M-ODHADY ZA PŘEDPOKLADU NEREGULÁRNÍ HUSTOTY

Jan SVATOŠ

MFF UK, KPMS

Abstract: Theory of M-estimators in linear regression model is well known. One of the classical regularity assumptions in the linear model $Y = X\beta + E$ is the existence of f'. This paper studies the case when f' does not exist in the sense, that $f(x) = |x - s_j|^{\alpha_j} q_j(x)$, where $j = 1, \ldots, k$ is finite number of "singular points" of the density f. It is shown, that in such a case rate of the convergence of M-estimators depends on the minimum of α_j.

Резюме: Проблематика M-оценок в линейной регресии хорошо известна и разработана. В работах, которые интересуются этой проблематикой учитывают, что плотность вектора ошибок E в модели $Y = X\beta + E$ регулярна. Это значит, что существует f'. Моя работа рассматривает ситуацию, когда f' не существует в том смысле, что $f(x) = |x - s_j|^{\alpha_j} q_j(x)$, где $j = 1, \ldots, k$ конечное число сингулярных точек плотности. Можно показать, что в таком случае степень скорости сходимости M-оценок изменяется в зависимости от минима α_j.

Předpokládejme platnost lineárního modelu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \tag{1}$$

s podmínkou na matici

$$\sum_1^n ||\mathbf{x}_i||^4 = O(n), \tag{2}$$

kde $||(\cdot)||$ označuje Euklidovskou normu, $\mathbf{x}_i$ označuje i-tý řádek matice a $\mathbf{E} = (\varepsilon_i)_1^n$, ε_i *iid*. M-odhadem neznámého parametru $\boldsymbol{\beta}$ je libovolné řešení minimalizace součtu

$$\sum_1^n \varrho(y_i - \mathbf{x}_i'\mathbf{b}) \text{ vzhledem k } \mathbf{b}. \tag{3}$$

Pokud existuje první derivace funkce ϱ a pokud ϱ je konvexní, pak lze definici (3) zjednodušit na tvar:

(4)
$$\sum_1^n \mathbf{x}_i \psi(y_i - \mathbf{x}_i'\mathbf{b}) \stackrel{!}{=} 0.$$

Uvažujme množinu $S = \{s_1, s_2, \ldots, s_k\}$ kde $-\infty < s_1 < s_2 < \ldots s_k < \infty$, a k ní příslušnou množinu $\alpha_1, \ldots, \alpha_k$, přičemž $\alpha_l \in (0,2)$, $l = 1, \ldots, k$. Označme dále $\alpha_0 = \min_{1 \leqq l \leqq k} \alpha_l$ jako . Nechť hustota chybových složek existuje a nabývá tvaru

(5)
$$f(e) = q_l(e)|e - s_l|^{(\alpha_l - 1)}.$$

Pro $\alpha_l \neq 1$ nechť $q_l(e)$ splňuje Lipschitzovu podmínku řádu $\alpha_0/2$.
Předpoklady na funkci ψ: $\psi = \psi_1 + \psi_2$ přičemž ψ_1 je neklesající, absolutně spojitá funkce, ψ_2 je neklesající skoková funkce. Dále předpokládejme splnění jedné ze dvou následujících podmínek:

(i) ψ_1' je absolutně spojitá a zároveň

(6)
$$\psi_1'' \text{ splňuje podmínku } \int_{\mathbb{R}} \big(\psi_1''(e+t)\big)^2 dF(e) \leqq C$$

při $|t| < \delta$ pro nějaké reálné konstanty $C, \delta > 0$

(ii) ψ_1 je konstantní mimo omezený interval (a,b) a zároveň ψ_1', ψ_1'' jsou omezené uvnitř (a,b).
Předpokládejme navíc, že ψ_1 a ψ_1' jsou obě integrovatelné se čtvercem vzhledem k $F(e)$.

