BAYESOVSKÝ ODHAD PARAMETRŮ MODELU METODAMI MCMC S APLIKACÍ NA MODELOVÁNÍ REGRESNÍCH KŘIVEK

Petr VOLF
ÚTIA AV ČR

Abstract: In the paper, the Markov chain Monte Carlo (MCMC) procedures are reviewed, namely the Gibbs and Metropolis–Hastings algorithms. It is shown how the procedures apply to Bayesian estimation of parameters of a probabilistic model. The MH algorithm is then used for adaptive construction of regression function from a chosen functional basis, e.g. from the B-splines.

Резюме: В этой статье описиваются методы Markov chain Monte Carlo (MCMC). Специально, изучается алгоритм Метрополиса-Астингса. Этот алгоритм используется для адаптивного построения регрессионной функции исползуя Б-сплайны.

Cílem statistické analýzy dat je poskytovat modely pozorovaných systémů. Mnohdy jsme v situaci, kdy je úloha specifikovat model převedena na úlohu odhadu (někdy velmi mnoha) parametrů. Bayesovský přístup nabízí jednu možnost, jak problém odhadu parametrů řešit. Jenže, v mnoha případech, zvláště těch složitějších, je konstrukce Bayesova aposteriorního rozdělení nezvládnutelným úkolem a nebývá to snadné ani s pomocí numerických metod. Pro tyto případy jsou nyní k dispozici postupy, které využívají počítačové simulace a namísto výpočtu aposteriorního rozdělení generují jeho reprezentaci. Přesněji, metody MCMC (Markov Chain Monte Carlo) generují Markovovu náhodnou posloupnost, jejíž rozdělení se blíží hledanému aposteriornímu.

V této práci nejprve zopakujeme Bayesovo pravidlo a jeho využití pro odhad parametrů modelu. Pak popíšeme některé metody umožňující generovat výběry, které se řídí (alespoň v limitě, jako je tomu pro MCMC metody) aposteriorním rozdělením. Jednu z metod, konkrétně algoritmus Metropolise–Hastingse, pak využijeme k modelování regresních křivek, a to k jejich konstruování z polynomiálních splinů (případně z jiné baze "elementárních" funkcí).

1 Bayesovský odhad parametrů

Představme si, že data $\boldsymbol{y}$, která měříme, jsou realizace náhodných veličin $\boldsymbol{Y} = Y_1, \ldots, Y_n$, model "vzniku" dat nechť je popsán pomocí pravděpodobnosti

(řekněme hustoty) $f(\boldsymbol{y};\boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \Theta \subset R^p$ je neznámý parametr. V Bayesovském 'světě' se neurčitost hodnoty $\boldsymbol{\theta}$ popisuje také pomocí pravděpodobnosti, na začátku se zvolí apriorní rozdělení – označme je (jeho hustotu) $p_0(\boldsymbol{\theta})$. Bayesovo pravidlo pak udává aposteriorní rozdělení jako podmíněné rozdělení hodnot parametru při napozorovaných datech, tj.

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = C(\boldsymbol{y}) \cdot f(\boldsymbol{y};\boldsymbol{\theta}) \cdot p_0(\boldsymbol{\theta}). \tag{1}$$

Zde $C(\boldsymbol{y})$ je normující člen, nezávisící na $\boldsymbol{\theta}$.

Tímto způsobem dostaneme tedy rozdělení pravděpodobnosti hodnot parametru, které je ovlivněno pozorovanými daty $\boldsymbol{y}$. Pokud chceme bodový odhad parametru $\boldsymbol{\theta}$, bere se nejčastěji modus aposteriorního rozdělení, $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$ (– všimněme si, že při rovnoměrném apriorním $p_0(\boldsymbol{\theta})$ je tento odhad totožný s maximálně věrohodným), nebo $\tilde{\boldsymbol{\theta}}(\boldsymbol{y}) = \mathsf{E}(\boldsymbol{\theta}|\boldsymbol{y})$, tj. střední hodnota z aposteriorního rozdělení.

Problém samozřejmě bývá s výpočtem členu $C(\boldsymbol{y})$, ten ale k zjištění $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ vlastně nepotřebujeme. Jenže často (v případech mnohorozměrného $\boldsymbol{\theta}$) není snadné získat ani maximum, ani další charakteristiky aposteriorního rozdělení (mohly by nás zajímat především rozptyl a kvantily, ke zjištění "šíře" aposteriorního rozdělení a tím i 'kredibility' (důvěryhodnosti? – Bayesovské varianty konfidence) odhadu parametru). A tady nastupují metody, které jsou schopny aposteriorní rozdělení nasimulovat.

