K-sign depth

Erik Mendroš, Marek Bedřich

PMSE - Matematická statistika
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

10.5.2024

Definition

K-sign depth

Let $K \geq 2$ be a natural number and $u_1, \ldots, u_N$ a sequence of residuals. Then their K-sign depth, or shortly K-depth, is defined as

$$d_K(u_1, \ldots, u_N) = \frac{1}{\binom{N}{K}} \sum_{1 \leq n_1 < \ldots < n_K \leq N} \left(\prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k u_{n_k} > 0\right\} + \prod_{k=1}^{K} \mathbb{1}\left\{(-1)^k u_{n_k} < 0\right\} \right)$$

i.e. the relative number of K-element subsets with alternating signs.

Assumptions

In this presentation, we will be working with a standard linear regression model

$$Y_i = \mathbf{X}_i^T \beta_X + \varepsilon_i, \quad i = 1, \ldots, N, \quad \beta_X \in \mathbb{R}^p,$$

although K-depth can be extended to any models of the form

$$Y_i = g(\mathbf{X}_i, \beta_X) + \varepsilon_i$$

where ε_i are mutually independent, absolutely continuous, and satisfy

$$\mathrm{Med}\left[\varepsilon_i\right] = 0, \quad i = 1, \ldots, N.$$

Basic properties

For a chosen value of β, denote the residuals

$$U_i(\beta) = Y_i - \mathbf{X}_i^T \beta, \quad i = 1, \dots, N.$$

Then the expectation of the K-depth is given by

$$\mathbb{E}\, d_K(U_1(\beta_X), \dots, U_N(\beta_X)) = \left(\frac{1}{2}\right)^{K-1}$$

and it holds that

$$d_K(U_1(\beta_X), \dots, U_N(\beta_X)) \xrightarrow[N \to \infty]{a.s.} \left(\frac{1}{2}\right)^{K-1}.$$

Computational complexity

There are multiple ways to calculate K-depth with various degrees of computational complexity. From slowest to fastest:

- Naive method using the definition. – $\Theta(N^K)$
- Approximation method (Muller 2021). – $\Theta(N)$
- Block implementation (Muller 2023). – $\Theta(N)$
- Dynamic programming (Nagy 2024). – $\Theta(N)$

Hypothesis testing

K-depth can be used as an alternative goodness-of-fit measure to test the regression model parameter, i.e. to test

$$H_0 : \beta_X \in \Theta^0.$$

This so-called K-depth test is based on the test statistic

$$T_K(\beta) = T_K(U_1(\beta), \ldots, U_N(\beta)) = N\left(d_K(U_1(\beta), ..., U_N(\beta)) - \left(\frac{1}{2}\right)^{K-1}\right).$$

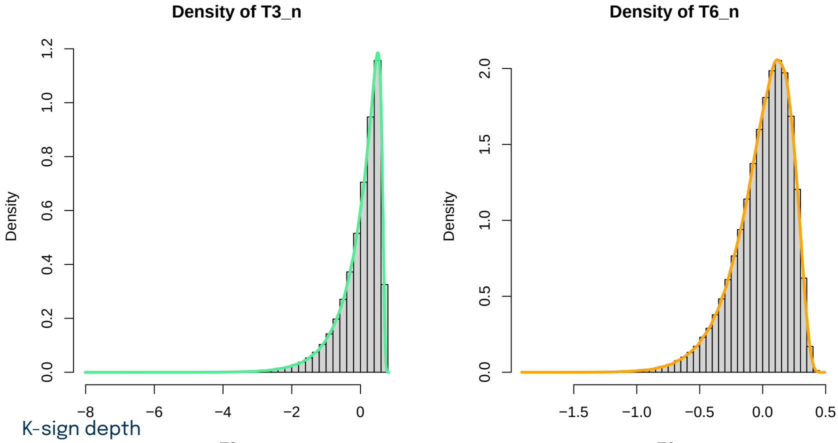
Low values of K-depth indicate that the assumption on the median has been violated, so consequently low values of the test statistic testify against the parameter β.

Limiting distribution

It holds that

$$T_K(\beta_X) \xrightarrow[N \to \infty]{d} \Psi_K(W),$$

where $\Psi_K(W)$ is a special function of a standard Brownian motion.



Goodness of fit testing

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, N$$

where $\varepsilon_i \sim N(0, 1)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $x_1, \ldots, x_N$ represent realized regressors.

Goodness of fit testing

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, N$$

where $\varepsilon_i \sim N(0,1)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $x_1, \dots, x_N$ represent realized regressors.
Denote

$$\hat{\beta}^{LAD} = \operatorname{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{N} |y_i - \beta_0 - \beta_1 x_i| \qquad \hat{\beta}^{LSE} = \operatorname{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Goodness of fit testing

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, N$$

where $\varepsilon_i \sim N(0,1)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $x_1, \dots, x_N$ represent realized regressors.
Denote

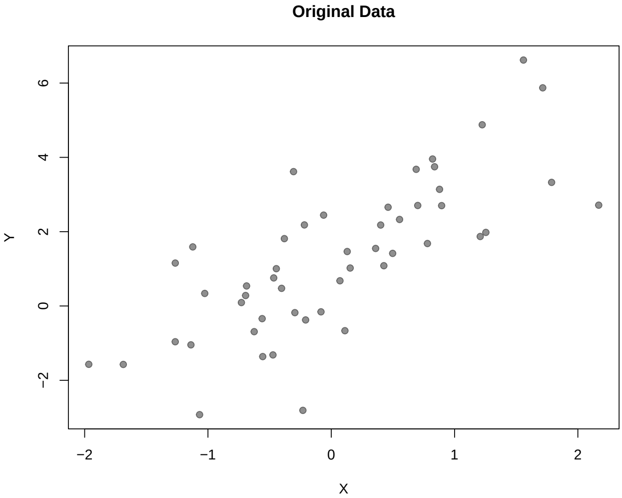
$$\hat{\beta}^{LAD} = \mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{N} |y_i - \beta_0 - \beta_1 x_i| \qquad \hat{\beta}^{LSE} = \mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Do the estimators $\hat{\beta}^{LAD}$ and $\hat{\beta}^{LSE}$ provide a “good” fit?

