APROXIMACE APOSTERIORNÍCH DISTRIBUCÍ

Martin JANŽURA[1]
ÚTIA ČSAV

Abstract: A posterior distribution in the Bayesian estimation can be approximated by a density of the exponential type with a similar asymptotic behaviour.

Резюме: Апостериорное распределение в задаче Баесовского оценивания позволяет приближение при помощи распределения экспоненциального типа у которого аналогичные асимптотитеские свойства.

1 Úvod

V bayesovské úloze odhadu parametru je stěžejním problémem vyčíslení aposteriorního rozdělení parametru. V případě velkého množství dat a omezeného času i prostoru na jejich zpracování, což je typické např. v problematice řízení, je vhodné, jestliže aposteriorní rozdělení závisí na datech pouze prostřednictvím rekurzívně aktualizovatelné postačující statistiky. Pokud tomu tak není, např. není-li parametrická rodina exponenciálního typu, je vhodné pokusit se o projekci do rodiny exponenciálního typu při zachování základních asymptotických vlastností (Kulhavý (1996), kapitola 4).

Ukážeme, že tento postup lze bezprostředně zobecnit i pro závislá pozorování, a navíc odvodíme velmi obecný postup, jak exponenciální aproximace konstruovat.

2 Aposteriorní distribuce

Uvažujme parametrickou rodinu

$$\{q_\theta\}_{\theta\in\Theta}$$

pravděpodobnostních rozdělení na konečné množině X. Pro jednoduchost budeme také množinu parametrů Θ považovat za konečnou. (Tento předpoklad má však pouze teoretický význam, prakticky bude množina Θ obvykle veliká.) Předpokládáme dále, že je dáno apriorní rozdělení

$$p(\theta)$$

[1] S podporou grantu GA AV ČR č. A2075603.

na množině Θ.

Potom bude aposteriorní rozdělení, získané na základě náhodného výběru

$$x_1, \ldots, x_n$$

rovno

$$p(\theta|x_1, \ldots, x_n) = \frac{\prod_{i=1}^n q_\theta(x_i)\, p(\theta)}{\sum_{\tau \in \Theta} \prod_{i=1}^n q_\tau(x_i)\, p(\tau)}.$$

Podívejme se nejprve na asymptotické chování aposteriorní distribuce. Je-li splněna elementární podmínka identifikace, tj. $\theta^1 = \theta^2$ je ekvivalentní s $q_{\theta^1} = q_{\theta^2}$, a označíme-li q_θ^∞ příslušnou součinovou míru, snadno ukážeme, že

$$p(\theta|x_1, \ldots, x_n) \longrightarrow \delta(\theta = \theta^0) \quad \text{s. j. } [q_{\theta^0}^\infty]$$

pro každé θ, $\theta^0 \in \Theta$.

Platí totiž (podle silného zákona velkých čísel)

$$n^{-1} \sum_{i=1}^n \log \frac{q_\theta(x_i)}{q_{\theta^0}(x_i)} \longrightarrow -H(q_{\theta^0}|q_\theta) \quad \text{s. j. } [q_{\theta^0}^\infty]$$

kde $H(\cdot|\cdot)$ je I-divergence (Kullback–Leiblerova vzdálenost), daná výrazem

$$H(q_{\theta^0}|q_\theta) = \sum_{x \in X} \log \frac{q_{\theta^0}(x)}{q_\theta(x)}\, q_{\theta^0}(x) \geq 0$$

s rovností právě tehdy, když $\theta = \theta^0$.

Dostaneme tedy přibližný asymptotický výraz

$$p(\theta|x_1, \ldots, x_n) \doteq \frac{e^{-n\, H(q_{\theta^0}|q_\theta)}\, p(\theta)}{\sum_{\tau \in \Theta} e^{-n\, H(q_{\theta^0}|q_\tau)}\, p(\tau)},$$

odkud vidíme, že $p(\theta|x_1, \ldots, x_n)$ konverguje k $\delta(\theta = \theta^0)$ exponenciálně rychle.