Podmínky na ψ_2:

(7)
$$\psi_2(e) = \begin{cases} a_0 & -\infty < e \leqq s_1 \\ a_1 & s_1 < e \leqq s_2 \\ \ldots & \\ a_k & s_k < e < \infty \end{cases}$$

a

(8)
$$\sum_{l \in K_0} (a_l - a_{l-1}) q_l(s_l) > 0.$$

kde

$$K_0 = \{l : 1 \leqq j \leqq k, \alpha_l = \alpha_0\}.$$

Označme

(9a)
$$\gamma_1 = \int_{\mathbb{R}} \psi_1'(e) dF(e) > 0.$$

(9b)
$$\gamma_2 = \frac{1}{\alpha_0} \sum_{l \in K_0} (a_l - a_{l-1}) q_l(s_l) > 0.$$

Tvrzení dané v tomto článku bude ještě nutno rozšířit na obecnější ψ_1, konkrétně ψ_1' = absolutně spojitá + skoková funkce. Důkaz pro tento případ bude podán v další práci. Problém asymptotického chování M-odhadu studovala v práci [1] Jurečková pro případ parametru polohy jednorozměrného rozdělení. Vzniká přirozený problém, které techniky důkazů projdou i pro regresní parametr a kdy je nutno zesílit podmínky na rozdělení a na ψ funkci. Prvním úkolem je zobecnit výsledek Lemmatu 3.1 z [1].
Za výše uvedených podmínek platí: $\forall\tau \geqq \frac{1}{2\alpha_0}$:

$$\begin{aligned}&\lim_{n\to\infty} P_0\{\max_{||\mathbf{t}||\leqq C} \Big|\Big|n^{-\frac{1}{2}}\sum_1^n \mathbf{x}_i[\psi(\varepsilon_i + n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi(\varepsilon_i)] \\ &- n^{\frac{1}{2}-\tau\alpha_0}\gamma_2\frac{1}{n}\sum_1^n \mathbf{x}_i\mathbf{x}_i'(|t_j|^{\alpha_0}\operatorname{sign} t_j)_{j=1}^p \\ &- n^{\frac{1}{2}-\tau}\gamma_1\frac{1}{n}\sum_1^n \mathbf{x}_i\mathbf{x}_i'\mathbf{t}\Big|\Big| \geqq \varepsilon\} = 0 \quad \forall\varepsilon > 0, C > 0\end{aligned}$$

Pro pevné $\mathbf{t}$ důkaz prochází podobně jako v [1]. Za prvé je nutno ukázat, že rozptyl výrazu je asymptoticky zanedbatelný pro spojitou i diskrétní část. Za druhé je zapotřebí ukázat, že aproximace střední hodnoty se od skutečné hodnoty liší pro $n \to \infty$ pouze zanedbatelně. Pro řadu nerovností a omezení, které budou ukázány, je postačující podmínkou $\sum ||\mathbf{x}_i||^4 = O(n)$, jelikož $\tau \geqq \frac{1}{4}$. Nechť platí $n^{-\tau}\mathbf{x}_i'\mathbf{t} > 0$. Pro případ opačného znaménka platí následující technika stejným způsobem, pouze se prohodí integrační meze a na pravé straně poslední nerovnosti bude výraz $n^{-\tau}\mathbf{x}_i'\mathbf{t} > 0$ v absolutní hodnotě.

$$\begin{aligned}&var\,[\psi_1(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_1(\varepsilon_i)] \\ &\leqq E\left[\int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}}\int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} \psi_1'(\varepsilon_i - u)\psi_1'(\varepsilon_i - v)dudv\right] \\ &\leqq \int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} \sqrt{E[\psi_1'(\varepsilon_i - u)]^2[\psi_1'(\eta_i - v)]^2}dudv = \\ &= \left\{\int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} \sqrt{E[\psi_1'(\varepsilon_i - u)]^2}du\right\}^2 \\ &\leqq n^{-\tau}\mathbf{x}_i'\mathbf{t}\int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} E[\psi_1'(\varepsilon_i - u)]^2du \leqq (n^{-\tau}\mathbf{x}_i'\mathbf{t})^2\end{aligned}$$

Z toho plyne, že $\forall j \in \{1,\dots,p\}$ platí:

$$var\, n^{-\frac{1}{2}} \sum_i x_{ij}[\psi_1(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_1(\varepsilon_i)]$$
$$\leqq \frac{1}{n}\sum_i x_{ij}^2 (n^{-\tau}\mathbf{x}_i'\mathbf{t})^2$$
$$\leqq C_1 n^{-2\tau} n^{-1} ||t||^2 \sum_i ||\mathbf{x}_i||^4$$

