HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Data-driven robust and efficient mathematical modeling of biochemical processes

Carolin Loos

February 2019

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Data-driven robust and efficient mathematical modeling of biochemical processes

Carolin Loos

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Christina Kuttler

Prüfer der Dissertation:

1. TUM Junior Fellow Dr.-Ing. Jan P. Hasenauer
2. Prof. Dr. Oliver Junge
3. Prof. Dr. Ruth Baker

Die Dissertation wurde am 07.02.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 20.05.2019 angenommen.

Für Mama

Acknowledgments

First of all, I would like to thank Jan Hasenauer for his great supervision, guidance and encouragement and Carsten Marr for his patience and outstanding support. You are the best supervisors a PhD student could imagine! I want to thank Fabian Theis for giving me the opportunity to join the stimulating environment of the Institute of Computational Biology. I enjoyed my time here and the amazing working atmosphere.

I especially want to thank all current and former members of the groups Data-driven Computational Modeling, Quantitative Single Cell Dynamics, as well as the thesis writer club. In particular, Anna Fiedler, Atefeh Kazeroonian, Benjamin Ballnus, Corinna Maier, Dennis Rickert, Elba Raimúndez-Álvarez, Emmanuel Klinger, Fabian Fröhlich, Felix Buggenthin, Daniel Weindl, Dantong Wang, Jakob Vanhoefer, Justin Feigelman, Linda Krause, Lisa Amrhein, Lisa Bast, Lea Schuh, Leonard Schmiester, Mark Sinzger, Maren Büttner, Michael Strasser, Paul Stapor, Sabine Hug, Sabrina Krause, Sabrina Hross, Susanne Pieschner, Philipp Angerer, Thomas Blasi, Valerio Lupperger and Yannik Schälte. Big thanks to Marianne Antunes, Elisabeth Noheimer, Sabine Kunz, Anna Sacher and Nina Fischer for always patiently helping with administrative tasks.

Next, I want to thank my collaborators, Katharina Möller, Tim Hucho, Helge Hass, Clemens Kreutz, Jens Timmer, Constance Alabert, Moritz Völker-Albert, Axel Imhof, Ignasi Forne, Anja Groth and Frederik Graw. I always enjoyed our discussions and loved to work with you!

I also would like to thank the Sander lab, especially Augustin Luna and Chris Sander for the great time I could spend in the lab. I learned a lot and enjoyed many fruitful scientific discussions and input.

Furthermore, I want to thank my thesis committee Oliver Junge, Ruth Baker and Christina Kuttler for examining my thesis and chairing the committee.

I thank the Global Challenges for Women in Math Science program of the Technische Universität München for funding many stays abroad throughout my whole PhD studies.

And finally, I want to thank Per von Soosten, my friends and my family, especially my parents, Walter and Susanne Loos, and my siiisters Pamela Müller and Stefanie Loos – what would I have done without you and your continuous support!

Abstract

In systems biology, mathematical models are often employed to study the dynamics of biological processes. Model parameters, such as kinetic rate constants or measurement noise parameters, cannot usually directly be measured and need to be estimated from experimental data. To test biological hypotheses and thus gain a better understanding of the biological system, different models representing different hypotheses are fitted to experimental data and compared using statistical concepts. Increasing amounts and resolution of experimental data yield more detailed information, but also challenge the building and calibration of mathematical models.

In this thesis, we first considered data which provide only information about the average behavior of a cell population, often modeled by ordinary differential equations. We studied the calibration of these models on outlier-corrupted data using heavier tailed distributions for the measurement noise. We exploited the structure of the optimization problem and calculated optimal values for the parameters which do not contribute to the dynamics of the model analytically for different noise distribution assumptions. This enabled the robust and efficient fitting of computationally demanding models. We applied the developed methods to study histone methylation, where we performed model calibration and selection to gain new insights into the dynamics of the restoration of epigenetic marks in cycling cell populations.

Afterwards, we focused on data collected at the single-cell level, which requires models that are able to capture cellular heterogeneity. We proposed the hierarchical population model which allowed us to mechanistically describe multiple levels of cellular heterogeneity. It combines mixture modeling with mechanistic modeling of the statistical moments of cellular subpopulations. To ensure robustness of the model, we incorporated different distribution assumptions and investigated their influence on the optimization results. We applied the hierarchical population model to gain new knowledge about pain sensitization in primary sensory neurons.

The concepts and methods developed in this thesis enable the reliable, robust and computationally efficient calibration of comprehensive models even in the presence of outliers. Their application to experimental data facilitates a deeper understanding of and new mechanistic insights into biological systems.

Zusammenfassung

In der Systembiologie werden oft mathematische Modelle verwendet, um die Dynamiken von biologischen Prozessen zu untersuchen. Modellparameter, wie zum Beispiel kinetische Ratenkonstanten oder Parameter für das Messrauschen, können oft nicht direkt gemessen werden und müssen aus experimentellen Daten geschätzt werden. Um biologische Hypothesen zu testen und ein besseres Verständnis des biologischen Systems zu erhalten, werden unterschiedliche Modelle, welche unterschiedliche Hypothesen repräsentieren, an experimentelle Daten angepasst und mithilfe statistischer Methoden verglichen. Wachsende Datenmengen und genauere Auflösung der Daten stellen detaillierte Informationen bereit, erschweren allerdings auch die mathematische Modellbildung und -kalibrierung.

In dieser Arbeit betrachten wir zunächst Daten, die nur Informationen über das durchschnittliche Verhalten der Zellpopulation enthalten. Dies wird meist mit gewöhnlichen Differentialgleichungen modelliert. Wir betrachten die Modellkalibrierung anhand von Daten, die mit Ausreißern behaftet sind. Hierfür verwenden wir für das Messrauschen Verteilungen mit Rändern, die schwerer sind als die der üblicherweise genutzten Normalverteilung. Darüber hinaus nutzen wir die Struktur des Optimierungsproblems und berechnen die optimalen Werte der Parameter, die nicht zur Dynamik des Modelles beitragen, analytisch für unterschiedliche Verteilungen für das Messrauschen. Diese Ansätze ermöglichen die robuste und effiziente Anpassung von rechenintensiven Modellen. Wir wenden unsere Methoden an, um Histonmethylierung zu untersuchen, für welche wir Modelle kalibrieren und diese mit Modelselektionskriterien vergleichen. Hierbei gewinnen wir neue Einblicke in die Dynamik der Restorierung von epigenetischen Markierungen in einer sich teilenden Zellpopulation.

Anschließend konzentrieren wir uns auf Daten, welche auf der Einzelzellebene erhoben wurden. Die Untersuchung dieser Daten erfordert Modelle, die in der Lage sind, zelluläre Heterogenität zu beschreiben. Wir stellen ein hierarchisches Populationsmodell vor, welches uns erlaubt, mehrere Heterogenitätslevel mechanistisch zu beschreiben. Das Modell kombiniert Mischmodelle mit mechanistischer Modellierung der statistischen Momente der zellulären Subpopulationen. Wir integrieren mehrere Verteilungsannahmen, um die Robustheit des Modells sicherzustellen, und untersuchen deren Einfluss auf die Optimierungsergebnisse. Wir wenden unser Populationsmodell an, um neue Kenntnisse über die Schmerzsensitivierungen von sensorischen Neuronen zu erhalten.

Die Konzepte und Methoden, die in dieser Arbeit entwickelt wurden, ermöglichen die robuste und recheneffiziente Kalibrierung von komplexen Modellen, sogar wenn Ausreißer in den Daten vorhanden sind. Die Anwendung der Methoden auf experimentelle Daten ermöglicht ein tiefgehenderes Verständnis von und neue Einblicke in biologische Systeme.

Contents

Chapter 1

Introduction

Mathematical models are widely used in systems biology to gain a mechanistic understanding and extend the knowledge about biological processes (Baker et al., 2018; Cho and Wolkenhauer, 2005; Kitano, 2002). The models facilitate the unraveling of system properties that cannot directly be assessed with experiments (Aderem, 2005). They can also be used to compare and reject hypotheses about biological mechanisms (Crauste et al., 2017; Hross and Hasenauer, 2016) or to perform *in-silico* studies to predict how a system would respond to perturbations (Fröhlich et al., 2018; Molinelli et al., 2013). Mathematical models are used to study many biological processes and are employed to investigate, among others, epigenetics (Zheng et al., 2012), immunology (Buchholz et al., 2013), and cancer (Hass et al., 2017). This broad applicability has rendered them a powerful tool for developing treatment strategies for various diseases and for studying biological systems in general (Isensee et al., 2018; Merkle et al., 2016).

There exist many different types of mathematical models in systems biology, including deterministic (Klipp et al., 2005; von Foerster, 1959) and stochastic models (Wilkinson, 2009). The choice of modeling framework depends on the particular question and biological system under consideration. Many studies focus on the average behavior of a cell population and thus employ deterministic ordinary differential equation (ODE) models (Bachmann et al., 2011; Schöberl et al., 2009). However, the importance of differences between individual cells has gained enormous attention during the last decades (Elowitz et al., 2002; Regev et al., 2017). Even isogenic cells can behave differently upon stimulation (Tay et al., 2010). Heterogeneity has been shown to have important implications for cell fate (Spencer et al., 2009) or differentiation (Gerlach et al., 2013), and occurs for instance in cancer (Michor and Polyak, 2010) or pain sensitization (Hucho and Levine, 2007). Investigating heterogeneity requires measurements at the single-cell level, because the heterogeneity is concealed in the population averages (Figure 1.1). The single-cell data then needs to be combined with models which are able to delineate heterogeneity (Hasenauer et al., 2014; Zechner et al., 2012). However, developing appropriate modeling frameworks is still an open topic of current research.

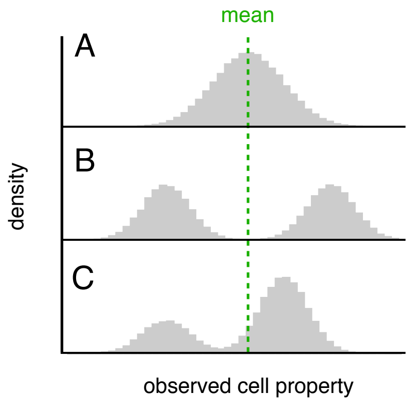

Figure 1.1: Heterogeneous cell populations. The displayed cell populations all have the same mean value of the measured cellular property. The population in (A) does not have subpopulations, while (B) and (C) comprise two distinct subpopulations with different subpopulation sizes and mean values for the subpopulations.

Using mathematical models to obtain a mechanistic understanding of the biological processes mostly requires the parametrization of the models (Tarantola, 2005), namely estimating the model parameters, such as kinetic rate constants, from experimental data. The goal is to obtain reliable results and draw sound conclusions even for complex models in a realistic and reasonable amount of time. Therefore, model calibration should be robust, reliable and (computationally) efficient. This is hindered and complicated by many factors, among others, outliers in the data, high number of parameters which need to be estimated and the lack of appropriate modeling frameworks. In this thesis, we developed and assessed robust and efficient methods for the calibration of ODE models on population average data. We exploited the structure of the optimization problem for robust distributions, which provide reliable results even in the presence of outliers, to split up and thus speed up the overall optimization problem. Additionally, we introduced a hierarchical population model for the analysis of single-cell snapshot data, which is a modeling framework that enables the description of heterogeneity at multiple levels. For this model, we also assessed its robustness with respect to the choice of incorporated distributions. We applied the developed models and methods to gain new insights into the dynamics of histone methylation and pain sensitization. The remainder of this chapter gives an introductory overview to the research topic and highlights the key issues and challenges in data-driven modeling, which are then addressed in the following chapters of this thesis.

1.1 Data-driven, dynamic modeling of biological processes

Different types of biological data carry different information. This ranges from providing measurements of the average behavior of the cells to providing the dynamics of individual cells. The first type of data, referred to as population average data, is typically collected by techniques such as Western blotting (Renart et al., 1979) or microarrays (Malone and Oliver, 2011). The latter, single-cell data, can be collected by fluorescent microscopy (Herzenberg et al., 2006), flow cytometry (Herzenberg et al., 2006) or mass cytometry (Giesen et al., 2014). It can further be differentiated between single-cell time-lapse and snapshot data, for which the first provides the dynamics of individual cells while the latter does not provide this information.

For both types of data, population average and single-cell data, often fluorescent markers or antibodies are employed to collect measurements, which yields measurements which are proportional to the cellular quantity of interest. Additionally, the measurements can be obscured by outliers, which arise due to errors in the data collection and processing (Ghosh and Vogt, 2012). The mathematical models which are used to interpret the data thus need to account for these discrepancies between the collected measurement and the examined cellular property.

The unknown parameters of the model, e.g., kinetic rate constants, scaling parameters, parameters encoding measurement noise or initial conditions, are estimated from experimental data. How the discrepancy between mathematical model and data is evaluated is crucial and influences the estimation results (Loos et al., 2015). The parameters are usually estimated by maximizing a likelihood function, which provides the probability of observing the data given the model and corresponding parameters. This optimization problem is generally non-convex, which has to be accounted for in the calibration. For this, multi-start local optimization has shown to be a good globalization strategy (Raue et al., 2013; Villaverde et al., 2018). The best found function value is assumed to be the potentially global optimum, with higher reliability if it is found more often. Methods for model calibration are assumed to be more robust if they find the global optimum more repeatedly, while a model is assumed to be robust if it provides parameter estimates close to the true parameter values even in the presence of outliers.

1.1.1 Mathematical modeling and model calibration for population averages

For modeling population average data, the most commonly used modeling technique are ODE models. These are reaction rate equations (RREs) and describe, e.g., the synthesis, degradation or transition of biochemical species and interactions between these. Challenges here arise in the definition of the model structure. Different structures represent different biological hypotheses.

Further challenges in the modeling of population averages arise due to the complexity of the models themselves or the number of models which need to be calibrated to answer the considered question. To understand complex processes, detailed and large-scale models with increasing numbers of parameters are developed (Bouhaddou et al., 2018; Fröhlich et al., 2018; Hass et al., 2017). In addition, experimental techniques allow the collection of large amounts of data which can be integrated by mathematical models. This yields an increasing number of observation parameters, i.e., parameters used to map the biochemical species to the measurable output. Furthermore, with more measured data points also the probability that outliers occur in the data increases, which changes the measurement noise distribution. All these factors can substantially hinder model calibration. Methods for model calibration need to be adapted in order to cope with the increasing complexity, i.e., by speeding up the gradient evaluation required for the optimization (Fröhlich et al., 2017) or exploiting the structure of the optimization problem (Weber et al., 2011). If there are many hypotheses to be tested, efficient techniques for model selection need to be employed in addition to the efficient calibration of a single model (Steiert et al., 2016).

1.1.2 Mathematical modeling of cell populations

A difficult task when modeling a cell population is the choice of modeling framework. Single-cell data carry information about the cellular heterogeneity, which needs to be captured by the employed model in order to obtain a mechanistic understanding of the heterogeneity. Heterogeneity in a cell population can arise on different levels: (i) differences between subpopulations or cell-types that can be caused by the cellular micro-environment (Ebinger et al., 2016) or stable epigenetic markers which are acquired during cell differentiation (Reik, 2007); and (ii) differences between cells of the same subpopulation or cell-type that arise, e.g., from differences in the cell state (Buettner et al., 2015) or from stochastic fluctuations in gene expression (Elowitz et al., 2002). The differences on both levels can be caused by intrinsic or extrinsic noise (Elowitz et al., 2002). Intrinsic noise emerges due to the stochastic nature of gene expression, while extrinsic noise refers to fluctuations in other cellular components, e.g., differences in protein levels of individual cells. Diverse modeling frameworks exist which capture intrinsic noise (Gillespie, 2000), extrinsic noise (van der Merwe, 2004), intrinsic and extrinsic noise (Zechner et al., 2012), or subpopulation structures (Hasenauer et al., 2014). However, a unifying framework is still missing.

1.2 Overview and contribution of this thesis

In this thesis, we focus on the following problems and bottlenecks:

(i) The assumption about the distribution of the measurement noise influences the estimation results. Yet, robust distributions have not been adapted to fitting procedures for dynamical models. Moreover, the influence of outliers on the estimation results and the optimization performance is unclear.

(ii) ODE models often comprise not only parameters which influence the dynamics, but also scaling parameters, which are used model relative data, and measurement noise parameters. The scaling and noise parameters are estimated along with the dynamic parameters. This increases the dimension of the optimization problem and complicates model calibration and analysis.

(iii) Cell populations often comprise heterogeneous subpopulations. Yet, no computationally tractable modeling framework is available which can incorporate a mechanistic description of variability between and within subpopulations.

(iv) The assessment of distribution assumptions is not only missing for modeling population average data (i), but also for population models that rely on distribution assumptions. Since the data types usually have quite different properties, e.g., number of measured data points, results for studying population averages cannot directly be transferred to single-cell data.

These problems arise when studying various biological processes and hinder a reliable computational analysis of the biological systems. In this thesis, the following two biological questions are addressed:

(v) Methylations at histone tails play an important role for epigenetic regulation. In a proliferating cell population, new unmodified histones are incorporated and the epigenetic marks need to be restored. To understand this process and the influence of parental histone modification, a deeper understanding of the establishment of H3K27K36 methylated chromatin is required.

(vi) Primary sensory neurons are highly heterogeneous cells which are involved in pain sensitization. So far, the influence of extracellular scaffolds, which are often highly altered in painful conditions such as wounds or tumors, on pain signaling has not been studied in a mechanistic way.

In this thesis, we addressed the aforementioned issues and in the following the contributions are delineated. For modeling population average data with ODE models, the main contributions are:

- **Robust calibration of ODE models with outlier-corrupted data.** In the context of dynamical systems, we assessed various noise distributions which have heavier tails than the generally used Gaussian distribution. We derived the equations for the likelihood functions and their gradients which is required for the efficient calibration of the distributions. In addition, we investigated the robustness and performance of these distributions in the absence and presence of outliers for different outlier scenarios. We found an improved robustness by using heavier tailed distributions. This contribution addressed aforementioned problem (i).

- **Efficient calibration of ODE models employing hierarchical optimization.** We split the overall optimization problem into subproblems of smaller dimension. The inner problem includes parameters which do not influence the dynamics of the ODE for the biochemical species, e.g., scaling and measurement noise parameters. Under two different distribution assumptions, we derived the analytical expressions for the inner subproblem. The hierarchical approach for optimization achieved a higher number of optimization runs which converged to the potentially global optimum than the standard approach in less computation time. This contribution addressed problem (ii).

- **Studying the dynamics of histone H3 methylation.** The developed approaches were applied to study the dynamics of histone H3 methylation. We developed two mathematical models describing the temporal evolution of the relative abundance of K27 and K36 methylation states. Performing model selection and validating the predictions of the models, we found that a model assuming that a fraction of histones can only be methylated up to a defined final state seems to be most reasonable. This contribution addressed problem (v).

For single-cell snapshot data, the main contributions of this thesis are the following:

- **Developing the hierarchical population model, which captures multiple levels of heterogeneity.** We developed a modeling framework which incorporates inter- and intra-subpopulation variability. This framework combines mixture modeling with modeling of the statistical moments of individual subpopulations. This enables a mechanistic description of various levels of heterogeneity and correlation structures of multiplexed measurements. We provided the theoretical concepts and equations for an efficient model calibration and model selection. This contribution addressed problem (iii).

- **Robust calibration of hierarchical population model.** We assessed and incorporated different distribution assumptions in the hierarchical population modeling framework. This enables a robust calibration of these models in the presence and absence of outliers. This contribution addressed problem (iv).

- **Unraveling sources of heterogeneity in NGF-induced Erk signaling.** We applied the introduced framework of hierarchical population models to study pain signaling in primary sensory neurons exposed to different extracellular scaffolds. Our analysis revealed that differences in the response to NGF stimulation of cells cultured on different extracellular scaffolds could be explained by altered intracellular signaling but not a shift in the subpopulation size. This contribution addressed problem (iv).

Some of these contributions are already part of peer-reviewed publications, currently submitted to peer-reviewed journals or in preparation. Parts of the work in this thesis thus correspond or are to some extent identical with the following publications:

- **Loos, C.***, Krause, S.*, & Hasenauer, J. (2018). Hierarchical optimization for the efficient parametrization of ODE models. *Bioinformatics*, 34(24), 4266–4273. (*equal contribution)

- **Loos, C.***, Völker-Albert, M.*, Forne, I., Hasenauer, J., Imhof, A., Marr, C., Groth, A., Alabert, C. Efficient K27me3 establishment requires naive histone substrates and pre-existing K27me3 on old histones. *in preparation.*

- **Loos, C.***, Moeller, K.*, Fröhlich, F., Hucho, T., & Hasenauer, J. (2018). A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Systems*, 6(5), 593-603.

- **Loos, C**., Fiedler, A., & Hasenauer, J. (2016). Parameter estimation for reaction rate equation constrained mixture models. In *International Conference on Computational Methods in Systems Biology* (pp. 186-200). Springer International Publishing.

- **Loos, C.**, & Hasenauer, J. Robust calibration of hierarchical population models on single-cell snapshot data. *in preparation.*

- Fröhlich, F., **Loos, C.**, & Hasenauer, J. (2019). Scalable inference of ordinary differential equation models of biochemical processes. In *Gene Regulatory Networks* (pp. 385-422). Humana Press, New York, NY.

- Maier, C., **Loos, C.**, & Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5), 718-725.

In the sixth publication, I contributed Section 3: *Inference of Model Structure*, which is in parts included in the Background Chapter 2 of this thesis. I wrote the main part of the last publication, which is based on a master's thesis (Maier, 2016), which I supervised during my doctoral research.

Other contributions of my doctoral research which are not included in this thesis are:

- Hass, H.*, **Loos, C.***, Raimúndez-Álvarez, E., Timmer, J., Hasenauer, J., & Kreutz, C. (2019). Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, btz020.

- Sinzger, M., Vanhoefer, J., **Loos, C.**, & Hasenauer, J. (2019). Comparison of null models for combination drug therapy reveals Hand model as biochemically most plausible. *Scientific Reports*, 9(3002).

- Stapor, P., Weindl, D., Ballnus, B., Hug, S., **Loos, C.**, Fiedler, A., Krause, S., Hroß, S., Fröhlich, F., & Hasenauer, J. (2018). PESTO: Parameter EStimation TOolbox. *Bioinformatics*, 34(4), 705-707.

1.3 Outline

This thesis is structured as follows: In Chapter 2, the background knowledge and notation for the methods employed in this thesis are introduced. The considered data types are explained and the corresponding possible modeling approaches are outlined. The chapter also provides the background of model calibration and selection. In Chapter 3, the robustness and computational efficiency of methods for calibrating ODE models on population average data are addressed. Different noise distributions and their influence on the estimation performance and results are assessed. This is followed by the introduction of a hierarchical optimization approach for calibrating ODE models on relative data. The established methods of this chapter are then applied to study the dynamics of histone H3 methylation to obtain a better understanding of epigenetic regulation. In Chapter 4, a hierarchical population model is introduced which incorporates multiple levels of heterogeneity. The developed modeling framework is applied to study NGF-induced Erk signaling in nociceptive neurons and provides new biological insights into mechanisms of pain signaling, which can have important implications for the treatment of pain sensitization. Afterwards, the framework, which relies on distribution assumptions, is extended by incorporating and analyzing further distributions to enable a more robust calibration of the population model. In Chapter 5, we conclude the thesis by briefly summarizing the main results and providing an outlook for potential future scientific directions.

Chapter 2

Background

In this chapter, the considered types of experimental data are described as well as the mathematical concepts to model the temporal evolution of biochemical species. Furthermore, we outline the main methods which are used to calibrate these models on experimental data and to perform model selection.

2.1 Experimental data

There are various experimental techniques which can be employed to collect measurement data. We mainly distinguish these techniques by the level of information they provide, i.e, whether a high number of cells is combined and the average properties of this cell population are measured or whether the properties of the individual cells are measured and tracked over time.

2.1.1 Population average data

Experimental techniques such as microarrays (Malone and Oliver, 2011) or Western blotting (Renart et al., 1979) provide information about cellular properties, such as protein or RNA levels, averaged over a cell population. Many techniques rely on antibodies or fluorescent markers and provide only measurements which are proportional to the quantity of interest.

We denote population average data by

$$\mathcal{D} = \left\{ \left\{ \{ \bar{\mathbf{y}}_{k,d,e}, t_{k,d,e}, \mathbf{u}_{d,e} \}_k \right\}_d \right\}_e \ , \tag{2.1}$$

with indices for time point k, experiment e and condition d. The vector $\bar{\mathbf{y}}$ includes jointly measured quantities $y_i, i = 1, \ldots, n_y$. The vector $\mathbf{u}_{d,e}$ gathers the inputs of experiment e, which can be concentrations of stimulating agents, such as growth factors, or drugs, such as kinase inhibitors.

2.1.2 Single-cell snapshot data

During the last decades the importance of heterogeneity in cell populations became clear (Altschuler and Wu, 2010; Regev et al., 2017). Information about this heterogeneity is hidden in population average data and higher resolution techniques need to be employed which provide measurements for individual cells. Common techniques are flow cytometry (Davey and Kell, 1996), mass cytometry (Bodenmiller et al., 2012), image cytometry (Ozaki et al., 2010), single-cell microscopy (Miyashiro and Goulian, 2007) or scRNA-seq (Kolodziejczyk et al., 2015). Further techniques have been developed which not only provide information about cellular properties, such as protein abundance, but also measure spatial information (Lin et al., 2015). For most of these experiments, measurements of a cell can only be collected once due to, e.g., fixation of the cells. Therefore, these data only provide snapshots and do not measure the same cells across different time points or drug dosages. Mostly, either protein levels or gene expression is measured for an individual cell, but also some techniques have recently been developed to measure both simultaneously (Frei et al., 2016; Lane et al., 2017).

In this thesis, we denote single-cell snapshot data as

$$\mathcal{D} = \left\{ \left\{ \left\{ \{ \bar{\mathbf{y}}^c_{k,d,e}, t_{k,d,e}, \mathbf{u}_{d,e} \}_c \right\}_k \right\}_d \right\}_e , \tag{2.2}$$

with cell index c. Other types of single-cell data such as single-cell time-lapse data, which keep track of individual cells over time, or data for the cell population size, e.g., measured by persistent cell labeling (Lyons and Parish, 1994), are not discussed in this thesis.

While Chapter 3 focuses on the estimation of model parameters based on population average data, Chapter 4 considers the calibration of mathematical models based on single-cell snapshot data.

2.2 Mathematical modeling of biological systems

Mathematical models are valuable tools for studying biological processes. Depending on the studied data type and particular biological question, different modeling techniques are used. We first introduce models for individual cells. A model assuming deterministic behavior for each cell can be used to model the population average. This is followed by a discussion of models for cell populations.

The parts addressing sigma-point and moment-closure approximation in Section 2.2.2 are modified versions of the corresponding sections of the author's publication (Loos et al., 2018b).

2.2.1 Models for individual cells

Models for individual cells describe the temporal evolution of biochemical species $\mathbf{x} = (x_1, \ldots, x_{n_x})$ depending on a vector $\boldsymbol{\psi}$ of dynamic parameters, e.g., kinetic rate constants, and input $\mathbf{u}$. Models for individual cells can principally account for intrinsic noise. In the following, we describe Markov jump processes and reaction rate equations, for which the first is a stochastic and the latter a deterministic modelling approach.

Markov jump processes (MJPs) have a discrete state space but continuous time. They are often employed to describe the dynamics of individual cells if it is important to capture the discreteness of the molecule numbers and account for intrinsic noise. Changes in the state $\mathbf{x} \in \mathbb{N}^{n_x}$ of the cell occur due to reactions

$$R_l : \sum_{i=1}^{n_x} \nu_{i,l}^- \, x_i \xrightarrow{k_l(\boldsymbol{\psi},\mathbf{u})} \sum_{i=1}^{n_x} \nu_{i,l}^+ \, x_i \, , \tag{2.3}$$

with reaction index l, parameter-dependent reaction rate constant $k_l(\boldsymbol{\psi}, \mathbf{u})$ and stoichiometry $\boldsymbol{\nu}_l = \boldsymbol{\nu}_l^+ - \boldsymbol{\nu}_l^-$. The propensity of a reaction indexed by l is denoted by $a_l(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u})$ and the probability of the reaction to occur in time interval dt is $a_l(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u})dt + o(dt)$. For example, for a zero-order reaction which does not depend on the number of molecules, i.e., $\boldsymbol{\nu}_l^- = \mathbf{0}$, the propensity is $a_l(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = k_l(\boldsymbol{\psi}, \mathbf{u})$. The reaction changes the state of the system from $\mathbf{x}$ to $\mathbf{x} + \boldsymbol{\nu}_l$. A method to simulate MJPs has been developed by Gillespie (1977) and is called the stochastic simulation algorithm (SSA). However, this algorithm can be computationally expensive, especially if a high number of reactions needs to be simulated. For this, more computationally efficient approximations for the SSA have been developed, e.g., tau-leaping (Gillespie, 2001) and multi-level methods (Anderson and Higham, 2012; Lester et al., 2015). An approximation of the MJP is given by the chemical Langevin equation (Gillespie, 2000) which is based on stochastic differential equations. It is a continuous approximation with continuous state space $\mathbf{x} \in \mathbb{R}^{n_x}$.

The **reaction rate equation** (RRE) is given by the ordinary differential equation (ODE)

$$\dot{\mathbf{x}} = f(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}), \quad \mathbf{x}(0) = \mathbf{x}_0(\boldsymbol{\psi}, \mathbf{u}), \tag{2.4}$$

with vector field $f : \mathbb{R}_{+}^{n_x} \times \mathbb{R}^{n_\psi} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$. A unique solution to (2.4) exists if f is Lipschitz continuous. The RRE assumes that each cell behaves in the same, deterministic way, not accounting for intrinsic noise. Since RREs can also be interpreted as modeling the behavior of the average cell, they are frequently used to model population averages (Klipp et al., 2005). This modeling approach will be used in Chapter 3.

We denote the overall vector of parameters which is estimated from the data as $\boldsymbol{\theta} \in \Theta$, for which Θ is the biologically reasonable regime of parameter values. Usually, $\boldsymbol{\theta}$ includes the dynamic parameters $\boldsymbol{\psi}$, scaling parameters $\mathbf{s}$ and distribution parameters $\boldsymbol{\varphi}$ used to describe measurement noise. We obtain the observables by an observation function $h\colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_y}$, which maps the states, parameters and inputs to the observables via

$$\mathbf{y}(t, \boldsymbol{\theta}, \mathbf{u}) = h(\mathbf{x}(t, \boldsymbol{\psi}, \mathbf{u}), \boldsymbol{\theta}, \mathbf{u})\,. \tag{2.5}$$

The observables are the properties of the system which are measured. Since measurements are mostly subject to measurement noise, the model also needs to account for this by a noise model

$$\bar{y}_{i,k,d,e} \sim p\left(\bar{y}_{i,k,d,e} | y_i(t_{k,e}, \boldsymbol{\theta}, \mathbf{u}_{d,e}), \boldsymbol{\varphi}_i(t_{k,e}, \boldsymbol{\theta}, \mathbf{u}_{d,e})\right)\,, \tag{2.6}$$

with indices introduced in (2.1, 2.2). For this, often Gaussian noise is assumed.

