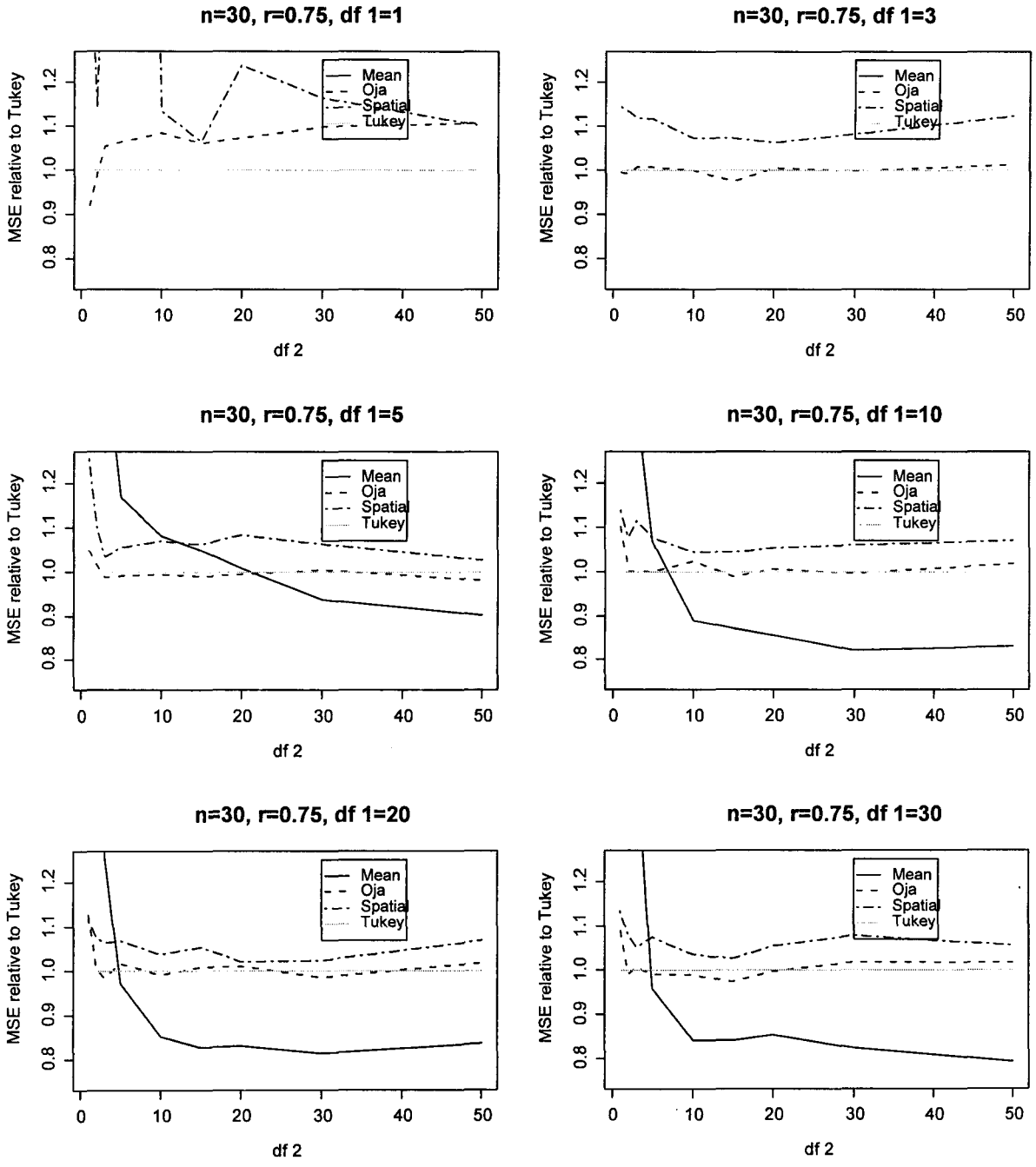

그림 A.2: n=30, r=0.75에 대한 결과

