2010 © JMF

LOKÁLNÍ REGRESNÍ HLOUBKA

Lukáš Kotík
kotik@utia.cas.cz
Ústav teorie informace a automatizace, AV ČR, v. v. i.
Oddělení stochastické informatiky

Abstrakt: *Poster představí definici Rousseeuwovy regresní hloubky a "hloubkových" regresních kvantilů. Oba tyto pojmy vychází z klasické definice jednorozměrných kvantilů a jsou jejím přirozeným zobecněním na lineární regresní model za účelem odhadu kvantilů odezvy při dané hodnotě regresorů, tedy podmíněných kvantilů. Nejhlubší regresní přímka je robustním odhadem v regresi a oproti odhadu založenému na minimalizaci L_1 normy (kvantilová regrese) je velice málo citlivá na odlehlá pozorování. Stejnou vlastnost mají i regresní kvantily založené na této hloubce. S přihlednutím k dalším vlastnostem jako je ekvivariance vůči afinním transformacím, knzistence, atd. jsou tyto metody velmi zajímavou alternativou k dnes velmi často používané kvantilové regresi. Druhá část posteru je věnována lokalizaci regresní hloubky, která nám umožní odhad podmíněných kvantilů v neparametrické regresi. Na tuto metodu můžeme nahlížet jako na rozšíření jádrových odhadů kvantilů.*

ÚVOD

Předpoklady: Uvažujme náhodný vektor (Y, X) se spojitým rozdělením P_{YX}, kde $Y \in \mathbb{R}$ a $X \in \mathbb{R}^p$. Dále mějme data

$$(Y_1, X_1), \ldots, (Y_n, X_n), \quad \text{kde } (Y_i, X_i) \sim \mathsf{P}_{YX},\ i = 1, \ldots, n.$$

Cíl: Chceme na základě dat odhadnout $\xi_\tau(x)$ podmíněný τ-kvantil Y při dané hodnotě $X = x$. Hledáme tedy pro všechna $x \in \mathbb{R}^p$ takovou hodnotu $\xi_\tau(x)$ pro kterou platí

$$\mathsf{P}(Y \le \xi_\tau(x) | X = x) = \tau.$$

REGRESNÍ HLOUBKA, REGRESNÍ KVANTILY

V lineárním modelu

$$Y_i = \beta_0 + \beta_1^T X_i + e_i, \quad i = 1, \ldots, n,$$

kde e_i jsou náhodné velčiny, máme hned několik možností. První z nich je kvantilová regrese, viz např. [2]. Klasický jednorozměrný τ-kvantil pro Y lze získat minimalizací

$$\min_{a \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - a), \quad \text{kde} \quad \rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\}).$$

Přenesením této definice do lineární regrese, tedy řešením minimalizace

$$\hat\beta(\tau) := \arg \min_{(b_0, b_1) \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_\tau(Y_i - b_0 - b_1^T X_i),$$

získáme odhad podmíněného kvantilu při daném $X = x$ jako $\xi_\tau(x) = \hat\beta_0(\tau) + \hat\beta_1^T(\tau)x$. Tento odhad má mnoho pěkných vlastností, jeho největší nevýhodou je poměrně vysoká citlivost na odlehlá pozorování.

Regresní hloubka:

Druhou možností jak definovat jednorozměrný τ-kvantil je jako takovou hodnotu pod kterou leží $\lfloor n\tau \rfloor$ pozorování. Z této definice vychází regresní hloubka. V jednorozměrném případě lze medián definovat jako bod u kterého musíme z výběru odstranit největší počet pozorování, aby se ocitl na okraji výběru (odstraníme polovinu pozorování). Tuto "krajní" pozici budeme nazývat **nonfit poloha**. Uveďme si nejdříve jak rozšíření této definice bude vypadat pro $X_i \in \mathbb{R},\ i = 1, \ldots, n$.

