Robust statistics

Tomáš Pevný

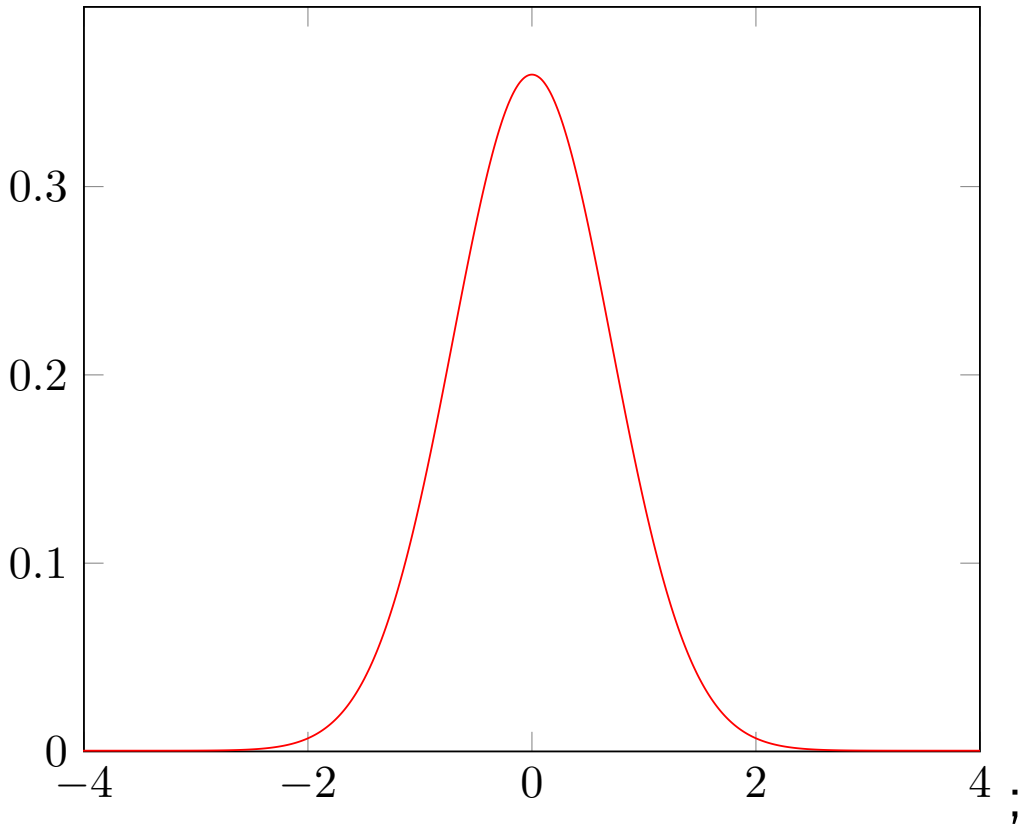
November 25, 2024

Goals of robust statistics

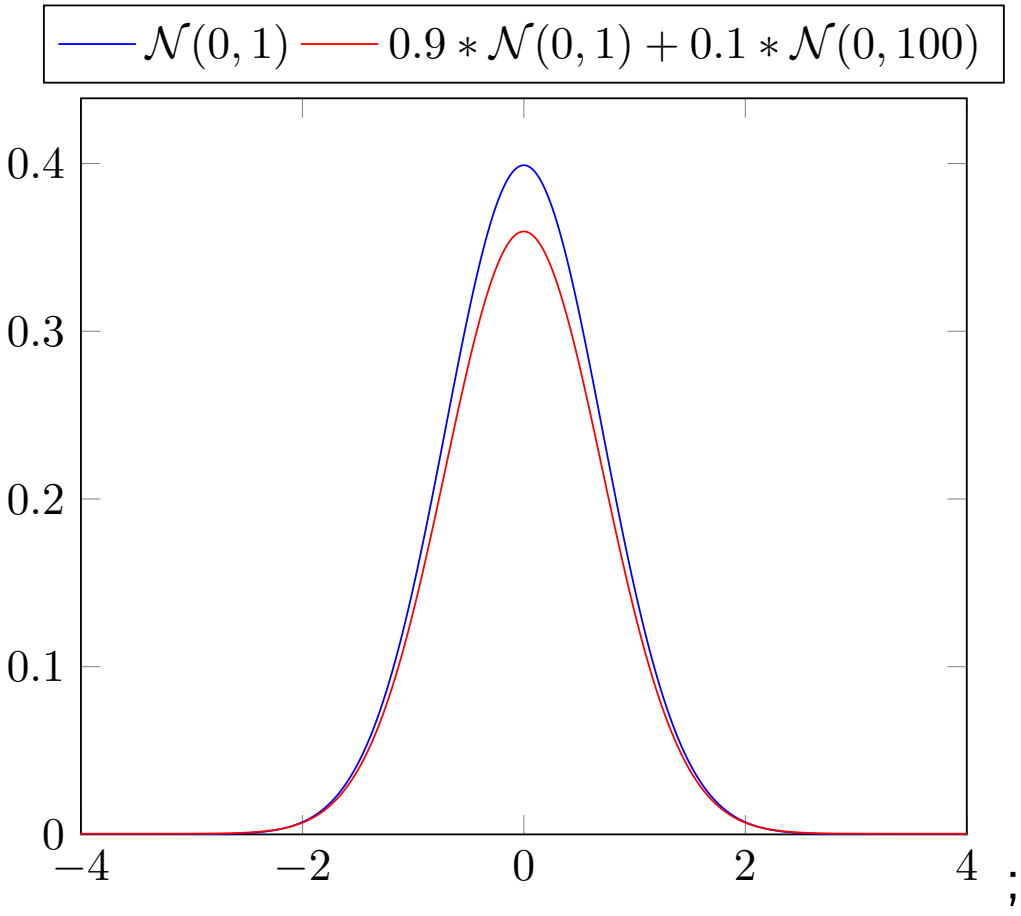
It should not be affected by

- the presence of outliers (even malicious)
- or in-correctness of assumed probability distribution.

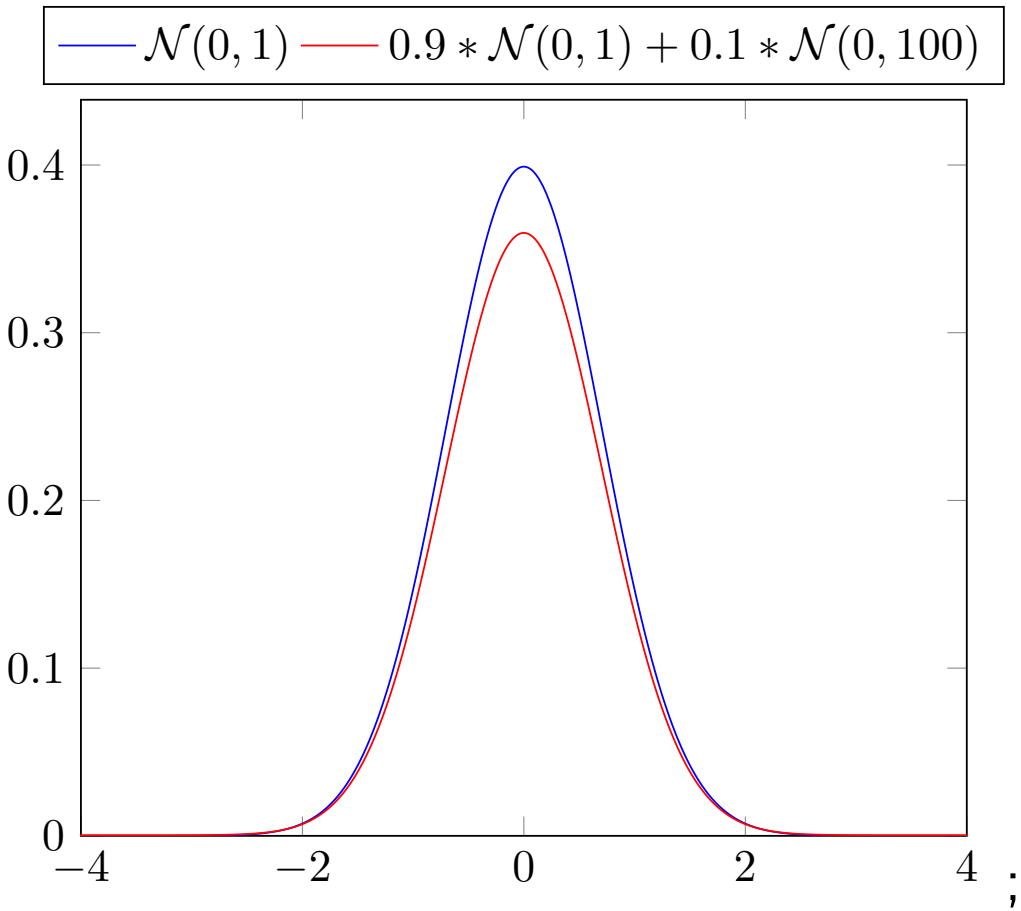
Which distribution is this?



