Supplement:

Robust parameter estimation for dynamical systems from outlier-corrupted data

Corinna Maier[1,2], Carolin Loos[1,2], and Jan Hasenauer[1,2,*]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany,

[2]Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany

[*]corresponding author (e-mail: jan.hasenauer@helmholtz-muenchen.de)

Contents

1 Statistical models for distribution assumptions

In this section the likelihood and the log-likelihood function for the different distribution assumptions are presented. In addition, the corresponding gradients and Hessian matrices of the log-likelihoods are listed.

1.1 Normal distribution

Under the assumption of independent normally distributed measurement noise, the likelihood is

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \left[\frac{1}{\sqrt{2\pi}\ \sigma_i(\theta)} \exp\Big(-\frac{1}{2} \frac{(\bar{y}_{ik} - y_i(t_k, \theta))^2}{\sigma_i^2(\theta)} \Big) \right].$$

Accordingly, the log-likelihood function is given by

$$\log \mathcal{L}_{\mathcal{D}}(\theta) = -\frac{1}{2} \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\log(2\pi\sigma_i^2(\theta)) + \Big(\frac{\bar{y}_{ik} - y_i(t_k, \theta)}{\sigma_i(\theta)} \Big)^2 \right].$$

The gradient of the log-likelihood for $l = 1, \dots n_\theta$ is given by

$$\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\theta)}{\partial \theta_l} = -\frac{1}{2} \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\frac{1}{\sigma_i^2(\theta)} \Big(1 - \frac{(\bar{y}_{ik} - y_i(t_k, \theta))^2}{\sigma_i^2(\theta)} \Big) \frac{\partial \sigma_i^2(\theta)}{\partial \theta_l} - 2\ \frac{\bar{y}_{ik} - y_i(t_k, \theta)}{\sigma_i^2(\theta)} \frac{\partial y_i(t_k, \theta)}{\partial \theta_l} \right]$$

and the Hessian matrix for $l, m = 1, \dots, n_\theta$ by

$$\begin{aligned}
\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\theta)}{\partial \theta_l \partial \theta_m} = -\frac{1}{2} \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Bigg[& -\frac{1}{\sigma_i^4(\theta)} \left(1 - 2\frac{(\bar{y}_{ik} - y_i(t_k, \theta))^2}{\sigma_i^2(\theta)} \right) \frac{\partial \sigma_i^2(\theta)}{\partial \theta_l} \frac{\partial \sigma_i^2(\theta)}{\partial \theta_m} \\
& + \frac{1}{\sigma_i^2(\theta)} \left(1 - \frac{(\bar{y}_{ik} - y_i(t_k, \theta))^2}{\sigma_i^2(\theta)} \right) \frac{\partial^2 \sigma_i^2(\theta)}{\partial \theta_l \partial \theta_m} \\
& + 2\frac{(\bar{y}_{ik} - y_i(t_k, \theta))}{\sigma_i^4(\theta)} \left(\frac{\partial \sigma_i^2(\theta)}{\partial \theta_l} \frac{\partial y_i(t_k, \theta)}{\partial \theta_m} + \frac{\partial \sigma_i^2(\theta)}{\partial \theta_m} \frac{\partial y_i(t_k, \theta)}{\partial \theta_l} \right) \\
& + 2\frac{1}{\sigma_i^2(\theta)} \frac{\partial y_i(t_k, \theta)}{\partial \theta_l} \frac{\partial y_i(t_k, \theta)}{\partial \theta_m} \\
& - 2\frac{\bar{y}_{ik} - y_i(t_k, \theta)}{\sigma_i^2(\theta)} \frac{\partial^2 y_i(t_k, \theta)}{\partial \theta_l \partial \theta_m} \Bigg].
\end{aligned}$$

For the optimization, we approximated the Hessian by neglecting the terms which depend on the second-order derivative of the outputs with respect to the parameters, yielding the Fisher Information Matrix (FIM).

1.2 Laplace distribution

Under the assumption of independent Laplace distributed measurement noise, the likelihood is

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \left[\frac{1}{2b_i(\theta)} \exp\Big(\frac{-|\bar{y}_{ik} - y_i(t_k, \theta)|}{b_i(\theta)} \Big) \right].$$

Accordingly, the log-likelihood function is given by

$$\log \mathcal{L}_{\mathcal{D}}(\theta) = -\sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\log(2b_i(\theta)) \ + \frac{|\bar{y}_{ik} - y_i(t_k, \theta)|}{b_i(\theta)} \right].$$

The gradient of the log-likelihood for $l = 1, \dots n_\theta$ is given by

$$\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\theta)}{\partial \theta_l} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\Big(-\frac{1}{b_i(\theta)} + \frac{|\bar{y}_{ik} - y_i(t_k, \theta)|}{b_i^2(\theta)} \Big) \frac{\partial b_i(\theta)}{\partial \theta_l} + \frac{\operatorname{sgn}(\bar{y}_{ik} - y_i(t_k, \theta))}{b_i(\theta)} \frac{\partial y_i(t_k, \theta)}{\partial \theta_l} \right],$$

and the Hessian matrix for $l, m = 1, \ldots, n_\theta$ by

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}_\mathcal{D}(\theta)}{\partial\theta_l\theta_m} = \sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\Bigg[&\left(-\frac{1}{b_i(\theta)} + \frac{|\bar{y}_{ik} - y_i(t_k,\theta)|}{b_i^2(\theta)}\right)\frac{\partial^2 b_i(\theta)}{\partial\theta_l\partial\theta_m} \\ &+ \left(\frac{1}{b_i^2(\theta)} - \frac{2|\bar{y}_{ik} - y_i(t_k,\theta)|}{b_i^3(\theta)}\right)\frac{\partial b_i(\theta)}{\partial\theta_l}\frac{\partial b_i(\theta)}{\partial\theta_m} \\ &- \frac{\mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta))}{b_i^2(\theta)}\left(\frac{\partial b_i(\theta)}{\partial\theta_l}\frac{\partial y_i(t_k,\theta)}{\partial\theta_m} + \frac{\partial b_i(\theta)}{\partial\theta_m}\frac{\partial y_i(t_k,\theta)}{\partial\theta_l}\right) \\ &+ \frac{\mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta))}{b_i(\theta)}\frac{\partial^2 y_i(t_k,\theta)}{\partial\theta_l\partial\theta_m}\Bigg],\end{aligned}$$

Note that the term including the second-order sensitivities has an influence on the Hessian even for small deviations of the measurement and observable. This required simulation of the second-order sensitivities slows down the computation of the Hessian and we therefore used an algorithm that does not rely on a user-supplied Hessian e.g. the interior-point algorithm.