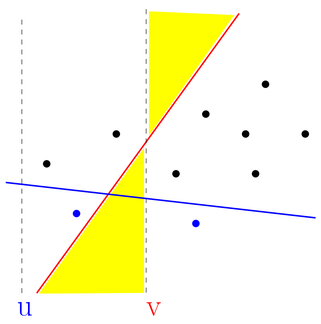
NONFIT POLOHA PRO $X_i \in \mathbb{R}$

Pro dané koeficienty $\theta = (\theta_0, \theta_1)^T$ označme residua regresního modelu jako $\theta = (\theta_0, \theta_1)^T$ jako $r_i(\theta) = Y_i - \theta_0 - \theta_1 X_i,\ i = 1, \ldots, n$. Pak $\theta = (\theta_0, \theta_1)$ nazveme v nonfit poloze pokud existuje číslo $v \in \mathbb{R}$, které je různé od všech X_i a takové, že buď

$$r_i(\theta) < 0 \quad \forall x_i < v \quad \text{a zároveň} \quad r_i(\theta) > 0 \quad \forall x_i > v$$

nebo

$$r_i(\theta) > 0 \quad \forall x_i < v \quad \text{a zároveň} \quad r_i(\theta) < 0 \quad \forall x_i > v.$$



Obr. 1: červená přímka je v nonfit poloze, modrá přímka má regresní hloubku 2.

JINÝMI SLOVY:
Přímka určená koeficienty (θ_0, θ_1) je v **nonfit poloze** pokud při její rotaci do vertikální polohy neprojdeme žádným bodem.

Na obrázku 1 je červeně zobrazena přímka v nonfit poloze. Nalevo od bodu v jsou jen kladná rezidua, napravo jen záporná. To je ekvivalentní s tím, že ve žlutě vyznačených oblastech, které musíme projít při její rotaci do horizontální polohy, neleží žádný bod. Nyní můžeme definovat regresní hloubku.

Modrá přímka na obrázku 1 má regresní hloubku 2. Je nutné odebrat 2 modře vyznačené body, aby zaujmula nonfit polohu. Poté napravo od bodu u budou jen kladná rezidua. Formálně lze definici hloubky zapsat

REGRESNÍ HLOUBKA
Regresní hloubka přímky dané koeficienty θ je definována jako nejmenší počet pozorování (Y_i, X_i), které je potřeba odebrat, aby θ byla v nonfit poloze.

$$D(\theta) = \min_{v \in \mathbb{R}} \min \left\{ \sum_{i:\ r_i(\theta)<0,\ X_i \le v} 1 + \sum_{i:\ r_i(\theta)>0,\ X_i > v} 1, \sum_{i:\ r_i(\theta)>0,\ X_i \le v} 1 + \sum_{i:\ r_i(\theta)<0,\ X_i > v} 1 \right\}$$

Je okamžitě vidět velmi nízká citlivost na odlehlá pozorování, jelikož bereme v potaz jen počty pozorování nad / pod regresní přímkou a nikoliv jejich vzdálenosti od přímky.

Pro vícerozměrné regresory $X_i \in \mathbb{R}^p$ je **definice regresní hloubky stejná**. Jen je nutné nějakým způsobem zobecnit definici nonfit polohy pro jednorozměrné regresory na regresory o více složkách.

NONFIT POLOHA PRO $X_i \in \mathbb{R}^p$

$\theta = (\theta_0, \theta_1^T)^T \in \mathbb{R}^{p+1}$ nazveme ležící v nonfit poloze, pokud existuje nadrovina V v prostoru hodnot náhodné veličiny X taková, že v ní neleží žádné X_i a taková, že $r_i(\theta) = Y_i - \theta_0 - \theta_1^T X_i > 0$ pro všechna X_i ležící v jednom poloprostoru prostoru hodnot X, který odděluje nadrovina V a $r_i(\theta) < 0$ pro všechna X_i ležící v poloprostoru druhém.

JINÝMI SLOVY:
Plocha určená koeficienty θ je v **nonfit poloze** pokud při její rotaci do polohy kolmé k prostoru hodnot X neprojdeme žádným bodem.

V teoretickém případě je regresní hloubka definována jako nejmenší pravděpodobnostní masa, která musí být odebrána, aby se regresní plocha dostala do horizontální polohy. Např. pokud $\text{med}\, e_i = 0,\ i = 1, \ldots, n$ potom je nejhlubší regresní plocha rovná ploše určené koeficienty (β_0, β_1^T).

Nejhlubší regresní plocha nejlépe vystihuje data, naopak plochy s malou hloubkou data moc dobře nevystihují. Vlastnosti jako je konzistence, robustnost, ekvivariance apod. jsou dokázané v [3].

