Параметрическое и непараметрическое оценивания

Центр биоэлектрических интерфейсов, 13 ноября 2018 г.

Денис Деркач, Влад Белавин

Оглавление

Параметрическое оценивание

Предыдущая лекция

- Оценка метода моментов (Method of moments, MOM):
$$\widehat{\alpha}_n = \alpha_n(\widehat{\theta}).$$
- Оценка максимально правдоподобия (ОМП, Maximum Likelihood Estimate, MLE):
$$\widehat{\theta} : \mathcal{L}_n(\theta) \to \max .$$

Оценка апостериорного максимума

Оценка апостериорного максимума, MAP

Формально, ОМП определяет значения параметров, при которых наши данные наиболее вероятны:

$$f(X;\theta) \sim f(X|\theta).$$

На самом деле, мы обычно задаёмся вопросом, какие значения параметров наиболее вероятны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)},$$

где f, g и h — соответствующие функции распределения.

Оценка MAP

Определение

Оценка апостериорного максимума (MAP) определяется как такое значение $\widehat{\theta}_n$ параметра θ, которое максимизирует $f(\theta|X)$.

Связь с MLE

MAP и MLE очевидно связаны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)} = \frac{\prod_{i=1}^{n} f(X_i;\theta)g(\theta)}{h(X)} \sim \mathrm{const} \prod_{i=1}^{n} f(X_i;\theta)g(\theta)$$

Логарифмируем:

$$\log f(\theta|X) = \log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Получается, что значение MAP оценки и значение оценки MLE совпадают с точностью до априорной оценки $\log(g(\theta))$.

Сопряжённые априорные оценки

Какую $g(\theta)$ выбрать?

- любую;
- но лучше выбирать сопряжённое априорное распределение, для которого функциональная форма совпадает с апостериорным.

Значения параметров сопряжённых распределений имеют смысл предыдущих измерений.

Список из Википедии.

Пример

См. задачу 0, про монетку.

Комментарий о MAP

- позволяет учесть предыдущие знания;
- выдаёт точечную оценку (не совсем байесовский);
- зависит от параметризации;
- при относительно больших n совпадает с MLE (а также в случае $g(X) = \mathrm{const}$!).

О переходе к байесовскому представлению

Для перехода к Байесовскому методу нам надо оценить знаменатель выражения:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)}$$

Так как $h(X)$ — распределение данных при любых значениях параметров, можем записать:

$$h(X) = \int\limits_{\Theta} f(X|\theta)g(\theta)d\theta.$$

И оценивать интервалы.

Информация Фишера

Мотивирующий пример

Предположим, что у нас есть монетка с вероятностью выпадения орла θ. Мы бросаем монетку 10 раз, получившаяся выборка, рассматривая случайную величину $X = \{0; 1\}$,

$$x_{\text{obs}}^n = \{1, 0, 0, 1, 1, 1, 1, 0, 1, 1\}$$

Количество возможных комбинаций при этом будет 1024.
Введём другую случайную величину (функцию от выборки, статистику):

$$Y = \sum_{i=1}^{n} X_i$$

Здесь количество возможных вариантов будет всего 11.

Мотивирующий пример

Интересный факт, для такого примера.

$$\mathbb{P}(X^n|Y = y, \theta) = 1/\binom{n}{y}\Bigg|_{n=10}$$

То есть, условная вероятность не зависит от параметра θ.

Фактически, это означает, что нам достаточно изучать статитстику $Y = \sum_{i=1}^{n} X_i$, если мы хотим знать θ.

Достаточные статистики

Определение

Статистика $T_n = T_n(X_1, \dots, X_n)$ назвается достаточной для параметра θ, если условное распределение выборки $X^n = (X_1, \dots, X_n)$ при условии того, что $T_n = a$, не зависит от параметра θ для всех $a \in \mathbb{R}$.

NB: Достаточные статистики существуют для ограниченного числа распределений. Список из Википедии.

NB2: Достаточно рассматривать только несмещённые оценки, которые являются функциями от достаточной статистики (при условии, что такая существует для данной задачи).

NB3: (Несмещенная) эффективная оценка параметра всегда является достаточной статистикой.

Информация Фишера

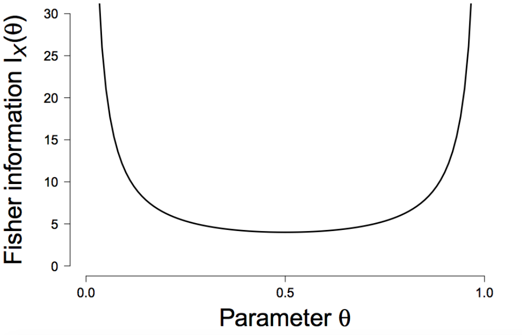
Определение

$$I(\theta) = \mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2.$$

NB: $\frac{\partial \log f(X;\theta)}{\partial \theta}$ показывает чувствительность модели к данному параметру.

NB2: $I_{X^n(\theta)} \geq I_{T(\theta)}$, причём $I_{X^n(\theta)} = I_{T(\theta)}$ тогда и только тогда, когда $T(\theta)$ - достаточная статистика.

