The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.3.397
27(3), 397–406

Minimum Density Power Divergence Estimation for Normal-Exponential Distribution

Ro Jin Pak[a,1]

[a]Department of Applied Statistics, Dankook University

(Received January 24, 2014; Revised March 12, 2014; Accepted April 9, 2014)

Abstract

The minimum density power divergence estimation has been a popular topic in the field of robust estimation for since Basu *et al.* (1988). The minimum density power divergence estimator has strong robustness properties with the little loss in asymptotic efficiency relative to the maximum likelihood estimator under model conditions. However, a limitation in applying this estimation method is the algebraic difficulty on an integral involved in an estimation function. This paper considers a minimum density power divergence estimation method with approximated divergence avoiding such difficulty. As an example, we consider the normal-exponential convolution model introduced by Bolstad (2004). The estimated divergence in this case is too complicated; consequently, a Laplace approximation is employed to obtain a manageable form. Simulations and an empirical study show that the minimum density power divergence estimators based on an approximated estimated divergence for the normal-exponential model perform adequately in terms of bias and efficiency.

Keywords: Efficiency, Laplace approximation, microarray, robustness.

1. 서론

마이크로어레이(microarray) 실험에서 관심 있는 유전자 혹은 단백질 시료의 발현값 만을 정확히 측정하기 위해 배경보정(background correction)이 통계적인 방법으로 수행되어야 한다 (Irizarry 등, 2003; Ritchie 등, 2007). 정규-지수밀도함수(normal-exponential density function; normexp)는 Bolstad (2004)에 의해 고안된 한 가지 밀도함수로서 배경보정(background correction)을 통계적으로 수행하기 위해 사용된다. Normexp는 기본적으로 정규밀도함수와 지수밀도함수의 합성곱(convolution)으로 정의되고 세 개의 모수를 갖는다. 자세한 형태를 제2장에 기술하겠다.

지금까지 normexp의 모수들을 추정하기 위해 최우추정법이 시도되었다 (Bolstad, 2004; McGee와 Chen, 2006). 유전자 발현 데이터를 생산해내기 위한 마이크로어레이 방법은 cDNA와 oligonucleotide를 장착한 Affymetrix GeneChip으로 대변된다. 본 논문 Affymetrix GeneChip에 의해 발현된 마이크로어레이 데이터를 분석하는 한 가지 방법인 정규분포와 지수분포의 합성곱에 기인한 방법에 초점을 맞추고 있다. 형광물질의 농도를 통해 측정되는 마이크로어레이 데이터는 다양한 잡음에 의해 영향을 받게 되고 참값을 얻기위해 배경잡음을 제거해야할 필요가 있다.

[1]Department of Applied Statistics, Dankook University, 152 Jukjeonro, Suji-Gu, Yongin 448-701, Korea.
E-mail: rjpak@dankook.ac.kr

Irizarry 등 (2003)이 robust multi-array average (RMA)라고 이름하는 마이크로어레이 분석 알고리즘은 (1) 배경잡음교정, (2) 정규화, 그리고 (3) 요약화(summarization)으로 나뉘는데 배경잡음제거 과정은 일종의 본격적 분석을 의한 사전 정비 단계라고 하겠다. 정규-지수밀도함수를 활용한 방법은 RMA 알고리즘의 첫 단계로 제시한 배경잡음교정의 과정의 한 부분이다. 이 방법론과 관련하여 Bolstad (2004), McGee와 Chen (2006)이 기본적인 추정방법으로 최우추정법을 제안하였다. 이후, Ritchie 등 (2007)은 안장점근사(saddle point approximation)을 통한 최우추정법이 일반적인 최우추정법 보다 계산상의 안정된 해를 얻을 수 있음을 보였다. 최근들어 Silver 등 (2009)이 수치해석적으로 Ritchie 등 (2007)의 방법 보다 정확한 계산이 가능한 최우추정법을 제안하였다.

본 논문에서는 normexp의 모수들에 대한 로버스트 추정을 시도하고자 한다. 시도하고자 하는 방법은 최소밀도함수승간격추정법(minimum density power divergence estimation; MDPDE)으로서 Basu 등 (1998)이 로버스트 추정을 위해 개발하였다. 관찰값이 참값과 배경잡음의 합으로 가정한 상태에서 참값을 추정하려는 과정은 모델이 불명확한(misspecification) 상황에서 로버스트 추정을 시행하는 과정과 닮았다고 하겠다. 추정법은 거리 혹은 간격을 이용한 기존의 최소헬링거거리추정법(minimum Helliger distance estimation) 혹은 최소절대값거리추정법(minimum absolute distance estimation)과 맥락을 같이 하면서 나름의 독특한 거리 혹은 간격을 사용한다. 이 추정법의 특별한 점은 밀도함수추정량(density estimator)에 수반되는 띠너비(bandwidth)를 찾아야하는 고충이 필요 없다는 것이다. 이러한 장점과 더불어 추정량이 로버스트하고 참모델 하에서 점근효율성(asymptotic efficiency)의 손실이 최우추정량(maximum likelihood estimator; MLE)에 비해 미미하다는 것이다.