2.2.2 Models for heterogeneous cell populations

One approach for modeling a whole-cell population are ensemble models, which use models for individual cells, e.g., MJPs or RREs as discussed in Section 2.2.1, and model the overall population as a collection of many individual cells (Henson, 2003; Kuepfer et al., 2007). The dynamics of the individual cells might also be stochastic. However, the computational complexity limits the practicability of ensemble models. In contrast to ensemble models, density based models do not describe the dynamics of each individual cell of the population, but model the temporal evolution of the cell density (Gillespie, 1992; Hasenauer et al., 2011b). An ensemble model can also be interpreted as a sampling-based approximation of a density based model (Waldherr, 2018). To incorporate measurement noise, the density is convolved with a density for the measurement noise. Similar to the models for individual cells, density based models differ in the discreteness/continuity of the state space, as well as the incorporation of intrinsic noise. Since these models describe the whole-cell population, they can also include extrinsic noise. This is often done by assuming certain parameters, which represent the quantities responsible for the extrinsic variability, to differ between cells. The parameters for cell c are then assumed to be distributed according to

$$\boldsymbol{\psi}^c \sim p_\psi(\boldsymbol{\psi})\,, \tag{2.7}$$

where $\boldsymbol{\beta}$ is the mean and $\mathbf{D}$ the covariance matrix of probability distribution p_ψ:

$$\mathbb{E}[\boldsymbol{\psi}^c] = \boldsymbol{\beta}, \quad \mathrm{Cov}[\boldsymbol{\psi}^c] = \mathbf{D}, \quad \boldsymbol{\xi} = (\boldsymbol{\beta}, \mathbf{D})\,. \tag{2.8}$$

We assume that extrinsic noise is encoded in L parameters of the parameter vector $\boldsymbol{\psi} \in \mathbb{R}^{n_\psi}$. For parameters that are considered to be homogeneous, i.e., not variable across the cells, it is assumed that $\beta_i = \psi_i$ and $D_{ii} = D_{ij} = D_{ji} = 0, \forall j$. In the case of cell population models which comprise extrinsic noise, $\boldsymbol{\theta}$ can also contain means, variances and parameters to parametrize the covariance matrix of the parameters which encode properties of extrinsic noise. Also methods exist which do not parametrize (2.7) and infer the density by, e.g., using maximum entropy approaches (Dixit et al., 2019).

The whole density which would be obtained with an ensemble model for a large number of individual cells can be described with a **population balance equation** (PBE). Assuming neither intrinsic nor extrinsic noise, heterogeneity occurs only due to differences in the initial state. If not the population average or an individual cell is described, the temporal evolution of the density of the cell states can be described by a PBE (see, e.g., (Waldherr, 2018)):

$$\frac{\partial p(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t)}{\partial t} = -\mathrm{div}_x(f(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) p(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t)), \quad p(0, \mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) = \mathbf{x}_0(\boldsymbol{\psi}, \mathbf{u})\,. \tag{2.9}$$

In (2.9) only cellular dynamics such as signaling are included and proliferation of the cells is neglected. These would be included as additional terms. Extensions address also the incorporation of intrinsic and extrinsic noise (Hasenauer et al., 2011b). However, the numerical simulation of these models requires, e.g., hierarchical simulation schemes (Pinto et al., 2007), grid-based approaches (Mantzaris et al., 2001) or characteristic-based approaches (Küper et al., 2019). These are often computationally expensive and do not scale well with high-dimensions, limiting the applicability of PBEs.

If intrinsic variability is neglected and the only source of biological variability is the distribution in the parameters $p_\psi(\boldsymbol{\psi})$ as introduced in (2.7), this distribution is mapped to a distribution of cell states and observables. A detailed analysis of this image requires sampling from $p_\psi(\boldsymbol{\psi})$ and subsequent evaluation of the state and observable vectors by simulation. This procedure is, however, computationally demanding. The **sigma-point approximation** addresses this issue and gives an approximation of the statistical moments of the image, mean and covariance matrix and their dynamics in time, using a small number of simulations (Filippi et al., 2016; Silk, 2013; van der Merwe, 2004). The sigma-point approximation uses only the image of deterministically chosen parameter vectors and can be seen as an approximation to an ensemble model. These parameter vectors, the so called sigma-points, are chosen to represent the mean $\boldsymbol{\beta}$ and the covariance $\mathbf{D}$ of p_ψ.

Following van der Merwe (2004), the $2L+1$ sigma-points $\{v_l, \boldsymbol{\mathcal{S}}_l\}$ are defined as

$$\begin{aligned}
&\boldsymbol{\mathcal{S}}_0 = \boldsymbol{\beta}, && v_0^{(m)} = \frac{\zeta_3}{L+\zeta_3}, && \text{for } l = 0\\
&\boldsymbol{\mathcal{S}}_l = \boldsymbol{\beta} + \left(\sqrt{(L+\zeta_4)\,\mathbf{D}}\right)_l, && v_l^{(c)} = \frac{\zeta_3}{L+\zeta_3} + 1 - \zeta_1^2 + \zeta_2\,, && \text{for } l = 1\ldots, L\\
&\boldsymbol{\mathcal{S}}_l = \boldsymbol{\beta} - \left(\sqrt{(L+\zeta_4)\,\mathbf{D}}\right)_l, && v_l^{(m)} = v_l^{(c)} = \frac{1}{2\,(L+\zeta_3)}, && \text{for } l = L+1\ldots, 2L\,.
\end{aligned} \tag{2.10}$$

For the hyperparameters, van der Merwe (2004) proposes to use $\zeta_2 = 2$ and $\zeta_3 = \zeta_1^2(L + \zeta_4) - L$, with $\zeta_1 = 0.7$ and $\zeta_4 = 0$. The superscripts for v_l indicate whether it is used for the calculation of the mean $^{(m)}$ or the covariance $^{(c)}$.

The dynamics of individual cells can, e.g., be described by the RRE (2.4). Accordingly, the images of the sigma-points in the state and the observation space, $\boldsymbol{\mathcal{X}}_l$ and $\boldsymbol{\mathcal{Y}}_l$, are computed as

$$\begin{aligned}
\dot{\boldsymbol{\mathcal{X}}}_l &= f(\boldsymbol{\mathcal{X}}_l, \boldsymbol{\mathcal{S}}_l, \mathbf{u})\,, \quad l = 0, \ldots, 2L\,,\\
\boldsymbol{\mathcal{Y}}_l &= h(\boldsymbol{\mathcal{X}}_l, \boldsymbol{\mathcal{S}}_l, \mathbf{u})\,.
\end{aligned} \tag{2.11}$$

The mean and covariance matrix of the species are computed as

$$\begin{aligned}
\mathbf{m}^x &\approx \sum_{l=0}^{2L} v_l^{(m)} \boldsymbol{\mathcal{X}}_l\,,\\
\mathbf{C}^x &\approx \sum_{l=0}^{2L} v_l^{(c)} \left(\boldsymbol{\mathcal{X}}_l - \mathbf{m}^x\right)\left(\boldsymbol{\mathcal{X}}_l - \mathbf{m}^x\right)^T\,.
\end{aligned}$$

The mean and covariances of the observables read

$$\begin{aligned}
\mathbf{m}^y &\approx \sum_{l=0}^{2L} v_l^{(m)} \boldsymbol{\mathcal{Y}}_l\\
\mathbf{C}^y &\approx \sum_{l=0}^{2L} v_l^{(c)} \left(\boldsymbol{\mathcal{Y}}_l - \mathbf{m}^y\right)\left(\boldsymbol{\mathcal{Y}}_l - \mathbf{m}^y\right)^T\,.
\end{aligned} \tag{2.12}$$

The temporal evolution of the cell density is governed by the **chemical master equation** (CME) (Gillespie, 1992)

$$\dot{p}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t) = \sum_{l} \left[p(\mathbf{x} - \boldsymbol{\nu}_l, \boldsymbol{\psi}, \mathbf{u}, t) a_l(\mathbf{x} - \boldsymbol{\nu}_l, \boldsymbol{\psi}, \mathbf{u}) - a_l(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}) p(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t) \right] , \tag{2.13}$$

with definitions as in (2.3). The CME accounts for the stochasticity of the biological processes and models intrinsic noise. Extrinsic noise can be included by extending the state space with the heterogeneous parameters and assuming that these additional states only influence reactions, but are not altered by any reaction and thus have no dynamics. For many biological systems, the state space is infinite and the CME an infinite dimensional system of coupled ODEs. Therefore, the CME is often studied by simulating trajectories of the system with the SSA. This can be computationally expensive, especially if high number of molecules are involved. To address this issue of computational complexity, many concepts for the approximation of the CME have been developed, such as the finite state projection (Munsky and Khammash, 2006), the moment-closure approximation (Engblom, 2006; Lee et al., 2009), or the system size expansion (van Kampen, 2007), which gives, e.g., the Fokker-Plank equation (Risken, 1996) or the linear noise approximation (Komorowski et al., 2009). In this thesis, the moment-closure approximation is employed.

The **moment-closure approximation** (MA) provides equations for the temporal evolution of moments of the species, i.e., of the mean

$$\dot{m}_i^x = \sum_{\mathbf{x} \in \Omega} x_i \, \dot{p}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t), \quad i = 1, \ldots, n_x , \tag{2.14}$$

of species x_i, with $\dot{p}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t)$ defined by (2.13), and higher-order moments such as the covariance

$$\dot{C}_{ij}^x = \sum_{\mathbf{x} \in \Omega} (x_i - m_i^x)(x_j - m_j^x) \dot{p}(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t), \quad i, j = 1, \ldots, n_x \tag{2.15}$$

between species x_i and x_j. Here, Ω denotes the set of possible states. Given the moments of the species, the moments of the observables are calculated by

$$\begin{aligned} m_i^y &= \sum_{\mathbf{x} \in \Omega} h_i(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u}) p(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t) \\ C_{ij}^y &= \sum_{\mathbf{x} \in \Omega} \left(h_i(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u}) - m_i^y \right) \left(h_j(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u}) - m_j^y \right) p(\mathbf{x}, \boldsymbol{\psi}, \mathbf{u}, t) . \end{aligned} \tag{2.16}$$

Here, h_i denotes the ith component of the observation function h defined in (2.5). For nonlinear systems, the dynamics of the moments of order k depend on moments of order

$k+1$. For this, moment-closure schemes are applied which introduce an approximation error (Lee et al., 2009). These schemes describe higher-order moments as functions of lower-order moments. In the MA, extrinsic noise can be included by extending the state space by the heterogeneous parameters and assuming that they are constant (Zechner et al., 2012).

Often, cell populations have subpopulation structures (Altschuler and Wu, 2010; Buettner et al., 2015; Nester and Stocker, 1963). The aforementioned models do not explicitly model these structures. **RRE-constrained mixture modeling** (Hasenauer et al., 2014) mechanistically models the mean of n_s subpopulations by RREs (2.4). The whole-cell population is then the mixture of normal or log-normal distributions, for which the subpopulations are the mixture components. Parameters are either homogeneous or assume distinct values for different subpopulations

$$p(\psi_i^c) = \begin{cases} \delta(\psi_i^c - \beta_i) & \text{homogeneous} \\ \sum_s w_s \, \delta(\psi_i^c - \beta_{s,i}) & \text{subpopulation variable} \end{cases}$$

in which δ denotes the Dirac delta distribution. For measurement $\bar{y}_{i,k}$, it then reads

$$\begin{aligned} p(\bar{y}_{i,k}|\boldsymbol{\theta}) &= \sum_{s=1}^{n_s} w_s(\boldsymbol{\theta}) \, \phi\left(\bar{y}_{i,k}|\mu_{s,i}(t_k,\boldsymbol{\theta},\mathbf{u}), \sigma_{s,i}^2(t_k,\boldsymbol{\theta},\mathbf{u})\right) \\ \text{with } \dot{\mathbf{x}}_s &= f\left(\mathbf{x}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right), \; \mathbf{x}_s(0) = \mathbf{x}_0(\boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}), \\ \boldsymbol{\mu}_s &= g_\varphi\left(\mathbf{x}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right) . \end{aligned} \tag{2.17}$$

Here and in the following, we neglect for convenience the indices for experiment and condition. The mean for the subpopulation obtained by RREs is linked to the distribution parameter μ by function g_φ. The distribution parameter σ is estimated from the data. No intrinsic or extrinsic noise within a subpopulation is included. This approach is a rough simplification of the underlying single-cell dynamics but computationally efficient.

The appropriate modeling framework needs to consider the importance of intrinsic and extrinsic noise upon population dynamics (Waldherr, 2018). The CME, the MA and variants of the PBE can account for intrinsic noise, while the sigma-point approximation only allows for variability due to extrinsic noise. The differences between measurements for cells in the RRE-constrained mixture model only occur due to measurement noise and cells belonging to different subpopulations. The CME, PBE and RRE-constrained mixture model describe the whole density of the population. However, the latter is a parametrized approximation to the cell population using mixture distributions. The MA and sigma-point approximation provide only the moments of the system.

2.3 Parameter inference

The mathematical models for population average and single-cell snapshot data introduced above depend on parameters, e.g., kinetic rate constants, initial conditions and measurement noise. These parameters generally cannot be directly experimentally measured and the corresponding, often ill-posed inverse problem needs to be solved, which means the parameters are inferred from experimental data. In this section, we describe how models introduced in Section 2.2 can be calibrated to data introduced in Section 2.1. General approaches to estimate the parameters of a model include frequentist (Raue et al., 2013), Bayes (Wilkinson, 2007) or set-based (Rumschinski et al., 2010) approaches. In this thesis, we mainly consider the frequentist approach.

The likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})\,, \tag{2.18}$$

is the conditional probability of observing data $\mathcal{D}$ given a model and its corresponding parameters $\boldsymbol{\theta}$. The optimal parameters are obtained by solving the optimization problem

$$\min_{\boldsymbol{\theta}\in\Theta} \; J(\boldsymbol{\theta}) \quad \text{with} \quad J(\boldsymbol{\theta}) = -\log\mathcal{L}(\boldsymbol{\theta})\,. \tag{2.19}$$

Due to numerical reasons, the negative log-likelihood is minimized rather than the negative likelihood. Both functions have the same minima, but the negative log-likelihood does not suffer from problems occurring when small probabilities and their products are numerically evaluated to zero. Solving (2.19) yields the maximum likelihood estimate (MLE)

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\; \mathcal{L}(\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\; J(\boldsymbol{\theta}).\end{aligned} \tag{2.20}$$

This optimization problem can efficiently be solved employing the gradient of the log-likelihood function (Nocedal and Wright, 2006; Raue et al., 2013). The local optimization which employs gradient information can, e.g., perform a line-search to determine the next parameter vector, or choose the next parameter vector based on an approximation of the likelihood function within a trust-region (see, e.g., (Nocedal and Wright, 2006) for more details). Most biological meaningful parameters are non-negative and thus can be log-transformed. In a comprehensive study, we showed that this substantially improves optimization (Hass et al., 2019). We use multi-start local optimization, which starts the local optimization from randomly sampled initial points (Raue et al., 2013). Using a likelihood function to calibrate the models in contrast to pure norm-based approaches has the advantage that statistical methods, as will be elaborated in more detail later in this chapter, can be employed to perform model selection.

In the following, we specify the structure of the likelihood functions for different data types and mathematical models. We again distinguish the cases of population average (2.1) and single-cell snapshot data (2.2).

2.3.1 Likelihood function for population average data

Since experiments are mostly subject to measurement noise, linking of observables $\mathbf{y}$, defined in (2.5), and measurements $\bar{y}$, defined in (2.1), requires a noise model

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i,k} p(\bar{y}_{i,k}|y_i(t_k,\boldsymbol{\theta},\mathbf{u}),\boldsymbol{\varphi}_i(t_k,\boldsymbol{\theta},\mathbf{u}))\,, \tag{2.21}$$

$$J(\boldsymbol{\theta}) = -\log\mathcal{L}(\boldsymbol{\theta}) = -\sum_{i,k}\log p(\bar{y}_{i,k}|y_i(t_k,\boldsymbol{\theta},\mathbf{u}),\boldsymbol{\varphi}_i(t_k,\boldsymbol{\theta},\mathbf{u})), \tag{2.22}$$

with distribution parameters $\boldsymbol{\varphi}$. For this, independence of measurement noise for different time points and observables is assumed. Most commonly, a Gaussian distribution is assumed with standard deviation $\sigma = \boldsymbol{\varphi}$, which is also estimated from the data. The calculation of the gradient of the negative log-likelihood function with respect to θ_j requires the derivative of $y_i(t_k,\boldsymbol{\theta},\mathbf{u})$ with respect to θ_j. This can reliably be computed by extending the ODE system by forward sensitivities (Raue et al., 2013; Sengupta et al., 2014) or calculating adjoint sensitivites (Fröhlich et al., 2017). This thesis is concerned with the choice of noise distribution, and different assumptions for the noise model are employed and analyzed in Chapter 3.

2.3.2 Likelihood function for single-cell snapshot data

As stated above, we focus on likelihood-based calibration of the mathematical models, since norm-based calibration as, e.g., used in (Dixit et al., 2019; Hasenauer et al., 2010, 2011b; Munsky and Khammash, 2010) does not allow for a statistical comparison of different models as it will be introduced in Section 2.4, or uncertainty analysis as it will be introduced in Section 2.3.3. If moments of the distributions are fitted rather than the whole distribution, the likelihood comprises terms for each of the considered moments (Fröhlich et al., 2016; Zechner et al., 2012). For fitting mean and covariances, this likelihood is given by

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \prod_{i,k} p(\bar{m}_{i,k}|m_i(t_k,\boldsymbol{\theta},\mathbf{u}),\boldsymbol{\theta}) \cdot \prod_{i,j,k} p(\bar{C}_{i,j,k}|C_{i,j}(t_k,\boldsymbol{\theta},\mathbf{u}),\boldsymbol{\theta})\,, \tag{2.23}$$

with sample mean and sample covariance matrix

$$\begin{aligned}\bar{m}_{i,k} &= \frac{1}{n_c}\sum_c \bar{y}^c_{i,k}\,,\\ \bar{C}_{i,j,k} &= \frac{1}{n_c-1}\sum_c \left(\bar{y}^c_{i,k}-\bar{m}_{i,k}\right)\left(\bar{y}^c_{j,k}-\bar{m}_{j,k}\right)\,.\end{aligned}$$

In this thesis, we consider a likelihood function, for which the probability density provided by the model is evaluated for each cell

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \prod_{i,k,c} p(\bar{y}^c_{i,k}|\boldsymbol{\theta})\,. \tag{2.24}$$

This is employed, for example, in RRE-constrained mixture modeling (2.17), where p is a normal or log-normal mixture distribution, by Filippi et al. (2013), where p is given by a log-normal distribution, and by Fox and Munsky (2019), where p is obtained by the finite state projection. Other approaches, for example, employ the area between empirical cumulative density functions of the observed data and simulated cumulative density function (Fischer et al., 2019).

2.3.3 Uncertainty analysis

The maximum likelihood estimate provides a single point estimator for the model parameters, but does not provide information about the uncertainties of the parameter estimates. These are especially important for model predictions. The uncertainty of parameters can be assessed using profile likelihoods or Bayesian sampling.

The **profile likelihood** is given by

$$\mathrm{PL}(\theta_i) = \max_{\theta_{j\neq i},\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta})\,, \tag{2.25}$$

for parameter θ_i (Raue et al., 2009). For a fixed value of θ_i, the maximum over all remaining parameters is computed. Based on the profiles, the confidence intervals (CIs) of parameter θ_i for significance level α can be defined as

$$\mathrm{CI}_{i,\alpha} = \left\{ \theta_i \middle| \frac{\mathrm{PL}(\theta_i)}{\mathcal{L}(\hat{\boldsymbol{\theta}})} > \exp\left(-\frac{\Delta_\alpha}{2}\right) \right\}, \tag{2.26}$$

(Meeker and Escobar, 1995). Here, Δ_α is the αth percentile of the χ^2 distribution with one degree of freedom. The profiles can be calculated by repeated optimization (Raue et al., 2015) or by an integration-based approach (Boiger et al., 2016; Stapor et al., 2018a).

Alternatively to profile likelihoods, **posterior distributions** can be assessed. The posterior is defined according to Bayes' theorem by

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}. \tag{2.27}$$

The numerator comprises the likelihood and the prior $p(\boldsymbol{\theta})$, which encodes prior knowledge about the parameter values. If the prior is a uniform distribution on the interval defined by the parameters boundaries, the posterior corresponds to the likelihood subject to constraints. The denominator consists of the evidence $p(\mathcal{D})$, which in most cases is difficult to compute, but can be ignored for parameter inference and sampling since it is independent of $\boldsymbol{\theta}$. Maximizing the posterior distribution gives the maximum a posteriori (MAP) estimate. Often samples of the posterior distributions are studied to directly assess the parameter uncertainty and report credible intervals. For sampling, Markov chain Monte Carlo (MCMC) approaches can be employed such as the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953), the adaptive Metropolis algorihm (Haario et al., 2001) or algorithms with multiple chains like parallel tempering (Lacki and Miasojedow, 2015; Neal, 1996) or parallel hierarchical sampling (Rigat and Mira, 2012).

2.3.4 Evaluation of parameter estimation results and distribution assumptions

In this thesis, new methods for the estimation of the parameters are introduced and different distribution assumptions assessed. To compare state-of-the-art and newly developed methods as well as to compare the appropriateness of different distribution assumptions, the following criteria are considered: We evaluate a method for optimization based on the number of converged starts. The number of converged starts is defined as the number of starts for which the distance between the final objective function value and the best found value across all methods is below a certain threshold (Figure 2.1A). This threshold can for example be motivated by a likelihood ratio test (Hross and Hasenauer, 2016). Using the number of converged starts as measure for the performance of the algorithm can be

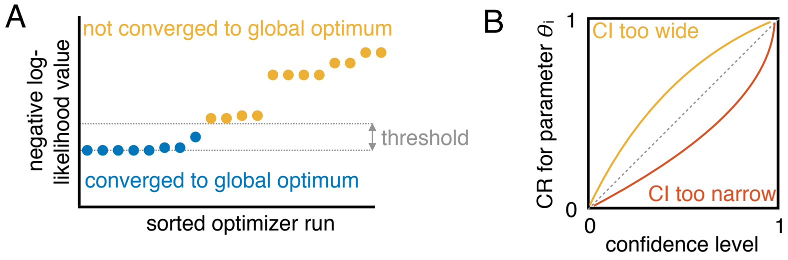

Figure 2.1: Illustration of the evaluation of optimization results and distribution assumptions. (A) Likelihood waterfall plot. Optimization runs are considered to be converged if the distance between the negative log-likelihood values to the minimal value found is below a certain threshold. The runs are sorted with respect to their final negative log-likelihood values. (B) Appropriateness of CIs is assessed by comparing the coverage ratio (CR) with the confidence level. Ideally, the CR should be close to the corresponding confidence level, yielding the dashed line.

misleading, because it does not take into account the computation time. However, computation resources are often the limiting factor. Thus performance can be quantified by the number of converged starts per given time unit.

In simulation studies, the true parameters, which were used to generate the data, are known and can be used to assess the ability of the method to obtain the true parameters. The accuracy of the MLE, apart from non-identifiabilities, for different distribution assumptions can be evaluated using the mean squared error (MSE)

$$\begin{aligned}\mathrm{MSE}\left[\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^{\mathrm{true}}\right] &= \mathbb{E}\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{true}}\right)^2\right] \\ &= \underbrace{\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}]\right)^2\right]}_{\mathrm{Var}(\hat{\boldsymbol{\theta}})} + \Big(\underbrace{\mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^{\mathrm{true}}}_{\mathrm{Bias}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^{\mathrm{true}})}\Big)^2 .\end{aligned} \tag{2.28}$$

It incorporates the variance and the bias of the estimator. A small (norm of the) MSE indicates a good agreement of the true and estimated parameters. However, the uncertainty of the parameter estimates is not taken into account when analyzing the MSE.

Uncertainty is included in the analysis by studying the CIs, defined in (2.26). For given confidence level α, the CIs should cover the true parameter with a frequency of $1 - \alpha$. Accordingly, if the true parameter is known, the appropriateness of the CIs can be evaluated by computing the coverage ratio (CR), which is the probability that the true parameter is contained in the CI (Figure 2.1B). It can be calculated by repeating the optimization procedure and profiling for many generated data sets. The CR should be close to the desired confidence level (Schelker et al., 2012).

2.4 Model selection

This section is modified from Section 3 of (Fröhlich et al., 2019) which I contributed to the book chapter.

In many applications, it is not apparent which biochemical species and reactions are necessary to describe the dynamics of a biochemical process. In this case, the structure of the ODE model (2.4) has to be inferred from experimental data. The selection should compromise between goodness-of-fit and complexity. Following the concept of Occam's razor (Blumer et al., 1987), one tries to control variability associated with over-fitting while protecting against the bias associated with under-fitting.

2.4.1 Model selection criteria

Given a set of candidate models $M_1, M_2, \ldots, M_{n_M}$, the aim of model inference is to find a model or a set of models which (i) describe the data available and (ii) generalize to other data sets (Hastie et al., 2009). The choice of model can be made based on several selection criteria, differing among others in asymptotic consistency (Shibata, 1980), asymptotic efficiency (Fisher, 1922) and computational complexity. If the true model is included in the set of candidate models, a consistent criterion will asymptotically, for an increasing number of data points, select the true model with probability one and an efficient criterion will select the model that minimizes the MSE of the prediction.

One popular criterion is the **Bayes factor** (Kass and Raftery, 1995), which has been shown to be asymptotically consistent for a broad range of models (Choi and Rousseau, 2015; Wang and Sun, 2014). However, for the case of general ODE models, no proofs for asymptotic efficiency and consistency are available for all the criteria presented in this section. Bayes' theorem yields the posterior model probability

$$p(M_m|\mathcal{D}) = \frac{p(\mathcal{D}|M_m)p(M_m)}{p(\mathcal{D})}\,, \tag{2.29}$$

with marginal likelihood

$$p(\mathcal{D}|M_m) = \int_{\Theta_m} p(\mathcal{D}|\boldsymbol{\theta}_m, M_m)p(\boldsymbol{\theta}_m|M_m)d\boldsymbol{\theta}_m\,, \tag{2.30}$$

with model prior $p(M_m)$ and marginal probability $p(\mathcal{D}) = \sum_j p(\mathcal{D}|M_j)P(M_j)$. The Bayes factor of models M_m and M_l is the ratio of the corresponding marginal likelihoods

$$B_{ml} = \frac{p(\mathcal{D}|M_m)}{p(\mathcal{D}|M_l)}. \tag{2.31}$$

The Bayes factor describes how much more likely it is that the data are generated from M_m instead of M_l. A Bayes factor $B_{ml} > 100$ is often considered decisive for rejecting model M_l (Jeffreys, 1961). The Bayes factor intrinsically penalizes model complexity by integrating over the whole parameter space of each model. Bayes factors can be approximated by Laplace approximation, which has a low computational complexity but provides only a local approximation. To enable a more precise computation of the Bayes factors, thermodynamic integration can be employed to evaluate (2.30) (Lartillot and Philippe, 2006). This approach uses the tempered posterior

$$p_\tau(\boldsymbol{\theta}_m|\mathcal{D}, M_m) = \frac{p(\mathcal{D}|\boldsymbol{\theta}_m, M_m)^\tau p(\boldsymbol{\theta}_m|M_m)}{\int_{\Theta_m} p(\mathcal{D}|\boldsymbol{\theta}_m, M_m)^\tau p(\boldsymbol{\theta}_m|M_m) d\boldsymbol{\theta}_m}, \tag{2.32}$$

with parameter $\tau \in [0, 1]$. For $\tau = 0$, (2.32) corresponds to the prior $p(\boldsymbol{\theta}_m|M_m)$ and for $\tau = 1$ it corresponds to the untempered posterior. It can then be shown that (Lartillot and Philippe, 2006)

$$\log p(\mathcal{D}|M_m) = \int_0^1 \mathbb{E}_{p_\tau} \left[\log p(\mathcal{D}|\boldsymbol{\theta}_m, M_m)\right] d\tau\,. \tag{2.33}$$

This integral can numerically easier be solved by choosing a discretization $0 = \tau_0 < \tau_1 < \ldots < \tau_{n_\tau - 1} = 1$, evaluating the integrand in (2.33) for the n_τ discretization steps by Monte Carlo sampling and applying, e.g., the trapezoidal or Simpsons' rule. The number of temperatures n_τ can also be chosen adaptively (Hug et al., 2016). Other methods which can be employed for the evaluation of the marginal likelihood are, e.g., bridge sampling (Meng and Wong, 1996), nested sampling (Skilling, 2006) or related methods. As the approaches require a large number of function evaluations, the methods are usually computationally demanding and the computational complexity is highly problem-dependent. Thus, efficient sampling methods are required.

For high-dimensional or computationally costly problems, the calculation of Bayes factors can be intractable and computationally less expensive model selection criteria need to be employed. A model selection criterion which is based on the MLE instead of a marginal likelihood (an integral over the whole parameter space), is the **Bayesian Information**

Criterion (BIC) (Schwarz, 1978). The BIC value for model M_m is

$$\mathrm{BIC}_m = -2\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m, M_m) + \log(n_{\mathcal{D}})\, n_{\theta_m}\,, \tag{2.34}$$

with $n_{\mathcal{D}}$ data points. For structurally identifiable models, the BIC provides in the limit of large sample sizes information about the Bayes factors (Kass and Raftery, 1995),

$$\lim_{n_{\mathcal{D}}->\infty} \frac{-2\log B_{ml} - (\mathrm{BIC}_m - \mathrm{BIC}_l)}{-2\log B_{ml}} = 0\,. \tag{2.35}$$

From information theoretical arguments, the **Akaike Information Criterion** (AIC)

$$\mathrm{AIC}_m = -2\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m, M_m) + 2n_{\theta_m}\,, \tag{2.36}$$

has been derived (Akaike, 1973). Low BIC and AIC values are preferable and differences above 10 are assumed to be substantial (see Table 2.1) (Burnham and Anderson, 2002; Kass and Raftery, 1995). For model selection in many problem classes, the AIC is asymptotically efficient, but not consistent, while the BIC is asymptotically consistent, but not efficient (Acquah, 2010; Kuha, 2004; Shibata, 1981). When incorporating prior information about parameters, the priors can conceptually be treated as additional data points and, thus, be part of the likelihood to still allow the use of BIC and AIC. While Bayes factors are proven to be valid for non-identifiable parameters, the use of AIC and BIC can be problematic for these cases.

