한국통계학회논문집
2010, 17권, 6호, 811–827

DOI: 10.5351/CKSS.2010.17.6.811

이항 선택 모형에서의 절단 모수 선택

김광래[a], 조규동[a], 구자용[1,a]

[a]고려대학교 통계학과

요 약

본 논문에서는 통계적 역문제로서 이항 선택모형에서의 밀도추정 방법에 대하여 연구하였다. 밀도함수의 추정을 위하여 직교열 기저를 이용하였으며, 모형의 복잡성과 예측의 정확성을 반영한 적절한 절단모수의 선택에 대하여 고려하였다. 이항 선택 모형에서 데이터에 의존하는 절단모수를 선택하는 방법에 대해 제안하고 모의실험, 실자료를 통해 제안한 방법의 성능을 규명하였다.

주요용어: 선택 확률, 밀도 추정, 역문제, 르장드르 다항식, 분광 분해, 평균적분제곱오차.

1. 서론

사회 현상에서 얻어진 데이터는 많은 경우에 있어서 연속형이 아닌 몇 가지 범주 안에서의 값으로 관측되는데, 특히 2가지 범주에서만 값을 가지게 되는 경우를 이진 데이터(binary data)라고 한다. 이진 데이터는 경제학, 의학 분야, 산업 공학등 사회 전반에서 다루어지고 있는데 반응변수가 이진 데이터인 경우에 대한 분석 기법으로 이항 선택 모형(binary choice model; BCM)을 들 수 있다.

이항 선택 모형은

$$Y = \mathbb{I}\left\{X^\top \beta \geq 0\right\} \tag{1.1}$$

으로 나타낼 수 있는데, 여기에서 $\mathbb{I}$는 지표(indicator) 함수를 나타내며, X와 β는 임의의 d차원 랜덤 벡터이다. X의 마지막 원소를 1로 가정하면,

$$X = (Z, 1), \quad Z \in \mathbb{R}^{d-1}$$

으로 표현할 수 있다. $\|\cdot\|$를 $\mathbb{R}^d$에서 유클리드 노름(norm)을 나타낸다면, 식 (1.1)은

$$Y = \mathbb{I}\left\{\frac{X^\top}{\|X\|}\frac{\beta}{\|\beta\|} \geq 0\right\}$$

으로 표현할 수 있으므로, 자료의 특성상 X와 β는 $(d-1)$차원 구면 $\mathbb{S}^{d-1}$에서 정의되었다고 가정한다 (Gautier와 Kitamura, 2009).

이분법 선택 모형에 대한 연구는 다양하게 진행되고 있는데 최근 연구로는 Chesher와 Santos Silva (2002), Harding과 Hausman (2007), Athey와 Imbens (2007), Bajari, 등 (2007), Gautier와 Kitamura (2009) 등을 들 수 있으며 참고 도서로는 Train (2003)이 있다.

이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (KRF-2008-313-C00127).

[1] 교신저자: (136-701) 서울 성북구 안암동 5-1 고려대학교 통계학과, 교수. E-mail: jykoo@korea.ac.kr

본 논문에서는 Gautier와 Kitamura (2009)와 같이 다음의 세 가지 가정하에서 (X, Y)의 분포에서 얻어진 랜덤 표본 $(X_1, Y_1), \ldots, (X_N, Y_N)$에 근거한 공변량(covariate) X와 랜덤계수 β의 밀도함수 추정문제를 고려하였다.

가정 1. β와 X는 독립이다.

가정 2. f_X의 지지(support)는 반구 H^+ 전체이다.

가정 3. β의 지지는 반구의 부분이다.

여기서 반구란 특정 $v \in \mathbb{S}^{d-1}$에 대해 $\{x : x \in \mathbb{S}^{d-1}, X \cdot v \geq 0\}$ 로 정의된다.

밀도함수 f_X와 f_β를 추정하는 문제는 $(d-1)$차원 구면 $\mathbb{S}^{d-1}$에서 정의된 밀도함수 추정문제로 간주할 수 있으며 (Huh, 등, 2004), 이 문제는 비유클리드 공간(non-Euclidean space)에서의 역문제(inverse problems)로 간주할 수 있다 (Gautier와 Kitamura, 2009). 비유클리드 역문제에 대한 연구로는 Healy 등 (1998), Healy와 Kim (1996), Kim (1998), Kim과 Koo (2000, 2002), Kim 등 (2004), Koo와 Kim (2005, 2008a, 2008b) 등을 들 수 있다. 특히, 비유클리드 공간에서의 폘침(deconvolution)에 대한 연구는 최근 들어 이루어지고 있는데 구면 폘침 및 방향 혼합 밀도 추정에 대한 연구로는 Healy 등 (1998), Healy와 Kim (1996), Kim과 Koo (2000) 등이 있다.

본 논문의 구성은 다음과 같다. 2장에서 $\mathbb{S}^{d-1}$에서 계수 벡터 밀도 추정에 대해 알아보고 3장에서는 이를 추정하기 위한 절단모수 T를 선택하는 알고리즘을 제안한다. 그리고 4장에서는 모의실험을 통해 이를 규명하며, 5장에서 결론으로 논문을 마무리 하고자 한다.

2. 역문제 관점에서의 계수 벡터 밀도 추정

2.1. $\mathbb{S}^{d-1}$에서 정의된 직교정규 기저

H^+ 위에서 정의된 선택 확률 함수(choice probability function)를 $r(x)$라고 했을 때 이것은 $X = x$로 주어졌을 때 $Y = 1$의 기대값을 의미하며

$$r(x) = \mathbb{E}(Y = 1 \,|\, X = x) \tag{2.1}$$

로 정의된다.

통계적 선형 역문제는 알려진 선형 작용소(linear operator) K에 의해 정의된다. 이항 선택모형의 선형 반구(hemispherical) 작용소는

$$Kf(x) = \int_{b \in \mathbb{S}^{d-1}} \mathbb{I}\left\{x^\top b \geq 0\right\} f(b)db = \int_{H(x)} f(b)db \tag{2.2}$$

로 규정할 수 있는데, 여기서 $H(x) = \{b : x^\top b \geq 0\}$이다. 반구 작용소 및 구면에서의 푸리에 해석에 관한 이론은 Groemer (1996)을 참조할 수 있다.

n차 구면 조화(harmonic) 함수들의 개수 h_n은

$$h_n = \frac{(2n + d - 2)(n + d - 2)!}{n!(d-2)!(n + d - 2)} \tag{2.3}$$

로 주어진다. $\mathcal{J} = \{(l, n) : l = 1, \ldots, h_n, n = 0, 1, 2, \ldots\}$일 때, $\mathbb{S}^{d-1}$에서 정의된 직교정규 기저(orthonormal basis)를 $\{\psi_n^l : (l, n) \in \mathcal{J}\}$라 하면, $L^2(\mathbb{S}^{d-1})$에 속하는 함수 f는

$$f = \sum_{n=0}^{\infty} \sum_{l=1}^{h_n} \left\langle f, \psi_n^l \right\rangle \psi_n^l$$

와 같이 표현된다.