그런데 MDPDE를 normexp에 적용하는 과정에서 적분이 대수적으로 쉽게 풀리지 않은 문제에 봉착하였다. 매우 수학적으로 능통한 연구자가 적분을 직접적으로 풀어 낼 수도 있겠지만, 본 논문에서는 라플라스근사(Laplace approximation)를 이용하여 간접적으로 문제를 해결하려 하였다. 근사법을 이용하면 추정량은 몇 가지 조건하에서 로버스트 특성을 갖고 있으며 효율성도 어느 정도 확보됨을 이론과 모의실험을 통해 확인하고 실제 자료에 적용하여 나름대로 가치가 있음을 보였다.

논문은 제 2장에서 MDPDE와 normexp에 대해 각각 간단히 기술하고 제 3장에서 MDPDE를 normexp에 적용했을 때의 이론적 결과와 실제적 결과를 정리하는 순서로 구성하겠다.

2. 최소밀도함수승간격추정법과 정규-지수밀도함수

2.1. 최소밀도함수승간격추정법

먼저 최소밀도함수승간격추정법에 대하여 간단하게 기술하겠다. 아래 내용은 Basu 등(1998)에 자세하게 소개되어있다. 밀도함수 g와 f의 간격 $d_\alpha(g,f)$를 다음과 같이 정의하자:

$$d_\alpha(g,f) = \int \left\{ f^{1+\alpha}(x) - \left(1+\frac{1}{\alpha}\right) g(x) f^\alpha(x) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} dx \quad (\alpha > 0). \tag{2.1}$$

만일 $\alpha = 0$, $d_0(g,f)$는

$$d_0(g,f) = \lim_{\alpha \to 0} d_\alpha(g,f) = \int g(x) \log \left\{ \frac{g(x)}{f(x)} \right\} dx$$

와 같이 정의된다. $d_0(g,f)$를 간혹 Kullback-Leiber 간격이라 부르고 α를 조절모수(tuning parameter)라고 한다.

미지의 모수(혹은 모수벡터) θ로 색인되어지는 밀도함수 f_θ를 갖는 확률분포함수족 $\{F_\theta\}$를 생각해보자. 랜덤표본 $X_1, \ldots, X_n$가 확률분포 G (G는 $\{F_\theta\}$에 속하지 않을 수 있음)를 따른다고 하고 g를 G의

밀도함수라고 하자. 식 (2.1)의 마지막 항은 모수와 상관이 없고 g를 경험적 분포(empirical distribution)로 대체하여 얻은 추정간격(estimated divergence),

$$\int f_\theta^{1+\alpha}(x)dx - \left(1+\frac{1}{\alpha}\right) n^{-1} \sum f_\theta^{\alpha}(X_i), \tag{2.2}$$

을 θ에 대하여 최소로 만드는 $\hat{\theta}$을 최소밀도함수승간격추정량(이하 MDPDE)이라고 정의한다. 그렇다면 식 (2.2)를 모수들에 대하여 미분하여 얻게된 추정함수(estimating equation)는

$$U_n(\theta) \equiv n^{-1} \sum u_\theta(X_i) f_\theta^{\alpha}(X_i) - \int u_\theta(x) f_\theta^{1+\alpha}(x)dx, \quad u_\theta(x) = \frac{\partial \log f_\theta(x)}{\partial \theta}$$

와 같은 형태를 갖는다. 여기서 $u_\theta(x)$는 점수함수(score function)를 의미한다.

사실, MDPDE는 일종의 M-추정량으로서

$$\psi(x,\theta) = u_\theta(x) f_\theta^{\alpha}(x) - \int u_\theta(x) f_\theta^{1+\alpha}(x)dx. \tag{2.3}$$

와 같은 M-추정함수 ψ를 갖는다고 할 수 있다. Hampel 등 (1986)은 ψ-함수가 유계이고 연속이면 주어진 분포에 대한 M-추정량은 소위 질적 로버스트성(robustness)을 가짐을 보였다. 따라서 MDPDE도 M-추정의 관점에서 본원적으로 로버스트성을 갖는다고 하겠다. Basu 등 (1998)은 특별히 조절모수 α가 로버스트성과 효율성을 적절하게 조절(trade-off)하는데 최우추정량과 비교해서 α가 0에 가까울수록 효율성이 커지고 α가 1에 가까울수록 로버스트성이 커짐을 보였다.