The **log pointwise predictive density** for model M_m (Gelman et al., 2014)

$$\mathrm{lppd}_m = \sum_{i=1}^{n_{\mathrm{subs.}}} \log\left(\frac{1}{n_{\mathrm{sam.}}} \sum_{j=1}^{n_{\mathrm{sam.}}} p(\mathcal{D}_i|\boldsymbol{\theta}_m^{i,j}, M_m)\right)\,, \tag{2.37}$$

evaluates the predictions of model M_m. For the calculation in (2.37), (i) the full data set is split into $n_{\mathrm{subs.}}$ subsets $\mathcal{D}_i$, (ii) for each combination of $n_{\mathrm{subs.}} - 1$ subsets, samples from the posterior distribution are collected and (iii) for each subset the logarithm of the average likelihood is calculated based on samples $\{\boldsymbol{\theta}_m^{i,j}\}$ from the posterior distribution for all remaining subsets. The superscript i here indicates the samples of the posterior for all subsets but the one indexed by i.

Further model selection criteria exist, such as the corrected AIC (Hurvich and Tsia, 1989), which provides a correction for finite sample sizes, cross-validation (Arlot and Celisse, 2010) or the likelihood ratio test, which is efficient and can only be applied to nested models (Neyman and Pearson, 1992; Wilks, 1938).

Table 2.1: Decisions based on the Bayes factor and differences in BIC and AIC values (Burnham and Anderson, 2002; Jeffreys, 1961; Kass and Raftery, 1995).

B_{lm}	$\mathrm{BIC}_m - \min_l \mathrm{BIC}_l$	$\mathrm{AIC}_m - \min_l \mathrm{AIC}_l$	decision
$1 - 3$	$0 - 2$	$0 - 4$	do not reject model M_m
$3 - 100$	$2 - 10$	$4 - 10$	-
> 100	> 10	> 10	reject model M_m

2.4.2 Reduction of the number of models

For most models, computing Bayes factors is computationally demanding compared to optimization and the evaluation of AIC, BIC or likelihood ratio. Yet, if the number of candidate models n_M is large, even the evaluation of AIC or BIC can become limiting as n_M optimization problems have to be solved. For non-nested models, the model selection criterion of choice needs to be calculated for each model to determine the optimal model. For a nested set of candidate models all candidate models are a special case of a comprehensive model and can be constructed by fixing a subset of the parameters to specific values (Figure 2.2A). We assume that we can split the model parameters $\boldsymbol{\theta}$ into general parameters $\boldsymbol{\eta} \in \mathbb{R}^{n_\eta}$, which are present in all models, and difference parameters $\boldsymbol{\kappa} \in \mathbb{R}^{n_\kappa}$, which encode the nesting between models and could for example be the kinetic rates of hypothesized reactions (Klimovskaia et al., 2016) or scaling factors for possibly cell-type or condition-specific parameters (Steiert et al., 2016). Such settings yield a total of 2^{n_κ} candidate models, where n_κ is limited by n_θ. Thus, for models with a high number of parameters, also a high number of nested models is possible. When n_κ and n_θ are both large, the inference of model parameters and thus the inference of model structure is challenging.

In statistics, step-wise regression is an often-used approach to reduce the number of models that need to be tested. This comprises **forward-selection and backward-elimination** (Hastie et al., 2009) and combinations of both (Kaltenbacher and Offtermatt, 2011). Forward-selection is a bottom-up approach which starts with the least complex model and successively activates individual difference parameters (i.e., setting $\kappa_i \neq 0$) until a sufficient agreement with experimental data is achieved, evaluated using a model selection criterion (Figure 2.2B). In contrast, backward-elimination is a top-down approach starting with the most complex model, successively deactivating individual difference parameters (i.e., setting $\kappa_i = 0$) that are not required for a good fit to the data. Forward-selection and backward-elimination reduce the number of models that need to be compared with the model selection criteria described before from 2^{n_κ} to at most $0.5 \cdot n_\kappa(n_\kappa + 1)$. However,

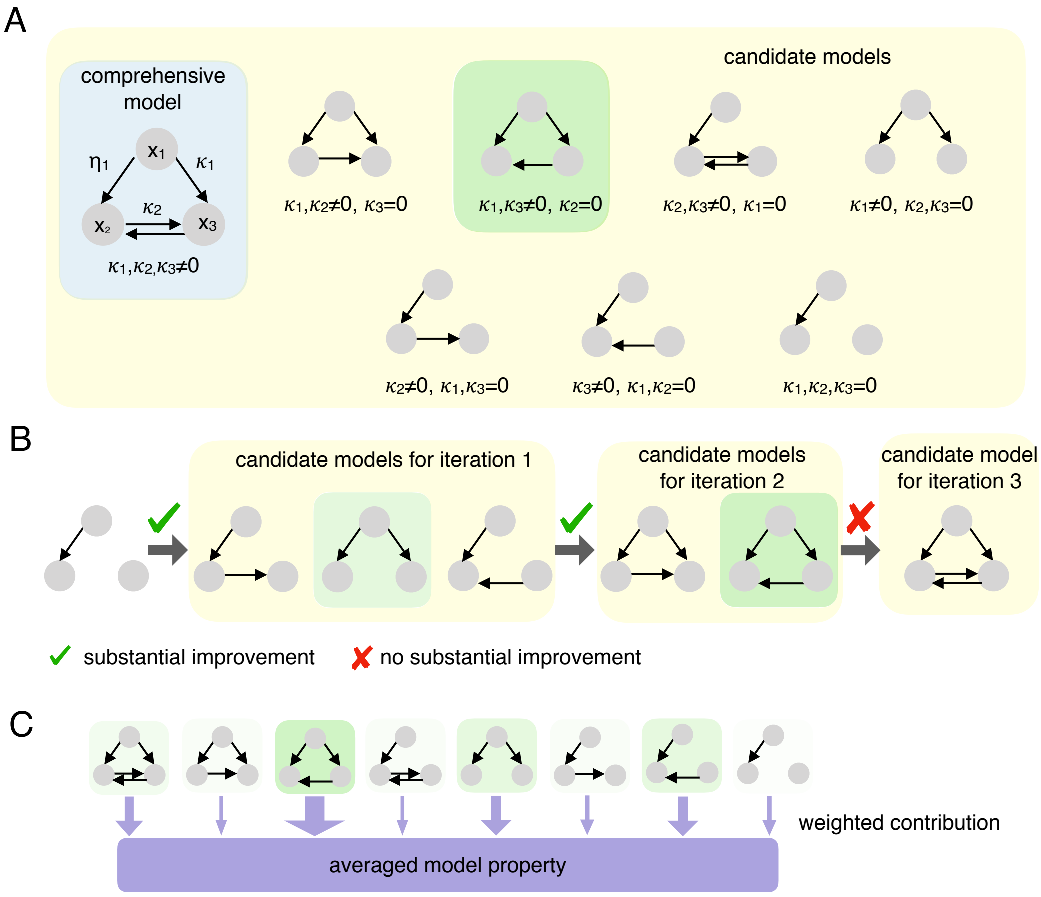
A
comprehensive model
η_1
x_1
κ_1
κ_2
x_2
x_3
$\kappa_1,\kappa_2,\kappa_3\neq0$
candidate models
$\kappa_1,\kappa_2\neq0,\ \kappa_3=0$
$\kappa_1,\kappa_3\neq0,\ \kappa_2=0$
$\kappa_2,\kappa_3\neq0,\ \kappa_1=0$
$\kappa_1\neq0,\ \kappa_2,\kappa_3=0$
$\kappa_2\neq0,\ \kappa_1,\kappa_3=0$
$\kappa_3\neq0,\ \kappa_1,\kappa_2=0$
$\kappa_1,\kappa_2,\kappa_3=0$
B
candidate models for iteration 1
candidate models for iteration 2
candidate model for iteration 3
✓ substantial improvement
✗ no substantial improvement
C
weighted contribution
averaged model property

Figure 2.2: Illustration of methods for model reduction. (A) Set of candidate models, varying in the existence of connections between nodes x_1, x_2 and x_3. In total, there are $2^{n_\kappa} = 2^3$ models with at least $n_\eta = 1$ parameters. (B) Illustration of forward-selection starting from minimal model. In the first iteration, the model with $\kappa_1 \neq 0, \kappa_2, \kappa_3 = 0$ is selected (green) and in the second iteration the model with $\kappa_1, \kappa_3 \neq 0, \kappa_2 = 0$. The full model is rejected based on the selection criteria. (C) Illustration of model averaging. The thickness of the arrows corresponds to the posterior probability, Akaike weight or BIC weight and indicates the contribution of the model to the averaged model properties. This figure and its legend is modified from Figure 2 of the author's publication (Fröhlich et al., 2019).

they are both greedy approaches and do not guarantee to find the globally least complex candidate model that explains the data. An alternative approach is to penalize the number of parameters in the objective function to enforce sparsity of the parameters (see our review (Fröhlich et al., 2019) for a more detailed discussion).

2.4.3 Model averaging

For large sets of candidate models and limited data, it frequently happens that not a single model is chosen by the model selection criterion. Instead, a set of models is plausible, cannot be rejected and should be considered in the subsequent analysis. In this case, model averaging can be employed to predict the behavior of the process (Figure 2.2C). Given that a certain parameter is comparable between models, an average estimate can be derived as

$$\mathbb{E}[\theta_i] = \sum_m \omega_m \hat{\theta}_{i,m} \,, \tag{2.38}$$

with ω_m denoting the weight of model M_m and $\hat{\theta}_{i,m}$ denoting the MLE of parameter θ_i for model M_m (Burnham and Anderson, 2002; Wassermann, 2000). The weights capture the plausibility of the model. An obvious choice for the weights is the posterior probability $p(M_m|\mathcal{D})$. Alternatively, BIC weights

$$\omega_m = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_m)}{\sum_{i=1}^{n_M} \exp(-\frac{1}{2}\mathrm{BIC}_i)} \,, \tag{2.39}$$

or Akaike weights, where the BIC is replaced by the AIC in (2.39), can be employed. The weights for models that are not plausible are close to zero and, thus, these models do not influence the output of the averaged model.

2.5 Implementation

In this section, the toolboxes are listed which were employed for the analyses described in the following chapters.

- AMICI (Fröhlich et al., 2017) provides an interface to the SUNDIALS solver suite (Hindmarsh et al., 2005) and enables the efficient simulation of the ODEs obtained by the RREs (2.4) and the corresponding sensitivities. In particular, AMICI employs the solver CVODES (Serban and Hindmarsh, 2005).
- CERENA derives and provides the simulation functions for the moment-closure approximation (2.14, 2.15) (Kazeroonian et al., 2016).
- SPToolbox provides the sigma-point approximation (2.10, 2.11).

- PESTO was employed for parameter estimation and uncertainty analysis (Stapor et al., 2018b). It provides an interface to the MATLAB optimization routine `fmincon`. PESTO also comprises algorithms for parallel tempering.

- ODEMM provides the implementation for RRE-constrained mixture models (2.17) (Hasenauer et al., 2014) and the hierarchical population model which will be introduced in Chapter 4.

All these toolboxes are available on GitHub under `https://github.com/ICB-DCM`. The author of this thesis developed the toolbox ODEMM and co-developed the toolbox PESTO. For the latter, the author contributed in particular the implementations for the hierarchical optimization as introduced in Section 3.2. The implementations for the analyses performed in this thesis are also freely available on GitHub and/or the Supplementary Information of the corresponding publications.

Chapter 3

Robust and efficient calibration of ODE models on population average data

ODE models are valuable tools to study biological processes. Their model parameters are often estimated based on population average data (Section 2.1.1). This requires the definition of the differential equations (2.4), the observable function (2.5) and the measurement noise model (2.6). The combination of these parts of the model should reflect the biological process as well as the process of data collection. This can require the introduction of observable parameters, e.g., scaling parameters which map the states of the biochemical species to the measurable output if only relative measurements can be collected. In addition, the measurements are often corrupted by outliers and follow a distribution which is a mixture of the unknown outlier-generating distribution, smaller measurement noise and the true biological process. This should also be captured by the measurement noise model. However, with increasing numbers of measurements and experiments which are integrated and fitted simultaneously, the models become more and more complex, which hinders their calibration and requires efficient methods.

In this chapter, we address the open problems which have been stated in Section 1.2: (i) the influence and handling of outliers; (ii) the high number of observation parameters; and (v) the dynamics of histone methylation. In particular, we investigate the measurement noise model by incorporating heavier tailed distributions in the maximum likelihood framework used for model calibration. We then study the choice of the distribution p in (2.22) and assess the properties of the distributions for simulated data of a conversion process. Afterwards, we provide an approach for efficiently estimating the model parameters by splitting the optimization problem into smaller subproblems, using the analytically calculated optimal values for the observable parameters and noise parameters in each optimization step. This is followed by a study on the kinetics of histone H3 methylation to address problem (v), where we apply the developed methods to compare models which differ in their equations and represent different biological hypotheses.

This chapter is based on and in part identical with these publications:

- Maier, C., **Loos, C.**, & Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5), 718-725.
- **Loos, C.***, Krause, S.*, & Hasenauer, J. (2018). Hierarchical optimization for the efficient parametrization of ODE models. *Bioinformatics*, 34(24), 4266–4273. (*equal contribution)
- **Loos, C.***, Völker-Albert, M.*, Forne, I., Hasenauer, J., Imhof, A., Marr, C., Groth, A., Alabert, C. Efficient K27me3 establishment requires naive histone substrates and pre-existing K27me3 on old histones. *in preparation.*

3.1 Robust calibration of ODE models in the presence of outliers

In addition to general measurement noise, various errors can occur in the process of data collection, including human errors, e.g., labeling or technical errors. Human errors usually yield bigger deviations, while measurement noise is generally smaller. The distorted measurements, often referred to as outliers, can influence further analyses and thus yield incorrect results (Hawkins, 1980; Tarantola, 2005). Using an assumption about measurement noise in the model calibration which cannot cope with the outliers can yield incorrect parameter estimates. This limits the validity of the models and hinders model predictions.

Methods which try to detect and remove the outliers from the data set exist in various fields (Ben-Gal, 2005; Hodge and Austin, 2004; Niu et al., 2011). However, this detection is potentially not easy due to noisy measurements and the increasing size and complexity of the data. The removal of data points which were wrongly assigned as outliers, as well as keeping true outliers in the data set can distort further analyses (Motulsky and Christopoulos, 2003). To circumvent the manual removal of the outliers, the employed method needs to be able to directly cope with the outliers.

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

The methods developed in the field of robust regression can in principle be applied across scientific fields. Each field has, however, its particularities regarding experimental data, e.g., noise levels, outlier generating mechanisms and mathematical models which influence the performance. For dynamical models of biological systems, the Huber M-estimator was already successfully applied, yielding more reliable parameter estimates (Cao et al., 2011; Qiu et al., 2016). A comprehensive evaluation of different methods in the field of quantitative biology is still missing. Furthermore, the standard formulation as a regression problem does not allow one to perform model selection using statistical methods such as the likelihood ratio test, the Akaike or the Bayesian information criterion in a straightforward way (Section 2.4.1). To facilitate model selection for the mechanistic as well as the statistical model, a formulation of robust estimation in terms of (normalized) probability distributions would be beneficial.

In this section, we provide the equations for the likelihood functions and its analytical gradients of distributions with heavier tails than the Gaussian distribution. We compare robustness and reliability of the models for these different distributions for the case of outlier-free and outlier-corrupted data sets.

Notation: In the following, we simplify the notation and define the index $j = 1, \ldots, n_{\mathcal{D}}$ for all measured data points, comprising observables, time points, dosages, experiments and replicates. We only explicitly note the dependence on the parameters $\boldsymbol{\theta}$. The dependence on, e.g., the states, input and time is implicitly captured by index j:

$$y_j(\boldsymbol{\theta}) = h_j(\boldsymbol{\theta})\,. \tag{3.1}$$

3.1.1 Equations for heavier tailed noise distributions

The distributions, along with their distribution parameters, are listed in Table 3.1 and visualized in Figure 3.1. While the Gaussian and Laplace distributions have well-defined moments for all values for the distribution parameters, the Students' t distribution only has a well-defined mean for $\nu > 1$ and a well-defined variance for $\nu > 2$. The variance of the Cauchy distribution is always infinite. We refer to the Laplace, Huber, Cauchy and Student's t distributions in the following as heavier tailed distributions, since their tails are heavier than the tails of a Gaussian distribution.

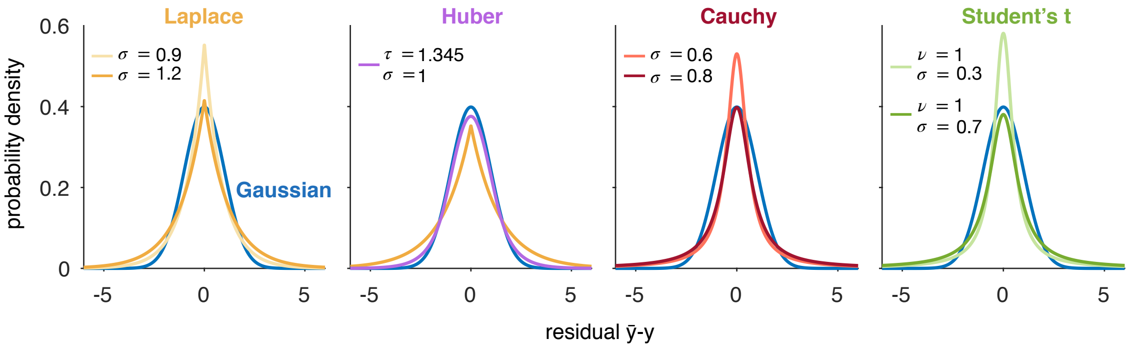

Figure 3.1: Distribution assumptions. Distributions with heavier tails than the Gaussian distribution are shown for different scale parameters in comparison with a standard Gaussian distribution $\mathcal{N}(0,1)$. For the Huber distribution, the corresponding Gaussian and Laplace distributions are indicated.

For parameter estimation, the distributions were used in the likelihood function (2.22). The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\varphi})$ comprises the dynamic parameters as well as the distribution parameters.

Under the generally used **Gaussian** distributed measurement noise, the log-likelihood function is given by given by

$$\log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_j \left[\log\left(2\pi\sigma_j(\boldsymbol{\theta})^2\right) \ + \left(\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})}\right)^2\right], \tag{3.2}$$

with standard deviation $\sigma_j(\boldsymbol{\theta}) > 0$. The gradient of the log-likelihood for $i = 1, \ldots, n_\theta$ is given by

$$\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_i} = -\frac{1}{2}\sum_j \left[\frac{1}{\sigma_j^2(\boldsymbol{\theta})}\left(1 - \left(\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})}\right)^2\right)\frac{\partial \sigma_j^2(\boldsymbol{\theta})}{\partial \theta_i} - 2\,\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})}\frac{\partial y_j(\boldsymbol{\theta})}{\partial \theta_l}\right].$$

Under the assumption of independent **Laplace** distributed measurement noise, the log-likelihood function is given by

$$\log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\sum_j \left[\log\left(2\sigma_j(\boldsymbol{\theta})\right) \ + \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j(\boldsymbol{\theta})}\right], \tag{3.3}$$

with scale $\sigma_j(\boldsymbol{\theta}) > 0$. The likelihood function is non-differentiable for parameters $\boldsymbol{\theta}$, for which $\exists j : \bar{y}_j = y_j(\boldsymbol{\theta})$. The gradient of the log-likelihood for $i = 1, \ldots, n_\theta$ is given by

$$\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_i} = \sum_j \left[\left(-\frac{1}{\sigma_j(\boldsymbol{\theta})} + \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j^2(\boldsymbol{\theta})} \right) \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial \theta_i} + \frac{\operatorname{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\sigma_j(\boldsymbol{\theta})} \frac{\partial y_j(\boldsymbol{\theta})}{\partial \theta_i} \right],$$

and we used $\operatorname{sgn}(0) = 0$.

The **Huber** M-estimator exploits a combination of squared 2-norm and 1-norm for penalization. Residuals with absolute value below τ are penalized quadratically while residuals with absolute values larger τ are penalized linearly. Under the assumption of independent Huber distributed measurement noise, the log-likelihood function is given by

$$\log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_j \left[\log(c_{\text{Huber},j}(\boldsymbol{\theta})) - \begin{cases} \frac{1}{2}(\text{res}_j(\boldsymbol{\theta}))^2 & |\text{res}_j(\boldsymbol{\theta})| \leq \tau(\boldsymbol{\theta}) \\ \frac{1}{2}(2\tau(\boldsymbol{\theta})|\text{res}_j(\boldsymbol{\theta})| - \tau^2(\boldsymbol{\theta})) & |\text{res}_j(\boldsymbol{\theta})| > \tau(\boldsymbol{\theta}) \end{cases} \right], \tag{3.4}$$

with $\text{res}_j(\boldsymbol{\theta}) = (\bar{y}_j - y_j(\boldsymbol{\theta}))/\sigma_j(\boldsymbol{\theta})$, scale $\sigma_j(\boldsymbol{\theta}) > 0$ and tuning parameter $\tau(\boldsymbol{\theta}) > 0$. The constant

$$c_{\text{Huber},j}(\boldsymbol{\theta}) = \left(\sqrt{2\pi}\sigma_j(\boldsymbol{\theta}) \text{erf}\left(\frac{\tau(\boldsymbol{\theta})}{\sqrt{2}} \right) + \frac{2\sigma_j(\boldsymbol{\theta})}{\tau(\boldsymbol{\theta})} \exp\left(-\frac{1}{2}\tau^2(\boldsymbol{\theta}) \right) \right)^{-1},$$

normalizes the function such that it is a probability density and possesses integral 1. Here,

$$\text{erf}(x) = \frac{2}{\pi} \int_0^x \exp\left(-t^2\right) dt,$$

denotes the error function. The gradient of the log-likelihood for $i = 1, \ldots, n_\theta$ is given by

$$\begin{aligned} \frac{\partial \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_i} = \sum_j \Bigg[& \frac{\partial y_j(\boldsymbol{\theta})}{\partial \theta_i} \cdot \begin{cases} \frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} & |\text{res}_j(\boldsymbol{\theta})| \leq \tau(\boldsymbol{\theta}) \\ \frac{\tau(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} \operatorname{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta})) & |\text{res}_j(\boldsymbol{\theta})| > \tau(\boldsymbol{\theta}) \end{cases} \\ & + \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial \theta_i} \left(-\frac{1}{\sigma_j(\boldsymbol{\theta})} + \begin{cases} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^3(\boldsymbol{\theta})} & |\text{res}_j(\boldsymbol{\theta})| \leq \tau(\boldsymbol{\theta}) \\ \frac{\tau(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} |\bar{y}_j - y_j(\boldsymbol{\theta})| & |\text{res}_j(\boldsymbol{\theta})| > \tau(\boldsymbol{\theta}) \end{cases} \right) \\ & + \frac{\partial \tau(\boldsymbol{\theta})}{\partial \theta_i} \left(\frac{\frac{2}{\tau^2(\boldsymbol{\theta})} \exp(-\frac{1}{2}\tau^2(\boldsymbol{\theta}))}{\sqrt{2\pi}\text{erf}\left(\frac{\tau(\boldsymbol{\theta})}{\sqrt{2}}\right) + \frac{2}{\tau(\boldsymbol{\theta})} \exp\left(-\frac{1}{2}\tau^2(\boldsymbol{\theta})\right)} \right. \\ & - \left. \begin{cases} 0 & |\text{res}_j(\boldsymbol{\theta})| \leq \tau(\boldsymbol{\theta}) \\ \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j(\boldsymbol{\theta})} - \tau(\boldsymbol{\theta}) & |\text{res}_j(\boldsymbol{\theta})| > \tau(\boldsymbol{\theta}) \end{cases} \right) \Bigg]. \end{aligned}$$

Table 3.1: Probability densities. The probability density functions for the Gaussian, Laplace, Huber, Cauchy and Student's t distribution are listed together with the parameters defining the distributions. The error function is denoted by erf and the gamma function by Γ. Note that the distribution parameter σ has a different meaning for the different distributions and is not comparable across these. This table has been modified from Table 1 from the author's publication (Maier et al., 2017).

	probability density $p(\bar{y}\|y,\boldsymbol{\varphi})$	distribution parameters $\boldsymbol{\varphi}$
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma}\right)^2\right)$	standard deviation $\sigma > 0$
Laplace	$\frac{1}{2\sigma}\exp\left(-\frac{\|\bar{y}-y\|}{\sigma}\right)$	scale $\sigma > 0$
Huber	$c_{\text{Huber}}\cdot\begin{cases}\exp\left(-\frac{1}{2}\left(\frac{\bar{y}-y}{\sigma}\right)^2\right), & \|\frac{\bar{y}-y}{\sigma}\| \leq \tau \\ \exp\left(-\frac{1}{2}\left(2\tau\|\frac{\bar{y}-y}{\sigma}\| - \tau^2\right)\right), & \|\frac{\bar{y}-y}{\sigma}\| > \tau\end{cases}$	scale $\sigma > 0$, tuning parameter $\tau > 0$
	with $c_{\text{Huber}} = \left(\sqrt{2\pi}\sigma\text{erf}\left(\frac{\tau}{\sqrt{2}}\right) + \frac{2\sigma}{\tau}\exp\left(-\frac{1}{2}\tau^2\right)\right)^{-1}$	
Cauchy	$\frac{1}{\pi\sigma}\frac{\sigma^2}{(\bar{y}-y)^2+\sigma^2}$	scale $\sigma > 0$
Student's t	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{\bar{y}-y}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$	scale $\sigma > 0$, degrees of freedom $\nu > 0$

Under the assumption of independent **Cauchy** distributed measurement noise, the log-likelihood function is given by

$$\log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_j \Big[-\log(\pi) + \log(\sigma_j(\boldsymbol{\theta})) - \log\left(\left(\bar{y}_j - y_j(\boldsymbol{\theta})\right)^2 + \sigma_j^2(\boldsymbol{\theta}) \right) \Big], \tag{3.5}$$

with scale $\sigma_j(\boldsymbol{\theta}) > 0$. The gradient of the log-likelihood for $i = 1, \ldots, n_\theta$ is given by

$$\begin{aligned} \frac{\partial \text{log} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_i} = \sum_j \Bigg[& \left(\frac{1}{\sigma_j(\boldsymbol{\theta})} - 2 \frac{\sigma_j(\boldsymbol{\theta})}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j^2(\boldsymbol{\theta})} \right) \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial \theta_i} + \\ & 2 \frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j^2(\boldsymbol{\theta})} \frac{\partial y_j(\boldsymbol{\theta})}{\partial \theta_i} \Bigg]. \end{aligned}$$

Under the assumption of independent **Student's t** distributed measurement noise, the log-likelihood function is given by

$$\begin{aligned} \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_j \Bigg[& \log \left(\frac{\Gamma\left(\frac{\nu_j(\boldsymbol{\theta})+1}{2}\right)}{\sqrt{\nu_j(\boldsymbol{\theta})\pi}\ \sigma_j(\boldsymbol{\theta})\ \Gamma\left(\frac{\nu_j(\boldsymbol{\theta})}{2}\right)} \right) - \\ & \frac{\nu_j(\boldsymbol{\theta}) + 1}{2} \log\left(1 + \frac{1}{\nu_j(\boldsymbol{\theta})} \left(\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} \right)^2 \right) \Bigg], \end{aligned} \tag{3.6}$$

with scale $\sigma_j(\boldsymbol{\theta}) > 0$ and degrees of freedom $\nu_j(\boldsymbol{\theta}) > 0$. The gradient of the log-likelihood for $i = 1, \ldots, n_\theta$ is given by

$$\begin{aligned} \frac{\partial \text{log} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_i} = \sum_j \Bigg[& \frac{1}{2} \Bigg(\Psi\left(\frac{\nu_j(\boldsymbol{\theta}) + 1}{2} \right) - \Psi\left(\frac{\nu_j(\boldsymbol{\theta})}{2} \right) - \log\left(1 + \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\nu_j(\boldsymbol{\theta})\sigma_j^2(\boldsymbol{\theta})} \right) \\ & - \frac{1}{\nu_j(\boldsymbol{\theta})} + \frac{\nu_j(\boldsymbol{\theta}) + 1}{\nu_j^2(\boldsymbol{\theta})\sigma_j^2(\boldsymbol{\theta})} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}} \Bigg) \frac{\partial \nu_j(\boldsymbol{\theta})}{\partial \theta_i} \\ & - \left(\frac{1}{\sigma_j(\boldsymbol{\theta})} - \frac{\nu_j(\boldsymbol{\theta}) + 1}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\nu_j(\boldsymbol{\theta})\sigma_j^3(\boldsymbol{\theta})} \right) \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial \theta_i} \\ & + \frac{\nu_j(\boldsymbol{\theta}) + 1}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})} \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}} \frac{1}{\nu_j(\boldsymbol{\theta})} \frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} \frac{\partial y_j(\boldsymbol{\theta})}{\partial \theta_i} \Bigg], \end{aligned} \tag{3.7}$$

where Ψ denotes the digamma function, which is the logarithmic derivative of the gamma function Γ. For $\nu = 1$, the Student's t distribution equals the Cauchy distribution, and for $\nu = \infty$, it equals the Gaussian distribution.

To obtain approximations for the Hessians of the distributions, it is often assumed that the deviation between measurement and observable is small and can be neglected. For the Gaussian, Cauchy and Student's t distributions, thus, the terms comprising the second-order sensitivities were neglected and approximations were obtained which only depend on the first-order sensitivities. However, this assumption is not appropriate for the Laplace and Huber distribution, since the terms including the second-order sensitivities have an influence on the Hessian even for small deviations of measurement and observable. Therefore, the Hessian requires the simulation of second-order sensitivities which slows down the computation. Thus, we employed the interior-point algorithm of the MATLAB routine `fmincon`, which uses the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method to approximate the Hessian (Fletcher and Powell, 1963; Goldfarb, 1970). The Hessian matrices can be found in Appendix A.

