

СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 16. Задачи робастной статистики
при обработке больших данных. Часть 2



Исследование робастности оценок

Вычисление характеристик оценок параметра положения

Производится сравнение различных оценок параметра положения как по смещению, так и по дисперсии оценки. При этом характеристики оценок вычисляются приближенно по методу Монте Карло:

$$bias \widehat{\theta}_n = \frac{1}{N} \sum_{j=1}^N \widehat{\theta}_{n,j} - \theta$$

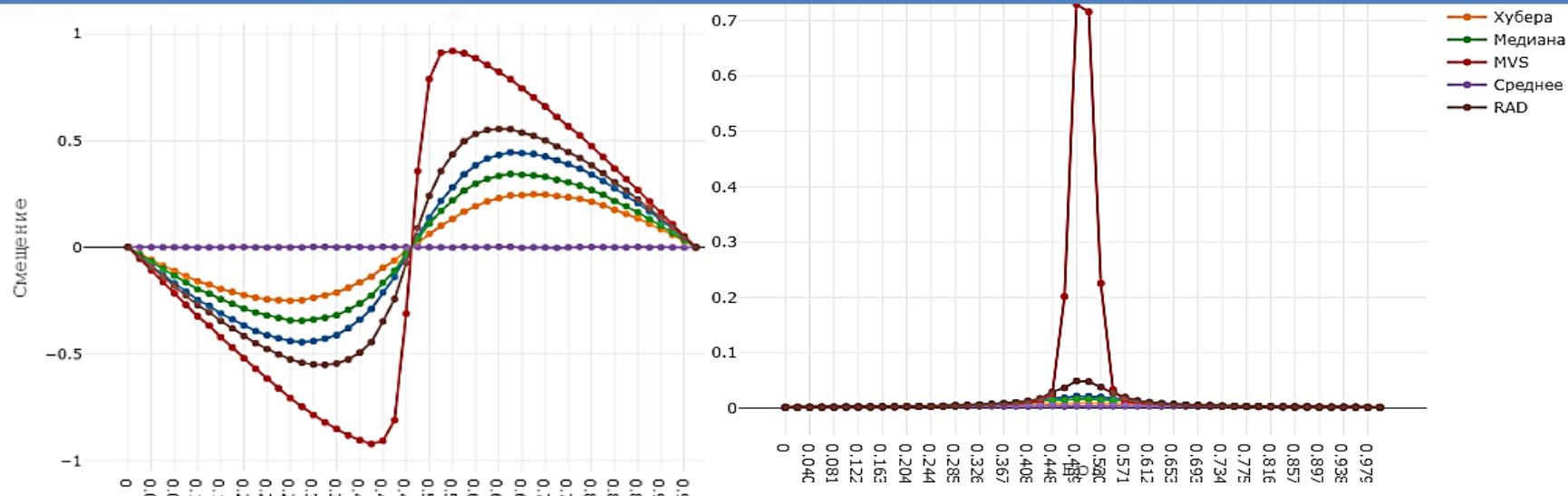
$$D\widehat{\theta}_n = \frac{1}{N} \sum_{j=1}^N \left(\widehat{\theta}_{n,j} \right)^2 - \left(\frac{1}{N} \sum_{j=1}^N \widehat{\theta}_{n,j} \right)^2.$$

Полагаем в оценке Хубера $k = 1.44$; в оценке Хампеля: $a = 1.31, b = 2.039, r = 4.16$.