Ke stejnému výsledku dospějeme, jestliže si uvědomíme (viz např. Kulhavý (1996), formule 2.24), že platí

$$p(\theta|x_1, \ldots, x_n) \doteq \frac{e^{-n\, H(\hat{q}^n|q_\theta)}\, p(\theta)}{\sum_{\tau \in \Theta} e^{-n\, H(\hat{q}^n|q_\tau)}\, p(\tau)},$$

kde $\hat{q}^n$ je empirická distribuce získaná z datového souboru $(x_1, \ldots, x_n)$, tj.

$$\hat{q}^n(y) = \frac{1}{n} \sum_{i=1}^n \delta(y = x_i)$$

pro každé $y \in X$.

Jelikož

$$H(\cdot|q_\theta)$$

je pro míry na konečné množině X spojitá funkce a opět podle silného zákona velkých čísel platí

$$\hat{g}^n \to g_{\theta^0} \quad \text{s.j. } [q_{\theta^0}^\infty],$$

máme tedy asymptoticky přibližně

$$H(\hat{q}^n|q_\theta) \doteq H(q_{\theta^0}|q_\theta)$$

a výše uvedený výsledek platí.

3 Exponenciální projekce

Aposteriorní distribuce, jak byla vyjádřena v předchozí části, závisí na datech prostřednictvím výrazů

$$H(\hat{q}^n|q_\theta)$$

pro všechna $\theta \in \Theta$. Z důvodů časových i paměťových je však žádoucí (viz opět např. Kulhavý (1996), kapitola 1), aby se tato závislost realizovala pouze prostřednictvím konečně rozměrné statistiky

$$\bar{h}^n = (\bar{h}_1^n, \ldots, \bar{h}_d^n)^\top,$$

kde

$$\bar{h}_j^n = \int h_j \, \mathrm{d}\hat{q}^n \quad \text{pro } j = 1, \ldots, d$$

a

$$h_1, \ldots, h_d : X \to R$$

jsou vhodně zvolené reálné funkce (1, $h_1, \ldots, h_d$ jsou lineárně nezávislé).

To je ovšem možné zejména tehdy, je-li

$$\{q_\theta^h\}_{\theta \in \Theta}$$

exponenciální rodina pravděpodobnostních rozdělení, tj. $\Theta = R^d$ a

$$q_\theta^h(x) = c(\theta) \exp\langle\theta,\, h(x)\} \, q_0(x),$$

kde $q_0(x)$ je "počáteční rozdělení" a

$$c(\theta) = \left[\sum_{x \in X} \exp\langle\theta,\, h(x)\rangle \, q_0(x)\right]^{-1}$$

je normalizační konstanta.

Potom obdržíme

$$p(\theta|x_1,\ldots,x_n) = p(\theta|\bar{h}^n) = \frac{c(\theta)^n\, e^{n\langle\theta,\bar{h}^n\rangle}\, p(\theta)}{\sum_{\tau\in\Theta} c(\tau)^n\, e^{n\langle\tau,\bar{h}^n\rangle}\, p(\tau)}$$

přičemž asymptotické výsledky předchozí části zůstávají pro tento speciální případ v platnosti.

Pokud však máme obecnou parametrickou rodinu $\{q_\theta\}_{\theta\in\Theta}$ s obecnou množinou parametrů Θ a přitom jsme schopni z dat efektivně extrahovat pouze statistiku

$$\bar{h}^n,$$

můžeme se pokusit aproximovat "skutečnou" aposteriorní distribuci

$$p(\theta|x_1,\ldots,x_n)$$

"exponenciální projekcí"

$$\tilde{p}(\theta|\bar{h}^n) = \frac{c(\theta)^n\, e^{n\langle\lambda(\theta),\bar{h}^n\rangle}\, p(\theta)}{\sum_{\tau\in\Theta} c(\tau)^n\, e^{n\langle\lambda(\tau),\bar{h}^n\rangle} p(\tau)},$$

kde $c(\theta) \in R$ a $\lambda(\theta) \in R^d$ jsou konstanty, které závisejí na původním parametru $\theta \in \Theta$. Problém zde ovšem spočívá ve vhodné volbě konstant $(c(\theta),\, \lambda(\theta)) \in R^{d+1}$.