3.1.2 Simulation example: Conversion reaction

To study the different distribution assumptions, we considered a simple conversion process, which often occurs in biological systems (Figure 3.2A). It involves two biochemical species A and B and the conversion between these species takes place with rate constants k_1 and k_2. The ODEs describing this process are given by

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 + k_2 x_2\,, \quad x_1(0) = 1\,, \\ \dot{x}_2 &= k_1 x_1 - k_2 x_2\,, \quad x_2(0) = 0\,, \end{aligned} \tag{3.8}$$

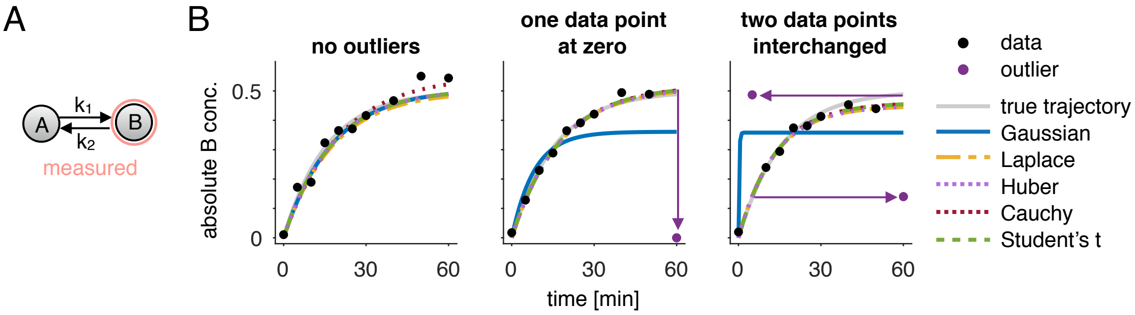

Figure 3.2: Data and fits for different scenarios and distribution assumptions. (A) Model of a conversion process of two species A and B. (B) The data points are generated by simulating the system with Gaussian distributed noise and generating outliers according to the defined scenarios. The fits corresponding to the different distribution assumptions, Gaussian, Laplace, Huber, Cauchy and Student's t distributions are plotted as lines. The true trajectories, which are noise- and outlier-free, are shown as gray lines.

with x_1 and x_2 denoting the concentrations of A and B, respectively. We assumed that only x_2 was measured yielding $y = s \cdot x_2$. The parameter s was introduced as scaling parameter to reflect the case of relative data. However, in this section it was set to 1 and not estimated from the data.

We generated 10^3 artificial data sets for three outlier scenarios (Figure 3.2B):

(i) *no outliers*: no outliers were included in the data.

(ii) *one data point at zero*: the measured concentration at a certain time point t_k is zero, e.g., due to a missing label or entry. Consequently, we measured $\bar{y}_j = 0$. In practice this might not be easy to spot due to background intensity and additional noise.

(iii) *two data points interchanged*: two data points in the data set were interchanged. This could have occurred due to labeling or entry errors. In the case of several observables ($n_y > 1$), the modification was applied to all n_y observables.

The data sets were generated for the dynamic parameters $\boldsymbol{\psi} = (k_1, k_2)^T = \left(10^{\text{-}1.5}, 10^{\text{-}1.5}\right)^T$ and Gaussian distributed measurement noise with standard deviation $\sigma = 0.02$. The dynamic parameters $\boldsymbol{\psi}$ were estimated from the data along with the distribution parameters $\boldsymbol{\varphi}$. For the dynamic parameters, we used the lower bound $10^{-3.5}$ and the upper bound 10. For the scale parameters, σ, we used the lower bound 10^{-5} and the upper bound 1. For the Huber distribution, the boundaries for τ were set to 10^{-1} and 10^5. For the degree of freedom ν of the Student's t distribution, we used boundaries such that it corresponds to the Cauchy distribution for the lower bound 1 and is similar to the Gaussian distribution for the upper bound 10^5. For optimization, we employed the MATLAB routine `fmincon` interfaced by PESTO (Section 2.5). For the Gaussian, Cauchy and Student's t distributions, we used the incorporated trust-region and for the Laplace and Huber distributions the incorporated interior-point algorithm.

Evaluation of estimation results for different distribution assumptions

Parameter estimation using the assumption of Gaussian distributed measurement noise allowed for the reconstruction of the systems trajectory in the absence of outliers (Figure 3.2B). However, if there were strongly deviating outliers, the fitted and the true trajectory differed, implying estimation errors. For the heavier tailed distributions the resulted trajectories were close to the true underlying trajectory of the system.

These findings were also reflected in the MSE for the parameter estimates of the dynamic parameters (Section 2.3.4 and Figure 3.3A). If *no outliers* were present in the data, all methods yielded a comparable MSE for both dynamic parameters. In the presence of outliers, the MSE which was achieved assuming Gaussian noise was much higher. This implies that the parameter estimates differ largely from the true parameters, which will result in wrong predictions. The heavier tailed distributions were able to provide reliable estimates of the parameters in the presence of outliers. Indeed, the MSE hardly increased, indicating that the influence of a small number of outliers could be compensated. Consequently, robust estimation methods reduced the MSE for outlier-corrupted data.

Using the 10^3 data sets per scenario, we calculated the percentage of how often a distribution assumption achieved the lowest BIC. The model employing the Gaussian distribution assumption was chosen for most of the *no outliers* data sets (Figure 3.3B). In the presence of outliers, heavier tailed distributions were preferred over the Gaussian noise model and model selection detected the presence of outliers in the data sets.

We found that for this simple example the convergence was comparable and above 75% for most distributions (Figure 3.3C). Merely the optimization using the Huber distribution yielded a slightly lower fraction of converged starts.

The mean time needed per start was similarly low for the Gaussian, Cauchy and Student's t distributions (Figure 3.3D). Only the Laplace and Huber distributions had a higher computation time, since no good approximation of the Hessians based on first-order sensitivities could be found. This verified that the use of robust methods did not increase the computation time significantly.

To assess the influence of outliers on parameter CIs, we computed profile likelihoods (Section 2.3.3). Based on these profile likelihoods, the CIs were computed for different confidence levels (Figure 3.3E). We found that in the case of *no outliers*, all distribution assumptions yielded similar CIs for parameter k_1. The CIs computed using the Gaussian distribution widened in the presence of outliers, yet not ensuring that the true parameter was covered. Also for the Laplace and Huber distributions the CIs became wider, but the true parameter remained covered. For the Cauchy and Student's t distributions, we observed that the CIs became even tighter, which is counter-intuitive as the information content in the data should be decreased. The presence of outliers shifted the probability mass often closer to the mode.

We evaluated the reliability of the CIs by determining the CR, which states how often the true parameter $\boldsymbol{\psi}^{\text{true}}$ was covered by the CIs for all 10^3 generated data sets per scenario (Section 2.3.4 and Figure 3.3F). Interestingly, the CR was lower than the confidence level for most of the cases, indicating that the size of the CIs was too narrow and therefore

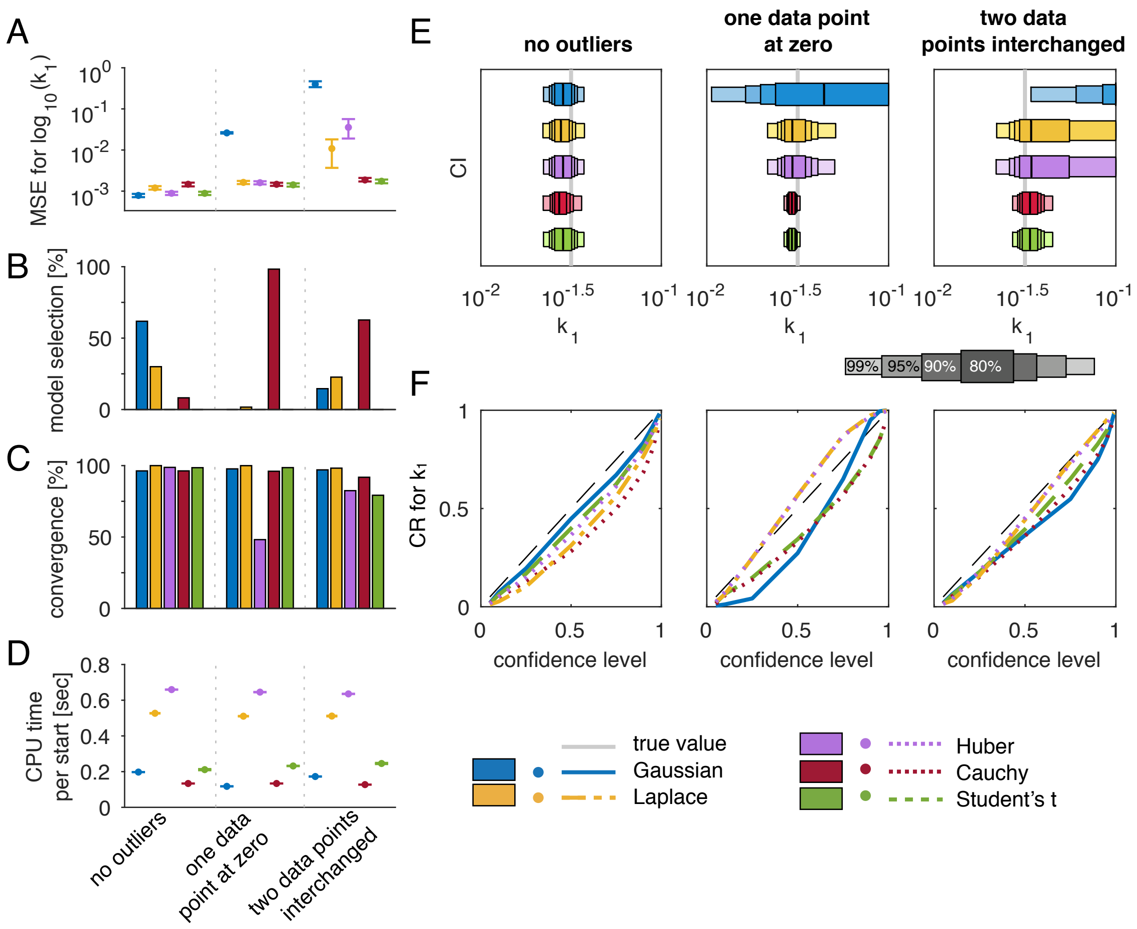

Figure 3.3: Evaluation of optimization results for the outlier scenarios. (A) MSE for the $\log_{10}$-transformed dynamic parameter k_1. The circles indicate the MSE over all 10^3 data sets per scenario, while the error bars represent the 95% percentile bootstrap CIs. (B) Model selection results using BIC. The percentage is given for how many times each statistical model is chosen for the 10^3 data sets per scenario. (C) Average percentage of converged starts over all data sets. (D) The mean computation time per optimizer start and the corresponding standard error of mean. (E) Example CIs for one data set per scenario (shown in Figure 3.2B), indicated by different bars for 80%, 90%, 95% and 99% from dark to light colors. The MLEs for the Gaussian, Laplace, Huber, Cauchy and Student's t distributions are displayed as vertical lines. The true parameter value for k_1 is displayed as vertical grey lines. (F) CRs for parameter k_1 for different confidence levels considering all 10^3 data sets per outlier scenario. Lines in the upper part of the panels indicate that the CI is too wide, lines in the lower part that it is too narrow. This figure is a modified version of Figures 3 and 4 of the author's publication (Maier et al., 2017).

the uncertainty in the parameter estimates was underrated. The Laplace and Huber distributions provided the best CR in the presence of outliers.

Sample size limitation of the Cauchy and Student's t distributions

The performance of estimators often depends strongly on the sample size. Therefore, we analyzed how different distributions perform for decreased sample sizes. To this end, we varied the number of data points ($n_t = 10, 4, 3$) for data sets of the conversion process without outliers. For a lower number of data points, the model could fit a higher percentage of the data points exactly, i.e., up to numerical accuracy. For the full data sets ($n_t = 10$), the obtained residual distributions for all combined data sets fitted the corresponding distributions (Figure 3.4A), visualized for the median scale parameters obtained with parameter estimation (Figure 3.4B). The scale parameters for the Gaussian, Laplace and Huber distributions did not become much smaller for lower numbers of data points. However, the scale parameters of the Cauchy and Student's t distributions were decreased and thus the mass of the distribution was concentrated on the exactly fitted data points. Other residuals were neglected, i.e., the model over-fitted single data points. For $n_t = 3$ these scale parameters were even estimated at the lower bound defined as 10^{-10}. Scale parameters close to zero yielded residual distributions which did not reflect the variation in the data.

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

We further used Laplace noise as a heavier tailed alternative to Gaussian noise. The Laplace distribution did not have problems with scale parameters becoming too small, like the Cauchy and Student's t distributions, and showed better convergence with less computation time than the Huber distribution.

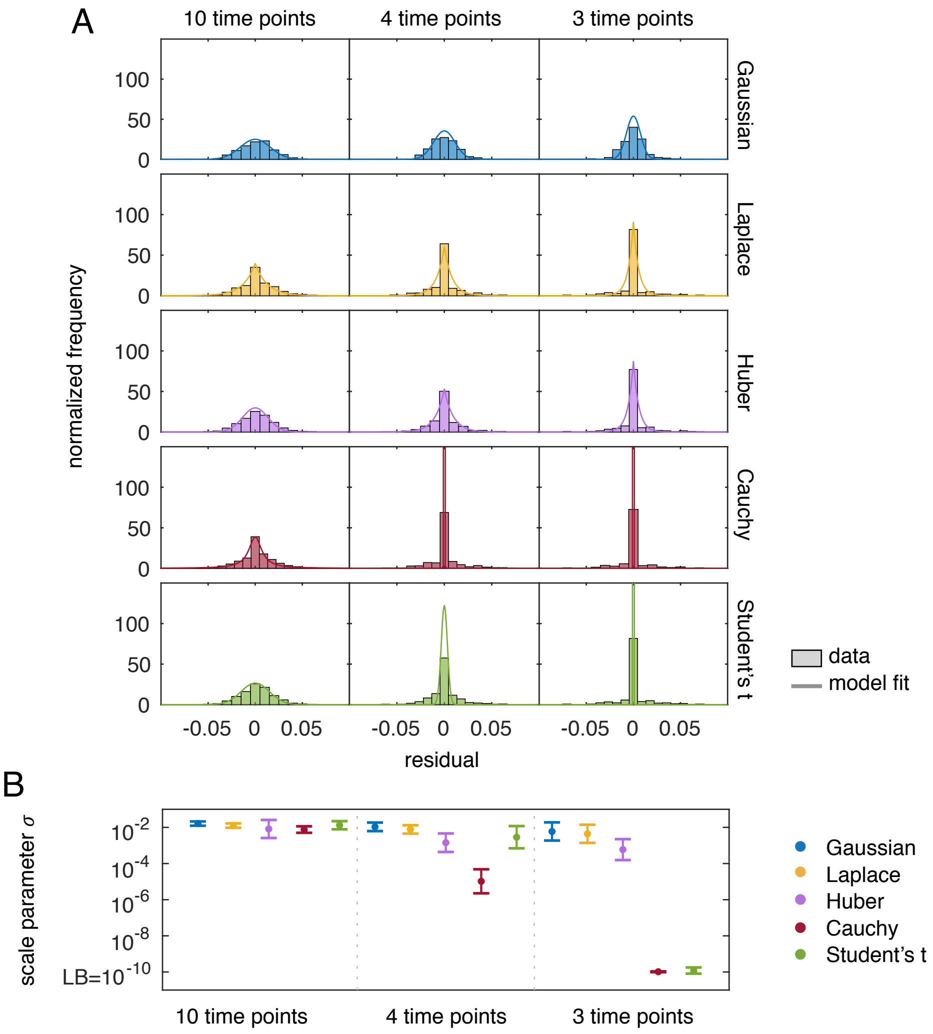

Figure 3.4: Sample size limitation of Cauchy and Student's t distribution. (A) Normalized histogram of the residuals of all 10^2 data sets when the parameter estimation is performed for $n_t = 10, 4, 3$. The curve represents the corresponding probability density of the Gaussian, Laplace, Huber, Cauchy or Student's t distribution using the estimated median value of the distribution-specific parameters over all 10^2 data sets. (B) Visualization of the corresponding scale parameters σ. The lower bound (LB) for optimization was set to 10^{-10}. This figure is a modified version of Figure 5 of the author's publication (Maier et al., 2017).

.

3.2 Hierarchical approach for calibrating ODE models on relative data

Experimental data often only provide information about the relative changes between conditions. In the literature, two methods are employed to link relative data to mathematical models: (i) evaluation of relative changes (Degasperi et al., 2017); and (ii) introduction of scaling parameters (Raue et al., 2013). In (i), relative changes between conditions are compared, and the differences between observed and simulated relative changes are minimized. While this approach is intuitive and does not alter the dimension of the fitting problem, the noise distribution is non-trivial and the residuals are not uncorrelated (Thomaseth and Radde, 2016), which is often disregarded (see, e.g., (Degasperi et al., 2017)). This can in principle result in incorrect confidence intervals. In (ii), scaling parameters are introduced to replace the calibration curves. The scaling parameters are often unknown and have to be inferred along with the remaining parameters of the model. While this increases the dimensionality of the optimization problem (see (Bachmann et al., 2011) for an example in which the number of parameters is doubled), the noise distribution is simple and the confidence intervals are consistent. To address the dimensionality increase, Weber et al. (2011) have proposed an approach for estimating the conditionally optimal scaling parameters given the dynamic parameters. This approach eliminates the scaling parameters, however, it is only applicable in the special case of additive Gaussian noise with known standard deviations. Estimating the noise parameters instead of providing the standard deviations has been shown to yield a statistically more accurate assessment of the model (Raue et al., 2013). Unknown noise parameters and outlier-corrupted data (Section 3.1) – as found in many applications – cannot be handled by the approach of Weber et al. (2011).

Here, we propose a hierarchical optimization approach for estimating the parameters for models which include scaling and noise parameters. Our approach restructures the optimization problem into an inner and outer subproblem. These subproblems possess lower dimensions than the original optimization problem, and the inner problem can often be solved analytically. Furthermore, the hierarchical approach can also be employed for optimization-based profile calculation (Section 2.3.3). We evaluate accuracy, robustness, and computational efficiency of the hierarchical approach by studying three signaling pathways.

The scaling parameters are usually incorporated in h, which is defined in (2.5). In this section, we factor-out the scaling parameters from the observable function h and write

$$y_j(\boldsymbol{\theta}) = s_j \cdot h_j(\boldsymbol{\psi}) \,. \tag{3.9}$$

Often, the scaling and noise parameters are assumed to not vary over time and thus the corresponding measured data points share these parameters. Also, these parameter might be shared across observables or experiments. We define $I_{s,1}, \dots, I_{s,n_s}$ as the index sets for the measured data points which share the same scaling parameter, and $I_{\sigma,1}, \dots, I_{\sigma,n_\sigma}$ as index sets for the measured data points which share the same noise parameter. This means that $s_j = s_{j_s}, \forall j \in I_{s,j_s}$ and $\sigma_j = \sigma_{j_\sigma}, \forall j \in I_{\sigma,j_\sigma}$.

For the standard approach, the dynamic parameters $\boldsymbol{\psi}$, the scaling parameters $\mathbf{s}$ and the noise parameters $\boldsymbol{\sigma}$ are estimated simultaneously. This optimization problem has dimension equal to number of dynamic parameters n_ψ + number of scaling parameters n_s + number of noise parameters n_σ. As described in Section 2.3, this optimization problem can be solved using multi-start local optimization. For each iteration, the objective function and its gradient are computed. If the objective function for these parameters fulfills certain criteria, e.g., the norm of the gradient is below a certain threshold, the optimization is stopped, otherwise the parameter is updated and the procedure is continued (Figure 3.5A).

Since the optimization problem (2.19) often possess a large number of optimization variables and can be difficult to solve, we exploited its structure. Instead of solving simultaneously for $\boldsymbol{\psi}$, $\mathbf{s}$, and $\boldsymbol{\sigma}$, we considered the hierarchical optimization problem (Figure 3.5B-D)

$$\min_{\boldsymbol{\psi}} \; J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi})) \tag{3.10}$$

$$\text{with} \quad (\hat{\mathbf{s}}, \hat{\boldsymbol{\sigma}}) = \operatorname*{argmin}_{\mathbf{s},\boldsymbol{\sigma}} \; J(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\sigma}) \,. \tag{3.11}$$

The inner problem (3.11) provides the optimal values $\hat{\mathbf{s}}(\boldsymbol{\psi})$ and $\hat{\boldsymbol{\sigma}}(\boldsymbol{\psi})$ of $\mathbf{s}$ and $\boldsymbol{\sigma}$ given $\boldsymbol{\psi}$. These optimal values are used in the outer subproblem to determine the optimal value for $\boldsymbol{\psi}$ denoted by $\hat{\boldsymbol{\psi}}$. It is apparent that a locally optimal point of the hierarchical optimization problem (3.10, 3.11) is also locally optimal for the standard optimization problem (2.19), given the same parameters boundaries for both problems. However, starting at the same initial values, the standard and hierarchical optimization might converge to different local optima, since the basin of attraction potentially differs.

The formulation (3.10) may appear more involved, however, it possesses several properties which can be advantageous:

(i) The individual dimensions of the inner and outer subproblems (3.10, 3.11) are lower than the dimension of the original problem (2.19).

(ii) The optimization of the inner subproblem does not require the repeated numerical simulation of the ODE model.

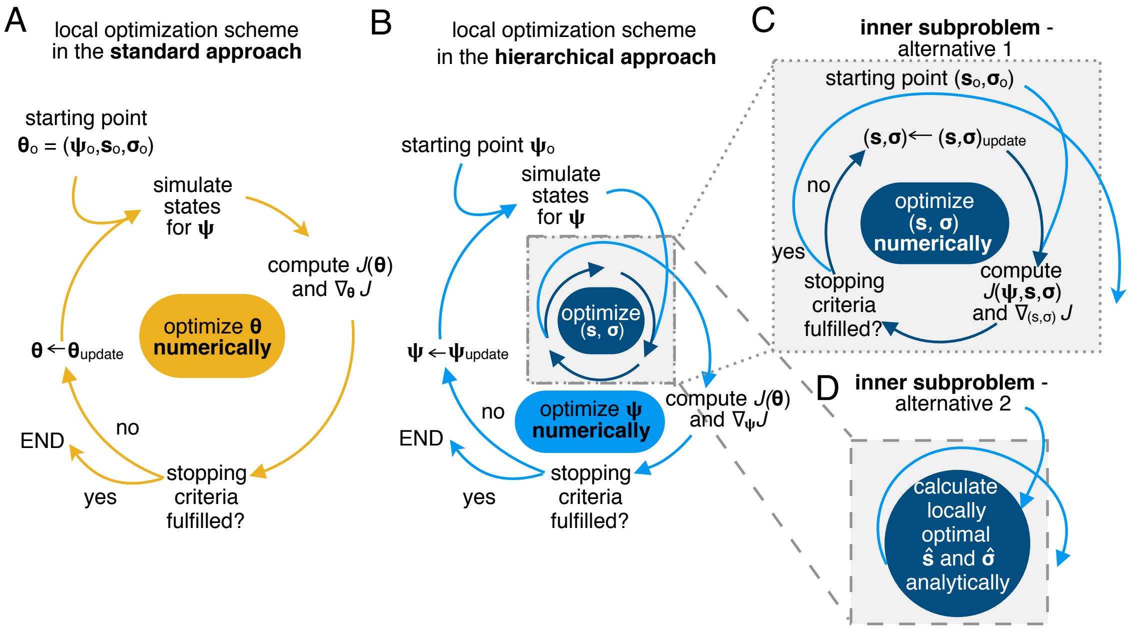
A
local optimization scheme
in the **standard approach**
starting point
$\boldsymbol{\theta}_o = (\boldsymbol{\psi}_o, \mathbf{s}_o, \boldsymbol{\sigma}_o)$
simulate
states
for $\boldsymbol{\psi}$
compute $J(\boldsymbol{\theta})$
and $\nabla_{\boldsymbol{\theta}} J$
optimize $\boldsymbol{\theta}$
numerically
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_{\text{update}}$
no
END
yes
stopping
criteria
fulfilled?
B
local optimization scheme
in the **hierarchical approach**
starting point $\boldsymbol{\psi}_o$
simulate
states
for $\boldsymbol{\psi}$
optimize
$(\mathbf{s}, \boldsymbol{\sigma})$
$\boldsymbol{\psi} \leftarrow \boldsymbol{\psi}_{\text{update}}$
no
optimize $\boldsymbol{\psi}$
numerically
compute $J(\boldsymbol{\theta})$
and $\nabla_{\boldsymbol{\psi}} J$
END
yes
stopping
criteria
fulfilled?
C
inner subproblem -
alternative 1
starting point $(\mathbf{s}_o, \boldsymbol{\sigma}_o)$
$(\mathbf{s}, \boldsymbol{\sigma}) \leftarrow (\mathbf{s}, \boldsymbol{\sigma})_{\text{update}}$
no
optimize
$(\mathbf{s}, \boldsymbol{\sigma})$
numerically
yes
stopping
criteria
fulfilled?
compute
$J(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\sigma})$
and $\nabla_{(\mathbf{s}, \boldsymbol{\sigma})} J$
D
inner subproblem -
alternative 2
calculate
locally
optimal
$\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$
analytically

Figure 3.5: Visualization of standard and hierarchical optimization schemes. (A) Local optimization in the standard approach with parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\sigma})$. (B) Outer local optimization in the hierarchical approach with parameters $\boldsymbol{\psi}$. (C,D) Inner (local) optimization in the hierarchical approach to find the optimal scaling and noise parameters, $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$, for given dynamic parameters $\boldsymbol{\psi}$. (C) Iterative local optimization to determine $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$. This does not require the numerical simulation of the model. (D) Calculating optimal parameters $\hat{\mathbf{s}}$ and $\hat{\boldsymbol{\sigma}}$ using analytic expressions for common noise distributions. This figure is a modified version of Figure 1 of the author's publication (Loos et al., 2018a).

(iii) For several noise models, e.g., Gaussian (3.2) and Laplace noise (3.3), the inner subproblem can be solved analytically.

If an analytical solution for the inner subproblem is available, the scaling parameters $\mathbf{s}$ and also the noise parameters $\boldsymbol{\sigma}$ can be calculated directly, and the number of parameters that need to be optimized iteratively reduces to n_ψ, which corresponds to alternative 2 in Figure 3.5D. In the following, the analytic expressions for scaling and noise parameters under Gaussian and Laplace noise are derived.

3.2.1 Analytical expressions for scaling and noise parameters for Gaussian and Laplace noise

The scaling and noise parameters for **Gaussian noise** were computed analytically. To derive the analytic expression for the optimal parameters, we exploited that the negative log-likelihood function for Gaussian noise,

$$J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma}) = \frac{1}{2} \sum_j \left[\log(2\pi\sigma_j^2) + \left(\frac{\bar{y}_j - s_j \cdot h_j(\boldsymbol{\psi})}{\sigma_j} \right)^2 \right], \tag{3.12}$$

is continuously differentiable, and that the gradient of J at a local minimum is zero. For the inner subproblem this implies

$$\nabla_s J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})\big|_{\hat{\mathrm{s}},\hat{\boldsymbol{\sigma}}} = \mathbf{0} \quad \text{and} \quad \nabla_\sigma J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})\big|_{\hat{\mathrm{s}},\hat{\boldsymbol{\sigma}}} = \mathbf{0}\,. \tag{3.13}$$

The derivative of the negative log-likelihood function with respect to a scaling parameter s_{j_s} reads

$$\frac{\partial J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})}{\partial s_{j_s}} = \frac{1}{2} \sum_{j \in I_{s,j_s}} \frac{2}{\sigma_j^2} \left(\bar{y}_j - s_{j_s} \cdot h_j(\boldsymbol{\psi})\right) \cdot \left(-h_j(\boldsymbol{\psi})\right) \overset{!}{=} 0\,,$$

which was set to zero to obtain the analytic expression for the optimal scaling parameter $\hat{s}_{j_s}$. If $\exists j \in I_{s,j_s} : h_j(\boldsymbol{\psi}) \neq 0$, the optimal scaling parameter is

$$\hat{s}_{j_s}(\boldsymbol{\psi}) = \frac{\sum_{j \in I_{s,j_s}} \bar{y}_j \cdot h_j(\boldsymbol{\psi}) \cdot \frac{1}{\sigma_j^2}}{\sum_{j \in I_{s,j_s}} h_j(\boldsymbol{\psi})^2 \cdot \frac{1}{\sigma_j^2}}\,. \tag{3.14}$$

It does not depend on the noise parameter if the following condition is fulfilled:

$$\forall j_s \; \exists j_\sigma : \; I_{s,j_s} \subset I_{\sigma,j_\sigma}\,. \tag{3.15}$$

This means that all observations which share a scaling parameters also need to share the noise parameters. This is a plausible restriction and should not pose a problem. It is

$$\frac{\partial^2 J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})}{\partial s_{j_s}^2} = \sum_{j \in I_{s,j_s}} \left(\frac{h_j(\boldsymbol{\psi})}{\sigma_j} \right)^2 > 0\,,$$

which yields that $\hat{s}_{j_s}(\boldsymbol{\psi})$ is the unique optimal scaling parameter which minimizes (3.12) for a given set of dynamic parameters $\boldsymbol{\psi}$. If $\forall j \in I_{s,j_s} : h_j(\boldsymbol{\psi}) = 0$, the scaling parameter does not have an effect on the objective function.

Given the optimal scaling parameter at each observation, $\hat{s}_j = \hat{s}_{j_s}$ if $j \in I_{s,j_s}$, it needs to hold for the noise parameter:

$$\begin{aligned}\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma}(\boldsymbol{\psi}))}{\partial \sigma_{j_\sigma}} &= \frac{1}{\sigma_{j_\sigma}} \sum_{j \in I_{\sigma,j_\sigma}} \left[1 - \left(\frac{\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})}{\sigma_{j_\sigma}} \right)^2 \right] \overset{!}{=} 0 \\ &\Rightarrow \hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) = \frac{1}{|I_{\sigma,j_\sigma}|} \sum_{j \in I_{\sigma,j_\sigma}} \left(\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}) \right)^2 , \end{aligned} \tag{3.16}$$

with $|I_{\sigma,j_\sigma}|$ denoting the cardinality of the index set. Since

$$\begin{aligned}\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \sigma^2_{j_\sigma}} &= \sum_{j \in I_{\sigma,j_\sigma}} \frac{3 \left(\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}) \right)^2 - \hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi})}{\hat{\sigma}^4_{j_\sigma}(\boldsymbol{\psi})} \\ &= \frac{1}{\hat{\sigma}^4_{j_\sigma}(\boldsymbol{\psi})} \left(3 \underbrace{\sum_{j \in I_{\sigma,j_\sigma}} \left[(\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}))^2 \right]}_{\hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) \cdot |I_{\sigma,j_\sigma}|} - \hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) \cdot |I_{\sigma,j_\sigma}| \right) \\ &= \frac{1}{\hat{\sigma}^4_{j_\sigma}(\boldsymbol{\psi})} \left(|I_{\sigma,j_\sigma}| \cdot \left(3\hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) - \hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) \right) \right) \\ &= \frac{2|I_{\sigma,j_\sigma}|}{\hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi})} > 0, \end{aligned} \tag{3.17}$$

the noise parameter $\hat{\sigma}^2_{j_\sigma}$ is the unique parameter minimizing (3.12). Consistent with the structure of the hierarchical problem (3.10), both equations for the optimal parameters (3.14, 3.16) depend only on the dynamic parameters $\boldsymbol{\psi}$.

