Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Systems Biology

Robust parameter estimation for dynamical systems from outlier-corrupted data

Corinna Maier [1,2], Carolin Loos [1,2], 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

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Dynamics of cellular processes are often studied using mechanistic mathematical models. These models possess unknown parameters which are generally estimated from experimental data assuming normally distributed measurement noise. Outlier corruption of datasets often cannot be avoided. These outliers may distort the parameter estimates, resulting in incorrect model predictions. Robust parameter estimation methods are required which provide reliable parameter estimates in the presence of outliers.
Results: In this manuscript, we propose and evaluate methods for estimating the parameters of ordinary differential equation (ODE) models from outlier-corrupted data. As alternatives to the normal distribution as noise distribution, we consider the Laplace, the Huber, the Cauchy and the Student's t distribution. We assess accuracy, robustness and computational efficiency of estimators using these different distribution assumptions. To this end, we consider artificial data of a conversion process, as well as published experimental data for Epo-induced JAK/STAT signaling. We study how well the methods can compensate and discover artificially introduced outliers. Our evaluation reveals that using alternative distributions improves the robustness of parameter estimates.
Availability: The MATLAB implementation of the likelihood functions using the distribution assumptions is available at *Bioinformatics* online.
Contact: jan.hasenauer@helmholtz-muenchen.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Quantitative dynamic models are widely used to gain a mechanistic understanding of biological processes (Ideker et al., 2001; Kitano, 2002). These dynamic models facilitate the integration of multiple experimental datasets and the analysis of system properties that are not within reach of biological experiments (Aderem, 2005). For this, the models need to be calibrated based on experimental data in order to determine the unknown parameters, e.g. initial values or kinetic rates (Tarantola, 2005).

Experimental data used for parameter estimation are collected using a broad spectrum of techniques. While measurement devices provide increasingly precise quantitative data (Chen et al., 2013), there are numerous potential sources of measurement errors during data collection and data processing (Ghosh and Vogt, 2012). These include technical limitations and human errors, such as pipetting errors or incorrect labeling, which result in potentially large errors (Motulsky and Christopoulos, 2003). Individual data points which are corrupted by large errors are usually denoted as outliers and assumed to be generated from a different mechanism as the remainder of the data points and might be misleading in the further analysis (Hawkins, 1980; Tarantola, 2005). Therefore, parameter estimation using outlier-corrupted data can result in large estimation errors and limits the validity of models.

Since outlier-corrupted data distorts results in various fields, many methods for the detection and subsequent removal of outliers have been developed (Hodge and Austin, 2004; Ben-Gal, 2005; Niu et al., 2011).

Most of the algorithms either assign a score for the degree of abnormality or a binary label to a data point. This labeling is usually based on a fit to a distribution or distance measure e.g. k-nearest neighbor distance (Ramaswamy et al., 2000). Eventually, it however remains a subjective decision on whether or not a data point is sufficiently abnormal to be removed (Aggarwal, 2015). Noisy measurements complicate the distinction even more and the increasing size and complexity of biological data make the removal of outliers a challenging task. Furthermore, the elimination of data points which are indeed no outliers, as well as the retention of outliers in the data, will yield less reliable results in the further analysis (Motulsky and Christopoulos, 2003).

In the fields of regression (Lange et al., 1989; Peel and McLachlan, 2000) and computer vision (Stewart, 1999) robust estimation methods are used to circumvent the removal of data points. These robust approaches exploit estimators that are less affected by outliers than the standard approach, the least squares estimator. Well known maximum-likelihood type estimators (M-estimators) (Press et al., 1988), which were found to be robust to outliers are, for example, the least absolute deviation estimator (Tarantola, 2005) and the Huber M-estimator (Huber et al., 1964). These estimators essentially use lower weights for data points with large residuals. In addition, Student's t regression models were studied, which assume Student's t distributed errors (Fernández and Steel, 1999).

The methods developed in the field of robust regression can in principle be applied across scientific fields. Each field has, however, its particularities regarding experimental data, e.g. noise levels, outlier generating mechanisms, and mathematical models which influence the performance. For dynamical models of biological systems, the Huber M-estimator was already successfully applied, yielding more reliable parameter estimates (Cao et al., 2011; Qiu et al., 2016). A comprehensive evaluation of different methods in the field of quantitative biology is, however, missing. Furthermore, the standard formulation as regression problem does not allow in a straightforward way to perform model selection using statistical methods such as the likelihood ratio test (Wilks, 1938), the Akaike (Akaike, 1973) or the Bayesian information criterion (Schwarz, 1978). To facilitate model selection for the mechanistic as well as the statistical model, a formulation of robust estimation in terms of (normalized) probability distributions would be beneficial.

In this manuscript we consider a comprehensive selection of statistical models for the residual distribution, assuming distributions with heavier tails than the generally used normal distribution. These statistical models correspond to a range of robust estimators. We derive the analytic gradients and Hessian matrices of the resulting objective functions, which are required for an efficient optimization (Raue et al., 2013; Hross and Hasenauer, 2016). The formulation in terms of probability distributions facilitates model selection and the estimation of tuning parameters, e.g. for the Huber M-estimator. We systematically assess and evaluate the properties of the resulting estimation in the absence and presence of outliers. The efficiency and robustness of the methods could statistically be evaluated for generated artificial data of a conversion process as the true parameters were known. Additionally, we applied our method to artificially perturbed experimental data of the JAK/STAT signaling pathway.