2.2. 정규-지수밀도함수

마이크로어레이 분석에서 X_f를 발현값(혹은 관찰값, foreground), X를 참값(signal) 그리고 X_b는 배경값(background)이라고 하면 $X_f = X + X_b$라고 정의된다. 확률변수 X와 X_b가 λ를 모수로 갖는 지수분포와 μ와 σ를 모수로 갖고 양수에서만 정의되는 정규분포를 갖는다고 하면 X_f의 밀도함수 (Bolstad, 2004; McGee와 Chen, 2006)는 아래와 같게 된다.

$$f(x_f;\mu,\sigma,\lambda) = \frac{1}{\lambda} \exp\left(-\frac{x_f-\mu}{\lambda} + \frac{\sigma^2}{2\lambda^2}\right) \left(1 - \Phi\left(0;\mu_{x_f},\sigma^2\right)\right), \tag{2.4}$$

여기서 $\Phi(\cdot)$는 평균 $\mu_{x_f}(= x_f - \mu - \sigma^2/\lambda)$와 분산 σ^2를 갖는 누적정규분포함수이다. 위의 f가 앞 장에서 언급한 정규-지수밀도함수가 된다. 참값 X에 대한 추정량은 X의 조건부 기댓값으로 정의된다. 즉,

$$E[X|X_f = x] = a + b\left(\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{x-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{x-a}{b}\right) - 1}\right),$$

여기서 $a = x - \mu - \sigma^2/\lambda$, $b = \sigma$ 그리고 $\phi(\cdot)$는 표준정규밀도함수이다.

3. 최소밀도함수승간격추정과 정규-지수밀도함수

구체적인 형태는 뒤에 기술하겠지만 정규-지수밀도함수에 최소밀도함수승간격추정을 시행하는 과정에서 적분이 대수적으로 쉽게 구해지지 않는데 그 적분에 대한 근삿값을 구하는 방법으로 소위 라플라스 근사법(Laplace approximation)을 이용할 것을 제안한다. 라플라스 근사법 내용은 MacKay (1998)의 논문을 기초로 정리하였다.

θ를 모수로 갖는 함수 $f_\theta(x)$가 있다고 하자. 그런데, 실제로는 로그를 취한 $h_\theta(x) \equiv \log f_\theta(x)$가 더 다루기 편하여 f 대신에 h를 어떤 점 x_0에 대하여 전개를 하여 지수함수를 취하면

$$f_\theta(x) \approx \exp\left\{h_\theta(x_0) + (x-x_0)h'_\theta(x_0) + \frac{(x-x_0)^2}{2}h''_\theta(x_0)\right\}$$

은 얻게 된다. 만일 $h'_\theta(\hat{x}) = 0$가 되는 $\hat{x}$를 x_0로 놓으면 보다 간단한 표현식

$$f_\theta(x) \approx \exp\left\{h_\theta(\hat{x}) + \frac{(x-\hat{x})^2}{2}h''_\theta(\hat{x})\right\}$$

을 얻을 수 있다. 그리고 양변에 적분을 취하면

$$\int f_\theta(x)dx \approx \int \exp\left\{h_\theta(\hat{x}) + \frac{(x-\hat{x})^2}{2}h''_\theta(\hat{x})\right\}dx.$$

가 되고 오른쪽 적분의 두 번째 항은 마치 평균이 $\hat{x}$ 그리고 분산이 $-1/h''_\theta(\hat{x})$인 정규밀도함수의 지수(exponent)와 유사함을 발견할 수 있다. 이제 $\hat{x}$가 최빈값(mode)라면 $h''_\theta(\hat{x})$가 음수가 되고 $-1/h''_\theta(\hat{x})$는 양수가 되어

$$\int f_\theta(x)dx \approx \exp\left\{h_\theta(\hat{x})\right\}\left(-\frac{2\pi}{h''_\theta(\hat{x})}\right)^{\frac{1}{2}}$$

임을 알 수 있다. 한걸음 더 나아가 본 연구에서 필요한

$$f_\theta^{1+\alpha}(x) \approx \exp\left\{(1+\alpha)h_\theta(\hat{x}) + (1+\alpha)\frac{(x-\hat{x})^2}{2}h''_\theta(\hat{x})\right\}$$

과 따라서

$$\int f_\theta^{1+\alpha}(x)dx \approx \exp\left\{(1+\alpha)h_\theta(\hat{x})\right\}\left\{-\frac{2\pi}{(1+\alpha)h''_\theta(\hat{x})}\right\}^{\frac{1}{2}}. \tag{3.1}$$

을 얻을 수 있다.