The gradient used for the outer optimization (3.10) for $i = 1, \ldots, n_\psi$ is given by

$$\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i} = -\sum_j \frac{\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})}{\hat{\sigma}^2_j(\boldsymbol{\psi})} \cdot \hat{s}_j(\boldsymbol{\psi}) \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} ,$$

for which $\partial h_j(\boldsymbol{\psi}) / \partial \psi_i$ is obtained by forward sensitivity equations. The Hessian with respect to the dynamic parameters for $i, l = 1, \ldots, n_\psi$ is

$$\begin{aligned}\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i \partial \psi_l} = \sum_j \Bigg[&\left(\frac{\hat{s}_j(\boldsymbol{\psi})}{\hat{\sigma}_j(\boldsymbol{\psi})} \right)^2 \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_l} - \\ &\underbrace{\frac{\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})}{\hat{\sigma}^2_j(\boldsymbol{\psi})} \cdot \hat{s}_j(\boldsymbol{\psi}) \cdot \frac{\partial^2 h_j(\boldsymbol{\psi})}{\partial \psi_i \partial \psi_l}}_{(*)} \Bigg] . \end{aligned}$$

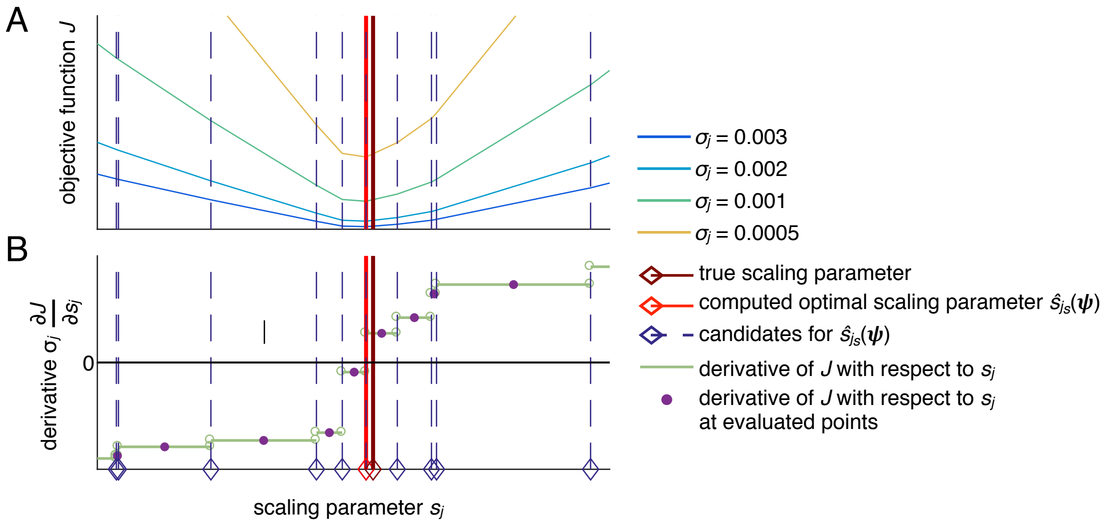

Figure 3.6: Illustration of the computation of an optimal scaling parameter $\hat{s}_{j_s}(\boldsymbol{\psi})$ for Laplace noise. (A) Objective function J for fixed $\boldsymbol{\psi}$ and different values of σ_j, showing that the kinks, i.e., the points of non-differentiability, which are indicated by the dashed lines, are independent of σ_j. (B) Derivative of the objective function with respect to the scaling parameter. The derivative is not defined at the kinks. The light red and dark red lines indicate the computed scaling parameter and the true optimal scaling parameter, respectively. This figure and its legend is modified from Figure 2 of the author's publication (Loos et al., 2018a).

We implemented an approximation of the Hessian neglecting the terms $(*)$ that include second-order sensitivities.

For **Laplace noise** the negative log-likelihood function is

$$J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma}) = \sum_j \left[\log(2\sigma_j) + \frac{|\bar{y}_j - s_j \cdot h_j(\boldsymbol{\psi})|}{\sigma_j} \right]. \tag{3.18}$$

This function is continuous, but not continuously differentiable. In this case, a sufficient condition for a local minimum is that the right limit value of the derivative is negative and the left limit value is positive. The derivative of (3.18) with respect to s_{j_s} can be written as

$$\frac{\partial J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})}{\partial s_{j_s}} = - \sum_{j \in I_{s,j_s}} \frac{1}{\sigma_j} \left(|h_j(\boldsymbol{\psi})| \cdot \operatorname{sgn}\left(\frac{\bar{y}_j}{h_j(\boldsymbol{\psi})} - s_{j_s} \right) \right).$$

As σ_j is positive and, if (3.15) holds, it is the same $\forall j \in I_{s,j_s}$, the locations of kinks in the objective function and the corresponding jumps in the derivative are independent of

σ_j (Figure 3.6). Accordingly, the problem of finding $\hat{s}_{j_s}(\boldsymbol{\psi})$ reduces to checking the signs of the derivative before and after the jump points

$$s_j(\boldsymbol{\psi}) = \frac{\bar{y}_j}{h_j(\boldsymbol{\psi})}, j \in I_{s,j_s}. \tag{3.19}$$

We sorted $s_j(\boldsymbol{\psi})$ in increasing order and evaluated the derivatives at the midpoints between adjacent jumps, a procedure which is highly efficient as the ODE model does not have to be simulated.

To obtain the optimal noise parameter, we require

$$\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma})}{\partial \sigma_{j_\sigma}} = \frac{1}{\sigma_{j_\sigma}} \sum_{j \in I_{\sigma,j_\sigma}} \left(1 - \frac{|\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})|}{\sigma_{j_\sigma}}\right) \overset{!}{=} 0. \tag{3.20}$$

This yields

$$\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi}) = |I_{\sigma,j_\sigma}|^{-1} \sum_{j \in I_{\sigma,j_\sigma}} |\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})|\,. \tag{3.21}$$

With

$$\begin{aligned}
\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \sigma_{j_\sigma}^2} &= \frac{1}{\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi})^2} \sum_{j \in I_{\sigma,j_\sigma}} \left(2 \cdot \frac{|\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})|}{\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi})} - 1\right) \\
&= \frac{1}{\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi})^2} \left(-|I_{\sigma,j_\sigma}| + \frac{2}{\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi})} \sum_{j \in I_{\sigma,j_\sigma}} |\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})|\right) \\
&= \frac{|I_{\sigma,j_\sigma}|}{\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi})^2} > 0\,,
\end{aligned}$$

we obtained that (3.21) minimizes (3.18).

The gradient used for optimization of the outer subproblem for $i = 1, \ldots, n_\psi$ is given by

$$\begin{aligned}
\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i} = -\sum_j \Bigg[& \frac{\operatorname{sgn}(\bar{y}_j - \hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}))}{\hat{\sigma}_j(\boldsymbol{\psi})} \cdot \hat{s}_j(\boldsymbol{\psi}) \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} + \\
& \underbrace{\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \hat{s}_j(\boldsymbol{\psi})}}_{(*)} \frac{\partial \hat{s}_j(\boldsymbol{\psi})}{\partial \psi_i} \Bigg],
\end{aligned}$$

for which $\partial h_j(\boldsymbol{\psi})/\partial \psi_i$ is obtained by forward sensitivity equations and the term $(*)$ is defined as the right limit value of the derivative.

3.2.2 Analytical expressions for log-transformed observables

In many studies (e.g., Bachmann et al. (2011)), observation functions of the form $\log(\bar{y}_j) = \log(s_j \cdot h_j(\boldsymbol{\psi})) + \epsilon_j$ are used. In the following, we provide the analytical expressions for the optimal values of the scaling and measurement noise parameters for log-transformed observables and measurement data.

For **Gaussian noise**, the objective function for the comparison on the logarithmic scale is given by

$$J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma}) = \frac{1}{2} \sum_j \left[\log(2\pi {\sigma_j}^2) + \left(\frac{\log(\bar{y}_j) - \log\left(s_j \cdot h_j(\boldsymbol{\psi})\right)}{\sigma_j} \right)^2 \right]. \tag{3.22}$$

Thus, the derivative with respect to the scaling parameters is

$$\frac{\partial J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma})}{\partial s_{j_s}} = \sum_{j \in I_{s,j_s}} \frac{\log(s_{j_s}) + \log(h_j(\boldsymbol{\psi})) - \log(\bar{y}_j)}{s_{j_s} \cdot {\sigma_j}^2}.$$

This yields the equation for the optimal scaling parameters

$$\hat{s}_{j_s}(\boldsymbol{\psi}) = \exp\left(\frac{\sum_{j \in I_{s,j_s}} \frac{\log(\bar{y}_j) - \log(h_j(\boldsymbol{\psi}))}{\sigma_j^2}}{|I_{s,j_s}|} \right), \tag{3.23}$$

which is independent of the noise parameter if (3.15) is fulfilled. The second-order derivative is

$$\begin{aligned} \frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma}}{\partial s_{j_s}^2} &= \frac{1}{\sigma_j^2 \cdot \hat{s}_{j_s}(\boldsymbol{\psi})^2} \sum_{j \in I_{s,j_s}} \Big[1 + \log(\bar{y}_j) - \log(h_j(\boldsymbol{\psi})) - \log(\hat{s}_{j_s}(\boldsymbol{\psi})) \Big] \\ &\Rightarrow |I_{s,j_s}| - |I_{s,j_s}| \cdot \log(\hat{s}_{j_s}(\boldsymbol{\psi})) + \sum_{j \in I_{s,j_s}} \Big[\log(\bar{y}_j) - \log(h_j(\boldsymbol{\psi})) \Big] \\ &= |I_{s,j_s}| > 0\,, \end{aligned}$$

and thus (3.23) minimizes (3.22).

Using

$$\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma})}{\partial \sigma_{j_\sigma}} = \frac{1}{\sigma_{j_\sigma}} \cdot \sum_{j \in I_{\sigma,j_\sigma}} \left(1 - \frac{\left(\log(\bar{y}_j) - \log\left(\hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi})\right)\right)^2}{\sigma_{j_\sigma}^2} \right),$$

we obtained the optimal noise parameter

$$\hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi}) = |I_{\sigma,j_\sigma}|^{-1} \sum_{j\in I_{\sigma,j_\sigma}} \left(\log\left(\bar{y}_j\right) - \log\left(\hat{s}_j\left(\boldsymbol{\psi}\right)\cdot h_j(\boldsymbol{\psi})\right)\right)^2 . \tag{3.24}$$

Similarly as in (3.17)

$$\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \sigma^2_{j_\sigma}} = \frac{2|I_{\sigma,j_\sigma}|}{\hat{\sigma}^2_{j_\sigma}(\boldsymbol{\psi})} > 0 .$$

The gradient used for the outer optimization problem for $i = 1, \ldots, n_\psi$ is given by

$$\frac{\partial J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i} = -\sum_j \frac{\log(\bar{y}_j) - \log(\hat{s}_j(\boldsymbol{\psi})\cdot h_j(\boldsymbol{\psi}))}{\hat{\sigma}^2_j(\boldsymbol{\psi})} \cdot \frac{1}{h_j(\boldsymbol{\psi})} \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} .$$

The Hessian with respect to the dynamic parameters for $i, l = 1, \ldots, n_\psi$ is

$$\begin{aligned}\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathrm{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i \partial \psi_l} = \sum_j \Bigg[\Bigg(& \frac{1 + \log(\bar{y}_j) - \log(\hat{s}_j(\boldsymbol{\psi})\cdot h_j(\boldsymbol{\psi}))}{\hat{\sigma}^2_j(\boldsymbol{\psi})\cdot h_j(\boldsymbol{\psi})^2} \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_l} \\ & - \underbrace{\frac{\left(\log(\bar{y}_j) - \log\left(\hat{s}_j\left(\boldsymbol{\psi}\right)\cdot h_j(\boldsymbol{\psi})\right)\right)}{\hat{\sigma}^2_j(\boldsymbol{\psi})\cdot h_j(\boldsymbol{\psi})} \cdot \frac{\partial^2 h_j(\boldsymbol{\psi})}{\partial \psi_i \partial \psi_l}}_{(*)} \Bigg) \Bigg] .\end{aligned}$$

An approximation can again be obtained by neglecting the terms (∗) that include second-order sensitivities.

If the data and simulation are compared on a $\log_{10}$ scale, the optimal scaling parameters are the same as when using the natural logarithm (3.23). For the optimal noise parameters the natural logarithm is replaced by $\log_{10}$ in (3.24).

For the **Laplace noise** including the logarithmic comparison,

$$J(\boldsymbol{\psi}, \mathrm{s}, \boldsymbol{\sigma}) = \sum_j \left[\log(2\sigma_j) + \frac{|\log(\bar{y}_j) - \log(s_j \cdot h_j(\boldsymbol{\psi})|}{\sigma_j} \right] ,$$

the same procedure can be applied for the logarithmic scale as for the linear scale, with the same set of candidate scaling parameters (3.19) as for the linear scale. However, one has to pay attention to adapt the derivative properly, for which the change of sign is checked.

For the calculation of the optimal noise parameter, we used

$$\frac{\partial J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma})}{\partial \sigma_{j_\sigma}} = \frac{1}{\sigma_{j_\sigma}} \sum_{j \in I_{\sigma,j_\sigma}} \left(1 - \frac{|\log(\bar{y}_j) - \log(\hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}))|}{\sigma_{j_\sigma}}\right),$$

and obtained

$$\hat{\sigma}_{j_\sigma}(\boldsymbol{\psi}) = |I_{\sigma,j_\sigma}|^{-1} \sum_{j \in I_{\sigma,j_\sigma}} |\log(\bar{y}_j) - \log(\hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}))|\,.$$

It also holds that

$$\frac{\partial^2 J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \boldsymbol{\sigma}(\boldsymbol{\psi}))}{\partial \sigma_{j_\sigma}^2} = \frac{|I_{\sigma,j_\sigma}|}{\hat{\sigma}_{j_\sigma}^2(\boldsymbol{\psi})} > 0\,.$$

The gradient used for the outer optimization for $i = 1, \dots, n_\psi$ is given by

$$\begin{aligned}\frac{\partial J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \psi_i} = -\sum_j \Bigg[& \frac{\operatorname{sgn}\left(\log(\bar{y}_j) - \log(\hat{s}_j(\boldsymbol{\psi}) \cdot h_j(\boldsymbol{\psi}))\right)}{\hat{\sigma}_j(\boldsymbol{\psi})} \cdot \frac{1}{h_j(\boldsymbol{\psi})} \cdot \frac{\partial h_j(\boldsymbol{\psi})}{\partial \psi_i} + \\ & \underbrace{\frac{\partial J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))}{\partial \hat{s}_j(\boldsymbol{\psi})}}_{(*)} \frac{\partial \hat{s}_j(\boldsymbol{\psi})}{\partial \psi_i} \Bigg],\end{aligned}$$

for which $\partial h_j(\boldsymbol{\psi})/\partial \psi_i$ is obtained by forward sensitivity equations and the term $(*)$ is defined as the right limit value of the derivative.

In summary, we reformulated the original optimization problem (2.19) as a hierarchical optimization problem (3.10, 3.11), and provided an analytic solution to the inner subproblem (3.11) for several relevant cases. Using the analytic solutions, the dynamic parameters can be inferred by solving a lower-dimensional optimization problem.

3.2.3 Simulation example: Conversion reaction

We compared the standard and hierarchical approach for optimization for the example of a conversion reaction (see (3.8) in Section 3.1.2) in the case of no outliers. Here, we assumed that both states could be measured up to some proportionality constant, yielding $\mathbf{y} = (y_1, y_2) = (s_1 x_1, s_2 x_2)$ (Figure 3.7A). Both observables had different noise levels $\sigma_1 = 0.02, \sigma_2 = 0.01$. This yielded six parameters for the standard optimization and two parameters for the hierarchical optimization (Figure 3.7B). The data was generated with Gaussian noise for the same parameter values as in Section 3.1.2 and with $s_1 = s_2 = 1$.

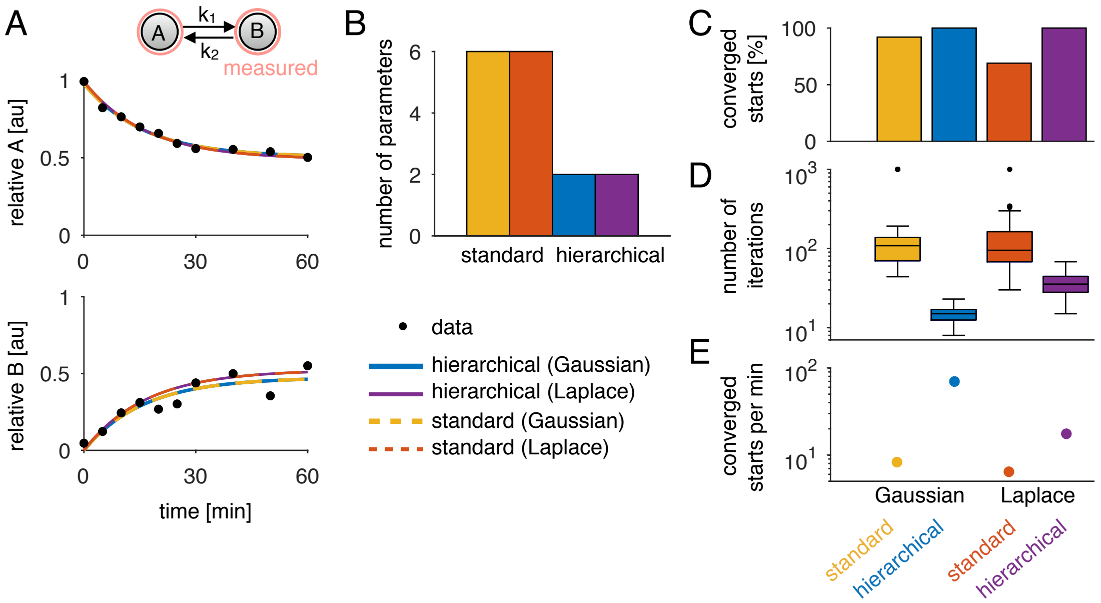

Figure 3.7: Comparison of the standard and hierarchical approach for optimization for the conversion reaction. (A) Simulated data and fitted trajectories employing Gaussian and Laplace noise. (B) Number of optimization variables in the outer subproblem. (C) Percentage of converged optimization starts. (D) Number of iterations for the optimization starts. (E) Converged starts per minute.

In this section, we employed the interior-point algorithm of `fmincon` for both distribution assumptions and optimization approaches. Estimating both noise distributions with the standard and the hierarchical approach for optimization showed that for this simple example, both approach achieved a high convergence rate (Figure 3.7C). The hierarchical approach converged for all optimization starts to the same optimal objective function value. To do so, it required less iterations (Figure 3.7D) and overall outperformed the standard approach in terms of optimization performance, which was measured as converged starts per minute (Figure 3.7E).

3.3 Evaluation of the methods for three signaling models

To study the proposed methods, the heavier tailed Laplace distribution (Section 3.1) and the hierarchical approach for optimization (Section 3.2), we calibrated mathematical models of signaling on published experimental data. In the following, we give a brief introduction to these models of the JAK-STAT (Bachmann et al., 2011; Swameye et al., 2003) and the RAF/MEK/ERK signaling pathway (Fiedler et al., 2016).

As first application example, referred to as **JAK-STAT signaling I**, we considered the model of Epo-induced JAK-STAT signaling introduced by Swameye et al. (2003) (Figure 3.8A). Epo yields the phosphorylation of signal transducer and activator of transcription 5 (STAT5), which dimerizes, enters the nucleus to trigger the transcription of target genes, gets dephosphorylated and is transported to the cytoplasm. We implemented the model which describes the phosphorylated Epo receptor concentration as a time-dependent spline (Schelker et al., 2012). The model parameters were estimated using immunoblotting data for the phosphorylated Epo receptor (pEpoR), phosphorylated STAT5 (pSTAT5) and the total amount of STAT5 in the cytoplasm (tSTAT5) (Figure 3.8B). In total 46 data points were available for 16 different time points. Since immunoblotting only provides relative data, the scaling parameters for the observables needed to be estimated from the data. As proposed by Schelker et al. (2012), the scaling parameter for pEpoR was fixed to avoid structural non-identifiabilities (Raue et al., 2009). The model with the reduced parameter vector is structurally identifiable. This yielded in total 16 parameters, which comprise $n_\psi = 11$ dynamic parameters, $n_s = 2$ scaling parameters and $n_\sigma = 3$ noise parameters.

The second application example is the model of **JAK-STAT signaling II** introduced by Bachmann et al. (2011). This model provides more details compared to the previous one. It includes, for instance, gene expression of cytokine-inducible SH2-containing protein (CIS) and suppressor of cytokine signaling 3 (SOCS3), and possesses more state variables and parameters (Figure 3.8C). The model parameters were estimated using 541 data points collected by immunoblotting, qRT-PCR and quantitative mass spectrometry (Figure 3.8D). To model the observables, Bachmann et al. (2011) used $n_s = 43$ scaling parameters, $n_\sigma = 11$ noise parameters and $n_\psi = 58$ dynamic parameters yielding in total 112 parameters. Some scaling and noise parameters were shared between experiments and some were shared between observables. For this model, most of the observables were compared at the $\log_{10}$ scale.

The third application example we considered is the model of **RAF/MEK/ERK signaling** introduced by Fiedler et al. (2016). The model describes the phosphorylation cascade and a negative feedback of phosphorylated ERK on RAF phosphorylation (Figure 3.8). Fiedler et al. (2016) collected Western blot data for HeLa cells for two observables, phosphorylated MEK, and phosphorylated ERK, with four replicates at seven time points giving 72 data points (Figure 3.8F,G). Each observable and replicate was assumed to have different scaling and noise parameters, yielding 16 additional parameters and in total 28 parameters in the standard approach compared to $n_\psi = 12$ in the hierarchical approach.

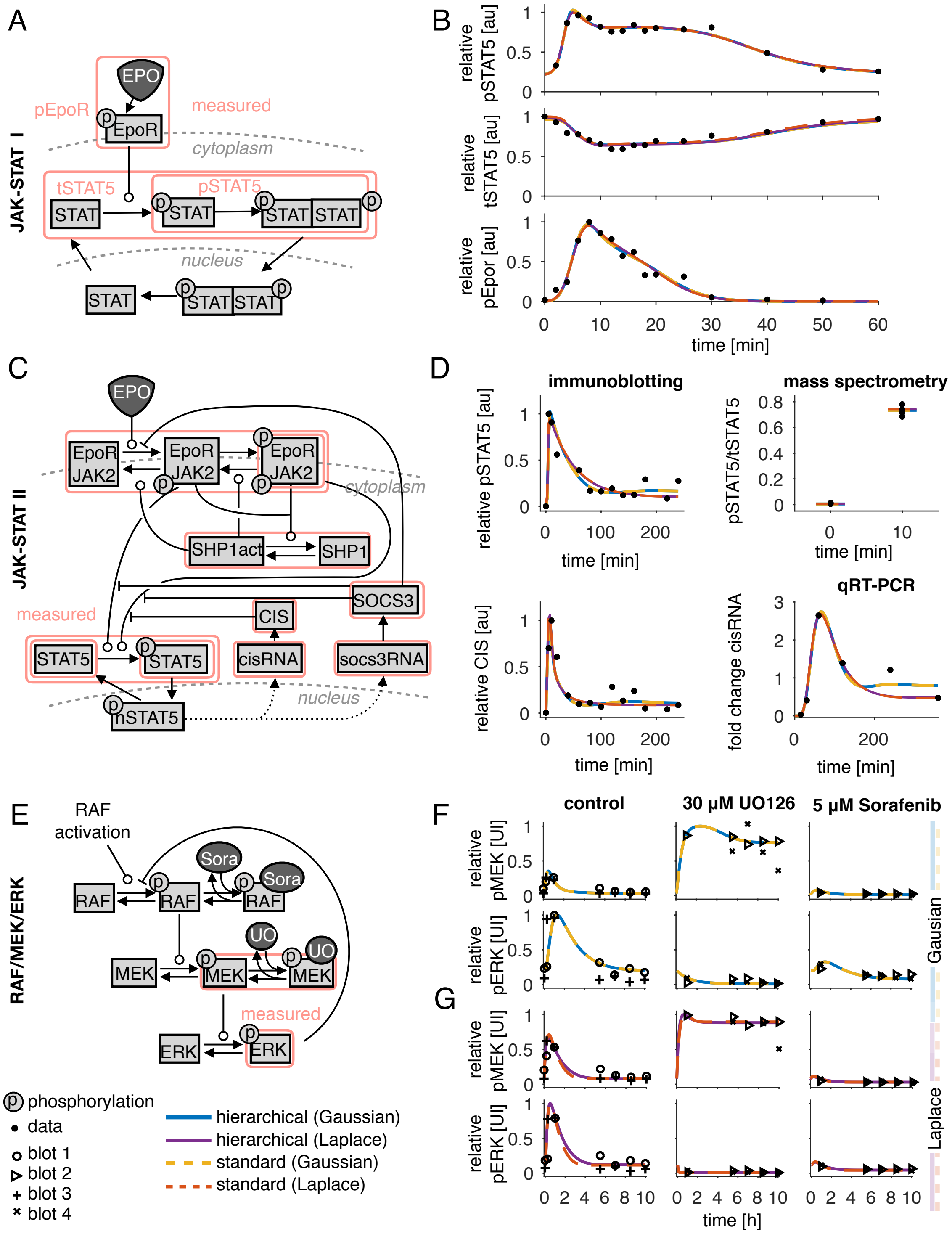

Figure 3.8: Models and experimental data. Figure caption on next page.

Figure 3.8: Models and experimental data. (A,B) JAK-STAT signaling I. (A) Illustration of the model according to Swameye et al. (2003). Arrows represent biochemical reactions, and the observables of the model used are highlighted by boxes. (B) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization with 100 starts. The results are shown for the standard (dotted lines) and hierarchical (solid lines) approach for optimization for Gaussian and Laplace noise. (C,D) JAK-STAT signaling II. (C) Illustration of the model according to Bachmann et al. (2011). (D) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization for 100 starts. 33 out of 541 data points are shown. (E-G) RAF/MEK/ERK signaling. (E) Illustration of the model according to Fiedler et al. (2016). (F,G) Experimental data and fitted trajectories for the best parameter found with multi-start local optimization for 500 and 1000 starts for Gaussian and Laplace noise, respectively. Different markers indicate the different blots. The data is scaled according to the estimated scaling parameters, yielding different visualizations for different parameters, as obtained with the Gaussian and the Laplace noise assumption. (F) Fitted trajectories for Gaussian noise for the standard (dotted line) and hierarchical (solid line) approach for optimization. (G) Fitted trajectories for Laplace noise. This figure and its legend is a modified version of Figure 3 of the author's publication (Loos et al., 2018a).

3.3.1 Comparison of standard and hierarchical approach for Gaussian and Laplace measurement noise

We performed parameter estimation for the application examples using the standard and the hierarchical approach. For each example, the case of Gaussian and Laplace noise was considered.

The trajectories of the optimizer for the dynamic parameters differed between the standard and the hierarchical approach (Figure 3.9). For the example shown in Figure 3.9, the hierarchical approach needed less than half of the iterations as the standard approach. The standard approach needed 50 iterations to reach the same objective function value the hierarchical approach reached after already two iterations. Both approaches, however, yielded the same dynamic, scaling and noise parameter values. This was true for these particular runs, but more generally also for multi-start optimization and other global optimization methods under the assumption of parameter identifiability.

As the standard and hierarchical approaches should in principle be able to achieve the same fit, we first studied the agreement of trajectories for the optimal parameters. We found that they coincide for the JAK-STAT models I and II, for both noise distributions, and the RAF/MEK/ERK for Gaussian noise. This indicated that the hierarchical approach was able to find the same optimal likelihood value as the standard approach. Also the best likelihood values which were found by the two approaches were the same. For the

RAF/MEK/ERK model with the assumption of Laplace distributed measurement noise, the fitted trajectories between the experimental data slightly deviated (Figure 3.8F), which could be explained by convergence issues and broad confidence intervals of the parameters (Figure 3.11C,D). As expected, there were differences between the results obtained with Gaussian and Laplace noise, which is visible in the fitted trajectories and the corresponding likelihood values. For all models, the Laplace noise yielded a better BIC, however, only for JAK-STAT model II and RAF/MEK/ERK the difference was substantial (Figure 3.10B).

The application examples varied with respect to the total number of parameters and the number of parameters which correspond to scaling or noise parameters (Figure 3.10A). While for the JAK-STAT model I only five out of 16 parameters could be optimized analytically, for the JAK-STAT model II almost half of the parameters correspond to scaling or noise parameters. Interestingly, even when the dimension of the outer optimization problem was only reduced by few parameters by solving the inner problem analytically, we observed a substantial increase of the percentage of converged multi-starts (Figure 3.10C). We found that the proposed hierarchical approach consistently achieved a higher fraction of converged starts than the standard approach. Local optimization using the hierar-

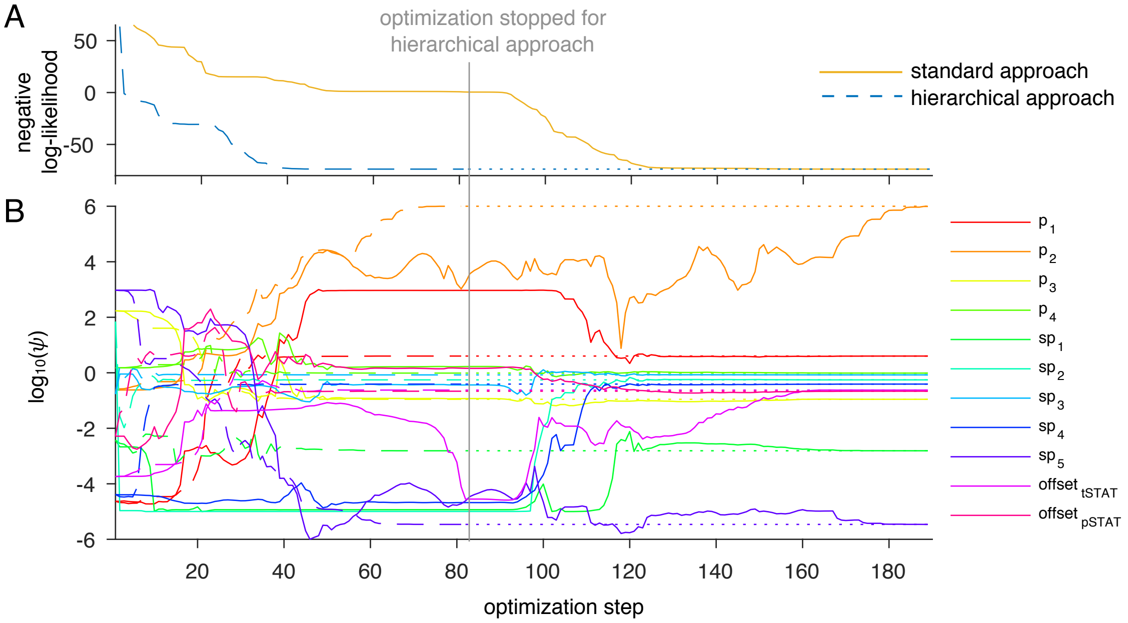

Figure 3.9: Optimization paths for JAK-STAT signaling I. (A) Objective function values and (B) parameter values at each optimization step. For a detailed description of the parameters $p_1, p_2, \ldots, \text{offset}_{\text{pSTAT}}$, we refer to Schelker et al. (2012). The trajectories for the standard approach are shown as solid lines and for the hierarchical approach as dashed lines. This figure is adapted from Figure S2 of the author's publication (Loos et al., 2018a).