A tedy postačující podmínkou pro konvergenci k nule je v tomto případě $n^{-\frac{3}{2}+\delta}\sum_i ||\mathbf{x}_i||^4 = O(1)$, $\delta > 0$.
Pro diskrétní složku budeme nejprve pracovat s

$$\psi_2(e) = \begin{cases} 0 & e \leqq r \\ 1 & e > r \end{cases}$$

Tedy

$$\psi_2(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_2(\varepsilon_i) = \begin{cases} 1 & r + n^{-\tau}\mathbf{x}_i'\mathbf{t} < \varepsilon_i < r \\ -1 & r + n^{-\tau}\mathbf{x}_i'\mathbf{t} > \varepsilon_i > r \\ 0 & \text{jinak} \end{cases}$$

Nechť $n^{-\tau}\mathbf{x}_i'\mathbf{t} > 0$. Pak

$$P\big(\psi_2(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_2(\varepsilon_i) = -1\big) = F(r + n^{-\tau}\mathbf{x}_i'\mathbf{t}) - F(r)$$
$$= \int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} f(r+u)du = \int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} q_l(r+u)u^{\alpha_l - 1}du$$
$$\leqq \int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} (|q_l(r+u) - q_l(r)| + |q_l(r)|)u^{\alpha_l-1}du$$
$$\leqq u^{\varepsilon_l + \alpha_l - 1}du + C_l \int_0^{n^{-\tau}\mathbf{x}_i'\mathbf{t}} u^{\alpha_l - 1}du \leqq K_l^*(n^{-\tau}\mathbf{x}_i'\mathbf{t})^{\alpha_l}$$

Tedy dostáváme, že platí

$$var\, n^{-\frac{1}{2}}\sum_i x_{ij}[\psi_2(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_2(\varepsilon_i)] \leqq Kn^{-1}\sum_i x_{ij}^2 n^{-\tau\alpha_l}|(\mathbf{x}_i'\mathbf{t})|^{\alpha_l}$$
$$\leqq Kn^{-\tau\alpha_0}n^{-1}||\mathbf{t}||^{\alpha_0}\sum x_{ij}^2||\mathbf{x}_i||^{\alpha_0} \leqq K_2 n^{-\tau\alpha_0}n^{-1}\sum ||\mathbf{x}_i||^{2+\alpha_0}$$

Vzhledem k definici τ platí $\frac{\mathbf{x}_i'\mathbf{t}}{n^{\tau}} \to 0$. Nyní je zapotřebí ukázat, jak malé jsou rozdíly mezi středními hodnotami a jejich aproximacemi.

$$n^{-\frac{1}{2}}\sum_i x_{ij}E[\psi_1(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_1(\varepsilon_i) - n^{-\tau}\gamma_1\mathbf{x}_i'\mathbf{t}]$$
$$\leqq Kn^{-\frac{1}{2}}\sum_i |x_{ij}|\,|\mathbf{x}_i'\mathbf{t}|^2 n^{-2\tau}$$
$$\leqq Kn^{-\frac{1}{2}}n^{-2\tau}||\mathbf{t}||^2\sum ||\mathbf{x}_i||^3$$

Pro diskrétní složku nabývá rozdíl tvaru:

$$
\begin{aligned}
&E[\psi_2(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_2(\varepsilon_i)] = \text{ asymptoticky} \\
&\sum_{j=1}^{k}(a_j - a_{j-1})\{F(s_j) - F(s_j - n^{-\tau}\mathbf{x}_i'\mathbf{t})\} \\
&= \sum_{j=1}^{k}(a_j - a_{j-1})\int_{-n^{-\tau}\mathbf{x}_i'\mathbf{t}}^{0} f(s_j+u)du \\
&= \sum_{j=1}^{k}(a_j - a_{j-1})\int_{-n^{-\tau}\mathbf{x}_i'\mathbf{t}}^{0} |u|^{\alpha_j-1}q_j(s_j+u)du
\end{aligned}
$$

Poznámka ke slovu "asymptoticky": Střední hodnota na levé straně rovnice ve skutečnosti nabývá hodnoty

$$\sum_k \sum_j (a_j - a_{j-k})P\{\varepsilon_i + n^{-\tau}\mathbf{x}_i'\mathbf{t} > s_j \& \varepsilon_i < s_{j-k+1}\}$$