$(l,n) \in \mathcal{J}$에 대해

$$K\psi_n^l = \lambda_n \psi_n^l \tag{2.4}$$

이며 특히 $f \in L^2(\mathbb{S}^{d-1})$이면

$$\left\langle Kf, \psi_n^l \right\rangle = \lambda_n \left\langle f, \psi_n^l \right\rangle \tag{2.5}$$

가 된다. 여기서 λ_n은

$$\lambda_n = \left|\mathbb{S}^{d-2}\right| \int_0^1 P_n(t)\left(1-t^2\right)^{\frac{d-3}{2}} dt$$

이며 P_n은 $[-1,1]$에서 정의된 르장드르 다항식(Legendre polynomials)을 나타내는데

$$P_n(t) = \frac{(-1)^n}{2^n(\vartheta+1)\cdots(\vartheta+n)}\left(1-t^2\right)^{-\vartheta}\frac{d^n}{dt^n}\left(1-t^2\right)^{\vartheta+n}, \quad \vartheta = \frac{d-3}{2}$$

로 정의된다.

2.2. f_X의 직교열 추정

많은 통계학의 추정 문제에서 직교열(orthogonal series)에 의한 근사 방법은 좋은 성능을 보였고 X의 밀도함수인 f_X를 다음과 같이 직교열의 합으로 전개(expansion)할 수 있다.

$$f_X = \sum_{(l,n)\in\mathcal{J}} a_n^l \psi_n^l. \tag{2.6}$$

이 때, a_n^l의 불편추정량으로

$$\hat{a}_n^l = \frac{1}{N}\sum_{j=1}^{N} \psi_n^l(X_j)$$

을 고려할 수 있고, f_X에 대한 직교열 추정량은

$$\hat{f}_X = \sum_{(l,n)\in\mathcal{J}_T} \widehat{a_n^l}\psi_n^l \tag{2.7}$$

가 된다. 여기에서 $\mathcal{J}_T = \{(l,n) : l = 1,\ldots,h_n,\ n = 0,1,\ldots,T\}$이다.

2.3. f_β의 직교열 추정

$\mathbb{S}^{d-1}$에서 정의된 함수 f에 대해 홀(odd)함수 부분은

$$f^-(b) = \frac{f(b)-f(-b)}{2}$$

로 나타내고, 짝(even)함수 부분은

$$f^+(b) = \frac{f(b)+f(-b)}{2}$$

로 나타낸다.

$$R(x) = \begin{cases} r(x), & x \in H^+, \\ 1 - r(-x), & x \in -H^+. \end{cases}$$

로 정의된 함수 R의 홀 함수부분 R^-은

$$Kf_\beta^- = R^- \tag{2.8}$$

를 만족하며, 식 (2.4), (2.5), (2.8)을 이용하면,

$$\left\langle R^-, \psi_n^l \right\rangle = \left\langle Kf_\beta^-, \psi_n^l \right\rangle = \lambda_n \left\langle f_\beta^-, \psi_n^l \right\rangle \tag{2.9}$$

가 되고, f_β^-의 푸리에 계수(Fourier coefficients)를 $(l,n) \in \mathcal{J}$에 대해

$$\left\langle f_\beta^-, \psi_n^l \right\rangle = \frac{1}{\lambda_n} \mathbb{E}\left[\frac{(2Y-1)\left\{\psi_n^l(X) - \psi_n^l(-X)\right\}}{2f_X(X)} \right] \tag{2.10}$$

을 얻을 수 있다 (Gautier와 Kitamura, 2009).

f_X와 마찬가지로 f_β^-의 직교열 합은

$$f_\beta^- = \sum_{(l,n)\in\mathcal{J}} b_n^l \psi_n^l = \sum_{(l,n)\in\mathcal{J}} \left\langle f_\beta^-, \psi_n^l \right\rangle \psi_n^l$$

와 같다. 식 (2.10)에 의해 b_n^l의 적절한 추정량으로서

$$\widehat{b_n^l} = \frac{1}{\lambda_n} \frac{1}{N} \sum_{j=1}^{N} \frac{(2Y_j - 1)\left\{\psi_n^l(X_j) - \psi_n^l(-X_j)\right\}}{2\max\left(\hat{f}_X(X_j), (\log N)^{-k}\right)}$$

를 고려할 수 있다. 그러면 f_β^-의 직교열 추정량은

$$\hat{f}_\beta^- = \sum_{(l,n)\in\mathcal{J}_T} \widehat{b_n^l} \psi_n^l \tag{2.11}$$

와 같다.

앞에서 β의 지지가 반구 위에 있다고 가정하였으므로, f_β는

$$f_\beta = 2f_\beta^- \mathbb{I}\left\{f_\beta^- \geq 0\right\} \tag{2.12}$$

와 같은 식으로 얻을 수 있다. 그러므로 f_β의 추정량은 식 (2.11)에서 구한 $\hat{f}_\beta^-$을 이용하여

$$\hat{f}_\beta = 2\hat{f}_\beta^- \mathbb{I}\left\{\hat{f}_\beta^- \geq 0\right\} \tag{2.13}$$

으로 구할 수 있다.

3. 자료에서의 절단모수 T선택

Gautier와 Kitamura (2009)에서 고려된 이항 선택모형에서는 추정량의 수렴속도와 모의실험을 하였으나 절단모수선택에 대한 연구는 이루어지지 않았다. 하지만 식 (2.7)과 (2.11)에서 T가 무한대이면 무한개의 계수를 추정하고 저장해야 하는 문제로 인하여 구현상 불가능하기에 적당한 절단모수를 고려하여야 한다. 실제 함수를 가장 잘 대표할 수 있는 적당한 개수의 직교열의 선택을 통하여 밀도함수를 추정할 수 있다. 일반적인 현상으로 절단모수 T가 증가하면 추정의 변동성이 증가하지만 편의가 감소하고, 반대로 절단모수가 감소하면 추정의 변동성이 감소하지만 편의가 증가하게 된다. 따라서 주어진 자료에 대해 절단모수를 선택하는 것이 편의와 분산의 상충관계를 해결해 주는 중요한 문제가 된다.

3.1. 평균적분제곱오차

적분제곱오차(integrated squared error; ISE)는

$$\text{ISE} := \int \left(\hat{f} - f\right)^2 \tag{3.1}$$

와 같이 정의되며, 적분제곱오차의 기대값인 평균적분제곱오차(mean integrated squared error; MISE)는

$$\text{MISE} := \mathbb{E} \int \left(\hat{f} - f\right)^2 \tag{3.2}$$

로 정의된다. MISE를 최소로 하는 절단 모수가 MISE관점에서 최적이라 할 수 있는데 MISE가 미지의 밀도함수 f에 의존하므로 이에 대한 추정량이 필요하다. 함수 $f = \sum_{(l,n)\in\mathcal{J}} f_n^l \psi_n^l$에서 파스발(Parseval)의 항등관계를 이용하면

$$\|f\|^2 = \int_{\mathbb{S}^{d-1}} f^2 = \sum_{(l,n)\in\mathcal{J}} \left(f_n^l\right)^2 \tag{3.3}$$

이다. 그리고 $\hat{f}_n^l = N^{-1}\sum_{j=1}^N Z_j$를 f_n^l의 적절한 추정량이라 하면, f의 추정량은

$$\hat{f} = \sum_{(l,n)\in\mathcal{J}_T} \hat{f}_n^l \psi_n^l$$

라 한다. 그러면 식 (3.1)의 ISE는

$$\text{ISE} = \sum_{(l,n)\in\mathcal{J}_T} \left(\hat{f}_n^l - f_n^l\right)^2 + \sum_{(l,n)\in\mathcal{J}_T^c} \left(f_n^l\right)^2$$

