Least-squares parameter estimation in the presence of outliers

Dmitry I. Kabanov

Book chapter presentations on Stochastic Numerics

17 June 2019

RWTH Aachen University

Motivation

- Least-squares regression is ubiquitously used procedure for parameter estimation
- However, it treats all data observations equally
- Therefore, outliers get the same weight as correct data
- Often leads to incorrect estimates

Problem setup

Given data

$$\boldsymbol{D} = \{D_k\} = \{x_k, Y_k\}, \quad k = 1, \ldots, N,$$

estimate parameter $\boldsymbol{\theta} \in \mathbb{R}^M$ of the data model

$$Y_k = f(x_k; \boldsymbol{\theta}) + \epsilon_k,$$

where $f(x_k; \boldsymbol{\theta})$ is ideal noiseless data and $\epsilon \in \mathbb{R}^N$ is observation noise.

Ordinary Least Squares in Bayesian setting

We reformulate the above problem using Bayes' theorem:

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{D}) \propto \mathbb{P}(\boldsymbol{D}|\boldsymbol{\theta}) \times \mathbb{P}(\boldsymbol{\theta})$$

Ordinary Least Squares in Bayesian setting

We reformulate the above problem using Bayes' theorem:

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{D}) \propto \mathbb{P}(\boldsymbol{D}|\boldsymbol{\theta}) \times \mathbb{P}(\boldsymbol{\theta})$$

Assumption 1: Uniform prior $\mathbb{P}(\boldsymbol{\theta}) = \text{constant}$

Assumption 2: Independent Gaussians for each datum. Then the likelihood:

$$\mathbb{P}(\boldsymbol{D}|\boldsymbol{\theta}) = \prod_{k=1}^{N} \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{R_k^2}{2}\right) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$\chi^2 = \sum_{k=1}^{N} R_k^2, \quad R_k = \frac{f(x_k;\boldsymbol{\theta}) - Y_k}{\sigma_k}$$

Ordinary Least Squares in Bayesian setting

We reformulate the above problem using Bayes' theorem:

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{D}) \propto \mathbb{P}(\boldsymbol{D}|\boldsymbol{\theta}) \times \mathbb{P}(\boldsymbol{\theta})$$

Assumption 1: Uniform prior $\mathbb{P}(\boldsymbol{\theta}) = \text{constant}$

Assumption 2: Independent Gaussians for each datum. Then the likelihood:

$$\mathbb{P}(\boldsymbol{D}|\boldsymbol{\theta}) = \prod_{k=1}^{N} \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{R_k^2}{2}\right) \propto \exp\left(-\frac{\chi^2}{2}\right)$$

$$\chi^2 = \sum_{k=1}^{N} R_k^2, \quad R_k = \frac{f(x_k;\boldsymbol{\theta}) - Y_k}{\sigma_k}$$

The goal is to maximize log-likelihood:

$$\boldsymbol{\theta} = \text{argmax}\left(\text{const} - \frac{\chi^2}{2}\right)$$

Extension 1: conservative formulation

Tolerate that the noise can be above some prescribed threshold:

$$\mathbb{P}\left(\sigma|\sigma_0\right) = \frac{\sigma_0}{\sigma^2} \text{ for } \sigma \geq \sigma_0 \text{ and zero otherwise}$$

Then the marginal likelihood for datum D_k is:

$$\mathbb{P}\left(D_k|\sigma_0\right) = \int_0^\infty \mathbb{P}\left(D_k|\sigma\right)\mathbb{P}\left(\sigma|\sigma_0\right)\,\mathrm{d}\sigma$$

