Odporna estymacja w wielowymiarowym rozkładzie normalnym

Agnieszka Kulawik

Doktorantka

Wydział Matematyki, Informatyki i Ekonometrii
Uniwersytet Zielonogórski

7 grudnia 2009 r.

1. Informacje ogólne

Załóżmy, że $J \in \mathbb{N}$. Zdefiniujmy następujące zbiory:

$$\mathcal{G} := \{F\colon \mathbb{R}^J \to \mathbb{R} \mid F \text{ jest dystrybuantą}\},$$

$$\mathcal{D} := \operatorname{lin}\{F - G \mid F, G \in \mathcal{G}\}.$$

1. Informacje ogólne

Załóżmy, że $J \in \mathbb{N}$. Zdefiniujmy następujące zbiory:

$$\mathcal{G} := \{F\colon \mathbb{R}^J \to \mathbb{R} \mid F \text{ jest dystrybuantą}\},$$

$$\mathcal{D} := \mathrm{lin}\{F - G \mid F, G \in \mathcal{G}\}.$$

Zbiór $\mathcal{G}$ będziemy rozpatrywać z metryką supremum:

$$||F - G|| = \sup_{x \in \mathbb{R}^J} |F(x) - G(x)|, \ \ F, G \in \mathcal{G}.$$

1. Informacje ogólne

Załóżmy, że $J \in \mathbb{N}$. Zdefiniujmy następujące zbiory:

$$\mathcal{G} := \{F \colon \mathbb{R}^J \to \mathbb{R} \mid F \text{ jest dystrybuantą}\},$$

$$\mathcal{D} := \operatorname{lin}\{F - G \mid F, G \in \mathcal{G}\}.$$

Zbiór $\mathcal{G}$ będziemy rozpatrywać z metryką supremum:

$$||F - G|| = \sup_{x \in \mathbb{R}^J} |F(x) - G(x)|, \; F, G \in \mathcal{G}.$$

Niech $m \in \mathbb{N}$, a $\Theta \subset \mathbb{R}^m$ jest przestrzenią parametrów. Interesować nas będzie model statystyczny postaci

$$\{F_\theta : \theta \in \Theta\} \subset \mathcal{G}.$$

Definicja

Funkcję $T : \mathcal{G} \to \Theta$ *nazywamy funkcjonałem statystycznym.*

Definicja

Funkcję $T : \mathcal{G} \to \Theta$ *nazywamy funkcjonałem statystycznym.*

Definicja

Funkcjonał statystyczny T *nazywamy zgodnym w sensie Fishera, gdy*

$$T(F_\theta) = \theta, \ \theta \in \Theta.$$

Definicja

Funkcję $T : \mathcal{G} \to \Theta$ nazywamy funkcjonałem statystycznym.

Definicja

Funkcjonał statystyczny T nazywamy zgodnym w sensie Fishera, gdy

$$T(F_\theta) = \theta,\ \ \theta \in \Theta.$$

Definicja

Funkcjonał statystyczny T nazywamy różniczkowalnym w sensie Fréchet w punkcie $F \in \mathcal{G}$, gdy istnieje taki funkcjonał liniowy $T' : \mathcal{D} \to \mathbb{R}^m$, że

$$|T(G) - T(F) - T'(G - F)| = o(||F - G||),$$

gdzie $|\cdot|$ jest metryką euklidesową w $\mathbb{R}^m$.

2. Wielowymiarowy rozkład normalny

Rozpatrujemy J-wymiarowy rozkład normalny

$$N(X\beta, \Sigma),$$

gdzie

- $X \in \mathbb{R}_a^J$ jest znaną macierzą o a niezależnych kolumnach;

2. Wielowymiarowy rozkład normalny

Rozpatrujemy J-wymiarowy rozkład normalny

$$N(X\beta, \Sigma),$$

gdzie

- $X \in \mathbb{R}_a^J$ jest znaną macierzą o a niezależnych kolumnach;
- $\Sigma = \sum_{i=1}^k \sigma_i^2 V_i$, gdzie $V_1, \ldots, V_k \in \mathbb{R}_J^J$ są znanymi symetrycznymi, nieujemnie określonymi i liniowo niezależnymi macierzami;

