LMD mokslo darbai, 359–365
1998, Vilnius

О скорости сходимости эмпирических байесовских оценок

В. Казакявичюс (ВУ)

В статье рассматривается задача эмпирического байесовского оценивания вещественного параметра при квадратической функции потерь. Доказано, что в широком классе параметрических моделей не существует эмпирических байесовских оценок, сходящихся к байесовской оценке со скоростью $O(n^{-1})$.

Сначала напомним необходимые понятия из теории эмпирического байесовского оценивания.

Пусть $(x_1, \theta_1), \ldots, (x_n, \theta_n)$, (x, θ) суть независимые одинаково распределенные случайные векторы со значениями в $X \times \Theta$, где X есть произвольное измеримое пространство, а $\Theta \subset \mathbb{R}$. На основании значений $x_1, \ldots, x_n$ и x нужно оценить ненаблюдаемое значение θ.

Пусть, далее, $(P_\theta | \theta \in \Theta)$ есть измеримое семейство распределений на X, описывающее условное распределение x относительно θ, а Π – некоторое множество априорных распределений на Θ. Каждое $\pi \in \Pi$ индуцирует интегральный оператор M_π в пространстве измеримых функций на $(X \times \Theta)^{n+1}$, действующий по формуле

$$\begin{aligned} &M_\pi f(x_1, \theta_1, \ldots, x_n, \theta_n, x, \theta) \\ &\quad = \int d\pi(\theta_1) dP_{\theta_1}(x_1) \ldots d\pi(\theta) dP_\theta(x) f(x_1, \theta_1, \ldots, x_n, \theta_n, x, \theta). \end{aligned}$$

Далее будем считать, что семейство (P_θ) доминировано некоторой σ-конечной мерой μ и через p_θ обозначим плотность P_θ относительно μ.

Хорошо известно, что среднеквадратический риск эмпирической оценки $t_n(x_1, \ldots, x_n, x)$ разлагается в сумму

$$M_\pi (t_n(x_1, \ldots, x_n, x) - \theta)^2 = M_\pi (t_n(x_1, \ldots, x_n, x) - \hat\theta_\pi(x))^2 + M_\pi (\hat\theta_\pi(x) - \theta)^2, \quad (1)$$

где

$$\hat\theta_\pi(x) = \frac{\int \theta p_\theta(x) d\pi(\theta)}{\int p_\theta(x) d\pi(\theta)}$$

есть байесовская оценка параметра θ, соответствующая априорному распределению π.

Поскольку π считается неизвестным, оценка $\hat\theta_\pi$ не может быть вычислена. Для оценивания неизвестного априорного распределения (или, точнее, функции $\hat\theta_\pi$) и предназначен так называемый архив $x_1, \ldots, x_n$. Из (1) следует, что эмпирическая оценка тем лучше, чем меньше она отклоняется от байесовской. Поэтому качество эмпирической оценки $t_n(x_1, \dots, x_n, x)$ естественно измерять величиной

$$\rho(t_n) = \sup_{\pi \in \Pi} M_\pi (t_n(x_1, \dots, x_n, x) - \hat{\theta}_\pi(x))^2.$$

Для любого $\theta \in \Theta$ обозначим

$$B(\theta) = \left\{ \theta' \in \Theta \middle| \int_{\{p_\theta \neq 0\}} p_{\theta'} d\mu = 1, \int_{\{p_\theta \neq 0\}} p^2_{\theta'}/p_\theta d\mu < \infty \right\}.$$

Основной наш результат содержится в следующей теореме.

Теорема. *Если $\theta_1, \dots, \theta_k \in B(\theta_0)$, $P_{\theta_i} \neq P_{\theta_j}$ при $i \neq j$ и Π содержит все априорные рапределения, сосредоточенные в $\{\theta_0, \dots, \theta_k\}$, то для любой эмпирической оценки t_n справедливо неравенство*

$$\liminf_{n \to \infty} n\rho(t_n) \geqslant \sum_{i=1}^{k} (\theta_i - \theta_0)^2.$$

Следствие. *Если $P_\theta \neq P_{\theta'}$ при $\theta \neq \theta'$, Π содержит все априорные рапределения, сосредоточенные в конечных подмножествах Θ, и для некоторого $\theta_0 \in \Theta$ множество $B(\theta_0)$ несчетно, то не существует эмпирических оценок, сходящихся к байесовской со скоростью $O(n^{-1})$.*

