Technische Universität München

Department of Mathematics

Master's Thesis

Parameter Estimation for Outlier Corrupted Data

Corinna Maier

Supervisor: Prof. Dr. Dr. Fabian Theis

Advisor: M.Sc. Carolin Loos, Dr.-Ing. Jan Hasenauer

Submission Date: April 29, 2016

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Garching,

Abstract

Mathematical models are a valuable tool to answer biological questions or evaluate competing hypotheses which are not within reach of experiments. Since commonly not all system parameters are known, models need to be calibrated based on experimental data. However, outlier corrupted data poses a serious threat to model calibration as outliers may lead to distorted parameters, which result in wrong model predictions. Detecting and removing those outliers is a challenging task with regard to the complexity and amount of biological data. A reasonable alternative approach constitutes robust parameter estimation. For parameter estimation it is commonly assumed that the deviation of the measurement from the predicted observable is normally distributed. This assumption is, however, strongly affected by large erroneous measurements. Heavier-tailed distributions, that have heavier tails than the normal distribution, are less susceptible to outliers and consequently, using a heavier-tailed distribution as distribution assumption for the deviation of the measurements from the predicted observables yields a robust approach to parameter estimation.

In the presented methods for estimating the parameters of ordinary differential equation (ODE) models, we propose the Laplace, Cauchy and Student's t distribution as heavier-tailed alternatives to the normal distribution assumption. The robustness of our novel methods was assessed for population average data, which was modified according to defined outlier scenarios. At first artificially generated data of a conversion reaction was studied and the results showed that the new methods are able to decrease the error of parameter estimates for outlier corrupted data. To support this finding an application study to artificially perturbed experimental data of the Jak/Stat signaling pathway was performed. Using heavier-tailed distribution assumptions constitutes indeed a robust approach to parameter estimation for outlier corrupted data that leads to reliable parameter estimates. Since model accuracy is a prerequisite for reliable predictions of the behavior of a biological process, the proposed methods will enhance the investigation of biological systems.

Zusammenfassung

Im Kontext von dynamischen Modellen für biologische Systeme werden in der vorliegenden Arbeit Methoden vorgestellt, die eine robuste Parameterschätzung für ausreißerbehaftete Datensätze ermöglichen. Dynamische Modelle erlauben es, biologische Fragen zu beantworten oder konkurrierende Hypothesen zu evaluieren, die nicht durch biologische Experimente auszuwerten sind. Im Allgemeinen sind jedoch nicht alle Systemparameter bekannt, die für die Modellierung nötig sind. Deswegen müssen unbekannte Modellparameter aus experimentellen Daten geschätzt werden. Ausreißerbehaftete Daten stellen allerdings ein großes Problem für die Modellkalibrierung dar, da sie zu verfälschten Parametern führen, die wiederum falsche Modellvorhersagen nach sich ziehen. Diese Ausreißer zu finden und aus dem Datensatz zu entfernen ist hinsichtlich der Komplexität und Größe biologischer Datensätze eine herausfordernde Aufgabe. Eine sinnvolle Alternative stellt die robuste Parameterschätzung dar. Für die Parameterschätzung wird üblicherweise angenommen, dass die Abweichungen der Messungen von den vorhergesagten Beobachtungen normalverteilt sind. Diese Annahme wird jedoch stark von fehlerhaften Messungen beeinflusst. Endlastige Verteilungen, die mehr Masse in ihren Randbereichen haben, sind weniger anfällig für Ausreißer und folglich führt die Verwendung dieser Verteilungen als Verteilungsannahmen zu einem robusteren Ansatz für die Parameterschätzung.

In den vorgestellten Parameterschätzmethoden für Differentialgleichungsmodelle wurden die Laplace-, Cauchy- und Studentsche t Verteilung als Alternativen zu der Normalverteilung eingeführt. Um die Robustheit unserer Methoden zu testen, untersuchten wir Datensätze denen künstlich Ausreißer hinzugefügt wurden, wie sie auch unter üblichen Laborbedingungen auftreten könnten. Die Eigenschaften der neuen Methoden wurden an künstlich generierten Daten eines Umwandlungsprozesses illustriert und bewertet. Die Resultate zeigen, dass die Methoden für ausreißerbehaftete Datensätze den Fehler der Parameterschätzer reduzieren. Dieses Ergebnis wird unterstützt durch eine Anwendungsstudie zu realen Daten des Jak/Stat Signalweges. Die Verwendung von endlastigen Verteilungen stellt folglich einen robusten Ansatz zur Parameterschätzung im Falle von ausreißerbehafteten Datensätzen dar. Da Modellgenauigkeit eine Vorrausetzung für zuverlässige Vorhersagen ist, tragen die neuen Methoden dazu bei die Untersuchung biologischer Systeme zu verbessern.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Quantitative dynamic models have become a beneficial tool to gain an integrated understanding of biological processes (Ideker et al. 2001; Kitano 2002). The key benefit of these dynamic models is the integration of multiple experimental data sets and the prediction of system properties that are not within reach of biological experiments (Aderem 2005; You et al. 2004). As the parameters of the model are generally unknown and need to be inferred from experimental data, the crucial step of quantitative dynamic modeling is the calibration of the model based on experimental data. Hence, the basis of modeling is a combination of experimental and computational methods (Wierling et al. 2007), which means that a model can be only as meaningful as the data it is based on. Experimental data used for parameter estimation is collected using a broad spectrum of quantitative measurement techniques. The measurement precision in quantification has increased (Chen et al. 2013), but there are still numerous ways in which errors are introduced during the process of data collection and data processing (Ghosh et al. 2012). Biological systems are highly sensitive to their environment, thus, unusual conditions in laboratories or inconsistent laboratory procedures have an distorting effect on the quality of measured data. In addition, technical limitations and human errors such as pipetting errors or incorrect data processing can result in large measurement errors (Pearson et al. 2004;Motulsky et al. 2004).

In the field of quantitative dynamic modeling it is commonly assumed for model calibration that the noise is normally or log-normally distributed. Erroneous data poses, however, a serious problem to model calibration, since false measurements are likely to distort parameter estimates (Tarantola 2005). Those individual data points, which are corrupted by errors, are in general denoted as outliers. Outliers are generated from a different mechanism as the remainder of the data points and may, therefore, be misleading (Hawkins 1980; Motulsky et al. 2004). As a result of inaccurate model calibration due to outliers the validity of the given model becomes limited, thus, its predictive power regarding biological questions decreases.

1.1 Treatment of outliers

Since the use of outlier corrupted data distorts results in various fields, as for example in regression analysis, many methods for the detection of outliers have been developed (Aggarwal 2015; Ben-Gal 2005; Hodge et al. 2004; Niu et al. 2011). Outlier detection algorithms either assign a score about the degree of abnormality or a binary label to a data point. This labeling is usually based on a fit to a distribution or a distance measure, e.g. k-nearest neighbor distance (Ramaswamy et al. 2000). Eventually, it remains but a subjective decision on whether or not a data point is sufficiently abnormal to be removed (Aggarwal 2015), which introduces a person dependent bias. In practice this distinction is even less straightforward if measurement noise is present, which complicates the identification of outliers. The increasing size and complexity of biological data makes the removal of outliers a challenging task, especially if the visualization is difficult due to high dimensionality (Tarantola 2005). The removal 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 et al. 2004). Furthermore, in the case of multiple outliers the distinction between outliers and non-outliers becomes even more ambiguous (Huber 2011). Thus, a robust parameter estimation method is needed that leads to reliable parameter estimates in the presence of outliers.

Robust approaches that do not alter outlier corrupted data are for example already used in regression (Lange et al. 1989) and computer vision (Stewart 1999). These approaches exploit estimators that are less affected by outliers, e.g.M-estimators (Huber 2011). To the best of our knowledge a robust parameter estimation approach in the field of quantitative dynamic models has not yet been introduced. By fitting the model to the majority of the data robust methods can even also be used to detect outliers. Outliers appear as large values in the residuals of the fit and consequently they can be identified by a distance measure, e.g. Z-value test (Aggarwal 2015). Those outliers are often hidden if a non-robust fit of the model to outlier corrupted data is used (Rousseeuw et al. 2005). In the present work, the unresolved issue of robust parameter estimation for ODE models for biological systems is addressed. We present an approach that exploits different distribution assumptions in the model to improve coping with outliers.

1.2 Contribution of the thesis

The work is composed of the following parts: at first, the fundamentals of ODE models and parameter estimation are outlined in Chapter 2. Subsequently, three biologically motivated outlier scenarios are introduced in Chapter 3 representing realistic ways in which outliers are introduced into data sets. Next, different statistical models for the residual distribution are presented in Chapter 4. These include, additionally to the standard approach, the normal distribution, three distributions with heavier tails than the normal distribution, namely the Laplace, Cauchy and Student's t distribution. These three distributions will be denoted in the following as heavier-tailed distributions. In order to illustrate and evaluate the properties of the resulting estimation in the absence and presence of outliers, artificial data for a conversion reaction with known parameters is generated according to the defined outlier scenarios in Chapter 5. Limitations, which arise if too many data points can be fitted exactly by the model, are discussed at the end of the application example. Following this, the methods are applied to artificially perturbed biological data of the Jak/Stat signaling pathway (Chapter 6). Finally, the results of the present thesis are summarized and strengths as well as weaknesses of the different distribution assumptions are discussed, providing directions for future work.

Chapter 2

Data-driven modeling of dynamic biological systems

The following chapter introduces the fundamentals of data-driven quantitative dynamic models and presents the type of data the models are based on. Starting from the translation of a biological process in ODEs, the typical work flow is described. Next, unknown parameters of the model need to be estimated from the data. This problem is approached by maximum likelihood estimation using multi-start local optimization. In addition, assessment criteria for the model performance are explained, including convergence, model accuracy and model selection. Finally, the uncertainty attached to the parameter estimates needs to be investigated.

2.1 Experimental data

In general, different types of biological data comprise single cell snapshot data, single cell time-lapse data and population average data. Whereas the first two types provide information about the cell-to-cell variability of a heterogeneous population, the latter describes an average cell in a homogeneous population (Hasenauer et al. 2011). In this work population average data

$$\mathcal{D} = \{(t_k, \bar{y}_{ik})\}, \qquad i = 1, \ldots, n_y, \ k = 1, \ldots, n_t$$

is studied, which contains the mean values $\bar{y}_{ik}$ of n_y measured molecular species at a certain number of time points n_t. By taking the mean over the whole population of cells, a representation for a typical behavior of a cell is given. Biological experiments that average over the population require the measurement of a high number of cells; possible techniques include e.g. western blotting (Renart et al. 1979) or micro arrays (Pirrung et al. 2014).

2.2 Quantitative modeling of dynamic biological systems

In a mathematical model $\mathcal{M}$ of a dynamic biological system the time evolution of the biological process is most commonly described by reaction rate equations (RREs),

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

with time-dependent states $x(t) \in \mathbb{R}_+^{n_x}$, vector field f, parameter-dependent initial state $x_0(\xi) \in \mathbb{R}_+^{n_x}$ and parameters $\xi \in \mathbb{R}_+^{n_\xi}$, comprising, for instance, reaction rates, binding affinities, and initial concentrations (Klipp et al. 2008). The states $x(t)$ correspond to the concentration of n_x molecular species, e.g. hormones, proteins or mRNA. The concentration of X molecules for a given reaction volume Ω is given by $x = X/\Omega$ (McNaught et al. 1997). Generally, not all of the parameters can be technically measured or were measured during the data collection. Thus, this description leads to a system of ODEs with unknown parameters. This deterministic setting is only valid under the assumption that stochastic effects can be neglected, i.e., molecules are present in high numbers and compartments are well-mixed (Gábor et al. 2015; Gillespie 1992). Biological measurement techniques allow the experimental assessment of an observable $y = h(x, \xi)$, which is a function of the states or parameters. In general, measurements $\bar{y}_{ik}$ do not exactly display the predicted observable $y(t, \xi)$, but are subject to uncertainty due to technical noise, technical limitations or human errors in the data collection and processing (Tarantola 2005). The residual vector $r_i(t, \xi)$ is defined as the deviation of the measurand $\bar{y}_i$ from the observable $y(t, \xi)$ for a parameter vector ξ:

$$r_i(t, \xi) = \bar{y}_i - y_i(t, \xi)\,, \qquad i = 1, \ldots, n_y\,.$$

For the distribution of the residuals a distribution p is assumed that describes the spread of the measurands around the observables

$$\bar{y}_i \sim p(\bar{y}_i | y_i, \varphi_i)\,, \tag{2.1}$$

with parameters φ_i accounting e.g., for the scale of the distribution; ($\varphi = (\varphi_1, \ldots, \varphi_{n_y})$). The measurement noise is most commonly assumed to be normally distributed with noise level σ, i.e., $p(\bar{y}_i | y_i, \varphi_i) = \mathcal{N}(y_i, \sigma_i^2)$. In the case of outlier corrupted data single observations are, however, drawn from an alternative distribution, which is difficult to assess due to small sample sizes. Since outliers result in large residual values, a normal distribution might not be an adequate representation. Possible assumptions for the distribution p are presented in Chapter 4, including next to the normal distribution three heavier-tailed distributions.

2.3 Parameter estimation

The inverse problem, to estimate model parameters based on experimental data, can be approached by maximum likelihood (ML) estimation (Raue et al. 2009; Weber et al. 2011). The goal is to find parameters $\theta = (\xi, \varphi)$ which maximize the conditional probability of observing the data $\mathcal{D}$ given the model $\mathcal{M}$, assuming a distribution p. Therefore, a cost function, the likelihood,

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathrm{P}(\mathcal{D}|\theta) = \prod_{k=1}^{n_t} \prod_{i=1}^{n_y} p\big(\bar{y}_{ik} | y_i(t_k, \xi), \varphi_i\big)\,, \tag{2.2}$$

is introduced that penalizes discrepancies of the model from the measured data, assuming independence. The data is recorded for n_y observables at time points t_k, $k = 1, \dots, n_t$. Instead of maximizing the likelihood, equally the negative log-likelihood $J(\theta) =$ -log$(\mathcal{L}_{\mathcal{D}}(\theta))$ can be minimized as objective function. Furthermore, the parameters of biological systems are usually non-negative by definition. Thus, in order to handle parameters of different magnitude, it is advised to log-transform the parameters for numerical computations (Maiwald et al. 2008). In the parameter estimation process the parameters φ of the distribution assumption p are estimated simultaneously with the parameters describing the model dynamics ξ. These formulations provide significant numerical advantages and lead to the optimization problem

$$\hat{\theta}^{\mathrm{ML}} = \arg \min_{\theta \in \Theta} J(\theta)\,, \tag{2.3}$$

where $\hat{\theta}^{\mathrm{ML}}$ denotes the maximum likelihood estimate (MLE). The problem can be solved efficiently by derivative-based multi-start local optimization (Raue et al. 2013). This optimization approach is based on the assumption that if the best optimum is found several times, starting from start points that are spread over the whole parameter space, this optimum corresponds to the global optimum. To guarantee a comprehensive coverage of the parameter space Θ, the start points for the local optimization are generated between the lower and upper parameter bounds by using a Latin hyper cube sampling scheme. Therefore, the parameter range in each dimension is equally divided in N intervals, where N is the number of start points. A uniformly distributed random number is drawn within each interval in each dimension, i.e., it is guaranteed that each interval is represented once (McKay et al. 1979). The required derivatives for the optimization of the objective function are most efficiently computed by employing the sensitivity equations, which are more reliable than a finite difference approximation (Sengupta et al. 2014).