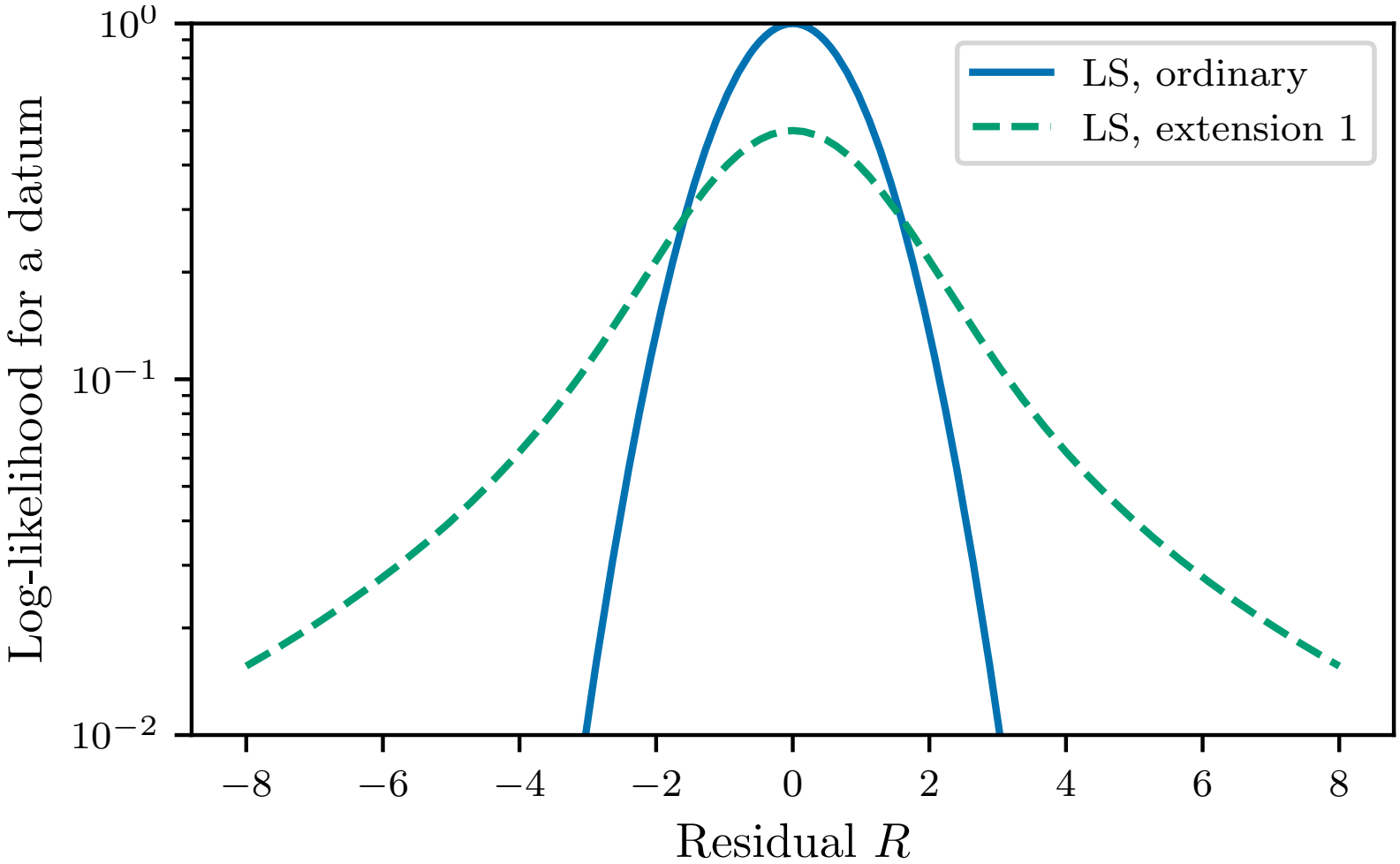
which gives

$$\mathbb{P}\left(D_k|\sigma_0\right) = \int_0^\infty \frac{\sigma_0}{\sigma^3\sqrt{2\pi}}\exp\left(-\frac{\left(F_k - D_k\right)^2}{2\sigma^2}\right)\mathrm{d}\sigma = \frac{1}{\sigma_0\sqrt{2\pi}}\frac{1-\exp\left(-R_k^2/2\right)}{R_k^2}$$

Extension 1: conservative formulation, cont.

Ordinary Least Squares: $P \sim \exp\left(-R^2/2\right)$ as $R \to \infty$

Conservative formulation: $P \sim \frac{1}{R^2}$ as $R \to \infty$ (less skewing effect)



Extension 1: conservative formulation, cont.

