M-оценки и ОМП

422 группа

Tue 28 Feb 2017

Содержание

1 M- и Z-оценки

По выборке $\mathbf{x} = (x_1, \ldots, x_n)$, оценим параметр $\theta \in \Theta$. Введем функцию-критерий

$$m_\theta : \mathfrak{X} \to \bar{\mathbb{R}},$$

показывающую насколько наблюдения соответствуют параметру и отображение

$$\theta \mapsto M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i),$$

сопоставляющее каждому параметру его «подходящесть».

Определение (M-оценка[1])**.** По принципу «наилучшей хорошести», оценка есть

$$\hat{\theta}_n = \operatorname*{argmax}_\theta M_n(\theta).$$

Замечание. Чтобы $\hat{\theta}_n$ было случайной величиной, требуем, чтобы $\hat{\theta}_n$ было измеримым, а для этого, чтобы Θ — польским (полным, измеримым, метрическим).

Замечание. Если взять $m = \log p_\theta$, то M-оценка есть ОМП.

Определение (Z-оценка[2])**.** Рассмотрим другие критерии $\psi_\theta(x)$ и отображение

$$\theta \mapsto \Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(x_i).$$

В качестве оценки найдем

$$\hat{\theta}_n = \operatorname*{root}_\theta \Psi_n(\theta).$$

[1] «Maximum»

[2] «Zero»

Замечание. M-оценка может быть сведена к этой оценке $\psi_\theta = \partial m_\theta / \partial\theta$.

Замечание. ψ_θ должна быть, конечно, дифференцируема.

Пример. Пусть $x_1, \dots, x_n$ i.i.d. на $\mathrm{U}[0, \theta)$, то $p_\theta = \theta^{-1}\mathbf{1}_{[0,\theta)}$ и

$$\theta \mapsto \sum_{i=1}^{n} (\log \mathbf{1}_{[0,\theta)}(x_i) - \log(\theta)).$$

Максимум z-оценки достигается на

$$\hat{\theta}_n = \max x_i.$$

ОМП оценку не найти, потому что не продифференцировать индикатор.

2 Состоятельность

Пусть есть

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(x_i) \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E} m_\theta = M(\theta) = \int m_\theta \, \mathrm{d}\mathcal{P}_x.$$

Навесим argmax на обе стороны:

$$M_n(\theta) \rightsquigarrow \hat{\theta}_n = \operatorname*{argmax}_\theta M_n(\theta).$$

Если бы argmax было непрерывным, то $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta^* = \operatorname{argmax} M(\theta)$. Нужно проверить следующие вещи:

1. Сходится ли $\hat{\theta}_n$ хоть к чему-нибудь.
2. Правда ли, что $\theta^* = \theta_0$.

argmax в принципе не является непрерывным. До тех пор, пока находимся в области притяжения параметра, всё хорошо, но как только из нее выходим, можем резко перескочить на другой экстремум.

2.1 Сходимость к чему-нибудь

Теорема (Вальд)**.** *Рассматриваем M-оценку. Пусть*

1. *m_θ полунепрерывна сверху по θ для почти всех x:*

   $$\limsup_{\theta\to\theta_0} \le m_{\theta_0}, \quad \text{для почти всех } x.$$

   (если экстремум в θ_0, то он хорошо выражен).

2. argmax *должно быть случайной величиной. Для этого отображение*

   $$\forall B_\delta \subset \Theta \ x \mapsto \sup_{\theta\in B_\delta} m_\theta(x)$$

   должно быть

   a) *измеримо*

   b) *ограничено:*

   $$\int \sup_{B_\delta} m_\theta(x) \, \mathrm{d}\mathsf{P} < \infty.$$

Замечание. Пусть $\Theta^* = \{\theta^* \in \Theta : M(\theta^*) = \sup_\theta M(\theta)\}$. Еще ослабим условие: максимум будет достигаться асимптотически. Пусть интересует

$$\hat{\theta}_n = \operatorname{argmax}(M_n(\theta) - o(1))$$

(не доходим до максимума — с точки зрения приложений ок, потому что ищем оценки численно). Т.е. интересуемся

$$M_n(\hat{\theta}_n) \geq c \sup_\theta M_n(\theta), \quad c \xrightarrow[n\to\infty]{} 1,\ 0 \leq c < 1.$$

Тогда для любой такой последовательности оценок, $\forall \epsilon > 0\ \forall K \subset \Theta$ (K — компакт), верно, что

$$\mathsf{P}\left(\operatorname{dist}\left(\hat{\theta}_n, \theta^*\right) \geq \epsilon, \hat{\theta}_n \in K\right) \xrightarrow[n\to\infty]{} 0.$$

Иными словами, это условие на то, как оценка не сходится — либо последовательность сходится, либо выходит за компакт.

Следствие. *Поэтому для $\mathbb{R}^n$ всё либо сходится по вероятности, либо уходит на бесконечности* — не может колебаться *на манер* $(-1)^n$. *Это очень здорово.*

2.2 Равенство $\theta^* = \theta_0$

Говорили, что ОМП получаются, когда

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i).$$

Совершим трюк: добавим константу, так что

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log \frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}.$$

На максимум не влияет, конечно. Тогда

$$\frac{1}{n}\sum_{i=1}^n \ln \frac{p_{\boldsymbol\theta}}{p_{\boldsymbol\theta_0}}(x_i) \xrightarrow{\mathsf{P}} \int_{\mathbb{R}} \ln\left(\frac{p_{\boldsymbol\theta}(x)}{p_{\boldsymbol\theta_0}(x)}\right) p_{\boldsymbol\theta_0}(x)\,\mathrm{d}x = \mathsf{E} \ln \frac{p_{\boldsymbol\theta}}{p_{\boldsymbol\theta_0}} \leq \ln \mathsf{E} \frac{p_{\boldsymbol\theta}}{p_{\boldsymbol\theta_0}} = \ln \int_{\mathbb{R}} \frac{p_{\boldsymbol\theta}}{p_{\boldsymbol\theta_0}}(x) p_{\boldsymbol\theta_0}(x)\,\mathrm{d}x = \ln 1 = 0$$

так что

$$M(\theta) \leq 0.$$

Интересуемся $\theta : M(\theta) = 0$. Это так, когда $p_\theta = p_{\theta_0}$ для почти всех x. В предположении свойства *идентифицируемости* задачи ($\theta_1 \neq \theta_2 \implies \mathcal{P}_{\theta_1} \neq \mathcal{P}_{\theta_2}$), получаем $\theta = \theta_0$.