1.3 Huber distribution

The Huber M-estimator exploits a combination of squared 2-norm and 1-norm for penalization. Residuals with absolute value below κ are penalized quadratically while residuals with absolute values larger κ are penalized linearly. For an individual data point, this can be expressed using the distribution

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

with $s = \left(\sqrt{2\pi}\sigma_h \mathrm{erf}\big(\frac{\kappa}{\sqrt{2}}\big) + \frac{2\sigma_h}{\kappa}\exp\big(-\frac{1}{2}\kappa^2\big)\right)^{-1}$, which we denote in this manuscript as Huber distribution. The constant s normalizes the distribution such that it possesses integral 1. Under the assumption of independent Huber distributed measurement noise, the likelihood is

$$\mathcal{L}_\mathcal{D}(\theta) = \prod_{k=1}^{n_t}\prod_{i=1}^{n_y}\left[s\cdot\begin{cases}\exp(-\frac{1}{2}(r_i(t_k,\theta))^2) & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \exp(-\frac{1}{2}(2\kappa|r_i(t_k,\theta)| - \kappa^2)) & |r_i(t_k,\theta)| > \kappa(\theta)\end{cases}\right],$$

with $r_i(t_k,\theta) = \frac{\bar{y}_{ik} - y_i(t_k,\theta)}{\sigma_i(\theta)}$. Accordingly, the log-likelihood function is given by

$$\log\mathcal{L}_\mathcal{D}(\theta) = \sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\left[\log(s) - \begin{cases}\frac{1}{2}(r_i(t_k,\theta))^2 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{1}{2}(2\kappa|r_i(t_k,\theta)| - \kappa^2) & |r_i(t_k,\theta)| > \kappa(\theta)\end{cases}\right].$$

The gradient of the log-likelihood for $l = 1, \ldots n_\theta$ is given by

$$\begin{aligned}\frac{\partial \log\mathcal{L}_\mathcal{D}(\theta)}{\partial\theta_l} = \sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\Bigg[&\frac{\partial y_i(t_k,\theta)}{\partial\theta_l}\cdot\begin{cases}\frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^2(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ (\frac{\kappa(\theta)}{\sigma_i(\theta)}\mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta)) & |r_i(t_k,\theta)| > \kappa(\theta)\end{cases} \\ &+ \frac{\partial\sigma_i(\theta)}{\partial\theta_l}\left(-\frac{1}{\sigma_i(\theta)} + \begin{cases}\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^3(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{\kappa(\theta)}{\sigma_i^2(\theta)}|(\bar{y}_{ik} - y_i(t_k,\theta))| & |r_i(t_k,\theta)| > \kappa(\theta)\end{cases}\right) \\ &+ \frac{\partial\kappa(\theta)}{\partial\theta_l}\left(\frac{\frac{2}{\kappa^2(\theta)}\exp(-\frac{1}{2}\kappa^2(\theta))}{\sqrt{2\pi}\mathrm{erf}\Big(\frac{\kappa(\theta)}{\sqrt{2}}\Big) + \frac{2}{\kappa(\theta)}\exp(-\frac{1}{2}\kappa^2(\theta))} - \begin{cases}0 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{|(\bar{y}_{ik} - y_i(t_k,\theta))|}{\sigma_i(\theta)} - \kappa(\theta) & |r_i(t_k,\theta)| > \kappa(\theta)\end{cases}\right)\Bigg],\end{aligned}$$

and the Hessian matrix for $l, m = 1, \dots, n_\theta$ by

$$
\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_\mathcal{D}(\theta)}{\partial \theta_l \partial \theta_m} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Bigg[& \frac{\partial^2 y_i(t_k,\theta)}{\partial \theta_l \partial \theta_m} \cdot \begin{cases} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^2(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ (\frac{\kappa(\theta)}{\sigma_i(\theta)} \mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta)) & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \\
& + \frac{\partial^2 \sigma_i(\theta)}{\partial \theta_l \partial \theta_m} \left(- \frac{1}{\sigma_i(\theta)} + \begin{cases} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^3(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{\kappa(\theta)}{\sigma_i^2(\theta)} |(\bar{y}_{ik} - y_i(t_k,\theta))| & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \right) \\
& + \frac{\partial^2 \kappa(\theta)}{\partial \theta_l \partial \theta_m} \left(\frac{\frac{2}{\kappa^2(\theta)} \exp(-\frac{1}{2}\kappa^2(\theta))}{\sqrt{2\pi} \mathrm{erf}\left(\frac{\kappa(\theta)}{\sqrt{2}}\right) + \frac{2}{\kappa(\theta)} \exp(-\frac{1}{2}\kappa^2(\theta))} - \begin{cases} 0 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{|(\bar{y}_{ik} - y_i(t_k,\theta))|}{\sigma_i(\theta)} - \kappa(\theta) & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \right) \\
& + \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \cdot \begin{cases} -\frac{1}{\sigma_i^2(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ 0 & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \\
& + \left(\frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} + \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \right) \cdot \begin{cases} -2\frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^3(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ -\frac{\mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta))\kappa(\theta)}{\sigma_i^2(\theta)} & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \\
& + \left(\frac{\partial \kappa(\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} + \frac{\partial \kappa(\theta)}{\partial \theta_m} \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \right) \cdot \begin{cases} 0 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{\mathrm{sgn}(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i(\theta)} & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \\
& + \frac{\partial \sigma_i(t_k,\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} \cdot \left(\frac{1}{\sigma_i^2(\theta)} - \begin{cases} 3\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^4(\theta)} & |r_i(t_k,\theta)| \le \kappa(\theta) \\ 2\frac{|\bar{y}_{ik} - y_i(t_k,\theta)|\kappa(\theta)}{\sigma_i^3(\theta)} & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \right) \\
& + \left(\frac{\partial \kappa(\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} + \frac{\partial \kappa(\theta)}{\partial \theta_m} \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \right) \cdot \begin{cases} 0 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ \frac{|\bar{y}_{ik} - y_i(t_k,\theta)|}{\sigma_i^2(\theta)} & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \\
& + \frac{\partial \kappa(\theta)}{\partial \theta_l} \frac{\partial \kappa(\theta)}{\partial \theta_m} \cdot \left(\frac{2\sigma_i^2(\theta) s^2 \exp(-\frac{1}{2}\kappa^2(\theta))}{\kappa(\theta)} \left(\frac{2\exp(-\frac{1}{2}\kappa^2(\theta))}{\kappa^3(\theta)} - \frac{(1 + \frac{2}{\kappa^2(\theta)})}{\sigma_i(\theta) s} \right) \right. \\
& \qquad \left. - \begin{cases} 0 & |r_i(t_k,\theta)| \le \kappa(\theta) \\ 1 & |r_i(t_k,\theta)| > \kappa(\theta) \end{cases} \right) \Bigg].
\end{aligned}
$$

Again the term including the second order sensitivities (line 1) cannot be neglected as it depends on the ratio of κ and σ for large residuals.

1.4 Cauchy distribution

Under the assumption of independent Cauchy distributed measurement noise, the likelihood is

$$
\mathcal{L}_\mathcal{D}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \left[\frac{1}{\pi} \frac{\gamma_i(\theta)}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \right].
$$

Accordingly, the log-likelihood function is given by

$$
\log \mathcal{L}_\mathcal{D}(\theta) = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Big[- \log(\pi) + \log(\gamma_i(\theta)) - \log\left(\left(\bar{y}_{ik} - y_i(t_k,\theta)\right)^2 + \gamma_i(\theta)^2 \right) \Big].
$$