Результаты исследования

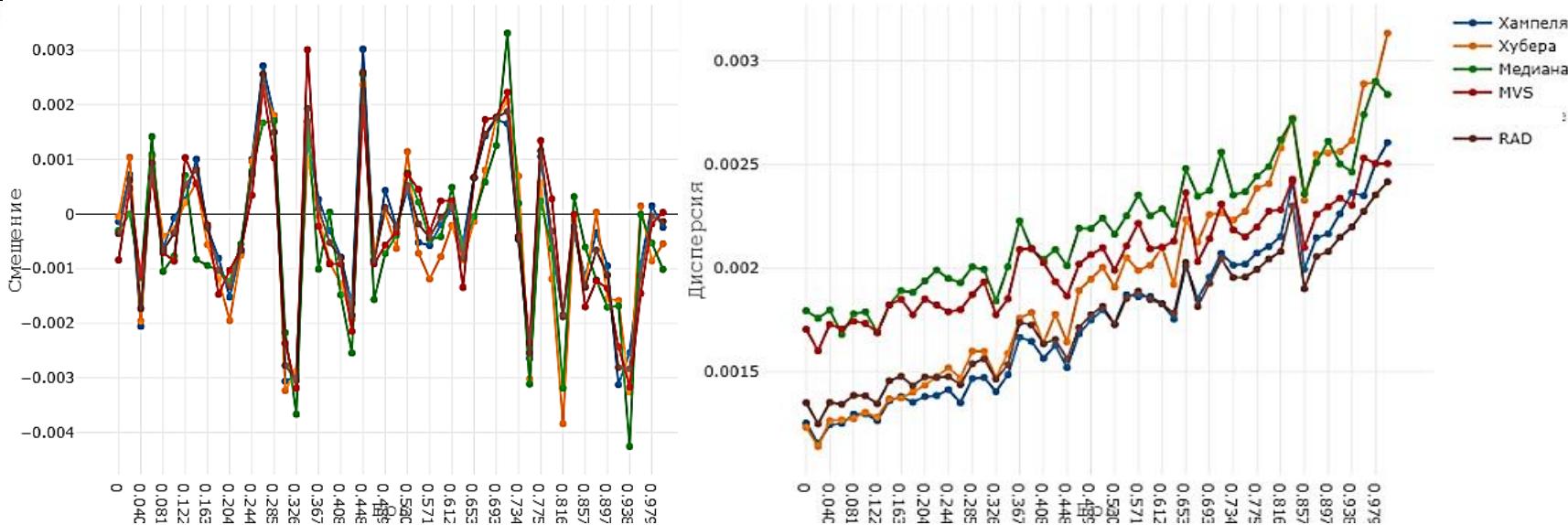
Поведение оценок параметра положения в модели смеси двух нормальных распределений в зависимости от ε



Здесь выборочное среднее является наилучшей оценкой как по дисперсии, так и по смещению. Второй по предпочтительности является оценка Хубера.

Результаты исследования

Поведение оценок параметра положения в модели смеси нормального распределения и распределения Коши в зависимости от ε