Poznámka: V dalším budeme s daty zacházet jako s "konstantou" – tj. jsme v situaci, kdy data jsou k dispozici a naším cílem je odhadnutí aposteriorního rozdělení parametru.

Mimochodem, při růstu rozsahu dat $n \to \infty$ se uplatní teorie konzistence, což v Bayesovském případě znamená, že aposteriorní rozdělení se soustřeďuje do jednoho bodu, do 'skutečné' hodnoty parametru $\boldsymbol{\theta}$. Tímto jevem se zde zabývat nebudeme.

2 MCMC procedury

Je vyvinuto několik metod jak vygenerovat výběr z určitého rozdělení a vyhnout se přitom výpočtu přesného tvaru tohoto rozdělení či alespoň jak generovat Markovovu posloupnost, jejíž rozdělení se blíží k onomu cílovému. Do první skupiny patří např.

Zamítací metoda [1], [2], [5]: Chceme získat (jednoho) reprezentanta rozdělení $p(\boldsymbol{\theta}|\boldsymbol{y})$. Mějme k dispozici jinou hustotu rozdělení pravděpodobnosti $q(\boldsymbol{\theta})$ na Θ, nedegenerovanou, takovou, že $p(\boldsymbol{\theta}|\boldsymbol{y})/q(\boldsymbol{\theta}) \leq U$ pro každé $\boldsymbol{\theta} \in \Theta$ (a pro naše data $\boldsymbol{y}$).

Vygenerujme nyní $\boldsymbol{\theta}^*$ z rozdělení $q(\boldsymbol{\theta})$ a "přijměme" jej s pravděpodobností $p^* = p(\boldsymbol{\theta}^*|\boldsymbol{y})/(q(\boldsymbol{\theta}^*) \cdot U) \leq 1$, tj. udělejme "nula–jedničkový" náhodný pokus, který má za výsledek "1" s pravděpodobností p^*, "0" s $1 - p^*$. Je snadné ukázat (viz zmíněná literatura), že takto získaná (a přijatá) veličina $\boldsymbol{\theta}^*$ má právě rozdělení $p(\boldsymbol{\theta}|\boldsymbol{y})$. Použití této metody je omezeno tím, že musíme znát konstantu U, která by neměla být příliš velká, abychom na získání dostatečného počtu reprezentantů nepotřebovali příliš mnoho "kandidátů", kteří by se při velkém U většinou zamítali.

Hned se nabízí následující zjednodušení. Generujme kandidáty $\boldsymbol{\theta}^*$ z apriorního rozdělení $q_0(\boldsymbol{\theta})$. Máme-li rozumnou konstantu V takovou, že $f(\boldsymbol{y};\boldsymbol{\theta}) \leq V$ pro každé $\boldsymbol{\theta}$,pak z (1) plyne, že $p(\boldsymbol{\theta}|\boldsymbol{y})/q_0(\boldsymbol{\theta}) = C(\boldsymbol{y})f(\boldsymbol{y};\boldsymbol{\theta}) \leq C(\boldsymbol{y})V$, což odpovídá konstantě U shora. Přijímací pravděpodobnost je pak

$$p^* = \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{C(\boldsymbol{y})V\ q_0(\boldsymbol{y})} = \frac{f(\boldsymbol{y};\boldsymbol{\theta})}{V}.$$

Gibbsův algoritmus [2], [4], [7] je MCMC metoda, která je vhodná pro případ mnohorozměrného parametru. Označme $p_j(\theta_j|\boldsymbol{\theta}_{(-j)}, \boldsymbol{y})$ hustoty podmíněných aposteriorních rozdělení, kde
$\boldsymbol{\theta}_{(-j)} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$.

Algoritmus začne z nějaké (zvolené) počáteční hodnoty $\boldsymbol{\theta}^{(0)}$ a postupně v každém kroku inovuje 1 složku $\boldsymbol{\theta}$, a to tak, že ji generuje právě z podmíněného rozdělení p_j. Takže například novou j-tou složku v $m+1$. cyklu dostaneme tak, že $\theta_j^{(m+1)}$ vygenerujeme pomocí hustoty
$p_j(\theta_j|\theta_1^{(m+1)}, \ldots, \theta_{j-1}^{(m+1)}, \theta_{j+1}^{(m)}, \ldots, \theta_p^{(m)})$. Zjevně dostáváme náhodnou posloupnost $\boldsymbol{\theta}^{(m)}$, která je Markovova. Není problém ukázat, [2], [7], že hustota invariantního rozdělení takovéto Markovovy posloupnosti je právě $p(\boldsymbol{\theta}|\boldsymbol{y})$, a že rozdělení $\boldsymbol{\theta}^{(m)}$ při $m \to \infty$ konverguje k aposteriornímu rozdělení.

