САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Кафедра прикладной математики

Робастные оценки параметра масштаба и коэффициента корреляции нормального распределения

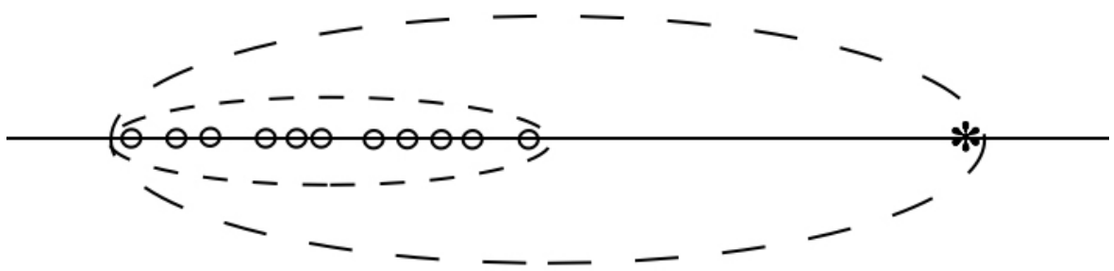
Диссертация на соискание степени магистра

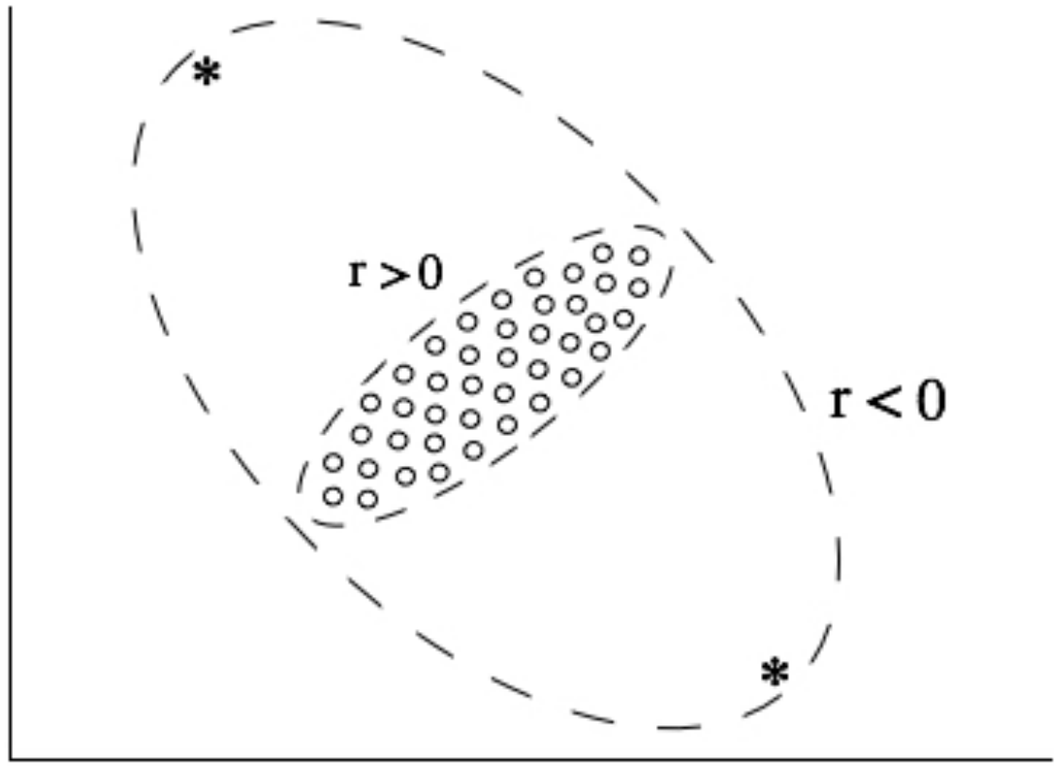
Выполнил студент группы 63601/3 Советников М.О.

Научный руководитель, д.ф.-м.н., проф. Шевляков Г.Л.

2015 г.