2 Methods

In this section, we propose methods for the robust estimation of parameters of biological processes from outlier-corrupted data. We introduce the considered dynamical and statistical models along with optimization and model selection methods. Additionally, we present three outlier scenarios.

Fig. 1. Three scenarios used for the data generation. No outliers: the data is not outlier-corrupted and any deviation is due to measurement noise. One data point at zero: the data has one outlier, which is a zero-entry. Two data points interchanged: the data has two outliers, which are two measurements that were interchanged. The arrows illustrate how the outliers were introduced in the dataset.

2.1 Data-driven modeling of dynamic biological systems

We consider biological processes, for which the dynamics are modeled by ODEs, e.g. reaction rate equations (Klipp et al., 2005). ODEs describe the temporal evolution of the concentration of molecular species and can be written as

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

with time-dependent states $x(t) \in \mathbb{R}_+^{n_x}$, vector field f, parameters $\xi \in \mathbb{R}_+^{n_\xi}$ (e.g. reaction rates), and parameter-dependent initial conditions $x_0(\xi) \in \mathbb{R}_+^{n_x}$. The states and parameters are mapped to the observables $y \in \mathbb{R}^{n_y}$ by an output function h,

$$y = h(x, \xi)\,.$$

We consider data $\mathcal{D} = \{(t_k, \bar{y}_k)\}_{k=1}^{n_t}$ at n_t time points with n_y observables, $\bar{y}_k = (\bar{y}_{1,k}, \ldots, \bar{y}_{n_y,k})^T$. The measurements $\bar{y}_k$ of the observables $y(t, \xi)$ are subject to measurement noise

$$\bar{y}_{i,k} \sim p(\bar{y}_{i,k} | y_i(t_k, \xi), \varphi_i)\,. \tag{1}$$

The noise is usually assumed to be normally distributed. In the presence of outliers, single observations are however drawn from an alternative distribution with heavier tails, which is difficult to assess due to small sample sizes.

2.2 Outlier scenarios

We studied three scenarios that differ in the outlier generating mechanism which are – in our own experience – practically relevant (Figure 1). In the first scenario (*no outliers*), no outliers are included in the data. In the second scenario (*one data point at zero*), the measured concentration at a certain time point t_k is zero, e.g. due to a missing label or entry. Consequently, we measure $\bar{y}_{i,k} = 0$. In practice this might not be easy to spot due to background intensity and additional noise. In the third scenario (*two data points interchanged*), two data points in the dataset were interchanged. This might have occurred due to labeling or entry errors. In the case of several observables ($n_y > 1$) the modification was applied to all n_y observables.

2.3 Distribution assumptions

For parameter estimation from outlier-corrupted data we study the standard assumption, the normal distribution, as well as distributions with heavier tails than the normal distribution, the Laplace, the Huber, the Cauchy and the Student's t distribution. The distributions used for p in (1) are listed in Table 1.

The considered distributions possess a range of properties regarding their moments. The Laplace distribution has well defined moments (e.g.

Table 1. Probability densities. The formulas for the normal, Laplace, Huber, Cauchy and Student's t distribution are listed together with the parameters defining the distributions. The error function is denoted by erf and the gamma function by Γ.

	probability density	distribution parameters φ
normal	$p(\bar{y}\|y,\sigma_n) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma_n}\right)^2\right)$	standard deviation $\sigma_n > 0$
Huber	$p(\bar{y}\|y,\sigma_h,\kappa) = s \cdot \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma_h}\right)^2\right), & \|\frac{\bar{y}-y}{\sigma_h}\| \leq \kappa \\ \exp\left(-\frac{1}{2}\left(2\kappa\|\frac{\bar{y}-y}{\sigma_h}\| - \kappa^2\right)\right), & \|\frac{\bar{y}-y}{\sigma_h}\| > \kappa \end{cases}$	scale $\sigma_h > 0$, tuning parameter $\kappa > 0$
	with $s = \left(\sqrt{2\pi}\sigma_h \mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right) + \frac{2\sigma_h}{\kappa}\exp\left(-\frac{1}{2}\kappa^2\right)\right)^{-1}$	
Laplace	$p(\bar{y}\|y,b) = \frac{1}{2b}\exp\left(-\frac{\|\bar{y}-y\|}{b}\right)$	scale $b > 0$
Cauchy	$p(\bar{y}\|y,\gamma) = \frac{1}{\pi\gamma}\frac{\gamma^2}{(\bar{y}-y)^2+\gamma^2}$	scale $\gamma > 0$
Student's t	$p(\bar{y}\|y,\sigma_t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma_t}\left(1+\frac{1}{\nu}\left(\frac{\bar{y}-y}{\sigma_t}\right)^2\right)^{-\frac{\nu+1}{2}}$	scale $\sigma_t > 0$, degrees of freedom $\nu > 0$

finite variance) for all parameter values. The Student's t distribution possesses an infinite variance for large degrees of freedom and the variance of the Cauchy distribution is always infinite. We refer to the Laplace, Huber, Cauchy and Student's t distribution in the following as heavier-tailed distributions. Note that the case of a log-normal distribution assumption, as used by Kreutz et al. (2007), is implicitly captured in the normal distribution assumption since this corresponds to log-transformation of the output and the use of a normal distribution assumption.