Motivation


$\mathcal{N}(0,1)$
$0.9 * \mathcal{N}(0,1) + 0.1 * \mathcal{N}(0,100)$
0.4
0.3
0.2
0.1
0
−4
−2
0
2
4
;

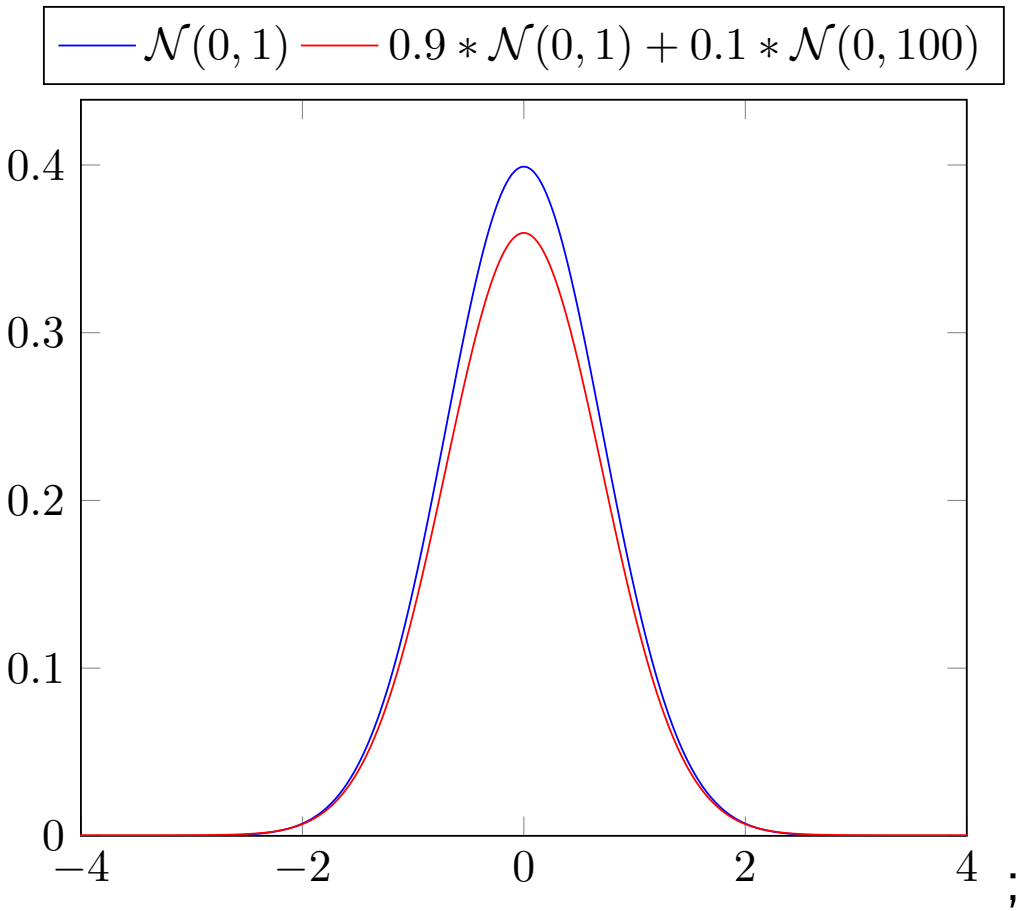
Motivation



;

mean estimated from 1000 samples: $-5 \cdot 10^{-3}$, 0.49

Motivation



;

median estimated from 1000 samples: -0.012, -0.013

Plan

Breakdown Point

Breakdown Point: the largest proportion of sample observations which may be given arbitrary values without taking the estimator to a limit uninformative about the parameter being estimated.

Example: Breakdown point

Breakdown point of

- mean is 0,
- median is 50%.

Gross error sensitivity

Influence function

$$\mathrm{IF}(x|p,\eta) = \lim_{\varepsilon \to 0} \frac{\eta((1-\varepsilon)p + \varepsilon\delta_x) - \eta(p)}{\varepsilon}$$

p — probability distribution
η — estimator
δ_x — dirac function at x

Gross error sensitivity

Influence function

$$\mathrm{IF}(x|p,\eta) = \lim_{\varepsilon \to 0} \frac{\eta((1-\varepsilon)p + \varepsilon\delta_x) - \eta(p)}{\varepsilon}$$

Gross error sensitivity

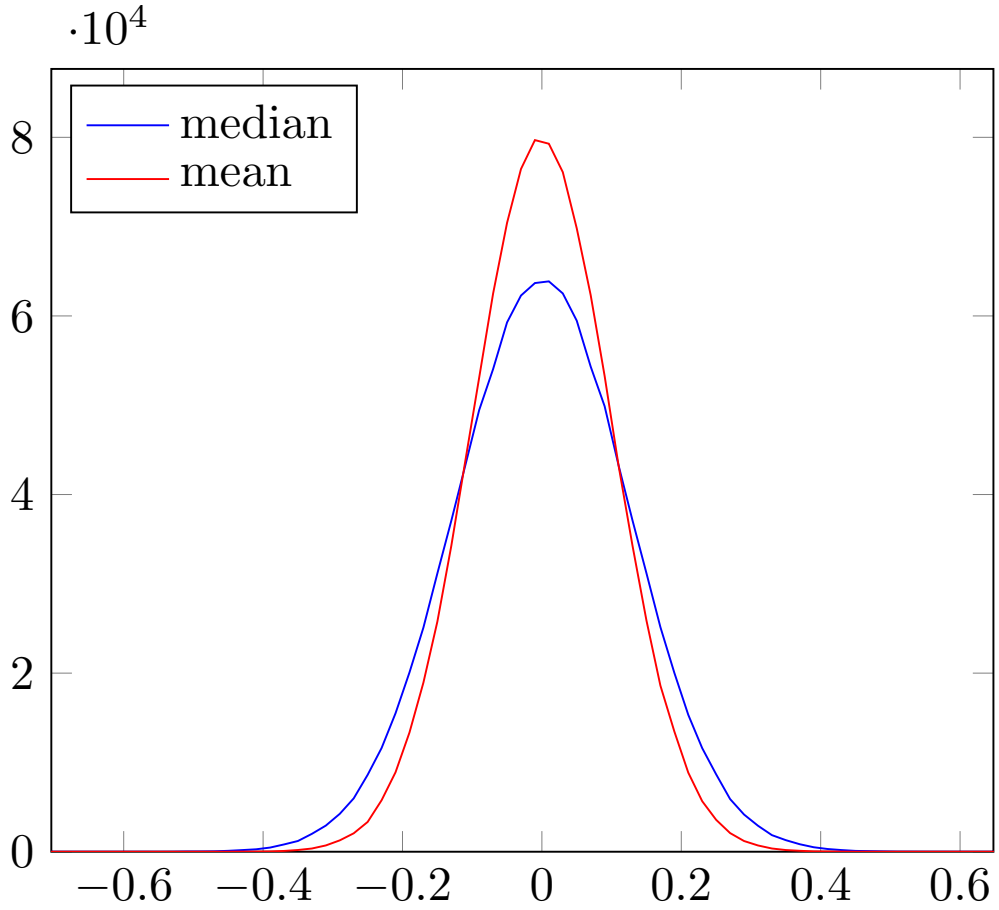
$$\mathrm{GES}(p,\eta) = \sup_x |\mathrm{IF}(x)|$$

p — probability distribution
η — estimator
δ_x — dirac function at x

How to measure efficiency

How is the sampling distribution of the estimator spread about the true value?