To je důsledek vět o konvergenci rozdělení nerozložitelných aperiodických Markovových řetězců k svému invariantnímu rozdělení (viz třeba kniha Feller, Probability Theory), zobecnění (neb zde již nejde o řetězce v pravém slova smyslu) je např. v [7]. Dalším důsledkem je také ergodická vlastnost, mimo jiné, že skoro jistě, pro $K \to \infty$,

$$\frac{1}{K} \sum_{m=1}^{K} \boldsymbol{\theta}^{(m)} \longrightarrow \int_{\Theta} \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta} = \mathsf{E}(\boldsymbol{\theta}|\boldsymbol{y}).$$

Samozřejmě, zdaleka ne vždy známe tvar podmíněných pravděpodobností p_j. Pak je možné Gibbsův algoritmus při generování každé nové složky kombinovat se zamítací metodou. A nebo použít proceduru Metropolise–Hastingse.

Algoritmus Metropolise–Hastingse [2], [6], [7] také vytváří Markovovu náhodnou posloupnost $\boldsymbol{\theta}^{(m)}$, a to následujícím způsobem. Nechť $\boldsymbol{\theta}^{(m)}$ je zatím poslední člen posloupnosti, nechť $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ je nějaká (zatím zcela libovolná) hustota rozdělení pravděpodobnosti (tj. může být podmíněná aktuální hodnotou parametru, ale nemusí). Vygenerujme s její pomocí "kandidáta" $\boldsymbol{\theta}^*$ na další člen posloupnosti. Položme pak

$$\pi(\theta^*, \theta^{(m)}) = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y}) \cdot q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(m)}|\boldsymbol{y}) \cdot q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m)})} \tag{2}$$

a přijměme $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^*$ s pravděpodobností $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(m)}) =$
$= \min\{1,\, \pi(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(m)})\}$. Pokud $\boldsymbol{\theta}^*$ nepřijmeme, pokládáme $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)}$.

Takto tedy vzniká Markovova posloupnost, jejíž (hustoty) přechodové pravděpodobnosti z $\boldsymbol{\theta}$ do $\boldsymbol{\theta}'$ jsou

$$h(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \begin{cases} q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \cdot \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) & \text{pro } \boldsymbol{\theta}' \neq \boldsymbol{\theta}, \\ 1 - \int_\Theta q(\boldsymbol{\theta}''|\boldsymbol{\theta})\, \alpha(\boldsymbol{\theta}''|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}'' & \text{pro } \boldsymbol{\theta}' = \boldsymbol{\theta}. \end{cases} \tag{3}$$

Funkce $q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \cdot \alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$ tedy tvoří "jádro" pro tyto pravděpodobnosti přechodů, a vlastnosti výsledné Markovovy posloupnosti do značné míry závisí na vlastnostech hustoty q. I tu můžeme chápat jako hustotu rozdělení pravděpodobnosti přechodů pro nějakou (jinou) Markovovu posloupnost na Θ (a to samozřejmě i v případě, kdy $q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = q(\boldsymbol{\theta}')$ – tj. q by generovala i.i.d. posloupnost).

Je dokázáno [6], že pokud q generuje nerozložitelnou a aperiodickou Markovovu posloupnost, tak pak i posloupnost $\boldsymbol{\theta}^{(m)}$ vzniklá algoritmem Metropolise–Hastingse je nerozložitelná a aperiodická. Z toho pak plynou ony příjemné vlastnosti: 1. Existence jediného invariantního rozdělení, 2. Konvergence rozdělení $\boldsymbol{\theta}^{(m)}$ k němu, 3. Ergodicita.

A je snadné ukázat ([2], [6]), že aposteriorní rozdělení $p(\boldsymbol{\theta}|\boldsymbol{y})$ je právě invariantním rozdělením posloupnosti $\boldsymbol{\theta}^{(m)}$. Nejprve se ukáže z (2) a (3), že pro libovolná $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ $h(\boldsymbol{\theta}'|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\boldsymbol{y}) = h(\boldsymbol{\theta}|\boldsymbol{\theta}') \cdot p(\boldsymbol{\theta}'|\boldsymbol{y})$ (pro $\boldsymbol{\theta} = \boldsymbol{\theta}'$ to platí), z toho pak vyvodíme $\int_\Theta p(\boldsymbol{\theta}|\boldsymbol{y}) \cdot h(\boldsymbol{\theta}'|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = p(\boldsymbol{\theta}'|\boldsymbol{y})$.