Пример



Для распределения Бернулли (броски монеток) мы можем честно подсчитать значение информации $I(\theta) = \frac{1}{\theta(1-\theta)}$. Когда $\theta = 0$ и $\theta = 1$, $I(\theta) \to \infty$

Свойства информации Фишера

Теорема

Имеет место равенство : $I_n(\theta) = nI(\theta)$, при этом

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right) = -\int\left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right) f(x;\theta)dx.$$

В примере с 10 бросками монетки,

$$I_{X^n} = nI_X = n\frac{1}{\theta(1-\theta)} = I_Y,$$

что подтверждает, что Y — достаточная статистика.

Использование: планирование эксперимента

Для ОМП $\widehat{\theta}$ можно показать, что по распределению, при $n \to \infty$:

$$(\widehat{\theta} - \theta) \to \mathcal{N}(0, I_X^{-1}(\theta))$$

Предположим, что мы хотим узнать количество бросков монет, которые нам надо совершить, чтобы 68% оценок лежало в пределах 0.1 от истинного значения θ, то есть $1/\sqrt{nI_X(\theta)} = 0.1$. Заметим, что $I_X(\theta)$ достигает минимума в точке $\theta = 0.5$. Таким образом: $1/\sqrt{nI_X(\theta)} \le 1/\sqrt{nI_X(\theta = 0.5)} = 1/2\sqrt{n} = 0.1$, или $n = 25$ То есть, для того, чтобы обеспечить 68% оценок в пределах 0.1, необходимо совершить минимум 25 экспериментов.

Неравенство Крамера-Рао

Определение

В случае оценки $\widehat{\theta}$

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{[1 + \frac{\partial b(\theta))}{\partial \theta}]^2}{I(\theta)}.$$

где $b(\theta) = \mathbb{E}(\widehat{\theta}) - \theta$, смещение.

Пример

Задача 1.

Дельта-метод

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений $\text{Bernoulli}(p)$ и $\text{Bernoulli}(r)$?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim$ Bernoulli(p). Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений Bernoulli(p) и Bernoulli(r)?
- $\frac{p}{1-p} / \frac{r}{1-r}$

Пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений $\text{Bernoulli}(p)$ и $\text{Bernoulli}(r)$.
- $\frac{p}{1-p} / \frac{r}{1-r}$

Обычно мы используем $\hat{p} = \sum_i X_i / N$ для оценки p. Кажется, что для других величин мы можем использовать похожие оценки: $\frac{\hat{p}}{1-\hat{p}}$.

Как при этом оценить дисперсию?

Ряд Тейлора

Пусть $\mathbf{T} = (\mathbf{T_1}, .., \mathbf{T_k})$ – случайные величины со средними $\theta = (\theta_\mathbf{1}, ..., \theta_\mathbf{k})$. Пусть задана дифференцируемая функция $g(\mathbf{T})$ (оценка какого-то параметра). Найти дисперсию этой оценки.

Будем называть $g_i'(\theta) = \frac{\partial}{\partial \mathbf{t_i}} \mathbf{g}(\mathbf{t})\Big|_{\mathbf{t_1}=\theta_\mathbf{1};...;\mathbf{t_k}=\theta_\mathbf{k}}$.

Разложим $g(t)$ в ряд Тейлора:

$$g(t) \approx g(\theta) + \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{g_i'}(\theta)(\mathbf{t_i} - \theta_\mathbf{i})$$

Из этого следует:

$$E_\theta g(T) = g(\theta).$$

Ряд Тейлора

Аналогично дисперсия:

$$\mathsf{Var}_\theta g(T) \approx E_\theta\left([g(\mathbf{T}) - \mathbf{g}(\theta)]^{\mathbf{2}}\right) \approx E_\theta\left(\left(\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i)\right)^2\right) =$$

$$= \sum_{i=1}^{k} [g_i'(\theta]^{\mathbf{2}} \mathsf{Var}_\theta \mathbf{T_i} + \mathbf{2} \sum_{\mathbf{i}>\mathbf{j}} \mathbf{g_i'}(\theta)\mathbf{g_j'}(\theta)\mathsf{Cov}_\theta(\mathbf{T_i}, \mathbf{T_j}).$$

Замечание: здесь мы не использовали почти никакой информации о функции $g(T)$.

Вернёмся к мотивирующему примеру, нас интересовала дисперсия оценки $\frac{\hat{p}}{1-\hat{p}}$. Здесь $g(p) = \frac{p}{1-p}$ Используя предыдущие выкладки несложно подсчитать, что:

$$\mathrm{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx [g'(p)]^2 \mathrm{Var}(\hat{p}) = \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}.$$

Пример

X - случайная величина, с ненулевым матожиданием μ.
Необходимо оценить матожидание и дисперсию для оценки:
$g(\mu) = 1/\mu$.
Используя:

$$E_\mu\left(g(X)\right) \approx g(\mu),$$

$$\mathsf{Var}_\mu(g(X)) \approx [g'(\mu)]^2 \mathsf{Var}_\mu(X).$$

Получаем:

$$E_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu},$$

$$\mathsf{Var}_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu^4}\mathsf{Var}_\mu(X).$$

(Продолжение следует)

Теорема (Теорема Слуцкого)

Если $X_n \to X$ по распределению и $Y_n \to a$ по вероятности, причём $a = \text{const}$, тогда:

- $Y_n X_n \to aX$ по распределению,
- $X_n + Y_n \to X + a$ по распределению.