2.4 Parameter estimation

The kinetic parameters ξ and distribution parameters φ are usually unknown. To estimate the unknown parameters $\theta = (\xi, \varphi)$ from the data, we use maximum likelihood estimation. The likelihood $\mathcal{L}_{\mathcal{D}}(\theta)$ is the conditional probability of observing some data $\mathcal{D}$ given the parameters θ,

$$\mathcal{L}_{\mathcal{D}}(\theta) = \prod_{k=1}^{n_t}\prod_{i=1}^{n_y} p\left(\bar{y}_{i,k}|y_i\left(t_k,\xi\right),\varphi_i\right), \tag{2}$$

with distributions p as listed in Table 1. For numerical reasons, the maximum of the likelihood is usually determined by minimizing the negative log likelihood,

$$\theta^{\mathrm{ML}} = \arg\min_{\theta}\left\{-\sum_{k=1}^{n_t}\sum_{i=1}^{n_y}\log p\left(\bar{y}_{i,k}|y_i\left(t_k,\xi\right),\varphi_i\right)\right\}. \tag{3}$$

Substitution of p from Table 1 in (3) reveals the relation of this formulation with least squares and M-estimators. For the normal distribution with known variances, (3) is a least squares problem. For the Laplace and Huber distribution with known parameters, we obtain the least absolute deviation estimator and the Huber M-estimator, respectively. The formulation can however also be employed if the parameters of the statistical models, e.g. the tuning parameter of the Huber distribution, are unknown. For details we refer to the *Supplementary Information*.

The optimization problem (3) is usually – independent of the distribution assumption – nonlinear and non-convex. We performed the minimization by multi-start local optimization (Raue et al., 2013) using the MATLAB toolbox PESTO (Hross and Hasenauer, 2016). To improve performance and convergence of the optimization an analytical description of the gradient and higher-order derivatives was derived for all distribution assumptions (see *Supplementary Information, Section 1*). The reaction rate equations and the sensitivities, needed for the calculation of the gradient, were simulated using the MATLAB toolbox AMICI (Kazeroonian et al., 2016). Moreover, we estimated the $\log_{10}$-transformed parameters due to better numerical properties.

The accuracy of the maximum likelihood estimate for different distribution assumptions were evaluated using the Mean Squared Error (MSE)

$$\mathrm{MSE}\left[\xi^{\mathrm{ML}},\xi^{\mathrm{true}}\right] = \mathbb{E}\left[\left(\xi^{\mathrm{ML}}-\xi^{\mathrm{true}}\right)^2\right].$$

A small MSE indicates a good agreement of the true and estimated parameters. The expectation is computed over several datasets. We only calculated the MSE for the kinetic parameters ξ, since the distribution parameters are not comparable.

The uncertainty of an individual estimate can be assessed by computing the confidence interval (CI) for a confidence level α using profile likelihoods (Raue et al., 2009). The CIs should cover the true parameter with a frequency of $1-\alpha$. Accordingly, if the true parameter is known, the appropriateness of the CIs can be evaluated by computing the coverage ratio (CR), which should be close to the desired confidence level (Schelker et al., 2012).

2.5 Model selection

We performed hypothesis testing for the statistical models including the distribution assumptions using the Bayesian Information Criterion (BIC),

$$\mathrm{BIC} = -2\log\left(p\left(\mathcal{D}|\theta^{\mathrm{ML}}\right)\right) + \log(n_{\mathcal{D}})n_\theta, \tag{4}$$

with $n_{\mathcal{D}}$ denoting the number of data points and n_θ denoting the number of parameters. Models with low BIC values are preferred and models with differences in BIC values to the minimal BIC value above 10 are commonly rejected (Raftery, 1999).

3 Results

To study the performance and robustness of parameter estimation using the different distribution assumptions, we applied the methods to artificial data of a conversion process as well as experimental data of the JAK/STAT signaling pathway.

3.1 Simulation study: Conversion reaction

For this simulation study, we considered a simple conversion process: $A \rightleftharpoons B$. This process is described by a reversible reaction, which converts a biochemical species A to a species B with rate k_1, and B to A with rate k_2. The corresponding ODEs are

$$\dot{x}_1 = -k_1 x_1 + k_2 x_2 \,,$$
$$\dot{x}_2 = k_1 x_1 - k_2 x_2 \,,$$

for which the state vector $x = (x_1, x_2)^T$ consists of the concentrations of A and B, respectively. We assumed that x_2 is measured yielding the observation model $y = h(x, \theta) = x_2$.

For the evaluation of the proposed method, we generated 10^3 artificial datasets for each of the three outlier scenarios described in Section 2.2. The datasets were generated with initial conditions $x_0 = (1, 0)^T$, kinetic parameters $\xi = (k_1, k_2)^T = \left(10^{-1.5}, 10^{-1.5}\right)^T$ and normally distributed measurement noise with standard deviation 0.02. Examples of datasets for the scenarios are depicted in Figure 2.

Mean squared estimation error for different distribution assumptions
To evaluate the differences in parameter estimation using different distribution assumptions, the kinetic parameters k_1 and k_2 together with the distribution-specific parameters were estimated from the 10^3 datasets per scenario using maximum likelihood estimation (*Supplementary Information, Section 2*). Parameter estimation using the assumption of normally distributed measurement noise allowed for the reconstruction of the systems trajectory in the absence of outliers (Figure 2). However, if there are strong outliers, the fitted and the true trajectory differ, implying estimation errors. In contrast, for the Laplace, Huber, Cauchy and Student's t distribution the fitting yielded systems trajectories close to the trajectory used to simulate the data.