2.4 Assessment criteria for model performance

The optimization problem (2.3) is in general non-convex and often has several locally optimal points (Banga 2008). For multi-start local optimization it is important to ensure the convergence to a global optimum and the generation of reproducible results. In this work the convergence was assessed statistically by a likelihood ratio test using the significance level $\alpha = 0.05$. A start s is said to be converged, if

$$-2\Big(\log(\mathcal{L}_{\mathcal{D}}(\theta_s)) - \log(\mathcal{L}_{\mathcal{D}}(\hat{\theta}^{\mathrm{ML}}))\Big) < \Delta_\alpha\,, \tag{2.4}$$

where Δ_α is the $100(1-\alpha)$th percentile of the χ^2 distribution with one degree of freedom and θ_s the parameter vector resulting from the optimization (Hross et al. 2016). If a certain number of starts, spread over the parameter space, have converged to the same best optimum it is assumed to have found the global optimum.

The key feature of a model is its accuracy, since only accurate parameter estimates allow meaningful model analysis and predictions. The model accuracy can be assessed by evaluating the mean squared error (MSE), given by

$$\begin{aligned}\mathrm{MSE}[\hat{\xi}^{\mathrm{ML}}, \xi^{\mathrm{true}}] &= \mathbb{E}[(\hat{\xi}^{\mathrm{ML}} - \xi^{\mathrm{true}})^2] & \text{(2.5a)}\\ &= \mathrm{VAR}[\hat{\xi}^{\mathrm{ML}}] + \mathrm{Bias}^2[\hat{\xi}^{\mathrm{ML}}, \xi^{\mathrm{true}}]\,. & \text{(2.5b)}\end{aligned}$$

The MSE measures the goodness of the estimate $\hat{\xi}^{\mathrm{ML}}$ by taking the mean of the squared difference of the estimate and the true parameter value. It combines the variance and the bias and thus it takes into account the random error, the variance of the estimates, as well as the systematic error, the difference of the true value and the expected value of the estimates (Hand et al. 2001). Consequently, it handles the well known variance-bias-tradeoff (Gábor et al. 2015). Note that the MSE is only reasonable for the parameters describing the model dynamics ξ, not for the parameters of the distribution assumptions φ, since these are not comparable.

In the modeling process often several possible models $\mathcal{M}_j$ arise that may describe an experimental data set $\mathcal{D}$. These models can differ in the RREs or distribution assumption p. A commonly used criterion for model selection is the Bayesian Information Criterion (BIC) (Schwarz 1978)

$$\mathrm{BIC}_j = -2\,\log(\mathcal{L}_{\mathcal{D}}(\hat{\theta}_j^{\mathrm{ML}})) + \log(n_{\mathcal{D}}) \cdot n_{\theta,j}\,, \tag{2.6}$$

for which $\hat{\theta}_j^{\mathrm{ML}}$ is the MLE for model $\mathcal{M}_j$, $n_{\mathcal{D}}$ the number of data points and $n_{\theta,j}$ the number of parameters of model $\mathcal{M}_j$. Note that $n_{\theta,j}$ comprises the number of parameters of the RREs and the parameters of the distribution assumption p_j used in model $\mathcal{M}_j$. The BIC is proportional to the log-likelihood and therefore rewards agreement between the model and the data but penalizes model complexity. It follows from Equation (2.6) that the model with smallest BIC value is most appropriate for the given data. In general, a difference in BIC to the minimum BIC value greater than 10 signifies a very strong evidence in favor of the model which yields the smaller BIC value. Since this difference evaluates how much more the data supports the second model over the first, this gives a reasonable criterion for model rejection (Raftery 1995).

2.5 Uncertainty analysis of parameter estimates

As already mentioned in the introduction, limitations in the data collection process affect the parameter estimation. Thus, parameter estimation has to be followed by an analysis of the uncertainty attached to the parameter estimates. Incomplete knowledge about the data or the model dynamics may lead to non-identifiable parameters. Identifiability comprises structural (Bellman et al. 1970) and practical identifiability (Raue et al. 2009). Whereas structural identifiability does not depend on the experimental data but solely on the model structure, practical identifiability takes into account the quality of the underlying data. To assess these parameter indeterminancies, profile likelihoods,

$$\mathrm{PL}(\theta_h) = \max_{\theta_g \neq \theta_h} \mathcal{L}_{\mathcal{D}}(\theta)\,, \tag{2.7}$$

can be used (Murphy et al. 2000; Raue et al. 2009). For each parameter θ_h a one-dimensional profile is calculated by fixing this parameter and maximizing over the remaining parameters $\theta_g\,, g \neq h$. A practically non-identifiable parameter for a data set is characterized by a flat profile, resulting in an infinite confidence interval (CI). Profile likelihood based confidence intervals, which contain the true parameter value with probability $1-\alpha$, are obtained by (Schelker et al. 2012)

$$\mathrm{CI}_\alpha(\theta_h|\mathcal{D}) = \{\theta_h| -2\log\big(\mathrm{PL}(\theta_h)\big) + 2\log\big(\mathcal{L}(\hat{\theta}^{\mathrm{ML}})\big) \leq \Delta_\alpha\}\,, \tag{2.8}$$

with the $100(1-\alpha)$th percentile Δ_α, computed by the inverse cumulative distribution of the χ^2 distribution with one degree of freedom. A criterion to evaluate the confidence in the parameter estimates is the coverage rate. The coverage rate provides information about how often the true value is in fact located within the confidence interval, see (Raue et al. 2009). In general, the coverage ratio (CR) is considered, which is given by the coverage rate divided by the number of simulations (Schelker et al. 2012). This ratio should be close to the desired level of confidence $CR \approx 1-\alpha$. If $CR < 1-\alpha$ the uncertainty in the parameter estimates is underrated. Contrary, if $CR > 1-\alpha$, the confidence intervals are more cautious than required (Raue et al. 2009). Thus, the coverage ratio provides a means to validate the appropriateness of confidence intervals.

Chapter 3

Outlier scenarios

Since quantitative models are calibrated to biological measurements, prospective predictions depend intrinsically on the quality of the data. Data quality is concerned with accuracy, incompleteness and reliability (Wang et al. 1996). Deficient quality is for example characterized by missing values and erroneous measurements. This thesis deals with data of poor quality, namely data which includes outliers. It is, however, difficult to get a notion of outliers initially. Hawkins (1980) gave an intuitive definition of an outlier:

> ”*An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*”

Thus, outliers are individual data points that deviate considerably from the remaining data points and in general, they are not generated by the same underlying mechanism as the remainder of the data set. In many applications outliers contain useful information, e.g. credit card fraud or medical diagnosis (Aggarwal 2015). In those cases an outlier is introduced into data sets by a disease or a fraud event, which constitutes an important discovery. Unfortunately, there are also many undesired causes of outliers. Biologically motivated mechanisms that produce unwanted outliers in data sets are, to name but a few, technical failures, inconsistent laboratory conditions, as well as inaccuracies in the data collection and data processing (Motulsky et al. 2004). Some concrete examples of problems that may arise in practice: Gassmann et al. (2009) have for instance outlined the difficulties arising in the quantification of Western Blots, comprising e.g. the inhomogeneous illumination of scanners. A labeling error in microarray data sets, which were made available to the participants of the CAMDA 2002 (Critical Assessment of Microarray Data Analysis), was revealed by Stivers et al. (2004). Inconsistent labels would have caused wrong biological conclusions, if this error had remained undetected (Pearson et al. 2004).

In the following we distinguish three scenarios; one reference scenario without outliers and two scenarios that represent outlier generating mechanisms. These outlier scenarios allow us to examine the robustness and accuracy of the methods in presence of outliers.

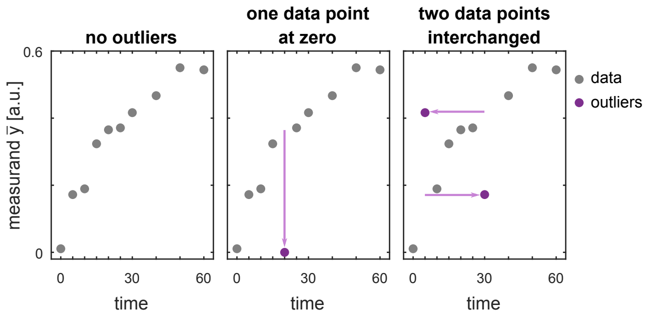

Figure 3.1: Scenarios with and without outliers. Data without outliers, corresponding to the scenario *no outliers*. Scenario *one data point at zero*. Randomly one data point is set to zero. For the third scenario two data points are randomly selected and interchanged, resulting in the scenario *two data points interchanged*. Data points that are no outliers are represented as grey circles, while outliers are highlighted in purple. The purple arrows represent the outlier generating mechanism.

1. *No outliers.* The first scenario, displayed in Figure 3.1, represents the noise realization of the output without any alteration. The data set does not include any outliers, any deviation from the observable is due to noise.

2. *One data point at zero.* Figure 3.1 shows also the second scenario, which describes the failure of a technical measurement device during the process of data collection. The device fails to record at a certain time point t_k resulting in a zero instead of a measured value $\bar{y}_{ik}$. To artificially generate a data set of this type a time point t_k is chosen randomly out of $t_1, \ldots, t_{n_t}$ and the corresponding measured value is set to zero
$$(t_k, \bar{y}_{ik}) \to (t_k, 0)\,, \qquad i = 1, \ldots, n_y\,.$$

3. *Two data points interchanged.* Another common mistake in data processing might be the interchange of data points. This typical entry or labeling error is covered in the scenario *two data points interchanged* of Figure 3.1. This modification is generated by randomly choosing two time points t_k and t_l out of $t_1, \ldots, t_{n_t}$ and interchanging their values $\bar{y}_{ik}$ and $\bar{y}_{il}$, i.e.,
$$\begin{aligned}(t_k, \bar{y}_{ik}) &\to (t_k, \bar{y}_{il})\\ (t_l, \bar{y}_{il}) &\to (t_l, \bar{y}_{ik}),\end{aligned}$$
where $k \neq l$ and $i = 1, \ldots, n_y$. Note that the degree of alteration of this scenario depends on the chosen time points. If, for example, time points next to each other are chosen, the modification might not lead to significant deviations from the main behavior of the data points.

In the case of several observables ($n_y > 1$) the modification is applied to all n_y observables.

Chapter 4

Distribution assumptions for the residuals

> ”*Everyone believes in the [normal] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.*” (Cramér 1945)

This aphorism given by Cramér, which is attributed to the mathematician Poincaré, is still valid as it is commonly assumed that the residuals, the deviations of the measured and predicted observables due to noise, are normally distributed. Kreutz et al. (2007) show that noise in immunoblotting data is adequately modeled as multiplicative log-normally distributed. This corresponds to a log-transformation of the data and using an additive normal distribution. Hence, usually the normal distribution is used for the distribution p, defined in Equation (2.1). In the case of outlier corrupted data, however, a distribution which only describes the technical noise might not be adequate. Outliers result in large values in the residual vector which cannot be captured by the normal distribution. In the following, we therefore introduce three distributions with heavier tails than the normal distribution in addition to the normal distribution; the Laplace, the Cauchy and the Student’s t distribution.

4.1 Normal distribution

Most commonly measurement noise is assumed to be normally distributed in parameter estimation for quantitative dynamic models (Raue et al. 2013). The probability density for the normal distribution $\mathcal{N}(\mu, \sigma^2)$ reads

$$p(z; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(z-\mu)^2}{2\sigma^2} \Big),$$

for which μ is the mean of the distribution and σ the standard deviation. The standard normal distribution $\mathcal{N}(0, 1)$ is shown in Figure 4.1A-C in blue. Assuming that the

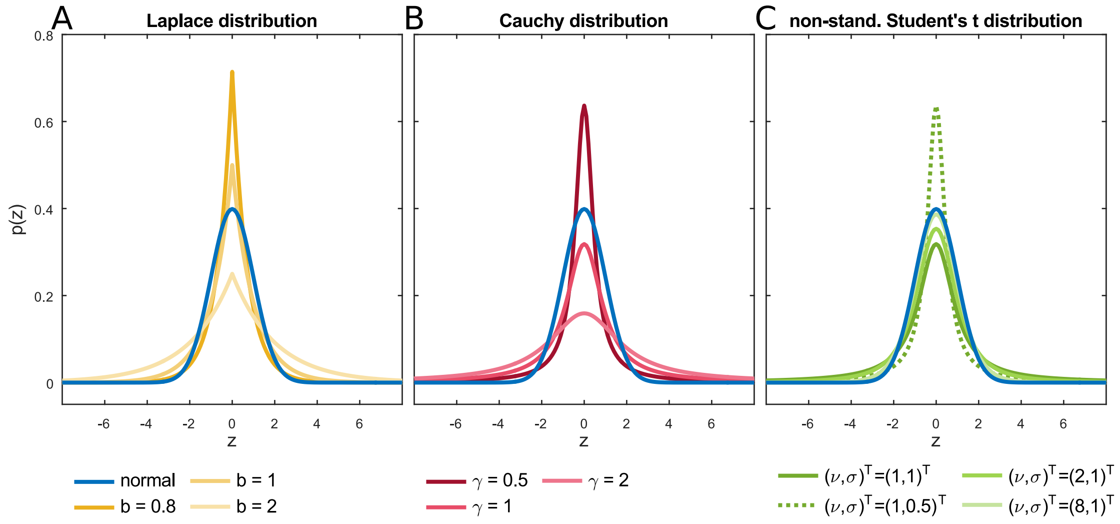

Figure 4.1: Visualization of the heavier-tailed distributions. (**A**) Comparison of the Laplace distribution with the standard normal distribution $\mathcal{N}(0,1)$ for diverse values of b. (**B**) The Cauchy distribution with differen γ values compared to the standard normal distribution. (**C**) Comparison of the non-standardized Student's t distribution with the standard normal distribution for various degrees of freedom ν and scale parameters σ. Note that the Cauchy distribution with $\gamma = 1$ coincides with the Student's t distribution with $\nu = 1$. All three distributions have heavier tails than the normal distribution.

measurements $\bar{y}$ are normally distributed around the observables y, $\bar{y} \sim \mathcal{N}(y, \sigma^2)$, the likelihood function, according to Equation (2.2), reads

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

with parameter $\varphi_i = \sigma_i$. The objective function is then given by the negative logarithm of the likelihood,

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

This approach coincides with minimizing the weighted sum of squared residuals. Outliers have, by definition, a large distance to the observable. Squaring this distance, as in the normal distribution assumption, gives greater weight to outliers. Thus, large errors have a relatively large contribution to the objective function compared to smaller errors (Willmott et al. 2005). The least-squares method is consequently not robust. In the field of regression Cornbleet et al. (1979) have for example reported that least-squared regression coefficients are incorrect if outliers are present. Minimizing the squared errors is a standard approach since it facilitates computation by avoiding the absolute value of the residuals compared to the least absolute deviations. The analytic gradient and Hessian matrix, which are used to improve the optimization, are provided in the Appendix A.1.

4.2 Laplace distribution

Taking into account the sum of absolute residuals (Edgeworth 1887) rather than the sum of squared residuals results in the Laplace distribution. The probability distribution of the Laplace distribution is given by

$$p(z;\mu,b)=\frac{1}{2b}\exp\left(\frac{-|z-\mu|}{b}\right)$$

with location parameter μ and scale parameter $b>0$. The distribution has mean and median μ and variance $2b^2$. Tarantola (2005) suggests to use the Laplace distribution as soon as a single outlier is included in a data set as the assumption of a normal distribution in this case leads to unacceptable results of the inverse problem. The Laplace distribution has longer tails than the normal distribution, see Figure 4.1A, and is therefore better suited to represent uncertainties in the data due to outliers. The likelihood function is defined as

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

with parameter $\varphi_i=b_i$. The negative log-likelihood used for the optimization is

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

Minimizing this objective function corresponds to minimizing the sum of absolute deviations. The method of least absolute deviations (L_1 method) is robust in contrast to least squared deviations (L_2 method) (Portnoy et al. 1997). Willmott et al. (2005) have demonstrated that the total squared error will become increasingly larger than the total absolute error, if the total error is contained in a small number of individual large errors. The authors suggest to use the absolute error, rather than the sum of squared errors, as the absolute error constitutes an less ambiguous measure. This is also supported by Huber (2011), who shows that already two erroneous observations out of 10^3 are enough to prefer the absolute deviations over the squared deviations in the case of unmodified data sets (without outlier detection and removal). In the context of biological data Purdom et al. (2005) showed that an asymmetric Laplace distribution fits errors in gene expression data better than a normal distribution. In robust regression using the least absolute regression estimator is referred to as L_1-regression (Rousseeuw et al. 2005). The gradient and Hessian matrix are to be found in Appendix A.2.