Теорема (Дельта-метод)

Пусть Y_n — последовательность случайных величин для которых $\sqrt{n}[Y_n - \theta] \to \mathcal{N}(0, \sigma^2)$ по распределению. Тогда для дифференциируемой в θ функции $g(.)$ с ненулевой производной, $\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$ по распределению.

Доказательство: Ряд Тейлора для $g(Y_n)$ около $Y_n = \theta$:

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + o(Y_n - \theta)$$

Третье слагаемое стремится к 0 по вероятности. Тогда мы сможем применить теорему Слуцкого:

$$[g(Y_n) - g(\theta)] = g'(\theta)(Y_n - \theta),$$

мы получили согласно условиям необходимую сходимость.

NB: В случае нулевой первой производной и ненулевой второй, оценка сходится к $\sigma^2 \frac{g''(\theta)}{2} \chi_1^2$ по распределению.

Пример

Продолжим предыдущий пример. Пусть есть выборка со средним $\bar{X}$, тогда:

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathcal{N}\left(0, \left(\frac{1}{\mu}\right)^4 \mathsf{Var}_\mu X_1\right) \text{ по распределению}$$

Если мы не знаем матожидание и дисперсию, можно ввести их оценку:

$$\widehat{\mathsf{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2,$$

то есть

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \mathcal{N}(0, 1)$$

Пример

Второй раз применив теорему Слуцкого, получим:

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\sigma/\mu^2} \to \mathcal{N}(0, 1)$$

.

Вспомним центральную предельную теорему!

Другая формулировка

Теорема

Если $\tau = g(\theta)$, где g — дифференцируема и $g'(\theta) \neq 0$, тогда

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{se}(\widehat{\tau})} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\widehat{\tau}_n = g(\widehat{\theta}_n)$ и $\widehat{se}(\widehat{\tau}_n) = |g'(\widehat{\theta})|\widehat{se}(\widehat{\theta}_n)$.
Таким образом, если

$$C_n = (\widehat{\tau}_n - z_{\alpha/2}\widehat{se}(\widehat{\tau}_n), \widehat{\tau}_n + z_{\alpha/2}\widehat{se}(\widehat{\tau}_n)),$$

тогда $\mathbb{P}(\tau \in C_n) \to 1 - \alpha$ и $n \to \infty$.

Пример

Пусть $X_1, \ldots, X_n \sim Bernoulli(p)$, $\psi = g(p) = \log(p/(1-p))$.
Информация Фишера равна

$$I(p) = \frac{1}{p(1-p)}.$$

Оценка стандартной ошибки

$$\widehat{se} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

ОМП величины ψ

$$\widehat{\psi} = \log \frac{\widehat{p}_n}{1-\widehat{p}_n}.$$

(далее на следующем слайде)

Пример

(продолжение, начало на предыдущем слайде)

Т.к. $g'(p) = 1/(p(1-p))$, то в соответствии с дельта-методом

$$\widehat{se}(\widehat{\psi}_n) = |g'(\widehat{p}_n)|\widehat{se}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

Таким образом, границы приближенного 95% доверительного интервала равны

$$\widehat{\psi}_n \pm \frac{2}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

Пример

Пусть $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Допустим, что μ известно, а σ неизвестно. Необходимо оценить $\psi = \log \sigma$. Логарифм функции правдоподобия

$$\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2,$$

значит

$$\widehat{\sigma}_n = \sqrt{\frac{\sum_i (X_i - \mu)^2}{n}}$$

(далее на следующем слайде)

Пример

(продолжение, начало на предыдущем слайде)

Для подсчета стандартной ошибки необходимо знать информацию Фишера.

$$\log f(X;\sigma) = -\log\sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

$$\frac{\partial^2(\log f(X;\sigma))}{\partial\sigma^2} = \frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4}$$

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}$$

(далее на следующем слайде)

Пример

(продолжение, начало на предыдущем слайде)

$$\widehat{se} = \frac{\widehat{\sigma}_n}{\sqrt{2n}}.$$

Пусть $\psi = g(\sigma) = \log \sigma$, тогда $\widehat{\psi} = \log \widehat{\sigma}_n$. Так как $g' = 1/\sigma$, то

$$\widehat{se}(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}_n} \frac{\widehat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}.$$

Границы приближенного 95%-ого доверительного интервала равны

$$\widehat{\psi}_n \pm 2/\sqrt{2n}.$$

Резюме параметрического оценивания

Способы параметрического оценивания

- Оценка метода моментов (Method of moments, MOM):

$$\widehat{\alpha}_n = \alpha_n(\widehat{\theta}).$$

- Оценка максимального правдоподобия (ОМП, Maximum Likelihood Estimate, MLE):

$$\widehat{\theta} : \mathcal{L}_n(\theta) \to \max.$$

- Оценка апостериорного максимума (Maximum A Posteriori Estimate, MAP):

$$\widehat{\theta} : f(\theta|x)g(\theta) \to \max.$$

- Метод Наименьших Квадратов (НМК, least squares):

$$\widehat{\theta} : \sum_{i=1}^{n}(\widehat{X}(\theta) - X)^2 \to \min.$$

Непараметрическое оценивание

Постановка задачи

Пусть задана выборка: $X_1, \ldots, X_n \sim F$.
F - абсолютно непрерывная функция распределения с неизвестной плотностью p. Необходимо оценить p в точке x, т.е. построить $\hat{p}_n(x) = \hat{p}_n(x; X_1, \ldots, X_n)$.