Regresní kvantily založené na hloubce:
Podobně jako při odhadu v kvantilové regresi můžeme pro výpočet kvantilů použít váhy τ a $1-\tau$. Počet kladných reziduí převážit váhou τ a počet záporných záporných reziduí váhou $1 - \tau$. Nejhlubší regresní plochou bude v tomto případě konzistentní (viz [1]) odhad podmíněného τ-kvantilu.
Formální definici dostaneme přepisem definice hloubky $D(\theta)$, místo váhy rovné 1 použijeme váhu τ, resp. $1 - \tau$:

REGRESNÍ KVANTILY
Pro regresní koeficienty θ a nadrovinu V v prostoru hodnot X označme $L(V)$ a $P(V)$ dva poloprostory, které tato nadrovina odděluje. Regresním τ-kvantilem nazveme takovou hodnotu koeficientu θ pro kterou je následující výraz maximální

$$D_\tau(\theta) = \min_V \min \left\{ \sum_{i:\ r_i(\theta)>0,\ X_i \in L(V)} \tau + \sum_{i:\ r_i(\theta)<0,\ X_i \in P(V)} (1-\tau), \sum_{i:\ r_i(\theta)>0,\ X_i \in P(V)} \tau + \sum_{i:\ r_i(\theta)<0,\ X_i \in L(V)} (1-\tau) \right\}$$

Očividně volbou $\tau = 1/2$ dostaneme nejhlubší regresní přímku. Pro úplnost ještě uveďme jak tato definice vypadá v případě výpočtu klasického jednorozměrného mediánu. τ-kvantilem bude hodnota $Y_{(k_0)}$, kde k_0 je řešením maximalizace

$$\max_{k=1,\ldots,n} (\min\{\tau(n-k),\ (1-\tau)k\}).$$

Tedy klasický jednorozměrný kvantil.

O ODHADU PODMÍNĚNÝCH KVANTILŮ

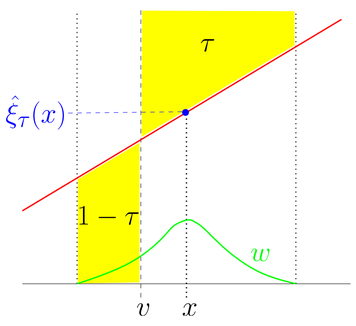
Nyní se oprostíme od předpokladu linearity modelu.
Učiníme pouze předpoklad: Funkce $\xi_\tau(x)$, která udává hodnotu podmíněného kvantilu veličiny Y při daném $X = x$ je spojitou funkcí.

Nejčastěji používané metody

Většina dnes rozšířených metod používá ideu jádrových odhadů, tedy pro odhad podmíněného kvantilu při dané hodnotě $X = x$ bereme převážně informaci jen z bodů v okolí x.

- První možností je použít jádrový odhad kvantilu či odhad kvantilu založený na k nejbližších sousedech . Tato jednoduchá a intuitivní metoda je popsána např. v [4]. Má mnoho dobrých vlastností (robustnost, ekvivariance vzhledem k monotonním transformacím odezvy, rychlost výpočtu, konzistence, ...). Je s podivem, že - zřejmě pro její přílišnou jednoduchost - není dnes téměř používána a dodnes není pořádně implementována ve statistických programech. Náš přístup lze chápat jako rozšíření této metody s ohledem na to, že pro odhad kvantilu v daném jádře místo konstanty použijeme polynom.
- Asi nejvíce používanou metodou je lokální polynomická kvantilová regrese, ta je implementována v programu R jako součást balíku *Quantile regression (quantreg)*. Autorem je R. Koenker. Idea je použít jádrový odhad na ztrátovou funkci ρ_τ z lineární kvantilové regrese. Obecně v každém okně odhadujeme kvantil jako polynom a za odhad ξ_τ vezmeme hodnotu tohoto polynomu v bodě x. My se budeme držet podobného postupu. Lokální polynomickou kvantilovou regresy lze opět chápat jako rozšíření jádrových odhadů kvantilů. Postup je popsán v [2].
- Další možností je využít nedávno vzniklého balíčku v R s názvem *kernlab*. O metodách používaných v tomto balíčku se můžeme dočíst v poměrně obsáhlém článku [5]. Postup staví na použití ztrátové funkce ρ_τ, kde se kvantilová funkce ξ_τ hledá mezi prvky Hilbertova prostoru s reprodukčním jádrem pro nějaké vhodně zvolené jádro. V kvalitě odhadu tato metoda zřejmě předčí lokální polynomickou kvantilovou regresi. V čem ji ale nepředčí je její i přes opačné tvrzení autorů výrazná časová a paměťová náročnost výpočtu a nutnost nastavení mnoha neintuitivních parametrů.