Treat all measurements independent of each other.

Use uniform prior distribution as in Ordinary Least Squares.

Then log-likelihood is

$$\mathcal{L} = \ln\left(\mathbb{P}\left(\boldsymbol{\theta}|\boldsymbol{D}\right)\right) = \text{constant} + \sum_{k=1}^{N} \ln\left(\frac{1-\exp\left(-\frac{R^2}{2}\right)}{R^2}\right)$$

Extension 1: conservative formulation, cont.

Treat all measurements independent of each other.

Use uniform prior distribution as in Ordinary Least Squares.

Then log-likelihood is

$$\mathcal{L} = \ln\left(\mathbb{P}\left(\boldsymbol{\theta}|\boldsymbol{D}\right)\right) = \text{constant} + \sum_{k=1}^{N} \ln\left(\frac{1-\exp\left(-\frac{R^2}{2}\right)}{R^2}\right)$$

Disadvantage: uncertainties are worse, then for ordinary least squares (to be seen later).

Extension 2: the good-and-bad data model

Distribution of noise admits two possibilities: either noise is within threshold or it is very large:

$$\mathbb{P}\left(\sigma_k|\sigma_0\right) = \boxed{(1-\beta)\,\delta(\sigma-\sigma_0)} + \boxed{\beta\,\delta(\sigma-\gamma\sigma_0)}$$

where $0 \leq \beta \ll 1$ and $\gamma \gg 1$, $\delta(\cdot)$ the Dirac delta.

β is the frequency of occurence of the outliers

γ is the scale of the outliers

Extension 2: the good-and-bad data model, cont.

Treat all measurements independent of each other.

Use uniform prior distribution as in Ordinary Least Squares.

Then the log-likelihood is:

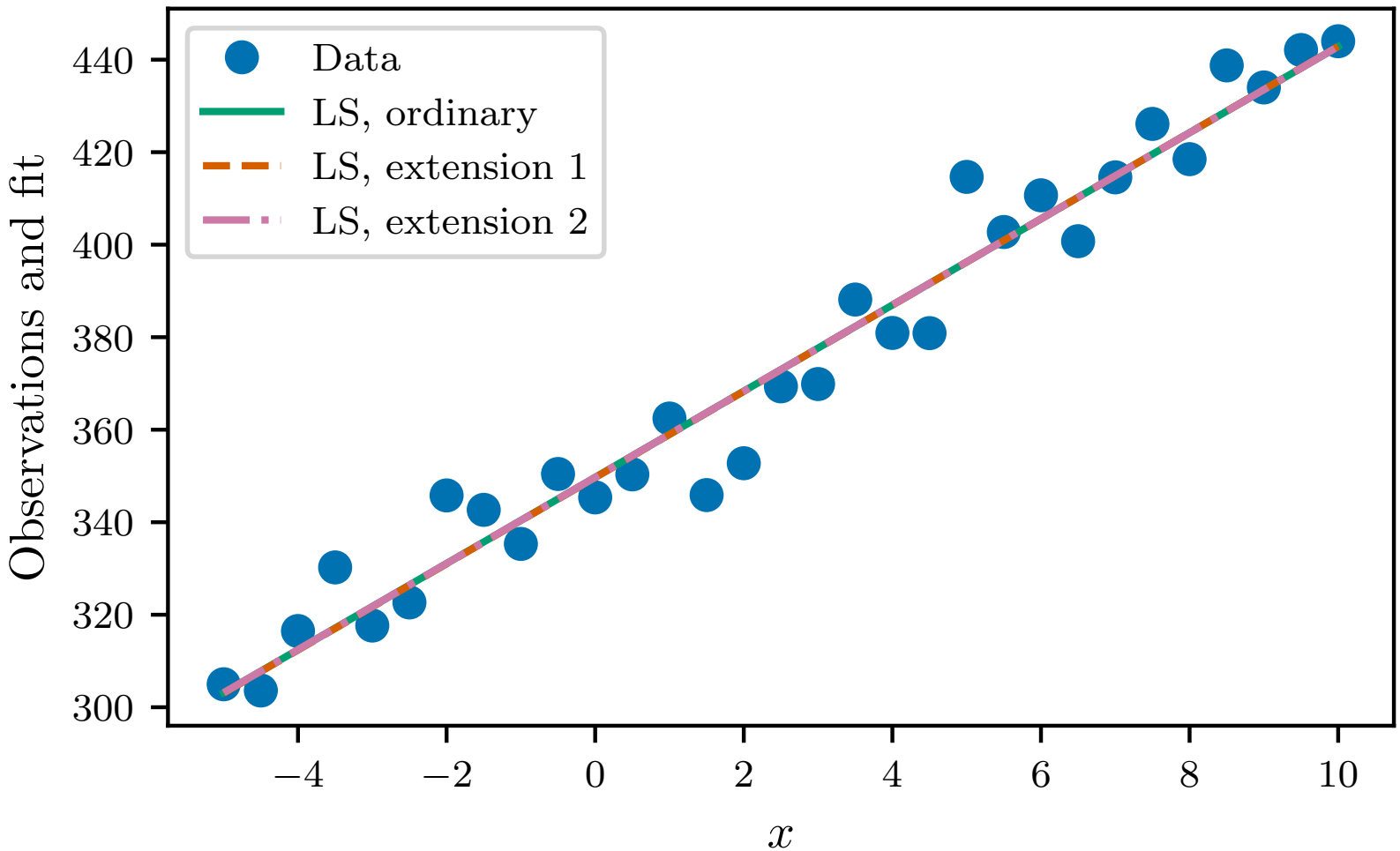
$$L = \text{constant} + \sum_{k=1}^{N} \ln \left[\frac{\beta}{\gamma} \exp\left(-\frac{R_k^2}{2\gamma^2} \right) + (1-\beta) \exp\left(-\frac{R_k^2}{2} \right) \right]$$