예컨대 실제로 f가 정규밀도함수($\phi(\cdot)$)라면 적분값과 근사값이 일치한다.

$$\begin{aligned}\int \phi^{1+\alpha}(x;\mu,\sigma)dx &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{\alpha}\frac{1}{\sqrt{1+\alpha}} \\ &= \exp\left\{(1+\alpha)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\right\}\left\{\frac{2\pi}{(1+\alpha)\frac{1}{\sigma^2}}\right\}^{\frac{1}{2}} \\ &= \exp\left\{(1+\alpha)h_\theta(\hat{x})\right\}\left\{-\frac{2\pi}{(1+\alpha)h''_\theta(\hat{x})}\right\}^{\frac{1}{2}}.\end{aligned}$$

Remark 3.1: (로버스트성) 앞서 정의한 추정간격 (2.2)은 근사식 (3.1)을 이용하여

$$\exp\left\{(1+\alpha)h_\theta(\hat{x})\right\}\left\{-\frac{2\pi}{(1+\alpha)h''_\theta(\hat{x})}\right\}^{\frac{1}{2}} - \left(1+\frac{1}{\alpha}\right)n^{-1}\sum f_\theta^{\alpha}(X_i) \tag{3.2}$$

와 같이 근사될 수 있고 따라서 근사식에 따르는 근사추정함수는

$$\psi_a(x,\theta) = u_\theta(x)f_\theta^{\alpha}(x) - d(\theta)$$

가 되는데, 여기서 $d(\theta) = (1+\alpha)^{-1}\nabla_\theta[\exp\{(1+\alpha)h_\theta(\hat{x})\}\{-2\pi/(1+\alpha)h_\theta''(\hat{x})\}^{1/2}]$이고 ∇_θ는 θ에 대한 편미분을 의미한다. 근사추정함수는 Basu 등 (1998)의 M-추정함수 (2.3)과 $d(\theta)$에서 차이가 나는데 이는 x를 포함하지 않는다는 점에서 상수와 같다. 따라서 Basu 등 (1998)이 증명한 모든 로버스트 성질을 그대로 만족한다고 하겠다.

Remark 3.2: (근사추정함수에 의한 추정량의 효율성) 만일 참분포(true distribution)가 함수족 $\{f_\theta(x)\}$에 속하고 Basu 등 (1998)의 Theorem 2에서 제시한 정칙조건(regular condition)을 만족한다면 $n^{1/2}(\hat{\theta}-\theta)$는 점근적으로 평균(벡터)이 $\int u_\theta(x)f_\theta^{1+\alpha}(x)dx - d(\theta)$이고 공분산행렬이 $J^{-1}KJ^{-1}$인 다변량 정규분포를 따른다. 위의 $J = J(\theta)$와 $K = K(\theta)$는

$$K = \int u_\theta(x)u_\theta^T(x)f_\theta^{1+2\alpha}dx - \xi\xi^T, \quad \xi = \int u_\theta(x)f_\theta^{1+\alpha}(x)dx$$

그리고

$$J = \int -i_\theta(x)f_\theta^{\alpha}(x)dx + \alpha\int u_\theta(x)u_\theta^T(x)f_\theta^{\alpha}dx - \nabla_\theta d(\theta), \quad i_\theta(x) = -\nabla_\theta u_\theta(x)$$

이다.

위의 결과는 Basu 등 (1998)의 Theorem 2의 증명을 따라하면 보일 수 있다.

만일 $\alpha \to 0$한다면 K는 $I(\theta)$ (Fisher's information)이 되고 J는 $-I(\theta)+\nabla_\theta d(\theta)$가 된다. 근사추정간격에 근거한 MDPDE의 점근 분산 행렬은 $(I(\theta)-\nabla_\theta d(\theta))^{-1}I(\theta)(I(\theta)-\nabla_\theta d(\theta))^{-1}$가 될 텐데, 불행히도 이것은 $I^{-1}(\theta)$와 일치하지 않고 따라서 근사추정간격에 근거한 MDPDE는 본래의 추정간격에 근거한 MDPDE에 비해 다소 효율성이 떨어질 것이다. 하지만 문제가 되는 $d(\theta)$와 $\nabla_\theta d(\theta)$에 속한 $h(\theta)|_{x=\hat{x}}$의 고차 미분값들이 무시할 정도라면 평균과 분산이 Basu 등 (1998)의 결과에 접근할 것이다.