The gradient of the log-likelihood for $l = 1, \dots n_\theta$ is given by

$$
\frac{\partial \log \mathcal{L}_\mathcal{D}(\theta)}{\partial \theta_l} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\left(\frac{1}{\gamma_i(\theta)} - 2\frac{\gamma_i(\theta)}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i^2(\theta)} \right) \frac{\partial \gamma_i}{\partial \theta_l} + 2\frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \right],
$$

and the Hessian matrix for $l, m = 1, \dots, n_\theta$ by

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}_\mathcal{D}(\theta)}{\partial \theta_l \theta_m} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Bigg[& \left(\frac{1}{\gamma_i(\theta)} - 2 \frac{\gamma_i(\theta)}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \right) \frac{\partial^2 \gamma_i(\theta)}{\partial \theta_l \partial \theta_m} \\ & + \left[\frac{4\gamma_i(\theta)^2}{\left((\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2\right)^2} - \frac{1}{\gamma_i(\theta)^2} - \frac{2}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \right] \frac{\partial \gamma_i(\theta)}{\partial \theta_l} \frac{\partial \gamma_i(\theta)}{\partial \theta_m} \\ & - 4 \frac{\gamma_i(\theta)(\bar{y}_{ik} - y_i(t_k,\theta))}{\left((\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i^2(\theta)\right)^2} \left(\frac{\partial \gamma_i(\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} + \frac{\partial \gamma_i(\theta)}{\partial \theta_m} \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \right) \\ & + \frac{2}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \left(\frac{2(\bar{y}_{ik} - y_i(t_k,\theta))^2}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} - 1 \right) \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \\ & + 2 \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{(\bar{y}_{ik} - y_i(t_k,\theta))^2 + \gamma_i(\theta)^2} \frac{\partial^2 y_i(t_k,\theta)}{\partial \theta_l \partial \theta_m} \Bigg],\end{aligned}$$

with $l, m = 1, \dots n_\theta$. Assuming that the deviation between measurement and observable is small, we can again neglect the second-order sensitivities. This provides an approximation which only depends on the first-order sensitivities.

1.5 Student's t distribution

Under the assumption of independent Student's t distributed measurement noise, the likelihood is

$$\mathcal{L}_\mathcal{D}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \left[\frac{\Gamma\left(\frac{\nu_i(\theta)+1}{2}\right)}{\sqrt{\nu_i(\theta)\pi}\ \sigma_i(\theta)\ \Gamma\left(\frac{\nu_i(\theta)}{2}\right)} \left(1 + \frac{1}{\nu_i(\theta)} \left(\frac{\bar{y}_{ik} - y_i(t_k,\theta)}{\sigma_i(\theta)} \right)^2 \right)^{-\frac{\nu_i(\theta)+1}{2}} \right].$$

The log-likelihood function is given by

$$\log \mathcal{L}_\mathcal{D}(\theta) = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \left[\log \left(\frac{\Gamma\left(\frac{\nu_i(\theta)+1}{2}\right)}{\sqrt{\nu_i(\theta)\pi}\ \sigma_i(\theta)\ \Gamma\left(\frac{\nu_i(\theta)}{2}\right)} \right) - \frac{\nu_i(\theta)+1}{2}\ \log \left(1 + \frac{1}{\nu_i(\theta)} \left(\frac{\bar{y}_{ik} - y_i(t_k,\theta)}{\sigma_i(\theta)} \right)^2 \right) \right].$$

The gradient of the log-likelihood for $l = 1, \dots n_\theta$ is given by

$$\begin{aligned}\frac{\partial \log \mathcal{L}_\mathcal{D}(\theta)}{\partial \theta_l} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Bigg[& \frac{1}{2} \Bigg[\psi\left(\frac{\nu_i(\theta)+1}{2}\right) - \psi\left(\frac{\nu_i(\theta)}{2}\right) - \log\left(1 + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_i(\theta)\sigma_i(\theta)^2} \right) \\ & - \frac{1}{\nu_i(\theta)} + \frac{\nu_i(\theta)+1}{\nu_i^2(\theta)\sigma_i^2(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \Bigg] \frac{\partial \nu_i(\theta)}{\partial \theta_l} \\ & - \left[\frac{1}{\sigma_i(\theta)} - \frac{\nu_i(\theta)+1}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_i(\theta)\sigma_i^3(\theta)} \right] \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \\ & + \frac{\nu_i(\theta)+1}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^2(\theta)} \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \Bigg],\end{aligned}$$

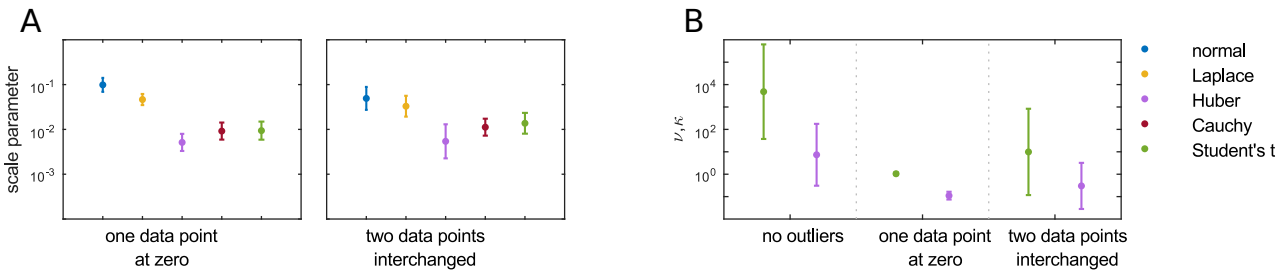
where ψ denotes the digamma function, which is the logarithmic derivative of the gamma function. The Hessian matrix is consequently for $l, m = 1, \dots, n_\theta$

$$
\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_{\mathcal{D}}(\theta)}{\partial \theta_l \partial \theta_m} = \sum_{k=1}^{n_t} \sum_{i=1}^{n_y} \Bigg[& \frac{1}{2} \Bigg[\psi\Big(\frac{\nu_i(\theta)+1}{2}\Big) - \psi\Big(\frac{\nu_i(\theta)}{2}\Big) - \log\left(1 + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_i(\theta)\sigma_i(\theta)^2}\right) \\
& - \frac{1}{\nu_i(\theta)} + \frac{\nu_i(\theta)+1}{\nu_i^2(\theta)\sigma_i^2(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \Bigg] \frac{\partial^2 \nu_i(\theta)}{\partial \theta_l \partial \theta_m} \\
& - \left[\frac{1}{\sigma_i(\theta)} - \frac{\nu_i(\theta)+1}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_i(\theta)\sigma_i^3(\theta)} \right] \frac{\partial^2 \sigma_i(\theta)}{\partial \theta_l \partial \theta_m} \\
& + \frac{\nu_i(\theta)+1}{1 + \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \frac{1}{\nu_i(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^2(\theta)} \frac{\partial^2 y_i(t_k,\theta)}{\partial \theta_l \partial \theta_m} \\
& + \frac{1}{2} \Bigg[\frac{1}{2} \psi_1\Big(\frac{\nu_i(\theta)+1}{2}\Big) - \frac{1}{2} \psi_1\Big(\frac{\nu_i(\theta)}{2}\Big) + \frac{1}{\nu_i^2(\theta)} \\
& \quad + \frac{1}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} \left(\frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2} - 1}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i(\theta)^2}} - \frac{1}{\nu_i(\theta)} \right) \\
& \quad \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_i(\theta)\sigma_i^2(\theta)} \Bigg] \frac{\partial \nu_i(\theta)}{\partial \theta_l} \frac{\partial \nu_i(\theta)}{\partial \theta_m} \\
& + \Bigg[\frac{1}{\sigma_i^2(\theta)} + \frac{\nu_i(\theta)+1}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^4(\theta)} \\
& \quad \left(2 \frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} - \frac{3}{\sigma_i(\theta)} \right) \Bigg] \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} \\
& + \frac{\nu_i(\theta)+1}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} \frac{1}{\sigma_i^2(\theta)} \left(2 \frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} - 1 \right) \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \\
& + 2 \frac{\nu_i(\theta)+1}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} \left(\frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}}{\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)}} - 1 \right) \\
& \quad \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^3(\theta)} \left(\frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} + \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \right) \\
& + \frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)} - 1}{(\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)})^2} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^3(\theta)} \left(\frac{\partial \nu_i(\theta)}{\partial \theta_l} \frac{\partial \sigma_i(\theta)}{\partial \theta_m} + \frac{\partial \nu_i(\theta)}{\partial \theta_m} \frac{\partial \sigma_i(\theta)}{\partial \theta_l} \right) \\
& + \frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)} - 1}{(\nu_i(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)})^2} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_i^2(\theta)} \left(\frac{\partial \nu_i(\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} + \frac{\partial \nu_i(\theta)}{\partial \theta_m} \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \right) \Bigg],
\end{aligned}
$$

