김일성종합대학학보 주체107(2018)년
수 학 제64권 제3호

반파라메터최량판별함수에 대한 추정

림 창 호

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《이미 일정한 토대가 있고 전망이 확고한 연구대상들에 힘을 넣어 세계패권을 쥐며 그 성과를 확대하는 방법으로 과학기술을 빨리 발전시켜야 합니다.》(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》 단행본 39페지)

반파라메터판별함수에 의한 판별분석문제는 패턴인식을 비롯한 현실문제들에서 많이 제기된다.

선행연구에서는 회귀분석문제로서 부분선형회귀모형과 반파라메터회귀모형의 추론문제[7-9]와 파라메터판별함수에 의한 판별분석문제[3, 4, 6]를 고찰하였다.

또한 비파라메터판별함수에 의한 판별분석문제[1, 2]와 여러반파라메터모형들사이의 판별분석문제[5], 극값판별분석과 선형판별분석모형의 일반화로서 반파라메터선형판별분석모형[10, 11]을 론의하였다.

판별변량 x에 관한 모집단 G_r $(r=1,\ 2)$의 밀도함수와 사전확률을 각각 $f_r^0(x),\ q_r^0$이라고 하자.

표본공간 R를 공통부분이 없는 2개의 구역 $R_1,\ R_2$로 나누고

$$x\in R_1 \Rightarrow x\in G_1,\ \ x\in R_2 \Rightarrow x\in G_2$$

로 판별할 때 오판별확률

$$P=q_1^0\int_{R_2} f_1^0(x)dx+q_2^0\int_{R_1} f_2^0(x)dx$$

가 최소로 되는 최량판별규칙은

$$2F^0(x)>1 \Rightarrow x\in G_1,\ \ 2F^0(x)\le 1 \Rightarrow x\in G_2 \tag{1}$$

이다. 여기서

$$F^0(x)=P\{G_1\mid x\}=\frac{q_1^0 f_1^0(x)}{q_1^0 f_1^0(x)+q_2^0 f_2^0(x)}$$

이며 이때 $F^0(x)$를 최량판별함수라고 부른다.

모집단 G_r $(r=1,\ 2)$의 밀도함수 $f_r^0(x)$ $(r=1,\ 2)$와 사전확률 q_r^0 $(r=1,\ 2)$들이 미지라고 하면 최량판별규칙 (1)은 미지판별함수 $F(x)=q_1f_1(x)/[q_1f_1(x)+q_2f_2(x)]$에 의한 판별규칙 $2F(x)>1\Rightarrow x\in G_1,\ 2F(x)\le 1\Rightarrow x\in G_2$로 된다. 그리고 미지판별함수 $F(x)$는 미지파라메터 $\boldsymbol{\theta}=(\theta_1,\ \cdots,\ \theta_s)^{\mathrm{T}}\in\Theta$와 미지함수 $\boldsymbol{g}=(g_1,\ \cdots,\ g_q)^{\mathrm{T}}\in H\equiv H(\mathbf{R},\ \mathbf{R}^q)$를 포함하는 반파라메터판별함수 $F(x)=F(x\,;\ \boldsymbol{\theta},\ \boldsymbol{g})$로 된다.

이때 $F(x\,;\boldsymbol{\theta},\ \boldsymbol{g})$에 의한 판별규칙 $2F(x\,;\boldsymbol{\theta},\ \boldsymbol{g})>1\Rightarrow x\in G_1,\ 2F(x\,;\boldsymbol{\theta},\ \boldsymbol{g})\le 1\Rightarrow x\in G_2$에서 오판별확률

$$P = q_1^0 \int_{R_2} f_1^0(x)dx + q_2^0 \int_{R_1} f_2^0(x)dx,\ \ R_1 = \{x \mid 2F(x;\ \boldsymbol{\theta},\ \boldsymbol{g}) > 1\},\ R_2 = \{x \mid 2F(x;\ \boldsymbol{\theta},\ \boldsymbol{g}) \le 1\}$$

이 최소로 되는 최량판별규칙

$$2F(x;\ \boldsymbol{\theta}^*,\ \boldsymbol{g}^*) > 1 \Rightarrow x \in G_1,\ 2F(x;\ \boldsymbol{\theta}^*,\ \boldsymbol{g}^*) \le 1 \Rightarrow x \in G_2 \tag{2}$$

를 반파라메터최량판별규칙, $F(x;\boldsymbol{\theta}^*,\ \boldsymbol{g}^*)$을 반파라메터최량판별함수라고 부른다.