Z těchto limitních vlastností tedy plyne, že pokud vygenerujeme dostatečně dlouhou posloupnost (a uřízneme její dostatečně dlouhý začátek), tak výsledný vzorek hodnot můžeme (s rezervou, že nejde o i.i.d. výběr) považovat za reprezentaci aposteriorního rozdělení. Rozhodně tak můžeme zacházet s výběrovými kvantily a s průměrem $\frac{1}{K} \sum \boldsymbol{\theta}^{(m)}$ (součet přes m, od nějakého $K_0 + 1$ do $K_0 + K$) jako aproximací pro $\mathsf{E}(\boldsymbol{\theta}|\boldsymbol{y})$ (použijeme jej tedy jako bodový odhad parametru $\boldsymbol{\theta}$).

Nyní si všimněme, jak se rozhodovací pravidlo zjednoduší, když budeme kandidáty na nové členy posloupnosti generovat přímo z apriorního rozdělení

$q_0(\boldsymbol{\theta})$. Ve (2) pak dostaneme (po dosazení za $p(\boldsymbol{\theta}|\boldsymbol{y})$ z Bayesova vzorce (1))

$$\pi(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \frac{f(\boldsymbol{y};\boldsymbol{\theta}^*)\cdot q_0(\boldsymbol{\theta}^*)}{f(\boldsymbol{y};\boldsymbol{\theta})\cdot q_0(\boldsymbol{\theta})} \cdot \frac{q_0(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta}^*)} = \frac{f(\boldsymbol{y};\boldsymbol{\theta}^*)}{f(\boldsymbol{y};\boldsymbol{\theta})}, \tag{4}$$

neboli věrohodnostní poměr.

Připomeňme si ještě další metodu patřící do této skupiny, zvanou **simulované žíhání**. Je to metoda vhodná na znáhodněné hledání maxima funkce. Je používána v úlohách rekonstrukce obrazů (viz i Janžura, Robust 90).

3 Úloha odhadu regresní funkce

Nyní budeme aplikovat algoritmus Metropolise–Hastingse na úlohu odhadu regresní funkce. Mějme *neparametrický* regresní model

$$Y = r(x) + e, \tag{5}$$

kde $r(x)$ je neznámá (předpokládáme, že 'hladká') funkce, e je náhodná veličina s $\mathcal{N}(0,\sigma^2)$ rozdělením, σ^2 neznámé. Data jsou pak nezávislé realizace dvojice (X,Y) : $(\boldsymbol{x},\boldsymbol{y}) = (x_1, y_1, \ldots, x_m, y_m)$. O jedné možnosti odhadu regresní funkce, o jádrových odhadech, existuje i v "Robustech" několik článků, např. Antoch (1986), Michálek (1994). Zde se budeme snažit odhad funkce $r(x)$ zkonstruovat jako lineární kombinaci z nějaké báze funkcí, tj. ve tvaru

$$r^*(x) = \boldsymbol{\alpha}'\,\boldsymbol{B}(x;\boldsymbol{\beta}) = \sum_{j=1}^{M} \alpha_j\, B_j(x;\boldsymbol{\beta}). \tag{6}$$

Funkce B_j jsou tedy vybrané "jednotky" ze zvolené báze funkcí, $\boldsymbol{\beta}$ jsou jejich vnitřní parametry. Zpravidla každá B_j závisí jen na jedné či několika málo komponentách z $\boldsymbol{\beta}$. Máme tedy před sebou úlohu optimální volby $\boldsymbol{\beta}$ (a také M – počtu použitých jednotek), zatímco $\boldsymbol{\alpha}$ se již dají (při ostatním známém) spočíst, zde přímo metodou nejmenších čtverců.

Jsme tedy v situaci, kterou jsme popsali na začátku, tj. máme co dělat s mnohorozměrným parametrem, který je velice obtížné odhadnout optimálně přímými metodami. Jsou k dispozici i adaptivní nebayesovské postupy, např. MARS [3] používá polynomiální spliny z "+" baze a tvoří spojitý model 'stromovitě', podobným způsobem, jako jsou vytvářeny (nespojité) klasifikační a regresní stromy (CART – viz Antoch, Robust'88). My budeme raději používat lépe lokalizovaných jednotek (než jsou "+" spliny), konkrétně kubické B-spliny, případně tzv. radiální bázové funkce. Parametr β_j tedy znamená "umístění" j-té funkce, tj. uzel regresního splinu, či střed RB funkce. Mimochodem, problém "učení" neuronových sítí je vlastně tentýž (optimalizace nastavení parametrů funkcionálních "jednotek" – a také počtu jednotek), množí se již pokusy řešit tuto úlohu v rámci Bayesova přístupu.