where ψ_1 is the trigamma function, the derivative of the digamma function. Assuming that the deviation between measurement and observable is small, we can again neglect the second-order sensitivities (line 4). This provides an approximation which only depends on the first-order sensitivities.

Supplement Table 1: Upper and lower parameter bounds used for the parameter estimation as well as the true parameter values for which the data was generated.

parameter	lower bound	upper bound	true value
k_1	$10^{-3.5}$	10	$10^{-1.5}$
k_2	$10^{-3.5}$	10	$10^{-1.5}$
σ_n	10^{-5}	1	$10^{-1.7}$
b	10^{-5}	1	-
γ	10^{-5}	1	-
σ_t	10^{-5}	1	-
ν	1	10^5	-
σ_h	10^{-5}	1	-
κ	10^{-1}	10^5	-



Supplement Figure 1: **Distribution parameters for the conversion reaction (A)** Mean ± standard deviation for the maximum likelihood estimates of the scale parameters in the outlier scenarios. **(B)** Estimated ν for the Student's t distribution assumption and estimated κ for the Huber distribution assumption.

2 Conversion reaction

In this section, we provide the details for the parameter estimation and statistical evaluation for the simulation study of a conversion process.

2.1 Parameter estimation

For solving the optimization problem we used a trust-region reflective algorithm for the normal, the Cauchy and the Student's t distribution employing the gradient and an approximation of the Hessian. As the second-order derivative for the Laplace distribution requires computationally expensive simulations of second-order sensitivities (see Suppl. Section 1.2), we used an interior-point algorithm without a user-supplied Hessian for the Laplace and Huber distribution assumption. The parameter bounds used for parameter estimation are provided in Suppl. Table S1. The bounds for the degrees of freedom for the Student's t distribution are chosen such that the distribution approaches the Cauchy distribution for the lower bound of ν and the normal distribution for the upper bound. The initial parameter values for the optimization were obtained by Latin hypercube sampling in the allowed parameter range. The estimates for the distribution parameters are visualized in Figure S1.

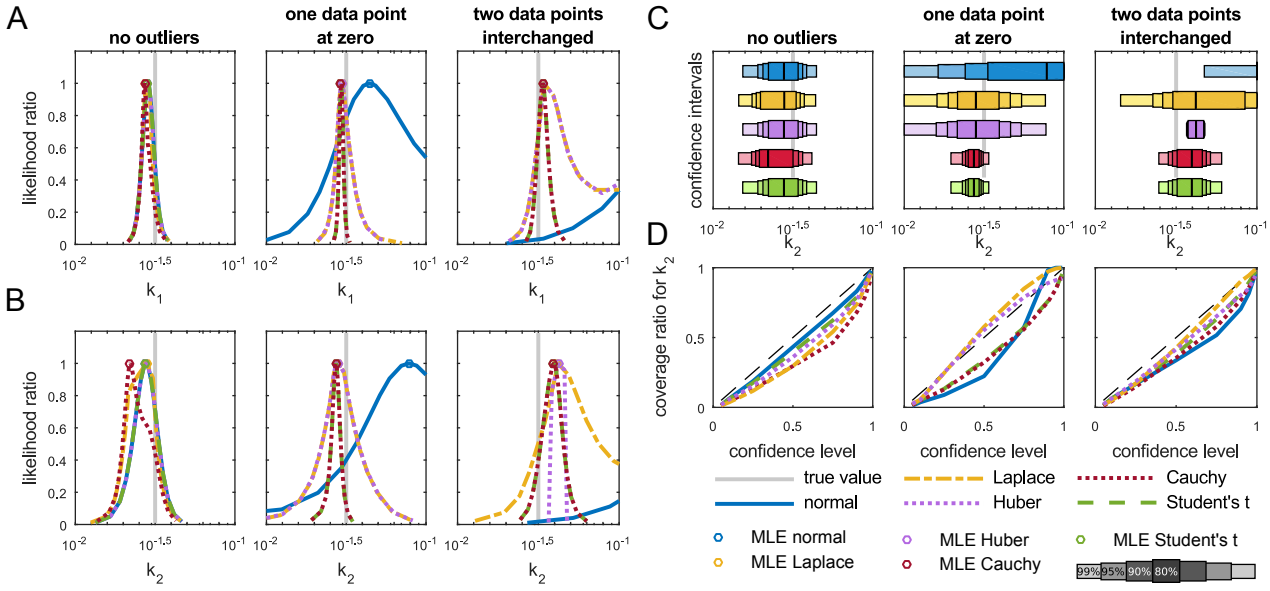
2.2 Confidence intervals

We used profile likelihoods to assess the uncertainty of the parameter estimates. The profile likelihoods were computed with the toolbox PESTO[1]. Figure S2A and B show the normalized profile likelihoods for parameters k_1 and k_2 for the cases displayed in Figure 2A,D and E. The true parameter value is indicated by a vertical grey line. For the *no outliers* scenario all profiles overlap and are close to the true parameter values for k_1 and k_2. The profiles for the Cauchy and Student's t distribution assumption remain similarly narrow for the two scenarios with outliers. The profiles for the Laplace distribution widen in the case of *two data points interchanged*, but the MLE is still close to the true value, whereas the profile of the normal distribution assumption broadens and the MLEs

[1]available at `https://github.com/ICB-DCM`

Supplement Table 2: Confidence Intervals that contain the true value $\log_{10}(k_1) = -1.5$

	no outliers		one data point at zero		two data points interchanged	
	$(1-\alpha)\%$	CI	$(1-\alpha)\%$	CI	$(1-\alpha)\%$	CI
normal	80%	[-1.5911,-1.4966]	80%	[-1.6217,-0.8848]	-	-
Laplace	90%	[-1.6059,-1.4966]	80%	[-1.5715,-1.4574]	80%	[-1.5297,-1.2549]
Huber	80%	[-1.5911,-1.4966]	80%	[-1.5716,-1.4572]	80%	[-1.5301,-1.2565]
Cauchy	95%	[-1.6159,-1.4874]	99%	[-1.5749,-1.4870]	80%	[-1.5150,-1.4298]
Student's t	80%	[-1.5911,-1.4966]	99%	[-1.5749,-1.4870]	80%	[-1.5150,-1.4298]



Supplement Figure 2: **Uncertainty analysis for the conversion reaction.** Profiles for k_1 **(A)** and k_2 **(B)**. **(C)** Confidence intervals $(80\%, 90\%, 95\%, 99\%)$ for parameter k_2 in all scenarios. **(D)** Coverage ratio for parameter k_2 compared to the confidence level.