론문에서는 표본자료 $x_1, \cdots, x_n$에 기초하여 반파라메터최량판별함수 $F(x;\boldsymbol{\theta}^*,\ \boldsymbol{g}^*)$의 미지파라메터 $\boldsymbol{\theta}^*$과 미지함수 $\boldsymbol{g}^*$을 추정하는 한가지 방법을 론의한다.

먼저 반파라메터최량판별함수를 추정하자.

표본자료 $x_1, \cdots, x_n$에 기초하여 반파라메터판별함수 $F(x;\ \boldsymbol{\theta}^*,\ \boldsymbol{g}^*)$의 $\boldsymbol{\theta}^*$과 $\boldsymbol{g}^*$을 추정하기 위하여 기준우연량 $p_k = \begin{cases} 1,\ x_k \in G_1 \\ 0,\ x_k \in G_2 \end{cases}$ $(k = 1,\ \cdots,\ n)$를 도입하면 반파라메터모형

$$p_k = F(x_k;\ \boldsymbol{\theta}^*,\ \boldsymbol{g}^*) + \varepsilon_k \tag{3}$$

$$\mathrm{E}\varepsilon_k = 0,\ \mathrm{Var}\varepsilon_k = P_k \cdot Q_k,\ \ P_k = P\{G_1 \mid x_k\} = F(x_k;\boldsymbol{\theta}^*,\ \boldsymbol{g}^*),\ \ Q_k = 1 - P_k\ \ (k = 1,\ \cdots,\ n)$$

가 얻어진다. 여기서 $\mathrm{E}p_k = P_k$, $\mathrm{Var}p_k = (1 - P_k)^2 \cdot P_k + (0 - P_k)^2 \cdot Q_k = P_k \cdot Q_k$이다.

이때 $F(x;\ \boldsymbol{\theta}^*,\ \boldsymbol{g}^*)$에서 x를 고정하면 F는

$$\boldsymbol{\alpha}^* = (\boldsymbol{\theta}^{*\mathrm{T}},\ \boldsymbol{g}^{*\mathrm{T}})^{\mathrm{T}} \in \Theta \times H = \mathbf{R}^{s+q}$$

에서 정의된 함수로 볼수 있다.

론문에서는 F가 $\boldsymbol{\alpha}^*$에 관하여 미분가능한 경우만을 론의한다.

$\mathbf{R}^{s+q}$에서 정의된 $F(x;\ \boldsymbol{\alpha}^*)$을 $\boldsymbol{\alpha}_0 = (\boldsymbol{\theta}_0^{\mathrm{T}},\ \boldsymbol{g}_0^{\mathrm{T}})^{\mathrm{T}}$에서 테일러전개하면 다음과 같다.

$$F(x;\ \boldsymbol{\alpha}^*) = F(x;\ \boldsymbol{\alpha}_0) + F'_{\boldsymbol{\alpha}^*}(x;\ \boldsymbol{\alpha}_0)(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0) + o(\| \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0 \|)$$

여기서 $o(\| \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0 \|) = r_2(\boldsymbol{\alpha}_0,\ l) = r_2(x_k;\ \boldsymbol{\alpha}_0,\ l) \equiv R(x)$ ($R(x)$는 $\boldsymbol{\theta}^*$, $\boldsymbol{g}^*$에 관계되는 반파라메터함수이다.)이며 $l = \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0$이다.

따라서 반파라메터모형 (3)은 부분선형모형

$$\widetilde{p}_k = F'_{\boldsymbol{\theta}^*}(x_k;\boldsymbol{\alpha}_0) \cdot \boldsymbol{\theta}^* + F'_{\boldsymbol{g}^*}(x_k;\boldsymbol{\alpha}_0) \cdot \boldsymbol{g}^* + R(x_k) + \varepsilon_k \tag{4}$$

$$\mathrm{E}\varepsilon_k = 0,\ \mathrm{Var}\varepsilon_k = P_k \cdot Q_k,\ \ P_k = F(x_k;\boldsymbol{\alpha}^*),\ \ Q_k = 1 - P_k$$

$$\widetilde{p}_k = p_k - F(x_k;\ \boldsymbol{\alpha}_0) + F'_{\boldsymbol{\alpha}^*}(x_k;\ \boldsymbol{\alpha}_0)\boldsymbol{\alpha}_0\ \ \ (k = 1,\ \cdots,\ n)$$

으로 되며 초기추정량 $\hat{\boldsymbol{\alpha}}_1$에 대하여 부분선형모형 (4)는 다음과 같다.