3.1 Aplikace algoritmu Metropolise–Hastingse

Uvažujme tedy regresní model (5) s Gaussovým šumem $e \sim \mathcal{N}(0, \sigma^2)$, zatím s jednorozměrnou kovariátou X v ohraničeném intervalu $[a, b] \subset R^1$. Hledejme odhad regresní funkce ve tvaru (6). Věrohodnostní funkce (či hustota sdružené pravděpodobnosti pro $\boldsymbol{y}$ při daných $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, M, $\boldsymbol{x}$), platí-li model (5) s funkcí (6), je

$$f(\boldsymbol{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, M, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - r^*(x_i))^2}{2\sigma^2}\right).$$

Máme tedy co dělat s parametry $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, M, kde M může nabýt kladných celých hodnot, prostor hodnot $\boldsymbol{\beta}$ je $a < \beta_1 < \beta_2 < \cdots < \beta_M < b$ (tj. závisí na M; v případě B-splinů jde o vnitřní uzly v $[a, b]$). Parametry $\boldsymbol{\alpha}$ by také mohly být získávány nějakým znáhodněným způsobem, ale protože jde o parametry "lineární", můžeme pro každé M a $\boldsymbol{\beta}$ (a daná data) ihned metodou nejmenších čtverců spočíst odhady $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{M} \alpha_j B_j(x_i; \boldsymbol{\beta})\right)^2$. V rámci Metropolisova–Hastingsova algoritmu budeme tedy generovat (po jednotlivých složkách) Markovovu posloupnost hodnot parametrů $\boldsymbol{\theta} = (M, \boldsymbol{\beta})$. Zvolíme apriorní rozdělení pro M, $Q_0(M)$, podmíněná apriorní rozdělení $q_{0j}(\beta_j | M, \boldsymbol{\beta}_{(-j)})$ – tj. se závislostí na M, dále přechodové rozdělení pro generování nových kandidátů M^*, např. jako symetrickou náhodnou procházku, s pravděpodobnostmi $Q(M^*|M) = 1/3$ pro $M^* = M - 1, M$, nebo $M + 1$, jinak $Q(M^*|M) = 0$. Pro generování nových kandidátů β_j^* při daném $\boldsymbol{\beta}_{(-j)}$ a M použijeme přímo q_{0j}. K vytvoření přijímacího pravidla ještě potřebujeme věrohodnostní funkci při daných $\boldsymbol{x}$, $\boldsymbol{\beta}, M$. Vezmeme prostě $f^*(\boldsymbol{y}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, M, \boldsymbol{x})$, tj. vždy pro $\boldsymbol{\beta}$, M dosadíme za $\boldsymbol{\alpha}$ příslušný odhad $\hat{\boldsymbol{\alpha}}$.

Jak nyní vypadá jeden dílčí krok MH algoritmu? Představme si, že poslední stav posloupnosti je M, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ a chceme zkusit inovovat j-tou složku $\boldsymbol{\beta}$. Nejprve vygenerujeme M^* z $Q(M^*|M)$. Je-li $M^* = M - 1$, prostě zkusíme β_j vynechat. Nové $\boldsymbol{\beta}^*$ je tedy $\boldsymbol{\beta}_{(-j)} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_M)$. Je-li $M^* = M$, generujeme β_j^* z $q_{0j}(\beta | M, \boldsymbol{\beta}_{(-j)})$, dostaneme tedy $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_{j-1}, \beta_j^*, \beta_{j+1}, \ldots \beta_M)$. Konečně, je-li $M^* = M + 1$, pak generujeme dvě hodnoty z $q_{0j}(\beta | M, \boldsymbol{\beta}_{(-j)})$, menší z nich označíme β_j^*, větší β_j^{**}, a máme tak M^* rozměrný vektor $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_{j-1}, \beta_j^*, \beta_j^{**}, \beta_{j+1}, \ldots \beta_M)$. Dále k M^* a $\boldsymbol{\beta}^*$ spočteme příslušné $\hat{\boldsymbol{\alpha}}^*$. Nyní nastupuje znáhodněné přijímací pravidlo. Podle (2) dostaneme

$$\pi(M^*, \boldsymbol{\beta}^*, M, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\beta}^*, M^* | \boldsymbol{y}, \boldsymbol{x}) \cdot q(\boldsymbol{\beta}, M | \boldsymbol{\beta}^*, M^*)}{p(\boldsymbol{\beta}, M | \boldsymbol{y}, \boldsymbol{x}) \cdot q(\beta^*, M^* | \beta, M},$$

kde ale rozdělení pravděpodobnosti, podle kterého jsme získali nové kandidáty $\boldsymbol{\beta}^*$, M^*, je