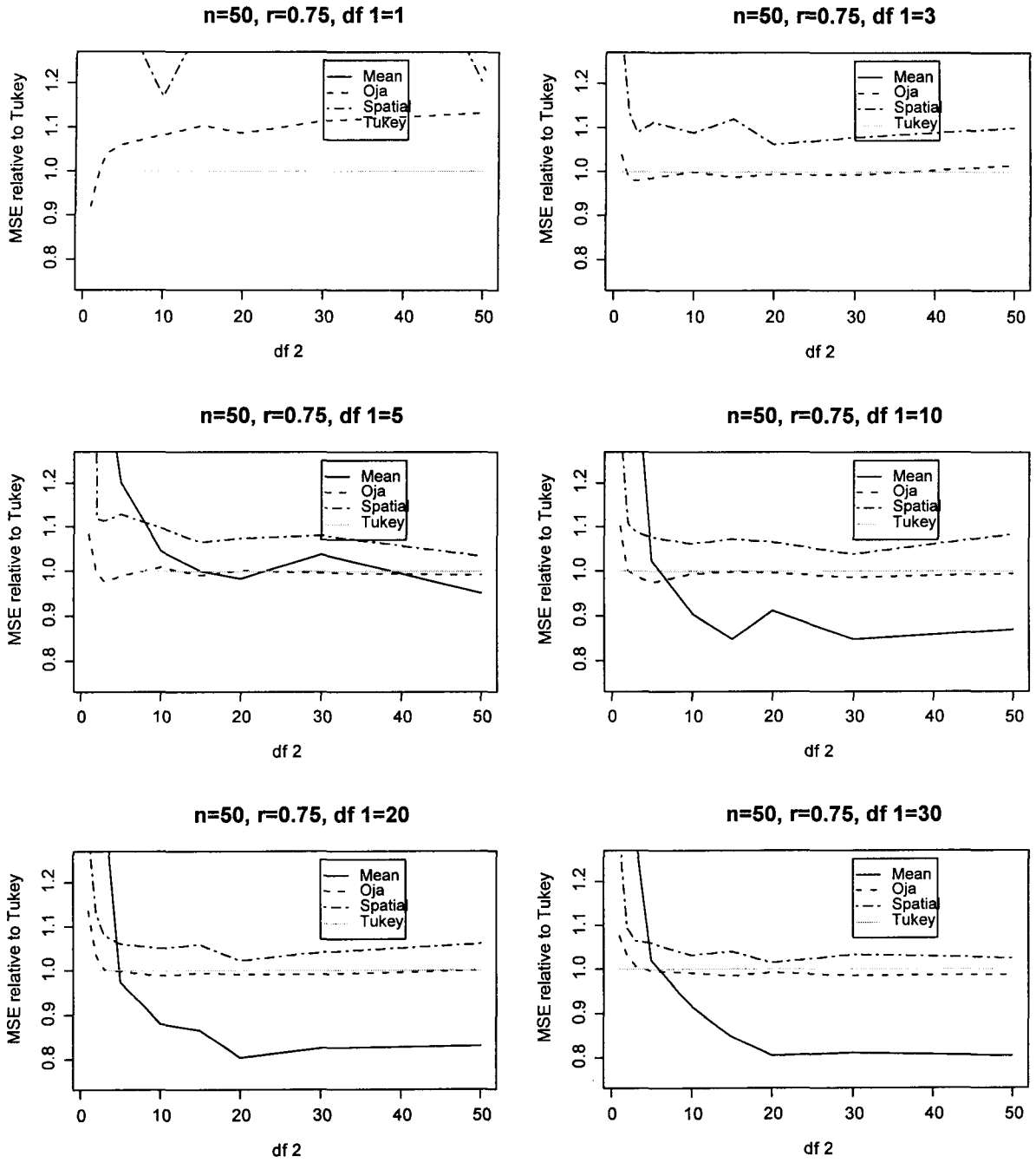

그림 A.3: n=50, r=0.75에 대한 결과