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\theta}^* + \boldsymbol{Z} \cdot \overline{\boldsymbol{g}}^* + \boldsymbol{\varepsilon} \tag{5}$$

$$\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0}_{n\times 1},\ \mathrm{Var}\,\boldsymbol{\varepsilon} \approx \mathrm{V\hat{a}r}\,\boldsymbol{\varepsilon} = \mathrm{diag}(\hat{P}_1 \cdot \hat{Q}_1,\ \cdots,\ \hat{P}_n \cdot \hat{Q}_n)_{n\times n}$$

$$\hat{P}_k = F(x_k;\hat{\boldsymbol{\alpha}}_1),\ \hat{Q}_k = 1 - \hat{P}_k\ \ (k = 1,\ \cdots,\ n)$$

$$\boldsymbol{X} = (X_1,\ \cdots,\ X_n)^{\mathrm{T}} = (F'_{\boldsymbol{\theta}^*}(x_1;\ \hat{\boldsymbol{\alpha}}_1)^{\mathrm{T}},\ \cdots,\ F'_{\boldsymbol{\theta}^*}(x_n;\ \hat{\boldsymbol{\alpha}}_1)^*)^{\mathrm{T}}_{n\times s}$$

$$\boldsymbol{Y} = (Y_1,\ \cdots,\ Y_n)^{\mathrm{T}} = (p_1 - F(x_1;\ \hat{\boldsymbol{\alpha}}_1) + F'_{\boldsymbol{\alpha}^*}(x_1;\ \hat{\boldsymbol{\alpha}}_1),\ \cdots,\ p_n - F(x_n;\ \hat{\boldsymbol{\alpha}}_1) + F'_{\boldsymbol{\alpha}^*}(x_n;\ \hat{\boldsymbol{\alpha}}_1))^{\mathrm{T}}_{n\times 1}$$

$$\boldsymbol{Z} = (Z_1,\ \cdots,\ Z_n)^{\mathrm{T}} = ((F'_{\boldsymbol{g}^*}(\boldsymbol{x}_1;\ \hat{\boldsymbol{\alpha}}_1),\ l)^{\mathrm{T}},\ \cdots,\ (F'_{\boldsymbol{g}^*}(\boldsymbol{x}_n;\ \hat{\boldsymbol{\alpha}}_1),\ l)^{\mathrm{T}})^{\mathrm{T}}_{n\times(q+1)}$$

$$\overline{\boldsymbol{g}}^* = (\boldsymbol{g}^{*\mathrm{T}},\ \mathbf{R})^{\mathrm{T}}{}_{(q+1)\times 1},\ \boldsymbol{\varepsilon} = (\varepsilon_1,\ \cdots,\ \varepsilon_n)^{\mathrm{T}}$$

모형 (5)의 미지파라메터 $\boldsymbol{\theta}^*$과 미지함수 $\overline{\boldsymbol{g}}^*$을 추정하기 위하여 측면최소두제곱법과 국부선형화수법을 적용하면 다음의 결론이 나온다.

보조정리 모형 (5)에서 오차들의 분산이 같다고 할 때 반파라메터최량판별규칙의 파라메터성분의 추정량은 $\widetilde{\boldsymbol{\theta}}_n' = \left(\sum_{k=1}^{n} \widetilde{X}_k \widetilde{X}_k^{\mathrm{T}}\right)^{-1} \left(\sum_{k=1}^{n} \widetilde{X}_k \widetilde{Y}_k\right)_{s\times 1}$ 이며 비파라메터성분의 추정량은

$$\widetilde{\overline{\boldsymbol{g}}}(x) = (\widetilde{\boldsymbol{g}}(x),\ \widetilde{R}(x)) = (\widetilde{g}_1(x),\ \cdots,\ \widetilde{g}_q(x),\ \widetilde{R}(x))^{\mathrm{T}} = (\boldsymbol{E}_{q+1},\ \mathbf{0}_{(q+1)\times(q+1)})(\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{D}_x)^{-1}\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\theta}}_n)$$

이다. 여기서 $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1,\ \cdots,\ \widetilde{Y}_n)^{-1} = \boldsymbol{Y} - \boldsymbol{SY}$, $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1,\ \cdots,\ \widetilde{X}_n)^{\mathrm{T}} = \boldsymbol{X} - \boldsymbol{SX}$,