Робастность (от лат. *robust:* крепкий, сильный, твёрдый) — **устойчивость** статистических выводов к отклонениям от принятых моделей распределений данных, в частности, к грубым ошибкам





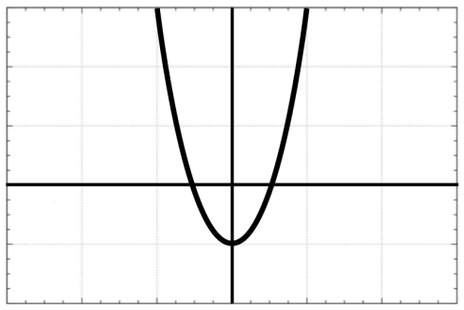
$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

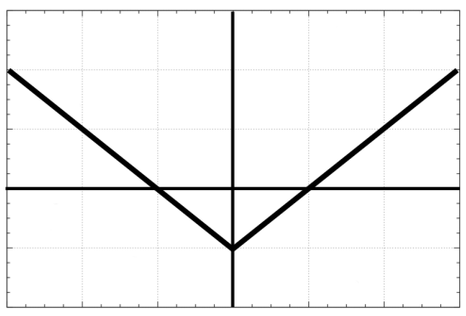
Оценивание **параметра масштаба**
(параметр положения считается известным и равным нулю)

$$x_1, \ldots, x_n \sim f(x;\theta) = \frac{1}{\theta}\, g\left(\frac{x}{\theta}\right), \quad \theta \text{ – параметр масштаба}$$

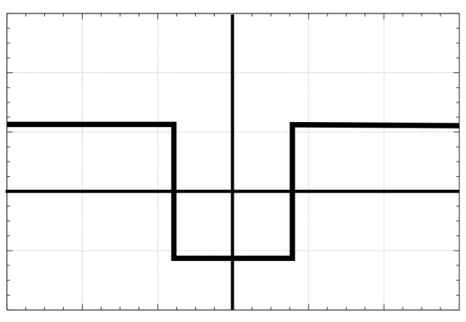
М-оценки: $$\sum_{i=1}^{n} \psi\left(\frac{x_i}{\widehat{\theta}}\right) = 0, \quad \psi(z) \text{ – оценочная функция}$$



$$\psi(z) = z^2 - 1 \quad \Longrightarrow \quad \widehat{\theta}_S = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$



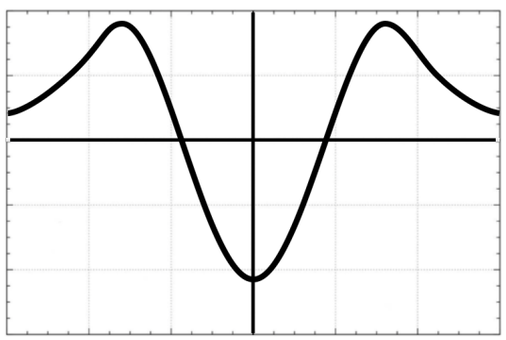
$$\psi(z) = |z| - 1 \quad \Longrightarrow \quad \widehat{\theta}_d = \frac{1}{n}\sum_{i=1}^{n} |x_i|$$



$$\psi(z) = \operatorname{sgn}(|z| - 1) \quad \Longrightarrow \quad \widehat{\theta}_{MAD} = \operatorname{med}|x_i|$$

Оценка **максимальной устойчивости** *MS*

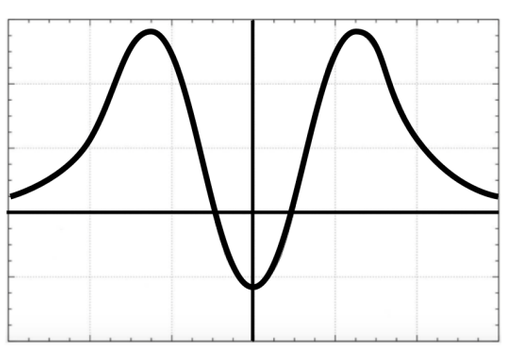
$$\psi(x,\theta) = c\left(\frac{\partial}{\partial\theta}f(x,\theta) + \beta f(x,\theta)\right), \quad c = const,\ \beta : \mathsf{M}\psi(x,\theta) = 0$$



$$\psi_{MS}(z) = [z^2 - 2/3]e^{-z^2/2} \quad \Longrightarrow \quad \widehat{\theta}_{MS}$$