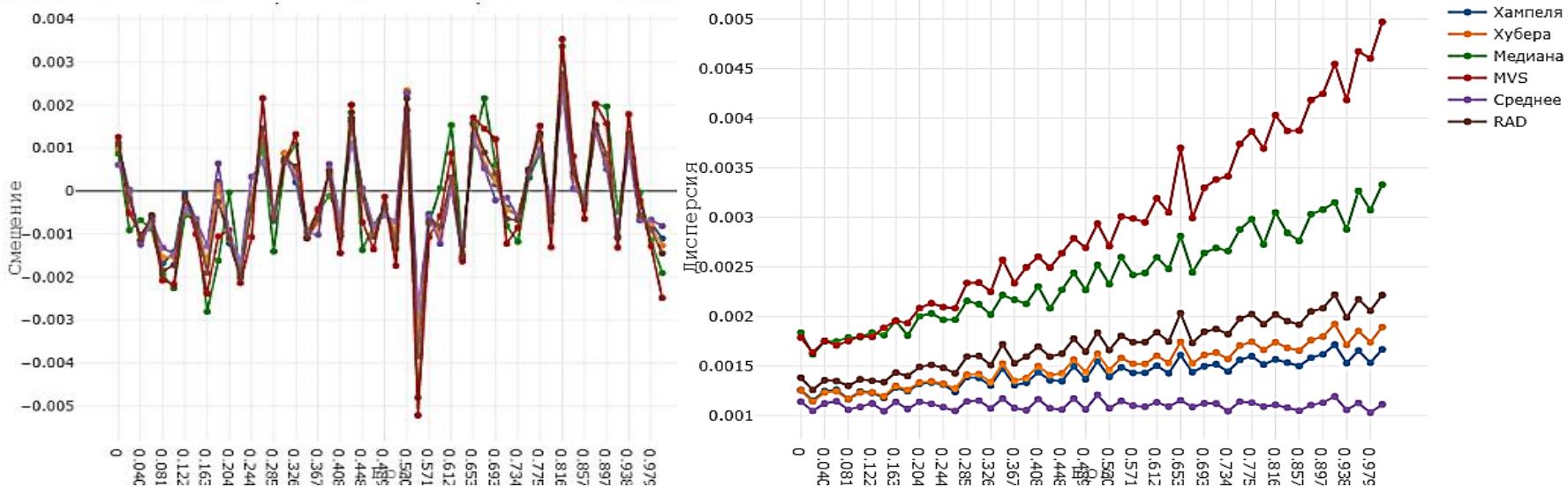


Оценка выборочного среднего ведет к резким скачкам. Приемлемым выбором в данной модели являются оценка Хампеля и радикальная (RAD) оценка Мешалкина-Шурыгина. Для всех оценок характерен рост дисперсии при увеличении ε .



Результаты исследования

Поведение оценок параметра положения в модели смеси нормального и равномерного распределения



Выборочное среднее является оценкой, обеспечивающей минимальную дисперсию.
Наблюдается низкое значение дисперсия оценок Хубера, Хампеля и RAD оценки
Мешалкина-Шурыгина.



Результаты исследования

Таблица. Рекомендации по выбору робастных оценок

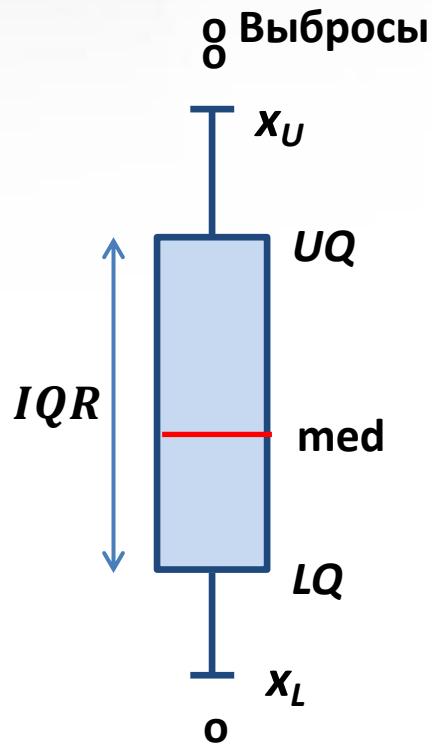
Модель данных	Рекомендуемые оценки параметра положения
Смесь двух нормальных распределений	Выборочное среднее, оценка Хубера
Смесь нормального распределения и Коши	Оценка Хампеля, радикальная оценка Мешалкина-Шурыгина
Смесь нормального и равномерного распределения	Выборочное среднее, оценка Хампеля, оценка Хубера
Распределение Стьюдента	Оценка Хубера, оценка Хампеля
Обобщённое гауссовское распределение при малых значениях параметра формы	Оценка Хампеля, оценка Хубера
Обобщённое гауссовское распределение при больших значениях параметра формы	Оценка Хубера, оценка Хампеля, Медиана



Двухэтапная оценка

Двухэтапная процедура оценивания состоит в следующем:

1. Отбраковка выбросов с помощью боксплота Тьюки.
2. Оценивание параметра положения путем вычисления выборочного среднего для оставшихся элементов выборки.