NB: Ранее в аналогичных задачах мы искали $p \in \{p(x;\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}^n$, где есть зависимость от параметров.

MSE, функция риска

Пусть в точке x_0 построена оценка $\hat{p}_n(x_0)$ плотности. Рассматривая квадратичную функцию потерь, приходим к следующему понятию.

Определение

Mean Square Error:

$$MSE(\hat{p}_n, p; x_0) = \mathbb{E}_p[(\hat{p}_n(x_0) - p(x_0))^2].$$

MISE

Если же построена оценка $\hat{p}_n(x)\ \forall x \in \mathbb{R}$, то

Определение

Mean Integrated Squared Error:

$$MISE(\hat{p}_n, p) = \mathbb{E}_p \left[\int_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx \right].$$

Bias

Определение

Смещение (bias)

$bias(x_0) = \mathbb{E}_p \hat{p}_n(x_0) - p(x_0)$

Разложение ошибки

Лемма

$$MSE(\hat{p}_n, p, x_0) = bias^2(x_0) + \mathrm{Var_p \hat{p}_n(x_0)} =$$
$$= [\mathbb{E}_p \hat{p}_n(x_0) - p(x_0)]^2 + \mathbb{E}_p[\hat{p}_n(x_0) - \mathbb{E}_p \hat{p}_n(x_0)]^2$$

Лемма

$$MISE(\hat{p}_n, p) = \int\limits_{\mathbb{R}} bias^2(x)dx + \int\limits_{\mathbb{R}} \mathrm{Var_p \hat{p}_n(x) dx}$$

Гистограмма

Простейший способ оценить плотность - построить гистограмму

Возьмём интервал $[a, b) \ni X_1, \dots X_n$

Поделим его на M равных частей Δ_i размера $h = \frac{b-a}{M}$:

$$\Delta_i = [a + ih, a + (i+1)h), i = 0, 1, \dots, M-1].$$

Пусть ν_i - число элементов выборки, попавших в Δ_i;

Определение

$$\hat{p}_n(x) = \begin{cases} \frac{\nu_0}{nh}, & x \in \Delta_0, \\ \dots \\ \frac{\nu_{N-1}}{nh}, & x \in \Delta_{M-1}; \end{cases} = \frac{1}{nh} \sum_{i=0}^{M-1} \nu_i \mathbb{I}\{x \in \Delta_i\}$$

При $x \in \Delta_i$ и малом h: $\mathbb{E}_p \hat{p}_n(x) = \frac{\mathbb{E}\nu_j}{nh} = \frac{\int_{\Delta_j} p(u)du}{h} \approx \frac{p(x)h}{h} = p(x)$

Гистограмма: определение параметра сглаживания

Рассмотрим выбор h - параметра сглаживания

Проведём вычисления для $x_0 \in \Delta_j$:

$$bias(x_0) = \mathbb{E}_p \hat{p}_n(x_0) - p(x_0) = \frac{1}{h} \int\limits_{\Delta_j} p(x)dx - \frac{1}{h} \int\limits_{\Delta_j} p(x_0)dx =$$

$$= \frac{1}{h} \int\limits_{\Delta_j} (p(x) - p(x_0))dx \approx \frac{1}{h} \int\limits_{\Delta_j} p'(x_0)(x - x_0)dx \approx$$

$$\approx p'(x_0)[a + (j + \frac{1}{2})h - x_0]$$

Определение параметра сглаживания

$$\int\limits_a^b bias^2(x_0)dx_0 = \sum_{j=0}^{N-1} \int\limits_{\Delta_j} bias^2(x_0)dx_0 =$$

$$= \sum_{j=0}^{N-1} \int\limits_{\Delta_j} [p'(x_0)]^2 [a + (j + \frac{1}{2})h - x_0]^2 dx_0 \approx$$

$$\approx \sum_{j=0}^{N-1} [p'(a + (j + \frac{1}{2})h)]^2 \int\limits_{\Delta_j} (a + (j + \frac{1}{2})h - x_0)^2 dx_0$$

$$= \sum_{j=0}^{N-1} [p'(a + (j + \frac{1}{2})h)]^2 \left(-\frac{(a + (j + \frac{1}{2})h - x_0)^3}{3} \right) \Bigg|_{\Delta_j} \approx$$

$$\approx \left(\int\limits_a^b [p'(x)]^2 dx \right) \frac{h^2}{12}.$$

Определение параметра сглаживания

$$\mathbb{V}ar_p \hat{p}_n(x_0) = \mathbb{V}ar_p \frac{\nu_j}{nh} = \frac{1}{(nh)^2} \mathbb{V}ar_p \nu_j =$$

$$= \frac{1}{(nh)^2} n \int\limits_{\Delta_j} p(x)dx(1 - \int\limits_{\Delta_j} p(x)dx) \approx \frac{1}{nh^2} \int\limits_{\Delta_j} p(x)dx$$

$$\int\limits_a^b \mathbb{V}ar_p \hat{p}_n(x_0) dx_0 = \sum_{j=0}^{N-1} \left(\frac{1}{nh^2} \int\limits_{\Delta_j} p(x)dx \right) h =$$