$$\boldsymbol{S} = \begin{pmatrix} (Z_1^{\mathrm{T}},\ \mathbf{0})(\boldsymbol{D}_{x_1}^{\mathrm{T}}\boldsymbol{W}_{x_1}\boldsymbol{D}_{x_1})^{-1}\boldsymbol{D}_{x_1}^{\mathrm{T}}\boldsymbol{W}_{x_1} \\ \vdots \\ (Z_n^{\mathrm{T}},\ \mathbf{0})(\boldsymbol{D}_{x_n}^{\mathrm{T}}\boldsymbol{W}_{x_n}\boldsymbol{D}_{x_n})^{-1}\boldsymbol{D}_{x_n}^{\mathrm{T}}\boldsymbol{W}_{x_n} \end{pmatrix}_{n\times n},\quad \boldsymbol{D}_x = \begin{pmatrix} Z_1^{\mathrm{T}} & (x_1 - x)/h\cdot Z_1^{\mathrm{T}} \\ \vdots & \vdots \\ Z_n^{\mathrm{T}} & (x_n - x)/h\cdot Z_n^{\mathrm{T}} \end{pmatrix}_{n\times 2(q+1)}$$

$$\boldsymbol{W}_x = \mathrm{diag}(K_h(x_1 - x),\ \cdots,\ K_h(x_n - x))_{n\times n},\ (Z_k^{\mathrm{T}},\ \mathbf{0}) = (Z_{k1},\ \cdots,\ Z_{kq},\ 0,\ \cdots,\ 0)_{1\times 2(q+1)}\ (k = 1,\ \cdots,\ n)$$

또한 $K_h(\cdot) = K(\cdot/h)/h$ 이며 $K(\cdot)$ 은 핵함수, $h = h_n$ 은 평활화파라메터, $\boldsymbol{E}_{q+1}$ 은 $q+1$ 차원단위행렬이다.

정리 1 반파라메터최량판별규칙 (2)의 파라메터성분의 추정량은

$$\hat{\boldsymbol{\theta}}_n = \left(\sum_{k=1}^{n} \gamma_k \widetilde{X}_k \widetilde{X}_k^{\mathrm{T}}\right)^{-1} \left(\sum_{k=1}^{n} \gamma_k \widetilde{X}_k \widetilde{Y}_k\right) \tag{6}$$

이며 비파라메터성분의 추정량은

$$\hat{\overline{\boldsymbol{g}}}_n(x) = (\hat{g}_1(x),\ \cdots,\ \hat{g}_q(x),\ \hat{R}(x))^{\mathrm{T}} = (\boldsymbol{E}_{q+1},\ \mathbf{0}_{(q+1)\times(q+1)})(\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{D}_x)^{-1}\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_n) \tag{7}$$

이다. 여기서 무게 $\gamma_k = 1/(\mathrm{Var}\,\varepsilon_k)\quad (k = 1,\ \cdots,\ n)$ 이다.

일반적으로 $\mathrm{Var}\,\varepsilon_k\ (k = 1,\ \cdots,\ n)$ 는 미지이므로 추정하여 얻은 $\hat{\gamma}_k = 1/\widehat{\mathrm{Var}}\,\varepsilon_k\ (k = 1,\ \cdots,\ n)$ 을 리용하면 추정량 (6), (7)은 다음과 같다.

$$\hat{\hat{\boldsymbol{\theta}}}_n = \left(\sum_{k=1}^{n} \hat{\gamma}_k \widetilde{X}_k \widetilde{X}_k^{\mathrm{T}}\right)^{-1} \left(\sum_{k=1}^{n} \hat{\gamma}_k \widetilde{X}_k \widetilde{Y}_k\right)$$

$$\hat{\hat{\overline{\boldsymbol{g}}}}_n(x) = (\hat{\hat{g}}_1(x),\ \cdots,\ \hat{\hat{g}}_q(x),\ \hat{\hat{R}}(x))^{\mathrm{T}} = (\boldsymbol{E}_{q+1},\ \mathbf{0}_{(q+1)\times(q+1)})(\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x\boldsymbol{D}_x)^{-1}\boldsymbol{D}_x^{\mathrm{T}}\boldsymbol{W}_x(\boldsymbol{Y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\theta}}}_n)$$

증명 오차분산이 상수가 아닌것을 고려하여 무게붙은 최소두제곱법을 적용하면 보조정리에 의하여 식 (6), (7)로 표시되는 추정량을 얻게 된다.(증명끝)

$F(x;\ \hat{\boldsymbol{\theta}},\ \hat{\boldsymbol{g}})$ 에 의한 판별규칙 (2)에서 오판별확률은 다음과 같이 표시된다.

$$P = q_1^0 \int_{R_2} f_1^0(x)dx + q_2^0 \int_{R_1} f_2^0(x)dx,\ R_1 = \{x \mid 2F(x;\ \hat{\boldsymbol{\theta}},\ \hat{\boldsymbol{g}}) > 1\},\ R_2 = \{x \mid 2F(x;\ \hat{\boldsymbol{\theta}},\ \hat{\boldsymbol{g}}) \le 1\}$$

다음으로 반파라메터최량판별함수추정량의 성질에 대하여 론의하자.