These findings are also reflected in the MSE for the parameter estimates of the kinetic parameters k_1 and k_2 (Figure 3A). If *no outliers* are present in the data, all methods yield a comparable MSE for both kinetic parameters. In the presence of outliers, the MSE achieved using the normal distribution is however much higher. This implies that the parameter estimates differ largely from the true parameters, which will result in wrong predictions. The heavier-tailed distributions were able to provide reliable estimates of the parameters in the presence of outliers. Indeed, the MSE hardly increased, indicating that the influence of a small number of outliers can be compensated. Consequently, robust estimation methods reduce the MSE for outlier-corrupted data.

Unraveling the presence of outliers using model selection
As parameter estimation using heavier-tailed distributions is robust with respect to outliers, we wondered if these methods can also be used to reveal outliers in datasets. To analyze this, model selection was performed regarding the statistical models using the BIC. Note that the models do not differ in the model dynamics but only in the distribution assumption. Using the 10^3 datasets per scenario the percentage was calculated how often a distribution assumption achieved the lowest BIC. The model employing the normal distribution assumption was chosen for most of the *no outliers* datasets (Figure 3B). In the presence of outliers, heavier-tailed distributions are preferred over the normal noise model. In particular for the *one data point at zero* scenario we observe a large difference of the average BIC values (Figure 3C). Accordingly, model selection detected the presence of outliers. This shows that heavier-tailed distributions can be used as diagnostic tools to test for the presence of outliers.

In addition to the distribution of the measurement noise and the outliers, also the structure of the biochemical reaction network might be unknown. In this case, the network structure has to be inferred from the experimental data along with the model parameters. If probability distributions with heavier tails are used, data points might (incorrectly) be considered as outliers due to the model's inability to describe them or because model selection methods prefer lower model complexity. At least in a simple toy problem this was, however, not observed. Furthermore, parameter estimation with heavier-tailed distribution tends to provide a good fit to a large fraction of the data instead of distributing the error equally (Figure 2B and C). Accordingly, a few measurements are sufficient to verify or falsify whether a data point is an outlier. For details we refer to the *Supplementary Information, Section 5*.

Optimizer convergence and computation time
In parameter optimization, critical aspects are optimizer convergence and computation time (Raue et al., 2013). We evaluated both properties by determining for each scenario and distribution assumption how many runs of the local optimizer converged by using a statistical approach (Hross and Hasenauer, 2016). We found that for this simple example the convergence is for most distributions comparable and above 75% (Figure 3D). Merely the optimization using the Huber distribution yields a slightly lower fraction of converged starts.

The mean time needed per start is similarly low for the normal, Cauchy and Student's t distribution (Figure 3E). Only the Laplace and Huber distribution have a higher computation time, since no approximation of the Hessian based on first-order sensitivities could be found (see *Supplementary Information, Section 1.2*). This verifies that the use of robust methods did not increase the computation time significantly.

Consistency of confidence intervals
To assess the influence of outliers on parameter CIs, we computed profile likelihoods. Examples for profiles are shown in Supplement Figure S2A and B. Based on these profile likelihoods, the CIs were computed for different confidence levels (Figure 4A).

For the artificial data (Figure 2) we find that in the case of *no outliers*, all distribution assumptions yield similar CIs for parameter k_1. The confidence intervals computed using the normal distribution widen in presence of outliers, yet not ensuring that the true parameter is covered. Also for the Laplace and Huber distribution the CIs become wider, the true parameter however remains covered. For the Cauchy and the Student's t distribution we observe that the CIs become even tighter, which is counter-intuitive as the information content in the data should be decreased. The presence of outliers shifts the probability mass often closer to the mode. The results for parameter k_2 are similar (Supplement Figure S2C).

We evaluated the reliability of the confidence intervals by determining the coverage ratio which states how often the true parameter ξ^{true} is covered by the confidence intervals for all 10^3 generated datasets per scenario (Figure 4B). Interestingly, the coverage ratio is lower than the confidence level for most of the cases indicating that the size of the confidence intervals is too narrow and therefore the uncertainty in the parameter estimates is underrated. For the normal distribution we tried to correct the coverage by applying the Bessel correction. The improvement was, however, minor (*Supplementary Information 2*). The Laplace and Huber distribution provide the best coverage ratio in the presence of outliers. It was shown that outliers have a greater influence on the confidence intervals when using the normal distribution assumption.

Sample size limitation of the Cauchy and the Student's t distribution
The performance of estimators often depends strongly on the sample size. Therefore, we analyzed how different distributions perform as the sample size is decreased. To this end, we varied the number of data points ($n_t = 10, 4, 3$) for datasets of the conversion process without outliers.