4.3 Cauchy distribution

The Cauchy distribution, also called Lorentz distribution, is an example of a pathological distribution, i.e., it has no finite moments (Haas et al. 1970). The probability density of the Cauchy distribution is defined as

$$p(z;\mu,\gamma) = \frac{1}{\pi} \frac{\gamma}{(z-\mu)^2 + \gamma^2} \, ,$$

with location parameter μ and scale parameter $\gamma > 0$. The influence of the scale parameter on the shape of the distribution is visualized in Figure 4.1B. The distribution is tending towards a dirac delta function for infinitesimal small scale parameter γ. Since the Cauchy distribution has neither defined mean nor variance, the distribution is characterized by the median, given by μ. Assuming the Cauchy distribution with median $y_i(t_k)$ for the distribution p, the likelihood function can be expressed as

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

with parameter $\varphi_i = \gamma_i$. The objective function used for the optimization is

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

The gradient and Hessian matrix are provided in Appendix A.3.

4.4 Non-standardized Student's t distribution

William S. Gosset derived the Student's t distribution and published his findings under the pseudonym Student (Student 1908). In statistics, the Student's t distribution is applied when estimating the mean for unknown standard deviation and small sample sizes (Fisher 1925). The Student's t distribution results from a scale mixture of a normal distribution with an inverse-gamma distributed mixing variable (Andrews et al. 1974). The derived distribution is still symmetric, but the tail behavior is altered. The probability density for the non-standardized Student's t distribution is defined by

$$p(z;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\,\sigma} \left(1 + \frac{1}{\nu}\left(\frac{z-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} ,$$

with location parameter μ, scale parameter $\sigma > 0$, and degrees of freedom $\nu > 0$ (Jackman 2009). Its mean and variance are defined by

$$\begin{aligned} \mathbb{E}[X] &= \ \mu && \text{for } \nu > 1 \\ \mathrm{VAR}[X] &= \begin{cases} \frac{\nu}{\nu-2} & \text{for } \nu > 2 \\ \infty & \text{for } 1 < \nu \leq 2 \, . \end{cases} \end{aligned}$$

The standardized version is given for $\mu = 0$ and $\sigma = 1$. In the following, the term Student's t distribution refers to the non-standardized version. As $\nu \rightarrow \infty$ the Student's t distribution tends to the normal distribution (Figure 4.1C) and for $\nu = 1$ the Student's t distribution coincides with the Cauchy distribution. Note that σ is not a standard deviation but determines the scaling of the distribution. The likelihood with non-standardized Student's t distribution assumption is written as

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

with parameter vector $\varphi_i = (\sigma_i, \nu_i)^T$. This means that the distribution has one parameter more than the normal distribution assumption that needs to be estimated. However, this additional parameter enables the Student's t distribution to approximate the normal distribution for large ν values in the case of outlier-free data, whereas for outlier corrupted data the degrees of freedom can take small values to put more weight in the tails. Consequently, the degree of robustness is strongly related to the degree of freedom (Fonseca et al. 2008). The objective function is computed as

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

In regression analysis the Student's t distribution was already introduced for robust statistical modeling (Fernández et al. 1999; Lange et al. 1989; Liu et al. 1995; Peel et al. 2000), allowing the exploitation of its ability to downweigh outliers. Fernández et al. (1999) discuss difficulties arising in global optimization for maximum likelihood estimation using the Student's t distribution, if the degrees of freedom ν are defined in $\mathbb{R}^+$. In this case, a global maximum does not exist if the model is able to fit too many data points exactly, since the likelihood function can reach arbitrarily large values for small scale parameters $\sigma \to 0$. However, this overfitting problem can be prevented by setting an appropriate lower bound for ν, for which the authors provide a criterion based on the percentage of exactly fitted data points. Thus, in the parameter estimation procedure the additional question arises where to set the lower bound for the degrees of freedom ν. This issue is further discussed in Section 5.8. The gradient and analytic Hessian of the log-likelihood can be found in Appendix A.4.

Note that we do not explicitly consider the log-normal distribution as this just corresponds to log-transformation of the output and using the normal distribution assumption.

Chapter 5

Illustrative model of a conversion reaction

As first application of the different distribution assumptions we studied a simple conversion reaction to illustrate the effect of different distribution assumptions (Chapter 4) on the parameter estimation in the case of outlier corrupted data. For this purpose artificial data was generated according to the three outlier scenarios described in Chapter 3. This artificial data generation allowed a statistical analysis of the accuracy of the novel methods since the true parameters are known. Conversion processes have a great biological relevance as they appear frequently in biochemical reaction networks, e.g. reversible phosphorylations.

5.1 Model description of a conversion process

A conversion process is a reversible reaction; biochemical species A converts to species B with reaction rate k_1 and B reconverts to A with rate k_2:

$$\mathrm{A} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{B}\,.$$

The RREs describing the time evolution of the conversion process are given by

$$\begin{aligned} \frac{dx_A(t)}{dt} &= -k_1 x_A(t) + k_2 x_B(t)\,, & x_A(0) &= x_{A_0}\,, \\ \frac{dx_B(t)}{dt} &= -k_2 x_B(t) + k_1 x_A(t)\,, & x_B(0) &= x_{B_0}\,, \end{aligned}$$

where x_A and x_B denote the concentration of species A and B, respectively. The states of the ODE system comprise the concentrations, $x(t) = (x_A(t), x_B(t))^T$ with initial state $x(t_0) = (x_{A_0}, x_{B_0})^T$, and the parameters are given by $\xi = (k_1, k_2)^T$. Assuming that the average concentration of B can be measured experimentally, the observable reads $y(t_k) = x_B(t_k)$.

5.2 Data generation

The artificial data generation was composed in three steps: The process was first simulated for ten time points $t \in \{0, 5, 10, 15, 20, 25, 30, 40, 50, 60\}$ minutes with true kinetic parameters $\log_{10}(\xi) = (-1.5, -1.5)^T$ and initial concentrations $x(t_0) = (1, 0)^T$. In accordance with the general applied assumption that measurement noise follows a normal distribution (Raue et al. 2013), normally distributed noise with constant noise level $\sigma = 0.02$ was added to the simulated observable $y(t_k)$ to obtain "realistic" measurements $\bar{y}_{ik} \sim \mathcal{N}(y(t_k), \sigma^2)$. As last step, the data set was modified according to one of the three outlier scenarios depicted in Figure 3.1. The above described procedure was repeated 10^3 times for each outlier scenario. This allows a statistical assessment of the difference in parameter estimation of the models involving the different distribution assumptions regarding robustness, performance and accuracy using the previously described assessment criteria, see Chapter 2.

5.3 Parameter estimation

Pretending that we had unknown parameters θ, but known initial concentrations $x(t_0)$, we performed maximum likelihood estimation of the model parameters for each generated data set. The objective function was minimized by multi-start derivative-based optimization in MATLAB (version R2015b) using the Parameter EStimation TOolbox PESTO (Hross et al. 2016) for each distribution assumption. The RREs and sensitivities for the derivatives were simulated using the toolbox CERENA (Kazeroonian et al. 2016). The parameter space Ξ for the dynamical parameters ξ was chosen as $\log_{10}(\Xi) = [-3.5, 1]^2$ and the parameter space of the distribution assumptions as $\log_{10}(\Phi^{(N,C,L)}) = [-5, 0]$ for the normal (N), Cauchy (C) and Laplace (L) distribution and $\log_{10}(\Phi^{(T)}) = [-5, 0] \times [0, 5]$ for the Student's t distribution (T). The parameter range for the degrees of freedom of the Student's t distribution was chosen in such a way that for the lower bound of ν the Student's t distribution corresponds to the Cauchy distribution, whereas for the upper bound it approaches the normal distribution. The choice of the lower bound of ν is explained in greater detail at the end of this application example, see Section 5.8. Within these bounds 100 start points were generated using Latin hypercube sampling. As local solver, starting from these start points, the MATLAB routine `fmincon.m` was used, utilizing a trust-region-reflective algorithm (Coleman et al. 1996; Nocedal et al. 2006) with provided gradient and approximation of the Hessian matrix, see Appendix A. For the Laplace distribution the optimization was performed using the interior-point algorithm rather than the trust-region-reflective algorithm due to better performance when the Hessian matrix is not user-supplied. Calculating the Hessian matrix for the Laplace distribution requires the calculation of the second order sensitivities, see Appendix A.2, which is avoided here. To ensure convergence to a global optimum, a convergence check pursuant to Equation (2.4) was included that doubled the number of start points in the

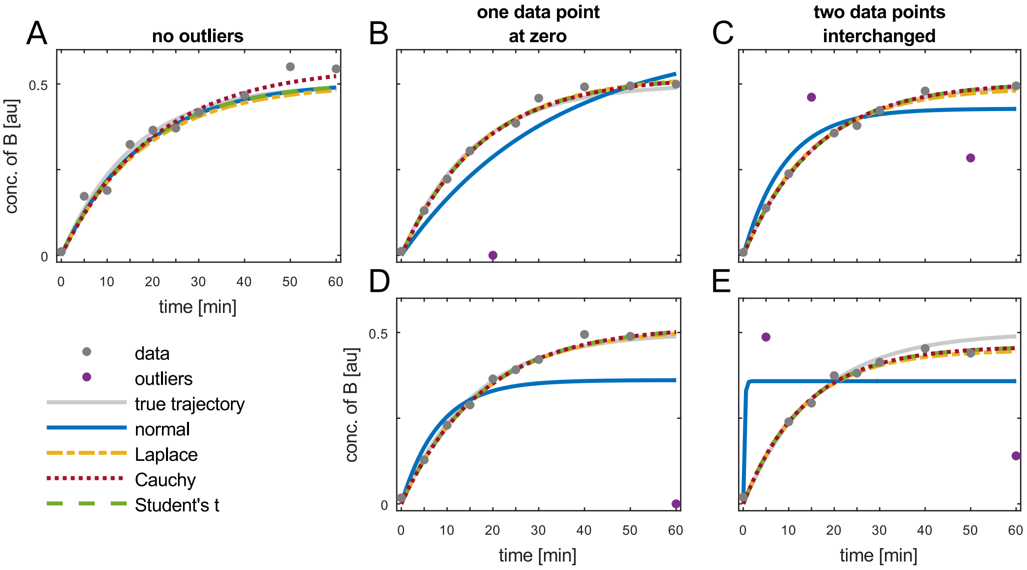

Figure 5.1: Example trajectories for the conversion process. The different distribution assumptions, normal (blue), Laplace (yellow), Cauchy (red) and Student's t (green) are used for model calibration based on artificial data of the conversion reaction with true model trajectory (grey), which was modified according to defined outlier scenarios. (**A**) *No outliers* scenario. Fits to an unaltered noise realization (grey circles). (**B**) Outlier scenario with *one data point at zero*. One outlier (purple circle) is introduced at time point $t = 20$ min. (**C**) Outlier scenario with *two data points interchanged*. The data points at time points $t = 15$ min and $t = 50$ min are interchanged. (**D**) A second example for the outlier scenario *one data point at zero*. The data point at time point $t = 60$ min is set to zero. (**E**) Another example for *two data points interchanged*. This time the data points at time points $t = 5$ min and $t = 60$ min are interchanged.

case of non-convergence. Non-convergence was assumed, if less than ten starts had converged to the best optimum. This multi-start optimization procedure was performed for the 10^3 data sets of each outlier scenario to obtain the MLEs of the parameters.

5.4 Qualitative analysis of the results

A first impression of the robustness of the new approach can be gained by simulating the model trajectories using the obtained MLEs of the model parameters and compare the reconstructed trajectories with the true trajectory. Figure 5.1 shows some example model trajectories for MLEs obtained by applying different distribution assumptions in the parameter estimation for the three outlier scenarios. In panel A the generated data with normally distributed noise but without outliers was considered for model calibration. The models with the different distribution assumptions are similarly able to describe the data sets, as all model trajectories are in good agreement with the data. It should be noted that the reconstructed trajectories using the Student's t distribution and normal distribution coincide and overlap with the true trajectory.

Since there are several possibilities for the outlier scenarios *one data point at zero* and *two data points interchanged*, as the outliers are chosen randomly, see Chapter 3, we demonstrate the effect of the distribution assumption with two examples for each scenario. In these examples, depicted in panels B-E, the fits assuming normally distributed residuals (blue) are clearly misled by the outliers. In all examples the fits deviate considerably from the true trajectory used for generating the data. Panel B and C have smaller deviations as the outliers are less extreme than in panels D and E. This shows that the normal distribution assumption is susceptible to outliers by putting too much weight on outlying observations. On the contrary, if the heavier-tailed distributions are applied, the ML estimation is less affected by the outliers. The fits using the heavier-tailed distributions are qualitatively the same for the scenarios including outliers as for the scenario without outliers. Remarkably, the model trajectories received for the Laplace, Cauchy and Student's t distribution coincide almost completely for the scenarios *one data point at zero* and *two data points interchanged* and are close to the true trajectory. Accordingly, the assumption of a heavier-tailed distribution leads to reliable fits for all cases displayed.

To gain an overview how these example fits are representative for the overall behavior, a plot displaying the fits to the first 100 data sets is shown in Figure 5.2. In the *no outliers* scenario all distribution assumptions deliver similar results. In presence of outliers the distributions with heavier tails than the normal distribution are less deceived by the introduced outliers. The Laplace, Cauchy and Student's t distribution lead in almost all cases to similar trajectories close to the trajectories of the *no outliers* case. Only in one of the cases for the scenario *two data points interchanged* all the heavier-tailed distributions show a larger deviation from the true trajectory. In the case of *one data point at zero* the trajectories gained with the normal distribution assumption are downward skewed and for *two data points interchanged* some trajectories do not even show a similar curvature as the true trajectory.

These first qualitative results of the obtained model trajectories revealed that the normal distribution is unsuitable in the presence of outliers and does not lead to reliable fits to the non-outlier data points, as it is misled by the outliers. Heavier-tailed distributions were shown to be advantageous in the case of outlier corrupted data as they are less affected by extreme outlying observations. They enable a robust parameter estimation which leads to appropriate fits to the non-outlier data points. For data without outliers all distributions are suitable and deliver similarly good results.