Pokud ovšem $n^{-\tau}\mathbf{x}_i'\mathbf{t} \to 0$, což je splněno z podmínky (2) vždy, když $\tau > 1/4$, pak $\exists n_0 : P\{\varepsilon_i + n^{-\tau}\mathbf{x}_i'\mathbf{t} > s_j \& \varepsilon_i < s_{j-k+1},\, k \geqq 2\} = 0\, \forall ||t|| \leqq C,\, n \geqq n_0$ a tedy není nutno uvažovat skoky většího než prvního řádu. Pro $\tau = 1/4$ však není podmínka (2) postačující ke splnění $n^{-\tau}\mathbf{x}_i'\mathbf{t} \to 0$ a musí tedy být zesílena podmínka na matici $\mathbf{X}$. Odečtením aproximace na konec dostáváme

$$
\begin{aligned}
&||n^{-\frac{1}{2}}\sum_i \mathbf{x}_i\{E[\psi_2(\varepsilon_i - n^{-\tau}\mathbf{x}_i'\mathbf{t}) - \psi_2(\varepsilon_i)] - \\
&-\frac{1}{\alpha_0}n^{-\tau\alpha_0}\sum_{j\in K_0}(a_j - a_{j-1})q_j(s_j)(\mathbf{x}_i'\mathbf{t})^{\alpha_0}\}|| \\
&\leqq K_1 n^{-\frac{1}{2}}\sum_i ||x_i||n^{-\frac{3}{4}}(\mathbf{x}_i'\mathbf{t})^{\alpha_0+\varepsilon} + \\
&K_2 n^{-\frac{1}{2}}\sum_i ||\mathbf{x}_i||n^{-\frac{\alpha_1}{2\alpha_0}}(\mathbf{x}_i'\mathbf{t})^{\alpha_1} \\
&\leqq K_1||\mathbf{t}||^{\alpha_0+\varepsilon}n^{-\frac{1}{4}}n^{-1}\sum_i ||\mathbf{x}_i||^{1+\alpha_0+\varepsilon} + \\
&+ K_2||\mathbf{t}||^{\alpha_1}n^{-\delta}n^{-1}\sum_i ||\mathbf{x}_i||^{1+\alpha_1}
\end{aligned}
$$

kde $\alpha_1 = \min_l\{\alpha_l \neq \alpha_0\}$.
Pro $\alpha_0 = \alpha_l\, \forall l$ druhý člen zmizí. Pro konvergenci celého výrazu k nule je nyní postačující pomínkou (2), protože bez újmy na obecnosti lze brát $\varepsilon = \alpha_0/2$.

Toto lemma je užitečným nástrojem pro důkaz hlavní věty, která ukazuje řád konvergence. Věta říká:

$$n^{\frac{1}{2\alpha_0}}||M_n - \boldsymbol{\beta}|| = O_p(1)$$

přičemž M_n je M-odhad definovaný ve (4), hustota splňuje (5) a funkce ψ vyhovuje podmínkám (6) až (9a,b). Zde nelze přímo použít techniku z práce [1], jelikož problém má obecně více než jeden rozměr. Proto je nutné k důkazu věty použít jiný způsob. Hlavní myšlenka tohoto postupu je ukázána v práci [2]. Nechť ψ je absolutně spojitá, tedy nechť $\psi = \psi_1$. Pak

$$P\{\sup_{||\mathbf{t}||=C} \mathbf{t}' \sum_i \mathbf{x}_i' \psi(\varepsilon_i - n^{-1/2\alpha_0}\mathbf{x}_i'\mathbf{t}) \geqq 0\} \to 0$$

dává, že řešení úlohy

$$\sum_i \mathbf{x}_i \psi(\varepsilon_i - n^{-1/2\alpha_0}\mathbf{x}_i'(\mathbf{M}_n - \boldsymbol{\beta})) \stackrel{!}{=} 0$$

leží v kouli se středem v $\mathbf{0}$ a poloměrem C. Pro dokončení důkazu je použita věta 6.3.4. z práce [3]. Konvergence se ukáže pomocí stejnoměrného omezení výrazu

$$P\{\sup_{||\mathbf{t}||=C} \mathbf{t}' \sum_i \mathbf{x}_i' \psi(\varepsilon_i - n^{-1/2\alpha_0}\mathbf{x}_i'\mathbf{t}) \geqq 0\}.$$

A nyní dostáváme výsledek ze skutečnosti, že řešení výše uvedeného problému je ekvivalentní skutečnosti, že $\mathbf{M}_n$ je M-odhad.

References

1. Jurečková, J., *Asymptotic Behavior of M-estimators of Location in Nonregular Cases*, Statistics & Decisions **1** (1983), 323-340.
2. Jurečková, J. and Sen, P.K., *Robust Statistical Procedures: Asymptotics and Interrelations*, 1996.
3. Ortega, J.M. and Rheinboldt, W. C., *Iterativa Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.