Сравнение методов

(обобщенное гауссовское распределение)

$\beta = 1$	n	Mean	Med	Huber	Hampel	MVS	RAD	TwoStep
	10	0,1786	0,1279	0,1310	0,1399	0,1535	0,1449	0,1419
	40	0,0484	0,0304	0,0338	0,0360	0,0365	0,0332	0,0375
	100	0,0198	0,0113	0,0136	0,0143	0,0145	0,0130	0,0151
$\beta = 1.5$		Mean	Med	Huber	Hampel	MVS	RAD	TwoStep
	10	0,1134	0,1255	0,1092	0,1154	0,1320	0,1271	0,1094
	40	0,0311	0,0355	0,0297	0,0313	0,0348	0,0335	0,0310
	100	0,0126	0,0141	0,0119	0,0126	0,0136	0,0131	0,0127
$\beta = 2$		Mean	Med	Huber	Hampel	MVS	RAD	TwoStep
	10	0,0920	0,1271	0,0992	0,1011	0,1346	0,1103	0,0967
	40	0,0246	0,0367	0,0263	0,0268	0,0354	0,0289	0,0260
	100	0,0099	0,0154	0,0107	0,0109	0,0146	0,0118	0,0104
$\beta = 5$		Mean	Med	Huber	Hampel	MVS	RAD	TwoStep
	10	0,0553	0,1095	0,0632	0,0595	0,3655	0,5861	0,0622
	40	0,0150	0,0359	0,0178	0,0164	0,2891	0,5338	0,0153
	100	0,0060	0,0152	0,0072	0,0066	0,2099	0,5037	0,0061
$\beta = 100$		Mean	Med	Huber	Hampel	MVS	RAD	TwoStep
	10	0,0321	0,0736	0,0327	0,0321	0,5295	0,6682	0,0362
	40	0,0089	0,0245	0,0089	0,0089	0,5811	0,8951	0,0090
	100	0,0035	0,0104	0,0036	0,0035	0,5995	0,9188	0,0035



Сравнение методов (распределение Стьюдента)

df = 1 n	Mean	Med	Huber	MVS	RAD	TwoStep
10	277307	0,297	1,17	6,39	5,02	0,4934
40	23634	0,064	0,096	15,3	8,47	0,0684
100	4583	0,025	0,039	30,2	18,0	0,0260
df = 1.5	Mean	Med	Huber	MVS	RAD	TwoStep
10	6,583	0,217	0,233	0,814	0,517	0,2460
40	5,700	0,053	0,055	0,448	0,188	0,0525
100	12,759	0,022	0,022	0,324	0,144	0,0202
df = 2	Mean	Med	Huber	MVS	RAD	TwoStep
10	1,43	0,178	0,180	0,272	0,211	0,1830
40	0,927	0,048	0,044	0,076	0,060	0,0444
100	0,255	0,020	0,018	0,034	0,021	0,0181
df = 5	Mean	Med	Huber	MVS	RAD	TwoStep
10	0,150	0,144	0,124	0,151	0,136	0,1245
40	0,041	0,042	0,033	0,039	0,035	0,0341
100	0,016	0,017	0,013	0,016	0,014	0,0136
df = 100	Mean	Med	Huber	MVS	RAD	TwoStep
10	0,092	0,126	0,098	0,130	0,109	0,0967
40	0,024	0,037	0,026	0,035	0,029	0,0255
100	0,010	0,015	0,011	0,014	0,012	0,0106



Выводы

1. Если отсутствуют априорные соображения о выборе модели, то в задаче оценки параметра положения рекомендуется использовать **оценку Хубера или двухэтапную оценку**. Они превосходят или не уступают в эффективности другим оценкам в большинстве случаев.
2. **Оценки Хампеля и Мешалкина-Шурыгина** рекомендуется использовать в моделях распределений с очень тяжелыми хвостами типа распределения Коши.
3. Как и непараметрические методы статистики, робастные методы оценивания являются методами математической статистики, пригодными для работы с большими данными в условиях изменчивости их распределений и наличия выбросов.



Контрольные вопросы и задания

- 1) Что такое свойство робастности статистических оценок и почему оно важно при обработке больших данных?
- 2) Положить количество испытаний 10000 в методе Монте Карло, сгенерировать выборки размера $n=100$ из следующих распределений: стандартного нормального $N(0,1)$, распределения Коши $C(0,1)$ и смеси $0.9N(0,1)+0.1C(0,1)$.
- 3) Получить средние и дисперсии по методу Монте-Карло для выборочного среднего, выборочной медианы, оценки Хубера и двухэтапной оценки.
- 4) Провести анализ полученных результатов и сделать выводы.