LOKÁLNÍ REGRESNÍ KVANTILY



Obr. 2: Jádrový (lokálně lineární) odhad regresních kvantilů

Myšlenka je totožná s jádrovými odhady kvantilů, jen místo konstanty budeme kvantil odhadovat polynomem.
Mějme váhovou funkci $w : \mathbb{R}^2 \to [0, +\infty)$. $w(X_i, x)$ přiřadí váhu bodu X_i vzhledem k poloze bodu x. Užitečnou volbou může být

$$w(X_i, x) = \begin{cases} 1, & \text{pokud } X_i \text{ patří mezi} \\ & k \text{ nejbližších sousedů } x \\ 0, & \text{jinak.} \end{cases}$$

Nebo Gaussovská váha

$$w(X_i, x) = e^{-\frac{1}{h}(X_i - x)^2}.$$

Pro odhad podmíněného kvantilu ξ_τ v bodě x použijeme modifikaci vzorce pro $D_\tau(\theta)$ pro výpočet kvantilu. Ke kladným, resp. záporným reziduím přidáme váhy v závislosti na vzdálenosti pozorování od bodu x.

JÁDROVÉ REGRESNÍ KVANTILY
Nechť $\hat\theta(\tau)$ je taková hodnota koeficientu $\theta \in \mathbb{R}^{p+1}$ pro kterou je následující výraz maximální

$$\min_V \min \left\{ \sum_{i:\ r_i(\theta)>0,\ X_i \in L(V)} \tau\ w(X_i, x) + \sum_{i:\ r_i(\theta)<0,\ X_i \in P(V)} (1-\tau)\ w(X_i, x), \right.$$
$$\left. \sum_{i:\ r_i(\theta)>0,\ X_i \in P(V)} \tau\ w(X_i, x) + \sum_{i:\ r_i(\theta)<0,\ X_i \in L(V)} (1-\tau)\ w(X_i, x). \right\}$$

Pak za odhad podmíněného kvantilu Y při daném $X = x$ vezmeme

$$\hat\xi_\tau(x) = \hat\theta_0(\tau) + \hat\theta^T(\tau)x.$$

V následujícím textu se zaměříme na odhad ξ_τ pro jednorozměrné regresory. Pro vícerozměrné proměnné se budeme muset spokojit s tvrzením, že následující postup lze rozšířit i pro regresory o více složkách.

Přímočarý výpočet má časovou náročnost $O(n^{p+2})$, existuje však přibližný algoritmus s náročností výrazně nižší.

Lokálně polynomické regresní hloubkové kvantily

Podmíněný kvantil budeme v jádře odhadovat polynomem stupně k.

Odhad kvantilu získáme aplikací vzorce pro výpočet jádrových regresních kvantilů pro odezvu Y_i a regresory $(X_i, X_i^2, \ldots, X_i^k,)$, $i = 1, \ldots, n$.

Je zřejmé, že volbou $k = 0$ získáme klasický jádrový odhad kvantilů. Pro většinu odhadů si bohatě vystačíme s volbou $k = 1$ či $k = 2$.

Odhad parametru podmíněných kvantilů není jednoznačný. Dvě plochy (resp. 2 různé vektory regresních koeficientů) mají stejnou hloubku (hodnotu výrazu pro výpočet kvantilů) pokud mezi nimi neleží žádný bod. Prakticky tedy při odhadu použijeme jen ty koeficienty θ které jsou určené všemi $(k+1)$-ticemi bodů $(X_i, X_i^2, \ldots, X_i^k,)$. To způsobí nespojitost $\hat\xi_\tau$ a tedy výslednou "kostrbatost" odhadu. Tomu lze zamezit přidáme–li si

podmínku na spojitost odhadů:
Mějme síť bodů $x_0, x_1, \ldots, x_l$ v kterých chceme získat odhad. Poté odhad v bodě x_i budeme hledat přes množinu všech koeficientů θ, které splňují

$$\hat\xi_\tau(x_{i-1}) = \theta_0 + \theta_1 x_{i-1} + \theta_2 x_{i-1}^2 + \ldots + \theta_k x_{i-1}^k.$$