Радикальная оценка *RAD*

$$\psi(x,\theta) = c\left(\frac{\partial}{\partial\theta}\ln f(x,\theta) + \beta\right)\sqrt{f(x,\theta)}, \quad c = const,\ \beta : \mathsf{M}\psi(x,\theta) = 0$$



$$\psi_{RAD}(z) = [z^2 - 2/3]e^{-z^2/4} \quad \Longrightarrow \quad \widehat{\theta}_{RAD}$$

$$f(x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right] \qquad \sigma \text{ – параметр масштаба}$$

Модель больших ошибок Тьюки: смесь распределений с плотностью

$$(1-\varepsilon) f(x,\sigma) + \varepsilon f(x, 3\sigma), \quad \varepsilon = 0.1$$

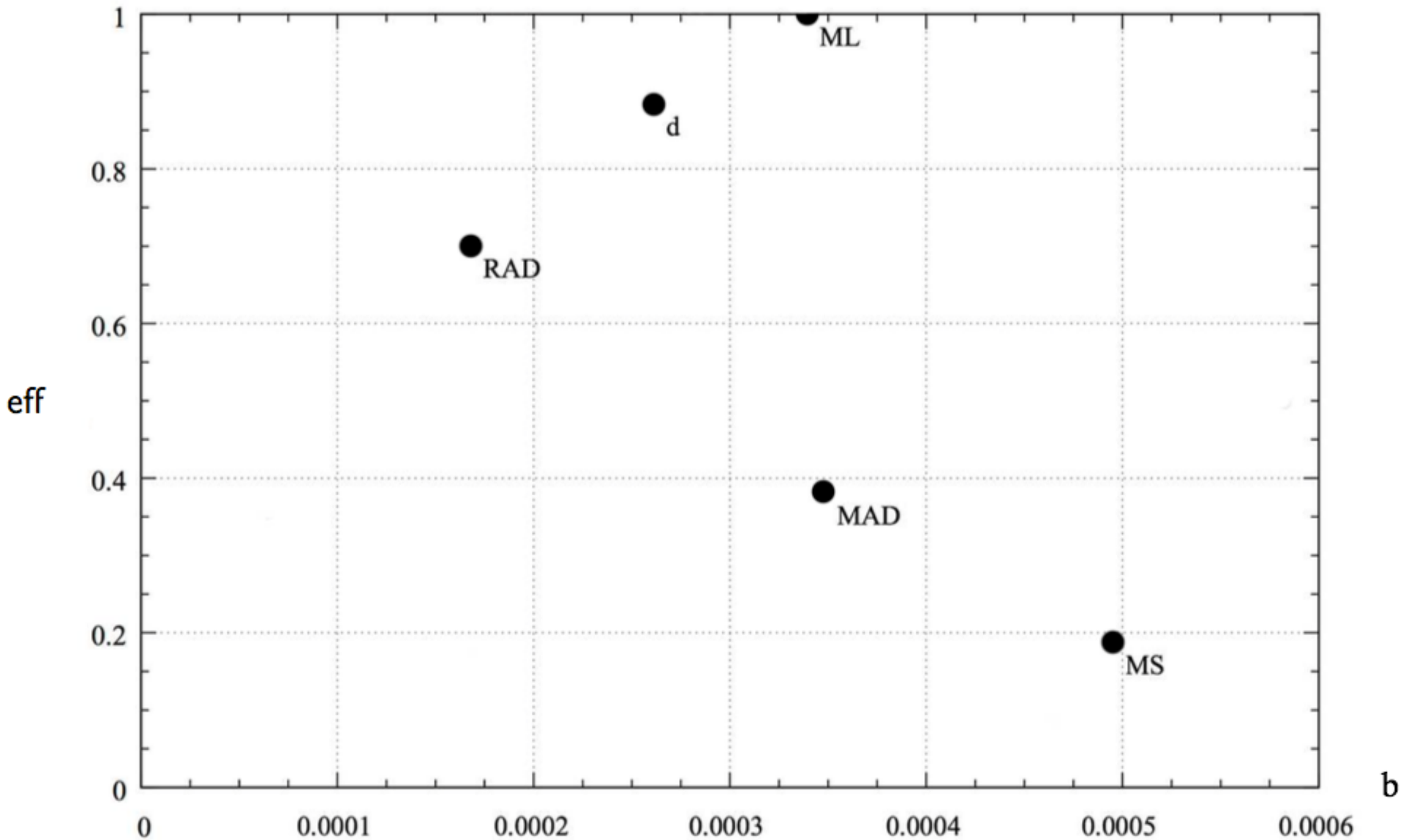
Эффективность оценок

$$\mathsf{eff}(\widehat{\sigma}) = \frac{\mathsf{D}(\widehat{\sigma}_{ML})}{\mathsf{D}(\widehat{\sigma})} \qquad \mathsf{D}(\widehat{\sigma}) = \frac{\int \psi^2(x,\sigma) f(x;\sigma) dx}{[\int \frac{\partial}{\partial\sigma}\psi(x,\sigma) f(x;\sigma) dx]^2}$$

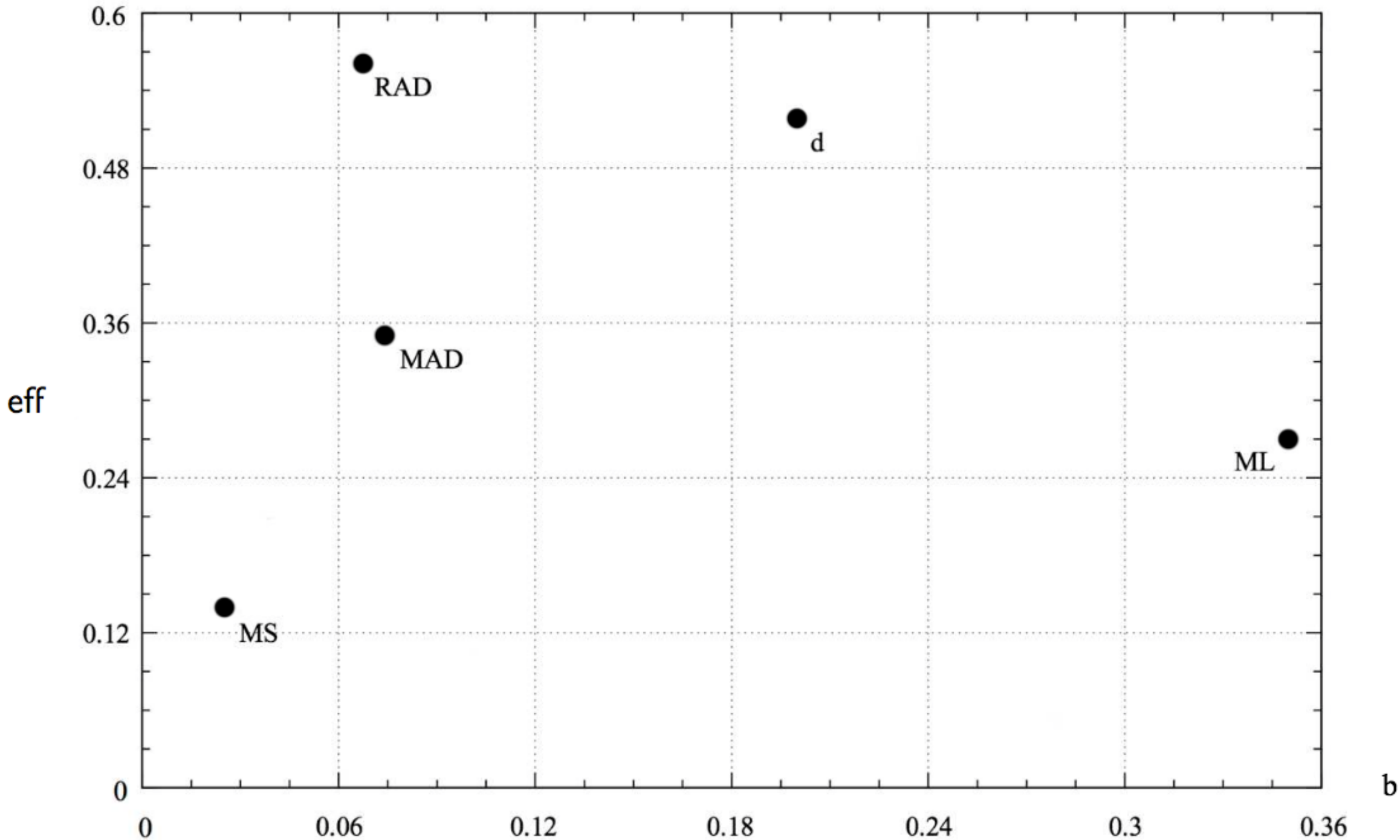
Сравнительный анализ классических и робастных оценок параметра масштаба на конечных выборках $(n = 20, 40, 100, 1000)$
и в асимптотике при $n \to \infty$

