HelmholtzZentrum münchen
German Research Center for Environmental Health

Robust parameter estimation for dynamical systems from outlier-corrupted data

Corinna Maier, Carolin Loos, and Jan Hasenauer

Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Munich, Germany
Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Munich, Germany

Problem statement

- Outliers in biological data arise due to human errors and technical limitations
- The distinction between outlier and no-outlier data points is complicated and therefore challenges the manual removal of outliers
- In the presence of outliers, parameter estimation, applied to calibrate mathematical models to the data, can result in large estimation errors limiting the validity of the models
- Robust estimation methods proposed in regression have not yet been applied and evaluated for dynamical systems

Data-driven modeling of dynamic biological systems

- Ordinary Differential Equation (ODE) model describing the temporal evolution of the concentrations of the species

$$\dot{x} = f(x,\xi)\,, \quad x(0) = x_0(\xi)$$

with time dependent states $x(t) \in \mathbb{R}_+^{n_x}$, vector field f and parameters $\xi \in \mathbb{R}_+^{n_\xi}$

- Mapping to observables $y = h(x,\xi)$
- Calibrate model for experimental data sets $\mathcal{D} = \{(t_k, \bar{y}_k)\}_{k=1}^{n_t}$
- Due to measurement noise, it is assumed that the measured value $\bar{y}_k$ of an observable is distributed

$$\bar{y}_{i,k} \sim p(\bar{y}_{i,k} | y_i(t_k,\xi), \varphi_i)$$

- Estimating kinetic and distribution specific parameters $\theta = (\xi, \varphi)$ by maximum likelihood estimation

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} p(\bar{y}_{i,k} | y_i(t_k,\xi), \varphi_i)$$

- In the presence of outliers, single observations are drawn from an alternative distribution with heavier tails which is difficult to assess due to small sample sizes

→ What distribution assumption should be used for outlier-corrupted data?

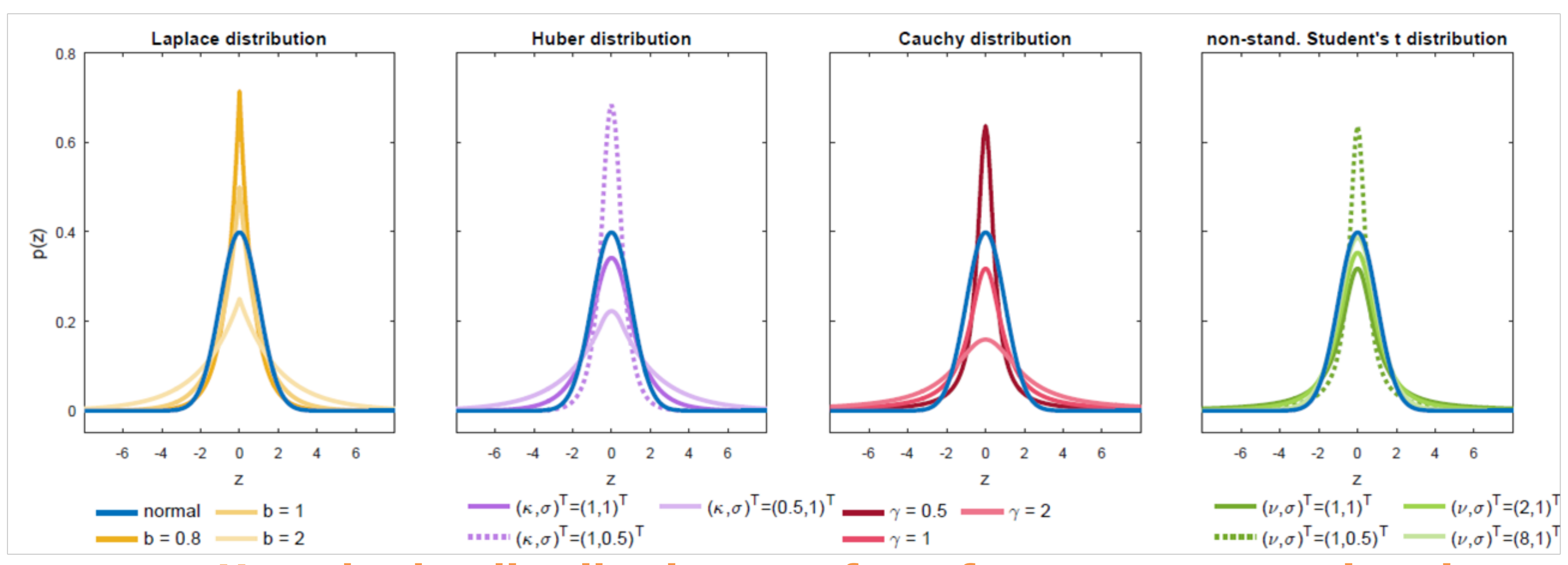
Distribution assumptions

- Normal distribution $p(\bar{y}|y,\sigma_n) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma_n}\right)^2\right)$
- Huber distribution