$$= \frac{1}{nh} \int\limits_a^b p(x)dx = \frac{1}{nh}$$

Определение параметра сглаживания

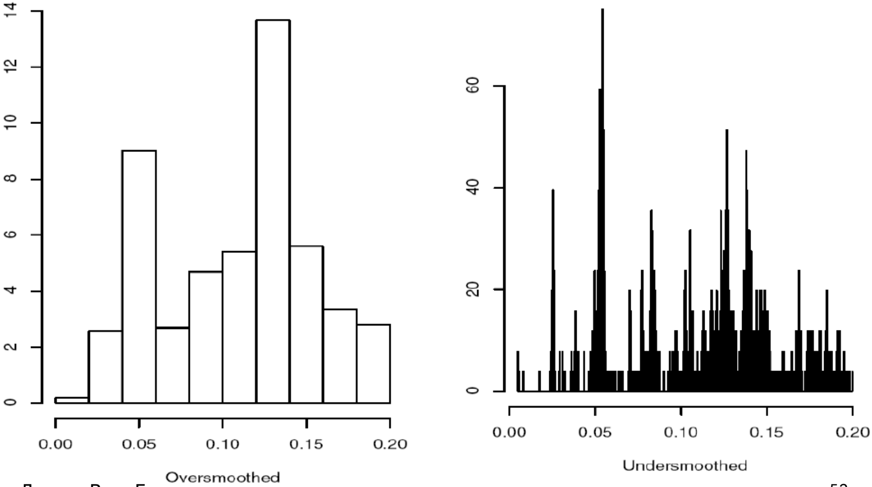
Таким образом,

$$MISE(\hat{p}_n, p) = \left(\int\limits_{\mathbb{R}} [p'(x)]^2 dx \right) \frac{h^2}{12} + \frac{1}{nh}$$

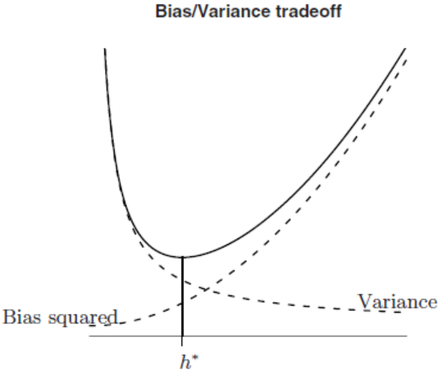
Чем больше h, тем больше смещение и меньше дисперсия, и наоборот. Это называется bias-variance tradeoff.

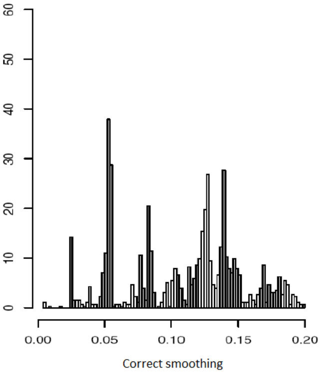
Ситуации с большим h - oversmoothing, с маленьким - undersmoothing.

Пример неоптимального сглаживания



Определение параметра сглаживания





Определение параметра сглаживания

Значение h, при котором $MISE$ минимальный

$$h^* = \frac{1}{n^{\frac{1}{3}}} \left(\frac{6}{\int\limits_{\mathbb{R}} [p'(x)]^2 dx} \right)^{\frac{1}{3}}.$$

При этом

$$MISE(\hat{p}_n, p) \approx \frac{C}{n^{\frac{2}{3}}}, \text{ где } C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int\limits_{\mathbb{R}} [p'(x)]^2 dx \right)^{\frac{1}{3}}.$$

Таким образом, при использовании гистограммы с оптимальным h, $MISE$ убывает как $n^{-\frac{2}{3}}$:

Определение параметров сглаживания

На практике h^* нельзя вычислить, так как h^* зависит от неизвестной истинной плотности.
Поэтому оценим $MISE$ и минимизируем по h оценку.
Так как:

$$\int\limits_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx = \int\limits_{\mathbb{R}} \hat{p}_n(x)^2 dx - 2 \int\limits_{\mathbb{R}} \hat{p}_n(x) p(x) dx + \int\limits_{\mathbb{R}} p(x)^2 dx,$$

то достаточно оценить и минимизировать только

$$\mathcal{J}(h) = \int\limits_{\mathbb{R}} \hat{p}_n(x)^2 dx - 2 \int\limits_{\mathbb{R}} \hat{p}_n(x) p(x) dx.$$

Определение параметра сглаживания

Определение

Оценка риска с помощью кросс-валидации:

$$\hat{\mathcal{J}}(h) = \int\limits_{\mathbb{R}} [\hat{p}_n(x)]^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{p}_{(-i)}(X_i),$$

где $\hat{p}_{(-i)}$ - оценка гистограммы по выборке без $i-$ого наблюдения.

Теорема

$\mathbb{E}\hat{\mathcal{J}}(h) \approx \mathbb{E}\mathcal{J}(h)$

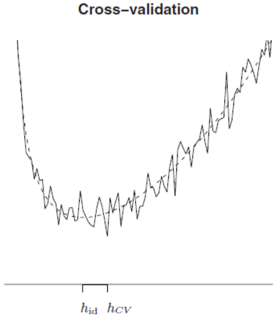
Теорема

Для гистограм оценка функции риска:

$$\hat{\mathcal{J}}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(h-1)h} \sum_{i=1}^{M} \left(\frac{\nu_j}{n}\right)^2$$

Определение параметра сглаживания

Типичное поведение $\hat{\mathcal{J}}(h)$ имеет вид:



Таким образом, вместо неизвестного $MISE$ можно минимизировать $\hat{\mathcal{J}}(h)$ и найти оптмимальное h_{cv}, которое будет недалеко от $h_{id} = h^*$.