where

$$R_k = \frac{f(x_k;\theta) - Y_k}{\sigma_0}$$

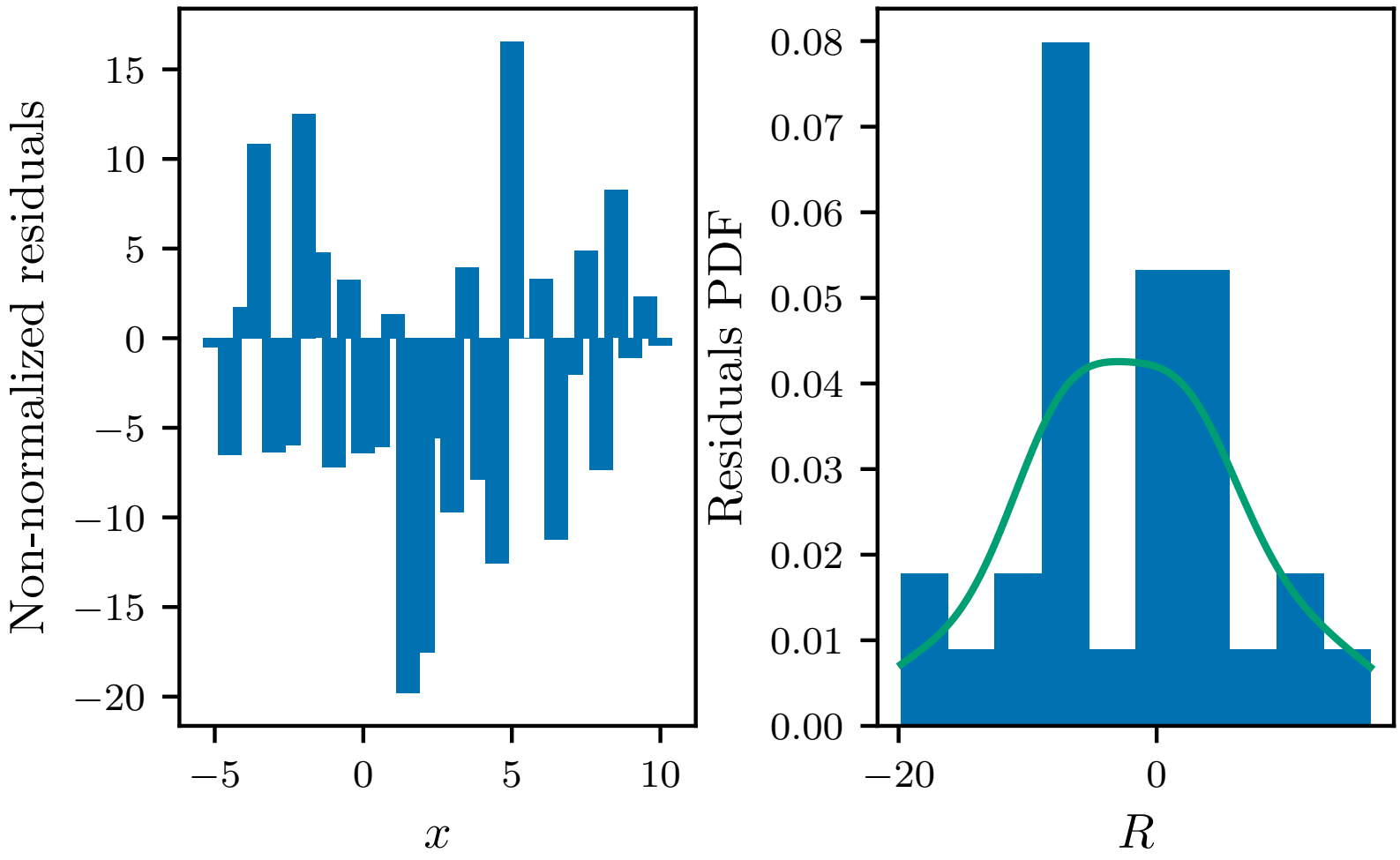
Example 1: no outliers

31 observations: data $= mx + b + 10N(0, 1)$



Example 1: no outliers, cont.

31 observations: data $= mx + b + 10N(0, 1)$



Example 1: no outliers, cont.