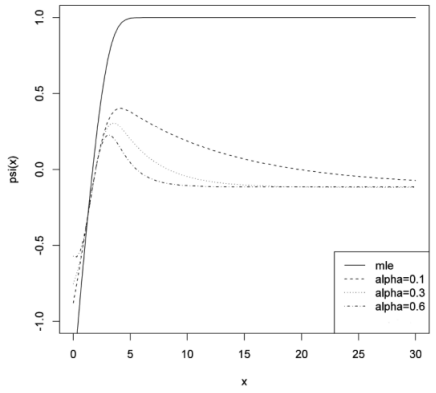
Remark 3.3: (Normexp의 경우) Normexp의 정확한 추정간격은 식 (2.4)의 f를 식 (2.2)에 넣으면

$$\int_0^\infty \frac{1}{\lambda^\alpha}\exp\left\{(1+\alpha)\left(-\frac{x-\mu}{\lambda}+\frac{\sigma^2}{2\lambda^2}\right)\right\}\left(1-\Phi\left(0;\mu_x,\sigma^2\right)\right)^{1+\alpha}dx - \left(1+\frac{1}{\alpha}\right)n^{-1}\sum_{i=1}^n f_\theta^\alpha(X_i)$$

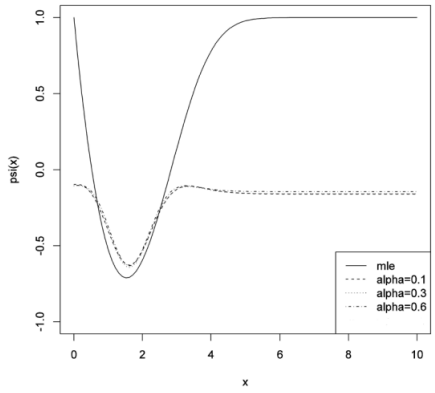
와 같은데 적분의 후반부에 나타나는 $(1+\alpha)$승이 포함된 부분의 적분을 직접적으로 해결할 수 없었다. 따라서 식 (3.2)의 근사식을 이용하면

$$\begin{aligned}
&\left\{\frac{1}{\lambda}\exp\left(-\frac{x-\mu}{\lambda}+\frac{\sigma^2}{2\lambda^2}\right)\left(1-\Phi\left(0;\mu,\sigma^2\right)\right)\Big|_{x=\hat{x}}\right\}^{1+\alpha} \\
&\times\left(\frac{2\pi}{1+\alpha}\right)^{-\frac{1}{2}}\left\{-\frac{\phi'(0;\mu_x,\sigma^2)\left(1-\Phi\left(0;\mu_x,\sigma^2\right)\right)-\phi^2\left(0;\mu_x,\sigma^2\right)}{\left(1-\Phi\left(0;\mu_x,\sigma^2\right)\right)^2}\Big|_{x=\hat{x}}\right\}^{\frac{1}{2}} \\
&-\left(1+\frac{1}{\alpha}\right)n^{-1}\sum_{i=1}^n f_\theta^\alpha(X_i) \qquad (3.3)
\end{aligned}$$

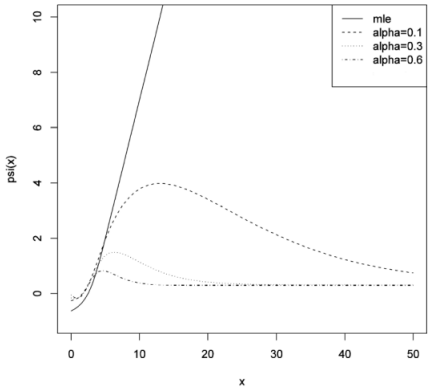
와 같이 구할 수 있고 normexp의 근사적 MDPDE는 위 식 (3.3)을 모수들에 대해 최소로 하는 값으로 구할 수 있다.



(a) μ



(b) σ



(c) λ

Figure 3.1. ψ-function for λ, μ, and σ for MLE and MDPDE with $\alpha = 0.1, 0.3$ and 0.6. The true values of all parameters are set to 1 for illustration.

위 식을 모수들에 대해 미분하여 ψ-함수를 구하고 Figure 3.1에 그려 보았다. Hampel 등 (1986)에 의하면 ψ-함수가 유계이고 연속일 때 추정량은 로버스트성을 갖는다고 하였는데 Figure 3.1을 보면 주어진 α에 따라 최우추정량의 ψ-함수 보다 x의 값에 대한 영향력이 더 많이 제한되고 있음을 볼 수 있다. 특히, λ의 경우 최우추정량의 ψ-함수는 유계를 갖지 않는 반면 제안된 추정법의 ψ-함수는 유계를 가짐을 볼 수 있다. 즉, 근사간격에 기초한 MDPDE도 원초적으로 로버스트하다고 하겠다.