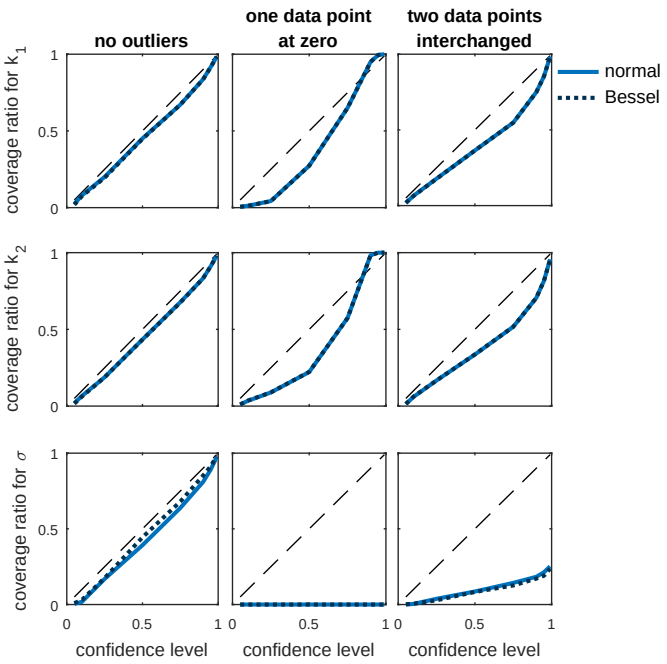
move away from the true parameter values. The smallest CIs that contain the true value for k_1 for each scenario are given in Suppl. Table S2. In the case of *two data points interchanged* the true value of k_1 is not located within one of the computed intervals for the normal distribution. Figure S2C illustrates the confidence intervals for k_2 and Figure S2D shows the corresponding coverage ratios.

2.3 Bessel correction

Since the coverage ratios were too small we considered the Bessel correction. The Bessel correction yields an unbiased estimate of the likelihood. The corresponding likelihood function is

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} \left[\frac{1}{\sqrt{2\pi}\ \sigma_i(\theta)} \exp\Big(- \frac{n_t}{2(n_t-1)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_i^2(\theta)} \Big) \right].$$

We estimated the parameters using this modified likelihood function and computed the confidence intervals. The correction did neither improve greatly the coverage ratios for the kinetic parameters nor for the distribution parameter σ_n (Figure S3). We note that the estimates for the standard deviation σ_n and the coverage of the corresponding confidence intervals is off in the scenarios one data point at zero and two data points interchanged as explanation of outliers requiring high noise levels. Therefore, we did not apply the Bessel correction.

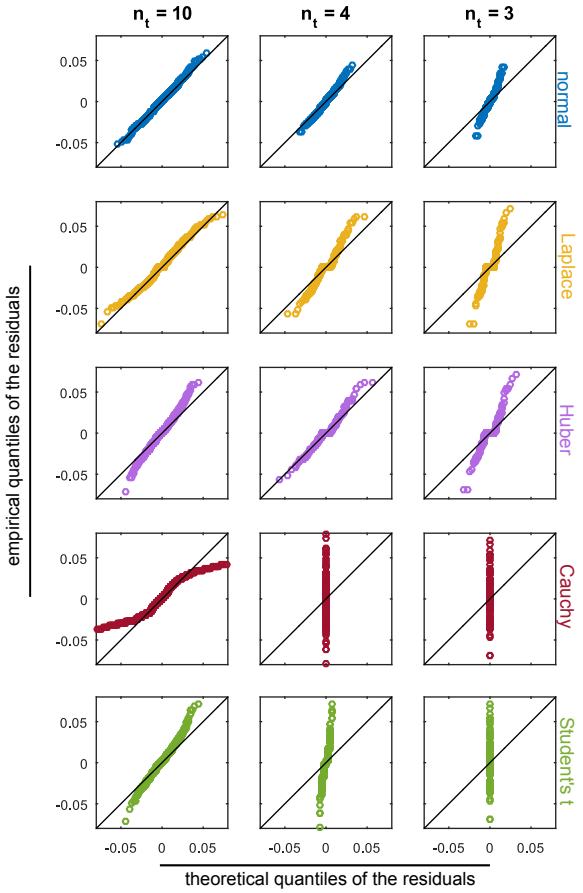


Supplement Figure 3: **Coverage ratios using the normal distribution assumption with and without Bessel correction.** Coverage ratios for the kinetic parameters k_1 and k_2 and for the distribution parameter σ_n.

2.4 Sample size limitation of the Cauchy and the Student's t distribution

In addition to the density plots provided in Figure 5, we evaluated sample size effects using a Q-Q plot (Figure S4). In the Q-Q plot (quantile-quantile-plot) the empirical quantiles of the residuals are compared with the theoretical quantiles, considering the quantiles $(i - 0.5)/n$, $i = 1, \ldots, n$, where n is the sample size. If the distributions correspond to each other the values are located at the diagonal line of the Q-Q plot. The theoretical quantiles were computed by means of the inverse cumulative distribution function of the respective distribution using the estimated mean value of the distribution parameters.

For 10 data points the empirical and theoretical quantiles coincide well for all distributions, except for the heavier-tailed distributions in the tails. This indicates that the sample has shorter tails than the theoretical distribution. In the cases of three and four data points the Q-Q plots for the Cauchy and the Student's t distribution show that the distributions do not reflect the spread in the residuals. In case of $n_t = 3$ also problems for the normal, Laplace and Huber distribution are visible. The sample size is apparently too small for reasonable parameter estimation.



Supplement Figure 4: **QQ-plots for the residuals.** The empirical quantiles of the residuals are plotted against the theoretical quantiles (using the mean scale parameters) for the normal, the Laplace, Huber, Cauchy and Student's t distribution for different numbers of data points in the *no outliers* datasets.

3 JAK/STAT signaling

3.1 ODE model for JAK/STAT signaling

The ODE system is given by (Schelker et al., 2012)

$$
\begin{aligned}
\frac{d[\text{STAT}]}{dt} &= \frac{1}{\Omega_{\text{cyt}}}(\Omega_{\text{nuc}}\,[\text{nSTAT5}]\,p_4 - \Omega_{\text{cyt}}\,[\text{STAT}]\,p_1\,u) \\
\frac{d[\text{pSTAT}]}{dt} &= -\frac{1}{[\text{STAT}_0]}(2\,p_2\,[\text{pSTAT}]^2 - [\text{STAT}]\,p_1\,u) \\
\frac{d[\text{pSTAT\_pSTAT}]}{=} & \frac{1}{[\text{STAT}]_0}(p_2\,[\text{pSTAT}]^2 - [\text{STAT}_0]\,p_3\,[\text{pSTAT\_pSTAT}]) \\
\frac{d[\text{nSTAT1}]}{dt} &= -\frac{p_4}{\Omega_{\text{nuc}}}\,(\Omega_{\text{cyt}}\,[\text{STAT}] - \Omega_{\text{cyt}}\,[\text{STAT}]_0 + 2\;\Omega_{\text{nuc}}\,[\text{nSTAT1}] \\
&\quad + \Omega_{\text{nuc}}\,[\text{nSTAT2}] + \Omega_{\text{nuc}}\,[\text{nSTAT3}] + \Omega_{\text{nuc}}\,[\text{nSTAT4}] \\
&\quad + \Omega_{\text{nuc}}\,[\text{nSTAT5}] + \Omega_{\text{cyt}}\,[\text{pSTAT}] + 2\;\Omega_{\text{cyt}}\,[\text{pSTAT\_pSTAT}]) \\
\frac{d[\text{nSTAT2}]}{dt} &= p_4\,([\text{nSTAT1}] - [\text{nSTAT2}]) \\
\frac{d[\text{nSTAT3}]}{dt} &= p_4\,([\text{nSTAT2}] - [\text{nSTAT3}]) \\
\frac{d[\text{nSTAT4}]}{dt} &= p_4\,([\text{nSTAT3}] - [\text{nSTAT4}]) \\
\frac{d[\text{nSTAT5}]}{dt} &= p_4\,([\text{nSTAT4}] - [\text{nSTAT5}])
\end{aligned}
$$

