Robust Discriminant Analysis using Minimum Disparity Estimators

조미정[1], 홍종선[2], 정동빈[3]

요약

Lindsay and Basu (1994)에 의해 소개된 최소차이추정량 (Minimum Disparity Estimators)들은 실제 자료 분석 도구로써 유용하다. 본 논문에서는 최소일반화음지수 차이추정량 (Minimum Generalized Negative Exponential Disparity Estimator, MGNEDE)이 최대가능도추정량 (Maximum Likelihood Estimator, MLE)와 최소가중 헬링거거리추정량 (Minimum Blended Weight Hellinger Distance Estimator, MBWHDE)에 비해 오염된 정규모형에서 효율적이고 로버스트하다는 것을 모의실험을 통하여 확인하였다. 또한 세 가지 추정량들에 의해 추정된 모수들을 이용하여 판별하였을 때 각 추정량들의 판별율을 비교함으로써 오염된 정규모형에서 MLE의 대안으로 MGNEDE와 MBWHDE를 사용할 수 있음을 보였다.

주요용어 : 최소차이추정량, 효율성, 로버스트성, 판별분석

1. 서론

지난 오랜 동안 모수추정에 대해 통계학자들이 근본적으로 중요하게 다룬 두 가지 논제가 있었다. 하나는 모형이 적절하게 주어져 있을 때 추정량의 효율성에 관한 것이고, 나머지는 참모형으로부터 벗어나 있을 때 추정량의 로버스트성에 관한 것이다. 두 논제가 개념상에 적지 않은 논란이 있는데 최대가능도추정량(Maximum Likelihood Estimator, MLE)는 추정량들 사이에서 점근효율이 가장 높은 반면에 좋은 로버스트 성질을 갖는 추정량은 될 수 없으나, 로버스트 M-추정량족은 모형에서 일차효율성을 희생하면서 로버스트성을 만족하였다(Hampel 외 1986).

효율성과 로버스트성 사이의 논란은 헬링거거리(Hellinger distance, HD)와 같은 밀도함수에 근거한 최소차이추정량(Minimum Disparity Estimator, MDE)과 이에 관련된 연구(Beran, 1977; Stather, 1981; Tamura와 Boos, 1986; Simpson, 1987, 1989; Lindsay, 1994; Basu와 Lindsay, 1994)에 의해 적어도 부분적으로 조정되었다. 이에 관한 첫 시도로써 Beran(1977)은 최소헬링거거리추정량(Minimum Hellinger Distance Estimator, MHDE)이 이차효율성과 로버스트성을 동시에 만족할 수 있다는 사실을 발견하였다. 주어진 모형 하에서 점근효율은 추정량이 MLE와 동일한 영향함수를 갖는 것을 의미한다. 앞에서 소개한 몇 개의 연구들은 MHDE가 이런 점에도 불구하고 로버스트 성질을 만족함을 나타낸다. Simpson(1987)은 포아송분포와 같은 몇몇의 이산분포에서 MHDE가 50% 붕괴(breakdown)가 있음을 보였다. Lindsay(1994)는 이산모형에서

1) Senior Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail : mjcho@skku.edu

2) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea

3) Associate Professor, Department of Information Statistics, Kangnung National University, Kangnung 210-702, Korea

소위 잔차조정함수(Residual Adjustment Function, RAF)라 부르는 거리의 함수를 특징화시켜 최소거리추정량에 근거한 최소차이추정량족에 대한 붕괴점을 일반화시킨 결과를 도출시켰다.

Lindsay(1994) 그리고 Basu와 Sarkar(1994)는 수정된 차이측도로써 가중헬링거거리(Blended Weight Hellinger Distance, BWHD)를 제시하였고 이를 최소화시켜 얻은 추정량을 최소가중헬링거거리추정량(MBWHDE)이라 하였다. 또한 Jeong과 Sarkar(2000)에 의해 제시된 일반화음지수차이(Generalized Negative Exponential Disparity, GNED)측도를 최소화함으로써 최소일반화음지수차이추정량(MGNEDE)을 얻게 된다.