$$\widehat{\sigma}_{ML}, \quad \widehat{\sigma}_{MAD}, \quad \widehat{\sigma}_{d}, \quad \widehat{\sigma}_{RAD}, \quad \widehat{\sigma}_{MS}$$

Абсолютное среднее смещение **b** и эффективность **eff** оценок параметра масштаба нормального распределения при $n = 1000$



Абсолютное среднее смещение **b** и эффективность **eff** оценок при 10%-ном загрязнении в модели больших ошибок Тьюки; $n = 1000$



$$(x_1, y_1), \ldots, (x_n, y_n) \sim f(x, y; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\},$$

ρ – коэффициент корреляции

$$\widehat{\rho}_r = \sum_{i=1}^{n} x_i y_i \Bigg/ \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

$$\widehat{\rho}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\widehat{\rho}_k}{1+\widehat{\rho}_k^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\right) \quad \Longrightarrow \quad \widehat{\rho}_{ML}$$

Алгоритм *MCD* (*Minimum Covariance Determinant*) $\Longrightarrow$ $\widehat{\rho}_{MCD}$

$$\psi_{RAD}(x,y,\rho) = c\left[\frac{\partial}{\partial\rho}\ln f(x,y,\rho) + \frac{\rho}{3(\rho^2-1)}\right]\sqrt{f(x,y,\rho)} \quad \Longrightarrow \quad \widehat{\rho}_{RAD}$$

Сравнительный анализ классических и робастных оценок коэффициента корреляции на конечных выборках $(n = 20, 40, 100, 1000)$
и в асимптотике при $n \to \infty$

$$\widehat{\rho}_r, \quad \widehat{\rho}_{ML}, \quad \widehat{\rho}_{MCD}, \quad \widehat{\rho}_{RAD}$$

$$\mathsf{D}(\widehat{\rho}_{RAD}) = \frac{81(9+10\rho^2)(1-\rho^2)^2}{512(1+\rho^2)^2}$$

$$\mathsf{eff}(\widehat{\rho}_{RAD}) = \frac{\mathsf{D}(\widehat{\rho}_{ML})}{\mathsf{D}(\widehat{\rho}_{RAD})} = \frac{512(1+\rho^2)}{81(9+10\rho^2)}$$

Эффективность оценок коэффициента корреляции двумерного нормального распределения; $\rho = 0$

n	$\widehat{\rho}_{ML}$	$\widehat{\rho}_{r}$	$\widehat{\rho}_{MCD}$	$\widehat{\rho}_{RAD}$
20	0.858	**0.948**	0.284	0.658
40	0.925	**0.980**	0.309	0.673
100	0.966	**0.990**	0.343	0.686
1000	0.995	**0.998**	0.400	0.703
∞	1.000	1.000	—	0.702