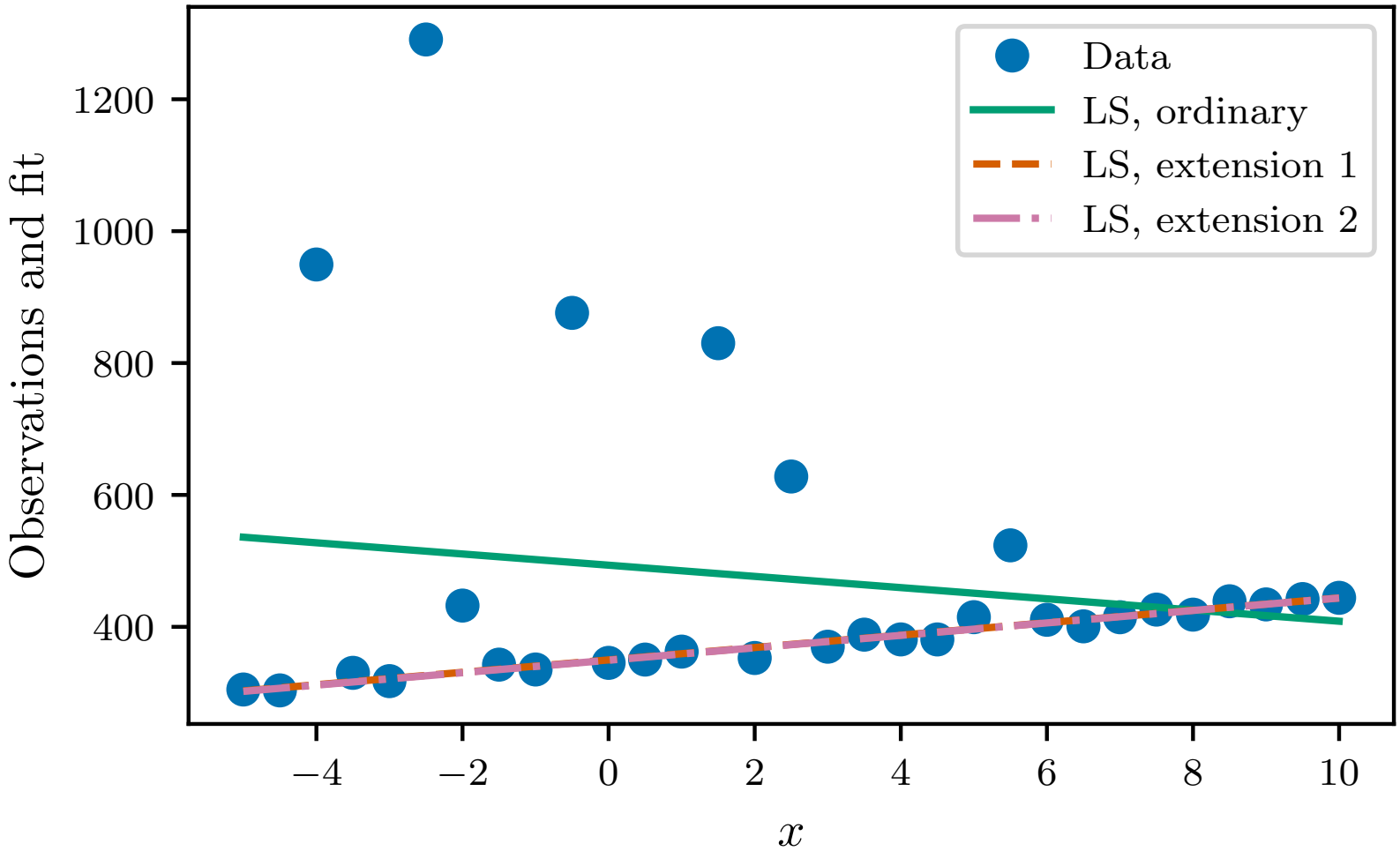
31 observations: $\text{data} = mx + b + 10N(0, 1)$

Type of method	m	se_m	b	se_b
Ideal	10.0	0	350.0	0
LS, ordinary	9.3	0.8	349.7	4.0
LS, extension 1	9.3	1.2	349.7	6.2
LS, extension 2	9.3	0.8	349.7	4.1