와 같이 되고, 식 (3.2)의 MISE는 Efromovich (1999)에서 처럼

$$\begin{aligned}
\text{MISE} &= \sum_{(l,n)\in\mathcal{J}_T} \mathbb{E}\left(\hat{f}_n^l - f_n^l\right)^2 + \sum_{(l,n)\in\mathcal{J}_T^c} \left(f_n^l\right)^2 \\
&= \sum_{(l,n)\in\mathcal{J}_T} \mathbb{E}\left(\hat{f}_n^l - f_n^l\right)^2 + \|f\|^2 - \sum_{(l,n)\in\mathcal{J}_T} \left(f_n^l\right)^2 \\
&= \sum_{(l,n)\in\mathcal{J}_T} \left\{\text{Var}\left(\hat{f}_n^l\right) - \left(f_n^l\right)^2\right\} + \|f\|^2
\end{aligned}$$

와 같이 나타낼 수 있다. 이 때 식 (3.4)에서 미지의 상수 $\|f\|^2$을 제외한 $M_0(T)$

$$M_0(T) = \sum_{(l,n)\in\mathcal{J}_T} \left\{ \mathrm{Var}\left(\hat{f}_n^l\right) - \left(f_n^l\right)^2 \right\}$$

를 작게 하는 것이 MISE를 작게 하는 것과 같다.

$$Z_1, \ldots Z_N : \mathrm{IID}\left(\mu,\, \sigma^2\right)$$

일 때,

$$\hat{\mu^2} = \bar{Z}^2 - \widehat{\mathrm{Var}(\bar{Z})} = \bar{Z}^2 - \frac{1}{N}\widehat{\mathrm{Var}(Z_1)}$$

이므로 $(f_n^l)^2$의 추정량 $\widehat{(f_n^l)^2}$을 사용하면 $M_0(T)$의 추정량 $M(T)$을

$$\begin{aligned} M(T) &:= \sum_{(l,n)\in\mathcal{J}_T} \left\{ \widehat{\mathrm{Var}\left(\hat{f}_n^l\right)} - \widehat{\left(f_n^l\right)^2} \right\} \\ &= \sum_{(l,n)\in\mathcal{J}_T} \left[\widehat{\mathrm{Var}\left(\hat{f}_n^l\right)} - \left\{ \left(\hat{f}_n^l\right)^2 - \widehat{\mathrm{Var}\left(\hat{f}_n^l\right)} \right\} \right] \\ &= \sum_{(l,n)\in\mathcal{J}_T} \left\{ 2\widehat{\mathrm{Var}\left(\hat{f}_n^l\right)} - \left(\hat{f}_n^l\right)^2 \right\} \end{aligned} \tag{3.4}$$

로 구할 수 있다. $\hat{f}_X$나 $\hat{f}_\beta^-$가 $\sum_{(l,n)\in\mathcal{J}_T}(\hat{f}_n^l)\psi_n^l$의 형태로 쓰여지므로 각각의 $M(T)$를 구하고 자료에 의존하는 절단모수 $\hat{T}$는

$$\hat{T} = \underset{T\le T_{\max}}{\mathrm{argmin}}\, M(T) \tag{3.5}$$

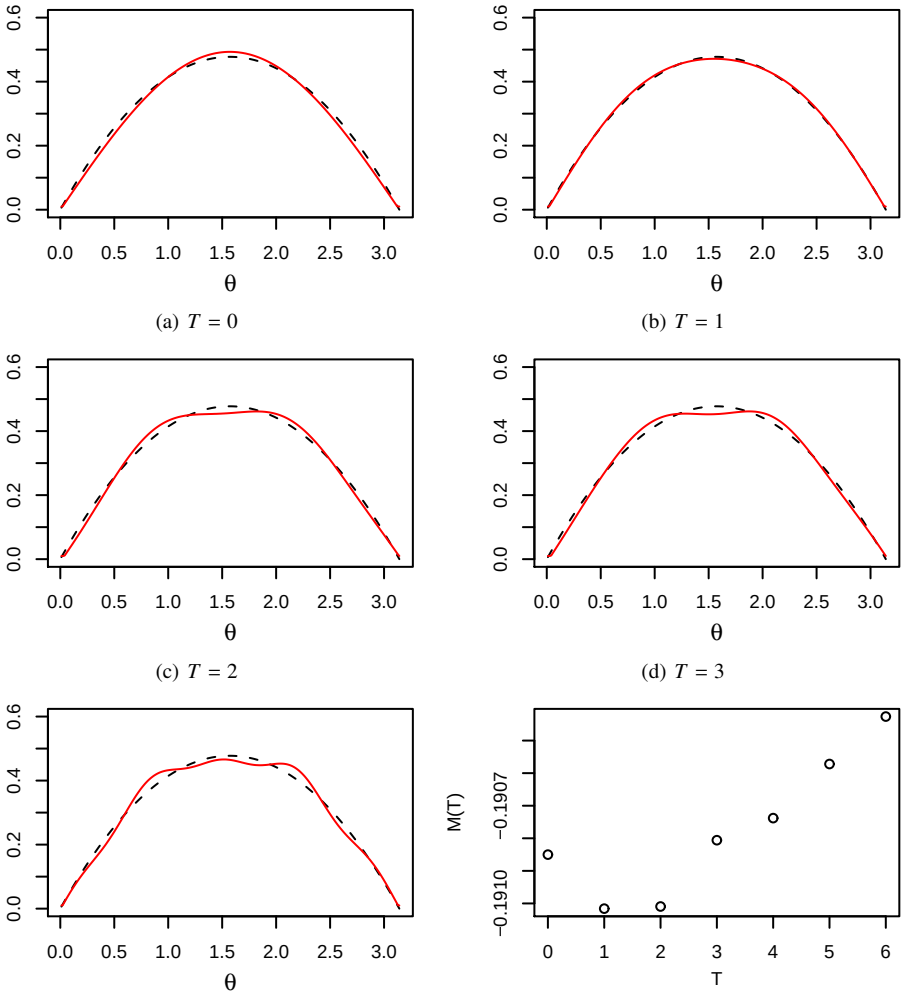
으로 선택할 수 있다. 이를 요약하면 다음과 같다.

1. $\hat{f}_X$을 위한 절단모수 $\hat{T}_X$을 선택한다.
2. $\hat{T}_X$을 이용하여 $\hat{f}_X$을 구한다.
3. f_X을 대신한 $\hat{f}_X$을 사용하여 $\hat{T}_\beta$을 선택한다.
4. $\hat{T}_\beta$을 사용하여 $\hat{f}_\beta^-$을 구한다.
5. 식 (2.13)을 이용하여 f_β의 최종추정치 $\hat{f}_\beta$을 구한다.