Эффективность оценок при 10%-ном внешнем противоположно коррелированном загрязнении в модели больших ошибок; $\rho = 0.9$

n	$\widehat{\rho}_{ML}$	$\widehat{\rho}_{r}$	$\widehat{\rho}_{MCD}$	$\widehat{\rho}_{RAD}$
20	0.036	0.012	0.127	**0.625**
40	0.023	0.010	0.235	**0.886**
100	0.010	0.009	0.340	**0.961**
1000	0.004	0.008	0.457	**0.974**

Эффективность оценок коэффициента корреляции двумерного распределения Коши; $\rho = 0$

n	$\widehat{\rho}_{ML}$	$\widehat{\rho}_{r}$	$\widehat{\rho}_{MCD}$	$\widehat{\rho}_{RAD}$
20	0.000	0.130	0.269	**0.401**
40	0.000	0.068	0.268	**0.351**
100	0.000	0.028	0.265	**0.352**
1000	0.000	0.003	0.265	**0.375**

Эффективность оценок коэффициента корреляции двумерного распределения Коши; $\rho = 0.9$

n	$\widehat{\rho}_{ML}$	$\widehat{\rho}_{r}$	$\widehat{\rho}_{MCD}$	$\widehat{\rho}_{RAD}$
20	0.000	0.020	0.102	**0.239**
40	0.000	0.011	0.171	**0.278**
100	0.000	0.004	0.227	**0.304**
1000	0.000	0.000	0.258	**0.322**

$$\vec{x}_1, \ldots, \vec{x}_n \sim \frac{1}{\sqrt{(2\pi)^p |C|}} \exp\left\{ -\frac{1}{2} (\vec{x})^\intercal C^{-1} (\vec{x}) \right\}$$

C – ковариационная матрица

Алгоритм *MCD* (*Minimum Covariance Determinant*) $\Longrightarrow$ $\widehat{C}_{MCD}$

$$\widehat{C}_{RAD} = \lim_{k\to\infty} \widehat{C}_k : \quad \widehat{C}_{k+1} = (3/2) \frac{\sum \vec{x}_i \vec{x}_i^{\,\intercal} \exp(-q_i/4)}{\sum \exp(-q_i/4)}, \quad q_i = \vec{x}_i^{\,\intercal} \widehat{C}_k^{-1} \vec{x}_i$$

$$\widehat{R} = D^{-1} \widehat{C} D^{-1}, \quad D = \sqrt{\operatorname{diag} \widehat{C}}$$

Сравнительный анализ классических и робастных оценок корреляционной матрицы на конечных выборках $(n = 20, 40, 100, 1000)$

$$\widehat{R}_{MCD}, \quad \widehat{R}_{RAD}$$

Дисперсия невязки $\delta = ||R^{-1}\widehat{R} - E||_F$ оценок корреляционной матрица трехмерного нормального распределения

n	$\widehat{R}_{MCD}$	$\widehat{R}_{RAD}$
20	1.653	0.768
40	0.339	0.130
100	0.065	0.033
1000	0.003	0.002

Дисперсия невязки $\delta = ||R^{-1}\widehat{R} - E||_F$ оценок при 10%-ном внутреннем загрязнении трехмерного нормального распределения в модели больших ошибок Тьюки

n	$\widehat{R}_{MCD}$	$\widehat{R}_{RAD}$
20	3.029	1.800
40	0.888	0.329
100	0.194	0.091
1000	0.015	0.012

Радикальная оценка *RAD* коэффициента корреляции:
- На сильно загрязненных данных **превосходит по эффективности** все современные классические и робастные оценки
- По скорости вычисления уступает некоторым робастным оценкам

Радикальная оценка *RAD* корреляционной матрицы:
- На сильно загрязненных данных **превосходит по эффективности** все современные классические и робастные оценки на малых выборках
- По скорости вычисления уступает некоторым робастным оценкам