$q(\boldsymbol{\beta}', M'|\boldsymbol{\beta}, M) = q_0(\beta'|M') \cdot Q(M'|M)$. Když nyní rozepíšeme i $p(\boldsymbol{\beta}, M|\boldsymbol{y}, \boldsymbol{x})$ podle Bayesova vzorce (1) jako $C(\boldsymbol{x}, \boldsymbol{y}) \cdot f(\boldsymbol{y}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, M, \boldsymbol{x}) \cdot q_0(\boldsymbol{\beta}|M) \cdot Q_0(M)$, dostaneme nakonec

$$\pi(M^*, \boldsymbol{\beta}^*, M, \beta) = \frac{f(\boldsymbol{y}; \hat{\boldsymbol{\alpha}}^*, \boldsymbol{\beta}^*, M^*, \boldsymbol{x}) \cdot Q_0(M^*)}{f(\boldsymbol{y}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, M, \boldsymbol{x}) \cdot Q_0(M)}. \tag{7}$$

Nyní tedy uděláme náhodný nula–jedničkový pokus a $(\boldsymbol{\beta}^*, M^*)$ s pravděpodobností $\min(1, \pi)$ přijmeme jako další člen Markovovy posloupnosti $\{\boldsymbol{\beta}^{(m)}, M^{(m)}, m = 0, 1, 2, \ldots\}$, v opačném případě je nový člen roven starému $(\boldsymbol{\beta}, M)$, tj. nedošlo k přechodu.

Vidíme, že rozhodovací pravidlo je znovu založeno na věrohodnostním poměru a také na poměru apriorních pravděpodobností pro "velikost" modelu M. To nám dává možnost vhodnou volbou $Q_0(M)$ usměrňovat volbu počtu použitých jednotek. Představme si například, že $Q_0(M)$ je klesající v M, pak pro $M^* = M + 1$ je tento poměr < 1 a vlastně tak "penalizujeme" prostý věrohodnostní poměr, abychom zabránili přílišnému růstu M během iterací. Připomíná to penalizující kriteria pro výběr řádu autoregresního modelu (používají se kriteria AIC, BIC, varianty krosvalidace a j.).

Zkoušeli jsme například volit $Q_0(M) \approx \exp(-M^2/n^\gamma)$, kde n byl rozsah dat a číslo $\gamma \in (0.5, 1)$. Dalším, spíše praktickým, opatřením bylo, že jsme rovnou zamítali takové nové uzly, které byly k některému starému blíže než na zvolenou vzdálenost. Podmíněná apriorní rozdělení $q_{0j}(\beta_j|\boldsymbol{\beta}_{-j}, M)$ jsme volili rovnoměrná, vždy pro β_j mezi v tu chvíli pevnými sousedními β_{j-1}, β_{j+1}. To pak odpovídá rovnoměrnému sdruženému rozdělení $q_0(\boldsymbol{\beta}|M)$ na $a < \beta_1 < \cdots < \beta_M < b$.

V případě Gaussova rozdělení odchylek e_i potřebujeme ve věrohodnostním poměru ještě odhad parametru σ. Ten jsme odhadovali z reziduí poslední předchozí iterace modelu.

Poznámka: Samozřejmě, pokud si předem stanovíme pevné M, které nechceme měnit, situace (a kriterium (7)) se zjednoduší.

4 Mnohorozměrná regrese

Jakmile je kovariáta mnohorozměrná, tj. $\mathbf{X} \in R^p$, naráží naše metoda na problém, jak v R^p náhodně volit jednotky. Za funkcionální jednotku v R^p zpravidla bereme součin jednorozměrných, např. $D(x, z) = B(x) \cdot C(z)$. Regresní funkci v R^2 bychom pak modelovali jako

$$(\alpha_{00}) + \sum_{j=1}^{M_1} \alpha_{j0} B_j(x, \boldsymbol{\beta}) + \sum_{k=1}^{M_2} \alpha_{0k} C_k(z, \boldsymbol{\gamma}) + \sum \sum \alpha_{jk} B_j(x, \boldsymbol{\beta}) \cdot C_k(z, \boldsymbol{\gamma}).$$

Takže "vnitřních" parametrů β_j, γ_k by bylo $M_1 + M_2$, zatímco parametrů α_{jk} by bylo $M_1 \cdot M_2 + M_1 + M_2 + (1)$. Změna jedné komponenty (uzlu) β_j by vyvolala změnu nejméně $M_2 + 1$ jednotek modelu. Snaha zmenšit dimenzi prostoru, ve kterém se v každém kroku rozhoduje, vedla k "stromovým" procedurám, viz CART, MARS [3]. Takovéto metody zkoušíme i v rámci MCMC přístupu, zatím nemáme uspokojivé řešení.