3.2. $d = 2$에서의 일반적 성질

$d = 2$일 때 원위의 점 $x = (x_1, x_2) \in \mathbb{S}^1$를 극좌표를 사용하여 $\theta = \arccos(x_1)$, $-\pi \le \theta < \pi$로 표현할 수 있고, $b = (b_1, b_2) \in \mathbb{S}^1$도 $\phi = \arccos(b_1)$을 이용해 구할 수 있다. 이러한 x와 b를 이용하면,

$$x^\top b = \cos(\theta)\cos(\phi) + \sin(\theta)\sin(\phi) = \cos(\theta - \phi)$$



(a) $T = 0$ (b) $T = 1$ (c) $T = 2$ (d) $T = 3$ (e) $T = 4$ (f) $M(T)$

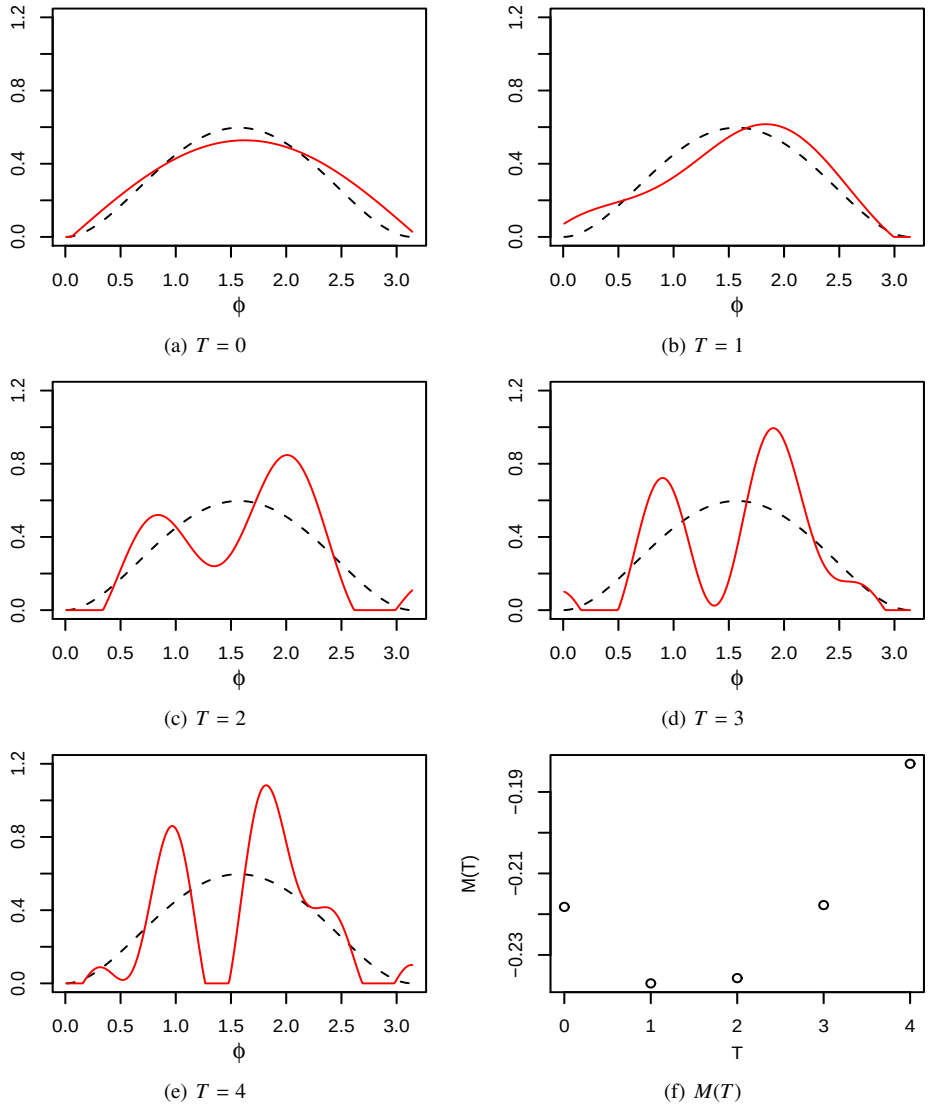
그림 1: *T에 따른 Θ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*

임을 구할 수 있고,

$$\begin{aligned} Kf_\beta(x) = Kf_\beta(\theta) &= \int_{\{b:x^\top b \geq 0\})} f_\beta(b)db \\ &= \int_0^{2\pi} \mathbb{I}\{\cos(\theta - \phi) \geq 0\} f_\beta(\phi)d\phi \end{aligned}$$

이다.

확률변수 X는 $X = (\cos\Theta, \sin\Theta)$로, β는 $\beta = (\cos\Phi, \sin\Phi)$로 나타내고자 하며 f_X는 f_Θ로, f_β는 f_Φ로 표현하고자 한다. $x = (\cos\theta, \sin\theta)$로 주어질 때, 식 (2.3)을 사용하면 $n \geq 1$에서 $h_0 = 1$와 $h_n = 2$임을 알 수 있고, 고유함수(eigenfunction)와 고유값(eigenvalues)은 다음과 같이 표현된다.



(a) $T = 0$ (b) $T = 1$

(c) $T = 2$ (d) $T = 3$

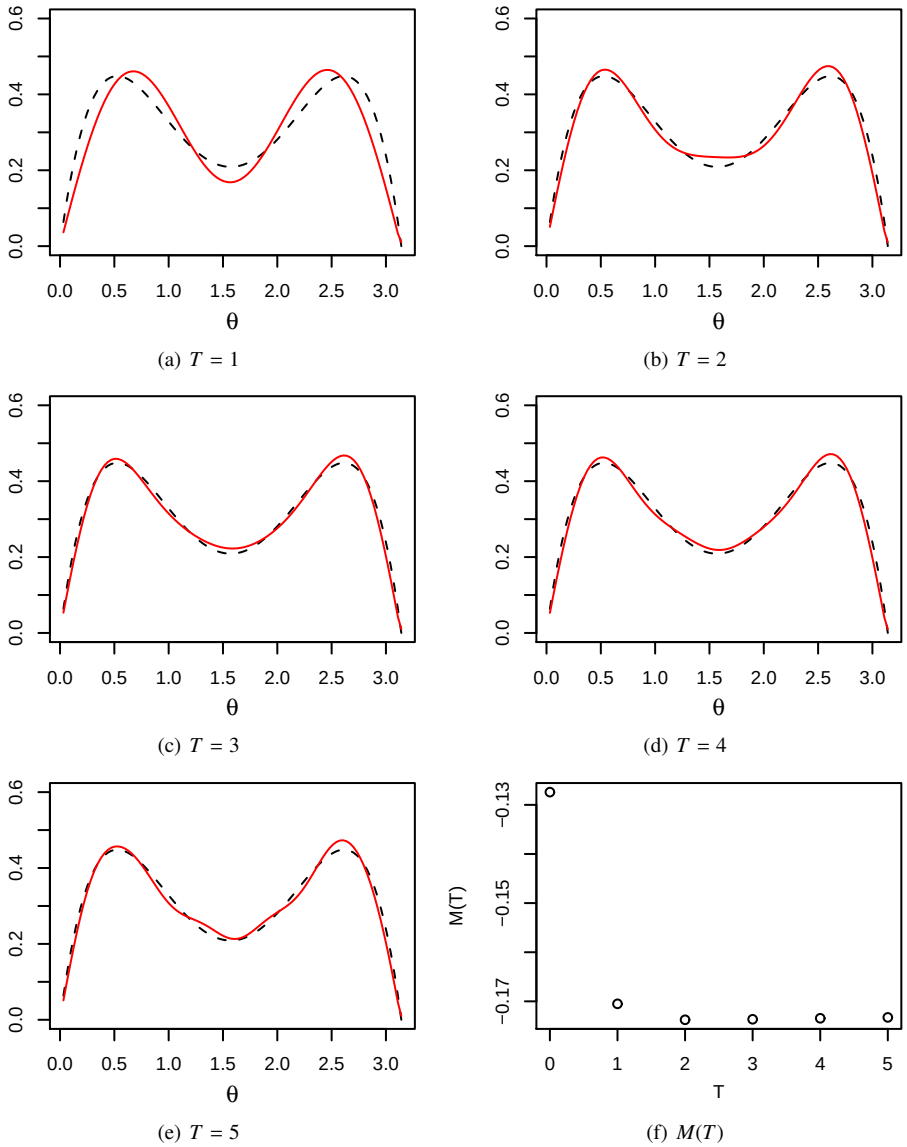
(e) $T = 4$ (f) $M(T)$

그림 2: *T에 따른 Φ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*

- 고유함수 :

$$\psi_n^l(x) = \psi_n^l(\theta) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}, & \text{if } n = 0,\ l = 0, \\ \dfrac{\cos(n\theta)}{\sqrt{\pi}}, & \text{if } n \geq 1,\ l = 1, \\ \dfrac{\sin(n\theta)}{\sqrt{\pi}}, & \text{if } n \geq 1,\ l = 2. \end{cases}$$