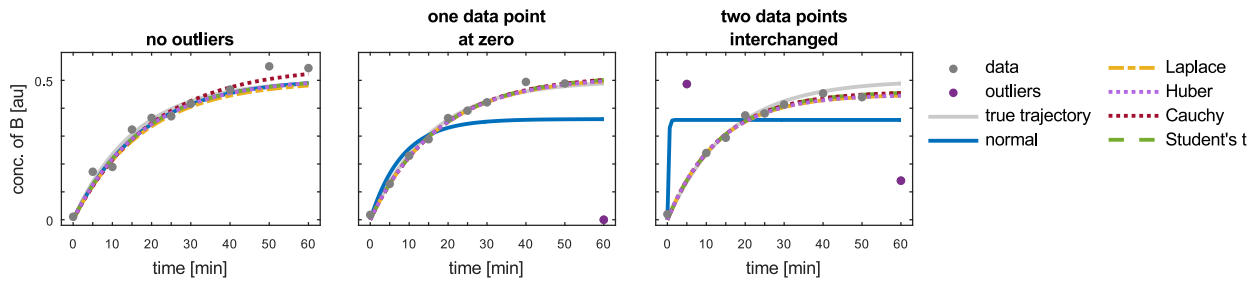

Fig. 2. Data and fits for different scenarios and distribution assumptions. The data points are generated by simulating the system with normally distributed noise and generating outliers according to the defined scenarios. The fits corresponding to the different distribution assumptions, normal, Laplace, Huber, Cauchy and Student's t distribution are plotted as lines.

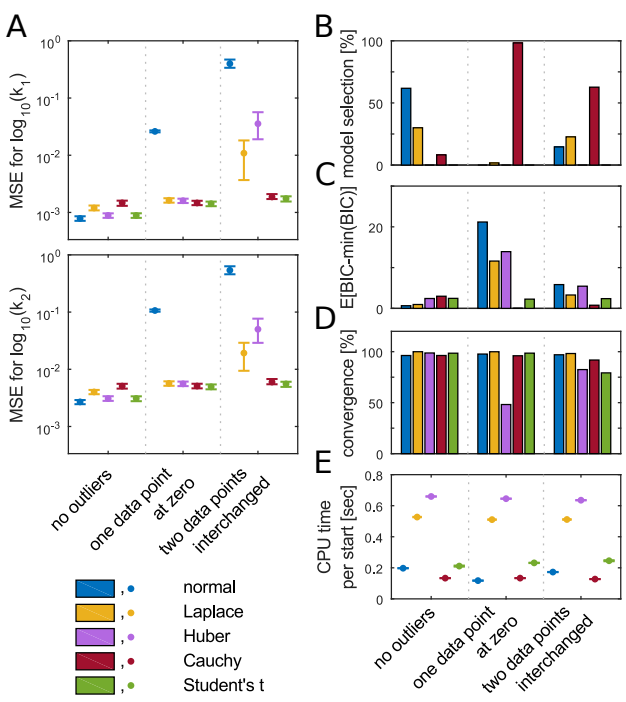

Fig. 3. Evaluation of optimization results for all three outlier scenarios. (A) MSE for the $\log_{10}$-transformed parameters. The circles indicate the MSE over all 10^3 datasets per scenario, while the error bars represent the 95% percentile bootstrap CIs. (B) Model selection results using BIC. The percentage is given for how many times each statistical model is chosen for the 10^3 datasets per scenario. (C) Difference of BIC value of a statistical model compared to the best statistical model. The difference is averaged over all datasets with the minimum computed for all datasets individually. (D) Average percentage of converged starts over all datasets. (E) The mean computation time per optimizer start and the corresponding standard error of mean.

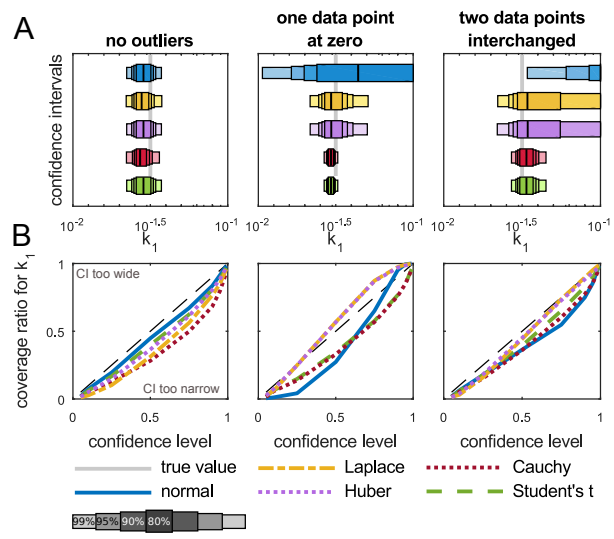

Fig. 4. Confidence intervals and coverage ratios. (A) Example CIs for one dataset per scenario (shown in Figure 2), indicated by different bars for $80\%, 90\%, 95\%$ and 99% from dark to light colors. The MLEs for the normal, Laplace, Huber, Cauchy and Student's t distribution are displayed as vertical lines. The true parameter value for k_1 is displayed as vertical grey line. (B) Coverage ratios for parameter k_1 for different confidence levels considering all 10^3 datasets per outlier scenario. Lines in the upper part of the panels indicate that the CI is too wide, lines in the lower part that it is too narrow.

For a lower number of data points, the model can fit an higher percentage of the data points exactly (= up to numerical accuracy). For the full datasets ($n_t = 10$), the obtained residual distributions for all combined datasets fit the corresponding distributions (Figure 5A), visualized for the median scale parameters obtained with parameter estimation (Figure 5B). The scale parameters for the normal, Laplace and Huber distribution do not become much smaller for lower number of data points. However, the scale parameters of the Cauchy and Student's t distribution are decreased and thus the mass of the distribution is concentrated on the exactly fitted data points, neglecting other residuals, i.e., the model overfits single data points. For $n_t = 3$ these scale parameters are even estimated at the lower bound defined as 10^{-10}. Scale parameters close to zero yield residual distributions which do not reflect the variation in the data (see also Figure S3).