$$H_0^{LAD} : (\beta_0, \beta_1)^\top = \hat{\beta}^{LAD} \qquad H_0^{LSE} : (\beta_0, \beta_1)^\top = \hat{\beta}^{LSE}$$

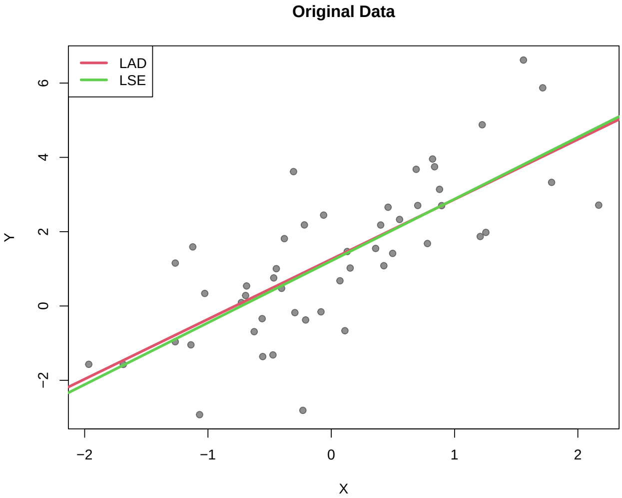
Simulation

Generated dataset of size $N = 50$, with true parameters set to $\beta_0 = 1, \beta_1 = 1.5$.



Simulation

Generated dataset of size $N = 50$, with true parameters set to $\beta_0 = 1, \beta_1 = 1.5$.



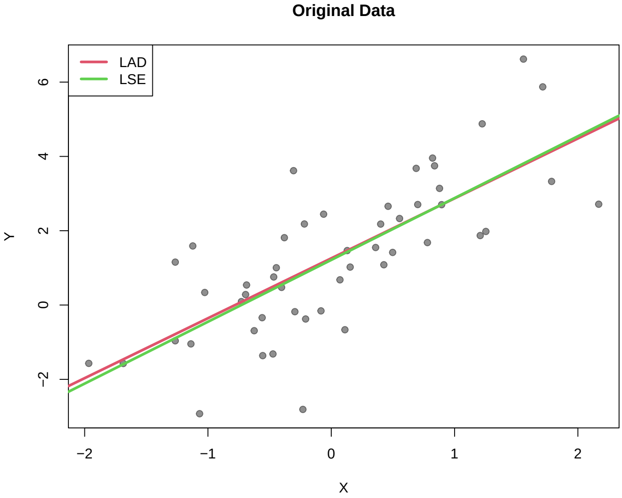
Estimated coefficients and the respective 3-depths

$$\hat{\beta}^{LAD} = (1.25, 1.61)^\top; \quad d_3^{LAD} = 0.27$$

$$\hat{\beta}^{LSE} = (1.21, 1.66)^\top; \quad d_3^{LSE} = 0.26$$

Simulation

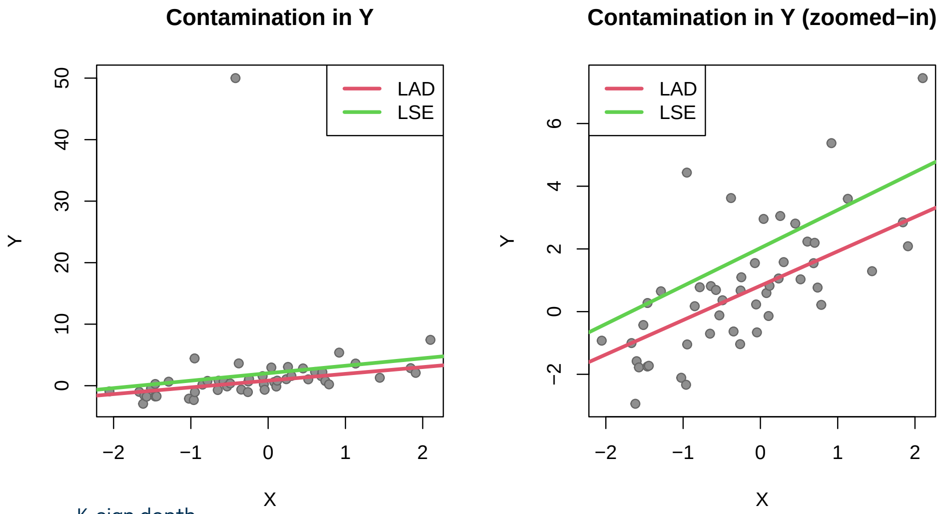
Null hypotheses not rejected for none of $K \in \{3, 4, 5, 6, 7\}$.



p-values	Estimates
p_3^{LAD}	0.8617
p_4^{LAD}	0.8296
p_5^{LAD}	0.7457
p_6^{LAD}	0.7425
p_7^{LAD}	0.7094
p_3^{LSE}	0.9450
p_4^{LSE}	0.8969
p_5^{LSE}	0.8452
p_6^{LSE}	0.9262
p_7^{LSE}	0.8489

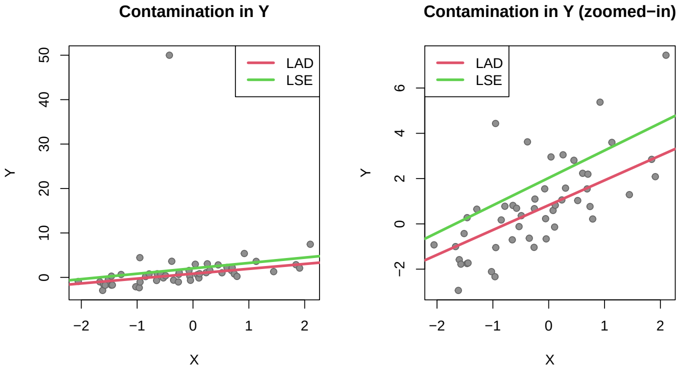
Outliers in dataset

We added one outlying point in terms of Y.