with initial conditions

$$x(t_0) = ([\text{STAT}]_0, [\text{pSTAT}]_0, [\text{pSTAT\_pSTAT}]_0, [\text{nSTAT1}]_0, [\text{nSTAT2}]_0, [\text{nSTAT3}]_0, [\text{nSTAT4}]_0, [\text{nSTAT5}]_0)^T\,,$$

kinetic parameters $p_1, \ldots, p_4$ and initial concentration $[\text{STAT}]_0$. The delay reaction of STAT binding to the DNA in the nucleus is modeled as linear chain approximation with intermediate steps $\text{nSTAT1}, \ldots, \text{nSTAT5}$. The volume of the two compartments, cytoplasm and nucleus, are constants; $\Omega_{\text{cyt}} = 1.4\,\text{pl}$ and $\Omega_{\text{nuc}} = 0.45\,\text{pl}$ (Raue et al., 2009). The observables are given by

$$
\begin{aligned}
y_1 &= \text{o}_{\text{pSTAT}} + \frac{\text{s}_{\text{pSTAT}}}{[\text{STAT}]_0}\,([\text{pSTAT}] + 2[\text{pSTAT\_pSTAT}]) \\
y_2 &= \text{o}_{\text{tSTAT}} + \frac{\text{s}_{\text{tSTAT}}}{[\text{STAT}]_0}\,([\text{STAT}] + [\text{pSTAT}] + 2[\text{pSTAT\_pSTAT}]) \\
y_3 &= u\,.
\end{aligned}
$$

The observable y_1 provides information about the total concentration of phosphorylated STAT in the cytoplasm (pSTAT), y_2 about the total concentration of STAT in the cytoplasm (tSTAT) and y_3 about the concentration of phosphorylated Epo receptors (pEpoR). The pEpoR concentration is modeled as time-dependent cubic spline function u with five parameters $sp_1, \ldots, sp_5$. The scale parameters s_{pSTAT} and s_{tSTAT} were introduced by (Swameye et al., 2003) because only relative protein amounts could be measured by the experimental setup. The offset parameters o_{pSTAT} and o_{tSTAT} account for the background noise. The initial concentration $[\text{STAT}]_0$ was set to 1 as in (Schelker et al., 2012) in order to tackle structural identifiability problems shown in (Raue et al., 2009). This leads to the structurally identifiable parameter vector

$$\theta = (p_1, p_2, p_3, p_4, sp_1, sp_2, sp_3, sp_4, sp_5, \text{o}_{\text{tSTAT}}, \text{o}_{\text{pSTAT}}, \text{s}_{\text{tSTAT}}, \text{s}_{\text{pSTAT}})^T.$$

3.2 Parameter estimation

The parameter space for the multi-start optimization is shown in Suppl. Table S3. The distribution parameters were chosen separately for each observable except of the degrees of freedom for the Student's t distribution and the tuning parameter for the Huber distribution. For the Student's t distribution the choice is based on model

selection performed for the *no outliers* scenario using the BIC. The model with three ν parameters (with lower bounds $(1.3, 0.8, 0.8)^T$) yields a BIC value of -87.95, whereas the model with one ν (with lower bound 2) gives a BIC of -91.91. Hence, the model with one parameter for ν is more appropriate as it describes the data equally well by employing less parameters. Only in four cases the lower bound of the degrees of freedom had to be increased to 2.2 according to the criterion by (Fernández and Steel, 1999) in order to address the overfitting problem. 100 multi-starts were generated with Latin hypercube sampling using the toolbox PESTO. As local solver the interior-point algorithm was used for the normal, Laplace and the Huber distribution and the trust-region algorithm for the Student's t distribution, which are included in the MATLAB function `fmincon`. If less than 5 starts have converged to the same optimum, the number of start points was increased by 100. The estimation results for the *no outliers* scenario are listed in Table S4. The estimated parameter values are similar for all used distribution assumptions.

Figure S6A-B displays the logarithm of the MSE for the spline parameters with errorbars indicating the 95% percentile bootstrap intervals. For *one data point at zero* the normal distribution leads to a higher MSE for all parameters. For *two data points interchanged* this is not true for all parameters. Note that not all cases of this scenario lead to clear outliers and consequently, the normal distribution is able to adequately describe some of the cases, which leads to a smaller overall error. The percentage of converged starts out of 100 start points of the different estimators is shown in Figure S6C. For the analysis only converged results were used. If less than five starts converged to the same optimum the number of starts was increased by 100. The number of start points never needed to be increased higher than 400, which is still in a reasonable range.

4 Pom1p gradient formation in fission yeast

As further application example, we analyzed the different distribution assumptions for experimental data of Pom1 gradient formation. Pom1 plays an essential role in cell growth and cell cycle. The data comprises single-cell imaging data, including a mean intensitiy curve and fluorescence recovery after photobleaching (FRAP) measurements (Saunders et al., 2012). We consider the partial differential equation model as introduced by (Hersch et al., 2015) and adapt the parameter estimation problem formulated by (Hross et al., 2016).

4.1 PDE model for Pom1p gradient formation

We used the multiple-site phosphorylation model from (Hersch et al., 2015) implemented in (Hross et al., 2016), which is given by

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial d^2} - \alpha u^2 + \frac{J}{\sqrt{2\pi}\rho} e^{-d^2/s\rho^2} \,,$$

with initial condition

$$u(0,d) = \begin{cases} 0 & \text{for } d \in Q \\ u^\infty(d;\theta) & \text{otherwise} \end{cases} , \tag{1}$$

where $u^\infty(d;\theta)$ is the steady state of the model and $Q \subseteq \Omega$ denotes the bleached region. Furthermore, no-flux boundary conditions are assumed at the division axis of the cells. The model was simulated using the method of lines, which yields systems of ODEs.