$$p(\bar{y}|y,\sigma_h,\kappa) = s \cdot \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma_h}\right)^2\right), & \left|\frac{\bar{y}-y}{\sigma_h}\right| \le \kappa \\ \exp\left(-\frac{1}{2}\left(2\kappa\left|\frac{\bar{y}-y}{\sigma_h}\right| - \kappa^2\right)\right), & \left|\frac{\bar{y}-y}{\sigma_h}\right| > \kappa \end{cases}$$

with $s = \left(\sqrt{2\pi}\sigma_h \operatorname{erf}\left(\frac{\kappa}{\sqrt{2}}\right) + \frac{2\sigma_h}{\kappa}\exp(-\frac{1}{2}\kappa^2)\right)^{-1}$

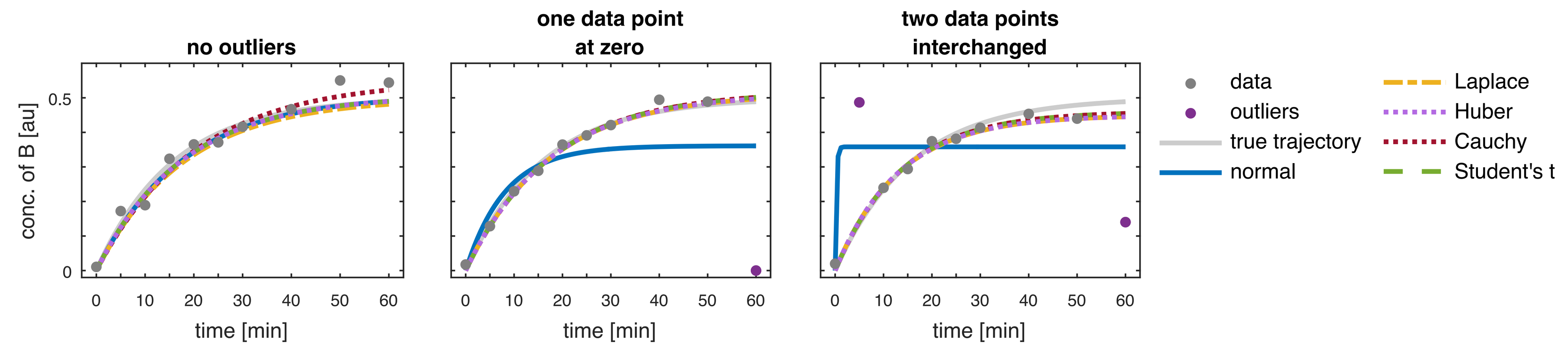
- Laplace distribution $p(\bar{y}|y,b) = \frac{1}{2b}\exp\left(-\frac{|\bar{y}-y|}{b}\right)$
- Cauchy distribution $p(\bar{y}|y,\gamma) = \frac{1}{\pi\gamma}\frac{\gamma^2}{(\bar{y}-y)^2+\gamma^2}$
- Student's t distribution $p(\bar{y}|y,\sigma_t,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_t}\left(1+\frac{1}{\nu}\left(\frac{\bar{y}-y}{\sigma_t}\right)^2\right)^{-\frac{\nu+1}{2}}$