How to measure efficiency



Distribution of mean and median estimates from 100 samples from $\mathcal{N}(0, 1)$.

Asymptotic relative efficiency

Asymptotic relative efficiency (ARE) is defined as

$$\mathrm{ARE}(\hat{\eta}_1, \hat{\eta}_2, p) = \frac{V_2}{V_1},$$

where $\frac{V_1}{n}$, $\frac{V_2}{n}$ are variances of estimators $\hat{\eta}_1$, $\hat{\eta}_2$ of a parameter μ of probability distribution p.

Example of comparison of efficiency:

Assuming x_i are i.i.d samples from $N(\mu, \sigma)$

- $\bar{X}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_1}\right)$
- $\mathrm{Med}(X_{n_2}) = \mathrm{med}\{x_1, \ldots, x_{n_2}\} \sim \mathcal{N}\left(\mu, \frac{1}{4p^2(\mu)n_2}\right)$

Example of comparison of efficiency:

Assuming x_i are i.i.d samples from $N(\mu, \sigma)$

- $\bar{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_1}\right)$
- $\mathrm{Med}(X_{n_2}) = \mathrm{med}\{x_1, \ldots, x_{n_2}\} \sim \mathcal{N}\left(\mu, \frac{1}{4p^2(\mu)n_2}\right)$
- Median and mean estimates are equally precise, iff

$$n_1 = \sigma_p^2 4p^2(\mu)n_2$$

- ARE of median to mean for $\mathcal{N}(0,1)$ is $\mathrm{ARE}(\mathrm{Med}, \bar{X}) = \frac{2}{\pi} = 0.6366.$

Plan

Estimators of location

- mean
- median
- q%-trimmed
- q%-winsorized
- Hodges-Lehmann

Mean

$$\{-39.61, -26.29, -1.07, -0.92, -0.85, -0.16, 0.93, 1.91, 2.18, 133.65\}$$

- mean $\frac{1}{n}\sum_i x_i = 6.97$
- Zero breakdown
- Optimal if samples follows Normal distribution.

Median

{-39.61, -26.29, -1.07, -0.92, -0.85, -0.16, 0.93, 1.91, 2.18, 133.65}

- median $\text{median}\{x_1, \ldots, x_{10}\} = -0.51$
- 50% breakdown
- ARE = 0.637 for Normal distribution

q%-trimmed

$$\{-39.61, -26.29, -1.07, -0.92, -0.85, -0.16, 0.93, 1.91, 2.18, 133.65\}$$

- calculate mean from samples $\left\{x | x_{q\%} \leq x \leq x_{1-q\%}\right\}$
- mean $\frac{1}{|\mathscr{X}_q|} \sum_{X \in \mathscr{X}_q} x_i = -0.41$
- q% breakdown
- ARE = 0.943 for Normal distribution

q%-Windsorized

$\{-1.07, -1.07, -1.07, -0.92, -0.85, -0.16, 0.93, 1.91, 1.91, 1.91\}$

- replace samples outside $\langle x_{q\%}, x_{1-q\%} \rangle$ by bounds, return mean.
- mean $\frac{1}{N} \sum_{X \in \tilde{\mathscr{X}}_q} x_i = 0.33$
- q% breakdown
- Robust alternative to mean, more dependant on the distribution then median.

Hodges-Lehman

$$\{-39.61, -26.29, -1.07, -0.92, -0.85, -0.16, 0.93, 1.91, 2.18, 133.65\}$$

- $\mathrm{HL} = \mathrm{med}\left\{\frac{x_i + x_j}{2} | i, j \in N\right\} = -0.03$
- 0.29 breakdown
- ARE = 0.955 for Normal distribution

Comparison of location estimators

estimator	location
mean	6.97
median	−0.51
q%-trimmed	−0.41
q%-winsorized	0.33
Hodges-Lehmann	−0.03

Plan

Estimators of scale

- sample standard deviation
- median absolute deviation
- S_n
- Q

Sample standard deviation

- (unbiassed) formula: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- (biassed) formula: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- breakdown point 0
- ARE=1 — optimal for Normal distribution

$\bar{x}$ denotes mean of samples

Median absolute deviation

- formula: $\mathrm{MAD} = \mathrm{med}\{|x_i - \mathrm{med}\{x_i\}|\}$
- breakdown point 50%
- For Normal distribution
  - ARE=0.37
  - $\hat{\sigma} = 1.4826 \cdot \mathrm{MAD}$

S_n

- formula: $S_n = \text{med}_i\{\text{med}_j|x_i - x_j|\}$
- breakdown point 29%
- For Normal distribution
  - ARE=0.86
  - $\hat{\sigma} = 1.0483 \cdot S_n$

Q

- sample standard deviation $Q = \{|x_i - x_j| | i < j\}_{q_{25}}$
- breakdown point 50%
- For Normal distribution
  - ARE=0.82
  - $\hat{\sigma} = 2.2219 \cdot Q$

Plan

Question

Why mean is so popular?

ML estimate of location of Normal distribution

Assuming $x \sim \mathcal{N}(\mu, \sigma^2)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

ML estimate of location of Normal distribution

Assuming $x \sim \mathcal{N}(\mu, \sigma^2)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

ML estimate of location of Normal distribution

Assuming $x \sim \mathcal{N}(\mu, \sigma^2)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$\arg\max_{\mu} \log L = -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 - \log \sqrt{2\pi}\sigma$$

ML estimate of location of Normal distribution

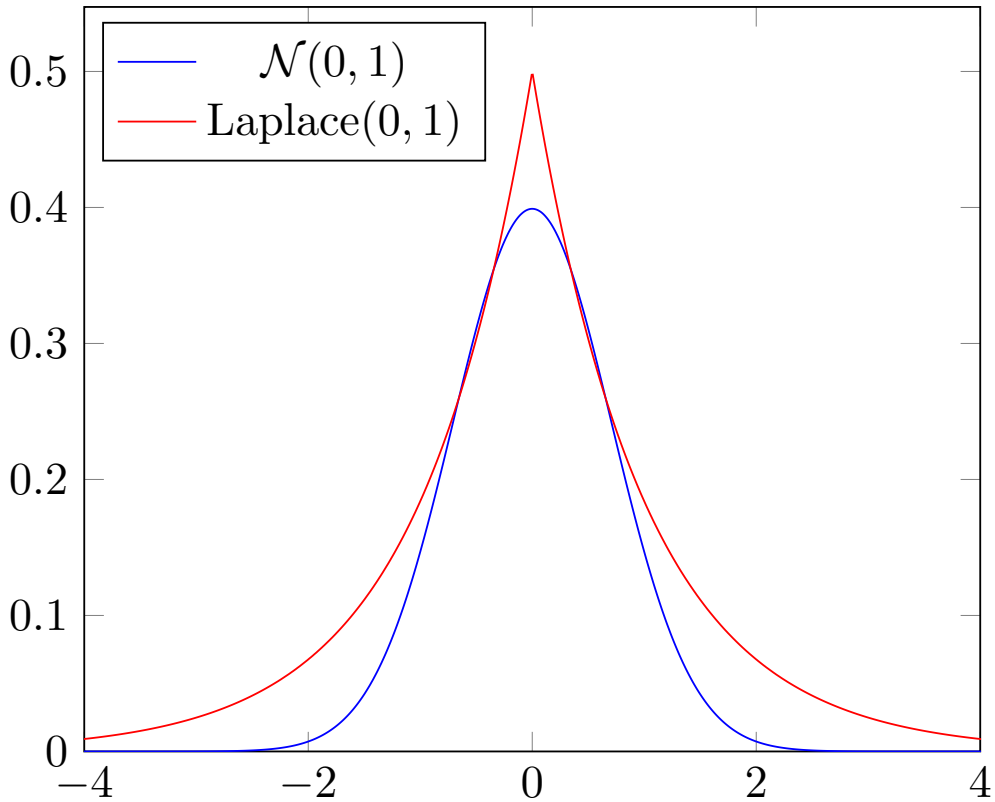
Assuming $x \sim \mathcal{N}(\mu, \sigma^2)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$\arg\max_{\mu} \log L = -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 - \log \sqrt{2\pi}\sigma$$

$$\mu = \frac{1}{n} \sum_i x_i$$

ML estimate of location of Laplace distribution



ML estimate of location of Laplace distribution

Assuming $x \sim \text{Laplace}(\mu, \sigma)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