Доверительная трубка для плотности

Пусть необходимо построить доверительные интервалы для p.
Для этого будем использовать гистограмму $\hat{p}_n(x)$, определенную ранее.
Определим
$\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \frac{\int_{\Delta_j} p(u)du}{h}$ для $x \in \Delta_j$.
По сути, $\overline{p_n}$ - "гистограммное" усреднение плотности p.

Определение

Пара функций $(p_-(x), p_+(x))$ является $1-\alpha$ доверительной областью (трубкой), если для любого x:

$$\mathbb{P}_p\left(p_-(x) \leq \overline{p_n}(x) \leq p_+(x)\right) \geq 1-\alpha$$

Доверительная трубка для плотности

Теорема

Пусть $M = M(n)$ - число ячеек в гистограмме $\hat{p}_n$, причем $M(n) \to \infty$ и $\frac{M(n)\log(n)}{n} \to \infty$ при $n \to \infty$.

Определим

$$p_-(x) = (max\{\sqrt{\hat{p}_n(x)} - C, 0\})^2, p_+(x) = (\sqrt{\hat{p}_n(x)} + C)^2,$$

где $C = \frac{1}{2} z_{\frac{\alpha}{2M}} \sqrt{\frac{M}{n(b-a)}}$

Тогда $(p_-(x), p_+(x))$ является $1-\alpha$ доверительным интервалом.

Доверительная трубка для плотности

Из центральной предельной теоремы

$$\frac{\nu_j}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{x_i \in \Delta_j\} \sim \mathcal{N}\left(\int\limits_{\Delta_j} p(x)dx, \frac{\int\limits_{\Delta_j} p(x)dx(1-\int\limits_{\Delta_j} p(x)dx)}{n}\right)$$

Согласно дельта-методу $\sqrt{\frac{\nu_j}{n}} \sim \mathcal{N}\left(\sqrt{\int\limits_{\Delta_j} p(x)dx}, \frac{1}{4n}\right)$. Более того, можно показать, что $\sqrt{\frac{\nu_j}{n}}$ приблизительно независимы. Тогда $2\sqrt{n}\left(\sqrt{\frac{\nu_j}{n}} - \sqrt{\int\limits_{\Delta_j} p(x)dx}\right) \approx \xi_j$, где $\xi_0, \ldots, \xi_{M-1} \sim \mathcal{N}(0,1)$.

Доверительная трубка

$$
\begin{aligned}
A &= \{p_-(x) \leq \overline{p_n}(x) \leq p_+(x) \forall x\} = \\
&= \{\sqrt{p_-(x)} - c \leq \sqrt{\overline{p_n}(x)} \leq \sqrt{p_+(x)} + c \forall x\} = \\
&= \{\max_x |\sqrt{\hat{p}(x)} - \sqrt{\overline{p_n}(x)}| \leq c\}
\end{aligned}
$$

Доверительная трубка для плотности

Тогда $\mathbb{P}(A^c) = \mathbb{P}\{\max_x |\sqrt{\hat{p}_n(x)} - \sqrt{\overline{p_n}(x)}| > c\} =$

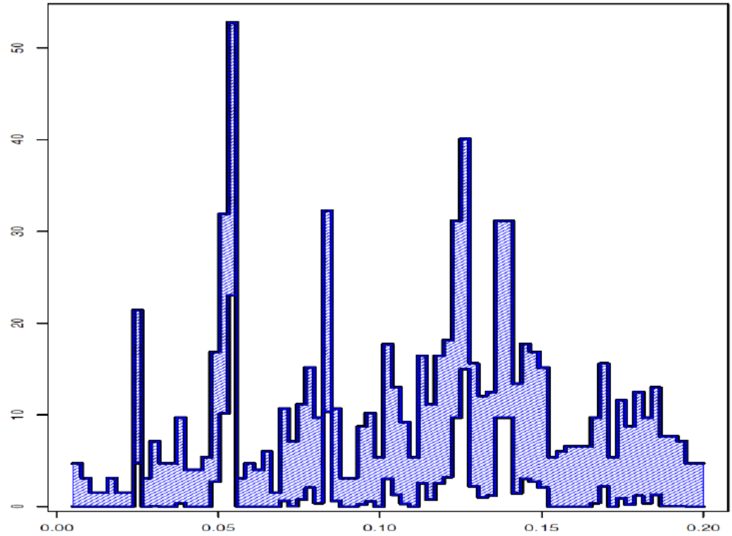
$$\mathbb{P}\left\{\max_{j=\overline{0,M-1}} \left|\sqrt{\frac{\nu_j}{nh}} - \sqrt{\frac{\int\limits_{\Delta_j} p(x)dxh}{h}}\right| > c\right\} \approx$$

$$\mathbb{P}\left\{\max_{j=\overline{0,M-1}} \frac{|\xi_j|}{2\sqrt{nh}} > \frac{z_{\frac{\alpha}{2n}}}{2}\sqrt{\frac{M}{n(b-a)}}\right\} = \mathbb{P}\{\max_{j=\overline{0,M-1}} |\xi_j| > z_{\frac{\alpha}{2M}}\} \leq$$

$$\sum_{j=0}^{M-1} \mathbb{P}\{|\xi_j| > z_{\frac{\alpha}{2M}}\} = \sum_{j=0}^{M-1} \frac{\alpha}{M} = \alpha,$$

т.е. для предъявленных $p_-(x), p_+(x)$ выполнено определение доверительной трубки.