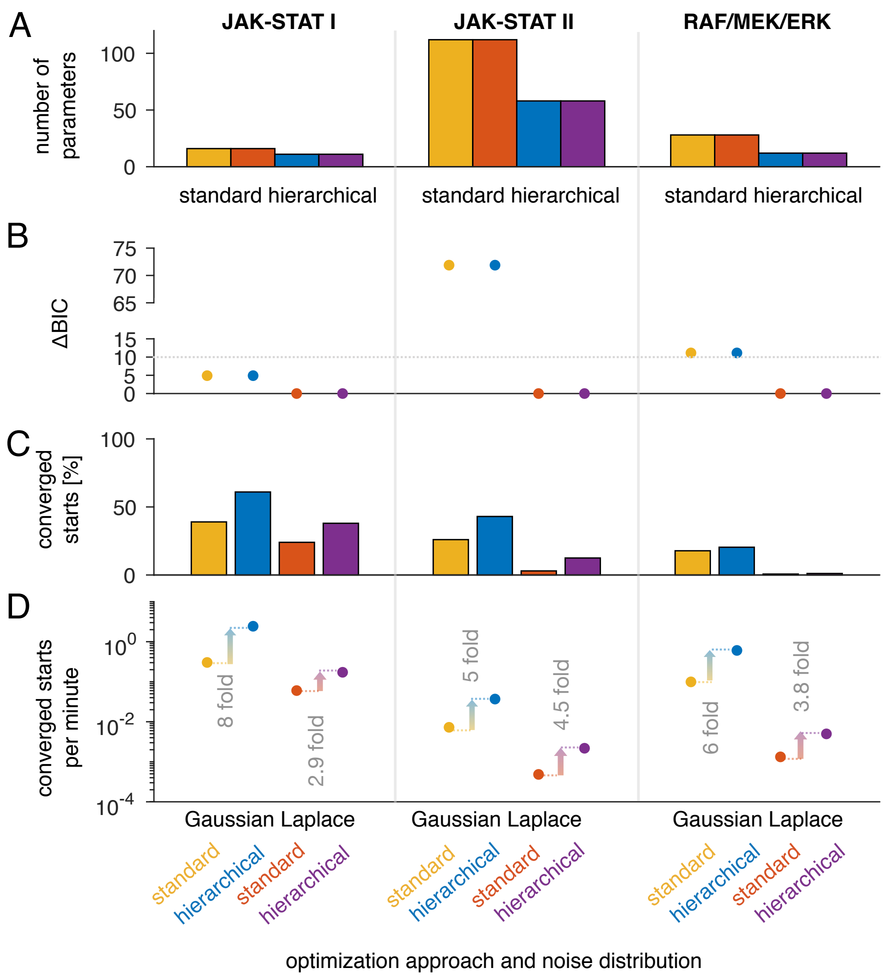

Figure 3.10: Evaluation of the standard and hierarchical approach for three signaling models. (A) Number of optimization variables in the outer problem. (B) Differences in BIC values compared to the minimal BIC for Gaussian and Laplace noise. (C) Percentage of converged starts over all performed local optimizations. (D) Number of converged starts per minute. This figure is a modified version of Figure 3 of the author's publication (Loos et al., 2018a).

chical approach converged on average in 29.3% of the runs while the standard approach converged on average in 18.4% of the runs.

We found that on average, the computation time per start was lower for the hierarchical approach than for the standard approach (Figure 3.10D). The hierarchical approach was faster than the standard approach for a high fraction of the starts. In combination with the improved convergence rate, this resulted in a substantially reduced computation time per converged start, i.e., a start which reached the minimal value observed across all starts (Figure 3.10E). Given a fixed computational budget, the hierarchical approach achieved on average 5.06 times more optimization runs which reached the best objective function values than the standard approach. The expected improvement in terms of CPU time per converged start when using the hierarchical approach was on average $3.4 \cdot 10^3$, $5.8 \cdot 10^2$, and $6.5 \cdot 10^4$ seconds for JAK-STAT I, JAK-STAT II, and RAF/MEK/ERK, respectively.

3.3.2 Profile likelihood-based uncertainty analysis employing hierarchical optimization

Employing the hierarchical approach for optimization for the calculation of the profile likelihood (2.25), we obtain

$$\begin{aligned}\mathrm{PL}(\psi_i) = & \exp\left(-\min_{\psi_{l\neq i}} J(\boldsymbol{\psi}, \hat{\mathbf{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi}))\right), \\ & \text{with } (\hat{\mathbf{s}}(\boldsymbol{\psi}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\psi})) = \underset{\mathbf{s},\boldsymbol{\sigma}}{\operatorname{argmin}}\ J(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\sigma}).\end{aligned}$$

The profile likelihoods with the hierarchical approach are the same as for the standard approach if the scaling and noise parameters are unconstrained, since the profile comprises only objective function values of optimal scaling and noise parameter. As shown before, the optimal values are the same for the standard and the hierarchical approach and also the objective function values and the resulting profile likelihoods are the same. If the gradient in the standard approach is zero, also the gradient in the hierarchical approach is zero, because it is a part of the gradient of the standard approach.

To compare the hierarchical and standard approaches for profiling, we used higher upper boundaries than used for optimization for the scaling and noise parameters in the standard approach. Employing the standard and the hierarchical approach within the routine for the profile likelihood calculation showed that the profiles coincide for the Gaussian distribution for JAK-STAT model I (Figure 3.11A). For RAF/MEK/ERK signaling, the standard approach underestimated the profiles due to convergence problems during optimization (Figure 3.11C). The profiles lay under the profiles calculated by the hierarchical approach. A similar problem has been observed by Stapor et al. (2018a) when using only first-order derivative information. Interestingly, the improved convergence of the hierarchical approach allowed to calculate the profiles even without the employment of second-order derivative information as proposed by Stapor et al. (2018a). The resulting profiles were also in good agreement with the profiles calculated by Stapor et al. (2018a), showing again also numerically that the identifiability of the model is not influenced by the use of the hierarchical approach. For profiling, we employed the trust-region and the interior-point algorithm of `fmincon` for JAK-STAT model I and RAF/MEK/ERK, respectively. These algorithms each provided more reliable results for the respective model.

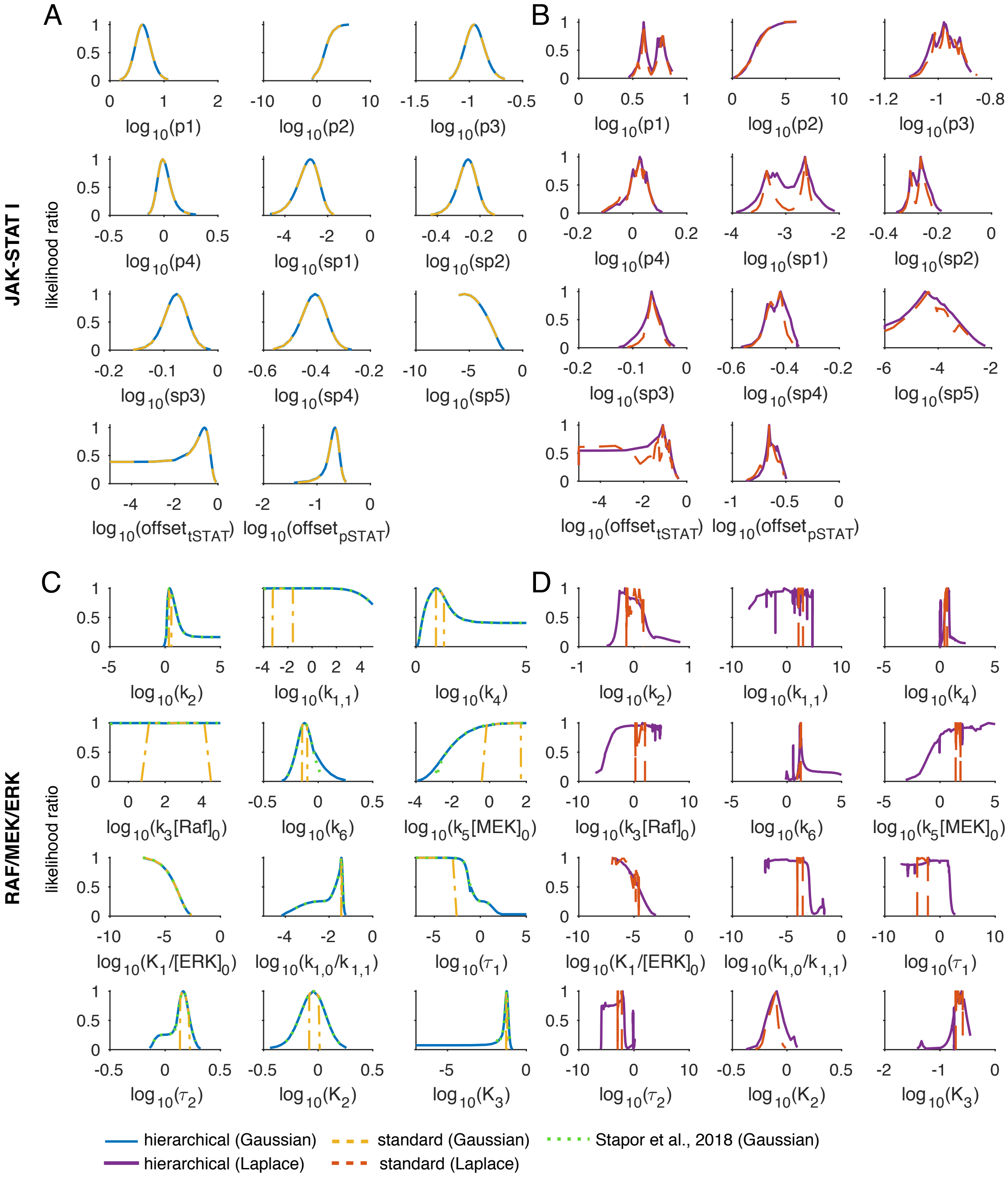

Figure 3.11: Profile likelihoods for (A,B) JAK-STAT signaling I and (C,D) RAF/MEK/ERK signaling for (A,C) Gaussian and (B,D) Laplace noise. This figure is adapted from Figures S4 and S8 of the author's publication (Loos et al., 2018a). For a detailed description of the parameters $p_1, \ldots, \text{offset}_{\text{pSTAT}}$ of JAK-STAT signaling I, we refer to Schelker et al. (2012) and for the description of parameters $k_2, \ldots, K_3$ of RAF/MEK/ERK signaling, we refer to Fiedler et al. (2016).

While optimization worked well for Laplace noise, the profile calculation for the models considered here showed difficulties, especially when using the standard approach (Figure 3.11B&D). Profile calculation for Laplace noise with the standard method failed to determine the true profile. Because of convergence problems, the profile dropped too early (as visible from the comparison of the standard and hierarchical approaches) and therefore underestimated the uncertainty. This demonstrates the relevance of the hierarchical approach. Further analysis and method development is required to enable a robust profile calculation with Laplace noise. However, employing the hierarchical approach for optimization is already a substantial improvement.

In summary, the application of our hierarchical approach to parameter estimation from relative data to the models showed consistently that our approach yielded parameter estimates of the same quality as the standard method, while achieving better convergence and reducing the computation time substantially. More reliable results were also observed when employing the optimization approach within profile likelihood-based uncertainty analysis.

3.4 Application example: Modeling the kinetics of histone H3 methylation

The evaluation of the methods proposed in Sections 3.1 and 3.2 is complemented by its application in an ongoing research project about histone methylation dynamics. The core of the nucleosome around which the DNA is wrapped consists of different histone proteins, H2A, H2B, H3 and H4. These can possess post-translational modifications (PTMs) at their N-terminal tail, at which the amino acids can be, e.g., methylated, acetylated, phosphorylated or ubiquinated. PTMs, for example, change DNA accessibility and thus play an important role in epigenetic gene regulation (Kouzarides, 2007; Orkin and Hochedlinger, 2011). Specific modifications at certain sites of the histone tails are associated with different functions: H3K27me3, i.e., trimethylation of lysine 27 of histone H3, is known to have a repressive function of transcription (Ferrari et al., 2014), while H3K36me3 has an activating function (Wagner and Carpenter, 2012). Methylation modifications are catalyzed by a group of enzymes called methyltransferases. Deregulation of methyltransferases thus yields epigenetic deregulation, which is often found in cancer cells. For example, the methyltransferase EZH2 (enhancer of zeste homolog 2), which is responsible for H3K27me3, is known to be overexpressed in many forms of cancer (Kim and Roberts, 2016). In a proliferating cell population, PTMs are diluted by continuous chromatin replication. Histones which are newly integrated into the chromatin thus need to restore the PTMs. However, the mechanisms that underlie PTM inheritance and restoration, and in particular the influence of the parental histones are so far not well understood and remain to be studied in detail.

3.4.1 Experimental data

Our experimental collaboration partners collected quantitative data of H3 methylation dynamics at K27 and K36 in mouse embryonic stem cells (mESCs) using triple-SILAC (stable isotope labeling with amino acids in cell culture) mass spectrometry. In triple-SILAC cells are cultured in three different media that contain different stable isotopes of amino acids (Figure 3.12A). We used labeled arginine (R) and lysine (K), which are incorporated into the proliferating cells. In particular, we used R0K0, R6K4 and R10K8, with changes at -3 and 0 h. Based on these isotopes, the peptides could then be distinguished in the mass spectrometer and assigned to the corresponding medium. This allowed us to track the histone modifications during DNA replication and to distinguish histones newly incorporated into the chromatin at different time points of the experiment. Histones which were incorporated on the different media belong to different generations of histones. The abundance of individual modifications was measured using LC-MSMS (liquid chromatography-mass spectrometry/mass spectrometry) which does not provide spatial information. It provides the relative abundance of each modification in a generation, i.e., all modifications in a generation at each time point sum up to 100%. We obtained data for two different scenarios (Figure 3.12A): Untreated cells ($\mathcal{D}_{\text{untr}}$) and inhibition of EZH2 ($\mathcal{D}_{\text{inh}}$) for which the inhibitor is added when changing the culture medium for the first time.

We were in particular interested in the modifications at K27 and K36, which both can be un-, mono-, di- or trimethylated, denoted by me0, me1, me2 and me3, respectively. For each histone generation, the relative abundances of 15 combinations of methylations were measured – excluding K27me3K36me3 (both lysines trimethylated) since it was not detected in any of the experiments. This yields in total 45 observables. In the following, we introduce the mathematical models which we used to describe these observables (Figure 3.12B,C) and fitted to the experimental data (Figure 3.12D).

3.4.2 Standard model

The first model we considered consists of 45 state variables, with 15 state variables for each generation. It describes the change in modifications due to methylation and demethylation as well as dilution, which occurs when the cells divide and new, unmodified histones are incorporated. This model is similar to the one proposed by Zheng et al. (2012). To obtain the model for the relative abundance of modifications, we first derived the model for the absolute number of histone modifications.

The ODE system for the absolute number of histone modifications $\tilde{\mathbf{x}}_g = (\tilde{x}_{g,00}, \dots, \tilde{x}_{g,23})$ reads for generation g

$$
\begin{aligned}
\dot{\tilde{x}}_{g,ij} = \quad & \chi_{\{(i,j)=(0,0)\}}(i,j)\, c_g(t) N \\
+ \; & \chi_{\{i>0 \wedge (i,j)\neq(3,3)\}}(i,j)\, k_{i-1\,j\to i\,j}\, \tilde{x}_{g,i-1\,j} \\
+ \; & \chi_{\{j>0 \wedge (i,j)\neq(3,3)\}}(i,j)\, k_{ij-1\to ij}\, \tilde{x}_{g,ij-1} \\
- \; & \chi_{\{i<3 \wedge (i,j)\neq(2,3)\}}(i,j)\, k_{ij\to i+1\,j}\, \tilde{x}_{g,ij} \\
- \; & \chi_{\{j<3 \wedge (i,j)\neq(3,2)\}}(i,j)\, k_{ij\to ij+1}\, \tilde{x}_{g,ij} \\
+ \; & \chi_{\{i<3 \wedge (i,j)\neq(2,3)\}}(i,j)\, d_{K27,i+1}\, \tilde{x}_{g,i+1\,j} \\
+ \; & \chi_{\{j<3 \wedge (i,j)\neq(3,2)\}}(i,j)\, d_{K36,j+1}\, \tilde{x}_{g,ij+1} \\
- \; & \chi_{\{i>0 \wedge (i,j)\neq(3,3)\}}(i,j)\, d_{K27,i}\, \tilde{x}_{g,ij} \\
- \; & \chi_{\{j>0 \wedge (i,j)\neq(3,3)\}}(i,j)\, d_{K36,j}\, \tilde{x}_{g,ij}\,, \\
\dot{N} = cN\,. &
\end{aligned}
\tag{3.25}
$$

The full equations can be found in Appendix B. The indicator function is denoted by χ, i denotes the number of methyl groups at K27, j the number of methyl groups at K36, $N(t) = \exp(c(t-t_0))N_0$ the total number of histone tails and $N(t_0) = N_0$ the number of histone tails at the beginning of the experiment (Figure 3.12B). Furthermore, $d_{K27,i}$ is the rate constant for demethylation, i.e., reducing the number of methyl groups at K27 from i to $i-1$. The special cases with $(i,j) \neq (2,3)$ and $(i,j) \neq (3,2)$ arise because we did not observe any K27me3K36me3 methylations. The newly incorporated histones are unmodified and belong to the generation of the corresponding current culture medium:

$$
c_g(t) = \begin{cases}
\chi_{\{t<-3\,\mathrm{h}\}}(t)\cdot c\,, & g=1\,,\\
\chi_{\{-3\,\mathrm{h}\leq t<0\,\mathrm{h}\}}(t)\cdot c\,, & g=2\,,\\
\chi_{\{t\geq 0\,\mathrm{h}\}}(t)\cdot c\,, & g=3\,.
\end{cases}
\tag{3.26}
$$

Here, $c > 0$ represents the cell division rate and is multiplied with the number of histone tails in (3.25), because the number of histone tails is proportional to the number of cells and thus duplicated at cell division.

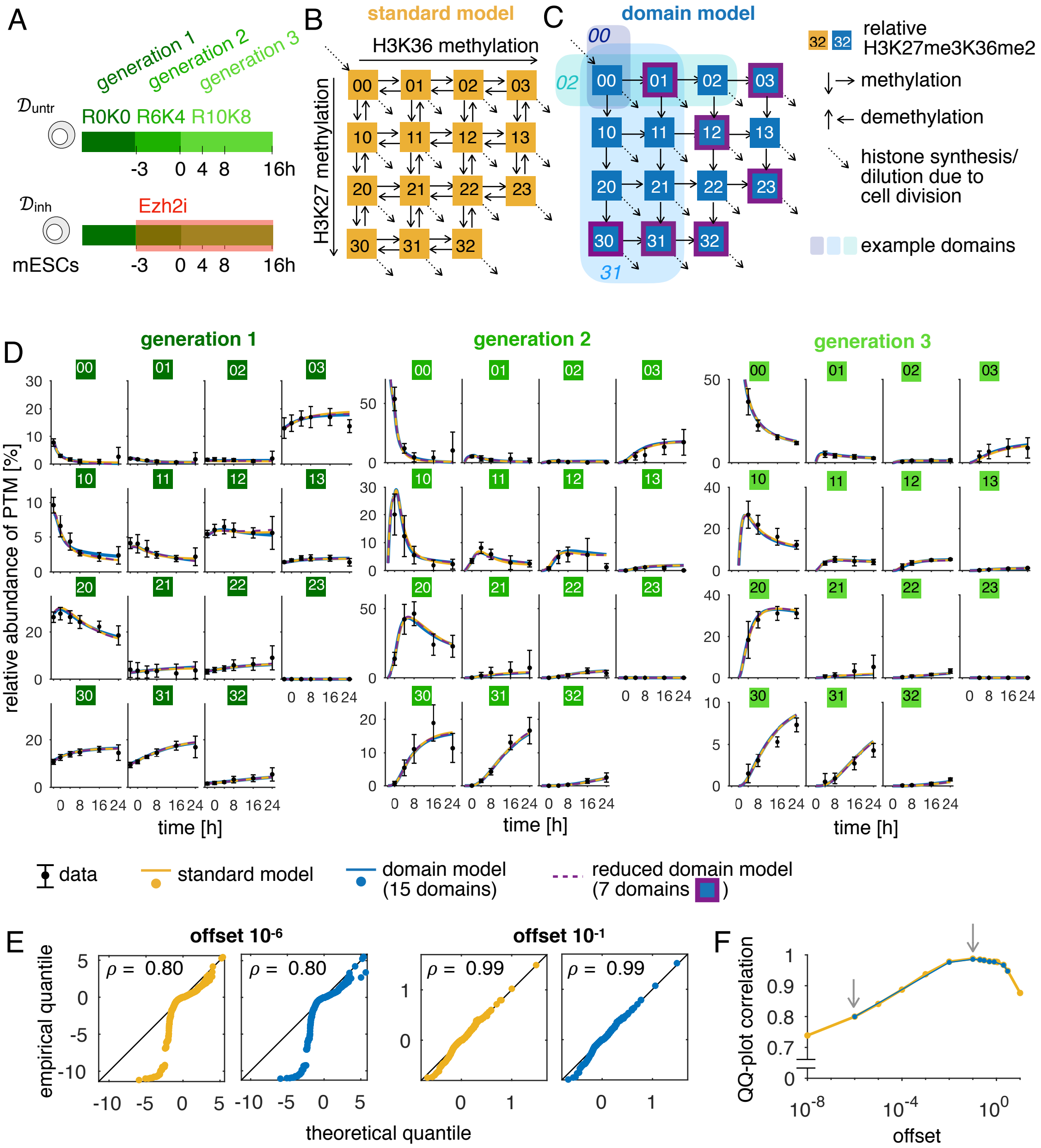

Figure 3.12: **Model and data for histone H3 methylation.** Figure caption on next page.

Figure 3.12: Model and data for histone H3 methylation. (A) Triple-SILAC mass spectrometry for mouse embryonic stem cells (mESCs) gives three generations of histones for untreated cells ($\mathcal{D}_{\text{untr}}$) and cells where an EZH2 inhibitor was added ($\mathcal{D}_{\text{inh}}$). (B,C) Model illustrations for the relative abundance of methylations at K27 and K36 for one generation. Differences between the generations are in the incorporation of unmodified histones and dilution of other modifications due to cell division (dotted diagonal arrows). This only occurs for the time where the cells are in the corresponding culture medium. (B) The standard model assumes methylation and demethylation for both lysines. (C) The domain model assumes no demethylation. Histone tails are methylated until they reach a defined final state, which depends on the domain the histone belongs to. Purple boxes indicate these final states of the domains obtained by performing model reduction. (D) Experimental data for the untreated case and model fits. (E) QQ-plots for the best model fits for two example offsets, with low and high correlation of theoretical and empirical quantiles. (F) Correlations for the QQ-plots for varying offsets. The maximum correlation is achieved for offset 10^{-1}. The gray arrows highlight the offsets for the QQ-plots shown in (E).

When changing the culture medium, initially no histones of this generation are present:

$$\tilde{x}_{g,ij}(t) = 0 \text{ for } \begin{cases} t < t_0\,, & g = 1\,, \\ t < -3\,\text{h}\,, & g = 2\,, \\ t < 0\,\text{h}\,, & g = 3\,, \end{cases} \qquad \forall i, j\,. \tag{3.27}$$

The model comprises $n_{\psi} = 29$ dynamic parameters

$$\begin{aligned} \boldsymbol{\psi} = (&c, d_{K27,1}, d_{K27,2}, d_{K27,3}, d_{K36,1}, d_{K36,2}, d_{K36,3}, k_{00\to01}, k_{00\to10}, k_{01\to02}, k_{01\to11}, \\ &k_{02\to03}, k_{02\to12}, k_{03\to13}, k_{10\to11}, k_{10\to20}, k_{11\to12}, k_{11\to21}, k_{12\to13}, k_{12\to22} \\ &k_{13\to23}, k_{20\to21}, k_{20\to30}, k_{21\to22}, k_{21\to31}, k_{22\to23}, k_{22\to32}, k_{30\to31}, k_{31\to32})\,. \end{aligned}$$

To bring the system to relative scale, we divided the total abundance of modifications by the number of histone tails

$$\begin{aligned} x_{g,ij} &= \frac{\tilde{x}_{g,ij}}{N}\,, \\ \dot{x}_{g,ij} &= \frac{\dot{\tilde{x}}_{g,ij}}{N} - \frac{\tilde{x}_{g,ij}\dot{N}}{N^2}\,. \end{aligned}$$

This yields for the relative scale

$$
\begin{aligned}
\dot{x}_{g,ij} = \quad & \chi_{\{(i,j)=(0,0)\}}(i,j)\, c_g(t) - c_g(t)\frac{\tilde{x}_{g,ij}}{N} \\
+ \quad & \chi_{\{i>0\wedge(i,j)\neq(3,3)\}}(i,j)\, k_{i-1\,j\to i\,j}\, \frac{\tilde{x}_{g,i-1\,j}}{N} \\
+ \quad & \chi_{\{j>0\wedge(i,j)\neq(3,3)\}}(i,j)\, k_{ij-1\to ij}\, \frac{\tilde{x}_{g,ij-1}}{N} \\
- \quad & \chi_{\{i<3\wedge(i,j)\neq(2,3)\}}(i,j)\, k_{ij\to i+1\,j}\, \frac{\tilde{x}_{g,ij}}{N} \\
- \quad & \chi_{\{j<3\wedge(i,j)\neq(3,2)\}}(i,j)\, k_{ij\to ij+1}\, \frac{\tilde{x}_{g,ij}}{N} \\
+ \quad & \chi_{\{i<3\wedge(i,j)\neq(2,3)\}}(i,j)\, d_{K27,i+1}\, \frac{\tilde{x}_{g,i+1\,j}}{N} \\
+ \quad & \chi_{\{j<3\wedge(i,j)\neq(3,2)\}}(i,j)\, d_{K36,j+1}\, \frac{\tilde{x}_{g,ij+1}}{N} \\
- \quad & \chi_{\{i>0\wedge(i,j)\neq(3,3)\}}(i,j)\, d_{K27,i}\, \frac{\tilde{x}_{g,ij}}{N} \\
- \quad & \chi_{\{j>0\wedge(i,j)\neq(3,3)\}}(i,j)\, d_{K36,j}\, \frac{\tilde{x}_{g,ij}}{N} \\
= \quad & \chi_{\{(i,j)=(0,0)\}}(i,j)\, c_g(t) - c_g(t) x_{g,ij} \\
+ \quad & \chi_{\{i>0\wedge(i,j)\neq(3,3)\}}(i,j)\, k_{i-1\,j\to i\,j}\, x_{g,i-1\,j} \\
+ \quad & \chi_{\{j>0\wedge(i,j)\neq(3,3)\}}(i,j)\, k_{ij-1\to ij}\, x_{g,ij-1} \\
- \quad & \chi_{\{i<3\wedge(i,j)\neq(2,3)\}}(i,j)\, k_{ij\to i+1\,j}\, x_{g,ij} \\
- \quad & \chi_{\{j<3\wedge(i,j)\neq(3,2)\}}(i,j)\, k_{ij\to ij+1}\, x_{g,ij} \\
+ \quad & \chi_{\{i<3\wedge(i,j)\neq(2,3)\}}(i,j)\, d_{K27,i+1}\, x_{g,i+1\,j} \\
+ \quad & \chi_{\{j<3\wedge(i,j)\neq(3,2)\}}(i,j)\, d_{K36,j+1}\, x_{g,ij+1} \\
- \quad & \chi_{\{i>0\wedge(i,j)\neq(3,3)\}}(i,j)\, d_{K27,i}\, x_{g,ij} \\
- \quad & \chi_{\{j>0\wedge(i,j)\neq(3,3)\}}(i,j)\, d_{K36,j}\, x_{g,ij}\,,
\end{aligned}
$$

for $g = 1, 2, 3$ and $i, j = 0, 1, 2, 3$. At relative scale, (3.26) can be seen as the dilution which occurs due to cell division.

The observables are obtained by

$$y_{g,ij} = \frac{x_{g,ij}}{\sum_{i,j} x_{g,ij}}.$$

We assumed that methylation rates did not change for different culture media and different dynamics were obtained due to presence/absence of dilution (3.26) and different initial conditions (3.27).

3.4.3 Domain model

As an alternative, we considered that demethylation is not required. For this, we proposed a model which assumes that certain domains of the chromatin are determined to acquire certain methylation patterns, e.g., due to particular transcription factor binding or parental histone context. For example, (i) histones of domain *00* do not get any methylations at all (Figure 3.12E). (ii) Histones of the domain *20* will get no additional methylations once the *20* state is reached. (iii) Histones of the domain *31* can acquire the methylation via different pathways. The histone composition tends towards the state where all domains acquired their determined state. Since newly incorporated histones are unmodified (*00*), the model still shows dynamics.

To construct this model, we denoted w_{lm} as the relative size of domain lm, with $\sum_{l,m} w_{lm} = 1$. Let $x^{lm}_{g,ij}$ be the relative abundance of histones tails with methylation K27me$i$K36me$j$ in domain lm for generation g. Then the ODEs are

$$\begin{aligned} \dot{x}^{lm}_{g,ij} = &\quad \chi_{\{(i,j)=(0,0)\}}(i,j)\, c_g(t) - c_g(t)\, x^{lm}_{g,ij} \\ &+ \chi_{\{(0<i\leq l\}}(i,j)\, k_{i-1\,j\to ij}\, x^{lm}_{g,i-1\,j} \\ &+ \chi_{\{0<j\leq m\}}(i,j)\, k_{ij-1\to ij}\, x^{lm}_{g,ij-1} \\ &- \chi_{\{i<l\}}(i,j)\, k_{ij\to i+1\,j}\, x^{lm}_{g,ij} \\ &- \chi_{\{j<m\}}(i,j)\, k_{ij\to ij+1}\, x^{lm}_{g,ij}\,, \end{aligned}$$

with $c_g(t)$ as defined in (3.26) and initial conditions

$$x^{lm}_{g,ij}(t) = 0 \text{ for } \begin{cases} t < t_0\,, & g = 1\,, \\ t < -3\,\mathrm{h}\,, & g = 2\,, \\ t < 0\,\mathrm{h}\,, & g = 3\,, \end{cases} \qquad \forall i,j,l,m\,. \tag{3.28}$$

The observables are obtained by

$$y_{g,ij} = \frac{\sum_{l,m} w_{lm} x^{lm}_{g,ij}}{\sum_{l,m} w_{lm} \sum_{i,j} x^{lm}_{g,ij}}\,.$$

We assumed that the methylation rate constants are shared between the domains, and estimated them together with the relative sizes of the domains from the data.