Additivní model. Jiná je situace, když se spokojíme s aditivním modelem vlivu kovariát, tj. hledáme regresní funkci ve tvaru $r(\mathbf{x}) = \sum_{j=1}^{p} r_j(x_j)$. Každou funkci r_j pak jednotlivě modelujeme z jednodimenzionálních funkcionálních jednotek. MCMC procedura probírá a iteruje opakovaně jednu funkci r_j za druhou.

Poznámka: Když už máme proceduru, která mění počet jednotek, podobným způsobem by mohla měnit i počet kovariát použitých v modelu, tj. měnit p. Vlastně by to znamenalo připustit i hodnoty $M_j = 0$.

5 Věrohodnostní modely

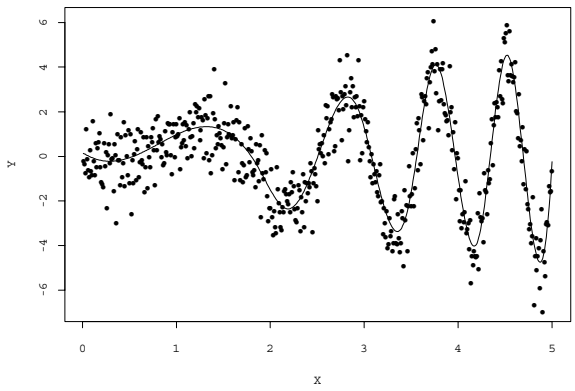
Věrohodnostním modelem myslíme takový, kde je neznámý parametr (v případě regrese tedy regresní funkce) parametrem ve věrohodnostní funkci, a je tedy zpravidla odhadován metodou maxima věrohodnosti. Samozřejmě, Gaussův model je speciálním případem. Další příklady jsou exponenciální modely (GAM – Hastie a Tibshirani), logistický regresní model, i Coxův model, kde je odhadování založeno na částečné věrohodnostní funkci (viz i Volf, Robust 92).

Naše pravidlo pro přijímání nových členů posloupnosti je založeno přímo na poměru věrohodností, můžeme je tedy použít pro jakýkoli věrohodnostní model. Změní se však způsob inovace parametrů $\boldsymbol{\alpha}$. Tenje teď třeba získat jako MVO, tj. maximalizací $\mathcal{L} = \ln f(\boldsymbol{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, M, \boldsymbol{x})$, pro daná data $\boldsymbol{y}, \boldsymbol{x}$ a dané hodnoty $M, \boldsymbol{\beta}$. Znamená to řešit rovnice $\partial\mathcal{L}/\partial\alpha_m = 0, m = 1, ..., M$.

Takové rovnice se většinou řeší iterací (algoritmem Newtona Raphsona třeba). V případě ortogonálních jednotek B_m s omezeným nosičem vystačíme v každém kroku s inovací jen několika α_j. Například, kdybychom použili aproximaci regresní funkce pomocí histogramu, změna jednoho uzlu histogramu má za následek změnu jen 2 sousedních 'políček' histogramu, a tedy jen 2 parametrů. Podrobněji viz i Linka a spol. [9].

Příklad: Zkoušeli jsme náš postup jak pro Coxův model, tak pro logistický klasifikační model, a samozřejmě pro model standardní, s Gaussovým šumem. Pro ilustraci uvedeme jen jednoduchý umělý příklad.

Vygenerovali jsme $x_i, i = 1, .., n$ pro $n = 500$, rovnoměrně v $(0, 5)$, k nim pak $y_i = x_i \sin(x_i^2) + e_i$, kde e_i byly i.i.d. z $\mathcal{N}(0, \sigma = 1)$. Jako bazi funkcí jsme použili kubické B–spliny. Začli jsme s třemi rovnoměrně umístěnými vnitřními uzly v $(0, 5)$. Apriorní rozdělení $\boldsymbol{\beta}$ bylo rovnoměrné při daném M, apriorní rozdělení M jsme zvolili tak, aby přidání jedné jednotky bylo v (7) penalizováno členem $\exp(-2M/n^{0.7})$, tj. apriorní rozdělení bylo úměrné $\exp(-M^2/n^{0.7})$. Generovali jsme Markovovu posloupnost $M^{(s)}, \boldsymbol{\beta}^{(s)}$ délky 200, z které jsme použili posledních 80 členů. Pro každý z nich jsme tedy obdrželi odhad regresní funkce $r_s^*(x)$ (ve tvaru (6)). Z těch jsme udělali průměr a to byl konečný odhad regresní funkce. Výsledek je na následujícím obrázku, kde body představují data a křivka je výsledný odhad.