본 논문에서는 오염된 정규모형에서 MGNEDE라 부르는 또 다른 밀도함수에 근거한 추정량이 MLE와 MBWHDE에 비해 어느 정도 효율성을 만족하는지 모의실험을 통해 알아보고 그 추정된 결과를 판별분석에 활용하는데 목적이 있다. 또한 2절에서는 GNED와 BWHD를 포함하는 차이측도(Disparity Measure, DM)로부터 MGNEDE와 MBWHDE를 도출하는 방법에 대해 간략히 살펴본다. 그리고 3절에서는 정규분포의 위치모수에 대한 효율성을 세 개의 추정량(MLE, MGNEDE, MBWHDE)들에 대해 살펴보기 위해서, 고려할 수 있는 여러 형태의 정규모형을 이에 적용시켰다. 끝으로 4절에서는 각 모의실험에서 얻은 결과를 구체적으로 나열하였으며, 5절은 4절에서 얻은 여러 결과에 근거한 전반적인 결론을 내린다.

2. 차이측도를 이용한 모수추정

모수분포족 $\Im_\Theta = \{F_\theta, \theta \in \Theta\}, \Theta \subset R^p$ 로부터 확률표본 $\{X_1, X_2, \cdots, X_n\}$ 을 추출하자. 분포들 F_θ 는 연속이며, 이에 대응하는 확률분포함수들을 f_θ 라고 하자. 이 때 최소차이추정값들에 근거한 밀도함수(Lindsay 1994; Basu와 Lindsay 1994)는 비모수적 밀도추정값과 모형밀도 f_θ 사이의 양의 값을 갖는 차이측도 ρ_G 를 최소화시켜 계산한다. 그리고 차이측도는 다음과 같이 정의된다.

$$\rho_G(f^*, f_\theta) \equiv \int G(\delta(f^*, \theta, x))dF_\theta(x) \tag{2.1}$$

여기에서 $\delta(x) = \delta(f^*, \theta, x) = \left(f^*(x) - f_\theta(x)\right)/f_\theta(x)$ 는 x 값에서 Pearson 잔차를 나타낸다. 그리고 G는 실수값을 갖고 세 번 미분가능하며, 구간 $[-1, \infty)$ 에서 $G(0) = 0$ 인 단조볼록함수이고 $\delta = 0$ 일 때에만 등호가 성립하고 항상 $G(\delta) \geq 0$ 이다. F는 경험적인 분포함수이고, w 는 평균이 y 이고 표준편차가 h 인 정규분포와 같은 핵함수의 평활족이라고 할 때, 다음 정의한 비모수적 핵밀도추정량 $f^*(x)$ 를 사용할 수 있다.

$$f^*(x) \equiv \int w(x; y, h)dF(y) \tag{2.2}$$

함수 G의 형태에 따라 다양한 차이측도가 생성된다. 즉, $G(\delta) = (\delta+1)\log(\delta+1) - \delta$ 는 가능도차이, $G(\delta) = \delta^2$ 은 Pearson의 카이제곱, $G(\delta) = \left\{(\delta+1)^{\lambda+1} - 1\right\}/\lambda(\lambda+1)$ 은 멱발산족(Cressie와 Read, 1984), $G(\delta) = 2\left\{(\delta+1)^{1/2} - 1\right\}^2$ 은 두 배의 제곱된 헬링거거리 그리고 $G(\delta) = e^{-\delta} - 2$ 는 음지수차이(NED)를 생성한다.

∇ 을 θ 에 대한 경사도라고 할 때, 차이측도 $\rho_G(f^*, f_\theta)$ 를 최소화함으로써 G에 대응하는 최소차이추정량을 얻을 수 있다. 모형의 미분가능 하에서 최소차이추정방정식은 다음과 같다.

