СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 15. Задачи робастной статистики при обработке больших данных. Часть 1

Введение

Тема данной лекции --- исследование устойчивости (робастности) и чувствительности основных оценок параметра положения, классических и робастных, при динамическом изменении моделей распределений данных.

Приведены результаты исследования робастности решений задач оценивания параметра положения распределений для тех же моделей изменения формы распределения данных, ранее использованных в предыдущей лекции:

- изменение параметров смеси распределений;
- изменение параметра формы обобщённого гауссовского распределения;
- изменения числа степеней свободы для распределения Стьюдента.

Изучались следующие М-оценки параметра положения: выборочные среднее и медиана, робастные оценки Хубера и Хампеля, стойкие оценки Мешалкина-Шурыгина.

Робастное оценивание

Допустим, что $x_1, x_2, \dots, x_n$ – одинаково распределённые, независимые реализации случайной величины с функцией распределения $F(x, \theta)$, где неизвестный параметр θ подлежит оцениванию. Рассматриваются параметры, которые могут быть представлены в виде функционалов от функции распределения: $\theta = T(F)$, например, $\theta = T(F) = \int_{-\infty}^{\infty} x dF(x)$ – математическое ожидание.

Эмпирическая функция распределения: $F_n(x) = \frac{k}{n},\ x \in \left(x_{(k)}, x_{(k+1)}\right], k \in \overline{0 \dots n}$;
$x_{(1)} \leq x_{(2)} \leq \cdots x_{(n)}$– порядковые статистики, $x_{(0)} = -\infty,\ x_{(n+1)} = \infty$.

Функция выборочных значений $T_n(x_1, \dots, x_n) = T(F_n)$ называется **статистикой.**

Статистика называется **состоятельной,** если $T(F_n) \xrightarrow{P} \theta$ при $F_n \to F,\ n \to \infty$.

Смещение статистики: $E[T(F_n)] - \theta$.

Робастное оценивание

Оценка характеристики распределения: строится статистика $\hat{\theta} = T(F_n)$, которая будет использована вместо параметра θ. Статистику $\hat{\theta}$ принято называть *оценкой* θ.

Робастность – свойство статистической процедуры быть устойчивой к неконтролируемым отклонениям от принятых моделей распределений данных.

Первый подход к построению робастных оценок был предложен Хубером (Huber, 1964), в основе которого лежит минимаксный принцип построения наилучшего решения в наихудшей ситуации.

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho(x_i, \theta)$$

$\rho(x)$- **функция потерь** (контраста):
$\rho(0) = 0, \rho(x) \geq 0 \ \forall x, \ \rho(x) = \rho(-x),$
$\rho(x) \geq \rho(y)$ при $|x| \geq |y|$

Оценки, полученные данным образом, называются ***M-оценками*** (Huber, 1964).

Робастное оценивание

Частные случаи *M*-оценок:

- оценки **метода наименьших квадратов** при $\rho(x) = x^2$,
- оценки **метода наименьших модулей** при $\rho(x) = |x|$,
- оценки **метода максимального правдоподобия** при $\rho(x) = -\ln f(x, \theta)$.

M-оценки можно рассматривать как класс оценок, обобщающий класс оценок максимума правдоподобия.

Пусть функция ρ имеет производную $\psi = \frac{\partial \rho}{\partial \theta}$, называемую оценочной функцией. В этом случае задача оценивания имеет в

$$\sum_{i=1}^{n} \psi(x_i, \hat{\theta}) = 0$$

В задаче оценивания параметра положения θ:

$$f(x, \theta) = f(x - \theta), \qquad \psi(x, \theta) = \psi(x - \theta)$$

Робастное оценивание

Функция влияния:

$$\mathrm{IF}(x;T,F)=\lim_{t\to 0}\frac{T\big((1-t)F+t\Delta_x\big)-T(F)}{t}.$$

Функция IF есть мера влияния бесконечно малого «загрязнения» в точке x на значение оценки. Для *M*-оценок с оценочной функцией ψ функция влияния пропорциональна оценочной функции:

$$IF(x;T,f)=\frac{\psi(x)}{\int_{-\infty}^{+\infty}\psi'(x)f(x)dx}$$

Асимптотическая дисперсия оценки:

$$V(T,F)=\int IF^2(x;T,F)dF(x),$$

$$V(\psi,f)=\frac{\int\psi^2(x)f(x)dx}{\left(\int\psi'(x)f(x)dx\right)^2}.$$

Робастное оценивание

Эффективность *M*-оценки с оценочной функцией ψ называется отношение дисперсий оценки максимума правдоподобия с оценочной функцией $\psi_{ML}(x) = -\frac{f'(x)}{f(x)}$ и этой оценки:

$$\mathrm{eff}(\psi, f) = \frac{V(\psi_{ML},\ f)}{V(\psi, f)} = \frac{1}{I(f)V(\psi, f)},$$

где $I(f) = \int_{-\infty}^{+\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx$ – информация Фишера

Робастное оценивание

Оценки параметра положения и их оценочные функции:

1. Выборочное среднее: $\psi(x) = x$

2. Выборочная медиана: $\psi(x) = \operatorname{sgn}(x)$

3. Оценка Хубера - компромиссная оценка между выборочным средним и выборочной медианой:

Линейная с ограничениями оценочная функция:

$$\psi(x) = \begin{cases} x, & |x| \le k \\ k\operatorname{sgn}(x), & |x| > k \end{cases}$$

Оценка Хубера есть наилучшая в смысле максимума правдоподобия оценка с оценочной функцией $\psi_{ML}^*(x)$ для наихудшей плотности $f^* = \arg\min_{f \in \mathcal{F}} V(\psi, f)$, минимизирующей информацию Фишера в классе приближенно-нормальных распределений (Huber, 1964).