- 고유값 : $\lambda_0 = 1$와 $\lambda_n = 2\sin(n\pi/2)/n$, $n = 1, 2, \ldots$.



(a) $T = 1$ (b) $T = 2$

(c) $T = 3$ (d) $T = 4$

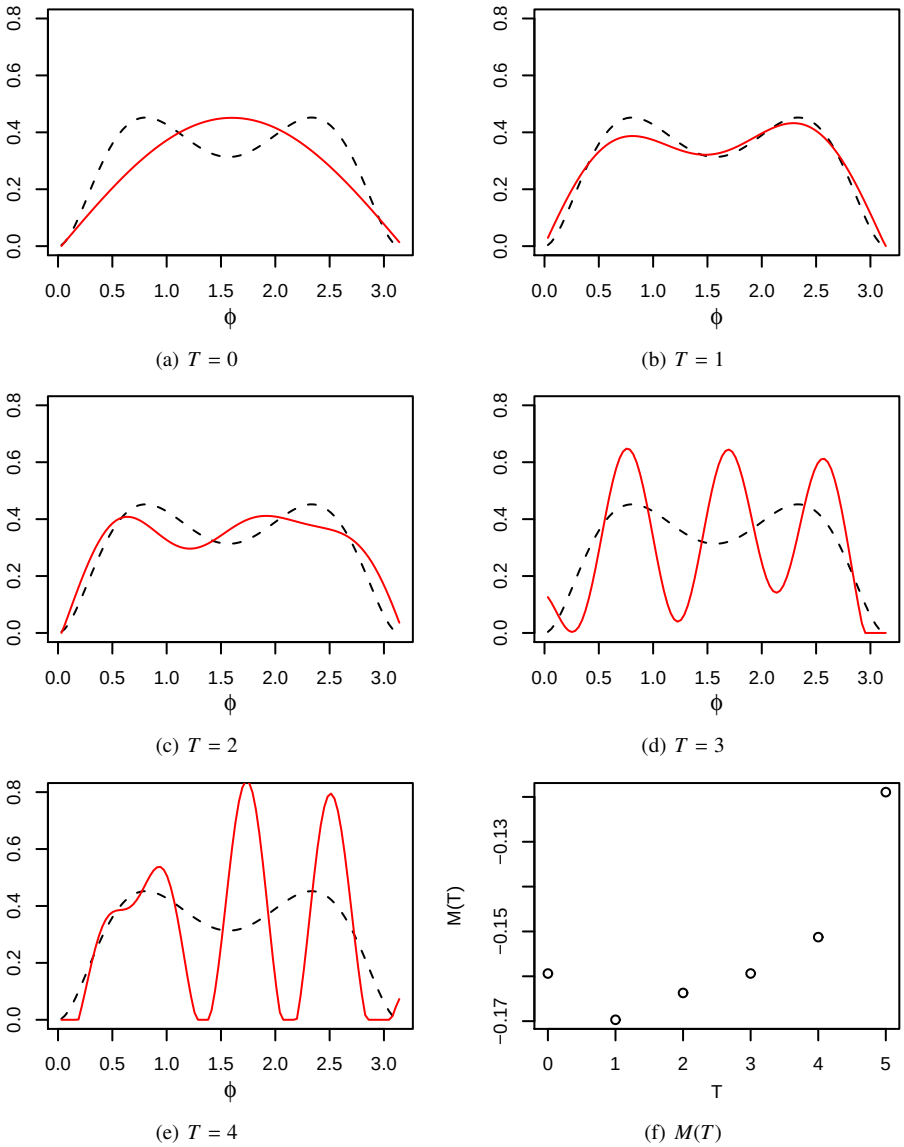
(e) $T = 5$ (f) $M(T)$

그림 3: *T에 따른 Θ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*

4. 모의실험

4.1. Θ의 생성

범위가 $(0,1)$인 베타분포의 범위를 $(0,\pi)$까지 확장하는 변수변환을 통하여, X에 해당하는 Θ를 생성할 수 있다. 본 모의실험은 베타분포의 모수가 각각 $(3,2)$와 $(2,3)$인 혼합분포를 사용하여 3000개의 확률 표본을 생성하여 Θ로 사용하였으며, $(3,4)$와 $(4,3)$을 이용하여 Φ를 생성하였다. 이때의 $Y = 1$의 개수는 2727개이고 $Y = 0$의 개수는 273개이다.



(a) $T = 0$ (b) $T = 1$

(c) $T = 2$ (d) $T = 3$

(e) $T = 4$ (f) $M(T)$

그림 4: *T에 따른 Φ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*

그림 1에서 (a)~(e)는 절단모수 T (0~4)에 대하여 f_Θ를 점선으로 $\hat{f}_\Theta$를 실선으로 나타내었고, (f)는 T가 0부터 6까지 변화할 때 $M(T)$의 변화를 나타낸다. 그림 1의 (f)를 통해 $M(T)$를 최소화하는 절단모수 T를 1로 선택할 수 있다. 그림 2는 Φ의 추정 결과로 T를 1로 선택한다.

그림 3과 4는 모수가 각각 $(6, 2)$와 $(2, 6)$인 혼합 베타분포를 사용하여 Θ를 생성하였고, 모수를 $(7, 3)$와 $(3, 7)$로 하여 β에 해당하는 Φ를 생성하였을 때의 결과를 보여준다. 생성한 확률 표본의 개수는 3000개이며, $Y = 1$의 개수는 2312개이고 $Y = 0$의 개수는 688개이다. 그림 3의 (f)를 통해 $M(T)$를 최소화하는 T를 2로 선택할 수 있다. 그림 4의 Φ의 밀도 추정에서는 T를 1로 선택한다.