$$-\nabla\rho_G = \int A(\delta(x))\nabla F_\theta(x) = 0 \tag{2.3}$$

여기에서 $A(\delta) \equiv (\delta+1)G'(\delta) - G(\delta)$ 이고 $G'(\delta)$ 는 $G(\delta)$ 의 일차미분계수이다. 함수 $A(\delta)$ 는 구간 $[-1, \infty)$ 에서 증가함수이며, 차이측도에 의해 생성된 추정값을 변화시키지 않고 표준화시킬 수 있다. 따라서 표준화된 $A(\delta)$ 에 대해 $A(0) = 0$ 과 $A'(0) = 1$ 을 얻게 된다. 이 때 표준화된 함수를 차이의 잔차조정함수(RAF)라 부르며, 추정량의 대부분의 이론적인 성질들을 결정하게

된다. 자료가 인라이어(inliers) 또는 이상값(outliers)들에 대해 헬링거거리와 음지수차이가 어떻게 반응하는지는 Lindsay(1994), Basu와 Lindsay(1994), Basu 외 2인(1997)을 참고하기 바란다. 헬링거거리와 음지수차이에 의해 생성된 추정량들은 모두 일차 효율적이지만, 이차 효율적인 것은 음지수차이추정량뿐이다.

Lindsay(1994)는 보다 로버스트한 추정량으로서 수정된 차이측도들을 제시하였다. 다음의 식 (2.4)는 가중헬링거거리측도이다.

$$\mathrm{BWHD}(\lambda) = \int \frac{\left[f^*(x) - f_\theta(x) \right]^2}{2\left[\lambda \sqrt{f^*(x)} + \bar{\lambda} \sqrt{f_\theta(x)}\right]^2} dx \tag{2.4}$$

여기서 $\lambda \in R$이고 $\bar{\lambda} = 1 - \lambda$ 이다. Basu와 Sarkar(1994)는 $\lambda = 1/9$ 일 때 즉, BWHD(1/9)를 최소화시켜 얻은 추정량(MBWHDE)이 멱발산족과 같이 다른 추정량의 비교에서 우수함을 보였다.

일반화음지수차이측도를 다음 식 (2.5)와 같이 정의할 수 있다.

$$\mathrm{GNED}(\lambda) = \int \frac{\exp\left\{-\lambda\left(\frac{f^*(x)}{f_\theta(x)} - 1\right)\right\} - 1 + \lambda\left(\frac{f^*(x)}{f_\theta(x)} - 1\right)}{\lambda^2} \cdot f_\theta(x)\, dx \tag{2.5}$$

Jeong과 Sarkar(2000)는 $\lambda = 4/3$ 일 때 즉, GNED(4/3)을 최소로 하는 추정량(MGNEDE)이 멱발산추정량보다 우수할 수 있음을 제안하였다.

따라서 본 연구에서는 식 (2.4)와 (2.5)에서 정의된 차이측도 BWHD(1/9)와 GNED(4/3)을 최소화함으로써 모수를 추정하고, 그 결과를 바탕으로 오염된 자료집단에서 세 가지 추정량 MLE, MBWHDE 그리고 MGNEDE를 이용하여 판별하였을 때 각 추정량들의 오분류율을 비교하였다.

3. 모의실험 연구

본 절에서는 정규분포 설정 하에 GNED(4/3), BWHD(1/9) 그리고 MLE에 대한 이론적 결과들을 경험적으로 조사하기 위하여 몬테칼로 방법을 사용하기로 한다.

3.1 밀도추정

MGNEDE와 MBWHDE를 구하기 위해서 밀도추정에서 최적의 성질을 갖는 biweight 핵을 사용하여(Simpson, 1989) 핵밀도함수 f^*를 계산한다.

$$f^*(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \frac{15}{16} \left\{1 - \left(\frac{x - X_i}{h_n}\right)^2\right\}^2 I_{\left|\frac{x-X_i}{h_n}\right| < 1}\left(\frac{x - X_i}{h_n}\right) \tag{3.1}$$

Parzen(1962)은 핵밀도추정값 f^*과 참밀도함수 f_θ 사이의 적분평균제곱오차를 최소화시키는 평활계수 h 값을 유도하였다. Devroye와 Gyorfi(1985)의 biweight 핵에 관련된 평균 L_1 기준과 정규밀도함수 f_θ를 고려하면 평활계수는 $h_n = (2.34\sigma n^{-1/5})$이다. 반면에 σ을 모르는 경우는 $(\mathrm{median} \mid X_i - \mathrm{median}(X_i) \mid)/0.6745$ 를 대신 사용할 수 있다.