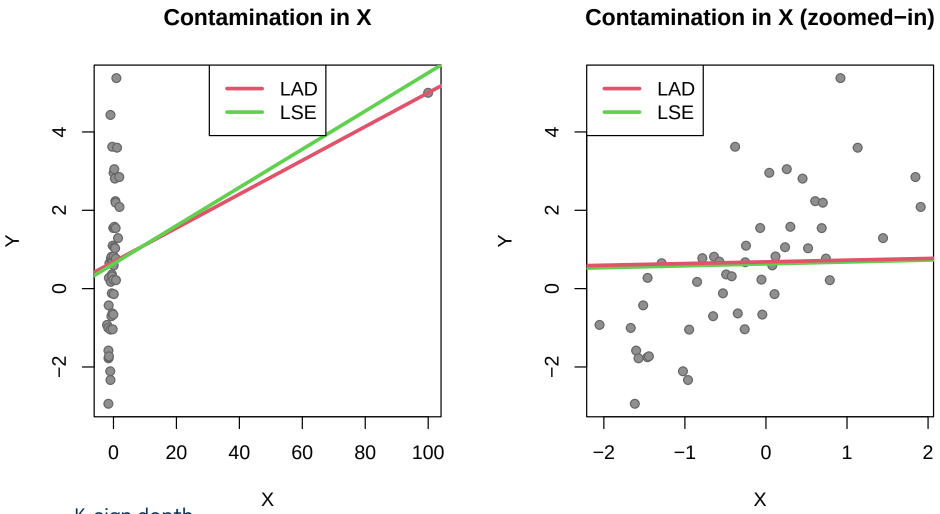
Outliers in dataset

- The estimator $\hat{\beta}^{LAD}$ is **robust** w.r.t. Y outliers, the estimator $\hat{\beta}^{LSE}$ is **not robust**.
- H_0's for $\hat{\beta}^{LAD}$ **not rejected** for all K with p-values ≈ 0.9.
- H_0's for $\hat{\beta}^{LSE}$ **rejected** for all K with p-values ≈ 0.002.



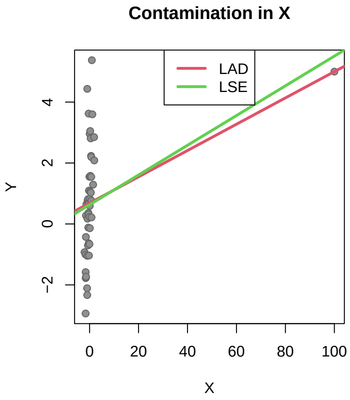
Outliers in dataset

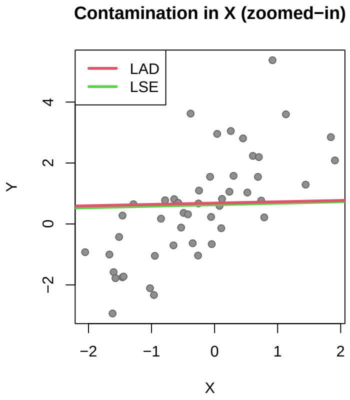
We added one outlying point in terms of X.



Outliers in dataset

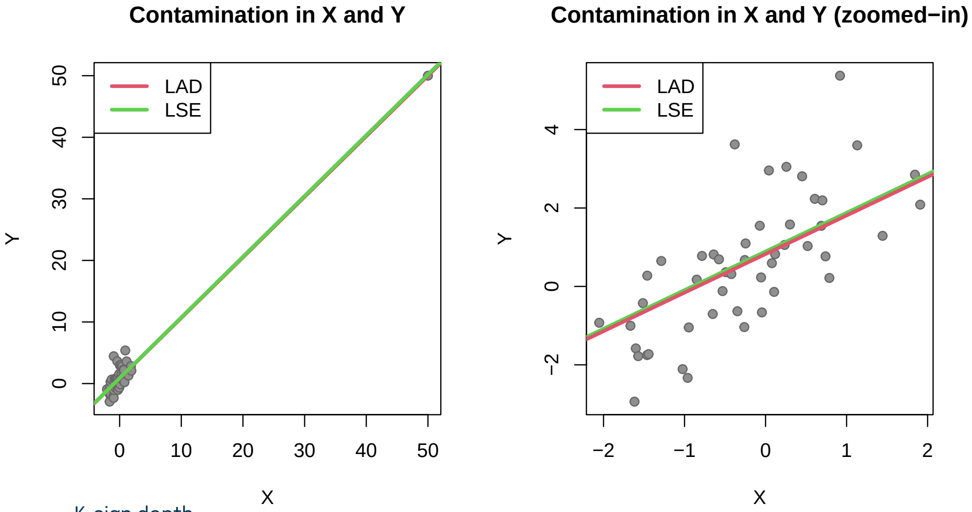
- The estimators $\hat{\beta}^{LAD}$ and $\hat{\beta}^{LSE}$ are **not robust** w.r.t. X outliers.
- H_0's for both $\hat{\beta}^{LAD}$ and $\hat{\beta}^{LSE}$ **rejected** for all K with p-values ≈ 0.00001.





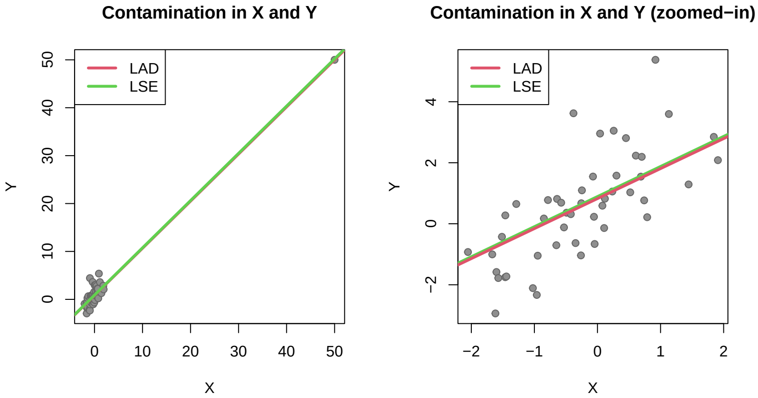
Outliers in dataset

We added one outlying point in terms of both X and Y.