Na rozdíl od Kulhavého (1996), který zvolil na základě geometrických představ přístup založený na projekci empirické distribuce $\hat{q}^n$ postupně pro každé $\theta \in \Theta$ vždy do exponenciální rodiny s počátečním rozdělením q_θ, my zde budeme naopak projektovat každou teoretickou distribuci

$$q_\theta$$

do exponenciální rodiny

$$\{q^h_\lambda\}_{\lambda\in R^d},$$

kde pro každé $\theta \in \Theta$ projekce $q^h_{\lambda(\theta)}$ minimalizuje vzdálenost danou I-divergencí, tj.

$$\lambda(\theta) \in \operatorname*{arg\,min}_{\lambda\in R^d} H(q_\theta|q^h_\lambda).$$

Potom $c(\theta) = c(\lambda(\theta))$ je příslušná normalizační konstanta exponenciálního rozdělení $q^h_{\lambda(\theta)}$. Je známo, že pokud $\lambda(\theta)$ existuje, potom vždy

$$\int h\, \mathrm{d}q^h_{\lambda(\theta)} = \int h\, \mathrm{d}q_\theta = h^\theta.$$

Potom též můžeme psát

$$\tilde{p}(\theta|\bar{h}^n) = \frac{e^{-n\,H(\hat{q}^n\,|\,q^h_{\lambda(\theta)})}\,p(\theta)}{\sum_{\tau\in\Theta} e^{-n\,H(\hat{q}^n\,|\,q^h_{\lambda(\theta)})}\,p(\tau)},$$

Jelikož

$$\bar{h}^n \to h^{\theta^0} = \int h\,\mathrm{d}q^h_{\lambda(\theta)} \quad \text{s.j. } [q^\infty_{\theta^0}],$$

máme asymptoticky přibližně

$$\tilde{p}(\theta|\bar{h}^n) \doteq \frac{e^{-n\,H(q^h_{\lambda(\theta^0)}\,|\,q^h_{\lambda(\theta)})}\,p(\theta)}{\sum_{\tau\in\Theta} e^{-n\,H(q^h_{\lambda(\theta^0)}\,|\,q^h_{\lambda(\theta)})}\,p(\tau)}.$$

Odtud obdržíme

$$\tilde{p}(\theta|\bar{h}^n) \longrightarrow \frac{\delta(\theta\in\mathcal{M}_{\theta^0})}{|\mathcal{M}_{\theta^0}|} \quad \text{s.j. } [q^\infty_{\theta^0}]$$

exponenciálně rychle, kde

$$\mathcal{M}_{\theta^0} = \{\theta\in\Theta; h^\theta = h^{\theta_0}\}.$$

Pokud $\mathcal{M}_{\theta^0} = \{\theta^0\}$, tedy h^θ stačí na identifikaci parametru $\theta\in\Theta$, obdržíme samozřejmě

$$\frac{\delta(\theta\in\mathcal{M}_{\theta^0})}{|\mathcal{M}_{\theta^0}|} = \delta(\theta=\theta^0).$$

Máme tedy v základním kvalitativním smyslu chování exponenciální projekce

$$\hat{p}(\theta|\bar{h}^n)$$

obdobné jako “skutečné” aposteriorní hustoty $p(\theta|x_1,\ldots,x_n)$.

Snaha o co “nejlepší” chování v kvantitativním smyslu vede k úloze “maximalizovat” rychlost $H(q^h_{\lambda(\theta^0)}|q^h_{\lambda(\theta)})$, přičemž jediný “volný parametr” je počáteční rozdělení exponenciální rodiny, tedy q_0. Důležité je odlišit zejména “blízké” hodnoty $\lambda(\theta)$, přičemž platí

$$H(q^h_{\lambda^0}|q^h_{\lambda^0+\Delta}) \doteq \frac{1}{2}\,\Delta^\top D_{q_0}(\lambda^0)\,\Delta,$$

kde $D_{q_0}(\lambda)$ je Hessián funkce

$$-\log c_{q_0}(\lambda) = \log\sum_{x\in X}\exp\langle\lambda,h(x)\rangle\,q_0(x).$$

Pro některé třídy lze nalézt takové q_0, které maximalizuje $D_{q_0}(\lambda(\theta))$ stejnoměrně pro všechna $\theta\in\Theta$ (Kulhavý (1996)).