3.2 모수추정과 판별분석

모의실험을 위한 표본추출집단의 모형은 표본크기, 오염수준, 오염된 분포의 표준편차에 따라 그 형태가 다양하다. 그 다양성을 반영하기 위해 <표 3-1>에 주어진 모형에 대해서 고려하기

로 한다. 여기에서 N은 표본크기, ϵ은 오염수준 그리고 σ는 표준편차를 나타낸다.

<표 3-1> 모의분포의 유형

		X	Y
모형 1	분포	$\epsilon N(5,1)+(1-\epsilon)N(8,1)$	$\epsilon N(15,1)+(1-\epsilon)N(12,1)$
	N	40, 60, 100	40, 60, 100
	ϵ	0.05, 0.1	0.05, 0.1
	σ	1	1
모형 2	분포	$\epsilon N(5,\sigma^2)+(1-\epsilon)N(8,1)$	$\epsilon N(15,\sigma^2)+(1-\epsilon)N(12,1)$
	N	100	100
	ϵ	0, 0.05, 0.1, 0.15	0, 0.05, 0.1, 0.15
	σ	1, 2, 3	1, 2, 3
모형 3	분포	$\epsilon N(8,\sigma^2)+(1-\epsilon)N(8,1)$	$\epsilon N(12,\sigma^2)+(1-\epsilon)N(12,1)$
	N	100	100
	ϵ	0, 0.05, 0.1, 0.15	0, 0.05, 0.1, 0.15
	σ	3	3

2절에서 제시된 GNED(4/3)과 BWHD(1/9)를 최소화하여 두 모수추정량 MGNEDE와 MBWHDE를 구하고 판별분석하는 과정은 다음과 같다.

먼저 <표 3-1>에 제시된 모의분포의 모형에 따라 표본크기, 오염수준 그리고 표준편차 등을 조합하여 자료 X는 소속집단을 1집단, Y는 2집단으로 모형구축자료를 생성한다. 생성된 자료로부터 모수 $\theta=(\mu,\sigma^2)'$을 추정하기 위해 뉴턴-랩슨 알고리즘을 이용한다. 여기에서 초기값은 $\tilde{\mu}^{(0)}=\text{median}(X_i)$ 와 $\tilde{\sigma}^{(0)}=(\text{median}\mid X_i-\text{median}(X_i)\mid)/0.6745$ 로 정한다. 최적의 모수를 구하기 위하여 뉴턴-랩슨 알고리즘을 반복실행하며 다음의 정지규칙을 따른다.

$$\mid\tilde{\theta}^{(j+1)}-\tilde{\theta}^{(j)}\mid=\sqrt{\{\tilde{\mu}^{(j+1)}-\tilde{\mu}^{(j)}\}^2+\{(\tilde{\sigma}^{(j+1)})^2-(\tilde{\sigma}^{(j)})^2\}^2}<10^{-5}$$

이제 표준편차가 1인 정규모형에서 평균을 변화시키며 소속집단이 1집단 또는 2집단인 모형검정자료를 생성하고 피셔의 선형판별함수에 의하여 분류한다. 이 때 모형검정자료의 오분류율을 조사하여 세 추정량들의 판별율을 비교한다.

4. 모의실험 결과

4.1 모형에 대한 모수추정

모형 1에서 오염수준 0.05와 0.1에 대하여 표본크기 N = 20, 40, 60 그리고 100 에 따라 추정된 모수 μ_1,μ_2와 표준편차, 편의, MSE 그리고 Eff를 구하였다. 그 결과 표본크기에 따른 편의는 MLE와 MBWHDE가 거의 차이가 없으나 표본의 크기가 증가할수록 MGENDE는 다른 추정량들보다 커짐을 알 수 있었다. 그러나 표본의 크기가 40이상인 경우에 MSE를 살펴보면 MGENDE가 표본이 증가할수록 MLE와 MBWHDE 보다 MSE의 값이 작아지므로 MLE에 대한 상대효율이 MGNEDE가 높아지고 효율적인 추정량이라 할 수 있다. MGNEDE가 다른 추정량들보다 편의는 크지만 MSE가 작다는 것은 그만큼 추정량의 분산이 상대적으로 작다는 것을 의미한다. 따라서 적당한 크기를 갖는 오염된 자료에서 MGNEDE가 로버스트 추정량이 됨을 알 수 있다.