Доверительная трубка для плотности



Комментарии о доверительных трубках

- Важным условием для предыдущего вывода является наличие большого количества семплов n. В случае малого количества семплов ситуация может отличаться, в зависисости от использованного метода оценки.
- Разные отрасли используют разные определения ширины доверительной трубки (от 68% до 100%).

Ядерная оценка плотности

Позволяет получить более гладкие по сравнению с гистограммной оценки, быстрее сходящиеся к плотности.

Определение

Ядро - функция K такая, что

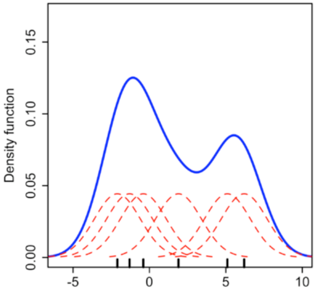
$$K(x) \geq 0, \int\limits_{\mathbb{R}} K(x)dx = 1, \int\limits_{\mathbb{R}} xK(x)dx = 0, \sigma_K^2 \equiv \int\limits_{\mathbb{R}} x^2 K(x)dx$$

Ядерная оценка плотности

Определение

Ядерная оценка плотности имеет вид:

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), h \text{ — } \underline{\text{ширина ядра}}$$



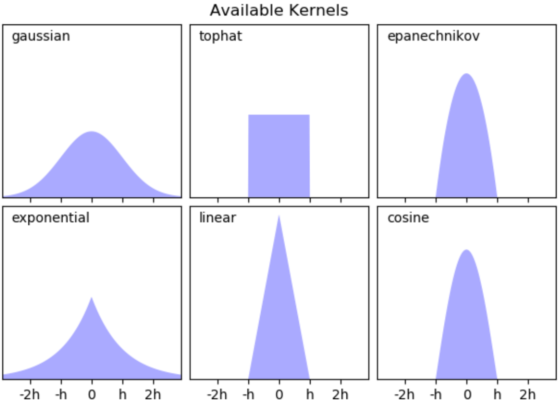
Виды ядер

Examples

- $K(x) = \frac{1}{2}\mathbb{I}\{|x| < 1\}$ — прямоугольное ядро
- $K(x) = (1 - |x|)\mathbb{I}\{|x| < 1\}$ — треугольное ядро
- $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ — Гауссовское ядро
- $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}\{|x| < 1\}$ — ядро Епанечникова

Далее мы будем рассматривать только гладкие ядра.

Примеры ядер



Вид ядерной функции K влияет на "качество" оценки не так сильно, как выбор ширины ядра h.

Picture credit:http://scikit-learn.org/stable/modules/density.html

Выбор ширины ядра

Теорема

$$MISE(\hat{p}_n, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int\limits_{\mathbb{R}} (p''(x))^2 dx + \frac{1}{nh} \int\limits_{\mathbb{R}} (K(x))^2 dx$$

Минимум достигается при $h = h^*$:

$$h^* = \left(\frac{1}{n} \frac{\int\limits_{\mathbb{R}} (K(x))^2 dx}{\left(\int\limits_{\mathbb{R}} x^2 K(x) dx \right)^2 \left(\int\limits_{\mathbb{R}} p''(x))^2 dx \right)} \right)^{\frac{1}{5}}$$

При этом $MISE(\hat{p}_n, p) = O\left(n^{-\frac{4}{5}}\right)$

Выбор ширины ядра

Воспользуемся bias-variance decomposition:

$$bias(x) = \mathbb{E}_p \hat{p}_n(x) - p(x) =$$

$$\int\limits_{\mathbb{R}} \left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right) p(x_1) \dots p(x_n) dx_1 \dots dx_n -$$

$$\frac{1}{n} \sum_{i=1}^{n} \int\limits_{\mathbb{R}} K(z) p(x) dz \approx \int\limits_{\mathbb{R}} K(z)[-p'(x)zh + p''(x)\frac{(zh)^2}{2}]dz =$$

$$\frac{1}{2}\sigma_K^2 h^2 p''(x)$$

$$\int\limits_{\mathbb{R}} (bias(x))^2 dx = \frac{1}{4}\sigma_K^4 h^4 \int\limits_{\mathbb{R}} [p''(x)]^2 dx$$

Выбор ширины ядра

$$\int\limits_{\mathbb{R}} \mathbb{V}ar_p \hat{p}_n(x)dx = \int\limits_{\mathbb{R}} \mathbb{V}ar_p[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{V}ar_p K(\frac{x-x_i}{h})dx \leq \frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{E}_p K(\frac{x-x_i}{h})^2 dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 p(x_i)dx_i dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 dx dx_i =$$

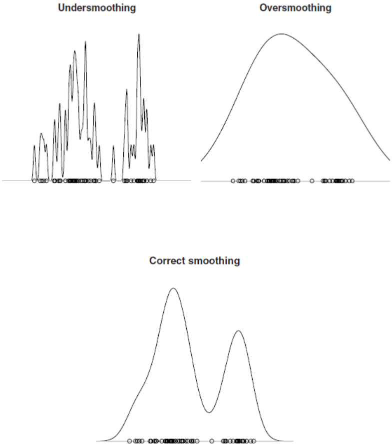
$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)dx_i h\int\limits_{\mathbb{R}} K^2(z)dz = \frac{1}{nh}\int\limits_{\mathbb{R}} K^2(z)dz$$

Минимум $MISE(\hat{p}_n, p)$ достигается в некотором h^*.