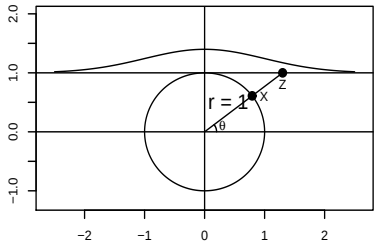
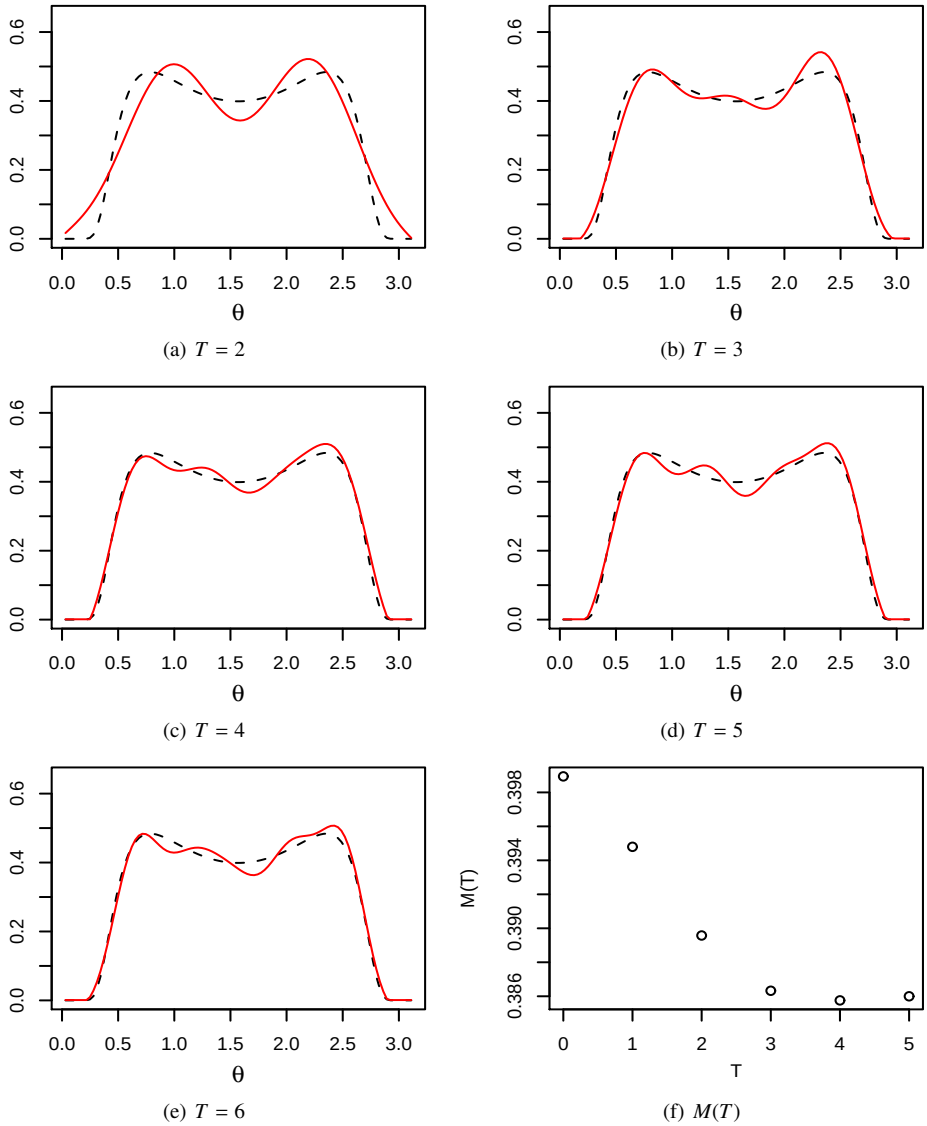