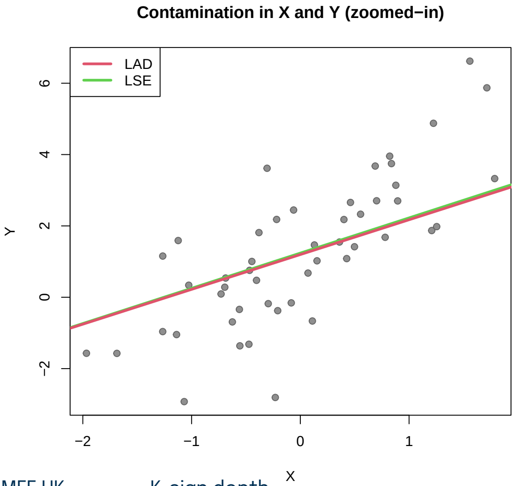
Outliers in dataset

- Clearly a bad fit for both $\hat{\beta}^{LAD} = (1.20, 0.98)^\top$ and $\hat{\beta}^{LSE} = (1.24, 0.99)^\top$
- Interestingly enough, null hypotheses **not rejected** for all K, except for $H_0 : (\beta_0, \beta_1)^\top = \hat{\beta}^{LSE}$ based on d_3 (that is $K = 3$).



Further examination

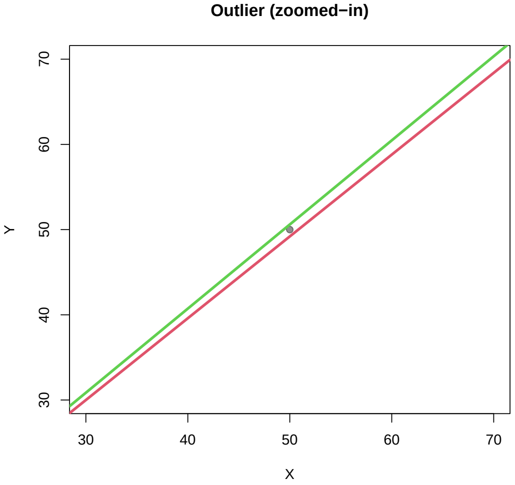
The estimated p-values are very different between $\hat{\beta}^{LAD}$ and $\hat{\beta}^{LSE}$, despite the two estimates being nearly identical.



p-values	Estimates
p_3^{LAD}	0.2472
p_4^{LAD}	0.5080
p_5^{LAD}	0.4508
p_6^{LAD}	0.5363
p_7^{LAD}	0.6201
p_3^{LSE}	**0.0333**
p_4^{LSE}	0.1477
p_5^{LSE}	0.0721
p_6^{LSE}	0.1680
p_7^{LSE}	0.1301

Further examination

A change in the sign of any residual can make a big difference in K-depth.



p-values	Estimates
p_3^{LAD}	0.2472
p_4^{LAD}	0.5080
p_5^{LAD}	0.4508
p_6^{LAD}	0.5363
p_7^{LAD}	0.6201
p_3^{LSE}	**0.0333**
p_4^{LSE}	0.1477
p_5^{LSE}	0.0721
p_6^{LSE}	0.1680
p_7^{LSE}	0.1301

K-depth estimator

What if instead $\hat{\beta}^{LAD}$ or $\hat{\beta}^{LSE}$ we define

$$\hat{\beta}^{d_k} = \mathrm{argmax}_{\beta_0,\beta_1} d_k((\beta_0, \beta_1)^\top)$$

K-depth estimator

What if instead $\hat{\beta}^{LAD}$ or $\hat{\beta}^{LSE}$ we define

$$\hat{\beta}^{d_k} = \operatorname{argmax}_{\beta_0,\beta_1} d_k((\beta_0, \beta_1)^\top)$$

- Difficult to find already for a simple linear model.
- Not unique

K-depth estimator

What if instead $\hat{\beta}^{LAD}$ or $\hat{\beta}^{LSE}$ we define

$$\hat{\beta}^{d_k} = \mathrm{argmax}_{\beta_0,\beta_1} d_k((\beta_0, \beta_1)^\top)$$

- Difficult to find already for a simple linear model.
- Not unique
- What if we only consider regression lines passing through the point $\bar{x}, \bar{y}$?

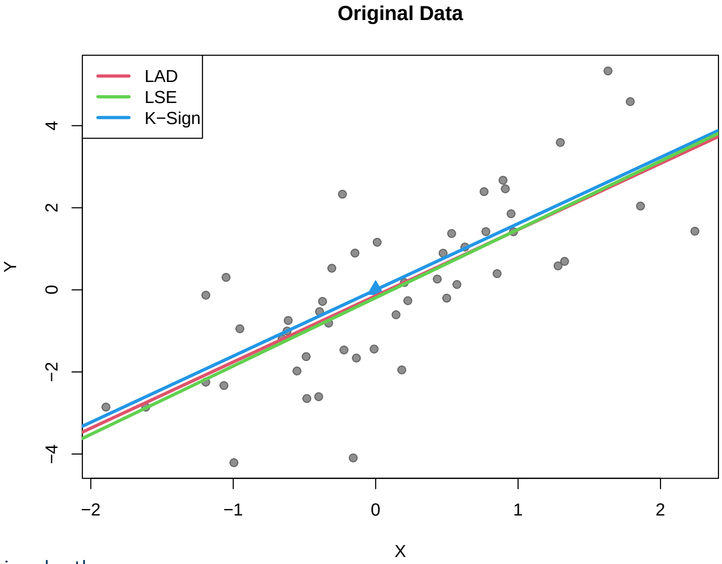
K-depth estimator

What if instead $\hat{\beta}^{LAD}$ or $\hat{\beta}^{LSE}$ we define

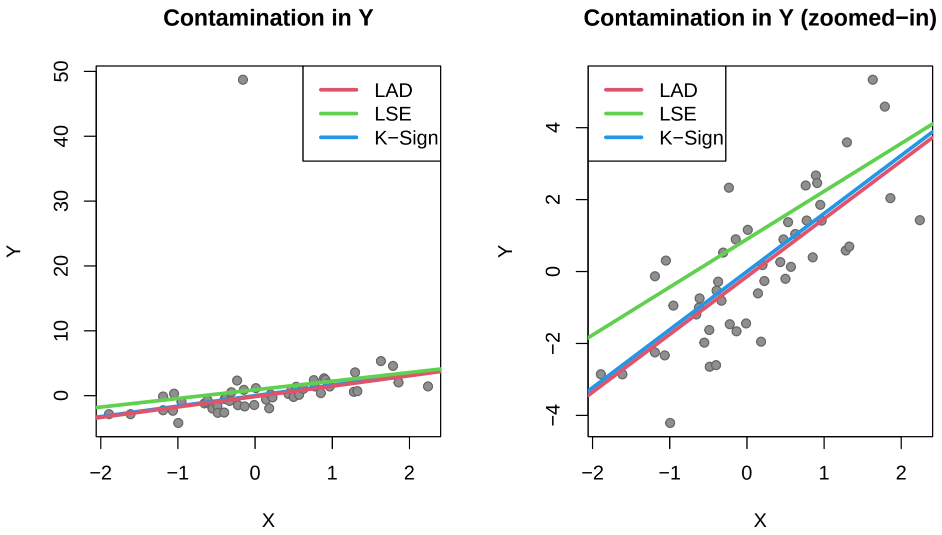
$$\hat{\beta}^{d_k} = \operatorname{argmax}_{\beta_0,\beta_1} d_k((\beta_0, \beta_1)^\top)$$