Выбор ширины ядра

Подставляя h^* в $\hat{p}_n$, получаем, что $MISE = O(n^{-\frac{4}{5}})$, т.е. сходимость ядерной оценки лучше, чем у гистограммы. Можно показать, что при достаточно общих условиях нельзя получить скорость лучше, чем $n^{-\frac{4}{5}}$.

Выбор ширины ядра



Как и в случае с гистограммной, при больших h имеет место oversmoothing, а при маленьких - undersmoothing из-за bias-variance tradeoff.

Доверительный интервал

Определим $\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \int\limits_{\mathbb{R}} \frac{1}{h} K(\frac{x-u}{h}) p(u) du$. Допустим, что $supp(p) \subset (a, b)$.

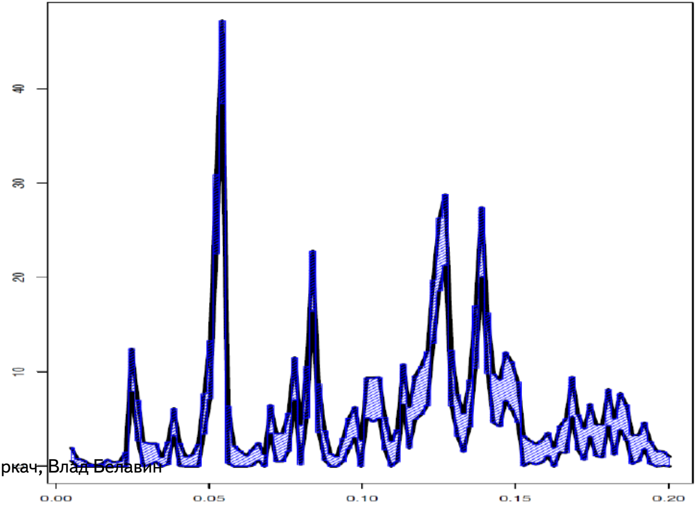
Тогда определим $(1-\alpha)$ доверительную трубку.

$$p_-(x) = \hat{p}_n(x) - \frac{z_\alpha}{\sqrt{n}} s(x),$$

$$p_+(x) = \hat{p}_n(x) + \frac{z_\alpha}{\sqrt{n}} s(x).$$

Где $s^2(x) = \frac{1}{n-1} \sum\limits_{i=1}^{n} [Y_i(x) - \overline{Y_n}(x)]^2, Y_i(x) = \frac{1}{h} K(\frac{x-X_i}{h})$, $z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, $\Phi-$ функция стандартного нормального распределения. w - эффективная ширина ядра.

Доверительный интервал для усредненной плотности



Ядерная оценка плотности: многомерный случай

Пусть теперь данные многомерные, то есть $i-$ое наблюдение - вектор размерности d:

$$X_i = [X_i^1, \dots X_i^d]^T.$$

Пусть $h = [h_1, \dots, h_d]^T$ - вектор ширины ядра вдоль каждого измерения.
Тогда:

$$\hat{p}_n(x) = \frac{1}{nh_1 \cdot ... \cdot h_d} \sum_{i=1}^{n} \left[\prod_{j=1}^{d} K\left(\frac{x_j - X_i^j}{h_j}\right) \right],$$

где $x = [x_1, \dots, x_d]^T$ — произвольная точка в $\mathbb{R}^d$

Ядерная оценка плотности: многомерный случай

Для такой оценки риск
$MISE(\hat{p}_n, p) \approx$
$$\frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int\limits_{\mathbb{R}^d} p_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int\limits_{\mathbb{R}^d} p_{jj}(x) p_{kk}^2(x)dx\right] +$$
$$\frac{\left(\int\limits_{\mathbb{R}^d} K^2(x)dx\right)^d}{nh_1 \cdot \ldots \cdot h_d},$$
где $p_{jj}(x) = \frac{\partial^2 p(x)}{\partial x_j^2}$

Оптимальная ширина ядра $h_i^* \approx cn^{-\frac{1}{4+d}}$

При этом риск имеет порядок: $MISE(\hat{p}_n, p) = O(n^{-\frac{4}{4+d}})$.

Проклятие размерности

Оптимальный порядок риска $O(n^{-\frac{4}{4+d}})$, т.е. наблюдаем "проклятье размерности" - при росте d скорость сходимости к истинной плотности падает.

Рассмотрим таблицу объёмов выборки, необходимых для того, чтобы средний квадрат ошибки в нуле был меньше 0.1 в зависимости от размерности наблюдений в случае многомерной нормальной плотности и оптимальной ширины ядра:

d	1	2	3	4	5	6	7	8	9
n	4	19	67	223	768	2790	10700	43700	187000

где d — размерность данных, n — необходимый объём выборки.