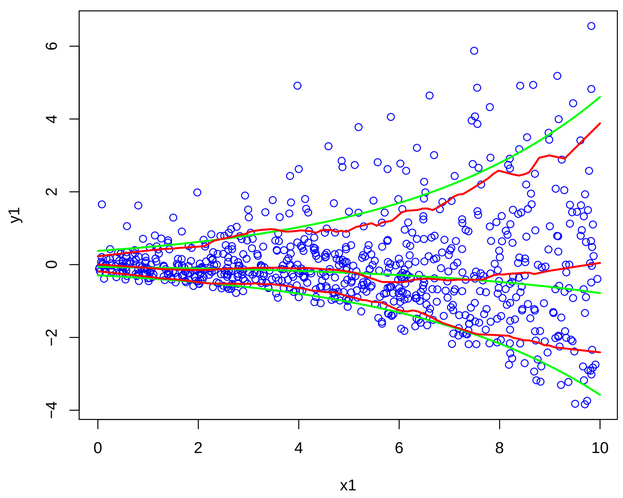
Nevýhodou je nutnost určení počátečního odhadu v bodě x_0, naopak výhodou je snížená výpočetní náročnost o jeden řád.

Vlastnosti: lokalizovaná verze podědí většinu vlastností verze nelokalizované, příkladem budiž:

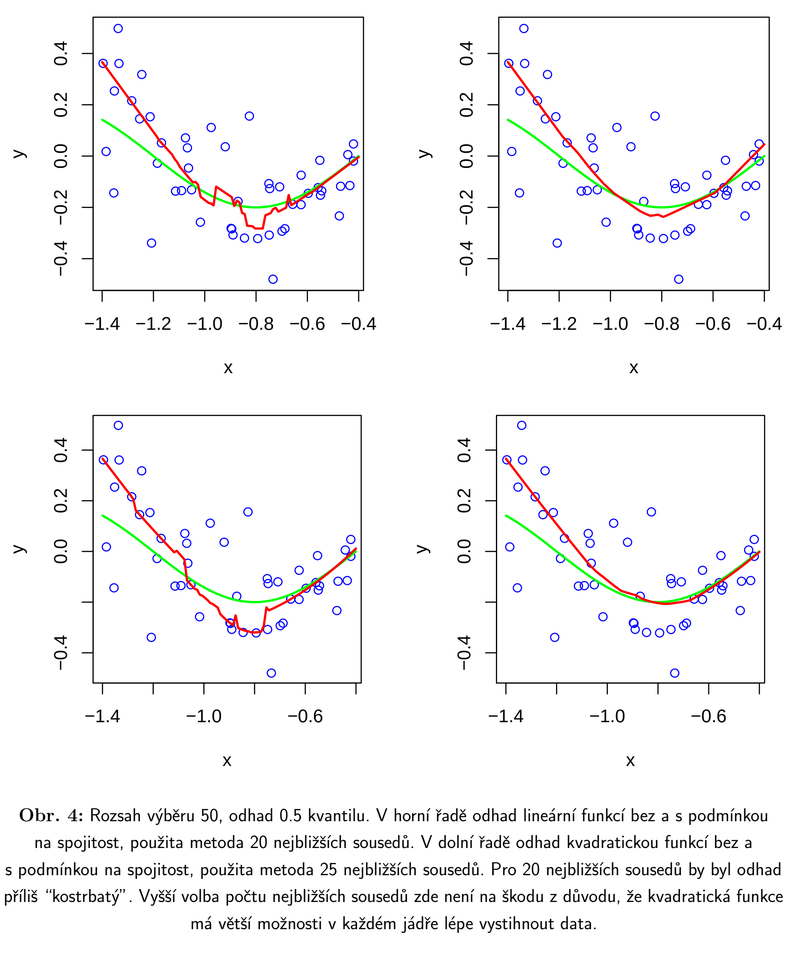
- Robustnost.
- Ekvivariance vzhledem ke změně měřítka odezvy.
- Ekvivariance vzhledem k afinním transformací při použití metody k nejbližších sousedů.
- ...