4 Zobecnění pro náhodné procesy

Uvažujme nyní parametrickou rodinu

$$\{q_\theta\}_{\theta\in\Theta}$$

pravděpodobnostních rozdělení na X^Z, $Z = (\ldots,-1,0,1,\ldots)$, tj. náhodných procesů. Budeme předpokládat, že tyto procesy jsou stacionární a ergodické.

Jestliže je opět dáno nějaké apriorní rozdělení $p(\theta)$, dostaneme aposteriorní rozdělení

$$p(\theta|x_1,\ldots,x_n) = \frac{q_\theta^n(x_{1.},\ldots,x_n)\,p(\theta)}{\sum_{\tau\in\Theta} q_\tau^n(x_{1.},\ldots,x_n)\,p(\tau)},$$

kde q_θ^n je příslušné marginální rozdělení příslušného procesu.

Za určitých podmínek (splněných okamžitě např. pro Markovovy procesy libovolného řádu – viz např. Föllmer (1973)) platí

$$n^{-1}\log\frac{q_\theta^n(x_1,\ldots,x_n)}{q_{\theta_0}^n(x_1,\ldots,x_n)} \longrightarrow -\mathcal{H}(q_{\theta^0}|q_\theta) \quad \text{s. j. } [q_{\theta^0}]$$

kde

$$\mathcal{H}(q_{\theta^0}|q_\theta) = \lim_{n\to\infty} n^{-1}\int \log\frac{q_{\theta^0}^n}{q_\theta^n}\,\mathrm{d}q_{\theta^0}$$

je asymptotická I-divergence (relativní rychlost entropie).

Obdobně jako pro nezávislý případ máme

$$p(\theta|x_1,\ldots,x_n) \doteq \frac{e^{-n\,\mathcal{H}(q_{\theta^0}|q_\theta)}\,p(\theta)}{\sum_{\tau\in\Theta} e^{-n\,H(q_{\theta^0}|q_\tau)}\,p(\tau)},$$

a pokud $\mathcal{H}(q_{\theta^0}|q_\theta) = 0$ právě když $\theta^0 = \theta$, potom také

$$p(\theta|x_1,\ldots,x_n) \longrightarrow \delta(\theta = \theta^0) \quad \text{s. j. } [q_{\theta^0}].$$

Jestliže chceme i nyní opět použít myšlenku exponenciální projekce, musíme nejprve definovat zobecnění exponenciálního rozdělení pro náhodné procesy. Definujme nejprve zobecnění empirického rozdělení, tedy empirický proces $\hat{q}^n$ odvozený z dat $x_1,\ldots,x_n$. Označme $\hat{x}^n \in X^Z$ periodické prodloužení vektoru $(x_1,\ldots,x_n)$ (tj. $(\hat{x}^n)_s = x_{[s-1 \bmod n]} + 1$ pro $s > 0$), a $\hat{x}^{n,t}$ jeho posunutí o $t \in Z$ (tj. $(\hat{x}^{n,t})_s = (\hat{x}^n)_{s+t}$). Nyní můžeme pro každou omezenou funkci $f : X^Z \to R$ položit

$$\bar{f}^n = \int f\,\mathrm{d}\hat{q}^n = \frac{1}{n}\sum_{t=1}^n f(\hat{x}^{n,t}).$$

Všimněme si, že takto definovaný empirický proces $\hat{q}^n$ je stacionární.

Zvolme nyní

$$g,\, h_1, \ldots, h_d : X^Z \to R,$$

přičemž předpokládejme, že všechny tyto funkce závisejí pouze na souřadnicích $1, \ldots, m$ (tj. např. existuje $g^* : X^{[1,m]} \to R$ tak, že $g(x_Z) = g^*(x_1, \ldots, x_m)$ pro všechna $x_Z \in X^Z$, atd.).

Definujme nyní náhodný proces g_λ^h předpisem

$$\left\| \frac{1}{n} \log q_\lambda^{h,n} - \bar{g}^n - \langle \lambda, \bar{h}^n \rangle - \log c(\lambda) \right\|_\infty \longrightarrow 0 \quad \text{pro } n \to \infty.$$

Takto jsou obecně definována Gibbsova náhodná pole ve statistické fyzice (Preston (1976)). Za našich předpokladů je q_λ^h vždy dáno jednoznačně a navíc je to Markovův řetězec řádu m.