4.2 Experimental data

For our study we calibrated the PDE model based on the normalized intensity profile along the membrane in the unperturbed system, modeled as

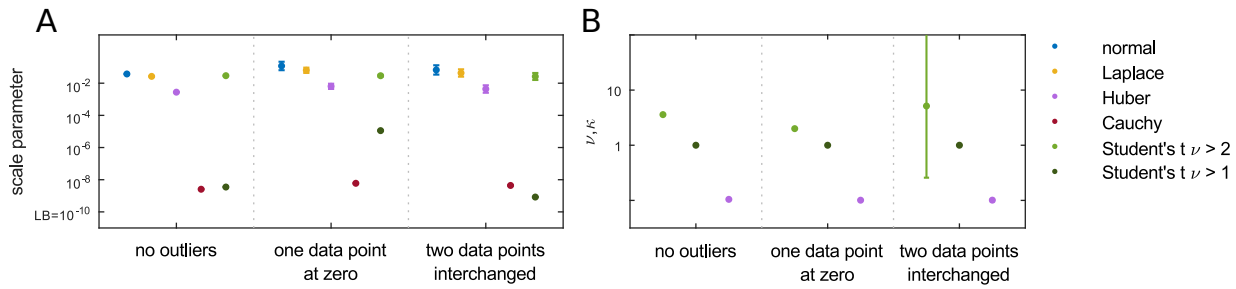
$$y_1(\theta, t_k) = s_1 u^\infty(d_k;\theta)$$

and the fluorescence recovery after photo bleaching (FRAP) of the full ($Q_2 = [-2.75, 2.75]\mu m$) and the half tip ($Q_3 = [0, 2.75]\mu m$), modeled as

$$y_i(\theta, t_k) = s_i \int_{Q_i} u(d, t_k; \theta, Q_i) dd, \quad i \in 2, 3 \,,$$

with u denoting the solution for u with (1) and $Q = Q_i$.

We neglect the reported standard error of means and estimated the scale parameters for each distribution. This yields the model parameters $\xi = (D, \alpha, J, \rho, s_1, s_2, s_3)$ that are estimated from the data.



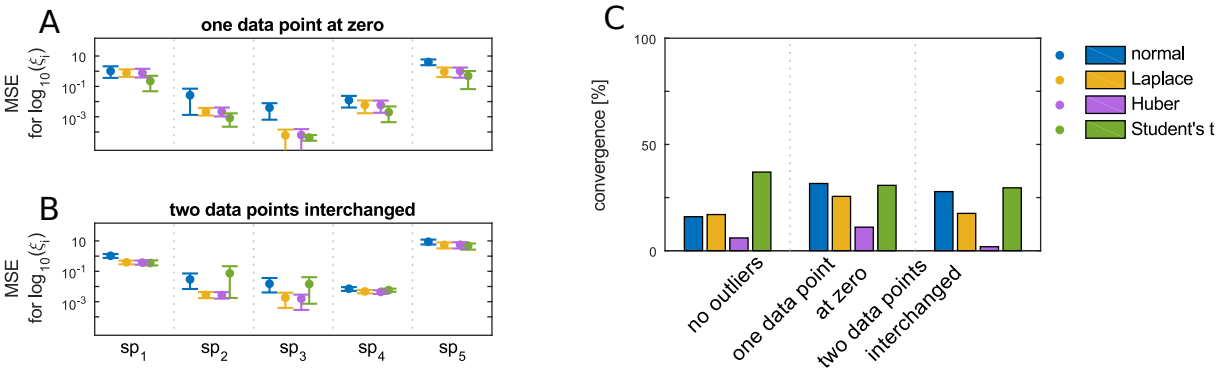
Supplement Figure 5: **Distribution parameter values for the JAK/STAT signaling pathway. (A)** Mean ± standard deviation of the maximum likelihood estimates for the scale parameter for observable pSTAT. **(B)** Maximum likelihood estimates for ν of the Student's t distribution for $\nu^{\text{LB}} = 1$ and $\nu^{\text{LB}} = 2$.

Supplement Table 3: **Parameter bounds for the model of JAK/STAT signaling.** Lower bounds (LB) and upper bounds (UB) for the estimated parameters are provided.

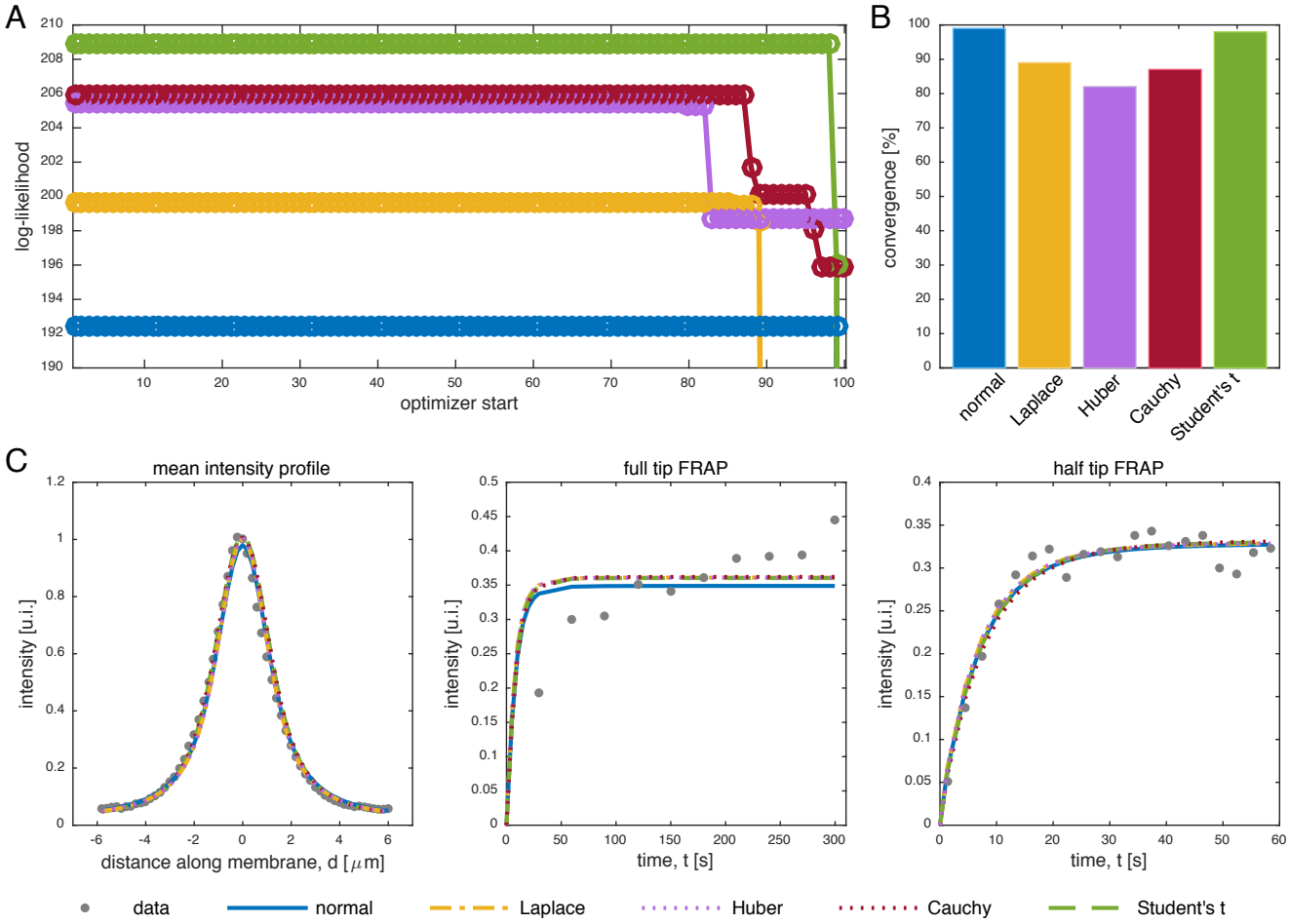
	LB	UB		LB	UB		LB	UB
p_1	10^{-5}	10^{3}	sp_4	10^{-5}	10^{3}	$\sigma_{n1}, b_1, \sigma_{h1}, \sigma_{t1}$	10^{-5}	10^{3}
p_2	10^{-3}	10^{6}	sp_5	10^{-6}	10^{3}	$\sigma_{n2}, b_2, \sigma_{h2}, \sigma_{t2}$	10^{-5}	10^{3}
p_3	10^{-5}	10^{3}	o_{tSTAT}	10^{-5}	10^{3}	$\sigma_{n3}, b_3, \sigma_{h3}, \sigma_{t3}$	10^{-5}	10^{3}
p_4	10^{-3}	10^{6}	o_{pSTAT}	10^{-5}	10^{3}	ν	2	10^{5}
sp_1	10^{-5}	10^{3}	s_{tSTAT}	10^{-5}	10^{3}	κ	10^{-1}	10^{5}
sp_2	10^{-5}	10^{3}	s_{pSTAT}	10^{-5}	10^{3}			
sp_3	10^{-5}	10^{3}						

Supplement Table 4: **Estimation results for the model of JAK/STAT signaling.** The individual columns provide the results for the *no outliers* scenario on the $\log_{10}$-scale obtained using the normal (N), the Laplace (L), the Huber (H) and the Student's t (T) distribution.