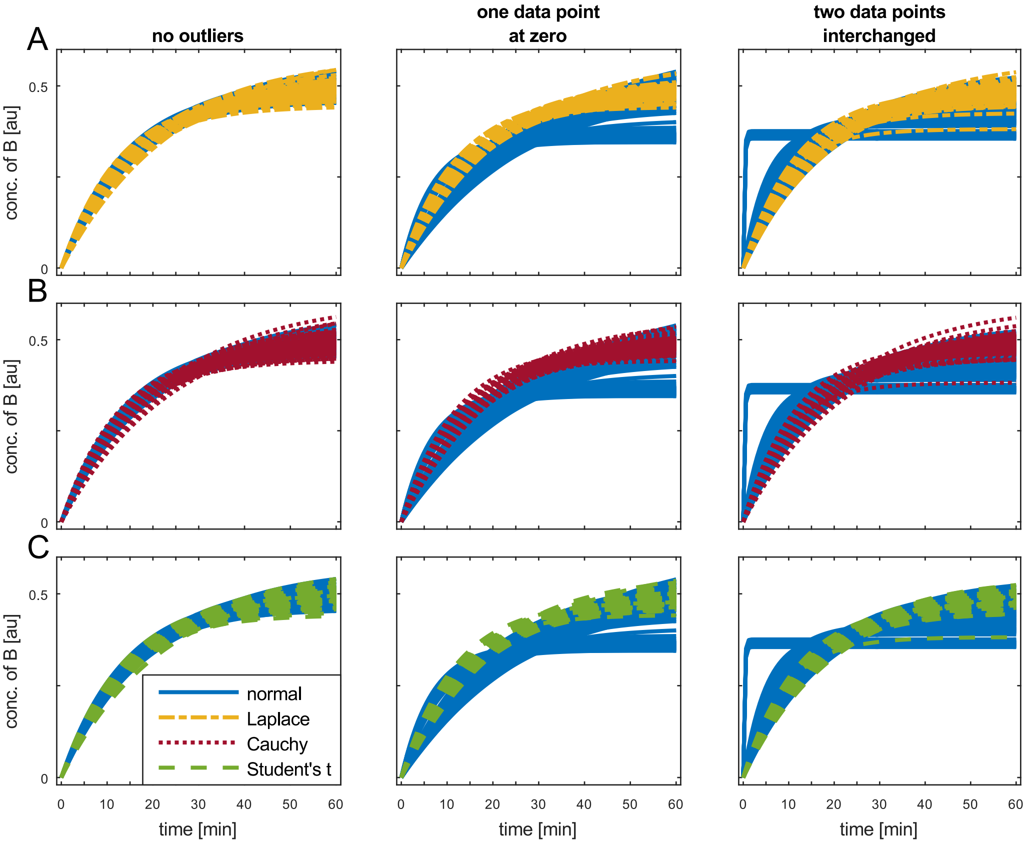

Figure 5.2: Estimation results for the first 100 data sets for each outlier scenario. Comparison of the model trajectories for MLEs obtained by model calibration using a normal distribution assumption (blue) to fits employing a Laplace distribution (yellow) (**A**), Cauchy distribution (red) (**B**) and Student's t distribution (green) (**C**).

5.5 Statistical analysis of the results

The artificial data generation enables a comprehensive assessment of the difference in parameter estimation with the different distribution assumptions using the previously described assessment criteria, see Chapter 2.

Estimation accuracy

One important feature of a model is its accuracy as inaccuracies propagate to prospective predictions. Therefore it needs to be determined how close the estimated parameter values are to the true parameters used for the data generation. This can be conducted by computing the MSE, see Equation (2.5). The logarithm of the MSE is visualized in Figure 5.3A with errorbars indicating the 95% percentile bootstrap confidence intervals,

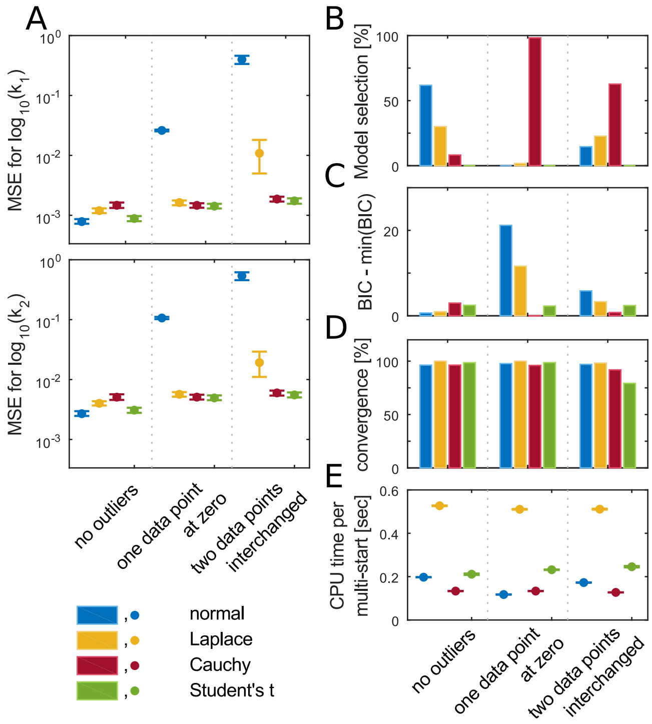

Figure 5.3: Statistical analysis for the conversion process. Statistical analysis of the parameter estimation results using the normal (blue), Laplace (yellow), Cauchy (red) and Student's t distribution (green). (**A**) Logarithmic plot of the MSE for $\log_{10}(k_1)$ and $\log_{10}(k_2)$ with errorbars obtained by bootstrapping. (**B**) Model selection based on BIC. (**C**) Differences in BIC values between the models. (**D**) Percentage of converged starts. (**E**) CPU time per multi-start in seconds.

see (Efron 1992; Rizzo 2007). A small value indicates good accordance of the estimate with the true parameter value. In the case of *no outliers* the assumption of a normal distribution represents the true model as the data was generated assuming normally distributed noise. Accordingly, the MSE is smallest for the normal distribution assumption, followed by the Student's t distribution, which approximates the normal distribution for high degree of freedom values. However, the MSE is equivalently small for the other distribution assumptions in the *no outliers* scenario, but highest for the Cauchy distribution. In the case of outlier corrupted data, the MSE assumes high values for the normal distribution assumption in the case of *one data point at zero* and *two data points interchanged*. This implies, that the parameter estimation employing the normal distribution is not able to infer the true parameter values in presence of outliers and is therefore not accurate. The Laplace distribution achieves low values for the first two outlier scenarios, but in the case of *two data points interchanged* the MSE rises. However, the value for the Laplace distribution is still much lower than the value obtained with the normal distribution. The MSE values for the models using the Cauchy and Student's t distribution stay at a small value for all scenarios and are consequently the most accurate in presence of outliers. In summary, the analysis of the estimation accuracy showed that choosing a heavier-tailed distribution assumption reduces the MSE for outlier corrupted data significantly and thus, leads to more reliable estimates than the normal distribution assumption. The robust methods are able to infer reasonable parameter values from the data, even in presence of outliers, unlike the normal distribution. The normal distribution assumption yields wrong parameter estimates for outlier corrupted data, which subsequently reduces the predictive power of the model.

Model selection

Using different distribution assumptions for the residuals leads to different models that describe a given data set. Hence, the question arises, which model should be selected for a given data set. In this work, model selection was performed via hypothesis testing using the BIC according to Equation (2.6). The model with the lowest BIC is selected to be the most appropriate one for the data set. In Figure 5.3B model selection is visualized for the three outlier scenarios for all 10^3 data sets. In the case of *no outliers* the model with normal distribution assumption is chosen in 61.8 % of the cases. This should be the case since the noise is generated following a normal distribution. The Laplace distribution is, however, also chosen almost half as many times. In scenario *one data point at zero* the Cauchy distribution is selected almost exclusively (98.3%). In the case of the Student's t distribution this can be explained by the higher number of parameters ($n_\theta^{(T)} = 4$) in the Student's t distribution, which yields a higher penalty term compared to normal, Laplace and Cauchy ($n_\theta^{(N,L,C)} = 3$). For the normal and Laplace distribution it was already shown that the MSE takes higher values for the scenarios including outliers than for the Student's t and Cauchy distribution, which results in smaller log-likelihood values. In the case of *two data points interchanged*, the Cauchy distribution is still selected most of the times, but sometimes model selection also favors the Laplace distribution as well as the normal distribution. This might be due to the fact that not all cases of *two data points interchanged* yield sufficiently large outliers, e.g. if data points next to each other are interchanged.

In Figure 5.3C the difference in BIC values is considered. For each data set the minimal BIC is computed and then subtracted from the BIC values found for the different distribution assumptions. This serves to reveal the actual difference in BIC of the models, which remains hidden in the previous analysis of model selection. In the case of *no outliers* this difference is smaller for all distribution assumptions than the common rejection value of 10, see Section 2.4. This shows that in this case all models explain the data almost equally well and none can be rejected. In the two scenarios containing outliers, modeling with Cauchy and Student's t distribution leads to similar BIC values. The Student's t distribution has a higher number of parameters and consequently, as seen in Figure 5.3C, model selection prefers the Cauchy over the Student's t distribution. In scenario *one data point at zero* the difference of the BIC values for the normal distribution and the minimal BIC value is in average greater than 10 (21.1). Thus, the model assuming a normal distribution is not appropriate in the scenario *one data point at zero* and can be rejected.

But also the Laplace distribution has to be rejected in some cases, as the mean value is above the threshold. For *two data points interchanged* all BIC differences are smaller than 10. None of the models can be rejected according to this criterion, but the normal distribution achieves the highest value. In conclusion, model selection revealed the presence of outliers. In absence of outliers mostly the normal distribution is chosen, but in both scenarios containing outliers, model selection favors the heavier-tailed distributions over the normal distribution.

Performance comparison

As explained in Section 2.4, the convergence is an important criterion for multi-start local optimization to yield a global optimum as well as reproducible results. The convergence of the methods for the multi-start local optimization approach was compared by counting the number of converged starts among the number of all start points, see Equation (2.4). In Figure 5.3D, the average percentage of converged starts is illustrated. The convergence of the models is for this simple model comparable and relatively high, only in the *two data points interchanged* scenario the convergence is little decreased for the Student's t distribution. However, in all scenarios and for all distribution assumptions enough starts converged to the optimum.

The average computation time for the different distribution assumptions per multi-start is the lowest for the Cauchy distribution with about 0.1 seconds in the case of *no outliers*, see Figure 5.3E. The Student's t distribution requires more than double the time, which might be due to the higher number of parameters, while the normal distribution is located within these two distributions. The value for the Laplace distribution is not directly comparable with the other values as a different local solver, the interior-point algorithm, was applied. As mentioned before, the interior-point was used as the computation of the Hessian matrix requires the second order sensitivities. All in all, the computation times as well as the convergence are in a reasonable range for all approaches.

5.6 Uncertainty analysis

As explained in the background section, the work flow of quantitative dynamic modeling does not end with the parameter estimates. A comprehensive report of MLEs requires an uncertainty analysis.

In order to assess the reliability of parameter estimates we used profile likelihoods, see Equation (2.7), which were computed with the toolbox PESTO (Hross et al. 2016). In Figure 5.4A, the normalized profile likelihoods of parameter k_1 obtained for the four distribution assumptions are compared for the cases displayed in Figure 5.1A,D and E. The true parameter value is indicated by a vertical grey line. In the scenario *no outliers* all profiles overlap and are close to the true parameter value of k_1. The profiles for the Cauchy and Student's t distribution assumption remain similarly tight for the two

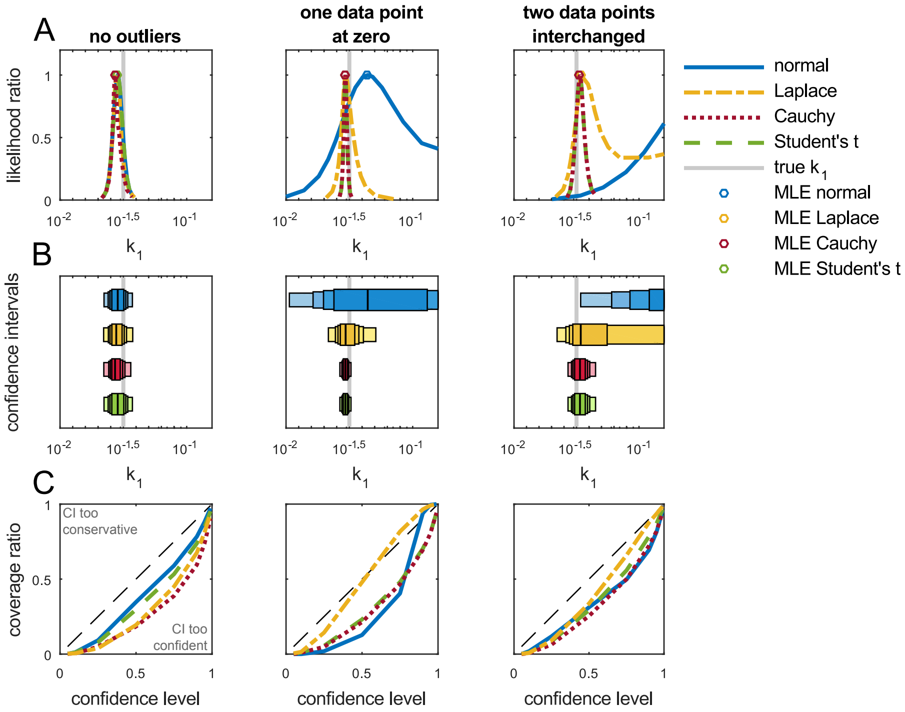

Figure 5.4: Uncertainty analysis for the conversion process. (**A**) Normalized profile likelihoods for the normal (blue), Laplace (yellow), Cauchy (red) and Student's t distribution (green). The vertical grey line displays the true value for $k_1 = 10^{-1.5}$. The profiles are computed for the cases displayed in Figure 5.1A,D and E. (**B**) The corresponding profile likelihood based CIs for different levels of confidence, $80\%, 90\%, 95\%$ and 99%, are indicated by bars from dark to light colors. (**C**) The coverage ratio is plotted against the confidence level. The dashed black line indicates identity. Lines above the identity line signify too conservative CIs (bigger CIs than necessary) and lines below the identity line indicate too confident CIs (CIs are too small). For the computation of the coverage ratio all 10^3 cases of each outlier scenario were considered.

scenarios with outliers. The profile for the Laplace distribution shows a bimodality in the case of *two data points interchanged*, but the MLE is still close to the true value. In contrast, the profiles of the normal distribution assumption broadens in both scenarios with outliers and the MLEs move away from the true parameter value. The profiles for k_2 are similar and are shown in Appendix Figure B.1A.

A parameter estimate should be reported along with its CI. We computed the CIs based on profile likelihoods according to Equation (2.8). The CIs obtained for the kinetic parameter k_1 for all outlier scenarios are displayed in Figure 5.4B also corresponding to the cases shown in Figure 5.1A,D and E. The CIs for parameter k_2 are to be found in Appendix B Figure B.1B. The differently colored bars represent the 80%, 90%, 95% and 99% confidence intervals from dark to light colors. In the case of *no outliers* the confidence intervals of

Table 5.1: CIs with confidence level (CL) that contain the true value $\log_{10}(k_1) = -1.5$, corresponding to the cases displayed in Figure 5.1A, D and E.

	no outliers		one data point at zero		two data points interchanged	
	CL	CI	CL	CI	CL	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]
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]

the four distribution assumptions span approximately the same parameter range. For the normal, Laplace and Student's t distribution the MLEs (vertical lines) are located at the same position close to the true value $10^{-1.5}$. Only for the Cauchy distribution the MLE is located at a lower value ($10^{-1.5657}$) within the 80% CI. If one data point is set to zero the CIs for the normal distribution become large, spanning more than the shown x-axes. There is a large uncertainty attached to the MLE, which assumes a value 40% larger ($10^{-1.3549}$) than the true value. The CIs for the Cauchy and Student's t distribution coincide and are tighter than for the data without outliers. They are closely located to the true value, but they do not account for the uncertainty the outlier introduced. Contrary, the Laplace distribution leads to larger CIs compared to the *no outliers* case, accounting for the additional uncertainty due to the outlier. Also, the MLE is located reasonably and close to the true value. For *two data points interchanged* the CIs of the normal distribution are not captured on the x-axes anymore. These CIs obviously no longer carry any useful information and do not even contain the true value. The MLEs found by Laplace, Cauchy and Student's t distribution are still located close to the true value. Again, the Laplace CIs are broader and take into account the uncertainty in outlier corrupted data, whereas the CIs of Cauchy and Student's t distribution are narrow, but with reasonable parameter values. The smallest CIs which contain the true value of k_1 are to be found in Table 5.1. Remarkably, the CIs of the Student's t distribution coincide in scenario *no outliers* with the CIs of the normal distribution and in the scenarios which include outliers they coincide with the CIs of the Cauchy distribution. This behavior is based on the degrees of freedom, as the Student's t distribution is able to adapt to both distributions. The appropriateness of the size of the CIs is analyzed in the following using the coverage ratio.