For regression, Fernández and Steel (1999) suggest to provide a lower bound for the degrees of freedom ν calculated with respect to the ratio of exactly fitted data points to other data points, thereby avoiding the regions of likelihood for which the problem occurs (Jones and Faddy, 2003; Taylor and Verbyla, 2004). However, such a restriction is not possible for the Cauchy distribution, which should, according to the formula of (Fernández and Steel, 1999), only be used if less than half of the data points can be fitted exactly. In general, the Cauchy and Student's t distribution should be applied carefully if the model is too flexible and overfitting is to be expected.

3.2 Application study: JAK/STAT signaling pathway

To assess the performance of the robust estimation methods under realistic conditions, we considered model and data from JAK/STAT signaling

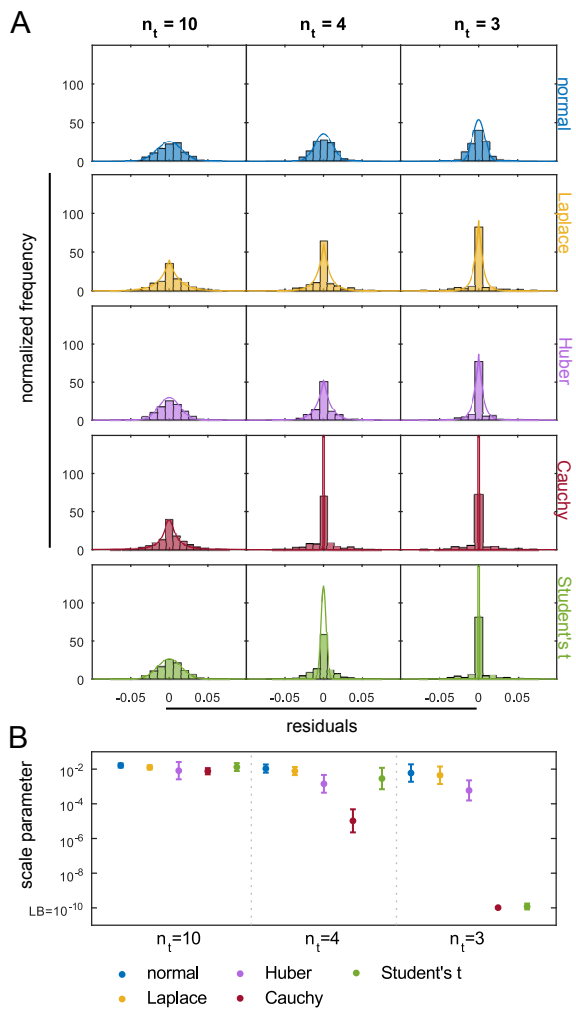

Fig. 5. Sample size limitation of Cauchy and Student's t distribution. (A) Normalized histogram of the residuals of all 10^2 datasets when the parameter estimation is performed for $n_t = 10, 4, 3$. The curve represents the corresponding probability density of the normal, Laplace, Huber, Cauchy or Student's t distribution using the estimated median value of the distribution-specific parameters over all 10^2 datasets. (B) Visualization of the corresponding scale parameters, σ_n for the normal, b for the Laplace, σ_h for the Huber , γ for the Cauchy and σ_t for the Student's t distribution

.

in response to Epo. The phosphorylated receptor (pEpoR) of the hormone Erythropoietin leads to JAK-mediated phosphorylation of STAT, which dimerizes and enters the nucleus to initiate the transcription of target genes (Figure 6A). We used the mathematical model introduced by Swameye et al. (2003), which is provided in the *Supplementary Information, Section 3.1*. It comprises kinetic parameters, spline parameters for modeling the input pEpoR concentration as well as scale and offset parameters.

Swameye et al. (2003) collected quantitative data of three observables at 16 time points. The dataset has been analyzed in a variety of studies and seems to be free of outliers. To evaluate the method, we introduced artificial outliers which led to 16 cases of *one data point at zero* and 120 cases of *two data points interchanged*. Subsequently, we studied how well the estimation results obtained for the original dataset are resembled. Examples of outlier realisations are visualized in Figure 6C and D along with the corresponding fit achieved using different noise models. We excluded the Cauchy distribution from our analysis and restricted the degrees of freedom for the Student's t distribution to $\nu > 2$, since overfitting of individual data points was an issue.

The fitting of the original dataset (*no outliers*) using the different distribution assumptions yields very similar trajectories (Figure 6B). This supports the hypothesis that the dataset is free of outliers. For the examples of the outlier-corrupted scenarios shown in Figure 6C and D, the trajectory obtained using the normal distribution is visibly influenced by the outliers, while the other distributions yield a similar behavior of the trajectories as in the original data without outliers.

The optimal parameter values found for *no outliers* were taken as reference for the MSE calculated for *one data point at zero* and *two data points interchanged* (see Figure 7A and B for the biologically meaningful parameters). For all parameters, the MSEs achieved using the Laplace, the Huber and the Student's t distribution are smaller than the MSE observed using the normal distribution. This implies that these statistical models yield more robust estimates in the presence of outliers even if the measurement noise might be normally distributed. Considering the MSE for the parameter vector instead of individual parameters, the normal distribution yields the highest error in the estimates for both outlier scenarios (Figure 7C). For the distribution parameters and the MSE of the biologically non-relevant parameters see Figure S5 and S6A-B. The convergence for the three scenarios is comparable for the normal, Laplace and Student's t distribution, but lower for the Huber distribution (Figure S6C). This application example demonstrates that the proposed approaches also yield promising results in a more realistic example.