	N	L	H	T		N	L	H	T
p_1	0.6026	0.5996	0.5996	0.6444	sp_4	-0.4073	-0.4217	-0.4217	-0.4445
p_2	5.9997	4.2804	4.2804	6.0000	sp_5	-5.4630	-4.3631	-4.3631	-4.9585
p_3	-0.9549	-0.9773	-0.9773	-0.9728	o_{tSTAT}	-0.7318	-1.1440	-1.1440	-0.8276
p_4	-0.0111	0.0271	0.0271	-0.0243	o_{pSTAT}	-0.6541	-0.6645	-0.6645	-0.6420
sp_1	-2.8096	-2.6355	-2.6355	-2.8835	s_{tSTAT}	-0.1086	-0.0324	-0.0324	-0.0782
sp_2	-0.2557	-0.2654	-0.2654	-0.2755	s_{pSTAT}	-0.0100	-0.0070	-0.0070	-0.0030
sp_3	-0.0765	-0.0652	-0.0652	-0.0618					



Supplement Figure 6: **MSE of spline parameters and convergence for outlier scenarios of JAK/STAT signaling.** The bootstrap confidence intervals for logarithm of the MSE for the spline parameters are shown for **(A)** *one data point at zero* and **(B)** *two data points interchanged*. Convergence for the different outlier scenarios is displayed in **(C)**.



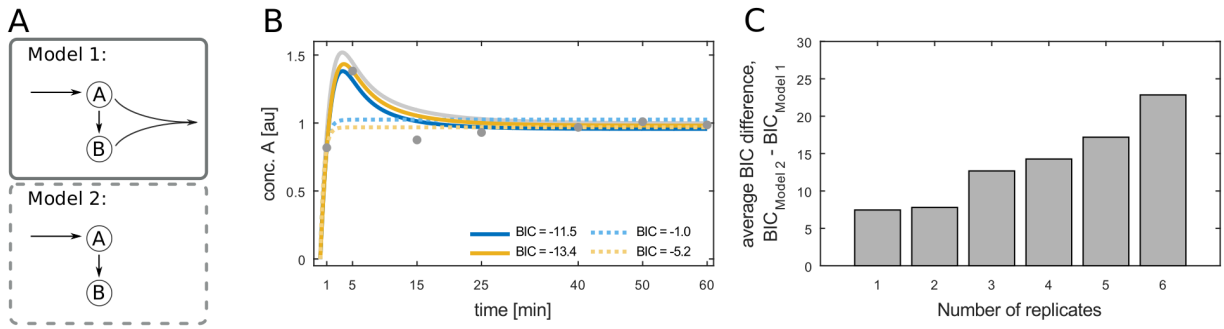
Supplement Figure 7: **Optimization results for the model of Pom1p gradient formation. (A)** Convergence using the different distribution assumptions for 100 multi-starts. **(B)** Percentage of converged starts. **(C)** Measurement data and fits: mean intensity profile, full and half tip FRAP.

4.3 Parameter estimation

We estimated the model parameters assuming the normal, the Laplace, the Huber, the Cauchy and the Student's t distribution. All distributions yield reasonable convergence (Figure S7A and B) and provide good fits to the measurement data (Figure S7C). Model selection rejects the normal distribution, as the difference in BIC is larger than 10 to all other statistical models, indicating that the normality assumption is not ideal for this dataset. This shows the applicability of the heavier-tailed distributions to more complex models, also in the absence of outliers.

5 Illustrative example for model selection using heavier tailed distributions

As we have observed in the case of the conversion reaction, heavier-tailed distributions are selected in presence of outliers and thus, model selection can be used to detect outlier corrupted datasets. Statistical modeling with heavier-tailed distributions tends to provide a fit to the majority of the data instead of trying to distribute the error equally. Therefore, the suspicion arises that a wrong model could be chosen if the model structure is unknown by treating data points as outliers. In this section we analyze this issue of model misspecification based on a simple example. We artificially generate a dataset for a model (Model 1, Figure S8A). To this dataset we calibrate the true model (Model 1) with three kinetic parameters as well as a simpler model (Model 2, Figure S8A) with two kinetic parameters. Although Model 2 is able to fit all but one data point, model selection favors in both cases, normal and Laplace, the more complex Model 1 (Figure S8B). For the Laplace distribution the difference in BIC is smaller than 10, thus the typical rejection criterion is not met. Hence, it is difficult to discriminate between outliers and a



Supplement Figure 8: **Model Selection with heavier-tailed distributions.** (**A**) Schematic representation of the model pathways. (**B**) Trajectories for one example dataset (•) with true trjectory (—) using the normal (—) and the Laplace distribution (—) for Model 1 (solid line) and Model 2 (dashed line). (**C**) The average difference in BIC values for 10 datasets with different number of replicates using the Laplace distribution.

wrong model.

However, the data points regarded as outliers possess a larger residual for the Laplace fit than in the normal fit. Consequently, it is more evident that the model does not account for these data points. These data points are good starting points for further examination, e.g. experimental validation of the measured value. If more replicates are available and used for model calibration, the average difference in BIC values increases and allows a more decisive model selection (Figure S8C).

References

Fernández, C. and Steel, M. F. (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika*, 86(1):153–167.

Hersch, M., Hachet, O., Dalessi, S., Ullal, P., Bhatia, P., Bergmann, S., and Martin, S. G. (2015). Pom1 gradient buffering through intermolecular auto-phosphorylation. *Mol. Syst. Biol.*, 11(7):818.

Hross, S., Fiedler, A., Theis, F. J., and Hasenauer, J. (2016). Quantitative comparison of competing pde models for pom1p dynamics in fission yeast. To appear in *Proceedings 6th IFAC Conference on Foundations of Systems Biology in Engineering*.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinf.*, 25(25):1923–1929.

Saunders, T. E., Pan, K. Z., Angel, A., Guan, Y., Shah, J. V., Howard, M., and Chang, F. (2012). Noise reduction in the intracellular pom1p gradient by a dynamic clustering mechanism. *Devel. Cell*, 22(3):558–572.

Schelker, M., Raue, A., Timmer, J., and Kreutz, C. (2012). Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinf.*, 28(18):i529—i534.

Swameye, I., Müller, T. G., Timmer, J., Sandra, O., and Klingmüller, U. (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci. U S A*, 100(3):1028–1033.