ML estimate of location of Laplace distribution

Assuming $x \sim \mathrm{Laplace}(\mu, \sigma)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|x_i - \mu|}$$

ML estimate of location of Laplace distribution

Assuming $x \sim \mathrm{Laplace}(\mu, \sigma)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|x_i - \mu|}$$

$$\arg\max_{\mu} \log \mathscr{L} = -\frac{1}{\sigma} \sum_i |x_i - \mu| - \log 2\sigma$$

ML estimate of location of Laplace distribution

Assuming $x \sim \text{Laplace}(\mu, \sigma)$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|x_i - \mu|}$$

$$\arg\max_{\mu} \log \mathscr{L} = -\frac{1}{\sigma} \sum_i |x_i - \mu| - \log 2\sigma$$

$$0 = \sum_i \text{sgn}(x_i - \mu)$$

Question

Can we generalize this?

ML estimate of location of exponential class

Assuming $x \sim \frac{1}{Z} e^{-\rho(\frac{x-\mu}{\sigma})}$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

ML estimate of location of exponential class

Assuming $x \sim \frac{1}{Z} e^{-\rho(\frac{x-\mu}{\sigma})}$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{Z} e^{-\rho(\frac{x_i-\mu}{\sigma})}$$

ML estimate of location of exponential class

Assuming $x \sim \frac{1}{Z} e^{-\rho(\frac{x-\mu}{\sigma})}$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{Z} e^{-\rho(\frac{x_i-\mu}{\sigma})}$$

$$\arg\max_{\mu} \log \mathscr{L} = -\sum_i \rho(\frac{x_i-\mu}{\sigma}) - \log Z$$

ML estimate of location of exponential class

Assuming $x \sim \frac{1}{Z} e^{-\rho(\frac{x-\mu}{\sigma})}$, then *maximum likelihood estimate* (ML) of μ from $\{x_1, \ldots, x_n\}$ is

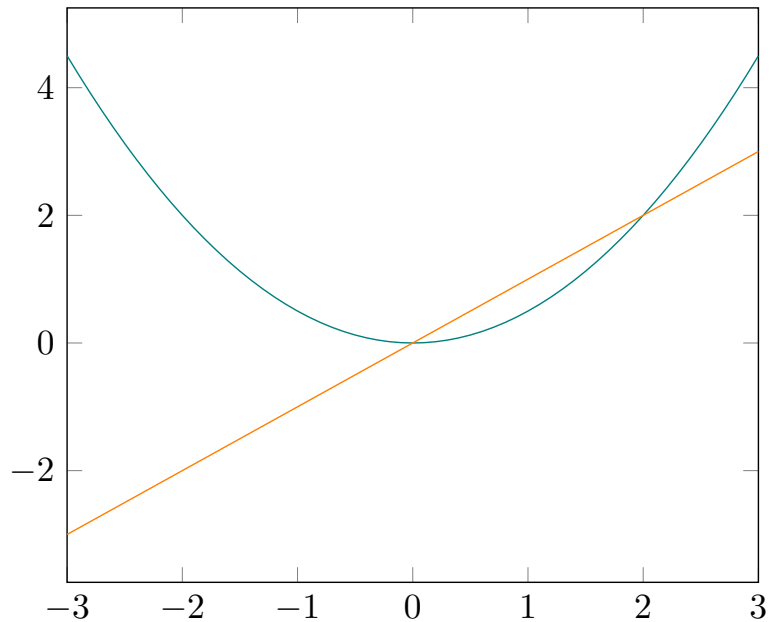
$$\arg\max_{\mu} \mathscr{L} = \prod_i \frac{1}{Z} e^{-\rho(\frac{x_i-\mu}{\sigma})}$$

$$\arg\max_{\mu} \log \mathscr{L} = -\sum_i \rho(\frac{x_i-\mu}{\sigma}) - \log Z$$

$$0 = \sum_i \rho'(\frac{x_i-\mu}{\sigma}).$$

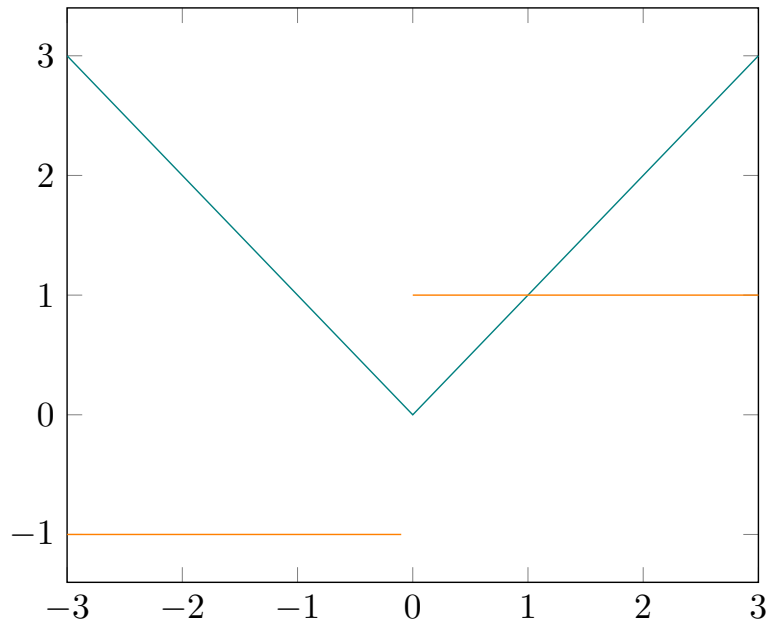
Normal distribution

- $\rho = \frac{1}{2}x^2$
- $\rho' = x$
- $\mu = \frac{1}{n}\sum_i x_i$



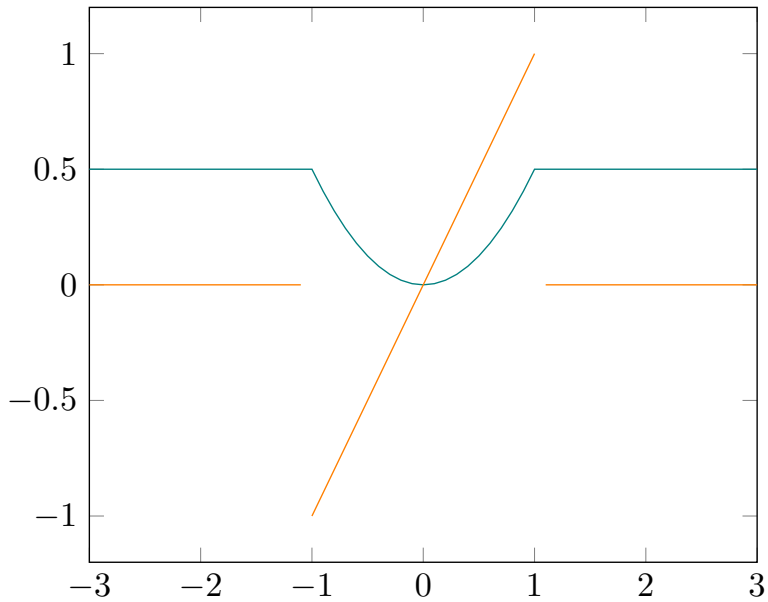
Laplace distribution

- $\rho = \frac{1}{2}|x|$
- $\rho' = \operatorname{sgn}(x)$
- $0 = \sum_i \operatorname{sgn}(x_i - \mu)$