편의는 모형 1에서는 표본의 크기나 오염 수준에 관계없이 MGNEDE가 다른 추정량들 보다 조금 크게 나타났으나 모형 2에서는 표본수가 20 또는 40으로 작은 경우에 MGNEDE의 편의가

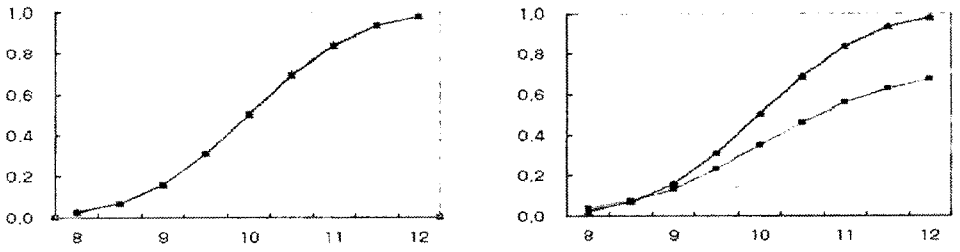
작게 나타남을 알 수 있다. 따라서 모형 2에서는 소표본인 경우 분산이 작고 편의도 작으므로 모형 1의 경우보다 MGNEDE의 MSE가 다른 추정량보다 더 작게 되며 따라서 상대적 효율성이 더욱 높아지게 된다.

모형 3의 경우는 오염수준이나 표본크기에 관계없이 차이측도의 하나인 MGNEDE가 추정값에 대한 분산과 편의가 모두 작으므로 다른 두 추정량 MLE와 MBWHDE 보다 MSE가 작다. 따라서 MLE에 대한 MGNEDE의 상대적 효율이 높다는 것이다. 또한 모형 1, 2와 다르게 모형 3과 같은 오염된 분포에 대하여 표본수가 작은 경우에도 MLE 보다 MGNEDE가 효율적이고 로버스트한 추정결과를 보여준다고 할 수 있다.

4.2 판별분석에의 활용

모형 1에서 MGNEDE는 표본 크기가 40 이상인 경우에 MLE와 대등한 결과의 분류율을 보였고, 표본의 크기가 100정도로 충분히 큰 경우는 MBWHDE도 고려해 볼 수 있다.

모형 2에서 오염수준이 0인 경우를 살펴보면 표준편차의 변화에 관계없이 세 가지 추정량의 분류율은 차이가 없는 것으로 나타났다. 특히 각 자료집단의 평균인 8과 12를 중심으로 그 사이에 존재하는 표본들의 분류율을 각 추정량과 오염수준에 수준에 따라 살펴보고 그래프로 나타낸 것이 <그림 4-1>이다. (a)는 오염수준이 0인 경우에 표준편차에 관계없이 공통적으로 적용되는 그래프이다. 8의 소속집단이 1집단이므로 2집단으로 분류될 확률은 0에 가깝고, 12는 소속집단이 2집단이므로 2집단으로 분류될 확률은 1에 가깝게 된다. (b)의 그래프는 오염수준이 0.05, 0.1 그리고 0.15에 공통적으로 적용되는 그래프이다. 여기서 MGNEDE와 MLE의 분류확률은 오염수준에 관계없이 모형 2의 경우에 거의 차이가 없음을 알 수 있다.