В качестве примера рассмотрим задачу оценивания среднего нормального распределения при известной дисперсии σ_0^2. Здесь $X = \Theta = \mathbb{R}$, μ – мера Лебега и

$$p_\theta(x) = \exp\{-(x - \theta)^2/(2\sigma_0^2)\}.$$

Поскольку

$$\int \frac{p^2_\theta(x)}{p_{\theta_0}(x)} dx = \int \exp\left\{ -\frac{x^2}{2\sigma_0^2} + \frac{2\theta - \theta_0}{\sigma_0^2} x - \frac{2\theta^2 - \theta_0^2}{2\sigma_0^2} \right\} dx < \infty,$$

то $B(\theta_0) = \Theta = \mathbb{R}$ и из сформулированного выше следствия вытекает, что в "непараметрическом случае" (т.е. когда множество априорных распределений достаточно велико) $n\rho(t_n) \to \infty$ для любой эмпирической оценки t_n.

Этот пример типичен для экспоненциальных семейств и поэтому наш результат подтверждает одно из предположений Сингха [5].

Теперь докажем основную нашу теорему.

Пусть

$$U = \left\{ u \in \mathbb{R}^k | u_1, \dots, u_k > 0, \sum_{i=1}^{k} u_i < 1 \right\}$$

и π_u (для $u \in \overline{U}$) есть такое распределение на $\{\theta_0, \dots, \theta_k\}$, что $\pi_u(\{\theta_i\}) = u_i$ для $i = 1, \dots, k$ (и, следовательно, $\pi_u(\{\theta_0\}) = 1 - \sum_{i=1}^k u_i$). Через $p(u, x)$ и $\hat{\theta}(u, x)$ обозначим безусловную плотность x и байесовскую оценку θ, соответствующие априорному распределению π_u, т.е.

$$p(u, x) = \Big(1 - \sum_{i=1}^k u_i\Big) p_{\theta_0}(x) + \sum_{i=1}^k u_i p_{\theta_i}(x)$$

и

$$\hat{\theta}(u, x) = \frac{\Big(1 - \sum_{i=1}^k u_i\Big)\theta_0 p_{\theta_0}(x) + \sum_{i=1}^k u_i \theta_i p_{\theta_i}(x)}{p(u, x)}.$$

Из неравенств

$$p(u, x) \geqslant u_i p_{\theta_i}(x) \quad (1 \leqslant i \leqslant k) \quad \text{и} \quad p(u, x) \geqslant (1 - \sum_{i=1}^k u_i) p_{\theta_0}(x)$$

следует, что для любого компакта $K \subset U$ существует такая константа $c(K) < \infty$, что

$$p_{\theta_i}(x) \leqslant c(K) p(u, x)$$

для всех $0 \leqslant i \leqslant k$ и $u \in K$. Если v есть другая точка из K, то

$$p(v, x) \leqslant (1 - \sum_{i=1}^k v_i) c(K) p(u, x) + \sum_{i=1}^k v_i c(K) p(u, x) = c(K) p(u, x).$$

Имея в виду это замечание, докажем непрерывную дифференцируемость $\sqrt{p(u, x)}$, как отображения из U в $L^2(\mu)$. Пусть $u^* \in U$ и V – такая выпуклая окрестность точки u^*, что $\overline{V} \subset U$. Положим

$$\psi_i(u, x) = \frac{\partial \sqrt{p(u, x)}}{\partial u_i} = \frac{p_{\theta_i}(x) - p_{\theta_0}(x)}{2\sqrt{p(u, x)}}$$

и $\psi(u, x) = (\psi_1(u, x), \dots, \psi_k(u, x))$. Очевидно, что $|\psi_i(u, x)| \leqslant c(\overline{V})^{3/2} \sqrt{p(u^*, x)}$ для любого $u \in V$. В силу теоремы о среднем значении для $u \in V$ имеем

$$\begin{aligned} |\psi_i(u, x) - \psi_i(u^*, x)| &= \frac{|\psi_i(u, x)|}{\sqrt{p(u^*, x)}} |\sqrt{p(u^*, x)} - \sqrt{p(u, x)}| \\ &\leqslant \frac{|\psi_i(u, x)|}{\sqrt{p(u^*, x)}} \sup_{u' \in V} \|\psi(u', x)\| \, \|u - u^*\| \leqslant \\ &\leqslant \sqrt{k} c(\overline{V})^3 \sqrt{p(u^*, x)} \|u - u^*\|. \end{aligned}$$