An important assessment criterion for the uncertainty of parameter estimates is the coverage ratio, see Section 2.4. The CR was computed based on profile based confidence intervals for various confidence levels $(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99)$, see Equation (2.8). We considered a multivariate CR, i.e., all parameters are investigated together, not separately. In Figure 5.4C, the CR is plotted against the confidence level as in Schelker et al. 2012. The CR is almost always lower than the confidence level indicating that the uncertainty in the estimates is underrated across all scenarios and distribution

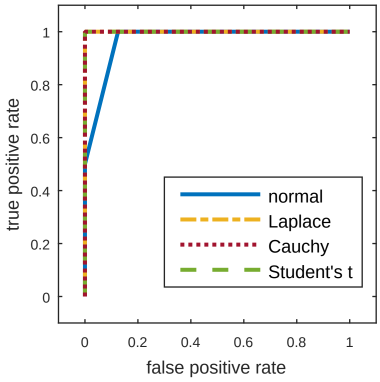

Figure 5.5: ROC curves. Detection of outliers with the different distribution assumptions considering the example displayed in Figure 5.1E.

assumptions. The true value is less often located within the interval as stated by the confidence level. This means that the size of the confidence intervals is underestimated; the CIs are overconfident. This can be explained by incomplete knowledge about the concrete outlier distribution in the case of outlier corrupted data. In presence of outliers the Laplace distribution provides the best coverage, indicating that the CIs have an appropriate size and are the most reliable throughout all outlier scenarios.

The uncertainty analysis showed that the confidence in the estimates is in general overestimated, even in the case of *no outliers*. Especially the Cauchy and Student's t distribution assumptions yield too small confidence intervals in presence of outliers. However, the MLEs are always close to the true parameter value. The Laplace distribution leads to more reasonable confidence intervals with an accurate MLE. Using the normal distribution assumption results in intervals that provide no useful information with wrong MLEs for outlier corrupted data.

5.7 Outlier detection

Robust parameter estimation can also be used as tool to identify outliers. Conveniently the percentiles of the residual distribution are used, which can be computed by using the inverse cumulative distribution function. As rule of thumb, the *three-sigma rule* or *two-sigma rule* is commonly applied for the normal distribution, which are special cases of a Z-value test (Aggarwal 2015). These rules state how many standard deviations the data is allowed to deviate from the mean, i.e., two and three standard deviations, which corresponds to the interval that contains 95% or 99.7% of the values, respectively. Data points outside of these intervals are classified as outliers. A standard method to compare different classifiers are receiver operating characteristic curves (ROC) (Metz 1978). In Figure 5.5 the ROC is shown for the example of *two data points interchanged* depicted

Table 5.2: True positive rates for varying confidence levels α considering the example displayed in Figure 5.1E.

	α				
	0.003/2	0.05/2	0.1	0.2	0.3
normal	0	0	1/2	1/2	1
Laplace	0	1	1	1	1
Cauchy	0	1/2	1	1	1
Student's t	0	1/2	1	1	1

in Figure 5.1E using $\alpha = (0, 0.003/2, 0.05/2, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)^T$. The corresponding percentiles are computed for the different distributions employing the inverse cumulative distribution function. To obtain the ROC the true positive rate, also called sensitivity is plotted against the false positive rate (1-specificity). The true positive rate states how many outliers have been correctly identified, whereas the false positive rate measures how many non-outliers are falsely classified as outliers. The computation of the area under the curve (AUC) gives for the Laplace, Cauchy and Student's t distribution in each case AUC = 1, whereas we obtain for the normal distribution AUC = 0.97. In Table 5.2 the true positive rates are listed to better show the actual difference between the methods better. For all the displayed α values the false positive rate is zero, i.e., none of the no-outliers is detected falsely as outlier, thus it is not visible in the ROC plot. If the *three-sigma rule*, the 99.7% percentile ($\alpha = 0.003/2$), is applied, none of the four distributions detects the outliers. The outliers are apparently not extreme enough. Using, however, the *two-sigma rule* (95%), the Laplace distribution assumption allows to identify both outliers (true positive rate equals 1). The Cauchy and Student's t distribution lead to the detection of the first outlier (true positive rate equals 0.5), whereas the normal distribution assumption does not allow the detection of either of the two outliers. The table indicates that a less conservative criterion could be more appropriate, as for example the 90% percentile. In this case Laplace, Cauchy and Student's t distribution identify the two outliers, while the normal distribution detects only one of the two outliers.

To conclude, the Laplace distribution seems to be best suited for the detection of outliers with robust parameter estimation. In the case of Cauchy and Student's t distribution it should be considered to choose a less conservative criterion as the *two-* or *three-sigma rule*, since the false-positive rate is still zero for less conservative criteria.

5.8 Limitations of Cauchy and Student's t distribution

As already mentioned in Section 4.4, Fernández et al. (1999) have revealed difficulties in global optimization assuming a Student's t distribution, if the model is able to fit too many data points "exactly" (up to numerical accuracy) and $\nu \in \mathbb{R}^+$. The theorem formulated by Fernández et al. for regression models can be translated to quantitative dynamic models. Defining the number of observations that can be fitted exactly, i.e., $\bar{y}_{i,k} = y_i(t_k, \theta_0)$ to be $s(\theta_0)$ for a parameter vector θ_0, the authors showed that if $\nu < s(\theta_0)/(n_t - s(\theta_0)) = d_0$ the likelihood function can take arbitrarily large values as $\sigma \to 0$. Choosing $\nu > s(\theta_0)/(n_t - s(\theta_0)) = d_0$ yields $\mathcal{L}_D(\theta_0) = 0$ as σ tends to zero. For small degrees of freedom ($\nu \leq 2$) the Student's t distribution has no finite variance, see Section 4.4. Scale parameters close to zero are not useful for modeling as the corresponding distribution does not reflect the variation in the data. The distribution concentrates all its mass on single data points, neglecting other residuals, i.e., the model overfits single data points. Thus, it is required to restrict the parameter space of the degrees of freedom to be greater than d_0 to avoid overfitting (Jones et al. (2003) and Taylor et al. (2004)). In our analysis we restricted the degrees of freedom to be greater or equal one, which is a more conservative choice than required. Up to 4 of the 10 data points can be fitted "exactly" without overfitting the data.

Since the Cauchy distribution corresponds to the Student's t distribution with one degree of freedom, the same problem arises for the Cauchy distribution, which has no defined variance. Employing the formulas from above with fixed $\nu = 1$ yields that maximum likelihood estimation is only reasonable for the Cauchy distribution if $d_0 < 1$. Consequently, the Cauchy distribution should not be applied if half or more of the data points can be fitted exactly by the model, i.e., $s(\theta_0) \geq n_t - s(\theta_0)$. If this is the case, however, there is already the problem of overfitting the data, meaning that the model describes the noise instead of the dynamics. This results in less predictive power although the model fit to the data might look appropriate (Villaverde et al. 2014). In practice it is not easy to determine when an observation is fitted "exactly" and approximations have to be made.

This problem was analyzed for the conversion example. In our analysis so far we used ten measurements for the parameter estimation. To analyze the above described problem, the number of measurements was varied to achieve different numbers of "exactly" (up to a threshold of $\epsilon = 10^{-4}$) fitted data points. If $n_t = 4$, for example, most of the times half of the data points, $s(\theta_0) = 2$, can be fitted "exactly". We consider the following data sets for the case of *no outliers*: $\mathcal{D}^{10}$ for time points $t \in \{0, 5, 10, 15, 20, 25, 30, 40, 50, 60\}$, $\mathcal{D}^4$ for time points $t \in \{0, 20, 30, 50\}$ and $\mathcal{D}^3$ for time points $t \in \{15, 30, 60\}$. All data sets comprise 100 noise realizations. Parameter estimation was again performed and the lower parameter bound for the scale parameter was set to 10^{-10} for all distributions. Histograms showing the residuals for the different distribution assumptions are displayed in Figure 5.6A. The histograms are normalized so that the area of the bars sums up to one. The curve represents the corresponding distribution using the mean value of the

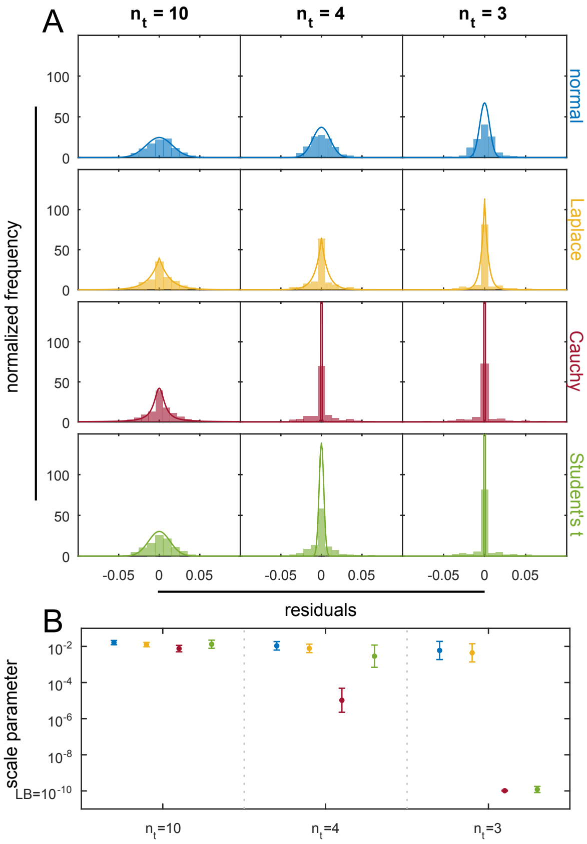

Figure 5.6: Limitations of Cauchy and Student's t distribution. Illustration of the problem arising if too many data points can be fitted exactly by the model on data sets with different number of data points. (**A**) Normalized histogram of the residuals of all 100 data sets when the parameter estimation is performed assuming a normal (blue), Laplace (yellow), Cauchy (red) or Student's t (green) distribution in the case of ten, four and three data points. The curve represents the corresponding probability density of the normal (—), Laplace (—), Cauchy (—) or Student's t (—) distribution using the estimated mean value of the distribution specific parameters over all 100 data sets. (**B**) Visualization of the corresponding scale parameters, $\sigma^{(N)}$ for the normal (●), b for the Laplace (●), γ for the Cauchy (●) and $\sigma^{(T)}$ for the Student's t distribution (●).

estimated distribution parameters $(\sigma^{(N)}, b, \gamma, \sigma^{(T)})$, which are displayed in panel B. In the case of ten data points all distributions capture the whole variation in the residuals. If only three data points are used for the parameter estimation, most of the time two of the data points can be fitted exactly, which gives $d_0 = 2$. Here the described problem becomes apparent: By decreasing the scale parameter to small values close to the lower bound, the Cauchy and Student's t distribution concentrate their mass on the data points the model can fit "exactly". The remaining data points are not appropriately represented anymore. The likelihood function takes high values due to the distribution's peak behavior approximating a delta distribution. The amplitude of the Cauchy distribution is defined

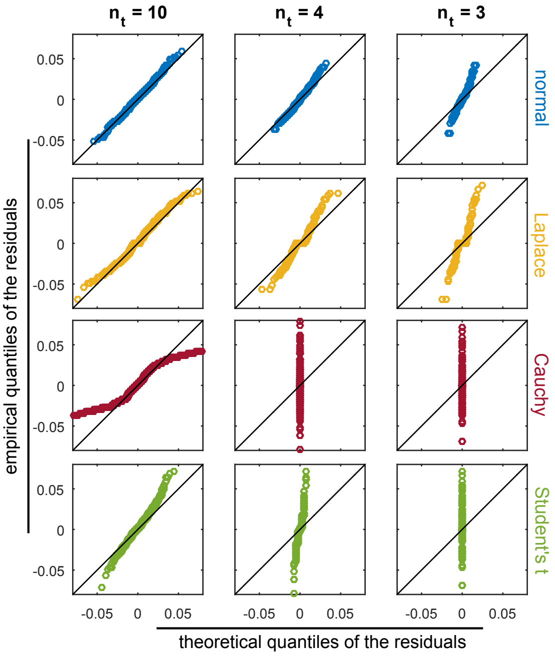

Figure 5.7: Quantile-Quantile plot of the residuals. Comparison of the empirical and theoretical quantiles for the normal distribution (blue), Laplace distribution (yellow), Cauchy distribution (red) and Student's t distribution (green).

as $1/(\pi\gamma)$, which yields large values for small scale parameter values, resulting in large likelihood values for those single data points that are fitted "exactly". Thus, according to the criterion of Fernández et al. (1999), ν needs to be larger than two to resolve the issue for the Student's t distribution, and the Cauchy distribution cannot be applied. The case of four data points constitutes the transition case as $d_0 = 1$ in most cases. The values for the distribution parameters in the other outlier scenarios (for $n_t = 10$) are to be found in the Appendix Figure B.2A.

In Figure 5.7 the empirical quantiles of the residuals are compared with the theoretical quantiles in a Q-Q plot (quantile-quantile-plot), considering the quantiles $(i - 0.5)/n$ for all $i = 1, \ldots, n$, 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 ten data points, the empirical and theoretical quantiles coincide well for all distributions, except for the heavier-tailed distributions in the tails. This is expected for the *no outliers* case and 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 Student's t distribution show that the distribution does not reflect the spread in the residuals. In the case of $n_t = 3$ problems for the normal and Laplace distribution are visible as well. The sample size is apparently too small for reasonable parameter estimation.

Thus, attention must be paid when applying the Cauchy and Student's t distribution if the model is too flexible and overfitting is to be expected. In these cases the distributions overfit single data points by neglecting the remainder of the data points. For the Student's t distribution this issue can be resolved by setting an appropriate bound for the degrees of freedom. The Cauchy distribution, however, can only be applied if the model does not allow the exact fit of half or more of the data points.

Chapter 6

The Jak/Stat signaling pathway

As real-world application, the Jak/Stat signaling pathway, is investigated. To assess the robustness of the methods in the case of real data, the experimental data of the Jak/Stat pathway was also modified in accordance with the outlier scenarios. The pathway plays a key role in the differentiation, proliferation and migration of cells in the erythropoietic system (Rawlings et al. 2004).

6.1 Biological overview of the signaling pathway

Three principal components are involved in the signaling cascade, the hormone Erythropoietin (Epo), the Janus family of kinases (JAK)-signal transducer and the activator of transcription 5 (STAT5). Intracellular activation is triggered by binding of the upstream activation factor Epo to its receptor (EpoR). This extracellular stimuli leads to phosphorylation of the EpoR cytoplasmic domain by the tyrosine Janus kinase 2 (JAK2). The latent transcription factor STAT is phosphorylated upon recruitment to the activated receptor (pEpoR). The cytoplasmic phosphorylated STAT (pSTAT) dimerizes and the dimer (pSTAT_pSTAT) enters the nucleus to initiate the transcription of target genes. Afterwards the STAT molecules are recycled to the cytoplasm (Bachmann et al. 2011; Rawlings et al. 2004; Swameye et al. 2003). A schematic representation of the Jak/Stat signaling pathway can be found in Figure 6.2A with arrows indicating biochemical reactions.

