EMPIRICKÝ PROCES

Petr Lachout
ÚTIA ČSAV, Pod vodárenskou věží 4, 182 08 Praha 8

V současné době se zabývám asymptotickými vlastnostmi empirických procesů typu

$$\left\{ \sum_{i=1}^{n} f_i(X_i,t);\ \alpha_j \le t_j \le \beta_j,\ j = 1,\dots,d \right\},$$

kde $X_1,X_2,\dots,X_n$ jsou iid náhodné vektory a
$f_1,f_2,\dots,f_n$ jsou reálné funkce.

Pro speciální skokovitý typ funkcí $f_1,f_2,\dots,f_n$ se mi podařilo odvodit konvergenční výsledky. Obecný výsledek je obsahem připravovaného článku. Mým příspěvkem do letošního ROBUSTu je aplikace dosažených výsledků na případ empirické distribuční funkce v modelu regrese.

Jsme v situaci $Y_i = X_i\beta + e_i$, $i \in N$, kde
$Y_1,Y_2,\dots$ jsou pozorované náhodné vektory
$X_1,X_2,\dots$ nenáhodné matice,
β je neznámý vektor parametrů,
$e_1,e_2,\dots$ jsou iid náhodné vektory s neznámou distribuční funkcí F.

Naším úkolem je odhadnout neznámou distribuční funkci F. Sestavíme empirickou distribuční funkci $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I[e_i<t]$. Chyby $e_1,e_2,\dots$ však neznáme a tak je musíme odhadnout. Sestrojíme nějaký odhad $\hat{\beta}_n$ vektoru β na základě $Y_1,X_1,\dots,Y_n,X_n$ a dostaneme

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\left[Y_i - X_i\hat{\beta}_n < t\right] = \frac{1}{n}\sum_{i=1}^{n} I\left[e_i < t + X_i(\hat{\beta}_n-\beta)\right].$$

Abychom vyšetřili chování procesu $\hat{F}_n$, je nutné se zabývat procesem s dalším parametrem $s \in R^k$, kde k je rozměr vektoru β.

$$F_n(t,s) = \frac{1}{n}\sum_{i=1}^{n} I\left[e_i < t + \psi(n)\, X_i s\right],$$

kde $\hat{\beta}_n = \beta = O_p(\psi(n))$.

Dále budeme předpokládat $\sup_{i=1}^{+\infty} \|X_i\| < +\infty$, $\lim_{n\to+\infty} \sqrt{n}\,\psi(n) = \tilde{\psi} \in R$
a symbolem ∇f budeme označovat gradient funkce f, který je chápán jako řádek.

Tvrzení 1: Nechť F je spojitá v každém bodě t, $a_j \le t_j \le b_j$, $j = 1,\dots,d$; kde $-\infty \le a_j < b_j \le +\infty$ jsou pevně zvolené. Potom

$$\left\{\sqrt{n}\,(F_n(t,s) - F(t)) - \tilde{\psi}\nabla F(t)\,\bar{X}_n s,\ a_j \le t_j \le b_j,\ j = 1,\dots,d,\ s \in R^k\right\} \xrightarrow[n\to+\infty]{d} W,$$

kde W je gaussovský proces s nulovou střední hodnotou a kovarianční funkcí

$$H(t_1,s_1;t_2,s_2) = F(t_1 \wedge t_2) - F(t_1)F(t_2)$$

a $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Konvergence procesů v distribuci však nedává žádnou informaci o chování trajektorií procesu jako celku. Proto je třeba vyšetřit silnější typ konvergence, konvergenci v prostoru D. Zde však nastává nepříjemná technická potíž. Proces $F_n(.,.)$ je prvkem prostoru D pouze v případě, že $d = k$, $X_i = \mathrm{diag}\,\varrho_i$ a

ϱ_i jsou nezáporné vektory. Takový předpoklad je velmi svazující a připouští de facto pouze regresi s jedním parametrem.

Použijme proto aproximaci

$$\frac{1}{n} \sum_{i=1}^{n} I\left[e_i < t + \psi(n)\ \bar{X}_i s - c\psi(n)\ Q_{i,n}\right] \le F_n(t,s) \le$$

$$\le \frac{1}{n} \sum_{i=1}^{n} I\left[e_i < t + \psi(n)\ \bar{X}_i s + c\psi(n)\ Q_{i,n}\right], \quad t \in R^d,\ -c \le s_j \le c,\ j = 1,\dots,k,$$

kde $\quad Q^j_{i,n} = \sum_{\ell=1}^{k} \left|X_i^{j,\ell} - \bar{X}_n^{j,\ell}\right|$.