- Difficult to find already for a simple linear model.
- Not unique
- What if we only consider regression lines passing through the point $\overline{x}, \overline{y}$?
- How about $\mathrm{median}(x_1, \ldots, x_N), \mathrm{median}(y_1, \ldots, y_N)$?

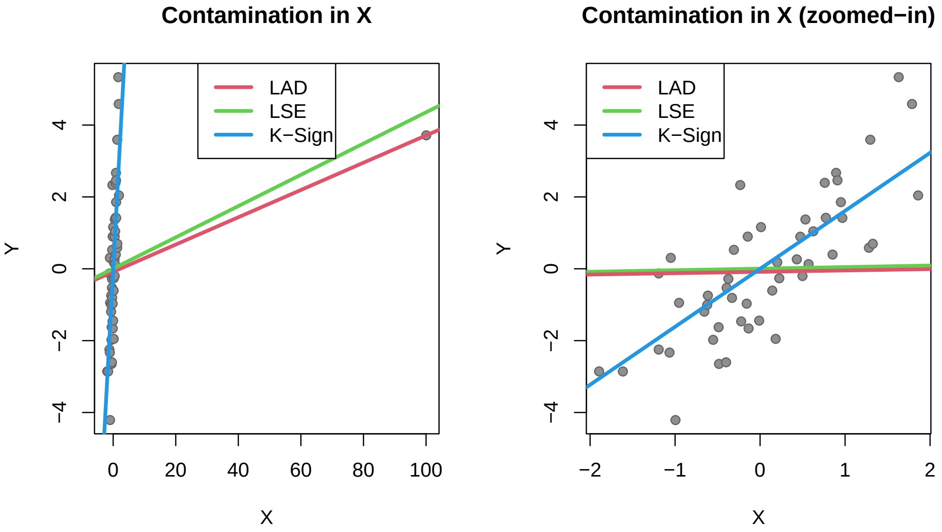
Outliers from before



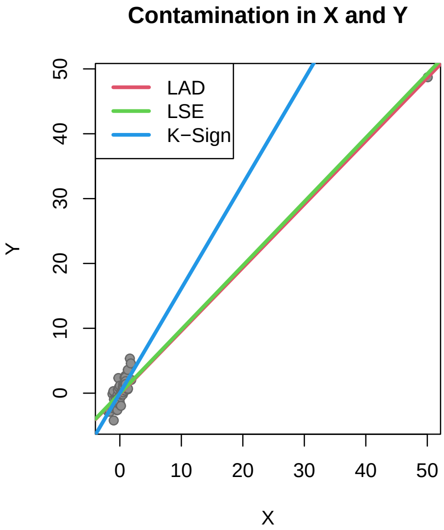
Outliers from before

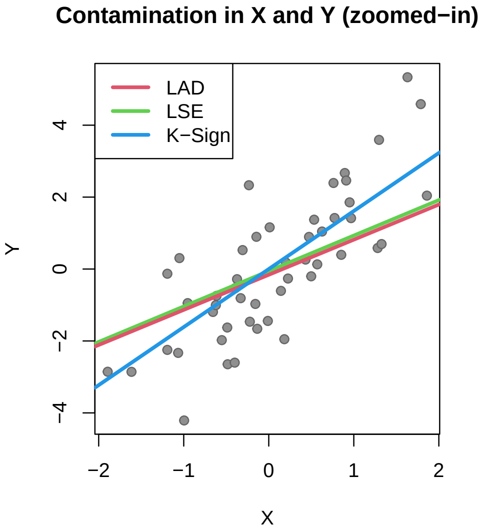


Outliers from before



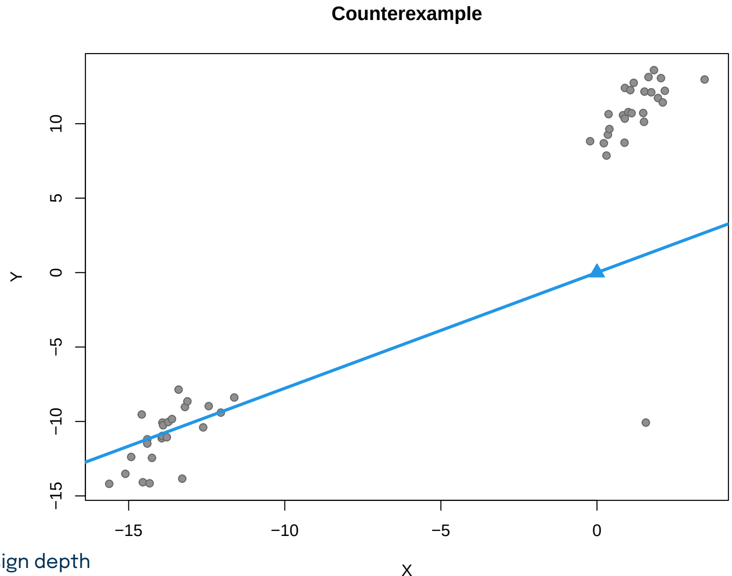
Outliers from before



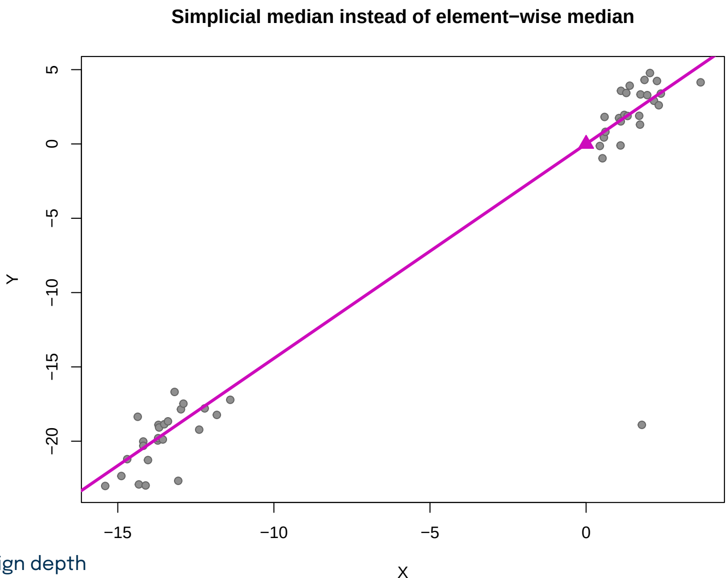


Is it unbreakable?