6.2 Experimental data

Swameye et al. (2003) have recorded average concentrations by quantitative immunoblotting of pEpoR, pSTAT (phosphorylated STAT as monomer and dimer) and tSTAT (unphosphorylated and phosphorylated STAT) in the cytoplasm, cf. orange boxes in Figure 6.2A. The data was recorded for time points $t \in \{0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 25, 30, 40, 50, 60\}$ minutes, see Figure 6.2B. The data does not include any outliers, hence, we artificially introduced outliers in accordance with the outlier scenarios, presented in Chapter 3, for all three observables in order to examine the robustness of the methods. It is assumed that, for example in the case of *one data point at zero*, the measurement device fails to record all three observables at the same time. In Figure 6.2C the resulting data is displayed, if the device fails to record at time point $t = 8$ min. As the observables were recorded at 16 time points, this procedure leads 16 possible *one data point at zero* scenarios and choosing two out of 16 gives 120 *two data points interchanged* scenarios, for an example see Figure 6.2D.

6.3 Model description

Based on these time-resolved measurements, Swameye et al. have introduced a mathematical model that describes the Epo induced Jak/Stat signaling cascade. This model was studied extensively in literature with regard to identifiability (Raue et al. 2009), sensitivity analysis (Kazeroonian et al. 2016) and comprehensive input estimation (Schelker et al. 2012). An extension of the core model was studied in Bachmann et al. (2011) considering also negative feedback regulators. The ODE system of the core model is based on the description in Kazeroonian et al. (2016),

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

with kinetic parameters p_1, p_2, p_3 and p_4 and initial concentration STAT_0. The delay reaction of STAT binding to the DNA in the nucleus is modeled as linear chain approximation with intermediate steps nSTAT1, …, nSTAT5. The volume of the two compartments, cytoplasm and nucleus, are constants $\Omega_{\text{cyt}} = 1.4$ pl and $\Omega_{\text{nuc}} = 0.45$ pl (Raue et al. 2009).

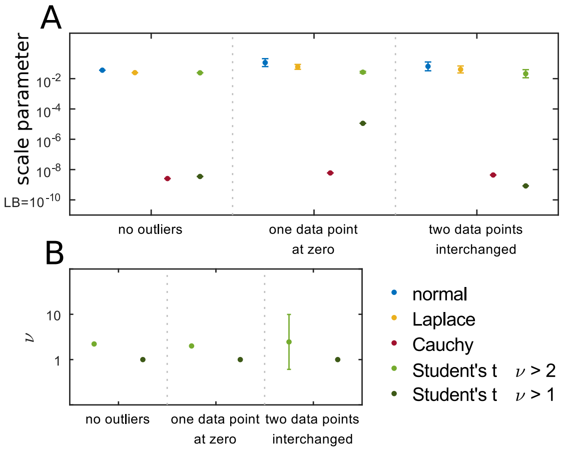

Figure 6.1: Distribution specific parameters. Mean values of the distribution specific parameters for the three outlier scenarios. For the Student's t distribution two cases are distinguished, $\nu > 1$ and $\nu > 2$.

The observables are given by

$$\begin{aligned} y_1 &= \text{offset}_{\text{pSTAT}} + \frac{\text{scale}_{\text{pSTAT}}}{\text{STAT}_0}(\text{pSTAT} + 2\text{pSTAT\_pSTAT}) \\ y_2 &= \text{offset}_{\text{tSTAT}} + \frac{\text{scale}_{\text{tSTAT}}}{\text{STAT}_0}(\text{STAT} + \text{pSTAT} + 2\text{pSTAT\_pSTAT}) \\ y_3 &= u(t)\,, \end{aligned}$$

for which y_1 is the total concentration of phosphorylated STAT in the cytoplasm (pSTAT), y_2 the total concentration of STAT in the cytoplasm (tSTAT) and y_3 the concentration of phosphorylated Epo receptors (pEpoR), see orange boxes in Figure 6.2. The pEpoR concentration is modeled as time-dependent cubic spline function u with five parameters $sp_1, \dots, sp_5$. Scale parameters were introduced by Swameye et al. (2003) because only relative protein amounts could be measured by the experimental setup. The initial concentration STAT_0 was set to one, as by Schelker et al. (2012), in order to tackle structural identifiability problems shown in (Raue et al. 2009). This leads to the parameter vector

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

For the optimization we considered the log-transformed parameters $\log_{10}(\theta) = \log_{10}(\xi, \varphi)$.

6.4 Parameter estimation

In the further analysis the Cauchy distribution is excluded. As in the end of the conversion reaction example presented, the Cauchy distribution is not appropriate if too many

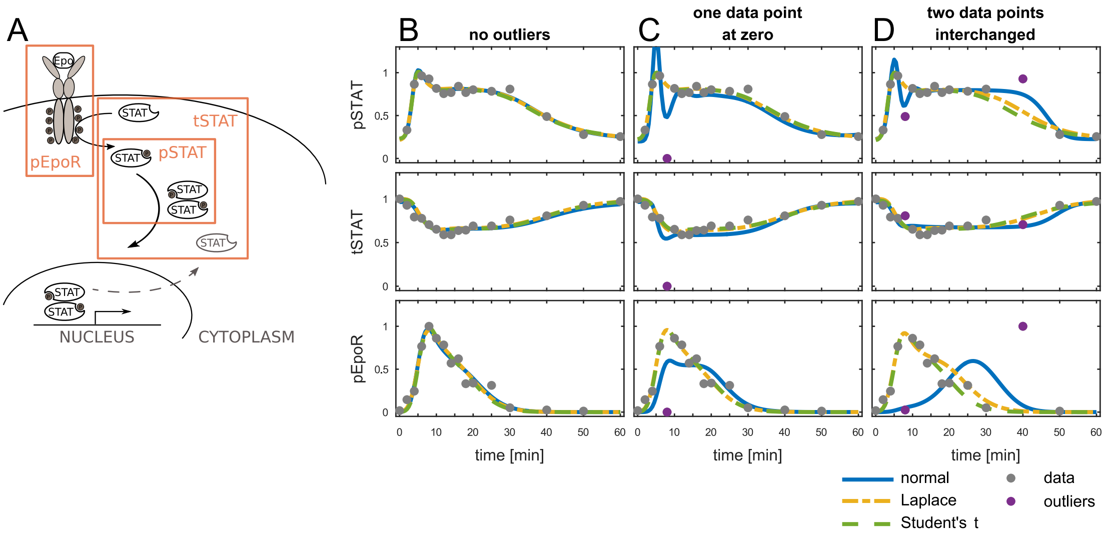

Figure 6.2: Jak/Stat model and trajectories. (**A**) Schematic representation of the Jak/Stat signaling pathway. The observables are shown in orange boxes and arrows represent biochemical reactions. (**B**) Parameter estimation results for the different distribution assumptions in the *no outliers* scenario. (**C**) Results for the scenario *one data point at zero*. The outlier (purple circle) is introduced at time point $t = 8$ min. (**D**) *two data points interchanged* at time points $t = 8$ and $t = 40$ min. For this application example the Cauchy distribution was excluded and the Student's t distribution was restricted to $\nu > 2$.

data points can be fitted exactly by the model. The analysis revealed that the model is able to fit more than half of the data points for one observable (pSTAT) exactly, i.e., $d_0 > 1$. This leads to unreasonable large likelihood values which are achieved by overfitting. As the objective function is computed as sum over all observables this leads to an arbitrarily large objective function. The scale parameters for all distributions are shown in Figure 6.1A. The scale parameter for the Cauchy distribution takes for all scenarios small values approaching the lower bound. The same holds true for the Student's t distribution with $\nu > 1$. The degrees of freedom of the Student's t distribution were, therefore, restricted to be larger than two. This restriction leads to appropriate scale parameter values (see Figure 6.1A). Only in four cases the lower bound of two was too small since one more data point could be fitted exactly, thus, in these cases the lower bound was increased to 2.2 in accordance with the criterion of Fernández et al. (1999). The corresponding degrees of freedom values are displayed in Figure 6.1B.

Note that we used the same ν for all three observables rather than a separate ν for each observable. This choice is based on model selection performed for the *no outliers* scenario using the BIC . The model with three ν parameters (with lower bounds $(1.3, 0.8, 0.8)^T$) leads to a BIC value of -87.95, whereas the model with one degree of freedom ($\nu > 2$) gives a BIC of -91.91. Hence, the model with one ν is more appropriate as it describes the data equally well by employing less parameters.

For the modified data sets multi-start local optimization was performed similarly as in the illustrative example. The parameter space for the multi-start approach was chosen for the dynamic parameters as $\Xi = [10^{-5}, 10^3] \times [10^{-3}, 10^6] \times [10^{-5}, 10^3] \times [10^{-3}, 10^6] \times [10^{-5}, 10^3]^4 \times [10^{-6}, 10^3] \times [10^{-5}, 10^3]^4$. The distribution specific parameters were chosen separately for each observable except for the degrees of freedom of the Student's t distribution. The scale parameters, $\sigma_i^{(N)}, b_i$ were searched within the interval $[10^{-5}, 10^3]$ and the parameters for the Student's t distribution, $\sigma_i^{(T)}, \nu$ within $[10^{-10}, 10^3] \times (2, 10^5]$. The lower bound for the scale parameter was decreased so that the overfitting problem could be easily detected. 100 multi-starts were generated with Latin hypercube sampling within the specified parameter bounds. If less than five starts had converged according to the likelihood ratio test (Equation (2.4)) the number of start points was increased by 100. As local solver the interior-point algorithm was used for the normal and Laplace distribution and the trust-region-reflective algorithm for the Student's t distribution. Only in three cases convergence problems for the Student's t distribution arose which could be resolved by using the interior-point algorithm. Parameter estimation was performed for the *no outliers* case, the 16 *one data point at zero* scenarios and the 120 *two data points interchanged* scenarios.

6.5 Qualitative analysis of the results

In Figure 6.2B-D the resulting fits for the three scenarios are displayed. For the *no outliers* scenario the three distribution assumptions deliver similar model trajectories close to the data points (panel B). In panel C the model is calibrated to data with an introduced measurement failure at time point $t = 8$ minutes. Using the normal distribution assumption in the parameter estimation leads to a different fit than to the data without outliers. The parameter estimation obviously tries to accommodate the outlier. Laplace and Student's t distribution assumption lead to trajectories that equal the trajectory for the *no outliers* case, even in presence of the outlier. The same holds true for the scenario *two data points interchanged* in panel D. The fit found by assuming a normal distribution is visibly drawn towards the outliers. For this case, the assumption of a Student's t distribution delivers a model trajectory closer to the trajectory in panel B as the assumption of a Laplace distribution. But this depends on the degree of abnormality of the outliers; the further away the outlier, the more the Laplace distribution is misled. Two more examples for the scenario *two data points interchanged* are presented in Figure 6.3. In the first case, displayed in panel A, the Student's t distribution is still able to find a trajectory similar to the trajectory for the data without any outliers. The Laplace distribution assumption shows, on the contrary, only a slightly better fit as the normal distribution assumption in terms of similarity to the model trajectory obtained for *no outliers*. The outliers in panel B are less extreme than in panel A and the Laplace distribution yields only a slightly more distorted fit than the Student's t distribution.

In summary, the qualitative analysis of the real data for the Jak/Stat pathway supports our finding for the artificial data of a conversion process. In presence of outliers, the

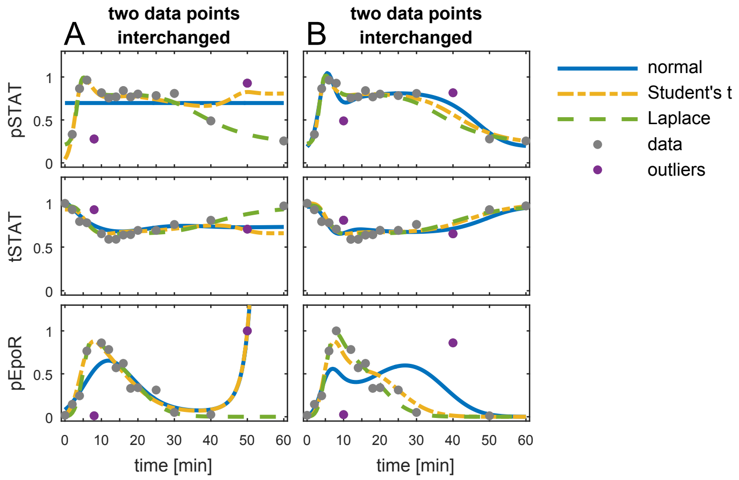

Figure 6.3: Jak/Stat example fits. Two more examples for the scenario *two data points interchanged*. (**A**) Values at time points $t = 8$ and $t = 50$ minutes were interchanged (purple circles). (**B**) Outliers at time points $t = 10$ and $t = 40$ minutes.

Laplace and Student's t distribution yield more reliable fits to the main behavior of the data than the normal distribution as they are closer to the fits obtained for the *no outliers* scenario.

6.6 Comparison of estimation accuracy

Since this is a real-world example we do not know the true parameter values. The model accuracy can still be analyzed by the MSE, using as "true" parameter values the values we obtained for the *no outliers* scenario for each distribution assumption (Figure 6.2B). The MSE was computed for each parameter separately. Figure 6.4A shows the logarithm of the MSE for the scenario *one data point at zero*. The errorbars display the 95% bootstrap percentile confidence intervals. For all parameters the MSE of the model using the normal distribution assumption is higher than for the Student's t and Laplace distribution. This supports our findings from the artificial data of a conversion reaction. Using a heavier-tailed distribution leads to more reliable estimates for outlier corrupted data, as they are closer to the estimates found for the corresponding data without outliers. In the case of *two data points interchanged* the difference is not so clear. This might be due to the fact, that in some of the cases the interchange of two data points does not lead to gross outliers. But still the Laplace and Student's t distribution obtain for most parameters smaller MSE values. In panel C the MSE for the parameter vector is shown. In total, the Laplace and Student's t distribution lead to a lower MSE for outlier corrupted data than the normal distribution.

As in the statistical analysis for the conversion reaction the heavier-tailed distributions

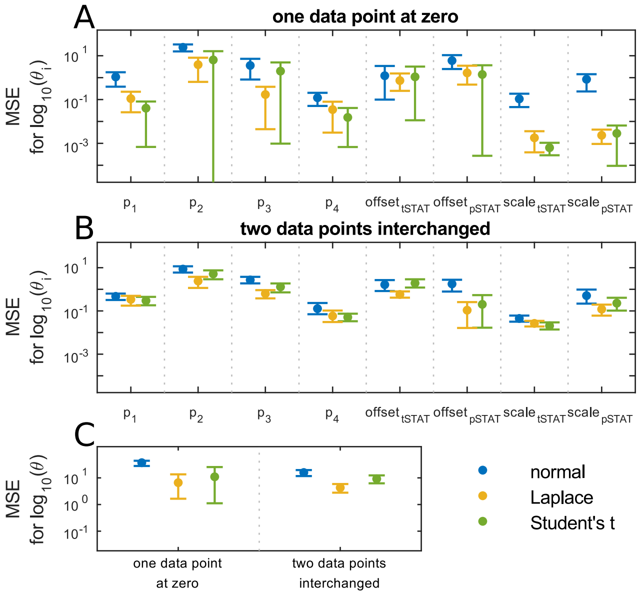

Figure 6.4: Estimation accuracy for the Jak/Stat example. Logarithm of the mean squared error for all 13 parameters describing the model dynamics for the scenario *one data point at zero* (**A**) and for *two data points interchanged* (**B**). As true parameters the MLEs for the case without outliers was used. Note that the parameters of the distributions ϕ cannot be compared.

yield a smaller MSE of the parameter estimates than the normal distribution for outlier corrupted data. The Cauchy distribution could not be used as the model can fit more than half of the data points exactly. Consequently, the degrees of freedom for the Student's t distribution had to be restricted to be larger than two.

Chapter 7

Summary

We proposed three heavier-tailed distributions, the Laplace, Cauchy and Student's t distribution for estimating parameters of ODE models from outlier corrupted data. These methods are well established in robust regression and it was shown that they are also beneficial in the context of dynamical models. To demonstrate and examine the properties of the new methods outlier corrupted was generated according to three outlier scenarios, describing biologically motivated mechanisms that produce outliers.