다음과 같은 조건들에 대하여 보자.

① x의 밀도함수 $f(x)$ 는 2계련속미분가능하고 $f(x) > 0$ 을 만족시킨다.

② $\forall x \in [0,\ 1],\ \Gamma(x) = \mathrm{E}(Z_1 Z_1^{\mathrm{T}} \mid x_1 = x)$ 는 불퇴화행렬이고

$$\mathrm{E}(X_1 X_1^{\mathrm{T}} \mid x_1 = x),\ \ \Gamma(x),\ \ \Phi(x) = \mathrm{E}(Z_1 X_1^{\mathrm{T}} \mid x_1 = x)$$

는 모두 립쉬츠련속이다. 특히 $\Sigma := (\mathrm{E}(\gamma_1(X_1 - \Phi(x_1)^{\mathrm{T}}\Gamma(x_1)^{-1}Z_1)^{\otimes 2}))^{-1}$ 은 정값행렬이다.

③ $\exists t > 2,\ \mathrm{E}\| X_1 \|^{2t} < \infty,\ \mathrm{E}\| Z_1 \|^{2t} < \infty,\ \exists s < 2 - t^{-1},\ n^{2s-1}h_n \to \infty\ (n \to \infty)$

④ $g_j(\cdot) \in C^2[0,\ 1]\ (j = 1, \cdots,\ q+1)\ \ nh_n^8 \to 0,\ nh_n^8/(\log n)^2 \to \infty$

⑤ 핵함수 $K(\cdot)$ 은 유계받침 $[-10,\ 1]$ 을 가지는 대칭밀도함수이다.

⑥ $\exists M_1,\ M_2\ ;\ 0 < M_1 < \sigma_i^2 < M_2 < +\infty,\ 1 \le i \le n$ (거의)

⑦ $\max\limits_{1 \le i \le n} |\hat{\gamma}_i - \gamma_i| = o_p(n^{-1/4})$

정리 2 조건 ①－⑥을 만족시키면 식 (6)으로 표시되는 $\hat{\boldsymbol{\theta}}_n$ 은 $\boldsymbol{\theta}$ 의 점근정규추정량이다. 즉 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}(\mathbf{0},\ \Sigma)$ 가 성립된다.

정리 3 조건 ①－⑦을 만족시키면 식 (7)로 표시되는 $\hat{\hat{\boldsymbol{g}}}_n(x) = (\hat{g}_1(x),\ \cdots,\ \hat{g}_{q+1}(x))^{\mathrm{T}}$ 에 대하여 $\max\limits_{1 \le j \le q+1} \sup\limits_{x \in (0,\ 1)} |\hat{g}_j(x) - g_j(x)| = o_p(c_n)$, $c_n = (\log h^{-1}/nh)^{1/2} + h^2$ 이 성립된다.

얻어진 추정량을 다시 초기추정량 $\hat{\boldsymbol{\alpha}}_1$ 로 놓고 반복한다.

참 고 문 헌

[1] G. Melnichenko; J. Multi. Anal., 101, 68, 2010.
[2] M. Mojirsheibani et al.; J. Multi. Anal., 98, 1051, 2007.
[3] M. S. Srivastava; J. Multi. Anal., 97, 2057, 2006.
[4] S. Velilla; J. Multi. Anal., 101, 1239, 2010.
[5] M. M. Seyam et al.; International Journal of Statistics and Probability, 2, 3, 96, 2013.
[6] D. R. Jeske et al.; J. Multi. Anal., 101, 1622, 2010.
[7] Jian Qing Fan et al.; Bernoulli, 11, 6, 1031, 2005.
[8] T. Otsu; J. Multi. Anal., 98, 1923, 2007.
[9] Jin Hong You et al.; J. Multi. Anal., 101, 1079, 2010.
[10] B. G. Manjunath et al.; J. Multi. Anal., 103, 107, 2012.
[11] Qing Mai et al.; J. Multi. Anal., 135, 175, 2015.

주체107(2018)년 3월 10일 원고접수

The Estimation for Semiparametric Optimal Discriminant Functions

Rim Chang Ho

In this paper, by using samples $x_1, \cdots, x_n$ we estimated the unknown parameter $\boldsymbol{\theta}^*$ and the unknown function $\boldsymbol{g}^*$ of semiparametric optimal discriminant function $F(x\,;\boldsymbol{\theta}^*,\ \boldsymbol{g}^*)$ and obtained statistical properties of the estimators.

Key words: semiparametric discriminant function, misclassification probability