6 Dodatek o polynomiálních splinech

Polynomiální spliny jsou jedním z nejpoužívanějších prostředků na aproximaci a modelování hladkých funkcí. Popíšeme zde kubické spliny, konstruované na intervalu $x \in [a, b] \subset R$.

Interval $[a, b]$ je rozdělen pomocí *uzlů* na podintervaly, na kterých je funkce modelována jako polynom třetího stupně, v uzlech pak přechází jeden polynom v druhý spojitě, i se spojitou 1. a 2. derivací. Nejznámější jsou dva způsoby konstrukce takovéto funkce (viz např. [8]).

První způsob používá funkcí, které jsou definovámy jako $(u)_+ = u$ pro $u \geq 0$, $= 0$ jinak. Mějme v (a, b) zvoleno M vnitřních uzlů $a < \beta_1 < ... <$

$\beta_M < b$. Pak kompletní spline je dán jako

$$r^*(x) \;=\; a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{j=1}^{M} \alpha_j (x - \beta_j)_+^3.$$

Tato funkce má požadované vlastnosti (spojitost derivací v uzlech) a vidíme, že obsahuje $M + 4$ 'lineární' parametry. Použité $'+'$ jednotky nejsou lokalizovány, tj. nemají omezený nosič.

Druhou možností je použít t.zv. B–spliny. Kromě M vnitřních uzlů definujme ještě formálně $\beta_{-j} = a - j(\beta_1 - a)$, $\beta_{M+1+j} = b + j(b - \beta_M)$ pro $j = 0, 1, 2, 3$. Máme nyní dohromady $M+8$ uzlů. Nyní pro $j = -1, 0, 1, ..., M, M+1, M+2$ definujme 'jednotky' – B-spliny:

$$B_j(x, \boldsymbol{\beta}) = 1[x \le \beta_{j+2}] \sum_{k=j-2}^{j+2} \{(x - \beta_k)_+^3 \;/ \prod_{s=j-2, s\neq k}^{j+2} (\beta_k - \beta_s)\}.$$

Výsledná funkce je pak $r^*(x) \;=\; \sum_{j=-1}^{M+2} \alpha_j B_j(x, \boldsymbol{\beta})$. Skládá se tedy z $M + 4$ jednotek a obsahuje opět $M + 4$ 'lineárních' parametrů α_j. Co je podstatné, funkce B_j jsou lokalizované, každá B_j je nenulová jen mezi uzly β_{j-2}, β_{j+2}. Změna jednoho uzlu β_k vede ke změně B_j pro $j = k-2, .., k+2$. Proto jsme si dovolili další aproximaci v inovaci modelu používajícím B–spliny, a to že k změněnému jednomu β_k jsme nepočítali celý nový vektor $\boldsymbol{\alpha}$, ale jen 5 komponent $\alpha_{k-2}, ..., \alpha_{k+2}$. Tato úspora je cenná zvlášť při modelování ve věrohodnostních modelech, kde α jsou počítána Newtonovou–Raphsonovou iterací.

References

[1] Antoch, J., Vorlíčková D. (1992). *Vybrané metody statistické analýzy dat*. Academia, Praha.

[2] Bernardo, J. M., Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

[3] Friedman, J. H. (1991). "Multivariate adaptive regression splines". *Annals Statist.* 19, 1–141.

[4] Geman, S., Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 724–741.

[5] Smith, A. F. M., Gelfand, A. E. (1990). "Sampling based approaches to calculating marginal densities". *J. Amer. Statist. Assoc.* 85, 398–409.

[6] Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* 57, 97–109.

[7] Roberts, G. O., Smith, A. F. M. (1994). "Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms". *Stoch. Processes and Applic.* 49, 207–216.

[8] Wold, S. (1974). "Spline functions in data analysis". *Technometrics* 16, 1–11.

[9] Linka, A., Picek, J., Volf, P. (1996). "Monte Carlo method for likelihood regression analysis". In: *Compstat'96*, Barcelona. Physica–Verlag, 343–348.

Petr Volf, ÚTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8,
e-mail: volf@utia.cas.cz