4. 자료 분석

본 장에서는 모의실험과 마이크로어레이 실험에서 얻은 실제 자료를 통한 결과를 기술하였다. MLE와 MDPDE를 구하여 편의(bias)와 상대효율(relative efficiency)을 비교하고 결과와 관련된 그림들도 첨부하였다. 계산은 R로 만들어진 limma라는 패키지에 속한 normexp.fit이라는 Ritchie 등 (2007)이 제안한 안장점근사(saddle-point approximation) 알고리즘을 활용한 함수를 사용하여 MLE를 구하였다. MDPDE는 'L-BFGS-B'라고 하는 최적화 알고리즘 (Byrd 등, 1995)을 구현하는 optim이라는 R함수

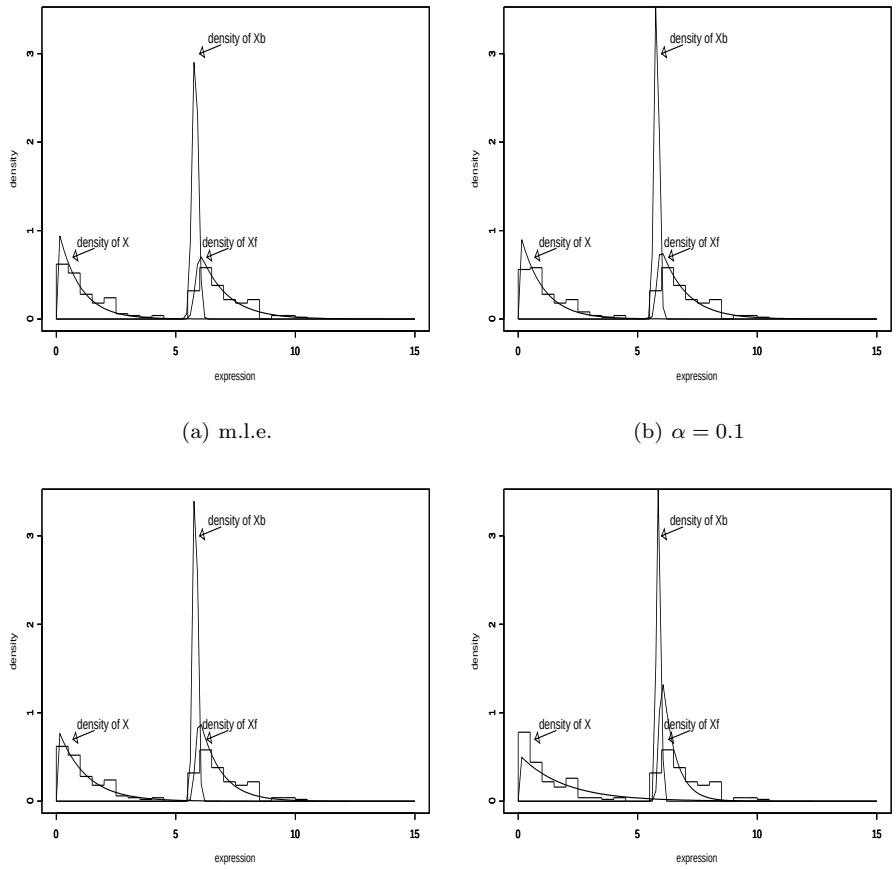
Table 4.1. The related statistics for the MLE and the MDPDE of μ, σ and λ under various contaminations.