Jelikož platí přibližně

$$q_\lambda^{h,n} \doteq e^{n[\bar{g}^n + \langle \lambda, \bar{h}^n \rangle]} \cdot c(\lambda)^n,$$

kde

$$\log c(\lambda) = -\lim_{n\to\infty} \frac{1}{n} \log \int e^{n[\bar{g}^n + \langle \lambda, \bar{h} \rangle]} \, \mathrm{d}q_\lambda^h,$$

můžeme toto rozdělení vskutku považovat za exponenciální rozdělení. (Zde $e^{n\,\bar{g}^n}$ má funkci “počátečního rozdělení”.)

Nyní platí analogicky všechna tvrzení předchozí části. Definujme $\lambda(\theta)$ tak, že

$$\int h \, \mathrm{d}q_{\lambda(\theta)}^h = \int h \, \mathrm{d}q_\theta = h^\theta,$$

a položme

$$\tilde{p}(\theta|x_1, \ldots, x_n) = \tilde{p}(\theta|\bar{h}^n) = \frac{c(\lambda(\theta))^n \, e^{n\langle \lambda(\theta), \bar{h}^n \rangle} \, p(\theta)}{\sum_{\tau\in\Theta} c(\lambda(\tau))^n \, e^{n\langle \lambda(\tau), \bar{h}^n \rangle} \, p(\tau)}.$$

Jelikož jsou všechny q_θ ergodické, platí

$$\bar{h}^n \to h^{\theta^0} \quad \text{s.j. } [q_{\theta^0}],$$

a tudíž přibližně

$$\tilde{p}(\theta|\bar{h}^n) \doteq \frac{e^{-n\,\mathcal{H}(q_{\lambda(\theta^0)}^h | q_{\lambda(\theta)}^h)} \, p(\theta)}{\sum_{\tau\in\Theta} e^{-n\,H(q_{\lambda(\theta^0)}^h | q_{\lambda(\tau)}^h)} \, p(\tau)}$$

neboť

$$\mathcal{H}\left(q_{\lambda(\theta^0)}^h | q_{\lambda(\theta)}^h\right) = \log \frac{c(\lambda(\theta^0))}{c(\lambda(\theta))} + \left\langle \lambda(\theta^0) - \lambda(\theta),\, h^{\theta^0} \right\rangle.$$

Máme tedy výsledek naprosto analogický jako pro případ nezávislých pozorování, tj.

$$\tilde{p}(\theta|\bar{h}^n) \longrightarrow \delta(\theta = \theta^0)$$

pokud $\mathcal{H}\left(q^n_{\lambda(\theta^0)}|q^n_{\lambda(\theta)}\right) = 0$ právě když $\theta^0 = \theta$. Problém může nastat s vyčíslováním konstanty $c(\lambda(\theta))$ (viz výše). Tomu se pokusíme čelit v následující části.

5 Obecná exponenciální aproximace

Mějme nadále třídu ergodických náhodných procesů

$$\{q_\theta\}_{\theta\in\Theta}$$

a hledejme aproximaci aposteriorní distribuce v exponenciálním tvaru

$$\tilde{p}(\theta|\bar{h}^n) = \frac{e^{n[\langle\lambda(\theta),\bar{h}^n\rangle - b(\lambda(\theta))]}\, p(\theta)}{\sum_{\tau\in\Theta} e^{n[\langle\lambda(\tau),\bar{h}^n] - b(\lambda(\tau))}\, p(\tau)}.$$

Víme, že

$$\bar{h}^n \to h^{\theta^0} \quad \text{s. j. } [q_{\theta^0}],$$

předpokládáme

$$h^{\theta^1} = h^{\theta^2} \quad \text{právě když } \theta^1 = \theta^2,$$

a požadujeme, aby platilo

$$\tilde{p}(\theta|\bar{h}^n) \longrightarrow \delta(\theta = \theta^0) \quad \text{s. j. } [q_{\theta^0}],$$

přičemž tato konvergence by měla být exponenciální s co největší rychlostí.