그림 5: *사영을 통한 데이터의 생성*

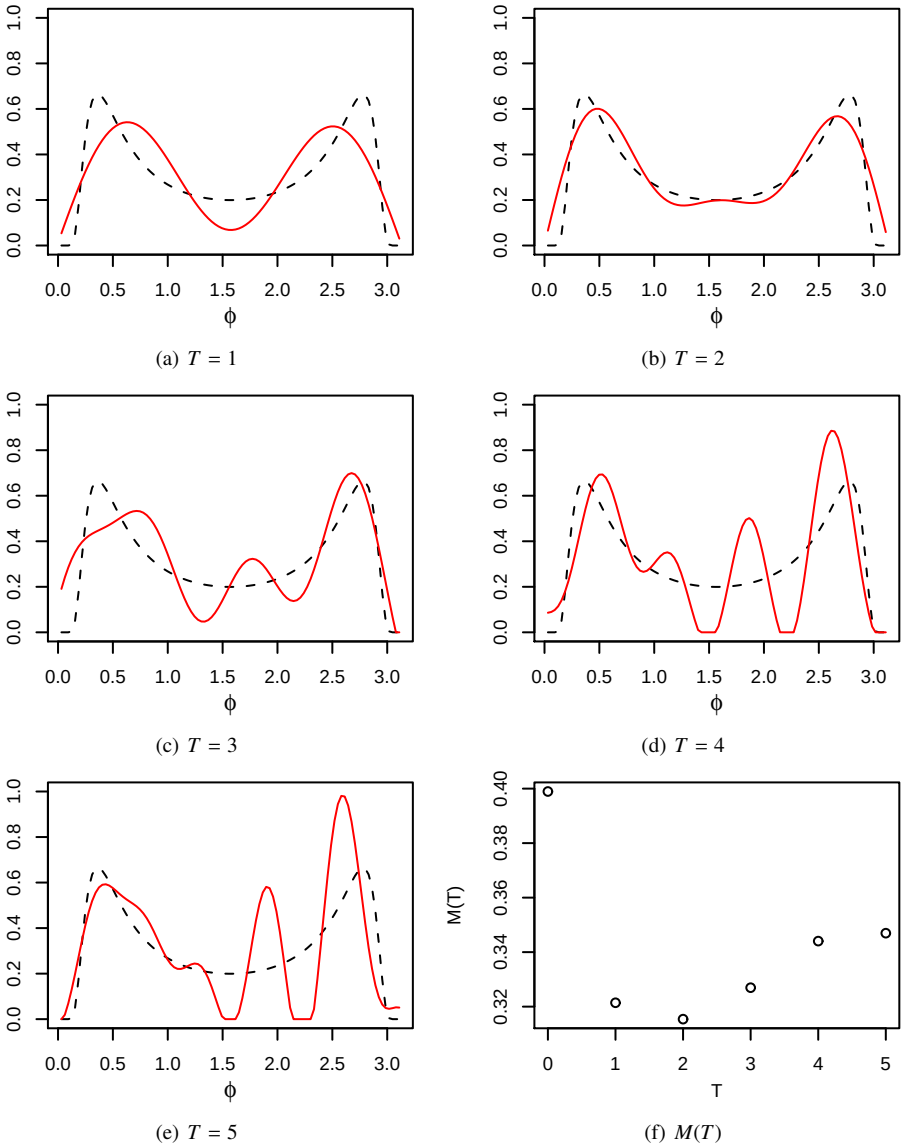


(a) $T = 2$ (b) $T = 3$

(c) $T = 4$ (d) $T = 5$

(e) $T = 6$ (f) $M(T)$

그림 6: *T에 따른 Θ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*



(a) $T = 1$ (b) $T = 2$

(c) $T = 3$ (d) $T = 4$

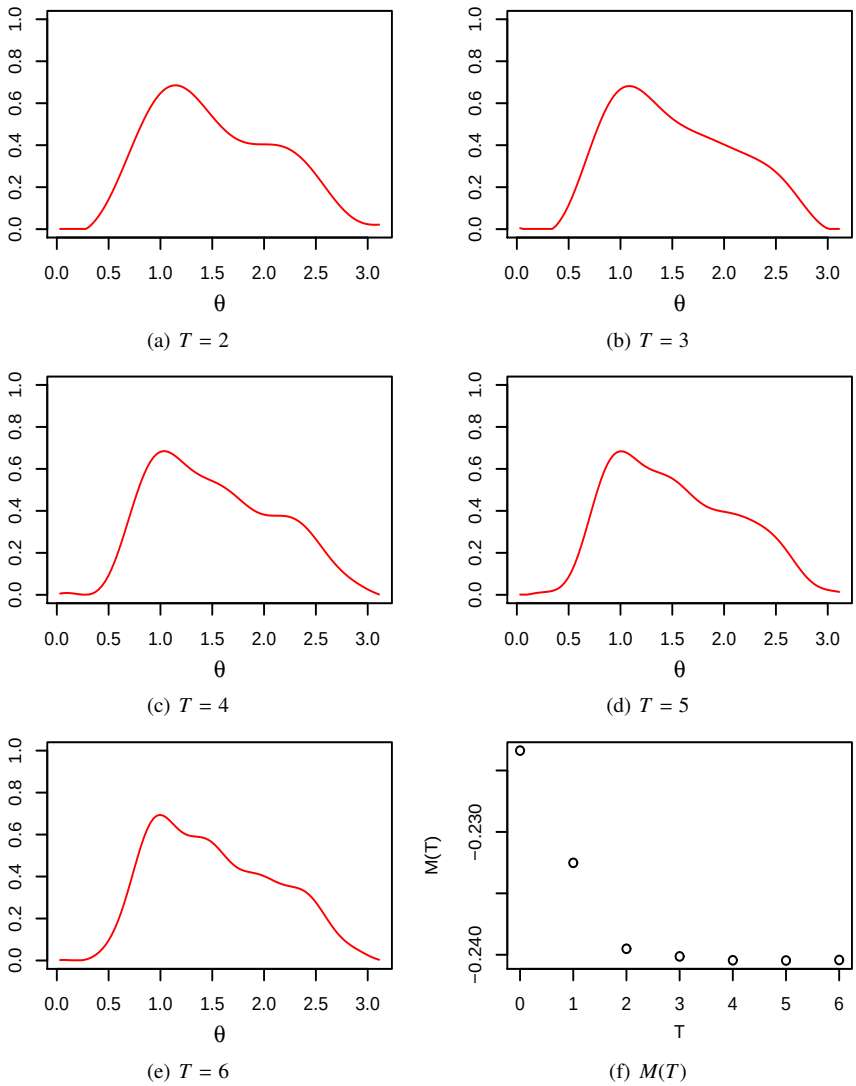
(e) $T = 5$ (f) $M(T)$

그림 7: *T에 따른 Φ의 밀도 추정 및 $M(T)$의 변화 그림으로 점선은 참값을, 실선은 추정값을 나타낸다.*

4.2. 사영을 통한 데이터의 생성

실선에서 존재하는 확률변수 Z를 가정했을 때, X와 Θ는 다음과 같이 Z를 $\mathbb{S}^1$에 사영함으로써 얻을 수 있다.

$$X = \frac{(Z,1)}{\|(Z,1)\|} = \left(\frac{Z}{\sqrt{1+Z^2}}, \frac{1}{\sqrt{1+Z^2}}\right),$$
$$\Theta = \arccos\left(\frac{Z}{\sqrt{1+Z^2}}\right).$$



(a) $T = 2$ (b) $T = 3$ (c) $T = 4$ (d) $T = 5$ (e) $T = 6$ (f) $M(T)$

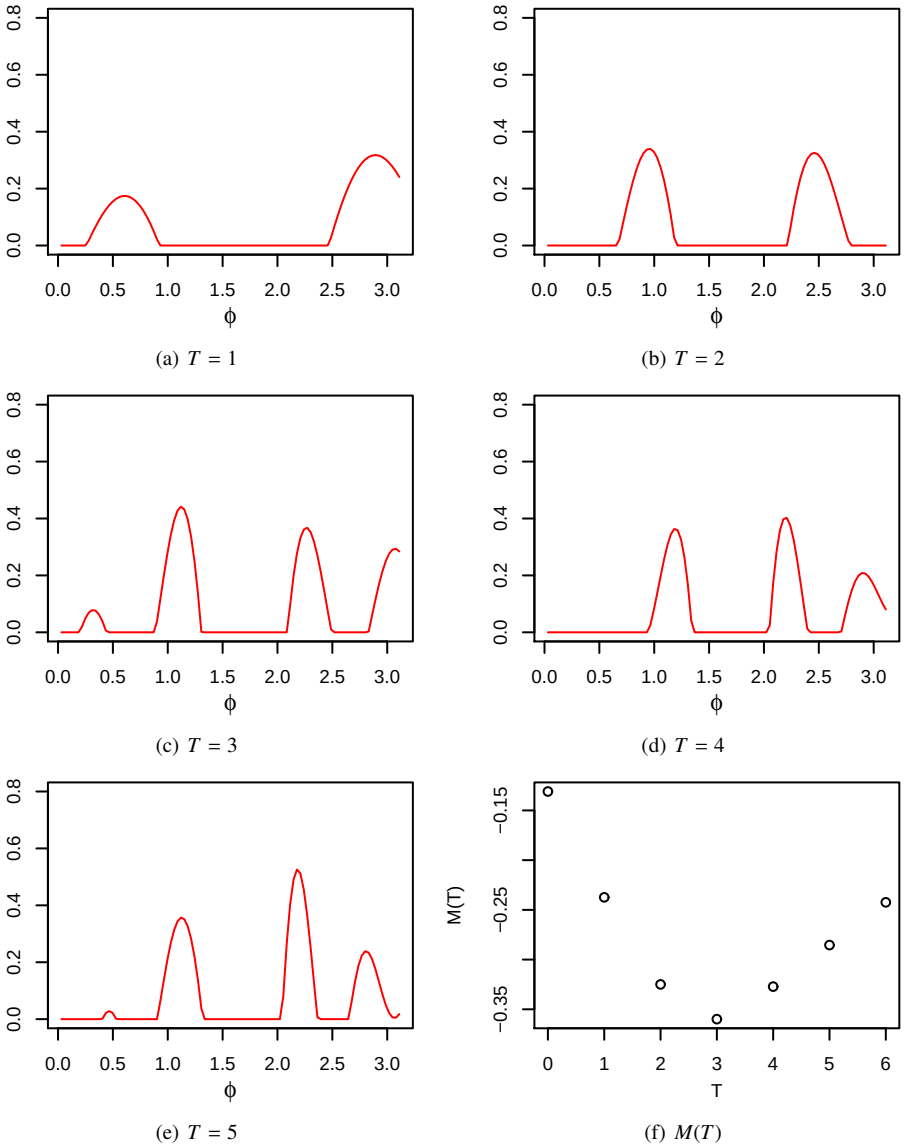
그림 8: *T에 따른 Θ의 밀도 추정 및 $M(T)$의 변화 그림*

β 와 Φ도 동일한 방법으로 얻게 되며, 목표변수 Y는

$$Y = \mathbb{I}\{\cos(\Theta - \Phi) \geq 0\}$$

을 통해 얻을 수 있다. 위의 과정을 도식화 하면 그림 5와 같다.