		0%		10%		20%		30%	
		$(\text{bias})^2$	RE	$(\text{bias})^2$	RE	$(\text{bias})^2$	RE	$(\text{bias})^2$	RE
μ	MLE	0.013	1.000	0.010	1.000	0.049	1.000	0.078	1.000
	$\alpha = 0.1$	0.011	0.786	0.001	0.754	0.013	0.959	0.023	0.778
	$\alpha = 0.2$	0.012	0.786	0.002	0.754	0.015	0.952	0.028	0.783
	$\alpha = 0.3$	0.012	0.783	0.002	0.759	0.018	0.952	0.033	0.794
	$\alpha = 0.4$	0.012	0.780	0.003	0.759	0.021	0.959	0.037	0.800
	$\alpha = 0.5$	0.011	0.776	0.004	0.759	0.023	0.966	0.041	0.811
	$\alpha = 0.6$	0.011	0.770	0.004	0.754	0.025	0.972	0.044	0.817
	$\alpha = 0.7$	0.011	0.767	0.005	0.750	0.026	0.979	0.046	0.822
	$\alpha = 0.8$	0.012	0.754	0.005	0.746	0.025	0.986	0.047	0.828
	$\alpha = 0.9$	0.013	0.745	0.004	0.737	0.024	0.993	0.047	0.833
	$\alpha = 1.0$	0.014	0.733	0.004	0.729	0.023	1.007	0.046	0.839
σ	MLE	0.013	1.000	0.011	1.000	0.056	1.000	0.075	1.000
	$\alpha = 0.1$	0.003	0.850	0.016	0.835	0.033	0.678	0.040	0.617
	$\alpha = 0.2$	0.002	0.853	0.017	0.839	0.034	0.681	0.041	0.623
	$\alpha = 0.3$	0.002	0.858	0.017	0.842	0.035	0.684	0.042	0.629
	$\alpha = 0.4$	0.002	0.864	0.018	0.853	0.036	0.711	0.043	0.635
	$\alpha = 0.5$	0.002	0.872	0.019	0.860	0.038	0.701	0.044	0.641
	$\alpha = 0.6$	0.001	0.878	0.020	0.867	0.039	0.707	0.045	0.643
	$\alpha = 0.7$	0.001	0.886	0.018	0.877	0.040	0.714	0.046	0.646
	$\alpha = 0.8$	0.001	0.894	0.021	0.884	0.041	0.720	0.047	0.652
	$\alpha = 0.9$	0.001	0.897	0.022	0.895	0.042	0.727	0.048	0.658
	$\alpha = 1.0$	0.001	0.903	0.022	0.898	0.043	0.734	0.048	0.664
λ	MLE	0.035	1.000	0.444	1.000	1.132	1.000	1.968	1.000
	$\alpha = 0.1$	0.004	0.919	0.469	0.985	0.972	0.838	1.695	0.852
	$\alpha = 0.2$	0.001	0.860	0.332	0.706	0.856	0.753	1.545	0.786
	$\alpha = 0.3$	0.000	0.833	0.297	0.660	0.817	0.730	1.506	0.771
	$\alpha = 0.4$	0.000	0.824	0.280	0.644	0.808	0.726	1.506	0.769
	$\alpha = 0.5$	0.000	0.824	0.270	0.635	0.803	0.725	1.501	0.769
	$\alpha = 0.6$	0.000	0.824	0.264	0.628	0.801	0.724	1.501	0.768
	$\alpha = 0.7$	0.000	0.824	0.259	0.625	0.792	0.721	1.496	0.765
	$\alpha = 0.8$	0.000	0.829	0.259	0.625	0.792	0.715	1.493	0.766
	$\alpha = 0.9$	0.000	0.833	0.253	0.613	0.792	0.717	1.486	0.764
	$\alpha = 1.0$	0.000	0.838	0.252	0.615	0.790	0.715	1.488	0.765

를 이용하여 구하였다. 이를 위해 초기치를 지정해 주어야하는데 Silver 등 (2009)과 같이 μ는 중간값(median), σ^2는 (자료-중간값)의 제곱의 평균 그리고 λ는 (산술평균-중간값)을 사용하였다.

4.1. 모의실험

모의실험에서는 (μ, σ, λ)의 참값이 $(1, 1, 1)$이라고 가정하고 normexp에서 원소의 개수가 500개인 표본 1000개를 추출하여 MLE와 MDPDE를 구하였다. 그 다음에 추출된 각 표본에서 10%, 20% 그리고 30%의 원소들을 균등분포 $U(0, 10)$에서 추출한 원소들로 대체하여 오염된 표본들을 만들어 MLE와 MDPDE를 구하였다. MDPDE는 $\alpha = 0.1, 0.2, \ldots, 0.9, 1.0$에 대하여 계산하였다. 그 결과가 Table 4.1에 기록되어있다.



(a) m.l.e. (b) $\alpha = 0.1$

(c) $\alpha = 0.3$ (d) $\alpha = 0.6$

Figure 4.1. The histogram on the left side is of the estimated signal X, while that on the right side is of foreground X_f for a particular sample. Estimated densities by MLE and by MDPDE with $\alpha = 0.1, 0.3$ and 0.6.

모의실험의 결과는 이론을 통해 짐작하고 있었던 대로 대부분의 경우 MDPDE의 편의가 MLE의 편의 보다 작게 나타나나 상대적 효율성은 MLE가 우수함을 확인할 수 있다. 상대적효율성(relative efficiency, RE)는 MDPDE의 표본분산을 MLE의 표본 분산으로 나누어 계산하였다. MDPDE는 상황에 따라 적당한 α에 대하여 편의와 상대적 효율성이 최적화되는 것을 확인할 수 있는데 전체적으로 보아서는 $\alpha \in (0.4, 0.6)$인 경우 주어진 데이터에 대해서 최적이라고 보인다. λ에 대한 MLE의 편의가 상대적으로 매우 큼을 확인할 수 있는데 λ에 대한 MLE의 ψ-함수가 유계가 아닌 것을 상기하게 된다.