Example 2: six outliers

31 observations: data $= mx + b + 10N(0, 1)$

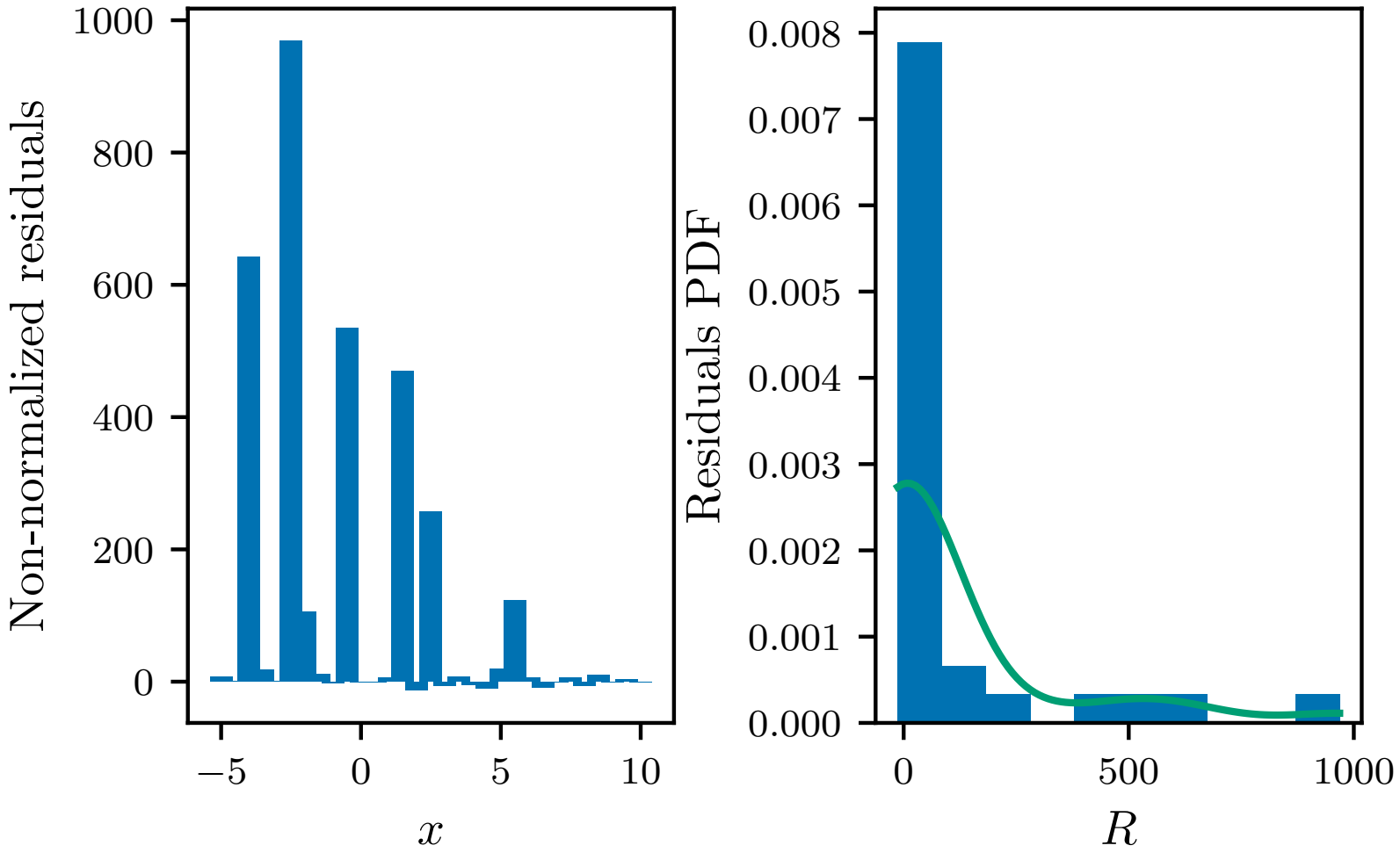
6 observations are increased 2–5 times



Example 2: six outliers, cont.