(a) $\sigma = 1,\ \epsilon = 0$ (b) $\sigma = 2, \epsilon = 0.05$

<그림 4-1> 국소구간의 판별율 (—●— MLE —■— MBWHDE —▲— MGNEDE)

판별분석을 위한 모형 3의 경우는 모수추정 결과가 다른 어떤 모형에 비해 효율성이 높고 오염수준과 표본의 크기에 관계없이 편의와 분산이 적었다. 오염정도에 관계없이 분류율이 MGNEDE와 MLE는 차이가 거의 없고 대등한 분류율을 갖는다고 할 수 있다.

5. 결론

본 논문에서는 정규모형의 위치모수에 대하여 차이측도 GNED와 BWHD를 사용하여 최소차이추정량 MGNEDE와 MBWHDE를 구하고, 모의실험을 통해 이 두 추정량과 최대가능도추정량간의 효율성을 비교하였다. 모형 3의 분포가 편의와 MSE가 모두 작고 이는 오염수준과 표본크기에 상관없는 결과이다. 그리고 모형 1인 경우에 추정된 모수는 MLE에 비해 편의가 조금 높게 나타나지만 그 대신 MGNEDE의 분산이 작아서 MSE가 MLE와 MBWHDE 보다 작게 되어 효율성이 높았다. 특히 오염수준 0.05와 0.1인 경우에 표본크기가 40이상의 자료집단에서의

추정된 모수는 MLE와 MBWHDE 보다 MGNEDE의 MSE가 작게 나타났다.

오염된 정규모형에서 추정된 위치모수를 이용한 판별율에 대한 실험에서 MGNEDE가 오염수준에 관계없이 MLE와 대등한 결과를 나타냈다. 특히 척도모수만이 다른 두 정규분포가 혼합되어 있는 모형 3에 대하여 이상값이 포함되지 않은 경우(즉, 오염수준=0)는 세 추정량의 분류율이 거의 같았고, 이상값이 포함되어 있는 경우에 MGNEDE가 MLE보다 양극에서 더 좋은 분류율을 보였다. 이와 같이 효율성과 분류율의 측면에서 자료에 이상값이 포함되어 있는지 여부에 관계없이 MGNEDE는 MLE의 대안으로 사용할 수 있다.

참고문헌

Basu, A. and Lindsay, B. G. (1994), Minimum disparity estimation for continuous models: efficiency, distributions and robustness, *Annals of the Institute of Statistical Mathematics*, Vol. 46, 683-705.

Basu, A. and Sarkar, S. (1994), The Trade-Off Between Robustness and Efficiency and The Effect of Model Smoothing in Minimum Disparity inference, *Journal of Statistical Computation and Simulation*, Vol. 50, 173-185.

Basu, A., Sarkar, S. and Vidyashankar, A. N. (1997), Minimum negative exponential disparity estimation in parametric models, *Journal of Statistical Planning and Inference*, Vol. 58, 349-370.

Beran, R. J. (1977), Minimum Hellinger distance estimates for parametric models, *Annals of Statistics*, Vol. 5, 445-463.

Cressie, N. and Read, T. (1984), Multinomial Goodness-of-fit Tests, *Journal of the Royal Statistical Society B*, Vol. 46, No. 3, 440-464.

Devroye, L. and Gyorfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, John Wiley, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley, New York.

Jeong, D. B. and Sarkar, S. (2000), Negative exponential disparity family based goodness-of-fit tests for multinomial models. *Journal of Statistical Computation and Simulation*, Vol. 65, 43-61.

Lindsay, B. G. (1994), Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *Annals of Statistics*, Vol. 22, 1081-1114.

Simpson, D. G. (1987), Minimum Hellinger distance estimation for the analysis of count data, *Journal of the American Statistical Association*, Vol. 82, 802-807.

Simpson, D. G. (1989), Hellinger Deviance Tests: Efficiency, Breakdown Points, and Examples, *Journal of the American Statistical Association*, Vol. 84, 107-113.

Stather, C. R. (1981), *Robust Statistical Inference using Hellinger Distance Methods*, Unpublished Ph. D. Dissertation, La Trobe University, Melbourne, Australia.

Tamura, R. N. and Boos, D. D. (1986), Minimum Hellinger Distance Estimation for Multivariate Location and Covariance, *Journal of the American Statistical Association*, Vol. 81, 223-229.