4 Conclusion

Outliers in biological data can arise through experimental errors or incorrect data processing and, by definition, deviate largely from the predicted observable. Using objective functions which exploit the squared distance, as the normal distribution, gives a great weight to outliers. Consequently, these outliers have a relatively large contribution to the objective function compared to other data points. We implemented efficient gradient-based parameter estimation for ODE models using heavier-tailed distributions to reduce the effect of outliers. These methods are well established in robust regression and we demonstrated that they are also beneficial in the context of dynamical systems.

We evaluated parameter estimation using heavier-tailed distributions from artificial data for a conversion process and from artificially perturbed experimental data for the JAK/STAT signaling pathway. The analysis revealed that in the absence of outliers, parameter estimation for the different methods performed similarly. In the presence of outliers, however the MSE is reduced by the use of heavier-tailed distributions. The heavier-tailed distributions yielded a reasonable optimizer convergence, even for a discretized PDE model of Pom1p gradient formation (*Supplementary Information 4*). The suggested model-based approaches facilitate an automatic, unbiased detection of outliers and can also be applied if no replicates for individual measurements are available. They allow a joint estimation of the kinetic and distribution parameters, which is in different formulations rather time-consuming (Qiu et al., 2016). Furthermore, only in this normalized description a thorough statistical evaluation is possible as it enables the use of statistical criteria.

A manual exclusion of outliers is time consuming and suffers from the lack of a universal definition of outliers, since extreme data points that are truly generated from the underlying mechanisms should not be removed as they carry important information. However, if the intention still is to exclude outliers from the data, the detection of the outliers might be more reliable when first fitting the full data with a robust method. Then

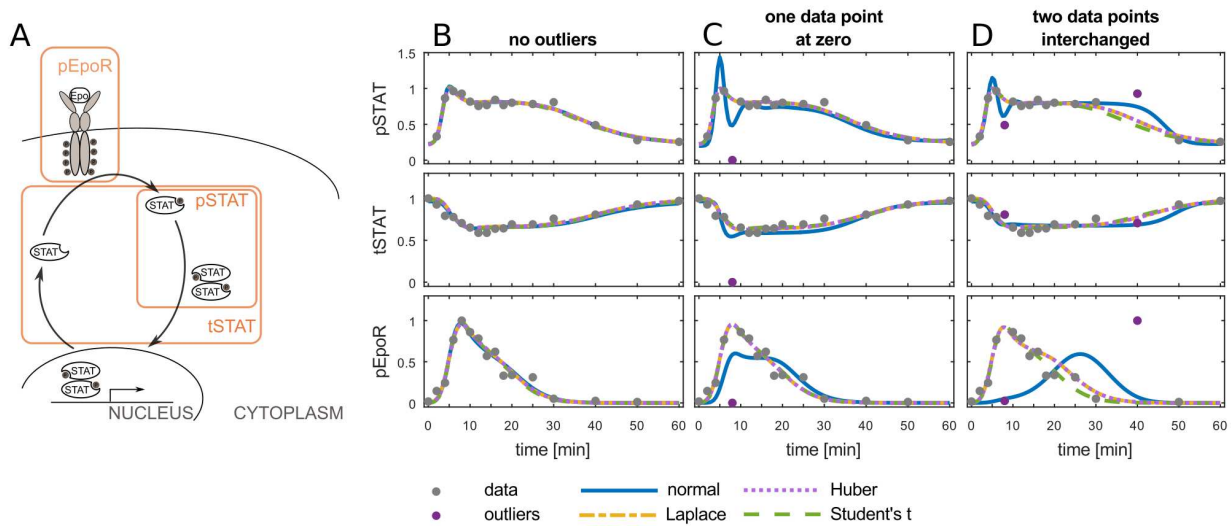

Fig. 6. Modeling and parameter estimation of JAK/STAT signaling. (A) Illustration of the pathway, for which arrows represent biochemical reactions and circles the species of the system. The observables are highlighted with boxes. (B) Experimental data without outliers and fitted trajectories obtained by normal, Laplace, Huber and Student's t distribution. (C) Example of the outlier scenario for which one data point is set to zero, in this case $t = 8$ min. (D) Scenario for which the two data points at $t = 8$ min and $t = 40$ min are interchanged.

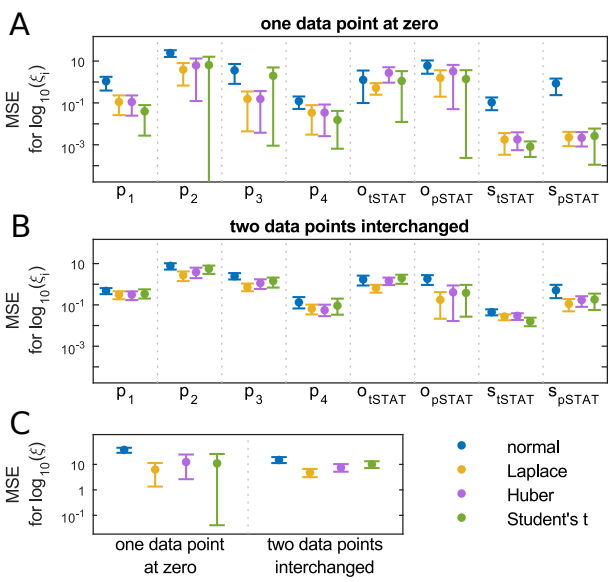

Fig. 7. Comparison of estimation accuracy for JAK/STAT signaling. Confidence intervals for the MSE, calculated by bootstrapping, for all $\log_{10}$-transformed parameters are shown for the normal, Laplace , Huber and Student's t distribution for the outlier scenario with (A) one data point at zero and (B) two data points interchanged. The parameter values are compared to those obtained by fitting the no outliers data. (C) MSE for the parameter vector including kinetic, offset and scale parameters.

data points that have a large distance from the corresponding simulated trajectory can be removed according to e.g. the three-sigma rule (Aggarwal, 2015).