31 observations: data $= mx + b + 10N(0, 1)$

6 observations are increased 2–5 times



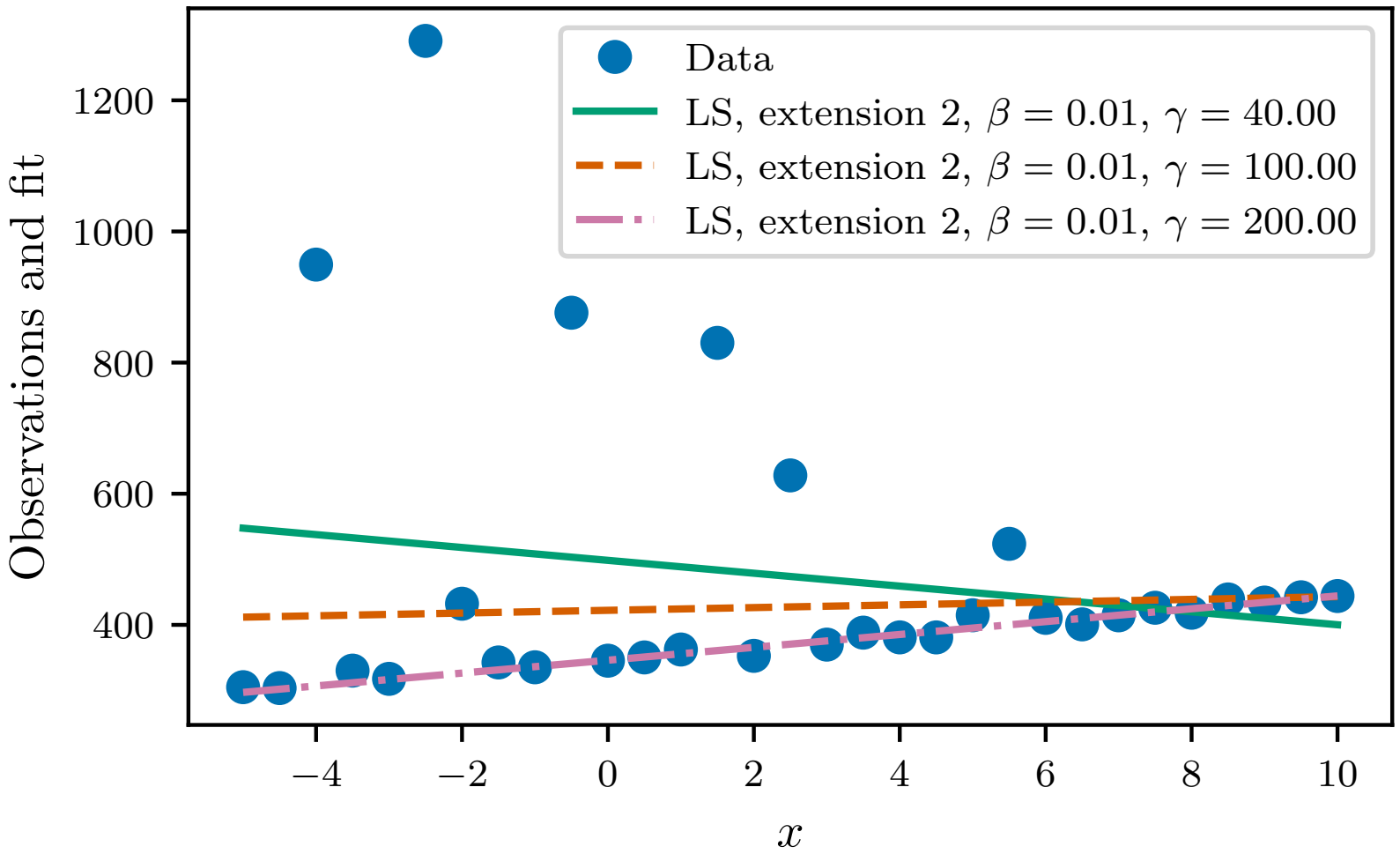
Example 2: six outliers, cont.

31 observations: data $= mx + b + 10N(0, 1)$

6 observations are increased 2–5 times

Type of method	m	se_m	b	se_b
Ideal	10.0	0	350.0	0
LS, ordinary	-8.5	0.8	493.5	4.0
LS, extension 1	9.4	1.3	349.8	7.4
LS, extension 2	9.4	0.9	349.3	4.9

LS extension 2 has disadvantage: sensitive to parameters



LS extension 2 has disadvantage: sensitive to parameters

Type of method	m	se_m	b	se_b
Ideal	10.0	0	350.0	0
LS, ext. 2, $\beta = 0.01$, $\gamma = 50$	−9.8	1.9	498.4	14.9
LS, ext. 2, $\beta = 0.01$, $\gamma = 100$	2.1	1.7	422.2	15.6
LS, ext. 2, $\beta = 0.01$, $\gamma = 200$	9.8	0.1	346.1	1.0

Conclusions

- Ordinary Least Squares (OLS) are prone to wrong estimation in the presence of outliers (not robust)
- Bayesian formulation of the least-squares algorithms simplifies adjustments needed to accommodate outliers
- Two extensions were considered: conservative formulation and bad-and-good formulation
- Both formulations require careful choice of parameters for good performance
- Conservative formulation gives worse uncertainties on the estimates, but requires less parameters than the bad-and-good formulation

References and codes

- Sivia D. S., Skilling J. *Data analysis: a Bayesian tutorial*, OUP Oxford, 2006
- Box, George E. P, Tiao G. C. *A Bayesian approach to some outlier problems*, Biometrika, 1968, pp. 119–129

- Codes for this presentation are available at: `https://github.com/dmitry-kabanov/stochastic-numerics-2019-robust-least-squares`

Thank you!