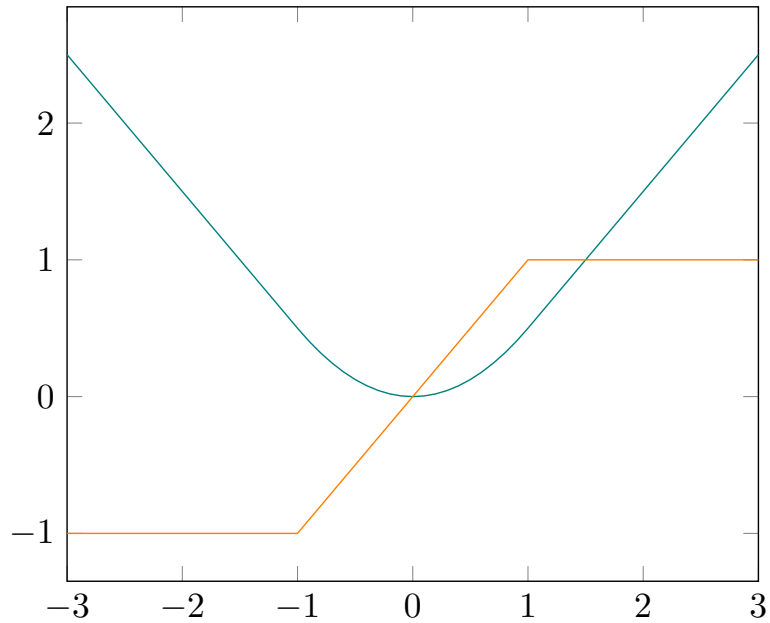
Huber loss (not called Huber)

- $\rho = \begin{cases} \frac{x^2}{2} & |x| < a \\ \frac{a^2}{2} & |x| \geq a \end{cases}$
- $\rho' = \begin{cases} x & |x| < a \\ 0 & |x| \geq a \end{cases}$
- $\mu = \frac{1}{n_{<a}} \sum_{i|abs(x_i)<a} x_i$
- trimming



Huber loss (called Huber)

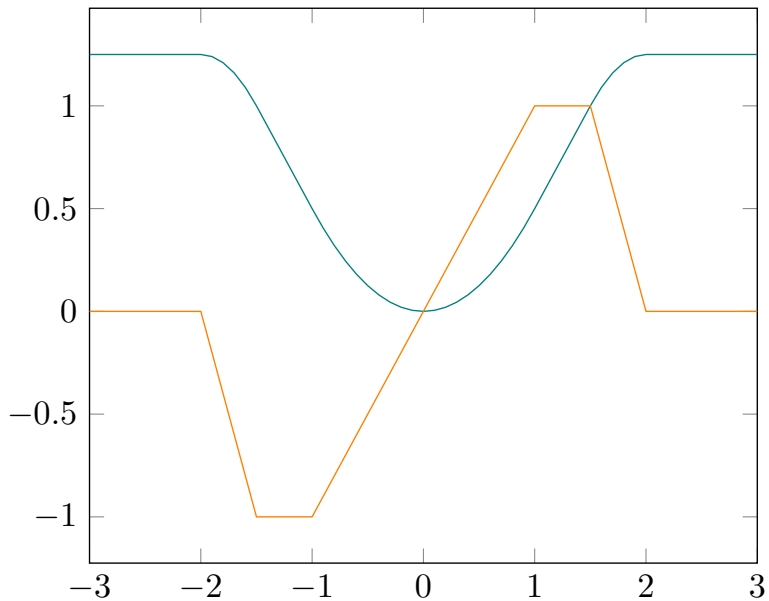
- $\rho = \begin{cases} \frac{x^2}{2} & |x| < a \\ a|x| - \frac{a^2}{2} & |x| \geq a \end{cases}$
- $\rho' = \begin{cases} x & |x| < a \\ a \cdot \mathrm{sgn}(x) & |x| \geq a \end{cases}$
- $\mu = \frac{1}{n}\left[\sum_{i|abs(x_i)<a} x_i + n_{>a} \cdot a\right]$
- Windsorizing



Hampel loss

$$\rho = \begin{cases} \frac{x^2}{2} & 0 \le x < a \\ ax - \frac{a^2}{2} & a \le x < b \\ \frac{a(x-c)^2}{2(b-c)} + \frac{1}{2}a(b+c-a) & b \le x < c \\ \frac{1}{2}a(b+c-a) & c \le x \end{cases}$$

$$\rho' = \begin{cases} x & 0 \le x < a \\ a & a \le x < b \\ \frac{a(x-c)}{b-c} & b \le x < c \\ 0 & c \le x \end{cases}$$



Caveats of robust losses

- To use them you need to set a scale — use robust estimate.
- Robust losses might have unfavourable efficiency.
- Sometimes you need to select parameters.
- Hampel loss is not convex — difficult to optimize.

Plan

Least-square regression is an M-estimator

Generative model behind OLS is

$$y = x^{\mathrm{T}}\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

$x, y \in \mathbb{R}^d$

Least-square regression is an M-estimator

Generative model behind OLS is

$$y = x^{\mathrm{T}}\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Therefore

$$p(y|x, \beta, \sigma^2) \sim \mathcal{N}(x^{\mathrm{T}}\beta, \sigma^2)$$

$x, y \in \mathbb{R}^d$

Least-square regression is an M-estimator

Generative model behind OLS is

$$y = x^{\mathrm{T}}\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Therefore

$$p(y|x, \beta, \sigma^2) \sim \mathcal{N}(x^{\mathrm{T}}\beta, \sigma^2)$$

and

$$\hat{\beta} = \arg\min_{\beta} \sum_i (x_i^{\mathrm{T}}\beta - y_i)^2.$$

$x, y \in \mathbb{R}^d$

Robust regression

Assume different distribution of noise

$$y = x^{\mathrm{T}}\beta + \varepsilon, \varepsilon \sim \mathrm{Laplace}(0, \sigma).$$

Robust regression

Assume different distribution of noise

$$y = x^{\mathrm{T}}\beta + \varepsilon, \varepsilon \sim \mathrm{Laplace}(0, \sigma).$$

and obtain median absolute regression

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} |x_i^{\mathrm{T}}\beta - y_i|.$$

Robust regression

Replace the mean estimate by robust alternatives

- Huber A: least median of squares (LMS)

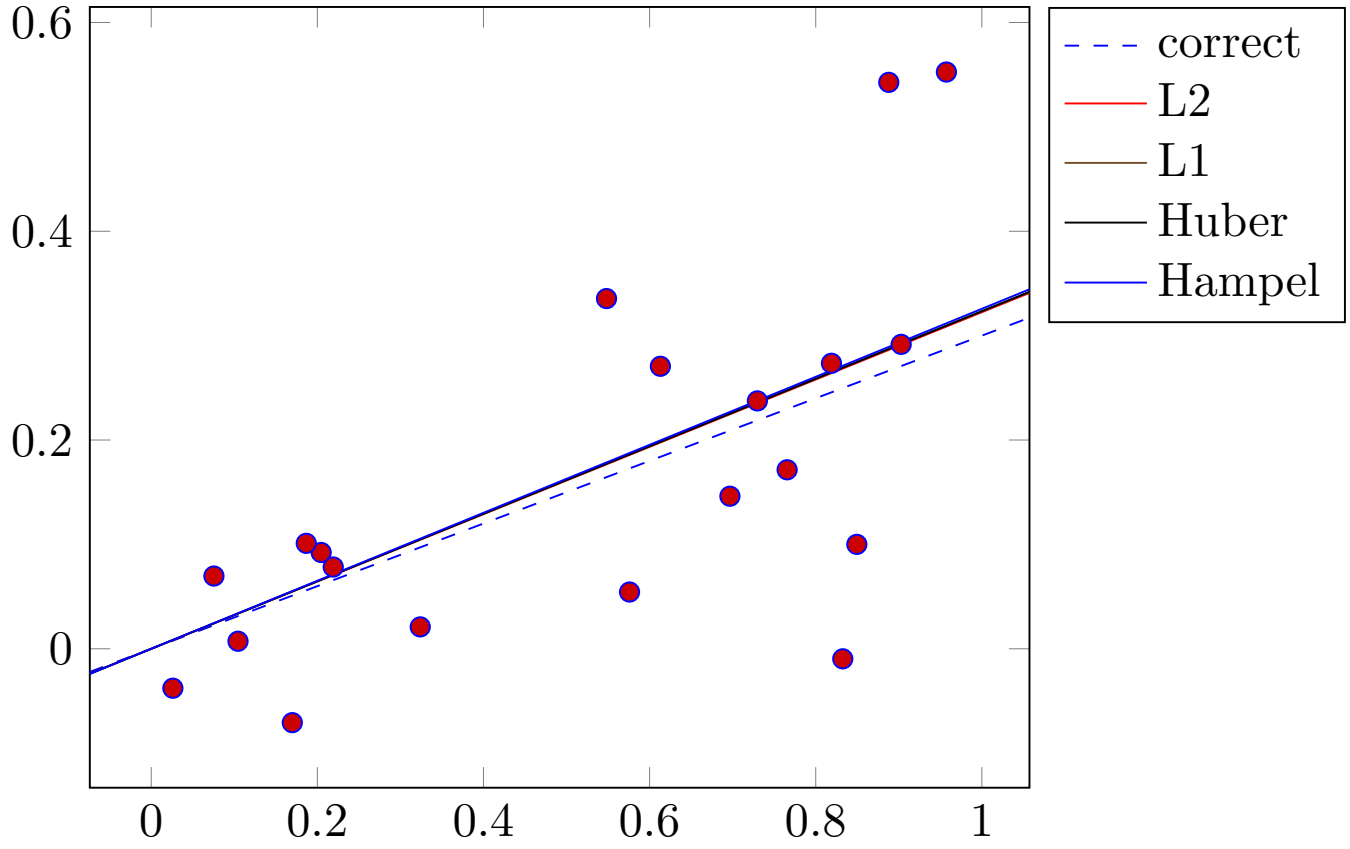
$$\hat{\beta} = \arg\min_{\beta} \operatorname{med} \left\{ (x_i^{\mathrm{T}} \beta - y_i)^2 \right\}$$