OBRÁZKY

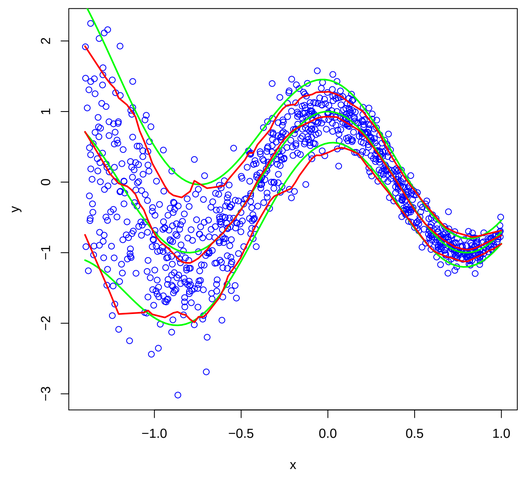
Na závěr pár obrázků na simulovaných datech. Na všech následujících obrázcích jsou zelenou barvou vyznačeny teoretické podmíněné kvantily a červeně jejich odhady.



Obr. 3: Odhad 0.1, 0.5, 0.9 kvantilů pro heteroskedastická data s asymetrickým rozdělením chyb. Rozsah výběru 500, použita metoda 100 nejbližších sousedů a polynom stupně 1 s podmínkou na spojitost.



Obr. 4: Rozsah výběru 50, odhad 0.5 kvantilu. V horní řadě odhad lineární funkcí bez a s podmínkou na spojitost, použita metoda 20 nejbližších sousedů. V dolní řadě odhad kvadratickou funkcí bez a s podmínkou na spojitost, použita metoda 25 nejbližších sousedů. Pro 20 nejbližších sousedů by byl odhad příliš "kostrbatý". Vyšší volba počtu nejbližších sousedů zde není na škodu z důvodu, že kvadratická funkce má větší možnosti v každém jádře lépe vystihnout data.



Obr. 5: Data, která v praxi zřejmě potkáte jen těžko. Rozsah výběru 1000. Odhad 0.05, 0.5, 0.95 kvantilů metodou 100 nejbližších sousedů, při odhadu opět použita podmínka na spojitost a lineární funkce.

ZÁVĚREM..

Pokud by se vytrvalý čtenář tohoto posteru zamýšlel nad tím, zda má smysl věnovat energii na aplikaci zde popsaných metod na nějaká data, může mu být nápomocno pár následujících řádků.

Kdy je to výhodné použít?

- Když máme podezření na odlehlá pozorování.
- Když se nám zdá jádrový odhad jako moc hrubý nástroj pro naše data díky tomu, že v okně / jádře používá pro odhad polynom stupně nula (konstantu).
- Když jsme zoufalí z honby za korektním a smysluplné výsledky vyplivujícím nastavením parametrů některých metod implementovaných ve statistických softwarech.
- Když v námi používaném softwaru není naprogramována žádná z metod na odhad podmíněných kvantilů a máme zrovna volné odpoledne.
- Když potřebujeme udělat poster na téma odhad podmíněných kvantilů...

Kdy je to nevýhodné použít?

- Když máme hodně dat a málo času.
- Když potřebujeme metodu s propracovaným teoretickým pozadím (asymptotika, testy, ...).
- Když v námi používaném statistickém programu máme rovnou k dispozici již implementovanou metodu pro odhad podmíněných kvantilů se kterou jsme spokojeni.
- ... Najde se toho jistě více.

Poděkování. Autor děkuje doc. Danielovi Hlubinkovi, jenž má nemalý podíl na formování myšlenek obsažených na tomto posteru. Poster je financován granty 1M0572 a MSM0021620839.

Literatura.

[1] He X., Portnoy S. (1998), *Asymptotics of the Deepest Line*, Statistical Inference and Related Topics: A Festschrift in Honor of A. K. Md. E. Saleh, New York: Nova Science.
[2] Koenker R. (2005), *Quantile Regression*, Cambridge University Press.
[3] Rousseeuw P.J., Hubert M. (1999), *Regression Depth*, Journal of the American Statistical Association, Vol. 94, No. 446.
[4] Stone, C. J. (1977), *Consistent Nonparametric Regression, with Discussion*, Annals of Statistics, 5, 595–645.
[5] Takeuchi I., Le Q.V., Sears T.D., Smola A.J. (2006), *Nonparametric Quantile Estimation*, Journal of Machine Learning Research 7 (2006) 1231–1264.