실험에 사용된 데이터는 다음과 같이 생성하였다. 확률 변수 Z를 표준정규분포로부터 2000개의 확률 표본을 생성하였고, β는 평균이 0이고 표준편차가 2인 정규분포로부터 생성하였다. $Y = 1$의 개수는 1582개이고 $Y = 0$의 개수는 418개이다. 그림 6은 $f_\Theta(x)$의 추정결과로, 부분그림 (f)로부터 최적의 T를 4로 선택할 수 있다. 그림 7은 $f_\Phi(x)$의 추정결과로, 부분그림 (f)로부터 최적의 T를 2로 선택할 수 있다.



(a) $T = 1$ (b) $T = 2$

(c) $T = 3$ (d) $T = 4$

(e) $T = 5$ (f) $M(T)$

그림 9: *T에 따른 Φ의 밀도 추정 및 $M(T)$의 변화 그림*

4.3. 실자료 분석

분석에 사용된 자료는 미국의 1990년부터 1993년까지 이주(migration)와 관련된 자료이다. PSID (Panel Study of Income Dynamics)에서 자료를 얻을 수 있으며, 본 연구에서는 Dong (2010)에서 사용한 4689개의 표본을 사용하였다. 목표변수는 3년간 이주를 하였으면 1의 값을, 이주를 하지 않았으면 0의 값을 갖는다. 사용된 데이터에서 이주를 한 사람은 807명, 이주를 하지 않은 사람은 3882명이고, 가정 2의 조건을 만족시키기 위해 로그수입액을 표준화하여 공변량으로 사용하였고, 사영을 통하여 밀도를 추정하였다.

그림 8은 Θ의 밀도함수 추정 결과이다. (a)~(e)에서 알 수 있듯이 Θ가 0.5 부근에서 밀도가 증가하여 1 근처에서 최대 밀도를 갖고 점차적으로 감소함을 볼 수 있다. 이때 적절한 T는 5로 선택할 수 있다. 그림 9는 Φ의 밀도함수 추정 결과를 나타내고 T는 3으로 선택할 수 있다. 또한 $\beta = (\beta_1, \beta_0) = (\cos\phi, \sin\phi)$이므로 $\beta_1 : \beta_0 = \cos\phi : \sin\phi$의 관계식을 얻을 수 있다. 선택된 (c)의 최빈값 3개를 살펴보면, $\phi \simeq 1$은 약 $\beta_1 : \beta_0 = 1 : \sqrt{3}$의 관계를 얻을 수 있고, $\phi \simeq 2.3$은 $\beta_1 : \beta_0 = 0.746 : -0.666$으로 해석할 수 있으며, $\phi = \pi$일때 $(-1, 0)$로 절편이 0, 기울기가 -1인 관계식을 생각할 수 있다.

5. 결론

본 연구는 Gautier와 Kitamura (2009)의 논문에서 언급한 반구상에서의 이항 선택 모형에 대한 확장으로 절단모수를 사용한 밀도함수의 추정에 대해 알아보았다. 절단모수의 값이 커지면 밀도함수의 편의가 작아지는 대신 변동성이 커지며, 반대로 절단모수의 값이 작아지면 추정의 변동성이 감소하지만 편의가 커짐을 실제 모의실험을 통해 확인하였다. 그렇기 때문에 이항 선택 모형에서 적절한 절단모수 T를 선택하는 것이 중요한 문제임을 알 수 있다.

현재 본 연구에서는 구면조화 $\mathbb{S}^{d-1}$에서 $d = 2$인 경우에 대해 연구를 하였다. 향후에는 본 논문의 연구결과를 $d > 2$로 확장하고자 하며 점근적 최소최대(minimax) 하한(lower bound)를 구하여 Gautier와 Kitamura (2009)에서 구한 상한과 일치하는지 규명해 보아야 할 것이다.

참고 문헌

Athey, S., and Imbens, G.W. (2007). Discrete choice models with multiple unobserved choice characteristics, Preprint.

Bajari, P., Fox, J. and Ryan, S. (2007). Linear regression estimation of discrete choice models with nonparametric distribution of random coefficients, *American Economic Review, Papers and Proceedings,* **97**, 459–463.

Chesher, A. and Santos Silva, J. M. C. (2002). Taste variation in discrete choice models, *Review of Economic Studies,* **69**, 147–168.

Dong, Y. (2010). Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration, *Economics Letters,* **107**, 33–35.

Efromovich, S. (1999). *Nonparametric curve estimation: methods, theory and applications,* Springer.

Gautier, E. and Kitamura, Y. (2009). Nonparametric estimation in random coefficients binary choice models, Manuscript.

Groemer, H. (1996). *Geometric applications of fourier series and spherical harmonics,* Cambridge University Press: Cambridge.

Harding, M. C. and Hausman, J. (2007). Using a laplace approximation to estimate the random coefficients logit model by nonlinear least squares, *International Economic Review,* **48**, 1311–1328.

Healy, D. M., Hendriks, H. and Kim, P. T. (1998). Spherical deconvolution, *Journal of Multivariate Analysis,* **67**, 1–22.

Healy, D. M. and Kim, P. T. (1996). An empirical Bayes approach to directional data and efficient computation on the sphere, *The Annals of Statistics,* **24**, 232–254.

Huh, J., Kim, P. T., Koo, J.-Y. and Park, J. H. (2004). Directional log-density estimation, *Journal of the Korean Statistical Society,* **33**, 255–269.

Kim, P. T. (1998). Deconvolution density estimation on SO(N). *Annals of Statistics,* **23**, 1083–1102.

Kim, P. T. and Koo, J.-Y. (2000). Directional mixture models and optimal estimation of the mixing density, *The Canadian Journal of Statistics,* **28**, 383–398.

Kim, P. T. and Koo, J.-Y. (2002). Optimal spherical deconvolution, *Journal of Multivariate Analysis,* **80**, 21–42.

Kim, P. T., Koo, J.-Y. and Park, H. J. (2004). Sharp minimaxity and sperical deconvolution for super-smooth error distributions, *Journal of Multivariate Analysis,* **90**, 384–392.

Koo, J.-Y. and Kim, P. T. (2005). Statistical inverse problems on manifolds, *The Journal of Fourier Analysis and Applications,* **11**, 639–653.

Koo, J.-Y. and Kim, P. T. (2008a). Asymptotic minimax bounds for stochastic deconvolution over groups, *IEEE Transactions on Information Theory,* **54**, 289–298.

Koo, J.-Y. and Kim, P. T. (2008b). Sharp adaptation for spherical inverse problems with applications to medical imaging, *Journal of Multivariate Analysis,* **99**, 165–190.

Train, K. E. (2003). *Discrete choice methods with simulation,* Cambridge University Press: Cambridge.

2010년 4월 접수; 2010년 8월 채택

Truncation Parameter Selection in Binary Choice Models

Kwang-Rae Kim[a], Kyu-Dong Cho[a], Ja-Yong Koo[1,a]

[a]Department of Statistics, Korea University

Abstract

This paper deals with a density estimation method in binary choice models that can be regarded as a statistical inverse problem. We use an orthogonal basis to estimate density function and consider the choice of an appropriate truncation parameter to reflect the model complexity and the prediction accuracy. We propose a data-dependent rule to choose the truncation parameter in the context of binary choice models. A numerical simulation is provided to illustrate the performance of the proposed method.

Keywords: Choice Probability, density Estimation, inverse Problem, legendre polynomials, spectral decomposition, MISE.

This research was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2008-313-C00127).

[1] Corresponding author: Professor, Department of Statistics, Korea University, Anam-dong 5-1, Seoul 136-701, Korea. E-mail: jykoo@korea.ac.kr