Из полученного неравенства следует непрерывность ψ, как функции из U в $L^2(\mu)$. С другой стороны, всилу известного следствия из теоремы о среднем значении [4, гл.3, параграф 5, теорема 13, следствие 1]

$$\begin{aligned} &|\sqrt{p(u,x)} - \sqrt{p(u^*,x)} - \psi^\top(u^*,x)(u-u^*)| \\ &\quad \leqslant \sup_{0<t<1} \|\psi((1-t)u + tu^*, x) - \psi(u^*,x)\| \, \|u-u^*\| \\ &\quad \leqslant kc(\overline{V})^3 \sqrt{p(u^*,x)} \|u-u^*\|^2. \end{aligned}$$

Следовательно, ψ является производной $\sqrt{p(u,x)}$ как отображения из U в $L^2(\mu)$.

Подобным образом доказывается непрерывная дифференцируемость $\hat\theta(u,x)$ как отображения из U в $L^2(p(u^*,\cdot)\cdot\mu)$. Производную этой функции обозначим через $\varphi(u,x) = (\varphi_1(u,x), \dots, \varphi_k(u,x))$; итак

$$\varphi_i(u,x) = \frac{\partial\hat\theta(u,x)}{\partial u_i} = \frac{\theta_i p_{\theta_i}(x) - \theta_0 p_{\theta_0}(x)}{p(u,x)} - \hat\theta(u,x)\frac{p_{\theta_i}(x) - p_{\theta_0}(x)}{p(u,x)}.$$

Через $J(u)$ обозначим информационную матрицу Фишера семейства плотностей $p(u,\cdot)$:

$$J(u) = 4\int \psi(u,x)\psi^\top(u,x)d\mu(x).$$

Если для некоторого $u \in U$ матрица $J(u)$ вырождена, то найдется такой ненулевой вектор $a \in \mathbb{R}^k$, что

$$0 = a^\top J(u)a = 4\int a^\top\psi(u,x)\psi^\top(u,x)a\,d\mu(x) = 4\int |a^\top\psi(u,x)|^2 d\mu(x).$$

Поэтому $a^\top\psi(u,x) = 0$ и, следовательно,

$$0 = \sum_{i=1}^n a_i(p_{\theta_i}(x) - p_{\theta_0}(x)) = -\sum_{i=1}^k a_i p_{\theta_0}(x) + \sum_{i=1}^k a_i p_{\theta_i}(x)$$

почти всюду по мере μ.

Положим для краткости $a_0 = -\sum_{i=1}^k a_i$. Тогда $\sum_{i=0}^k a_i P_{\theta_i} = 0$ и, следовательно,

$$\sum_{i=0}^k a_i^+ P_{\theta_i} = \sum_{i=0}^k a_i^- P_{\theta_i}, \tag{2}$$

где, как обычно, $a_i^+ = \max\{0, a_i\}$ и $a_i^- = \max\{0, -a_i\}$. Поскольку $\sum_{i=0}^k a_i = 0$, то $\sum_{i=0}^k a_i^+$ и $\sum_{i=0}^k a_i^-$ есть одно и то же число, а так как a определяется с точностью до постоянного множителя, это число можем считать равным 1.

Пусть π^+ и π^- суть вероятностные меры на $\{\theta_0, \dots, \theta_k\}$, точкам θ_i приписывающие массы a_i^+ и a_i^-. Равенство (2) тогда означает, что безусловные распределения x, соответствующие априорным распределениям π^+ и π^-, совпадают. В статье [3] доказано, что тогда не существует эмпирических оценок, сходящихся к байесовской. Поэтому утверждение доказываемой теоремы в этом случае очевидно.

Далее матрицы $J(u)$ будем считать невырожденными.

Фиксируем произвольные $u^* \in U, a, b > 0$ и положим

$$K_n(b) = \{u^* + n^{-1/2}J(u^*)^{-1/2}z \,|\, |z_1| \leqslant b, \dots, |z_k| \leqslant b\} \quad \text{и} \quad w_a(t) = \min\{t^2, a\}.$$