- Huber B: least trimmed squares (LTS)

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (x_i^{\mathrm{T}} \beta - y_i)^2_{(j)}$$

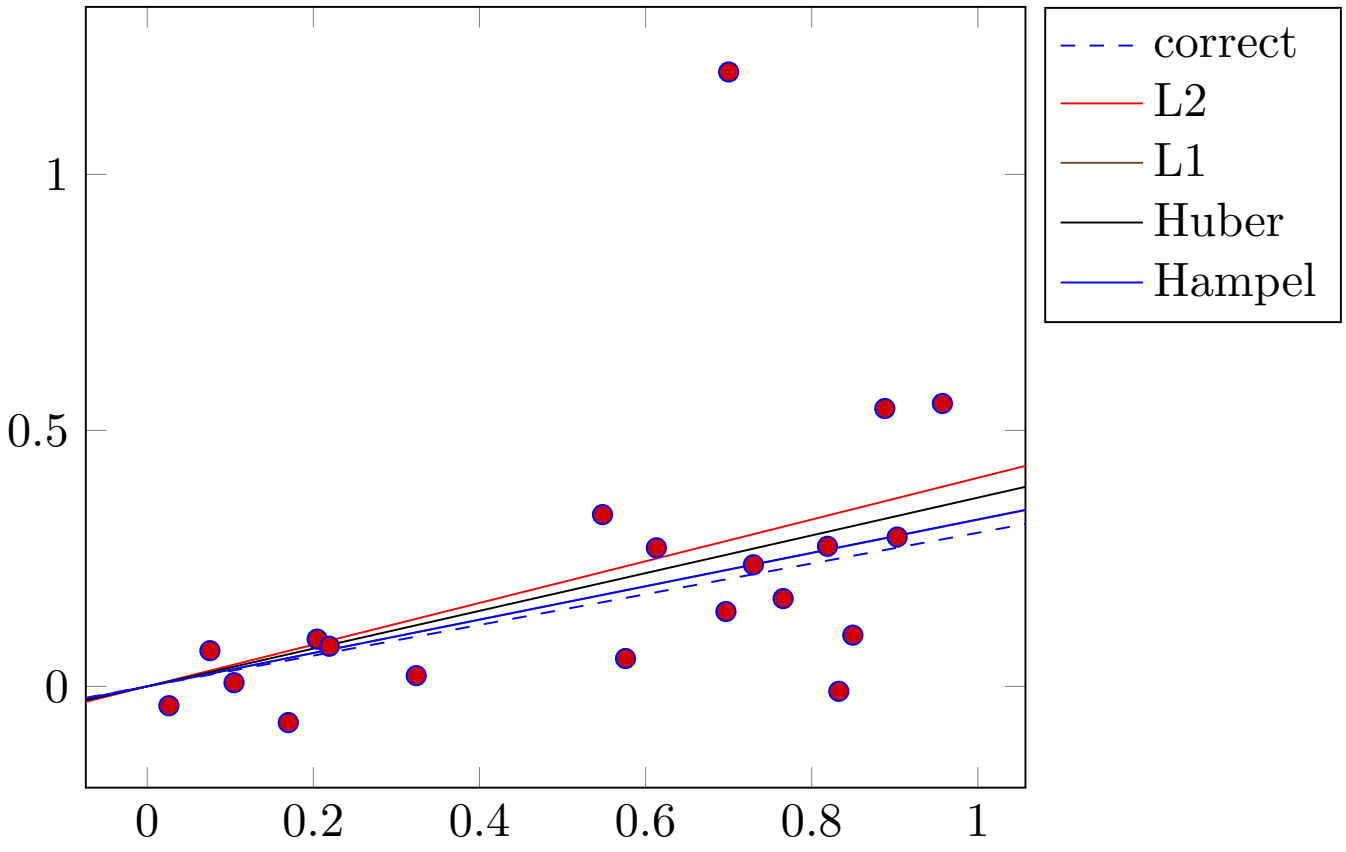
$\cdot_{(j)}$ represents an order statistics (quantile)

Examples of robust regression



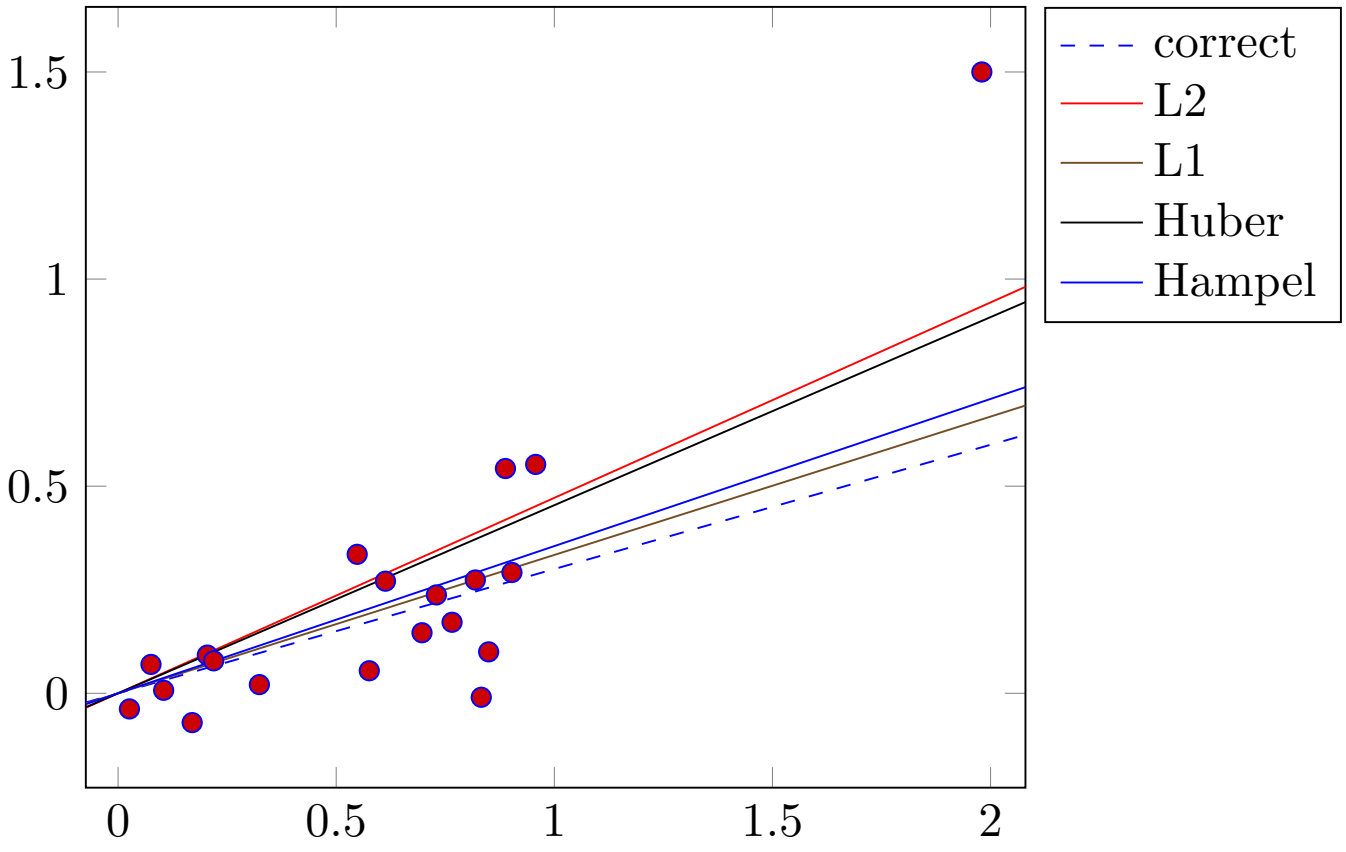
$y = 0.1x + \mathcal{N}(0, 0.1)$ without outliers