The evaluation of different statistical models revealed that sample size limitations can result in problems, e.g. overfitting. Furthermore, the coverage of confidence intervals might not be appropriate, as the distribution does not capture the true outlier distribution. Even in the case in which the true distribution is known, i.e. in the *no outlier* scenario, the coverage might be incorrect as the threshold value of the profile likelihoods holds only asymptotically or for linear problems. In other cases, the threshold values can be computed numerically using Monte-Carlo sampling (Kreutz et al., 2012). For the considered problem and threshold values derived from the χ^2-distribution, the Laplace and Huber distribution were found to provide the best balance as they lead to reliable fits by ensuring a good coverage. If, however, the convergence is an issue, e.g. for large models with many parameters and state variables, the Laplace distribution might be advantageous. Consequently, our recommendation for outlier-corrupted data is to employ the Laplace or Huber distribution as residual distribution in the parameter estimation. As part of future work, other distribution assumptions, e.g., the normal-Laplace distribution (Reed, 2006) , which also has an asymmetric version, could be examined for parameter estimation in dynamical systems.

In summary, we provided a first comprehensive evaluation of the different properties of heavier-tailed distributions when calibrating dynamic mathematical models to experimental data. Therefore, we derived the necessary gradients and Hessian matrices of the objective function to ensure an efficient optimization. The proposed approach has substantial practical value, since it allows to use statistical tools, such as model selection, and it yields robust parameter estimates in the presence of outliers. This facilitates more accurate and reliable predictions, which are important to gain a better understanding of the biological processes of interest.

References

Aderem, A. (2005). Systems biology: its practice and challenges. *Cell*, 121(4):511–513.

Aggarwal, C. C. (2015). Outlier analysis. In *Data Mining*, pages 237–263. Springer.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR*, volume 1, pages 267–281. Akademiai Kiado.

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer.

Cao, J., Wang, L., and Xu, J. (2011). Robust estimation for ordinary differential equation models. *Biometrics*, 67(4):1305–1313.

Chen, J.-Q., Heldman, M. R., Herrmann, M. A., Kedei, N., Woo, W., Blumberg, P. M., and Goldsmith, P. K. (2013). Absolute quantitation of endogenous proteins with

precision and accuracy using a capillary western system. *Analytical biochemistry*, 442(1):97–103.

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

Ghosh, D. and Vogt, A. (2012). Outliers: An evaluation of methodologies. In *Joint Statistical Meetings*, pages 3455–3460. American Statistical Association San Diego, CA.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer Netherlands.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

Hross, S. and Hasenauer, J. (2016). Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinf.*, 32(15):2321–2329.

Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2(1):343–372.

Jones, M. and Faddy, M. (2003). A skew extension of the t-distribution, with applications. *J. R. Stat. Soc. Series B Stat. Methodol.*, 65(1):159–174.

Kazeroonian, A., Fröhlich, F., Raue, A., Theis, F. J., and Hasenauer, J. (2016). CERENA: ChEmical REaction Network Analyzer–A toolbox for the simulation and analysis of stochastic chemical kinetics. *PLoS ONE*, 11(1):e0146732.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560):1662–1664.

Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2005). *Systems biology in practice*. Wiley-VCH, Weinheim.

Kreutz, C., Bartolome Rodriguez, M. M., Maiwald, T., Seidl, M., Blum, H. E., Mohr, L., and Timmer, J. (2007). An error model for protein quantification. *Bioinf.*, 23(20):2747–2753.

Kreutz, C., Raue, A., and Timmer, J. (2012). Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Syst. Biol.*, 6(120).

Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.*, 84(408):881–896.

Motulsky, H. and Christopoulos, A. (2003). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. GraphPad Software Inc., San Diego CA.

Niu, Z., Shi, S., Sun, J., and He, X. (2011). A survey of outlier detection methodologies and their applications. In *Artificial intelligence and computational intelligence*, pages 380–387. Springer Berlin Heidelberg.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Stat. Comput.*, 10(4):339–348.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA.

Qiu, Y., Hu, T., Liang, B., and Cui, H. (2016). Robust estimation of parameters in nonlinear ordinary differential equation models. *Journal of Systems Science and Complexity*, 29(1):41–60.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, 27(3):411–417.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, volume 29, pages 427–438. ACM.

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.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., and Timmer, J. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335.

Reed, W. J. (2006). *The Normal-Laplace Distribution and Its Relatives*, pages 61–74. Birkhäuser Boston, Boston, MA.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Stewart, C. V. (1999). Robust parameter estimation in computer vision. *SIAM Review*, 41(3):513–537.

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.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.

Taylor, J. and Verbyla, A. (2004). Joint modelling of location and scale parameters of the t distribution. *Statistical Modelling*, 4(2):91–112.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62.