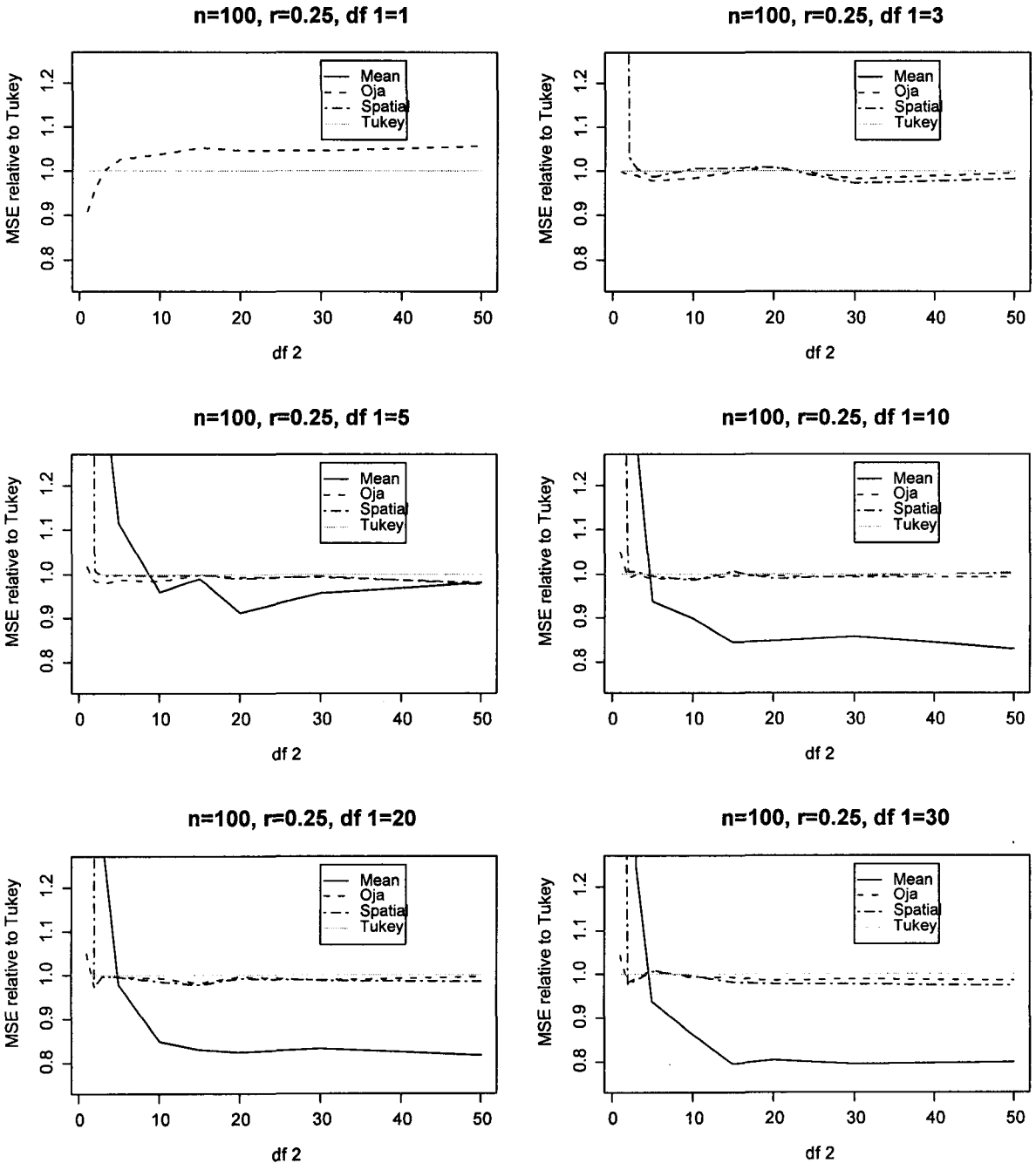

그림 A.4: n=100, r=0.25에 대한 결과