2. Wielowymiarowy rozkład normalny

Rozpatrujemy J-wymiarowy rozkład normalny

$$N(X\beta, \Sigma),$$

gdzie

- $X \in \mathbb{R}_a^J$ jest znaną macierzą o a niezależnych kolumnach;
- $\Sigma = \sum_{i=1}^k \sigma_i^2 V_i$, gdzie $V_1, \ldots, V_k \in \mathbb{R}_J^J$ są znanymi symetrycznymi, nieujemnie określonymi i liniowo niezależnymi macierzami;
- $\beta = (\beta_1, \ldots, \beta_a)^T \in \mathbb{R}^a$, przy czym $\theta = (\beta_1, \ldots, \beta_a, \sigma_1, \ldots, \sigma_k)^T \in \Theta \subset \mathbb{R}^m$, gdzie $m = a + k$, jest wektorem nieznanych parametrów.

Gęstością rozpatrywanego rozkładu $N(X\beta, \Sigma)$ jest

$$f(y|\theta) = (2\pi)^{-\frac{J}{2}}|\Sigma|^{-\frac{1}{2}}\exp[-\frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)],\ y \in \mathbb{R}^J,$$

Gęstością rozpatrywanego rozkładu $N(X\beta, \Sigma)$ jest

$$f(y|\theta) = (2\pi)^{-\frac{J}{2}}|\Sigma|^{-\frac{1}{2}}\exp[-\frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)], \ y \in \mathbb{R}^J,$$

która po zlogarytmowaniu i pewnym uproszczeniu przyjmie postać

$$L(y|\theta) = \ln|\Sigma|^{\frac{1}{2}} + \frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta), \ y \in \mathbb{R}^J.$$

Gęstością rozpatrywanego rozkładu $N(X\beta, \Sigma)$ jest

$$f(y|\theta) = (2\pi)^{-\frac{J}{2}}|\Sigma|^{-\frac{1}{2}}\exp[-\frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)], \; y \in \mathbb{R}^J,$$

która po zlogarytmowaniu i pewnym uproszczeniu przyjmie postać

$$L(y|\theta) = \ln|\Sigma|^{\frac{1}{2}} + \frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta), \; y \in \mathbb{R}^J.$$

W celu otrzymania estymatora odpornego parametru θ, uogólniamy drugi składnik powyższej sumy i otrzymujemy:

$$\Phi(y|\theta) = \ln|\Sigma|^{\frac{1}{2}} + \varphi[\frac{1}{c^2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)], \; y \in \mathbb{R}^J,$$

gdzie $c > 0$ jest stałą odpowiednio dobraną do funkcji φ: $\mathbb{R}_{\geqslant} \rightarrow \mathbb{R}$.

Funkcjonał statystyczny

Definiujemy funkcjonał statystyczny $T\colon \mathcal{G} \to \Theta$ wzorem

$$T(G) = \mathrm{argmin}_{\theta\in\Theta} \int \Phi(y|\theta)\, dG(y) \quad \text{dla } G \in \mathcal{G},$$

gdzie

$$\Phi(y|\theta) = \ln|\Sigma|^{\frac{1}{2}} + \varphi\left[\frac{1}{c^2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)\right],\ y \in \mathbb{R}^J.$$

Funkcjonał statystyczny

Definiujemy funkcjonał statystyczny $T \colon \mathcal{G} \to \Theta$ wzorem

$$T(G) = \operatorname{argmin}_{\theta \in \Theta} \int \Phi(y|\theta)\, dG(y) \quad \text{dla } G \in \mathcal{G},$$

gdzie

$$\Phi(y|\theta) = \ln |\Sigma|^{\frac{1}{2}} + \varphi[\frac{1}{c^2}(y - X\beta)^T \Sigma^{-1}(y - X\beta)],\ y \in \mathbb{R}^J.$$

(B1) Funkcja φ ma nieujemną pochodną na $[0, +\infty)$.
(B2) Funkcja $x\varphi'(x^2)$ ma nieujemną pochodną na $[0, +\infty)$ oraz istnieje takie $x_0 > 0$, że $2x_0^2\varphi'(x_0^2) > J$.
(B3) Funkcje φ' i φ'' są ograniczone.
(B4) Funkcje $x\varphi'(x^2)$ i $x^2\varphi''(x^2)$ są ograniczone.