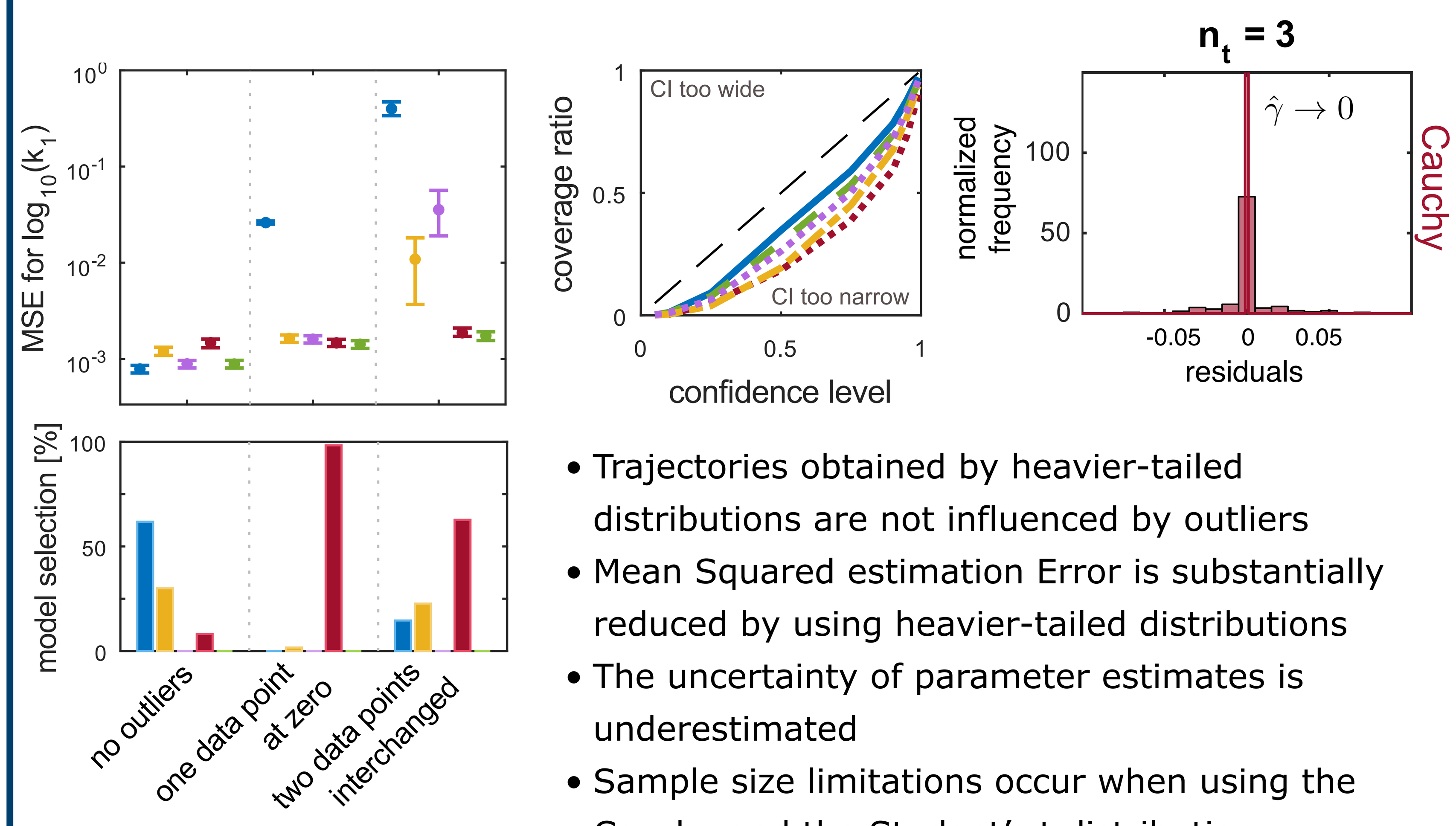


→ How do the distributions perform for parameter estimation in the presence and absence of outliers?

Simulation study: Conversion reaction $A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} B$

- Analysis using 100 data sets of a conversion reaction for each scenario
- Estimation using multi-start local optimization providing the analytical gradients
- Comparison of estimators for three outlier scenarios: i) no outlier ii) one data point at zero iii) two interchanged data points