At first a simulation study for a conversion process was performed and the investigation of the obtained model trajectories gave already a good impression that heavier-tailed distributions are advantageous in presence of outliers. The assumption of normally distributed residuals did not lead to reasonable results for outlier corrupted data. While heavier-tailed distributions enabled the reconstruction of the true trajectory, the normal distribution did not allow a reliable inference of the parameters from outlier corrupted data. The statistical results supported that finding. The mean squared error for outlier corrupted data was significantly smaller for the heavier-tailed distributions than for the normal distribution. It is therefore required to include a precedent outlier detection and removal when using the assumption of normally distributed residuals. For data without outliers all methods yielded similar results and the true trajectory was successfully reconstructed. Model selection revealed the presence of outliers, as the heavier-tailed distributions are favored in these cases. It was further shown that the use of robust methods does not decrease the performance, regarding convergence and computation time. The uncertainty analysis indicated that the obtained confidence intervals are too small to appropriately capture the uncertainty of the estimates throughout all methods, due to incomplete knowledge about the concrete outlier distribution. Furthermore, it was shown that the novel approach can also be used to identify outliers by exploiting the percentiles of the distributions. The overfitting problem for the Cauchy and the Student's t distribution was examined, which arises if too many data points can be fitted exactly by the model. For the Student's t distribution this problem can be resolved by increasing the lower bound for the degrees of freedom, but the Cauchy distribution can only be applied if less than half of the data points can be fitted exactly. Consequently, the Cauchy distribution had to be excluded in the analysis of the Jak/Stat signaling pathway and the lower bound for the degrees of freedom of the Student's t distribution was increased.

The study for the real data of the Jak/Stat signaling pathway affirmed the previous results. For the real data the heavier-tailed distribution assumptions were also less affected by the artificially introduced outliers than the normal distribution. Heavier-tailed distributions enabled a consistently accurate parameter estimation for data with and without outliers. The normal distribution assumption led to distorted trajectories in presence of outliers, which deviated considerably from the trajectory in absence of outliers. The mean squared error for the outlier corrupted data sets was considerably higher for the normal distribution than for the heavier-tailed distributions, Laplace and Student's t.

In conclusion, the use of heavier-tailed distributions constitutes indeed a robust approach to parameter estimation for ODEs. It is a reasonable alternative to outlier detection and removal that does not require an alteration of the data set. Thus, the method is a means to improve parameter estimation for outlier corrupted data that allows the inference of reliable parameter estimates in presence of outliers.

Chapter 8

Discussion and outlook

In the case of outlier corrupted data the assumption of a normal distribution was shown not to be reasonable. Using objective functions which consider the squared residuals gives greater weight to outliers. Consequently, larger deviations contribute considerably more to the objective function than smaller deviations. As seen in the application examples this results in bad fits to the main behavior of the data set in presence of outliers.

Exploiting the ability of heavier-tailed distributions to give less weight to outliers, leads to a robust approach to parameter estimation for outlier corrupted data. Three different heavier-tailed distributions were proposed that reduce the error in parameter estimation in presence of outliers significantly. The heavier-tailed distributions capture, however, not the true outlier distribution, which leads to confidence intervals that do not reflect the true coverage. Hence, the uncertainty in the parameter estimates is underestimated, resulting in too small confidence intervals. Attention should also be paid if the model is too flexible, as the Student's t and Cauchy distribution can be only used limited. However, if this problem arises the model itself should be reconsidered. Furthermore, the Laplace distribution is not differentiable at the location parameter, which might yield problems in the optimization. Although we have not experienced any difficulties in our analysis, this establishes the need for further studies.

As all of the presented distribution assumptions have their strengths and weaknesses, the choice remains problem-dependent. One approach is to apply all different distribution assumptions and compare the results, although this is only possible if the computational time for one parameter estimation is not too high. The analysis showed that heavier-tailed distributions are advantageous in the case of outlier corrupted data, but determining the best suited distribution remains not only case-dependent, but still open to further investigation. Other heavier-tailed distributions could be examined as part of future work. A further class reasonable next to heavier-tailed distributions are skewed distributions as errors are in general not symmetric, e.g. the skew t distribution (Jones et al. 2003). Furthermore, Reed (2006) introduced the normal-Laplace distribution, which results by convolution of a normal distribution with an asymmetric Laplace distribution. The symmetric version employs three parameters, whereas the asymmetric version requires four parameters. The distribution is differentiable and has longer tails, thus, it takes a reasonable intermediate position between the normal and Laplace distribution. However, the probability density function assumes a complicated form with more parameters, which might complicate the parameter estimation process. Kotz et al. (2012) provide a good starting point for the application of symmetric and asymmetric Laplace distributions.

In this work, an approach to robust parameter estimation for quantitative dynamic models was presented. Since biological systems are highly sensitive to their environments, measurement errors are common in biological data. Applying the standard approach, these errors have an distorting effect on the model calibration. The results of this work showed that in the novel method erroneous measurements propagate less to the parameter estimates. Therefore, the new proposed approach enhances model calibration and therefore improves the investigation of biological systems.

Appendix A

Derivatives of the likelihood function

In order to improve the derivative-based optimization approach the analytic gradient and an approximation of the Hessian matrix were provided to the local solver. In the following formulas the general notation for the distribution specific parameters is chosen, depending also on the time points t_k, i.e., $\varphi_{i,k}$ as an explicit time-dependence is possible, although not considered for this work.

A.1 Normal distribution

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

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

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

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

In the optimization an approximation was used. For this purposes the last term including the second order sensitivities was neglected, assuming that the difference between measurements and predicted observable $\bar{y}_{ik} - y_i(t_k, \theta)$ is small. This can be still considered valid for outlier corrupted data, as generally only a small number of outliers is included in a data set.

A.2 Laplace distribution

Using alternatively the Laplace distribution as assumption for the residual distribution the gradient of the log-likelihood function reads for $l = 1, \ldots, n_\theta$

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

In the following we assume that $\bar{y}_{ik} - y_i(t_k, \theta) \neq 0$. This is not contradictory to the previous assumption stated for the approximation of the Hessian in the normal distribution assumption case. The difference is small but will usually not be exactly zero. The Hessian of the log-likelihood function using the Laplace distribution is given by,

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

where $l, m = 1, \ldots, n_\theta$. Note that the term including the second order sensitivities cannot be neglected in this case. This makes the computation of the Hessian very slow and it is advised to use an algorithm that does not rely on a user-supplied Hessian.

A.3 Cauchy distribution

The gradient for the log-likelihood function assuming a Cauchy distribution is given for $l = 1, \dots, n_\theta$ by

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

In consequence the Hessian with $l, m = 1, \dots, n_\theta$ is calculated as

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

For the approximation of the Hessian it is again possible to neglect the term including the second order sensitivities.

A.4 Student's t distribution

Assuming a Student's t distribution leads to the following gradient for the log-likelihood for $l = 1, \ldots, n_\theta$:

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

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

$$
\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,k}(\theta)+1}{2}\Big) - \psi\Big(\frac{\nu_{i,k}(\theta)}{2}\Big) - \log\left(1 + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_{i,k}(\theta)\sigma_{i,k}(\theta)^2}\right) \\
& - \frac{1}{\nu_{i,k}(\theta)} + \frac{\nu_{i,k}(\theta)+1}{\nu_{i,k}^2(\theta)\sigma_{i,k}^2(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{1 + \frac{1}{\nu_{i,k}(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2}} \Bigg] \frac{\partial^2 \nu_{i,k}(\theta)}{\partial \theta_l \partial \theta_m} \\
& - \Bigg[\frac{1}{\sigma_{i,k}(\theta)} - \frac{\nu_{i,k}(\theta)+1}{1 + \frac{1}{\nu_{i,k}(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2}} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_{i,k}(\theta)\sigma_{i,k}^3(\theta)} \Bigg] \frac{\partial^2 \sigma_{i,k}(\theta)}{\partial \theta_l \partial \theta_m} \\
& + \left(\frac{\nu_{i,k}(\theta)+1}{1 + \frac{1}{\nu_{i,k}(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2}} \frac{1}{\nu_{i,k}(\theta)} \frac{(\bar{y}_{ik} - y_i(t_k,\theta))}{\sigma_{i,k}^2} \frac{\partial^2 y_i(t_k,\theta)}{\partial \theta_l \partial \theta_m} \right)^* \\
& + \frac{1}{2} \Bigg[\frac{1}{2}\psi_1\Big(\frac{\nu_{i,k}(\theta)+1}{2}\Big) - \frac{1}{2}\psi_1\Big(\frac{\nu_{i,k}(\theta)}{2}\Big) + \frac{1}{\nu_{i,k}^2(\theta)} \\
& \quad + \frac{1}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2}} \left(\frac{\frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2} - 1}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\sigma_{i,k}(\theta)^2}} - \frac{1}{\nu_{i,k}(\theta)} \right) \\
& \quad \cdot \frac{(\bar{y}_{ik} - y_i(t_k,\theta))^2}{\nu_{i,k}(\theta)\sigma_{i,k}^2(\theta)} \Bigg] \frac{\partial \nu_{i,k}(\theta)}{\partial \theta_l} \frac{\partial \nu_{i,k}(\theta)}{\partial \theta_m}
\end{aligned}
$$

$$
\begin{aligned}
&+ \Bigg[\frac{1}{\sigma_{i,k}^2(\theta)} + \frac{\nu_{i,k}(\theta)+1}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^4(\theta)} \\
&\quad \cdot \left(2 \frac{\frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} - \frac{3}{\sigma_{i,k}(\theta)} \right) \Bigg] \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_l} \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_m} \\
&+ \frac{\nu_{i,k}(\theta)+1}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} \frac{1}{\sigma_{i,k}^2(\theta)} \left(2 \frac{\frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} - 1 \right) \\
&\quad \cdot \frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \\
&+ 2 \frac{\nu_{i,k}(\theta)+1}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} \left(\frac{\frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}}{\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)}} - 1 \right) \\
&\quad \cdot \frac{(\bar{y}_{ik}-y_i(t_k,\theta))}{\sigma_{i,k}^3(\theta)} \left(\frac{\partial y_i(t_k,\theta)}{\partial \theta_l} \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_m} + \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_l} \right) \\
&+ \frac{\frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)} - 1}{(\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)})^2} \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^3(\theta)} \\
&\quad \cdot \left(\frac{\partial \nu_{i,k}(\theta)}{\partial \theta_l} \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_m} + \frac{\partial \nu_{i,k}(\theta)}{\partial \theta_m} \frac{\partial \sigma_{i,k}(\theta)}{\partial \theta_l} \right) \\
&+ \frac{\frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)} - 1}{(\nu_{i,k}(\theta) + \frac{(\bar{y}_{ik}-y_i(t_k,\theta))^2}{\sigma_{i,k}^2(\theta)})^2} \frac{(\bar{y}_{ik}-y_i(t_k,\theta))}{\sigma_{i,k}^2(\theta)} \\
&\quad \cdot \left(\frac{\partial \nu_{i,k}(\theta)}{\partial \theta_l} \frac{\partial y_i(t_k,\theta)}{\partial \theta_m} + \frac{\partial \nu_{i,k}(\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. It is again possible to neglect the term including the second order sensitivities (marked by *) because the difference of measurement and predicted observable is in general small.

Appendix B

Supplementary material for the conversion reaction

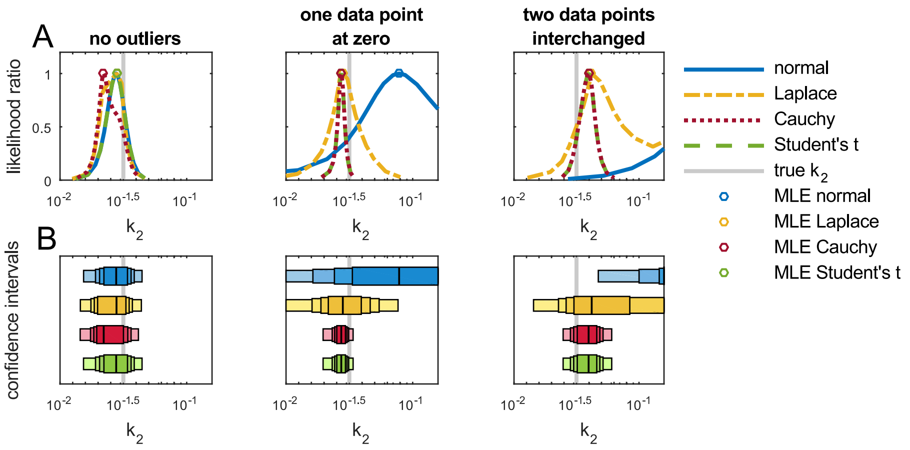

Figure B.1: Uncertainty analysis based on profile likelihoods for k_2. (**A**) Profiles for k_2. (**B**) Corresponding confidence intervals with different confidence levels (80%,90%,95% and 99%) indicated by the bars colored from light to dark. The vertical line indicates the MLE.

In the statistical analysis of the results the second parameter k_2 was neglected as the PLs and CIs are rather similar to the ones obtained for parameter k_1. In the case of *no outliers* the profiles of normal, Laplace and Student's t distribution overlap and are close to the true value, while the profile of the Cauchy distribution deviates from the others and is located further away from the true parameter value, see Figure B.1A. In scenario *one data point at zero* the profiles of Cauchy and Student's t distribution are very narrow and close to the true parameter value. The Laplace distribution shows a broader profile, yet the MLE is at the same position as for Cauchy and Student's t distribution. Whereas the profile of the normal distribution covers an extensive parameter range with a too large MLE. In the case of *two data points interchanged* the profile of the Laplace distribution has a similar bimodality as seen for the parameter k_1. Also for k_2 the profile

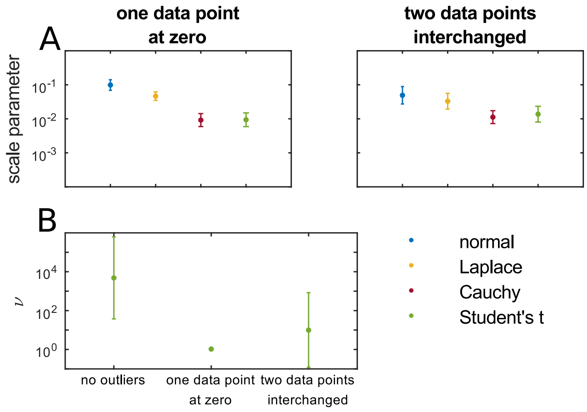

Figure B.2: Distribution specific parameters for the outlier scenarios. Mean values of the distribution specific parameters in the case of $n_t = 10$. (**A**) Scale parameters for the outlier scenarios *one data point at zero* and *two data points interchanged*. (**B**) Degrees of freedom of the Student's t distribution for the three outlier scenarios.

of the normal distribution covers not a reasonable parameter range far off the true value. The corresponding profile based confidence intervals are displayed in Figure B.1B.

The scale parameters assumed for all distributions reasonable values, see Figure B.2A. The described problem if too many data points are fitted "exactly" does not occur for ten data points. In Figure B.2B the estimated mean values of the degrees of freedom for the Student's t distribution are shown ± the standard deviation. In the case of *no outliers* the degrees of freedom assume larger values and consequently the Student's t distribution approximates the normal distribution. This is desired for the case without outliers as the normal distribution constitutes the "true" model. In presence of outliers ν becomes small, the distribution puts more weight in the tails, which is necessary to capture the outliers.