Таким образом, нужно иметь адекватную модель (выполнялось бы условие идентифицируемости). Для выполнимость теоремы Вальда достаточно компактности и $\sqrt{\log p}\ldots$ Компактность на практике выполняется — можно взять достаточно большой отрезок всегда.

Так что состоятельность выполняется.

3 Скорость сходимости

При выполнении некоторых условий регулярности (простые почти никогда не выполняются, сложные сложно проверить) ОМП асимптотически нормальны, т.е.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I^{-1}(\theta_0)).$$

Если условия регулярности не выполняются, может быть все что угодно. Обычно, если $\Theta = \mathbb{R}^n$, то оценки сходятся оч быстро. Иначе медленно.

4 Доверительные границы

4.1 Аппроксимация Вальда

Исходя из нормальной аппроксимации, доверительные границы можно

$$\left(\hat{\theta}_n - \theta_0\right)^{\mathrm{T}} \Sigma^{-1} \left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathrm{d}} \chi^2_d.$$

Проблемы:

- Границы лишь асимптотические.
- Нужно знать Σ, зависящую от оцениваемого параметра. Поэтому нужно еще и оценивать Σ^{-1}.

Так можно делать в низких размерностях и при больших выборках.

4.2 Метод профилей правдоподобия

Пусть есть $\boldsymbol{\theta}, \dim \boldsymbol{\theta} = k$. НУО, выделим $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \dim \boldsymbol{\theta}_1 = k_1$. Рассмотрим логарифм функции правдоподобия как функцию двух аргументов

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

Построим *профиль*

$$r(\boldsymbol{\theta}_2) = \ell(\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2),$$

где $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$ — MLE оценка $\boldsymbol{\theta}_1$ с данным $\boldsymbol{\theta}_2$: $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2) = \operatorname{argmax}_{\boldsymbol{\theta}_1} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Замечание. Если $\boldsymbol{\theta}_2 = (\hat{\boldsymbol{\theta}_2})_{\mathrm{MLE}}$, то $\hat{\boldsymbol{\theta}}_1(\hat{\boldsymbol{\theta}_2})_{\mathrm{MLE}} = (\hat{\boldsymbol{\theta}_1})_{\mathrm{MLE}}$.

Утверждение. Пусть истинный параметр $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20})$. Тогда

$$z^2 := -2\left(r(\boldsymbol{\theta}_{10}) - \ell(\hat{\boldsymbol{\theta}}_n)\right) \xrightarrow{\mathrm{d}} \chi^2(k_1).$$

Идея доказательства. Знаем, что $\hat{\boldsymbol{\theta}}_n$ имеет нормальное распределение, $\ell(\hat{\boldsymbol{\theta}}_n) \sim \chi^2$. Тогда в пределе ℓ должна выглядеть как сумма квадратов — можно разложить по Тейлору, будет что-то вроде

$$\dots + \frac{1}{2}\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n\right)^2 + \dots$$

Условия регулярности нужны чтобы убить все члены старше второго. □

Нужно уметь обращать r, чтобы

$$\boldsymbol{\theta}_{10} : \mathrm{qnt}_{\chi^2}(\gamma_1) \le -2\,(\dots) \le \mathrm{qnt}_{\chi^2}(\gamma_2).$$

Так что заменили задачу оценивания Σ^{-1} на задачу обращения детерминированной функции r. В качестве бонуса — профиль должен быть похож на квадратичную функцию, тогда доверительный интервал по квантилям будет правильно выбран (частный случай «локальной асимптотической нормальности»).

На глаз можно посмотреть график $|z|$. Когда всё хорошо, он будет выглядеть как модуль.

Пример. Пусть $\xi \sim \mathrm{N}(\mu_0, \sigma_0^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$. Пусть μ_0 известна; оценим по выборке σ^2. Т.к. $\chi^2(1) = (\mathrm{N}(0,1))^2$, то

$$z^2 = -2(\underbrace{r(\mu_0)}_{\max_{\sigma^2} \ell(\mu_0, \sigma^2)} - \ell(\mu_0, \widehat{\sigma^2})) \xrightarrow{\mathrm{d}} (\mathrm{N}(0,1))^2\,.$$

Значит $z \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$. Ясно, что в $\widehat{\sigma^2} = \widehat{\sigma^2}_{\mathrm{MLE}}$, $|z| = 0$. На остальных значениях это модуль-галочка, причем по значениям $|z|$ — квантилям $\mathrm{N}(0,1)$ — можно построить доверительный интервал для $\widehat{\sigma^2}$.

Замечание. В R — пакет `bbmle`, функция `mle2`. На выходе — объект типа `mle`. Это оценки + ... `summary(mle)`. Построить профиль можно при помощи функции `p <- profile(m)`. Вызывает `optim` и т.п. Профили можно рисовать: `plot(p)`. Подсчет доверительных интервалов: `confint(p)`.