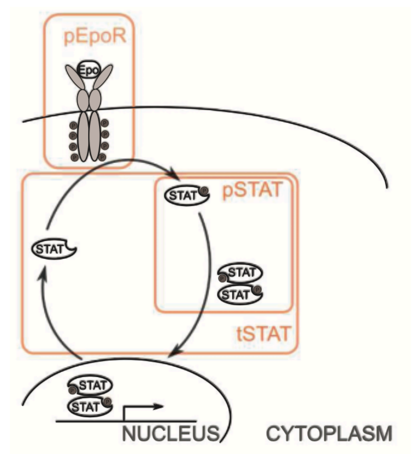


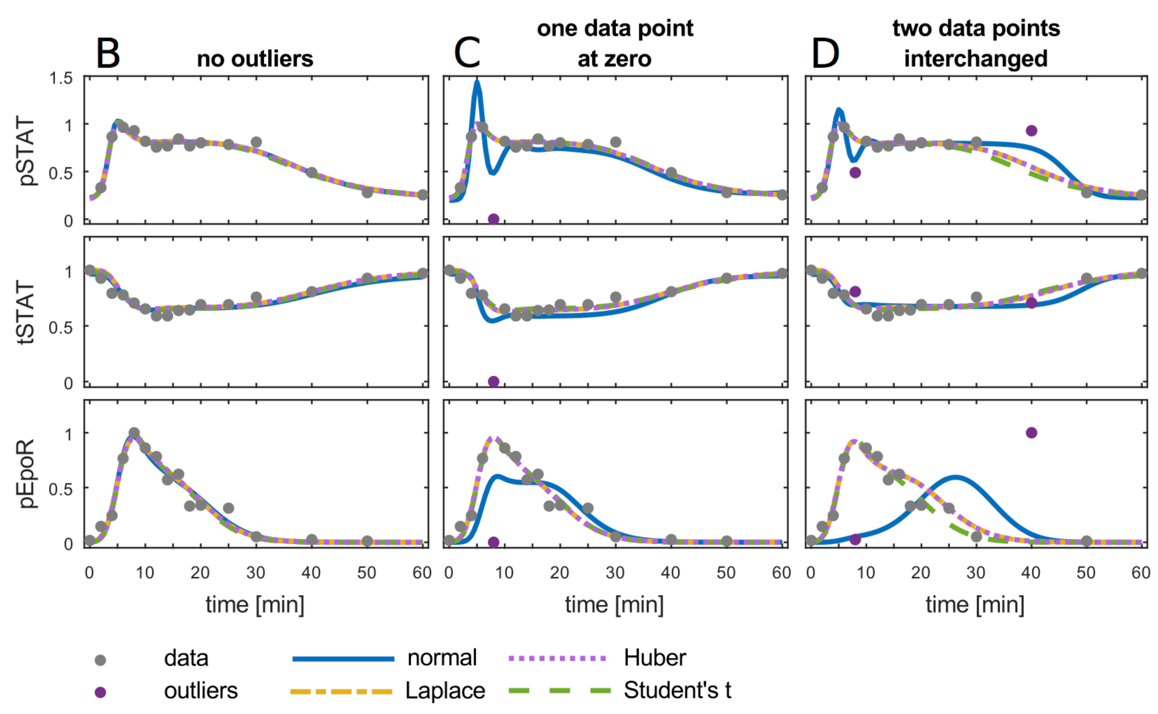
- Trajectories obtained by heavier-tailed distributions are not influenced by outliers
- Mean Squared estimation Error is substantially reduced by using heavier-tailed distributions
- The uncertainty of parameter estimates is underestimated
- Sample size limitations occur when using the Cauchy and the Student's t distribution

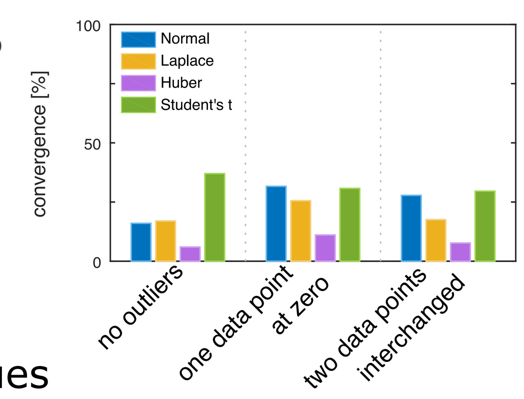
→ Heavier-tailed distributions provide more robust estimates, but need to be applied carefully

Application study: JAK/STAT signaling

- Data and model adapted from Swameye et al. (2003)
- Artificially introduced outliers according to outlier scenarios







- Exclusion of Cauchy distribution due to over-fitting issues
- Heavier-tailed distributions provide robust estimates for all scenarios
- Convergence and performance of heavier-tailed distributions comparable with the normal distribution

→ Heavier-tailed distributions should be considered when potential outliers are in the data

Summary

- We examined the use of the Laplace, Cauchy, Student's t and Huber distribution instead of the generally used normal distribution assumption
- In the presence of outliers, the heavier-tailed distribution assumptions yield more robust and more precise parameter estimates with similar performance and convergence
- Using model selection, the presence of potential outliers can be revealed and further experiments or model refinements can be applied to improve the analysis

For additional information, please contact:
Corinna Maier, corinna.maier@uni-potsdam.de

Current affiliation:
Graduate Research Training program PharMetrX, Germany
Institute of Mathematics, University of Potsdam

HELMHOLTZ ASSOCIATION