Appendix C

Supplementary material for the Jak/Stat signaling pathway

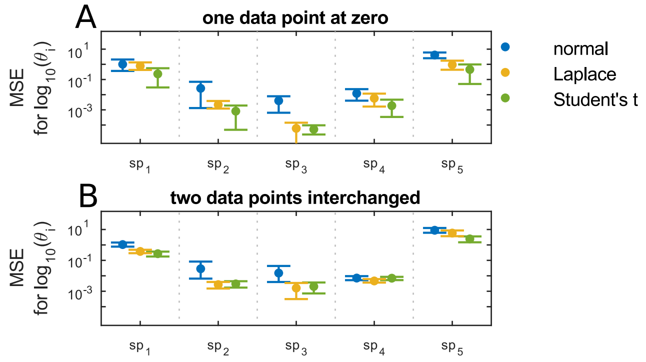

Figure C.1: MSE for spline parameters. Logarithm of the mean squared error for the spline parameters. The scenarios *one data point at zero* and *two data points interchanged* are compared to the scenario *no outliers*.

Since the spline parameters are of no biological interest, the MSE for those parameters was not shown in the main text. Figure C.1 displays the logarithm of the MSE for the spline parameters with errorbars indicating the 95% percentile bootstrap interval. In scenario *one data point at zero* the normal distribution leads to a higher MSE for all parameters. In the case of *two data points interchanged* this is not true for all parameters. As not all cases of this scenario lead to clear outliers, the normal distribution is able to adequately describe some of the cases which leads to a smaller overall error.

Bibliography

Aderem, A. (2005). "Systems biology: its practice and challenges". In: *Cell* 121.4, pp. 511–513. DOI: 10.1016/j.cell.2005.04.020.

Aggarwal, C. C. (2015). "Outlier analysis". In: *Data Mining.* Springer, pp. 237–263. DOI: 10.1007/978-3-319-14142-8_8.

Andrews, D. F. and C. L. Mallows (1974). "Scale mixtures of normal distributions". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102. URL: http://www.jstor.org/stable/2984774.

Bachmann, J., A. Raue, M. Schilling, M. E. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. D. Lehmann, J. Timmer, and U. Klingmüller (2011). "Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range". In: *Mol. Syst. Biol.* 7.1. DOI: 10.1038/msb.2011.50.

Banga, J. R. (2008). "Optimization in computational systems biology". In: *BMC Syst. Biol.* 2.1, p. 1. DOI: 10.1186/1752-0509-2-47.

Bellman, R. and K. J. Åström (1970). "On structural identifiability". In: *Mathematical biosciences* 7.3, pp. 329–339. DOI: 10.1016/0025-5564(70)90132-X.

Ben-Gal, I. (2005). "Outlier detection". In: *Data mining and knowledge discovery handbook.* Springer, pp. 131–146. DOI: 10.1007/978-0-387-09823-4.

Chen, J.-Q., M. R. Heldman, M. A. Herrmann, N. Kedei, W. Woo, P. M. Blumberg, and P. K. Goldsmith (2013). "Absolute quantitation of endogenous proteins with precision and accuracy using a capillary Western system". In: *Analytical biochemistry* 442.1, pp. 97–103. DOI: 10.1016/j.ab.2013.07.022.

Coleman, T. F. and Y. Li (1996). "An interior trust region approach for nonlinear minimization subject to bounds". In: *SIAM Journal on optimization* 6.2, pp. 418–445. DOI: 10.1137/0806023.

Cornbleet, P. J. and N. Gochman (1979). "Incorrect least-squares regression coefficients in method-comparison analysis." In: *Clinical chemistry* 25.3, pp. 432–438. URL: http://www.clinchem.org/content/25/3/432.short.

Cramér, H. (1945). *Mathematical methods of statistics.* Vol. 9. Princeton university press.

Edgeworth, F. Y. (1887). "On observations relating to several quantities". In: *Hermathena* 6.13, pp. 279–285. URL: http://www.jstor.org/stable/23036355.

Efron, B. (1992). *Bootstrap methods: another look at the jackknife.* Springer. DOI: 10.1007/978-1-4612-4380-9_41.

Fernández, C. and M. F. Steel (1999). "Multivariate Student-t regression models: Pitfalls and inference". In: *Biometrika* 86.1, pp. 153–167. DOI: 10.1093/biomet/86.1.153.

Fisher, R. A. et al. (1925). “Applications of Student’s t distribution”. In: *Metron* 5.3, pp. 90–104.

Fonseca, T. C., M. A. Ferreira, and H. S. Migon (2008). “Objective Bayesian analysis for the Student-t regression model”. In: *Biometrika* 95.2, pp. 325–333. DOI: 10.1093/biomet/asn001.

Gábor, A. and J. R. Banga (2015). “Robust and efficient parameter estimation in dynamic models of biological systems”. In: *BMC Syst. Biol.* 9.1, p. 74. DOI: 10.1186/s12918-015-0219-2.

Gassmann, M., B. Grenacher, B. Rohde, and J. Vogel (2009). “Quantifying Western blots: pitfalls of densitometry”. In: *Electrophoresis* 30.11, pp. 1845–1855. DOI: 10.1002/elps.200800720.

Ghosh, D. and A. Vogt (2012). “Outliers: An evaluation of methodologies”. In: *Joint Statistical Meetings*. American Statistical Association San Diego, CA, pp. 3455–3460. URL: http://www.amstat.org/sections/SRMS/Proceedings/y2012/files/304068_72402.pdf.

Gillespie, D. T. (1992). “A rigorous derivation of the chemical master equation”. In: *Physica A: Statistical Mechanics and its Applications* 188.1, pp. 404–425. DOI: 10.1016/0378-4371(92)90283-V.

Haas, G., L. Bain, and C. Antle (1970). “Inferences for the Cauchy distribution based on maximum likelihood estimators”. In: *Biometrika* 57.2, pp. 403–408. DOI: 10.1093/biomet/57.2.403.

Hand, D. J., H. Mannila, and P. Smyth (2001). *Principles of data mining.* MIT press.

Hasenauer, J., C. Hasenauer, T. Hucho, and F. J. Theis (2014). “ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics”. In: *PLoS Comput Biol* 10.7, e1003686. DOI: http://dx.doi.org/10.1371/journal.pcbi.1003686.

Hasenauer, J., S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer (2011). “Identification of models of heterogeneous cell populations from population snapshot data”. In: *BMC bioinformatics* 12.1, p. 1. DOI: 10.1186/1471-2105-12-125.

Hawkins, D. M. (1980). *Identification of outliers.* Vol. 11. Springer. DOI: 10.1007/978-94-015-3994-4.

Hodge, V. J. and J. Austin (2004). “A survey of outlier detection methodologies”. In: *Artificial Intelligence Review* 22.2, pp. 85–126. DOI: 10.1007/s10462-004-4304-y.

Hross, S. and J. Hasenauer (2016). “Analysis of CFSE time-series data using division-, age-and label-structured population models”. In: *Bioinformatics*, btw131. DOI: 10.1093/bioinformatics/btw131.

Huber, P. J. (2011). *Robust statistics.* Springer.

Ideker, T., T. Galitski, and L. Hood (2001). “A new approach to decoding life: systems biology”. In: *Annual review of genomics and human genetics* 2.1, pp. 343–372. DOI: 10.1146/annurev.genom.2.1.343.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Vol. 846. John Wiley & Sons.

Jones, M. and M. Faddy (2003). "A skew extension of the t-distribution, with applications". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 159–174. DOI: 10.1111/1467-9868.00378.

Kazeroonian, A., F. Fröhlich, A. Raue, F. J. Theis, and J. Hasenauer (2016). "CERENA: ChEmical REaction Network AnalyzerA Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics". In: *PloS one* 11.1, e0146732. DOI: http://dx.doi.org/10.1371/journal.pone.0146732.

Kitano, H. (2002). "Systems biology: a brief overview". In: *Science* 295.5560, pp. 1662–1664. DOI: 10.1126/science.1069492.

Klipp, E., R. Herwig, A. Kowald, C. Wierling, and H. Lehrach (2008). *Systems biology in practice: concepts, implementation and application.* John Wiley & Sons.

Kotz, S., T. Kozubowski, and K. Podgorski (2012). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance.* Springer Science & Business Media.

Kreutz, C., M. B. Rodriguez, T. Maiwald, M. Seidl, H. Blum, L. Mohr, and J. Timmer (2007). "An error model for protein quantification". In: *Bioinformatics* 23.20, pp. 2747–2753. DOI: 10.1093/bioinformatics/btm397.

Lange, K. L., R. J. Little, and J. M. Taylor (1989). "Robust statistical modeling using the t distribution". In: *Journal of the American Statistical Association* 84.408, pp. 881–896. DOI: 10.1080/01621459.1989.10478852.

Liu, C. and D. Rubin (1995). "ML estimation of the multivariate t distribution with unknown degrees of freedom". In: *Statistica Sinica* 5, pp. 19–39.

Maiwald, T. and J. Timmer (2008). "Dynamical modeling and multi-experiment fitting with PottersWheel". In: *Bioinformatics* 24.18, pp. 2037–2043. DOI: 10.1093/bioinformatics/btn350.

McKay, M. and R. Beckman (1979). "WJ, A comparison of three methods for selecting values of input variables in the analysis of output from a computer CodeTechnometrics". In: *Am Stat Assoc Am Soc Qual* 21, pp. 239–245. DOI: 10.1080/00401706.2000.10485979.

McNaught, A. D. and W. A. (1997). *Compendium of Chemical Terminology.* Vol. 1669. Blackwell Science Oxford.

Metz, C. E. (1978). "Basic principles of ROC analysis". In: *Seminars in nuclear medicine.* Vol. 8. 4. Elsevier, pp. 283–298. DOI: 10.1016/S0001-2998(78)80014-2.

Motulsky, H. and A. Christopoulos (2004). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* OUP USA.

Murphy, S. A. and A. W. Van der Vaart (2000). "On profile likelihood". In: *Journal of the American Statistical Association* 95.450, pp. 449–465. DOI: 10.1080/01621459.2000.10474219.

Niu, Z., S. Shi, J. Sun, and X. He (2011). "A survey of outlier detection methodologies and their applications". In: *Artificial intelligence and computational intelligence.* Springer, pp. 380–387. DOI: 10.1007/978-3-642-23881-9_50.

Nocedal, J. and S. Wright (2006). *Numerical optimization.* Springer Science & Business Media.

Pearson, R. K., G. E. Gonye, and J. S. Schwaber (2004). “Outliers in microarray data analysis”. In: *Methods of Microarray Data Analysis III.* Springer, pp. 41–55. DOI: 10.1007/0-306-48354-8_4.

Peel, D. and G. J. McLachlan (2000). “Robust mixture modelling using the t distribution”. In: *Statistics and computing* 10.4, pp. 339–348. DOI: 10.1023/A:1008981510081.

Pirrung, M. C. and E. M. Southern (2014). “The genesis of microarrays”. In: *Biochemistry and Molecular Biology Education* 42.2, pp. 106–113. DOI: 10.1002/bmb.20756.

Portnoy, S., R. Koenker, et al. (1997). “The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators”. In: *Statistical Science* 12.4, pp. 279–300. DOI: 10.1214/ss/1030037960.

Purdom, E. and S. P. Holmes (2005). “Error distribution for gene expression data”. In: *Statistical applications in genetics and molecular biology* 4.1. DOI: 10.2202/1544-6115.1070.

Raftery, A. E. (1995). “Bayesian model selection in social research”. In: *Sociological methodology* 25, pp. 111–164. DOI: 10.2307/271063.

Ramaswamy, S., R. Rastogi, and K. Shim (2000). “Efficient algorithms for mining outliers from large data sets”. In: *ACM SIGMOD Record.* Vol. 29. 2. ACM, pp. 427–438. DOI: 10.1145/342009.335437.

Raue, A., C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer (2009). “Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood”. In: *Bioinformatics* 25.15, pp. 1923–1929. DOI: 10.1093/bioinformatics/btp358.

Raue, A., M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer (2013). “Lessons Learned from Quantitative Dynamical Modeling in Systems Biology”. In: *PLoS ONE* 8.9, e74335. DOI: 10.1371/journal.pone.0074335.

Rawlings, J. S., K. M. Rosler, and D. A. Harrison (2004). “The JAK/STAT signaling pathway”. In: *Journal of Cell Science* 117.8, pp. 1281–1283. DOI: 10.1242/jcs.00963.

Reed, W. J. (2006). “The normal-Laplace distribution and its relatives”. In: *Advances in distribution theory, order statistics, and inference.* Springer, pp. 61–74. DOI: 10.1007/0-8176-4487-3_4.

Renart, J., J. Reiser, and G. R. Stark (1979). “Transfer of proteins from gels to diazobenzyl-oxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure”. In: *PNAS* 76.7, pp. 3116–3120.

Rizzo, M. L. (2007). *Statistical computing with R.* CRC Press.

Rousseeuw, P. J. and A. M. Leroy (2005). *Robust regression and outlier detection.* Vol. 589. John Wiley & Sons.

Schelker, M., A. Raue, J. Timmer, and C. Kreutz (2012). “Comprehensive estimation of input signals and dynamics in biochemical reaction networks”. In: *Bioinformatics* 28.18, pp. i529–i534. DOI: 10.1093/bioinformatics/bts393.

Schwarz, G. et al. (1978). “Estimating the dimension of a model”. In: *The annals of statistics* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Sengupta, B., K. Friston, and W. Penny (2014). "Efficient gradient computation for dynamical models". In: *NeuroImage* 98, pp. 521–527. DOI: `10.1016/j.neuroimage.2014.04.040`.

Stewart, C. V. (1999). "Robust parameter estimation in computer vision". In: *SIAM review* 41.3, pp. 513–537. DOI: `10.1137/S0036144598345802`.

Stivers, D. N., J. Wang, G. L. Rosner, and K. R. Coombes (2004). "Organ-specific differences in gene expression and UniGene annotations describing source material". In: *Methods of Microarray Data Analysis III.* Springer, pp. 59–72. DOI: `10.1007/0-306-48354-8_5`.

Student (1908). "The Probable Error of a Mean". In: *Biometrika* 6.1, pp. 1–25. DOI: `10.2307/2331554`.

Swameye, I., T. G. Müller, J. Timmer, O. Sandra, and U. Klingmüller (2003). "Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling". In: *PNAS* 100.3, pp. 1028–1033. DOI: `10.1073/pnas.0237333100`. eprint: `http://www.pnas.org/content/100/3/1028.full.pdf`.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation.* siam.

Taylor, J. and A. Verbyla (2004). "Joint modelling of location and scale parameters of the t distribution". In: *Statistical Modelling* 4.2, pp. 91–112. DOI: `10.1191/1471082X04st068oa`.

Villaverde, A. F. and J. R. Banga (2014). "Reverse engineering and identification in systems biology: strategies, perspectives and challenges". In: *Journal of The Royal Society Interface* 11.91, p. 20130505. DOI: `10.1098/rsif.2013.0505`.

Wang, R. Y. and D. M. Strong (1996). "Beyond accuracy: What data quality means to data consumers". In: *Journal of management information systems* 12.4, pp. 5–33. DOI: `10.1080/07421222.1996.11518099`.

Weber, P., J. Hasenauer, F. Allgöwer, and N. Radde (2011). "Parameter estimation and identifiability of biological networks using relative data". In: *Proc. of the 18th IFAC World Congress. Milano, Italy.* Vol. 18, pp. 11648–11653.

Wierling, C., R. Herwig, and H. Lehrach (2007). "Resources, standards and tools for systems biology". In: *Briefings in functional genomics & proteomics* 6.3, pp. 240–251. DOI: `10.1093/bfgp/elm027`.

Willmott, C. J. and K. Matsuura (2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In: *Climate research* 30.1, p. 79.

You, L., R. S. Cox, R. Weiss, and F. H. Arnold (2004). "Programmed population control by cell–cell communication and regulated killing". In: *Nature* 428.6985, pp. 868–871. DOI: `10.1038/nature02491`.