3.4.4 Model calibration and validation

To find the best measurement noise model, we first compared Gaussian and Laplace distributed measurement noise. We compared the model output and the observables on a log-scale and offsetted both to cope with zero measurements

$$\log(\bar{y}_{g,ij} + \text{offset}) \sim p\left(\log(\bar{y}_{g,ij} + \text{offset})\right| \log\left(y_{g,ij} + \text{offset}\right), \sigma\right) , \tag{3.29}$$

with noise distribution p. We chose offset $= 10^{-1}$ which provided the best fit with respect to the QQ-plots for the standard model and the domain model with 15 domains using Laplace noise (Figure 3.12D-F). We performed 100 local optimization runs for Gaussian and Laplace noise as well as the standard and the hierarchical approach for optimization. For the hierarchical approach, we analytically calculated the measurement noise parameter σ, which is shared for all time points, observables and generations. We found strong support for Laplace noise over Gaussian noise for both models ($\Delta\text{BIC} > 700$). This demonstrates the importance of employing a heavier tailed distribution assumption as proposed in Section 3.1. Even though the optimization problem was reduced by only one parameter using the hierarchical approach for optimization, the performance, i.e., the number of converged starts per minute, increased by 20% (standard model) and 30% (domain model) for the given realizations of multi-starts. Both models were able to describe the data well (Figure 3.12D).

We did not expect all 15 domains to be necessary to explain the data and thus the domain model could be overparametrized. Since it was unclear which domains exist a priori, we performed model selection to detect the present domains. If we would consider all potential combinations of domains, we would end up with 2^{15} models, which is computationally too expensive. Thus, we employed model reduction techniques described in Section 2.4.2. Since $\sum_{l,m} w_{lm} = 1$, the model reduction using a l_1 penalized objective function as done by Steiert et al. (2016) is not possible. The penalty would always be exactly one independent of the choice of w_{lm}. Performing forward-selection and backward-elimination (Section 2.4.2, Figure 3.13), we found seven domains which were necessary to explain the data (highlighted in Figure 3.12C). The corresponding fit of the reduced model is shown in Figure 3.12D.

To further test and validate the models, we used the data $\mathcal{D}_{\text{inh}}$ of the inhibitor experiment. EZH2 is the only known methyltransferase for K27 trimethylation (Kuzmichev et al., 2002) and this enzyme was inhibited by EPZ-6438. This inhibitor competes with S-adenosylmethionine (SAM) for the binding to the EZH2 SET domain and directly blocks the transfer of the methyl group to histone tails. Using our calibrated models, we predicted

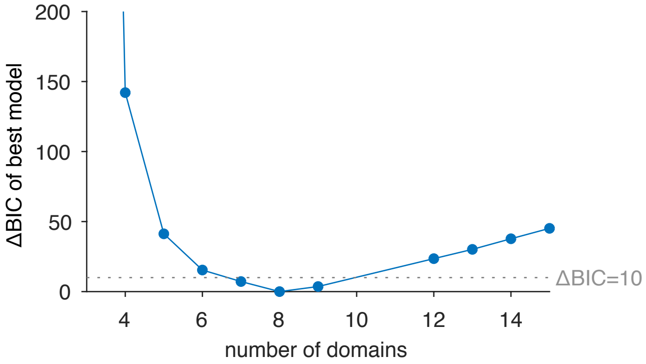

Figure 3.13: Model reduction for the domain model. The difference in BIC values is shown for the best tested model with given numbers of domains. A model with eight domains has the best BIC value, and there are models with seven and nine domains which cannot be rejected according to their BIC value. The dotted line shows the generally employed threshold of ΔBIC=10 (Table 2.1).

the total amount of K27me3 in generation 1 under inhibitor treatment (Figure 3.14). For this, we assumed that the trimethylation rate was inhibited by a factor κ:

$$k_{2j\to 3j,\text{inh}} = (1-\kappa)\cdot k_{2j\to 3j,\text{untr}}\,. \tag{3.30}$$

To obtain reasonable values for κ, we analyzed a simplified model which only considers K27 methylations of histones of generation 3, x_0, x_1, x_2, x_3 for un-, mono-, di-, and trimethylation at K27 and assumes independence between the methylation sites. The model reads

$$\begin{aligned}
\dot{x}_0 &= c - cx_0 - k_{0\to 1}\,x_0 + d_{K27,1}x_1\,,\\
\dot{x}_1 &= -cx_1 + k_{0\to 1}\,x_0 - k_{1\to 2}\,x_1 + d_{K27,2}x_2 - d_{K27,1}x_1\,,\\
\dot{x}_2 &= -cx_2 + k_{1\to 2}\,x_1 - k_{2\to 3}\,x_2 + d_{K27,3}x_3 - d_{K27,2}x_2\,,\\
\dot{x}_3 &= -cx_3 + k_{2\to 3}\,x_2 - d_{K27,3}x_3\,.
\end{aligned}$$

The steady states for K27me3 for untreated cells and cells in the inhibitory experiment are given by

$$x_{3,\text{untr}} = \frac{k_{2\to 3}\,x_{2,\text{untr}}}{d_{K27,3}+c}\,, \tag{3.31}$$

$$x_{3,\text{inh}} = \frac{(1-\kappa)k_{2\to 3}\,x_{2,\text{inh}}}{d_{K27,3}+c}\,. \tag{3.32}$$

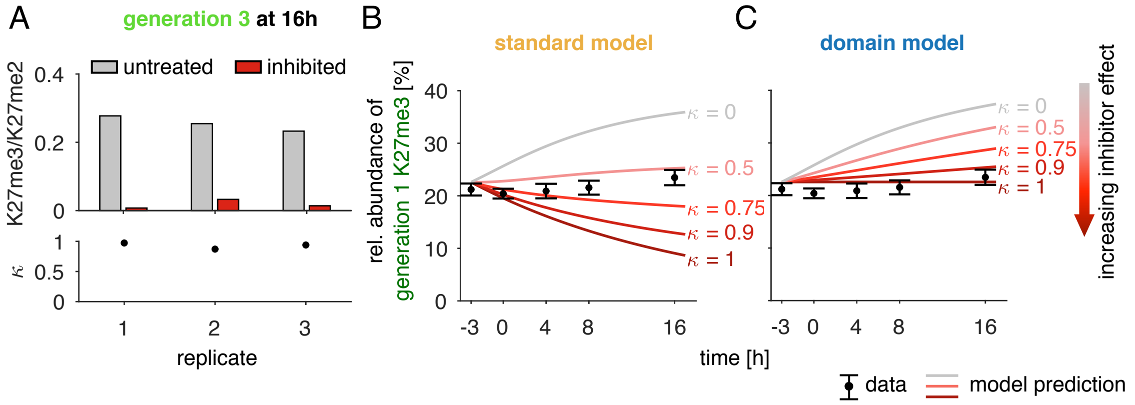

Figure 3.14: Model validation for inhibitor treatment data. (A) Ratio of K27me3 and K27me2 levels for untreated histones of generation 3 at 16 h for three replicates. From this, the inhibitor efficiency was estimated with a factor κ which changes K27 trimethylation $k_{2\to3,\mathrm{inh}} = (1-\kappa)k_{2\to3,\mathrm{untr}}$. The average value for κ from three replicates is 0.928. (B,C) Data for the K27me3 levels of generation 1 for the inhibitor treatment and model predictions for (B) the standard model and (C) the domain model. Both models were calibrated for the untreated case for all generations. The gray lines shows the model prediction for the case of no inhibition ($\kappa = 0$), the dark red line the case of full inhibition ($\kappa = 1$).

Thus, K27me3 only depends on the trimethylation rate $k_{2\to3}$, the demethylation $d_{K27,3}$, the dilution rate c, the amount of relative K27me2 and in the inhibitor case the factor κ (3.30) by which the trimethylation is inhibited. Thus we obtained

$$\frac{d_{K27,3}+c}{k_{2\to3}} = \frac{x_{2,\mathrm{untr}}}{x_{3,\mathrm{untr}}} = (1-\kappa)\frac{x_{2,\mathrm{inh}}}{x_{3,\mathrm{inh}}} \tag{3.33}$$

$$\Rightarrow \kappa = 1 - \frac{x_{2,\mathrm{untr}}\,x_{3,\mathrm{inh}}}{x_{3,\mathrm{untr}}\,x_{2,\mathrm{inh}}}\,. \tag{3.34}$$

Using the last time point for generation 3 from data $\mathcal{D}_{\mathrm{inh}}$, we obtained for three replicates a rough estimate $\kappa = 0.928 \pm 0.052$ (Figure 3.14A). The same expression for κ (3.34) is also valid for the domain model.

We compared our model predictions to the experimental data for K27me3 levels, i.e., summing all states with K27me3, to be robust against potential effects of the inhibitor on K36 methylations. Predicting the K27me3 levels of generation 1 with inhibition of EZH2 for different values of κ, we found that the standard model failed to explain the data for reasonable ranges of κ (Figure 3.14B). Here, we only assumed the trimethylation rate to change. However, if also mono-, or dimethylation changes, the model predictions would be even lower and the illustrated predictions in Figure 3.14B can be seen as rough estimates for the upper bound. However, the prediction of the domain model for the inhibitor treatment was in good agreement with the data (Figure 3.14C).

In summary, we found that a domain model is more suitable to describe the kinetics of histone H3 methylation at K27 and K36. The antagonism between the two methylation sites was reflected in the rates which were estimated from experimental triple-SILAC LC-MSMS data. For this study, we applied the methodological approaches which were developed in Sections 3.1 and 3.2, demonstrating their practical importance and applicability.

3.5 Summary and discussion

In this chapter, we evaluated different heavier tailed distributions for the calibration of dynamical models. We found that the Laplace distribution seems to be the most promising distribution. On the one hand, it provides reliable parameter estimates and confidence intervals in the absence and presence of outliers, on the other hand it still has a reasonable optimizer performance and does not yield problems when the model tends to over-fit the data. For this distribution and the generally used Gaussian distribution, we were able to derive analytical expressions for the optimal measurement noise parameters as well as for the optimal scaling parameters, which are required to model relative data. This consistently improved optimization, even if the dimension of the optimization problems is reduced by only one parameter, as it was the case for the considered application example of histone methylation. For this example, we compared different mathematical models explaining the restoration of histone modifications and gained new insights into the dynamics of histone methylation. Furthermore, the developed methods are currently used in other application examples which are not included in this thesis, for example, in a project about viral infection where we collaborate with Frederik Graw from BioQuant in Heidelberg.

For the considered models, we observed that the fraction of converged local optimization runs decreases as the model dimension increases. Potential reasons are that for larger models the region of attraction of the global optimum is potentially smaller. We also observed that the fraction of converged starts is lower for Laplace noise than for Gaussian noise. This most probably occurs due to non-differentiabilities in the objective function, which complicate the optimization procedure. When using Laplace priors for parameters, the optimization routine can be adapted (Steiert et al., 2016). However, this approach is not easily transferable to the use of Laplace measurement noise, as the switching points depend on the numerical solution of the ODE.

In addition to the scaling and noise parameters, other parameters which only contribute to the mapping from the states to the observables could also be optimized analytically. This includes offset parameters, which are used to model background intensities or unspecific binding. Extending our approach to also calculate these parameters analytically would decrease the parameters in the outer optimization even more.

When using gradient-based optimization, further improvements could be achieved by extending the approach to scalable approaches to calculate the objective function gradient. In this thesis, we employed forward sensitivities for the calculation of the gradient. However, it has been shown that for large-scale models with a high number of parameters, adjoint sensitivities can reduce the computation time needed for simulation (Fröhlich et al., 2017). Thus, a further promising approach is the combination of both complementary approaches for the handling of large-scale models.

In summary, the presented methods enabled a robust and efficient calibration of ODE models. The methods can cope with an increasing complexity of the data and corresponding models and, thus, will facilitate the in-depth mechanistic analysis of biological processes.

Chapter 4

Mechanistic modeling of heterogeneous cell populations based on single-cell snapshot data

Cellular heterogeneity is critical for cellular decision making and the formation of complex organisms (Balázsi et al., 2011). To study heterogeneity, experimental techniques (Section 2.1.2) have been developed which yield increasing amounts of data. A large number of powerful statistical methods have been developed for the analysis of single-cell data (Kharchenko et al., 2014; Lun et al., 2017). Unfortunately, these are unable to identify causalities and latent causes, or to reconstruct the governing equations of the process. Improved methods of data analysis are therefore required.

In this chapter, we address the problem (iii) stated in Section 1.2 that a unifying framework which facilitates a mechanistic description of the heterogeneity in the presence of subpopulations is missing. Furthermore, we tackle the problem (iv) that the influence of the incorporated distribution assumptions on the optimization results and performance is unknown (Section 1.2). We develop and analyze the framework and apply it to study pain sensitization in primary sensory neurons to address problem (vi).

In the case of homogeneous cell populations, the RREs (2.4) provide a description of the population behavior (Figure 4.1A). Stochastic fluctuations or latent differences between cells result in cell-to-cell variability and a distribution of cell states (Filippi et al., 2016; Hasenauer et al., 2011a; Yao et al., 2016; Zechner et al., 2012) (Figure 4.1B). The statistical moments of this distribution are described by moment-closure approximations (2.14, 2.15) and system size expansions (Fröhlich et al., 2016; van Kampen, 2007). These provide scalable approximations for a range of processes in which variability arises from different sources. However, the approximation could be crude, e.g., even negative variances can be predicted (Schnoerr et al., 2014). Additionally, they fail to provide an accurate description of the population heterogeneity when subpopulations are present and cannot be used to study the causal differences between cells and subpopulations.

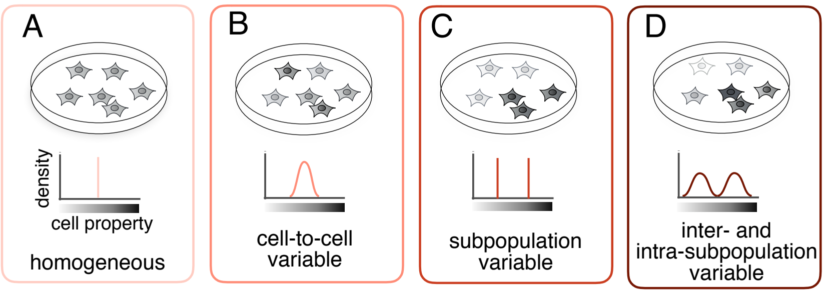

Figure 4.1: Cell populations exhibiting different levels of heterogeneity. Properties of cells, e.g., receptor levels or reaction rates, indicated by different gray shades for individual cells, can be (A) homogeneous: the property is the same for the entire cell population; (B) cell-to-cell variable: the property has a unimodal distribution across the cells; (C) subpopulation variable: the population can be separated into subpopulations, but within each subpopulation, the property does not vary; (D) inter- and intra-subpopulation variable: the property splits the population into subpopulations and also varies between cells within a subpopulation. This figure is adapted from Figure 1 of the author's publication (Loos et al., 2018b)

To address parameter differences between cellular subpopulations, RRE-constrained mixture modeling (2.17) combines mixture modeling and mechanistic RRE modeling of the subpopulation means (Figure 4.1C). Cell-to-cell variability within a subpopulation is treated naively as an additional parameter that has to be estimated. Thus, no mechanistic description of cell-to-cell variability within a subpopulation is possible. Moreover, the method can only be applied to one-dimensional measurements. When multivariate measurements are used, only marginal distributions can be analyzed and correlations between measurements are neglected. This may result in a substantial loss of information (Altschuler and Wu, 2010; Buchholz et al., 2013).

We introduce a non-trivial combination of mixture models that is able to capture subpopulation structures and models for individual subpopulations that account for differences between individual cells (Figure 4.1D). The approach therefore covers several levels of heterogeneity simultaneously (Figure 4.1A-D). We provide the equations which are required for the calibration of these models. Afterwards, we evaluate the capability of the framework to disentangle different sources of variability and handle multivariate data. We apply our framework to study pain sensitization based on quantitative single-cell microscopy of cultured sensory neurons. The presented modeling framework relies on parametric mixture distributions. Therefore, we incorporate different distribution assumptions in the framework and analyzed their robustness and performance.

The following chapter is based on and in part identical to these publications:

- **Loos, C.***, Moeller, K.*, Fröhlich, F., Hucho, T., & Hasenauer, J. (2018). A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Systems*, 6(5), 593-603. (*equal contribution)

- **Loos, C.**, Fiedler, A., & Hasenauer, J. (2016). Parameter estimation for reaction rate equation constrained mixture models. In *International Conference on Computational Methods in Systems Biology* (pp. 186-200). Springer International Publishing.

- **Loos, C.**, & Hasenauer, J. Robust calibration of hierarchical population models on single-cell snapshot data. *in preparation*

4.1 Hierarchical population model

For the hierarchical population model, the mechanistic description of individual subpopulations is combined with mixture models to describe the entire cell population. Each subpopulation itself is allowed to be heterogeneous.

4.1.1 Hierarchical model and its approximations

We considered heterogeneous cell populations consisting of multiple subpopulations, $s = 1, ..., n_s$. Assuming independence, the distribution of the states $\mathbf{x}$ and observables $\mathbf{y}$ in the overall population at time t, $p(\mathbf{x}, t)$ and $p(\mathbf{y}, t)$, is the weighted sum of the distribution of the states and observables in the subpopulations, $p_s(\mathbf{x}, t)$ and $p_s(\mathbf{y}, t)$. The weights $w_s(t)$ are the relative population sizes, with $\forall t : \sum_s w_s(t) = 1$. This yields the hierarchical population model for states and observables

$$\begin{aligned} p(\mathbf{x}, t) &= \sum_s w_s(t) p_s(\mathbf{x}, t)\,, \\ p(\mathbf{y}, t) &= \sum_s w_s(t) p_s(\mathbf{y}, t)\,. \end{aligned}$$

As the measurements $\bar{\mathbf{y}}$ are in general noise corrupted, $\bar{\mathbf{y}} \sim p(\bar{\mathbf{y}}|\mathbf{y})$, we also considered the distribution

$$\begin{aligned} p(\bar{\mathbf{y}}, t) &= \int p(\bar{\mathbf{y}}|\mathbf{y}) p(\mathbf{y}, t) d\mathbf{y} \\ &= \sum_s w_s(t) \underbrace{\int p(\bar{\mathbf{y}}|\mathbf{y}) p_s(\mathbf{y}, t) d\mathbf{y}}_{=: p_s(\bar{\mathbf{y}}, t)}\,. \end{aligned}$$

To ensure computational efficiency, we mostly worked with the statistical moments. For the measured observables, the computed statistical moments were encoded in $\boldsymbol{\varphi}_s$, yielding

$$p(\bar{\mathbf{y}}, t) = \sum_s w_s(t)\phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}_s(t)) \tag{4.1}$$

with parametric probability distribution ϕ.

In the following, we explain how the statistical moments of distribution ϕ are obtained and linked to the distribution parameters $\boldsymbol{\varphi}_s$. For this, we assumed that cells differ in their cellular properties. Each cell indexed by c has cellular properties which are encoded in the parameter vector $\boldsymbol{\psi}^c \in \mathbb{R}^{n_\psi}$. In the hierarchical framework (Figure 4.2), these parameters are considered to be drawn from a mixture distribution as follows:

$$\boldsymbol{\psi}^c \sim \sum_s w_s \mathcal{N}(\boldsymbol{\beta}_s, \mathbf{D}_s)\,,$$

with subpopulation weight w_s, mean $\boldsymbol{\beta}_s$ and covariance $\mathbf{D}_s$ for subpopulation $s = 1, \dots, n_s$. The subpopulation parameters $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \boldsymbol{D}_s)$ classify the variability of a property ψ_i: The subpopulation parameters $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \boldsymbol{D}_s)$ are given by

$$\beta_{s,i} = \begin{cases} \beta_i\,, & \text{homogeneous}\,,\\ \beta_i\,, & \text{cell-to-cell variable}\,,\\ \beta_{s,i}\,, & \text{subpopulation variable}\,,\\ \beta_{s,i}\,, & \text{inter- and intra-subpopulation variable}\,,\end{cases}$$

$$D_{s,ii} = \begin{cases} 0\,, & \text{homogeneous}\,,\\ D_{ii}\,, & \text{cell-to-cell variable}\,,\\ 0\,, & \text{subpopulation variable}\,,\\ D_{s,ii}\,, & \text{inter- and intra-subpopulation variable}\,,\end{cases}$$

and allow for correlated parameters, $D_{s,ij} \neq 0$, with indices $i, j = 1, \dots, n_\psi$ for the cellular property. The distribution of the parameters produces a distribution of cell states and observables (Figure 4.3).

The temporal evolution of the statistical properties of the cells of a subpopulation, including the mean and covariance, were computed using scalable methods. System size expansions and moment-closure approximations are used to account for stochastic single-cell dynamics, whereas sigma-points are used otherwise (Section 2.2.2). These approaches yield ODE models for the statistical moments, comprising the means and covariances

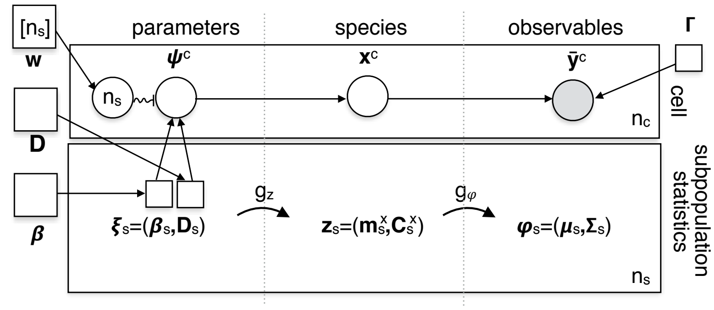

Figure 4.2: Plate notation for the structure of the single-cell system and approximation by the hierarchical population model. Squares indicate fixed parameters, whereas circles indicate random variables. Gray shading of the circles/squares indicates a known value, whereas the other values are latent. The upper plate illustrates the variables associated with a cell c. Each of the n_c cells has parameters $\boldsymbol{\psi}^c$ drawn from a distribution defined by $\boldsymbol{\xi}_s$ and $\mathbf{w}$. The states of the species $\mathbf{x}^c$, resulting from the single-cell dynamics, yield the observables $\bar{\mathbf{y}}^c$, additionally influenced by measurement noise $\boldsymbol{\Gamma}$. The bottom plate visualizes the statistics of the corresponding cells of a subpopulation. For each subpopulation, the subpopulation parameters $\boldsymbol{\xi}_s$ are mapped to the means and covariances of the species of a subpopulation $\mathbf{z}_s$, which then are mapped to the distribution parameters $\boldsymbol{\varphi}_s$. The observables at the population level are considered to be distributed according to (4.1). This figure is a modified version of Figure 2C of the author's publication (Loos et al., 2018b).

$\mathbf{z}_s = (\mathbf{m}_s^x, \mathbf{C}_s^x)^T$ of species $\mathbf{x}$. The models are simulated for each of the n_s subpopulations

$$\dot{\mathbf{z}}_s = g_z\left(\mathbf{z}_s, \boldsymbol{\xi}_s, \mathbf{u}\right), \quad \mathbf{z}_s(0) = \mathbf{z}_0\left(\boldsymbol{\xi}_s, \mathbf{u}\right), \tag{4.2}$$

with initial conditions $\mathbf{z}_0$ and experimental condition $\mathbf{u}$. The moments of the species in a subpopulation are then mapped to the distribution parameters

$$\boldsymbol{\varphi}_s = g_\varphi\left(\mathbf{z}_s, \boldsymbol{\xi}_s, \mathbf{u}\right), \tag{4.3}$$

of the distribution ϕ, including measurement noise $\boldsymbol{\Gamma}$, which is mostly assumed to be the same for all subpopulations.

In this section, we employed for ϕ the multivariate normal distribution

$$\mathcal{N}(\bar{\mathbf{y}}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{1}{(2\pi)^{\frac{n_y}{2}} \det\left(\boldsymbol{\Sigma}_s\right)^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{\mathbf{y}}-\boldsymbol{\mu}_s)^T(\boldsymbol{\Sigma}_s)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}_s)}, \tag{4.4}$$

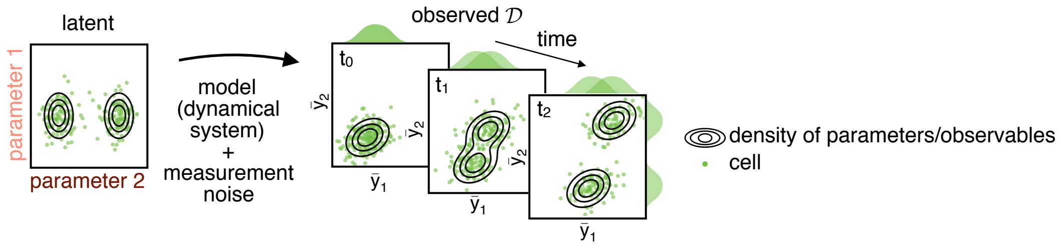

Figure 4.3: Illustration of the dynamics of a heterogeneous cell population and the mechanistic hierarchical population model. (A) Parameter distribution of a cell population consisting of two subpopulations. The contour lines illustrate the (approximated) parameter density of the cell-to-cell variable parameter 1 and the inter-and intra-subpopulation variable parameters 2. The heterogeneity of parameters is propagated from the latent parameter space to the observed measurement space. (B) Heterogeneity in parameters yields heterogeneous observables $\mathbf{y} = (y_1, y_2)^T$ that separate into two subpopulations after stimulation at time point t_0. This figure is a modified version of Figure 2A-B of the author's publication (Loos et al., 2018b)

and multivariate log-normal distribution

$$\log\mathcal{N}\left(\bar{\mathbf{y}}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right) = \frac{1}{(2\pi)^{\frac{n_y}{2}} \det\left(\boldsymbol{\Sigma}_s\right)^{\frac{1}{2}} \left(\prod_{i=1}^{n_y} \bar{y}_i\right)} e^{-\frac{1}{2}(\log(\bar{\mathbf{y}})-\boldsymbol{\mu}_s)^T(\boldsymbol{\Sigma}_s)^{-1}(\log(\bar{\mathbf{y}})-\boldsymbol{\mu}_s)}, \tag{4.5}$$

with distribution parameters $\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$.

For the multivariate normal distribution (4.4), the distribution parameters can be obtained by

$$\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = g_\varphi((\mathbf{m}_s^x, \mathbf{C}_s^x), \boldsymbol{\xi}_s, \mathbf{u}) = \left(\mathbf{m}_s^y, \mathbf{C}_s^y + \boldsymbol{\Gamma}\right), \tag{4.6}$$

including additive normally distributed measurement noise parametrized by

$$\boldsymbol{\Gamma} = (\Gamma_{ij})_{i,j=1,\ldots,n_y} = \begin{pmatrix} \sigma^2_{1,\text{noise}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{n_y,\text{noise}} \end{pmatrix}. \tag{4.7}$$

For the multivariate log-normal distribution (4.5), the distribution parameters can directly be simulated with the sigma-point approximation for the logarithm of the observable, yielding the relation

$$\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = g_\varphi((\mathbf{m}_s^x, \mathbf{C}_s^x), \boldsymbol{\xi}_s, \mathbf{u}) = \left(\mathbf{m}_s^{\log(y)}, \mathbf{C}_s^{\log(y)} + \boldsymbol{\Gamma}\right), \tag{4.8}$$

Table 4.1: Comparison of the hierarchical population model with existing methods.

method	mechanistic description of		subpopulations	multivariate data	reference
	dynamics	variability			
mixture model			✓	✓	e.g., Hastie et al. (2009)
moment-closure approximation	✓	✓		✓	e.g., Zechner et al. (2012)
RRE-constrained mixture model	✓		✓		Hasenauer et al. (2014)
hierarchical population model	✓	✓	✓	✓	Loos et al. (2018b)

accounting for multiplicative log-normally distributed measurement noise. Alternatively, the mean of the simulation can be linked to the mean of the log-normal distribution by

$$\begin{aligned}\mu_{s,i} &= \log(m^y_{s,i}) - \frac{1}{2}\Sigma_{s,ii}\,,\\ \Sigma_{s,ij} &= \log\left(\frac{C^y_{s,ij}}{m^y_{s,i}m^y_{s,j}} + 1\right) + \Gamma_{ij}\,,\end{aligned}$$

with observable indices i and j. In Section 4.4, we provide the equations for the incorporation of further distributions.

The sigma-point or moment-closure approximation (Section 2.2.2) provides time-dependent moments of the system and accounts for cell-to-cell variability. When combined with subpopulation variability, this yields both the inter- and intra-subpopulation variability (Figure 4.1D). For a comparison of the hierarchical population model with existing methods we refer to Table 4.1. In this thesis, we assumed a log-normal distribution of the parameters, i.e., $\boldsymbol{\beta}$ and $\mathbf{D}$ describe the median and scale matrix of the corresponding log-normal distribution, and the exponent of $\boldsymbol{\mathcal{S}}_l$ was used in (2.11).

4.1.2 Likelihood function for the hierarchical population model

The parameters of the hierarchical population model $\boldsymbol{\theta}$ comprise the means/medians of the single-cell parameters $\boldsymbol{\beta}, \boldsymbol{\beta}_s$ as well as the entries of the scale matrices $\mathbf{D}, \mathbf{D}_s$, the mixture weights w_s and entries of the measurement noise matrix $\boldsymbol{\Gamma}$. These parameters need to be estimated from experimental data. For this, we employed a maximum likelihood approach as introduced in Section 2.3.2.