Is it unbreakable? No.

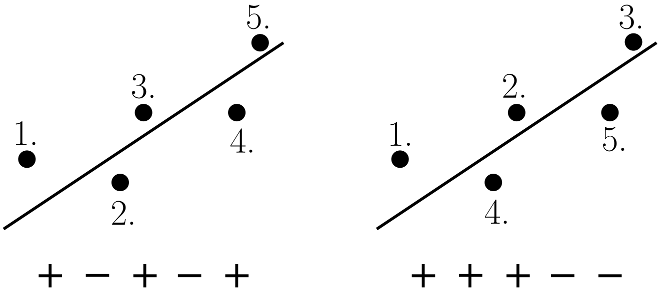


Some other possibilities



Multiple regression

The concept of K-sign depth relies heavily on the **order** of the residuals.



- Straightforward to define in a one-dimensional case.
- Can it be generalized to higher dimensions?

Ordering in the bivariate case

Consider a multiple linear regression model

$$Y_i = \beta_0 x_i^{(1)} + \beta_1 x_i^{(2)} + \varepsilon_i, \quad i = 1, \ldots, N.$$

What if we order residuals based on one of the regressors?

Ordering in the bivariate case

What if we order the residuals based on one of the regressors?

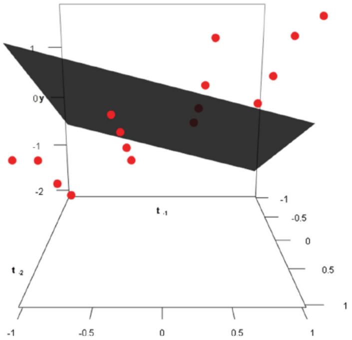
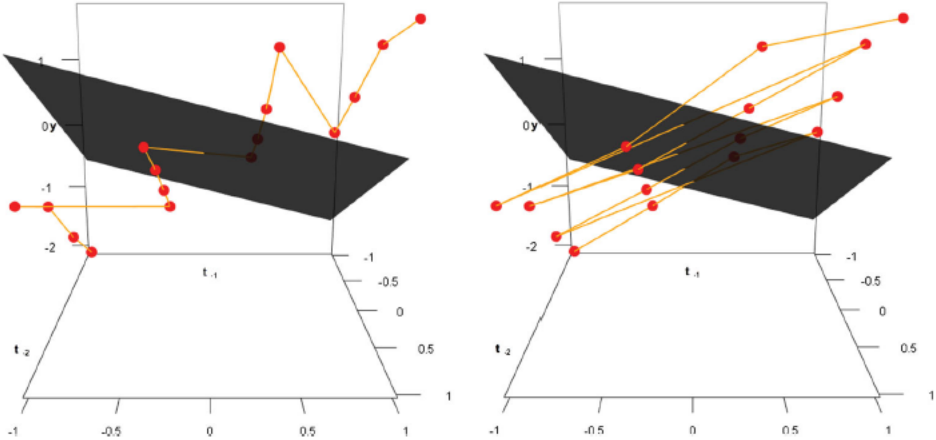


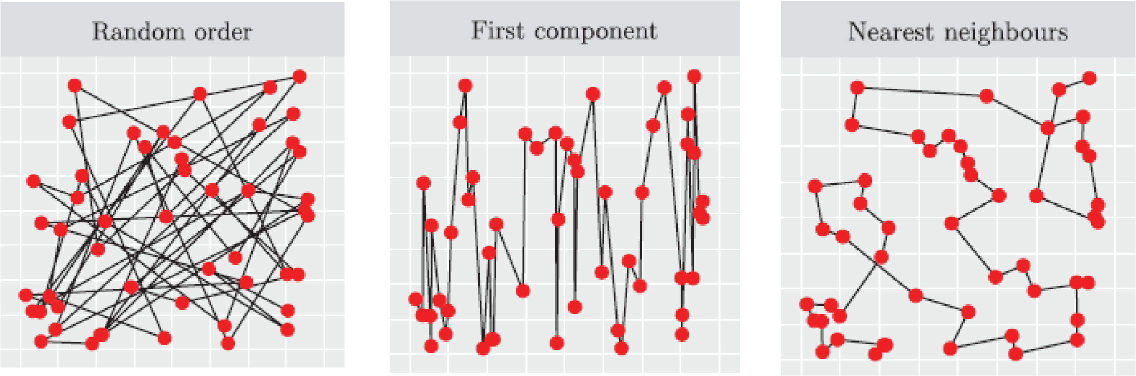
Figure 2. Visualization of the 16 data points and the fitted model (dark grey area).

Ordering in the bivariate case



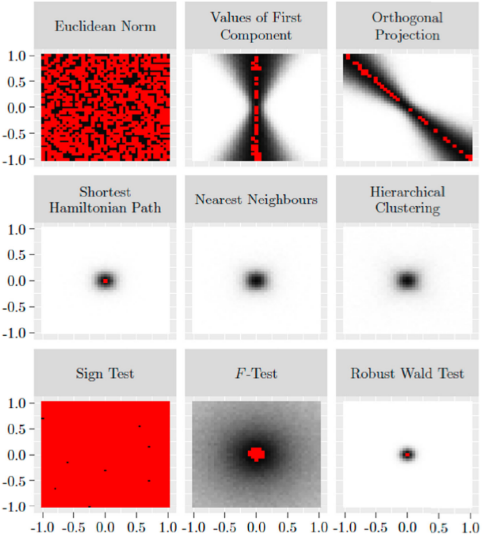
(a) (b)