Робастное оценивание

4. Оценка Хампеля

Оценочная функция Хампеля:

$$\psi(x) = \begin{cases} x, & 0 \le |x| \le a \\ a\,\mathrm{sgn}(x), & a \le |x| \le b \\ a\dfrac{r-|x|}{r-b}\mathrm{sgn}(x), & b \le |x| \le r \\ 0, & r \le |x| \end{cases}$$

где $0 < a \le b < r < \infty$ (Andrews et al., 1972).

Робастное оценивание

5. Оценка Мешалкина-Шурыгина минимальной чувствительности (MVS) с оценочной функцией (Шурыгин, 2000):

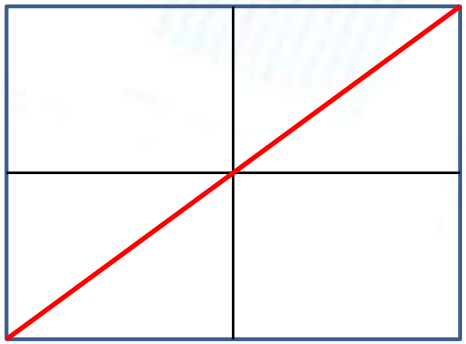
$$\psi_{MVS}(x) = \arg\min_{\psi \in C^1(\mathbb{R})} VS(\psi, f) = -f'(x)$$

где $\mathrm{VS}(\psi, f) = \dfrac{\int \psi^2(x)dx}{(\int \psi'(x)f(x)dx)^2}$ – чувствительность дисперсии оценки.

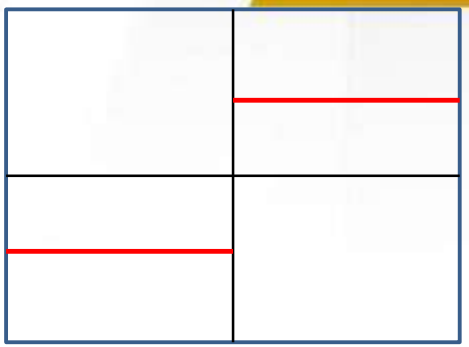
6. Радикальная оценка Мешалкина-Шурыгина с оценочной функцией:

$$\psi_{RAD}(x) = -\frac{f'(x)}{\sqrt{f(x)}}$$

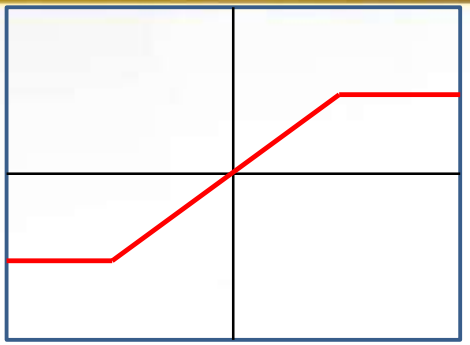
Оценочные функции



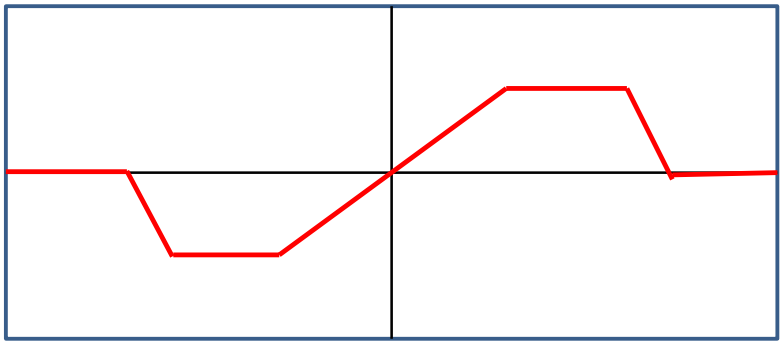
1.Выборочное среднее



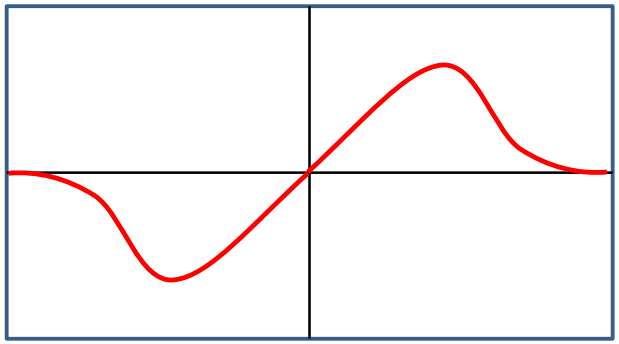
2.Выборочная медиана



3. Оценка Хубера



4. Оценка Хампеля



5,6. Оценки Мешалкина-Шурыгина

Робастное оценивание

1. Робастность оценки определяется ограниченностью ее оценочной функции. Отметим, все оценочные функции, приведенные здесь, ограничены, кроме оценочной функции выборочного среднего.
2. И именно выборочное среднее не является робастной оценкой параметра положения в силу неограниченного влияния больших по модулю ошибок.
3. Оценочные функции Хампеля и Мешалкина-Шурыгина кроме того, что они ограничены, близки к нулю при достаточно больших значениях аргумента. Т.е. большие ошибки учитываются в этих оценках с пренебрежимо малыми весами.