4.2. 마이크로어레이 자료

만성 림프구성 백혈병(chronic lymphocytic leukemia) 유전자 발현 자료를 사용하여 분석을 해보았다. HG-U95Av2 Affymetrix gene chips 자료는 12625개의 유전자에 대한 24개 샘플의 발현 자료를 포함하고 있다. 본 예제에서는 PM(perfect match)자료를 사용하였다. Figure 4.1에 예시로 유전자 크기가 100인 표본 하나를 택하여 MLE와 세 가지 MDPDE에 따라 참값(X)분포를 추정하여 그림을 그려보았

고 $\alpha = 0.3$일 때의 MDPDE에 의한 참값의 지수함수 추정함수가 자료에 가장 잘 맞는 것을 시각적으로 확인할 수 있다.

5. 결론

최소밀도함수승간격추정은 다소 다루기 힘든 적분식을 포함하고 있다. 대부분의 경우 수치적으로 다룰 수 있느나 대수적으로 해결이 어려운 경우가 있다. 본 논문에서 관심이 있는 정규-지수밀도함수의 경우에도 역시 대수적으로 다루는데 어려움이 있다. 이 문제를 해결하는 한 가지 방법으로 적분을 근사식으로 바꾸어 볼 것을 제안하였고 관련된 수학적 분석과 경험적 분석을 수행하였다. 나름 로버스트성과 효율성에서 효과적인 결과를 얻었음을 주장하였다. 본 논문이 특정한 분포에 초점이 맞추어져 있는 것은 사실이지만 이것을 기회로 좀 더 일반화된 연구가 이루어질 단초를 제공하고 있다고 생각된다. 또한, Basu 등 (1998)에서도 언급되었듯이 적절한 조절모수 α를 선택하는 범용적인(universal)방법은 현재 없음으로 경험적으로 이루어져야 하는 문제가 본 연구에서도 여전히 존재한다. 본 논문에서는 라플라스근사 이후에 발생하는 로버스트 성질을 다루었으나, 라플라스근사가 행해지는 과정에서의 로버스트 문제는 차후 연구되어야 할 부분이라고 하겠다. 기존에 효과적이고 편리한 방법들이 존재하지만 최소밀도함수승간격추정이라는 로버스트 추정법이 마이크로어레이 분석 같은 특정한 분야에도 사용될 수 있음을 보였다는 것은 의미가 있다고 하겠다.

References

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence, *Biometrika*, **85**, 549–559.

Bolstad, B. M. (2004). *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*, Dissertation, University of California-Berkeley.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization, *SIAM J. Sci Comput*, **16**, 1190-1208.

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

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249–264.

MacKay, D. J. C. (1998). Choice of basis for Laplace approximation, *Mach Learn*, **33**, 77–86.

McGee, M. and Chen, Z. (2006). Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data, *Stat Appl Genet Molec Biol*, 5:Article 24.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays, *Bioinformatics*, **23**, 2700–2707.

Silver, J., Ritchie, M. E. and Smyth, G .K. (2009). Microarray background correction: maximum likelihood estimation for the normal-exponential convolution model, *Biostatistics*, **10**, 352–363.

정규-지수분포에 대한 최소밀도함수승간격 추정법

박노진[a,1]

[a]단국대학교 응용통계학과

(2014년 1월 24일 접수, 2014년 3월 12일 수정, 2014년 4월 9일 채택)

요 약

최소밀도함수승간격 추정법은 Baus 등 (1998)에 의해 처음 소개된 이후 많은 관심의 대상이 되었다. 최소밀도함수승간격 추정량은 우수한 로버스트 성질을 갖고 효율성도 최우추정량에 필적한 것으로 알려져 있다. 본 논문에서는 생물정보학에서 사용되는 노말-지수 분포에 근거한 추정량을 최소밀도함수승간격 추정법을 사용하여 구하는 방법을 다루고자 한다. 그런데 그 과정에서 간격을 적분을 통해 구하는 것이 매우 어려움으로 인해 직접적인 적분 대신 라플라스 근사를 시도할 것을 제안한다. 그 결과 추정량이 다소 효율성이 줄어들지만 로버스트 성질을 갖고 있음을 수학적 방법과 모의실험을 통하여 보였다.

주요용어: 로버스트 추정법, 라플라스 근사, 마이크로어레이, 효율성.

[1](448-701) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과. E-mail: rjpak@dankook.ac.kr