Examples of robust regression



$y = 0.1x + \mathcal{N}(0, 0.1)$ with outlier

Examples of robust regression

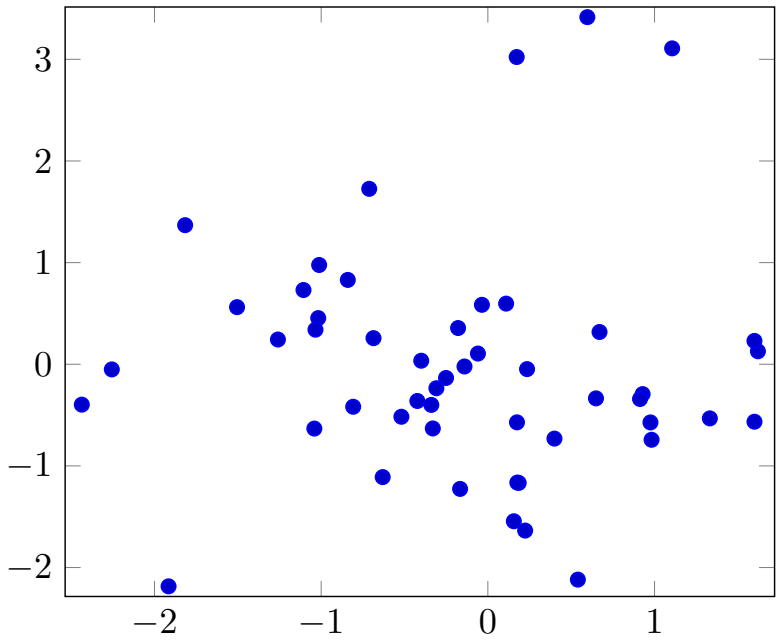


$y = 0.1x + \mathcal{N}(0, 0.1)$ with outliers

Plan

Pearson's correlation

- Assume pairs of samples $\{(x_i, y_i)\}_{i=1}^{n}$
- $\rho_{X,Y} = \frac{\frac{1}{n}\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_X \sigma_Y}$

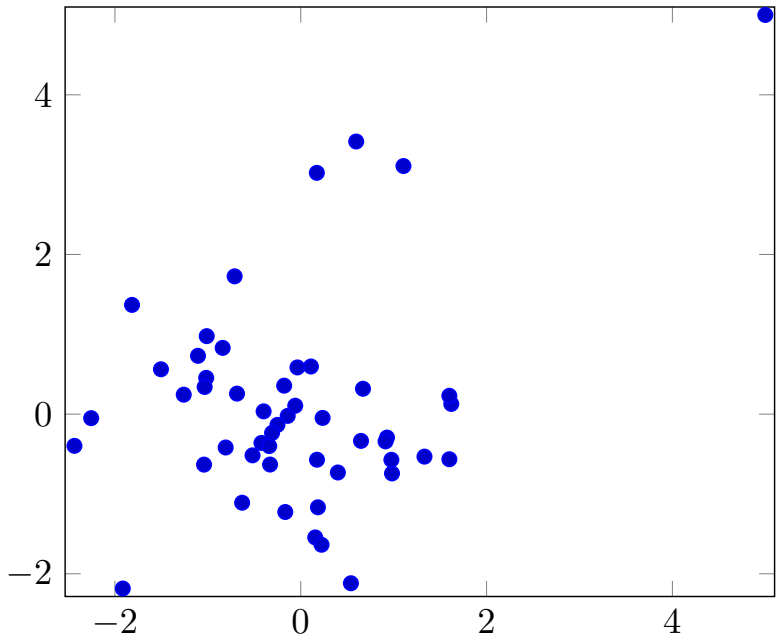


$\rho = -0.0008$

$\bar{x}$, $\bar{y}$ denotes a sample mean

Pearson's correlation

- Assume pairs of samples $\{(x_i, y_i)\}_{i=1}^{n}$
- $\rho_{X,Y} = \frac{\frac{1}{n}\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_X \sigma_Y}$
- Breakdown point is zero

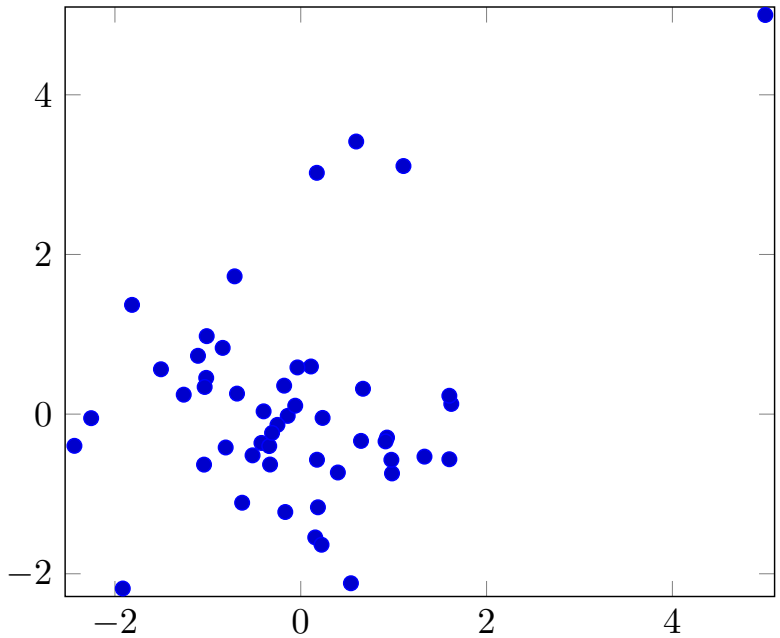


$\rho = 0.3245$

$\bar{x}$, $\bar{y}$ denotes a sample mean

Spearman's correlation

- Replaces $\{x_i, y_i\}_i$ by ranks $\{r_i^x, r_i^y\}_i$
- $r_s = \frac{\frac{1}{n}\sum_i [(r_i^x - \bar{r}_x)(r_i^y - \bar{r}_y)]}{\sigma_{r_x}\sigma_{r_y}}$