Всилу неравенства треугольника

$$\begin{aligned}\sqrt{n\rho(t_n)} &\geqslant \sup_{u\in K_n(b)} \left\{ nM_{\pi_u}\big(t_n(x_1,\dots,x_n,x) - \hat{\theta}(u,x)\big)^2 \right\}^{1/2} \\ &\geqslant \sup_{u\in K_n(b)} \left\{ nM_{\pi_u}\big(t_n(x_1,\dots,x_n,x) - \hat{\theta}(u^*,x) - \varphi^\top(u^*,x)(u-u^*)\big)^2 \right\}^{1/2} \\ &\quad - \sup_{u\in K_n(b)} \left\{ n\int \big(\hat{\theta}(u,x) - \hat{\theta}(u^*,x) - \varphi^\top(u^*,x)(u-u^*)\big)^2 p(u,x)d\mu(x) \right\}^{1/2}.\end{aligned} \tag{3}$$

Если n достаточно велико, то $p(u,x) \leqslant cp(u^*,x)$ для всех $u \in K_n(b)$ и из дифференцируемости $\hat{\theta}(u,x)$ следует, что второе слагаемое в правой части неравенства (3) есть $o(1)$. С другой стороны, квадрат первого слагаемого не меньше

$$\begin{aligned}&\sup_{u\in K_n(b)} M_{\pi_u} w_a\Big(\sqrt{n}\big[t_n(x_1,\dots,x_n,x) - \hat{\theta}(u^*,x) - \varphi^\top(u^*,x)(u-u^*)\big]\Big) \\ &\geqslant \sup_{u\in K_n(b)} M_{\pi_u} \int w_a\Big(\sqrt{n}\big[t_n(x_1,\dots,x_n,x) - \hat{\theta}(u^*,x) - \varphi^\top(u^*,x)(u-u^*)\big]\Big) \\ &\times p(u^*,x)d\mu(x) - a \sup_{u\in K_n(b)} \int |p(u,x) - p(u^*,x)|d\mu(x).\end{aligned}$$

Из неравенства

$$|p(u,x) - p(u^*,x)| \leqslant \sum_{i=1}^{k} |u_i - u_i^*|\,(p_{\theta_i}(x) + p_{\theta_0}(x))$$

следует, что и здесь второе слагаемое есть $o(1)$. С первым слагаемым произведя такие же преобразования как в доказательстве известной теоремы об асимптотической нижней границе для риска минимаксной оценки [2, гл. II, теорема 12.1] и переходя к пределу при $a, b \to \infty$, получим следующее неравенство:

$$\begin{aligned}\liminf_{n\to\infty} n\rho(t_n) &\geqslant (2\pi)^{-k/2}\int_{\mathbb{R}^k} dz \int \Big(\varphi^\top(u^*,x)J^{-1/2}(u^*)z\Big)^2 e^{-\|z\|^2/2} p(u^*,x)d\mu(x) \\ &= \int \varphi^\top(u^*,x)J^{-1}(u^*)\varphi(u^*,x)p(u^*,x)d\mu(x).\end{aligned} \tag{4}$$

Обозначим правую часть через $\sigma^2(u^*)$. Чтобы как то оценить эту величину, найдем ее предел при $u^* \to 0$.

Если $\sum_{i=1}^k u_i < \varepsilon$, то

$$|4\psi_i(u,x)\psi_j(u,x)| \leqslant \frac{(p_{\theta_i}(x)+p_{\theta_0}(x))(p_{\theta_j}(x)+p_{\theta_0}(x))}{(1-\varepsilon)p_{\theta_0}(x)} \in L^1(\mu).$$

Поэтому из теоремы Лебега о мажорированной сходимости следует, что

$$J_{ij}(u) = 4\int \psi_i(u,x)\psi_j(u,x)d\mu(x) \to 4\int \psi_i(0,x)\psi_j(0,x)d\mu(x) = J_{ij}(0)$$

при $u \to 0$. Но тогда $J(u) \to J(0)$ и поэтому $J^{-1}(u) \to J^{-1}(0)$ при $u \to 0$.

Подобным образом из неравенства

$$|\varphi_i(u,x)\varphi_j(u,x)p(u,x)| \leqslant \frac{(2\max\{|\theta_0|,\ldots,|\theta_k|\})^2}{1-\varepsilon}\frac{(p_{\theta_i}(x)+p_{\theta_0}(x))(p_{\theta_j}(x)+p_{\theta_0}(x))}{p_{\theta_0}(x)} \in L^1(\mu)$$

и той же теоремы Лебега вытекает, что

$$\sigma^2(u) = \sum_{i,j=1}^k J_{ij}^{-1}(u)\int \varphi_i(u,x)\varphi_j(u,x)p(u,x)d\mu(x) \to \sigma^2(0),$$