Tvrzení 2: Nechť $-\infty \le a_j < b_j \le +\infty$, $j = 1,\dots,d$ takové, že F má spojitý gradient v každém bodě t, $a_j \le t_j \le b_j$, $j = 1,\dots,d$ a když nějaké $a_j = -\infty$ nebo $b_j = +\infty$, pak $\lim_{\substack{\|t\|\to+\infty \\ a_j\le t_j\le b_j,\ j=1,\dots,d}} \nabla F(t) = 0$.

Existuje ε, $D > 0$ tak, že $\quad \forall\ B = \bigtimes_{j=1}^{d} (\alpha_j + \gamma_j\rangle$, $\alpha_j = -\infty$ nebo $\alpha_j \ge a_j$, $a_j \le \gamma_j \le b_j$ pro $j = 1,\dots,d$ a $\quad \forall\ t \in R^d$, $\|t\| < \varepsilon$ je $F(B+t) \le DF(B)$.

Nechť navíc ještě F má spojité všechny jednorozměrné marginály. Potom

$$\left(\sqrt{n}\ \left(\frac{1}{n} \sum_{i=1}^{n} I\left[e_i < t + c\psi(n)\ Q_{i,n}\right] - F(t)\right) + c\tilde{\psi}\nabla F(t)\bar{Q}_n;\ a_j\le t_j\le b_j,\ j=1,\dots,d\right)$$
$$\xrightarrow[n\to+\infty]{} W \quad v\ D\left(\bigtimes_{j=1}^{d} \langle a_j, b_j\rangle\right)$$

a

$$\left(\sqrt{n}\ \left(\frac{1}{n} \sum_{i=1}^{n} I\left[e_i < t + c\psi(n)\ Q_{i,n}\right] - F(t)\right) - c\tilde{\psi}\nabla F(t)\bar{Q}_n;\ a_j\le t_j\le b_j,\ j=1,\dots,d\right)$$
$$\xrightarrow[n\to+\infty]{} W \quad v\ D\left(\bigtimes_{j=1}^{d} \langle a_j, b_j\rangle\right)$$

kde $\bar{Q}_n = \frac{1}{n} \sum_{i=1}^{n} Q_{i,n}$

a W je gaussovský proces s nulovou střední hodnotou a kovarianční funkcí

$$H(t_1;t_2) = F(t_1 \wedge t_2) - F(t_1)F(t_2).$$

Konvergence odvozená v tvrzení 2 už dává jistou představu o chování trajektorií procesu $F_n(.,.)$ a to například

$$\sup_{\substack{a_j\le t_j< b_j \\ -c\le s_i\le c \\ j=1,\dots,d \\ i=1,\dots,k}} \sqrt{n}\ \left|F_n(t,s) - F(t)\right| = O_P(1)$$

když jsou splněny předpoklady tvrzení 2 pro $a'_j := a_j - \eta$, $b'_j = b_j + \eta$, pro nějaké $\eta > 0$.

Pro původní problém to znamená, že když si předepíšeme $\alpha \in (0,1)$, pak pro každé n najdeme $c > 0$ tak, že

$$P\left(|\hat{\beta}_n - \beta| > c\psi(n)\right) \ge \alpha$$

$$\sup_{\substack{a_t\le t_j\le b_j \\ j=1,\dots,d \\ |\hat{\beta}_n-\beta|\le c\psi(n)}} \sqrt{n}\ \left|\hat{F}_n(t) - F(t)\right| = O_P(1).$$

LITERATURA

[1] P. J. Bickel and M. J. Wichura: Convergence criteria for multiparameter stochastic processes and some application, Ann. Math. Statist 42 (1971), 1656-1670.

[2] P. Lachout: Billingsley-type tightness criteria for multiparameter stochastic processes, Kybernetika, 24, 5 (1988), 363-371.

[3] G. Neuhaust: On weak convergence of stochastic process with multidimensional time parameter, Ann. Math. Statist. 42 (1971), 1285-1295.

[4] M. L. Straf: A general Skorohod space and its applications to the weak convergence of stochastic processes with several parameters, Ph.D. Dissertation, Univ. of Chicago, 1969.