Ordering in the bivariate case



Cauchy versus Normal errors

Power of the tests - simulation



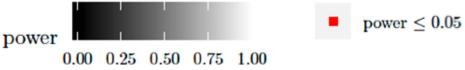
Model

$$Y_i = \beta_0 x_i^{(1)} + \beta_1 x_i^{(2)} + \varepsilon_i, \quad i = 1, \dots, N,$$

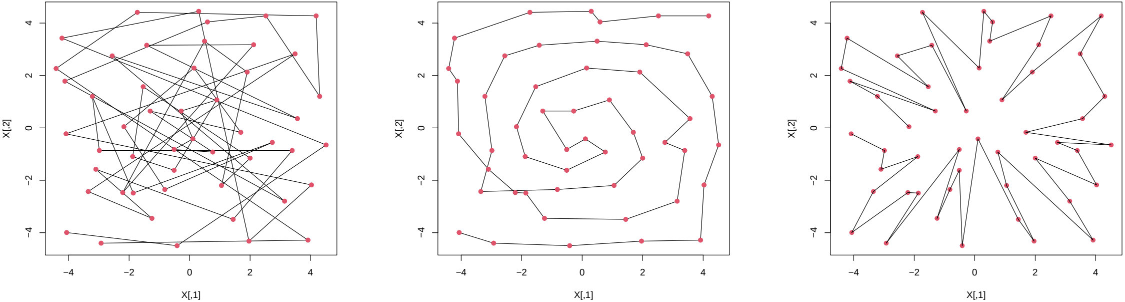
$$\varepsilon_i \sim \text{Cauchy}(0, \sigma^2); \quad x_i^{(1)}, x_i^{(2)} \sim \text{Unif}(-1, 1).$$