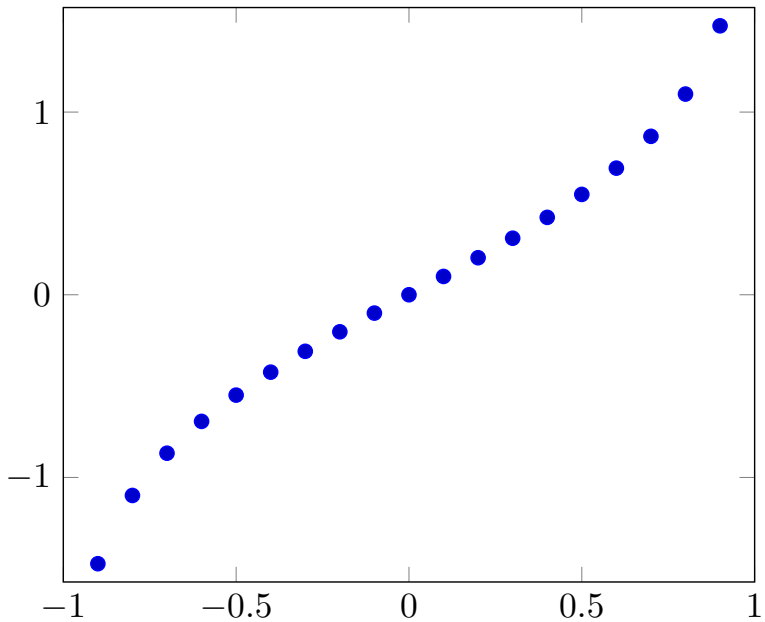


$r_s = -0.0008$, $\rho = 0.3245$

$\bar{r}_x$, $\bar{r}_y$ denotes a sample mean of ranks

Spearman's correlation

- Replaces $\{x_i, y_i\}_i$ by their ranks $\{r_i^x, r_i^y\}_i$
- $r_s = \frac{\frac{1}{n}\sum_i [(r_i^x - \bar{r}_x)(r_i^y - \bar{r}_y)]}{\sigma_{r_x}\sigma_{r_y}}$
- Statistic $r_s\sqrt{\frac{n-2}{1-r^2}}$ follows Student-t

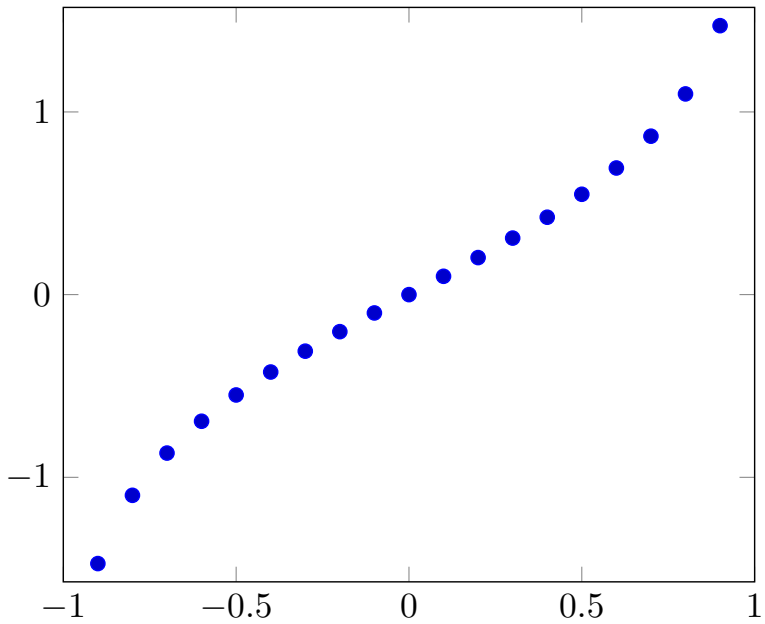


$r_s = 1.0,\ \rho = 0.98$

$\bar{r}_x$, $\bar{r}_y$ denotes a sample mean of ranks

Kendall correlation

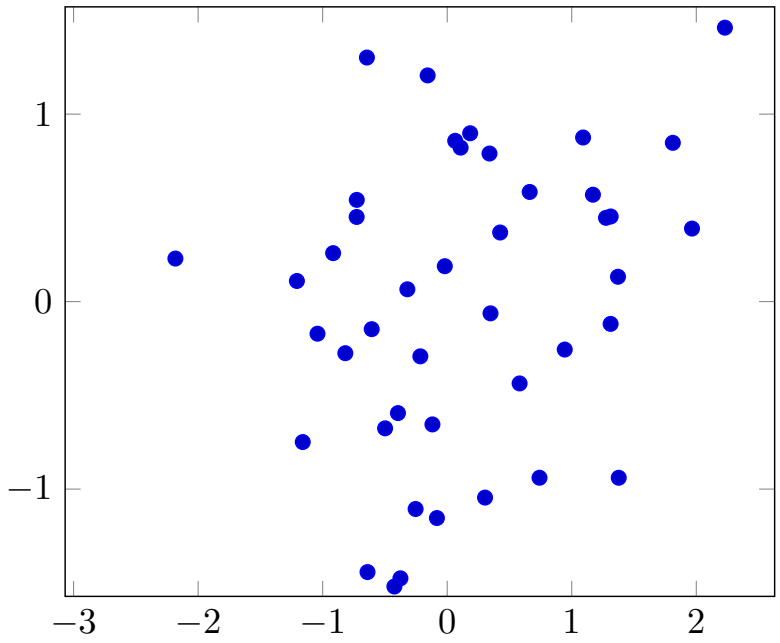
- Kendalls'τ removes all quantities and uses order
- Samples are concordant if
  - $x_i < x_j$ and $y_i < y_j$
  - $x_i > x_j$ and $y_i > y_j$
- $r_k = \frac{1}{\binom{n}{2}}(n_c - n_d)$
- $\tau \sim \mathcal{N}\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$



$r_k = 1.0,\ r_s = 1.0,\ \rho = 0.98$

Kendall correlation

- Kendalls'τ removes all quantities and uses order
- Samples are concordant if
  - $x_i < x_j$ and $y_i < y_j$
  - $x_i > x_j$ and $y_i > y_j$
- $r_k = \frac{1}{\binom{n}{2}}(n_c - n_d)$
- $\tau \sim \mathcal{N}\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$



$r_k = 0.34,\ r_s = 0.48,\ \rho = 0.55$

Plan

Sign test

Tests if differences between pairs of observations are consistent.

Sign test

Population of pairs $\{(x_i, y_i)\}_i$

1. discard samples for which $|y_i - x_i| = 0$
2. test statistic
$$W = \sum_{i=1}^{N_r} I(y_i > x_i)$$
3. under null hypothesis W follows binomial distribution $\mathrm{Bi}(N, 0.5)$

$I(\cdot)$ denotes the indicator function

Wilcoxon-signed rank test

Tests if population of two related (matched) samples have equal mean rank.

Test hypothesis of Wilcoxon-signed rank test

Difference between pairs follows a symmetric distribution around zero.

Wilcoxon-signed rank test

Population of pairs $\{(x_i, y_i)\}_i$

1. calculate $|y_i - x_i|$ and discard those with $|y_i - x_i| = 0$
2. rank remaining samples according to $|y_i - x_i|$
3. test statistic
$$W = \sum_{i=1}^{N} [\mathrm{sgn}(y_i - x_i) \cdot R_i]$$
4. under null hypothesis W has
   - zero mean
   - variance $\sigma_w^2 = \frac{N(N+1)(2N+1)}{6}$
5. For small N critical values are tabulated.
6. For large N with $z = \frac{W}{\sigma_W}, \sigma_W$

Discussion of sign and signed-rank test

- Sign test have less assumptions — needs only order relationship
- Signed rank test have higher power: ARE is 0.67.
- Would differences follows normal distribution, paired t-test is more appropriate; ARE is 0.95.
- Generalization of a sign test to n-tuples is a Friedman test.

Mann-Whitney U-test

Tests, whether a probability that a value from population X is greater than a value from population Y (and vice versa) is greater than 0.5.

Tests, whether the distributions of both populations are equal.

Mann-Whitney U-test

1. Assume we have $\{(x_i)\}_{i=1}^{n_1}$
2. Calculate ranks of all samples together.
3. Sum ranks of samples from the first population, R_1.
4. Sum ranks of samples from the second population, R_2.
5. Calculate $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$.
6. $U = \min\{U_1, U_2\}$
7. For small n_1, n_2 critical values are tabulated,
   for large n_1, n_2 $U \sim \mathcal{N}\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$.