Tento požadavek znamená

$$b(\lambda(\theta)) - b(\lambda(\theta^0)) - \langle\lambda(\theta) - \lambda(\theta^0),\, h^{\theta^0}\rangle > 0$$

pro $\theta \neq \theta^0$, což bude splněno zejména pokud

$$b : R^d \to R$$

bude striktně konvexní funkce a $\lambda(\theta)$ pro každé $\theta \in \Theta$ je určeno tak, aby vektor h^θ byl diferenciálem funkce b právě v bodě $\lambda(\theta)$, tedy

$$h^\theta = \nabla\, b(\lambda(\theta)).$$

Přímo z definice striktně konvexní funkce je totiž

$$b(\lambda) - b(\lambda^0) - \langle\lambda - \lambda^0,\, \nabla\, b(\lambda^0)\rangle > 0$$

pro každé $\lambda \neq \lambda^0$.

Jestliže je tedy $\nabla b(\lambda(\theta^0)) = h^{\theta^0}$, máme

$$b(\lambda) - b(\lambda(\theta^0)) - \langle \lambda - \lambda(\theta^0),\, h^{\theta^0} \rangle > 0$$

pro každé $\lambda \neq \lambda(\theta^0)$ a tím spíše pro $\lambda(\theta)$, $\theta \neq \theta^0$, pokud je ovšem funkce b hladká. Kdyby totiž hladká nebyla, mohlo by se stát $h^{\theta^1}, h^{\theta^2} \in \nabla b(\lambda^0)$ pro nějaké $\theta^1, \theta^2 \in \Theta$ a $x^0 \in R^d$. Potom bychom ztratili schopnost mezi θ^1 a θ^2 rozlišit. (Toto se může přihodit při postupu z předchozí části použitém na náhodná pole, kde mohou nastat fázové přechody (Preston (1976)). Pak θ^1 a θ^2 mohou reprezentovat dvě různé fáze, mezi kterými při exponenciální projekci nelze rozhodnout.)

Vzhledem ke zřejmé aproximaci

$$b(\lambda + \Delta) - b(\lambda) - \langle \Delta,\, \nabla b(\lambda) \rangle \doteq \frac{1}{2} \Delta^\top \nabla^2 b(\lambda)\, \Delta$$

pro "malá" Δ, je třeba požadovat, aby funkce f byla dokonce silně konvexní, což zajistí $\nabla^2 b(\lambda) > 0$ a rychlost konvergence zůstává exponenciální s kladným koeficientem.

Shrňme nyní navrhovaný postup:

I. Uvažujeme rodinu ergodických náhodných procesů $\{q_\theta\}_{\theta \in \Theta}$ splňujících $\theta^1 = \theta^2$ právě když $h^{\theta^1} = h^{\theta^2}$. Předpokládáme, že h^θ pro $\theta \in \Theta$ jsme schopni vyčíslit.

II. Zvolíme silně konvexní hladkou funkci

$$b : R^d \to R$$

tak, aby $\{h^\theta\}_{\theta \in \Theta} \subset \{\nabla b(\lambda)\}_{\lambda \in R^d}$.

III. Pro každé $\theta \in \Theta$ najdeme $\lambda(\theta) \in R^d$ tak, aby

$$h^\theta = \nabla b(\lambda(\theta)).$$

IV. Položíme

$$\tilde{p}(\theta | \bar{h}^n) = \frac{e^{n[\langle \lambda(\theta), \bar{h}^n \rangle - b(\lambda(\theta))]}\, p(\theta)}{\sum_{\tau \in \Theta} e^{n[\langle \lambda(\tau), \bar{h}^n \rangle - b(\lambda(\tau))]}\, p(\tau)}.$$

Za výše uvedeních předpokladů jsme dokázali následující

Tvrzení: Platí

$$\tilde{p}(\theta | \bar{h}^n) \longrightarrow \delta(\theta = \theta^0) \quad \text{s. j. } [q_{\theta^0}]$$

pro každé θ^0, $\theta \in \Theta$, přičemž tato konvergence je exponenciální.

Literatura

Föllmer, H. (1973): On entropy and information gain in random fields. Z. Wahrs. verw. Geb. 26, 207–217.

Kulhavý, R. (1996): Recursive Nonlinear Estimation. A Geometric Approach. Springer Verlag.

Preston, C. (1976): Random Fields. Springer Verlag, Lecture Notes in Math. 534.