Null Hypothesis

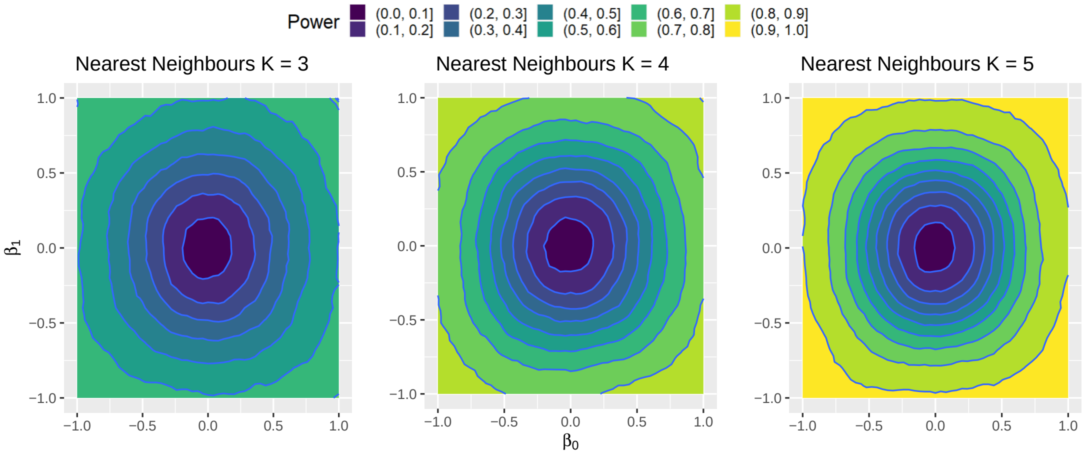
$$H_0 : \beta_0 = \beta_1 = 0.$$



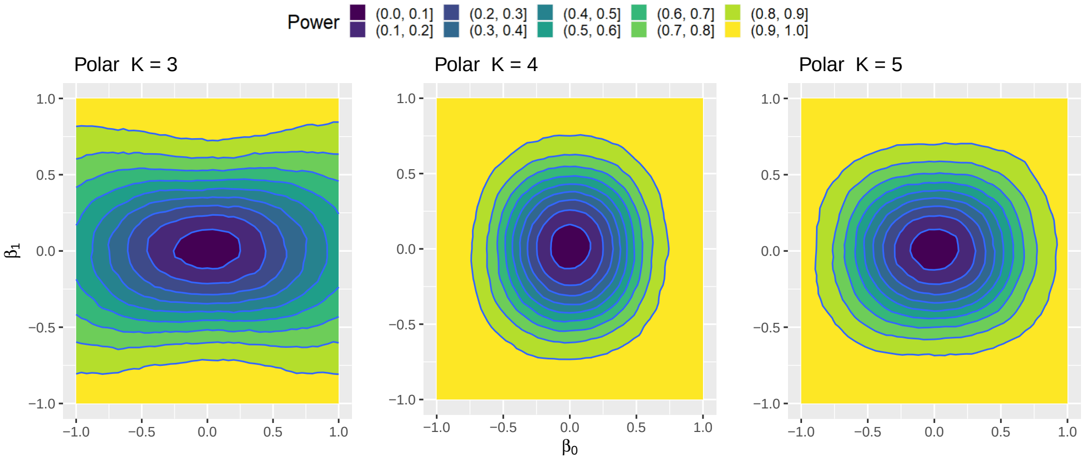
Experimenting with bivariate ordering



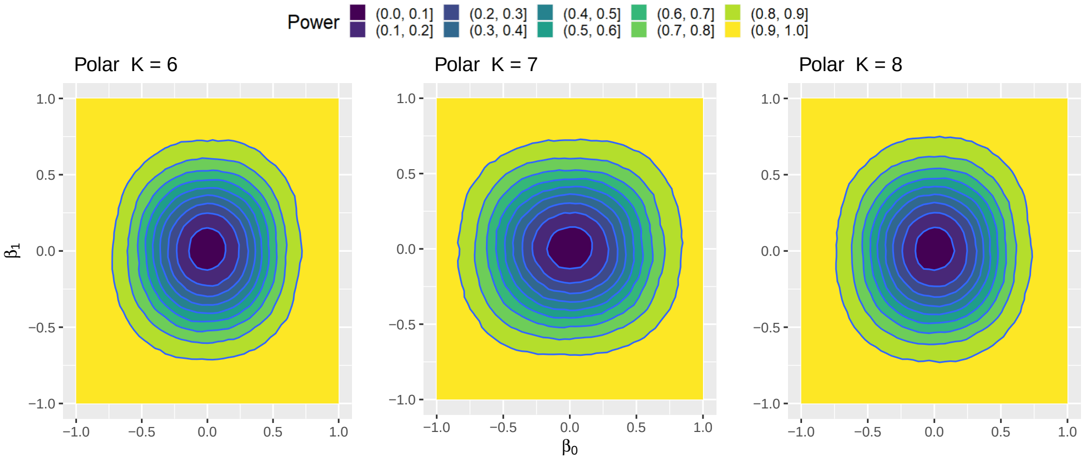
Power for different K and order



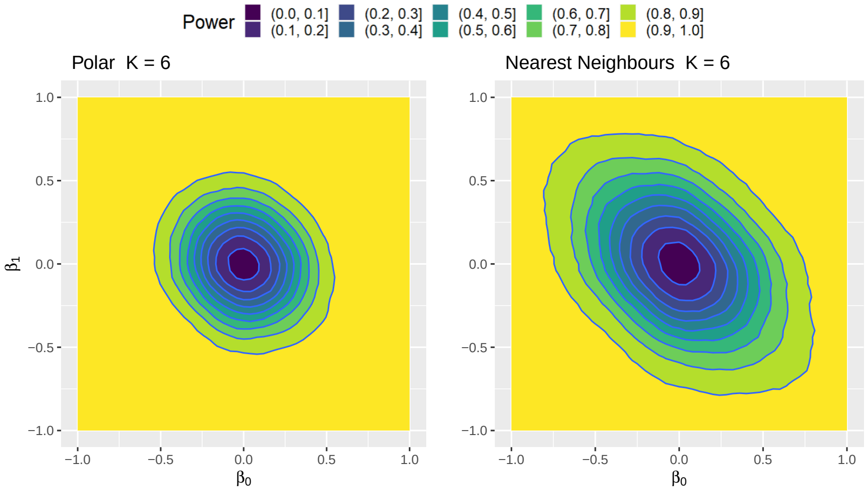
Power for different K and order



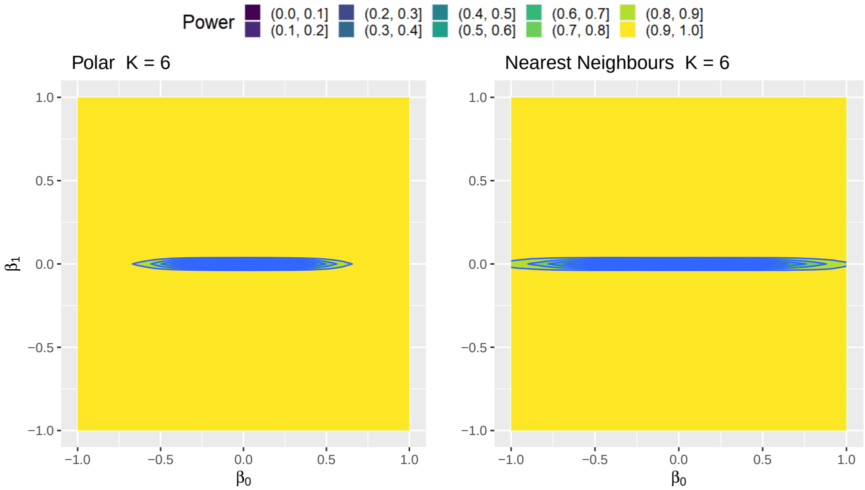
Power for different K and order



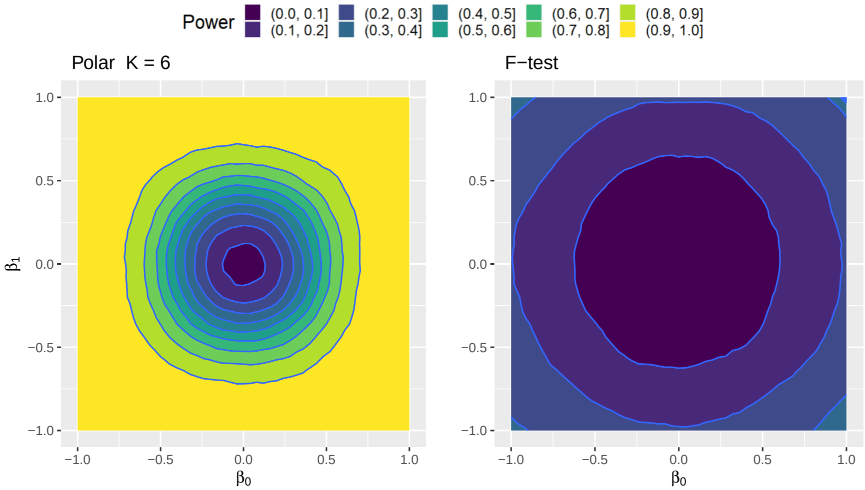
Exponentially distributed regressors



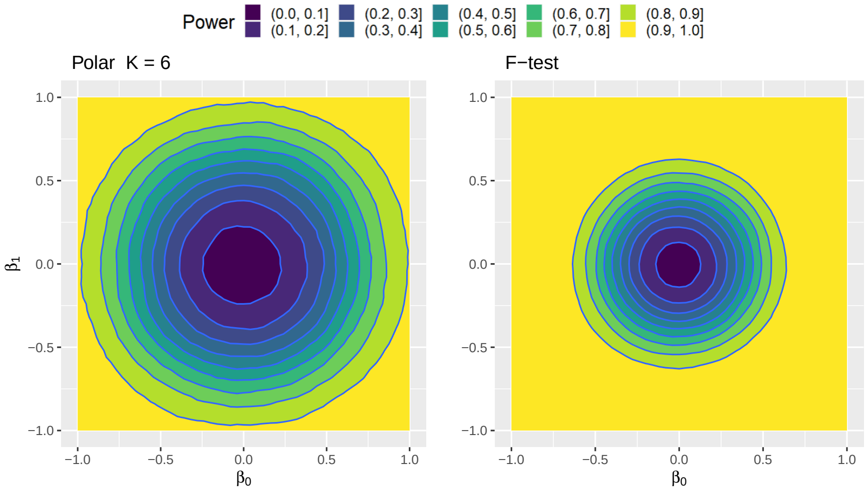
Bimodal regressors



K-sign depth versus F-test



K-sign depth versus F-test



Power of the tests - high dimensional data