Twierdzenie (Dobór stałej c)

Jeżeli funkcja $\varphi\colon \mathbb{R}_{\geqslant} \to \mathbb{R}$ spełnia warunki (B1) *i* (B2) *oraz $(Y_1, \ldots, Y_J)^T$ ma J-wymiarowy rozkład normalny $N(0_J, \mathbb{I}_J)$, to istnieje takie jednoznacznie wyznaczone $c_\varphi > 0$, w którym funkcja*

$$c \longmapsto J \ln(c) + \mathbb{E}\varphi\Big(\sum_{j=1}^{J} \frac{Y_j^2}{c^2}\Big), \; c > 0,$$

osiąga minimum globalne.

Twierdzenie (Zgodność w sensie Fishera)

Jeżeli funkcja $\varphi\colon \mathbb{R}_{\geqslant} \to \mathbb{R}$ *spełnia warunki* (B1) *i* (B2) *oraz* $c = c_{\varphi}$, *to funkcja*

$$\theta \longmapsto \int \Phi(y|\theta)\, dF_{\theta_0}(y), \quad \theta \in \Theta,$$

osiąga minimum globalne w punkcie $\theta = \theta_0$.

Twierdzenie (Zgodność w sensie Fishera)

Jeżeli funkcja $\varphi\colon \mathbb{R}_{\geqslant} \to \mathbb{R}$ *spełnia warunki* (B1) *i* (B2) *oraz* $c = c_\varphi$, *to funkcja*

$$\theta \longmapsto \int \Phi(y|\theta)\, dF_{\theta_0}(y), \quad \theta \in \Theta,$$

osiąga minimum globalne w punkcie $\theta = \theta_0$.

Twierdzenie (Różniczkowalność w sensie Fréchet)

Jeżeli funkcja $\varphi\colon \mathbb{R}_{\geqslant} \to \mathbb{R}$ *spełnia warunki* (B1)-(B4) *oraz* $c = c_\varphi$, *to*

$$T(G) - T(F_\theta) = -M(\theta)^{-1} \int \psi(y|\theta)\, dG(y) + o(||G - F_\theta||),$$

gdzie $\psi(\cdot|\theta) = \frac{\partial \Phi(\cdot|\theta)}{\partial \theta}$, *a* $M(\theta) = \int \{\frac{\partial \psi(y|\theta)}{\partial \theta}\}\, dF_\theta(y)$.

3. Wnioski

Jeżeli funkcja $\varphi\colon \mathbb{R}_{\geqslant} \to \mathbb{R}$ spełnia warunki (B1)-(B4) oraz $c = c_\varphi$, to funkcjonał statystyczny

$$T(G) = \mathrm{argmin}_{\theta\in\Theta} \int \Phi(y|\theta)\, dG(y) \quad \text{dla } G \in \mathcal{G},$$

gdzie

$$\Phi(y|\theta) = \ln|\Sigma|^{\frac{1}{2}} + \varphi[\frac{1}{c^2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)],\ y \in \mathbb{R}^J,$$

jest zgodny w sensie Fishera i różniczkowalny w sensie Fréchet w $F_\theta,\ \theta \in \Theta$.

Estymator odporny parametru $\theta \in \Theta$

Dla $n \in \mathbb{N}$ estymator parametru $\theta \in \Theta$ definiujemy następująco:

$$\hat{\theta}_n := T(\hat{F}_n),$$

gdzie $\hat{F}_n$ jest dystrybuantą empiryczną.

Dla tak zdefiniowanego estymatora mamy

$$\sqrt{n}(\hat{\theta}_n - \theta) = -M(\theta)^{-1}\sqrt{n}\int \Psi(y|\theta)\, d\hat{F}_n(y) + \sqrt{n}||\hat{F}_n - F_\theta||\frac{o(||\hat{F}_n - F_\theta||)}{||\hat{F}_n - F_\theta||},$$

a więc jego rozkład można przybliżać rozkładem normalnym z wartością oczekiwaną θ i macierzą kowariancji postaci

$$\frac{1}{n}M(\theta)^{-1}\{\int \Psi(y|\theta)\Psi(y|\theta)^T\, dF_\theta(y)\}[M(\theta)^{-1}]^T.$$

4. Literatura

- Bednarski, T., Zontek, S., *Robust estimation of parameters in a mixed unbalanced model*, The Annals of Statistics 24 (4), pp.1493-1510, 1996.
- Clarke, B.R., *Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations*, The Annals of Statistics 11 (4), pp.1196-1205, 1983.
- Zmyślony, R., Zontek, S., *Robust M-estimator of parameters in variance components model*, Discussiones Mathematicae Probability and Statistics 22, pp.61-71, 2002.