где

$$\sigma^2(0) = \sum_{i,j=1}^k J_{ij}^{-1}(0)\int \varphi_i(0,x)\varphi_j(0,x)p(0,x)d\mu(x) = \int \varphi^{\mathsf{T}}(0,x)J^{-1}(0)\varphi(0,x)dP_{\theta_0}(x).$$

Следовательно, после перехода в (4) к пределу при $u^* \to 0$ получаем неравенство

$$\liminf_{n\to\infty} n\rho(t_n) \geqslant \sigma^2(0).$$

Докажем, что $\sigma^2(0)$ есть квадрат нормы Гильберта–Щмидта некоторого линейного оператора. Для этого положим

$$\xi_i = p_{\theta_i}/p_{\theta_0} - 1$$

и в гильбертовом пространстве $L^2(P_{\theta_0})$ рассмотрим два конечномерных подпространства: E_0, порожденное множеством функций $\{\xi_1,\ldots,\xi_k\}$, и E, порожденное множеством E_0 и тождественной единичной функцией 1. Пусть T_0 есть линейный оператор из E_0 в E, переводящий ξ_i в $\varphi_i(0,\cdot)$. Такое определение корректно, так как семейство $(\xi_i | 1 \leqslant i \leqslant k)$ линейно независимо, а $\varphi_i(0,\cdot) = (\theta_i - \theta_0)p_{\theta_i}/p_{\theta_0} \in E$.

Квадрат нормы Гильберта–Шмидта оператора T_0 по определению равен

$$\|T_0\|_2^2 = \sum_{i=1}^{k} \|T_0\eta_i\|^2,$$

где $(\eta_i | 1 \leqslant i \leqslant k)$ есть произвольный ортонормированный базис в E_0. Пусть I означает единичную матрицу, $\xi = (\xi_1, \dots, \xi_k)$, $\eta = (\eta_1, \dots, \eta_k)$ и C – такая $(k \times k)$-матрица, что $\eta = C\xi$. Тогда

$$I = \int \eta\eta^\top dP_{\theta_0} = \int C\xi\xi^\top C^\top dP_{\theta_0} = CJ(0)C^\top,$$

т.е. $J^{-1}(0) = C^\top C$. С другой стороны,

$$\begin{aligned}\|T_0\|^2 &= \sum_{i=1}^{k} \int (T_0\eta_i)^2 dP_{\theta_0} = \int (T_0\eta)^\top (T_0\eta) dP_{\theta_0} = \int (C\varphi(0,\cdot))^\top (C\varphi(0,\cdot)) dP_{\theta_0} \\ &= \int \varphi(0,\cdot) C^\top C\varphi(0,\cdot) dP_{\theta_0} = \sigma^2(0).\end{aligned}$$

Продолжим T_0 до оператора T, определенного на всем E, полагая $T1 = 0$. Единичная функция ортогональна всем ξ_i, так как

$$\int \xi_i dP_{\theta_0} = \int (p_{\theta_i} - p_{\theta_0}) d\mu = 0.$$

Поэтому норма Гильберта–Шмидта продолженного оператора не изменится:

$$\|T\|_2^2 = \|T_0\|_2^2 = \sigma_0^2.$$

Но T переводит $p_{\theta_i}/p_{\theta_0}$ в $\varphi_i(0,\cdot) = (\theta_i - \theta_0)p_{\theta_i}/p_{\theta_0}$, т.е. числа $\theta_i - \theta_0$ являются собственными значениями оператора T. Поэтому утверждение теоремы вытекает из следующего факта функционального анализа : квадрат нормы Гильберта–Шмидта линейного оператора в конечномерном пространстве не больше суммы квадратов модулей его собственных значений [1, гл. 11, параграф 6, лемма 21 и следствие 3].

ЛИТЕРАТУРА

[1] Данфорд Н., Шварц Дж. Т., *Линейные операторы*, т. II, Мир, М., 1966.

[2] Ибрагимов И. А., Хасьминский Р. З., *Асимптотическая теория оценивания*, Наука, М., 1979.

[3] Казакявичюс В., Условия существования эмпирически байесовских оценок и оптимизация границ Большева, *Изв. ВУЗов. Матем.*, **11** (1983), 67–73.

[4] Шварц Л., *Анализ, т.* I, Мир, М., 1979.

[5] Singh R. S., Empirical Bayes estimation in Lebesgue-exponential families with rates near the best possible rate, *Ann. Statist.*, **7**(4) (1979), 890–902.