The likelihood function of the hierarchical population model for multivariate measurement data $\bar{\mathbf{y}}_k^c \in \mathbb{R}^{n_y}$ with time index k and single-cell index c is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{k,c} \sum_s w_s(t_k, \boldsymbol{\theta}, \mathbf{u})\, \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right) \\ \text{with} \quad \dot{\mathbf{z}}_s &= g_z\left(\mathbf{z}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right), \quad \mathbf{z}_s(0) = \mathbf{z}_0\left(\boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right), \\ \boldsymbol{\varphi}_s &= g_\varphi\left(\mathbf{z}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right),\end{aligned} \tag{4.9}$$

with means and covariances $\mathbf{z}_s = (\mathbf{m}_s^x, \mathbf{C}_s^x)^T$ of species $\mathbf{x}$. In general, the subpopulation weight w_s and the subpopulation parameters can also be experiment specific (see Section 4.3.3). Also other distributions can be incorporated in our modeling framework (Section 4.4). Due to numerical reasons, we used the negative log-likelihood function (Loos et al., 2016)

$$J(\boldsymbol{\theta}) = -\log \mathcal{L}(\boldsymbol{\theta}) = -\sum_{k,c} \log\left(\sum_s w_s(t_k, \boldsymbol{\theta}, \mathbf{u})\, \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right)\right). \tag{4.10}$$

To promote efficiency of the numerical optimization and robust convergence, we derived the gradient of the negative log-likelihood function

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_i} = -\sum_{k,c} & \frac{1}{\sum_s w_s(t_k, \boldsymbol{\theta}, \mathbf{u})\, \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right)} \cdot \\ & \sum_s \left(\frac{\partial w_s(t_k, \boldsymbol{\theta}, \mathbf{u})}{\partial \theta_i} \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right) + w_s(t_k, \boldsymbol{\theta}, \mathbf{u}) \frac{\partial \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right)}{\partial \theta_i}\right).\end{aligned} \tag{4.11}$$

The gradient of the negative log-likelihood function (4.11) comprises the gradient of the corresponding mixture distribution ϕ with respect to θ_i. For a simpler notation we only explicitly denote the dependence of the distribution parameters on $\boldsymbol{\theta}$. For the normal and log-normal distributions, it is

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \mathcal{N}\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right) = -\frac{1}{2} \mathcal{N}\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right) \cdot & \Bigg(\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)^{-1} \frac{\partial \boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial \theta_i}\right) \\ & + \left(\boldsymbol{\mu}_s(\boldsymbol{\theta}) - \bar{\mathbf{y}}_k^c\right)^T \left(\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)^{-1} \left(\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \\ & + \left(\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \left(\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)^{-1} \left(\boldsymbol{\mu}_s(\boldsymbol{\theta}) - \bar{\mathbf{y}}_k^c\right) \\ & + \left(\boldsymbol{\mu}_s(\boldsymbol{\theta}) - \bar{\mathbf{y}}_k^c\right)^T \frac{\partial \left(\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)^{-1}}{\partial \theta_i} \left(\boldsymbol{\mu}_s(\boldsymbol{\theta}) - \bar{\mathbf{y}}_k^c\right)\Bigg),\end{aligned} \tag{4.12}$$

and the relation

$$\log\mathcal{N}\left(\bar{\mathbf{y}}_k^c|\boldsymbol{\mu}_s(\boldsymbol{\theta}),\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)=\mathcal{N}\left(\log(\bar{\mathbf{y}}_k^c)|\boldsymbol{\mu}_s(\boldsymbol{\theta}),\boldsymbol{\Sigma}_s(\boldsymbol{\theta})\right)\left(\prod_{i=1}^{n_y}y_{i,k}^c\right)^{-1}.$$

Additionally, the sensitivities of the distribution parameters $\partial\boldsymbol{\mu}_s(\boldsymbol{\theta})/\partial\theta_i$ and $\partial\boldsymbol{\Sigma}_s(\boldsymbol{\theta})/\partial\theta_i$ were required, which were obtained by simulating the sensitivity equation for the sigma-point or the moment-closure approximation to obtain $\partial\mathbf{z}_s(\boldsymbol{\theta})/\partial\theta_i$, and linking it to the distribution parameters using g_φ.

A reformulation of the equations for the robust evaluation of the log-likelihood function and its gradient is given in the following. This reformulation prevents numerical problems when small probabilities occur (Murphy, 2012). For fixed k and c, we define $\phi_s=\phi\left(\bar{\mathbf{y}}_k^c|\boldsymbol{\varphi}_s(t_k,\boldsymbol{\theta},\mathbf{u})\right)$, $q_s:=\log(\phi_s)$, $w_s=w_s(t_k,\boldsymbol{\theta},\mathbf{u})$ and $\hat{s}=\text{argmax}_s q_s$, and reformulate

$$\log\left(\sum_s w_s e^{q_s}\right)=\log\left(1+\sum_{s\neq\hat{s}}\frac{w_s}{w_{\hat{s}}}\left(e^{q_s-q_{\hat{s}}}\right)\right)+\log(w_{\hat{s}})+q_{\hat{s}}\,. \tag{4.13}$$

Regarding the calculation of the gradient we obtain

$$\begin{aligned}\frac{\partial\log\sum_s w_s\phi_s}{\partial\theta_i}&=\frac{1}{\sum_{s^*}w_{s^*}\phi_{s^*}}\cdot\sum_s\frac{dw_s\phi_s}{d\theta_i}\\&=\frac{1}{\sum_{s^*}w_{s^*}e^{q_{s^*}-q_{\hat{s}}}}\cdot\sum_s e^{q_s-q_{\hat{s}}}\left(\frac{\partial w_s}{\partial\theta_i}+\frac{w_s}{\phi_s}\frac{\partial\phi_s}{\partial\theta_i}\right).\end{aligned} \tag{4.14}$$

The proposed reformulations (4.13) and (4.14) were used for the robust evaluation of the log-likelihood function and its gradient.

4.1.3 Calibration of the single-cell model

The calibrated hierarchical population model provides estimates for $\beta_{s,i}$, and $D_{s,ii}$ which can then be used as prior information for the single-cell parameters $\boldsymbol{\psi}^c$ of cell c:

$$p(\psi_i^c)=\begin{cases}\delta(\psi_i^c-\beta_i)\,, & \text{homogeneous}\,,\\ \mathcal{N}(\beta_i,D_{ii})\,, & \text{cell-to-cell variable}\,,\\ \sum_s w_s\,\delta(\psi_i^c-\beta_{s,i})\,, & \text{subpopulation variable}\,,\\ \sum_s w_s\,\mathcal{N}(\beta_{s,i},D_{s,ii})\,, & \text{inter- and intra-subpopulation variable}\,,\end{cases}$$

in which δ denotes the Dirac delta distribution. The posterior distribution for the parameters of cell c, $\boldsymbol{\psi}^c$, is given by

$$p(\boldsymbol{\psi}^c|\bar{\mathbf{y}}^c, \boldsymbol{\Gamma}) \propto p(\bar{\mathbf{y}}^c|\boldsymbol{\psi}^c, \boldsymbol{\Gamma})p(\boldsymbol{\psi}^c)\,,$$

in which $p(\bar{\mathbf{y}}^c|\boldsymbol{\psi}^c, \boldsymbol{\Gamma})$ denotes the likelihood of the single-cell measurement $\bar{\mathbf{y}}^c$ for single-cell parameters $\boldsymbol{\psi}^c$ and noise parameters $\boldsymbol{\Gamma}$. The likelihood is $p(\bar{\mathbf{y}}^c|\boldsymbol{\psi}^c, \boldsymbol{\Gamma}) = \mathcal{N}(\bar{\mathbf{y}}^c|\mathbf{y}^c, \boldsymbol{\Gamma})$ for additive normally distributed measurement noise and $p(\bar{\mathbf{y}}^c|\boldsymbol{\psi}^c, \boldsymbol{\Gamma}) = \log\mathcal{N}(\bar{\mathbf{y}}^c|\mathbf{y}^c, \boldsymbol{\Gamma})$ for multiplicative log-normally distributed measurement noise.

4.2 Evaluation of the hierarchical population model

We evaluated the capabilities of the proposed hierarchical population model for three simulation examples. For a simple conversion, we compared our model with RRE-constrained mixture modeling (2.17), assessed its ability to detect causal sources of variability and reconstruct the latent single-cell trajectory. Furthermore, for a model of stochastic gene expression, we analyzed the incorporation of intrinsic noise in the framework. For a model of protein expression, we assessed the model for multivariate data.

4.2.1 Unraveling sources of variability

To demonstrate the advantages of the hierarchical population model, which incorporates a mechanistic description of the means and variances, over RRE-constrained mixture modeling (2.17), we applied our approach to simulated data on a simple conversion process (3.8). The conversion process comprised two species A and B, with cell-to-cell variable conversions from B to A (Figure 4.4A), corresponding to different levels of phosphatase in the cells. Two subpopulations were assumed with different responses to stimulus u. This produced subpopulations with different rates of stimulus-dependent conversion from A to B. Artificial measurement noise was added to allow the capability of the framework to distinguish measurement noise from biological variability to be assessed.

The RRE for $(x_1, x_2) = ([\mathrm{A}], [\mathrm{B}])$ is given by

$$\begin{aligned}\dot{x}_1 &= k_3x_2 - (k_1u + k_2)x_1\,,\\ \dot{x}_2 &= (k_1u + k_2)x_1 - k_3x_2\,,\end{aligned}$$

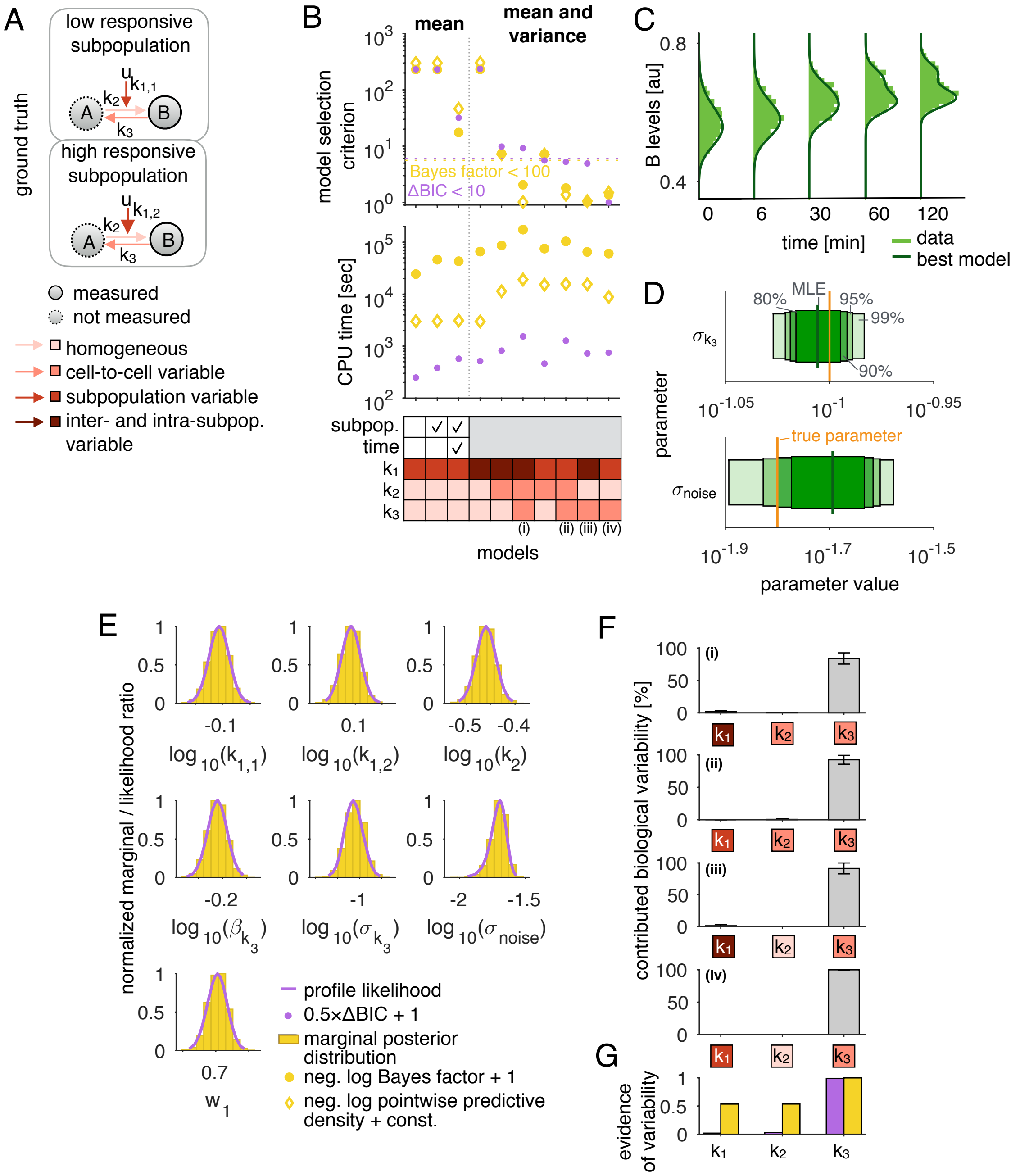

Figure 4.4: Inference of cell-to-cell variability using mechanistic models. Figure caption on next page.

Figure 4.4: Inference of cell-to-cell variability using mechanistic models. (A) Model of a conversion between two species A and B comprising two subpopulations differing in their response to stimulus u. Different colors indicate the variability of the reaction rates. (B) Model selection criteria and required computation times for all models. Lower values indicate a higher evidence for the corresponding model. The horizontal dotted lines indicate the cutoff corresponding to a BIC difference of 10 and a Bayes factor of 100. (C) Data on the conversion process (1000 cells per time point) and fit corresponding to the best and true underlying model. (D) CIs for the variability of k_3 and the measurement noise (σ_{noise}). Horizontal bars show the CIs corresponding to the 80%, 90%, 95%, and 99% confidence levels, and the vertical lines the MLE. (E) Normalized marginal posterior distribution computed from samples of the posterior distribution and likelihood ratio obtained by profile likelihoods for all parameters. (F) Contribution to overall cell-to-cell variability of the observable for the models with Bayes factor $<$ 100. The errorbars indicate deviation over time points. (G) Evidence for variability in parameters computed based on BIC weights (left, purple) and marginal likelihoods (right, yellow). This figure is a modified version of Figures 3 and S1 of the author's publication (Loos et al., 2018b)

with initial conditions

$$x_1(0) = \frac{k_3}{k_2}, \quad x_2(0) = 1 - \frac{k_3}{k_2},$$

accounting for mass conservation [A] + [B] = 1 and the assumption that the system was in steady state before the stimulus was added at 0 min. We assumed the conversion from B to A to be cell-to-cell variable,

$$k_3 \sim \log\mathcal{N}(\beta_{k_3}, \sigma_{k_3}^2), \tag{4.15}$$

yielding cell-to-cell variable initial conditions. The parameter k_1 was considered to differ between subpopulations and therefore was parametrized by $k_{1,1}$ and $k_{1,2}$. The weight w_1 indicated the proportion of the low responsive subpopulation. We generated artificial data for the parameters

$$\begin{aligned}\boldsymbol{\theta}^{\text{true}} &= (k_{1,1}, k_{1,2}, k_2, \beta_{k_3}, \sigma_{k_3}, \sigma_{\text{noise}}, w_1)\\ &= (10^{-0.1}, 10^{0.1}, 10^{-0.45}, 10^{-0.2}, 10^{-1}, 10^{-1.8}, 0.7).\end{aligned}$$

We observed the concentration of B, i.e., $y = x_2$. The data was created including 1000 cells at five time points for $u = 1$ by sampling from the distribution for k_3 (4.15) and simulating the corresponding RREs. Of the 1000 cells, 700 cells belonged to subpopulation 1 with low response to stimulation and 300 cells to the high responsive subpopulation 2. Additionally, the measurements of both subpopulations were assumed to be subject to logarithmic multiplicative measurement noise parametrized by σ_{noise}.

We assumed the parameters $\boldsymbol{\theta}$ to be unknown and estimated them from the data with

(i) RRE-constrained mixture modeling (2.17) using the means and

(ii) the hierarchical population model describing the means and covariances (obtained by the sigma-point approximation).

For both approaches, the underlying subpopulation structure was given, i.e., subpopulation variability of k_1.

Hierarchical model using RREs We considered a hierarchical model with subpopulation means that were described by the RRE. The distribution of the observables was assumed to be log-normal and the scale parameters were estimated from the data. This approach does not model the temporal evolution of the variance, requiring different parametrizations to be compared, i.e., constant, time-dependent, and time/subpopulation-dependent variability. We distinguished the following scenarios:

- one scale parameter that is shared across time points and subpopulations,
- one scale parameter for every subpopulation, which is shared between time points,
- 10 scale parameters that differ for each subpopulation and time-point.

These scale parameters were estimated along with $k_{1,1}, k_{1,2}, k_2, k_3$, and w_1 for this setting, which corresponds to the RRE-constrained mixture modeling (2.17). For optimization, the kinetic parameters k_i were assumed to be in the interval $[10^{-3}, 10^3]$, the weight w_1 in $[0, 1]$, and the scale parameters for the log-normal distribution were restricted to the interval $[10^{-2}, 10^2]$. For each model we performed 50 multi-starts at randomly drawn initial points.

Hierarchical model using sigma-point approximations For the hierarchical population model, the parameter vector for subpopulation s was given by $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \mathbf{D}_s)$ with

$$\boldsymbol{\beta}_s = \begin{pmatrix} k_{1,s} \\ k_2 \\ \beta_{k_3} \\ \sigma_{\text{noise}} \end{pmatrix} \begin{array}{l} \text{subpopulation variable}\,, \\ \text{homogeneous}\,, \\ \text{cell-to-cell variable}\,, \\ \text{homogeneous}\,, \end{array}$$

and

$$D_{s,ij} = \begin{cases} \sigma_{k_3}^2 \,, & \text{for } i = j = 3 \,, \\ 0 \,, & \text{otherwise} \,. \end{cases}$$

To describe the introduced cell-to-cell variability in k_3 (4.15), we used the sigma-point approximation for the log-parameters. For optimization, the dynamic parameters or their means (in case of cell-to-cell variability) were assumed to be in the interval $[10^{-3}, 10^3]$, the scale parameters σ_{k_i} and measurement noise σ_{noise} in $[10^{-3}, 10^2]$ and the weight w_1 in $[0, 1]$. As for the RRE model, we performed 50 multi-starts. For sampling and to facilitate a comparison of frequentist and Bayesian approaches, we considered uniform prior distributions.

We performed **model selection** for a range of hypotheses, including the number of variance parameters when using RREs, as well as additional cell-to-cell variability when using sigma-points. We employed BIC (2.34), log marginal likelihoods and log pointwise predictive density (2.37) (Figure 4.4B). The log marginal likelihood was determined using thermodynamic integration with the Simpsons' rule (Section 2.4.1). The log pointwise predictive density was determined by sampling the posterior distribution for a subset of the data, for the measurements for all but one time point, and computing the logarithm of the average likelihood on the remaining data. The log pointwise predictive density was robustly evaluated using the expressions in (4.13). The comparison of BIC values, log marginal likelihoods and log pointwise predictive densities revealed a good agreement. The Spearman's rank correlation coefficient between BICs and log marginal likelihoods is $r = 0.98$, and $r = 0.83$ between BICs and log pointwise predictive densities.

Model selection indicates that different parameters for each subpopulation at every time point are required to describe the data (Figure 4.4B). This demonstrates that the observed cell-to-cell variability changes over time but provides no information about the sources of the observed cell-to-cell variability. The mechanistic modeling of multiple levels of heterogeneity facilitates the prediction of its causal source via model selection. We considered a range of hypotheses and tested all possible combinations of cell-to-cell variability in $k_{1,s}, k_2$ or k_3. The sigma-point approximation was applied to the logarithm of the observable, to link the mean and variance of the simulation directly to the distribution parameters of the log-normal distribution. The case of no additional cell-to-cell variability corresponds to the RRE models and is therefore not covered here. All criteria suggest the rejection of the models which include only the mechanistic description of the mean but not the variance. For the remaining models the methods provided a sightly different ordering, but all of them indicate the importance of the variability of k_3. Interestingly, model complexity seems to be more penalized by the BIC. Given the subpopulation structure, the additional source of heterogeneity, namely, the conversion from B to A, was correctly predicted and the corresponding model provided a good fit to the data (Figure 4.4C). The model selection criteria for most of the hierarchical models were substantially better than that of the best model that incorporates only the mean. This confirms that a mechanistic description of the variability is more appropriate.

For **uncertainty analysis**, we first calculated CIs obtained using profile likelihoods (2.25) for the best model (Figure 4.4D). We analyzed the ability of the hierarchical model to predict the different contributions of cell-to-cell variability and measurement noise, as both are normally present in single-cell experiments. The uncertainty analysis suggested that the hierarchical modeling approach was able to distinguish between the two.

Furthermore, we evaluated the reliability of the CIs obtained using profile likelihoods. Therefore, we sampled the posterior distribution of the ground truth model using the parallel tempering algorithm implemented in the parameter estimation toolbox PESTO. The chains were initialized at the MLEs and their convergence was assessed using the Geweke test (Geweke, 1992). The comparison of the marginal posterior distributions and the profile likelihoods revealed an excellent agreement (Figure 4.4E). We note that the initialization of the parallel tempering algorithm using a sample from the prior instead of using the pre-computed MLEs, yielded substantially longer computation times and often did not result in a converged chain for $2 \cdot 10^5$ iterations (corresponding to roughly 4 CPU hours). This indicates that for this problem optimization is an important step.

As the model selection did not reject all models but the ground truth model, we evaluated the contribution of the variability of individual parameters to the variability of the observable. Therefore, we determined the reduction of the variability of the observable achieved by removing the variability in the parameter of interest. This analysis was performed for samples from the posterior distribution (Figure 4.4E). We performed this analysis for the models which cannot be rejected based on a Bayes factor cutoff of 100 (Table 2.1) and found that the main contribution to the variability clearly comes from variability in k_3. This means that even for plausible models which account for additional variability in k_1 or k_2, the main source of variability is k_3. To confirm this further, we computed the BIC weights for a certain variability by summing the BIC weights (2.39) for all models accounting for this variability. To detect the source of variability, we took the models for all possible combinations into account. Similarly, we calculated the evidence of a variability based on the computed marginal likelihoods. Both approaches agree in the presence of variability in k_3, confirming the agreement of the results. The BIC weights for the parameters k_1 and k_2 are higher than the evidences computed from the log marginal likelihoods, which, however, do not have a big contribution to the overall variability (Figure 4.4F&G).

4.2.2 Single-cell calibration

To evaluate the predictive power of the method for single-cell trajectories, we inferred the parameters of individual cells from the single data point available for each cell in combination with the calibrated hierarchical population model as a prior (Section 4.1.3). We found that the information about the behavior of a single-cell encoded in the measurement of the first time point was limited, e.g., the prediction is off (Figure 4.5A). However, using data from late time points, we obtained a good estimate of the (latent) single-cell trajectory (Figure 4.5B). The predictions of the trajectories for 100 single-cells from measurements at time point $t = 120\,\mathrm{min}$ (Figure 4.5C) reveal a correlation between true and predicted values > 0.9 for all but early time points.

4.2.3 Including intrinsic and extrinsic noise sources

To study the possibility of accounting for intrinsic noise in the hierarchical population model, we generated artificial data of a two stage gene expression (Figure 4.6A) using the SSA. The system comprises the following reactions

$$
\begin{aligned}
&\mathrm{R}_1: && \emptyset \to \mathrm{mA}\,, && \text{rate} = k_1\,,\\
&\mathrm{R}_2: && \emptyset \to \mathrm{mA}\,, && \text{rate} = uk_2\,,\\
&\mathrm{R}_3: && \mathrm{mA} \to \emptyset\,, && \text{rate} = k_3[\mathrm{mA}]\,,\\
&\mathrm{R}_4: && \mathrm{mA} \to \mathrm{A}\,, && \text{rate} = k_4[\mathrm{mA}]\,,\\
&\mathrm{R}_5: && \mathrm{A} \to \emptyset\,, && \text{rate} = k_5[\mathrm{A}]\,.
\end{aligned}
$$

Here, mA denotes the mRNA and A the protein, and we assumed that only A could be observed. The two subpopulations differed in their response to stimulus u yielding different rate constants $k_{2,1}$ and $k_{2,2}$. For this setting, we only accounted for homogeneous and subpopulation variable parameters. However, the intrinsic variability of the production and degradation of individual molecules gave cell-to-cell variability in the cellular states.

The ODEs for the temporal evolution of the means and covariances were provided by the toolbox CERENA (Kazeroonian et al., 2016). In particular, the means m_1 and m_2 and the variances C_{11} and C_{22} of mRNA mA and protein A, respectively, were described as

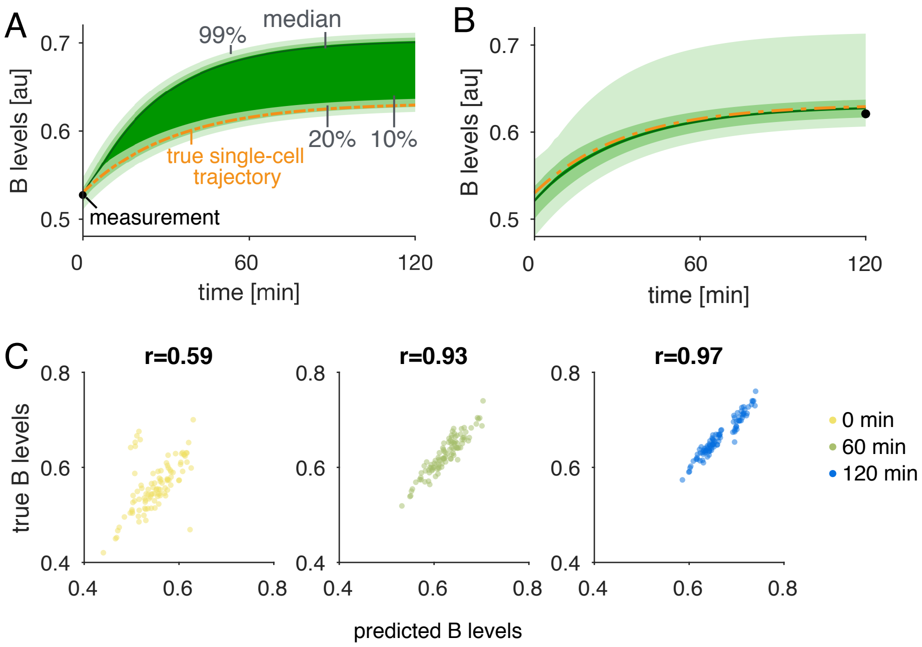

Figure 4.5: Single-cell calibration. (A) Single-cell trajectories inferred using a single measurement at (A) t=0 min and (B) t=120 min. The inference is regularized using the hierarchical population model as prior. Shaded areas indicate the CIs which were evaluated for samples of the posterior distribution and the dotted line indicates the single-cell trajectory from which the measurement point was generated. (C) Correlation of predicted and true level of B at 0, 60 and 120 min. True values were extracted from the (noise-free) simulation. Predictions are obtained using the single-cell data at time t=120 min. This figure is adapted from Figure 3E-G of the author's publication (Loos et al., 2018b).

well as the correlation C_{12} of mA and A. The ODE system reads

$$\begin{aligned}
\dot{m}_1 &= \frac{k_1}{\Omega} + \frac{uk_2}{\Omega} - k_3 m_1\,,\\
\dot{m}_2 &= k_4 m_1 - k_5 m_2\,,\\
\dot{C}_{11} &= \frac{k_1}{\Omega^2} + \frac{uk_2}{\Omega^2} - 2C_{11}k_3 + \frac{k_3 m_1}{\Omega}\,,\\
\dot{C}_{12} &= C_{11}k_4 - C_{12}(k_3 + k_5)\,,\\
\dot{C}_{22} &= 2C_{12}k_4 - 2C_{22}k_5 + \frac{k_4 m_1}{\Omega} + \frac{k_5 m_2}{\Omega}\,,
\end{aligned}$$

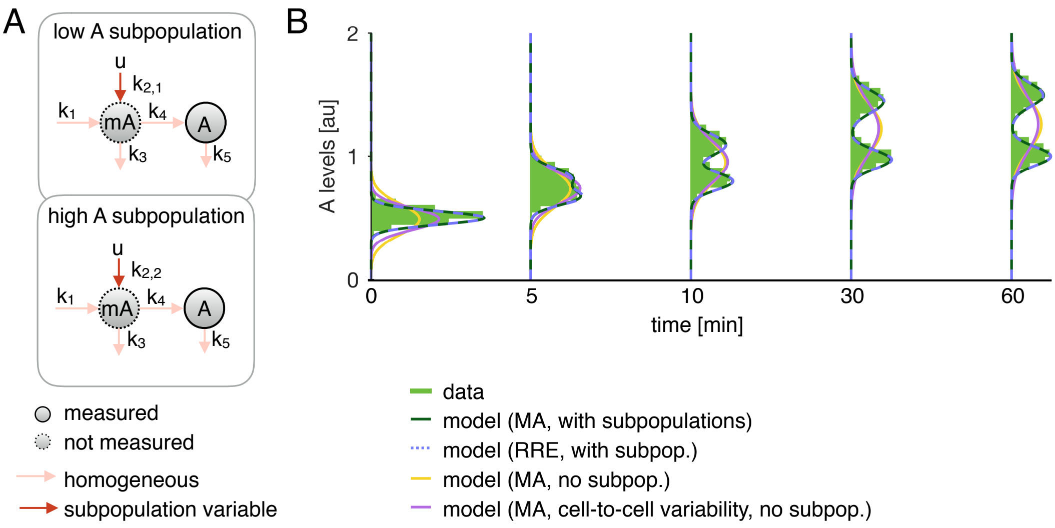

Figure 4.6: Incorporation of intrinsic noise. (A) Illustration of the system. (B) Data and fitted models for the MA, for the case of accounting for subpopulation structures and disregarding subpopulation structures, and RREs. This figure is adapted from Figure S6A-B of the author's publication (Loos et al., 2018b).

with system size $\Omega = 1000$. Under the assumption that the system was in steady state before stimulation with u, the initial conditions are

$$\begin{aligned}
m_1(0) &= \frac{k_1}{\Omega k_3}\,, \\
m_2(0) &= \frac{k_1 k_4}{\Omega k_3 k_5}\,, \\
C_{11}(0) &= \frac{k_1}{\Omega^2 k_3}\,, \\
C_{12}(0) &= \frac{k_1 k_4}{\Omega^2 k_3 (k_3 + k_5)}\,, \\
C_{22}(0) &= \frac{1}{\Omega^2}\left(\frac{k_1 k_4}{k_3 + k_5} + \frac{k_1 k_4^2}{k_3 k_5 (k_3 + k_5)}\right).
\end{aligned}$$

The true parameters used for the generation of the data were

$$\begin{aligned}
\boldsymbol{\theta}^{\text{true}} &= (k_1, k_{2,1}, k_{2,2}, k_3, k_4, k_5, w_1) \\
&= (10, 10, 20, 1, 5, 0.1, 0.5)\,.
\end{aligned}$$

In this example, we employed mixtures of normal distributions, for which the mean and variance were linked to the distribution parameters by $\boldsymbol{\mu}_s = \mathbf{m}_s$ and $\boldsymbol{\Sigma}_s = \mathbf{C}_s$. First, we compared a model accounting for the mean, which was obtained by the RRE-constrained

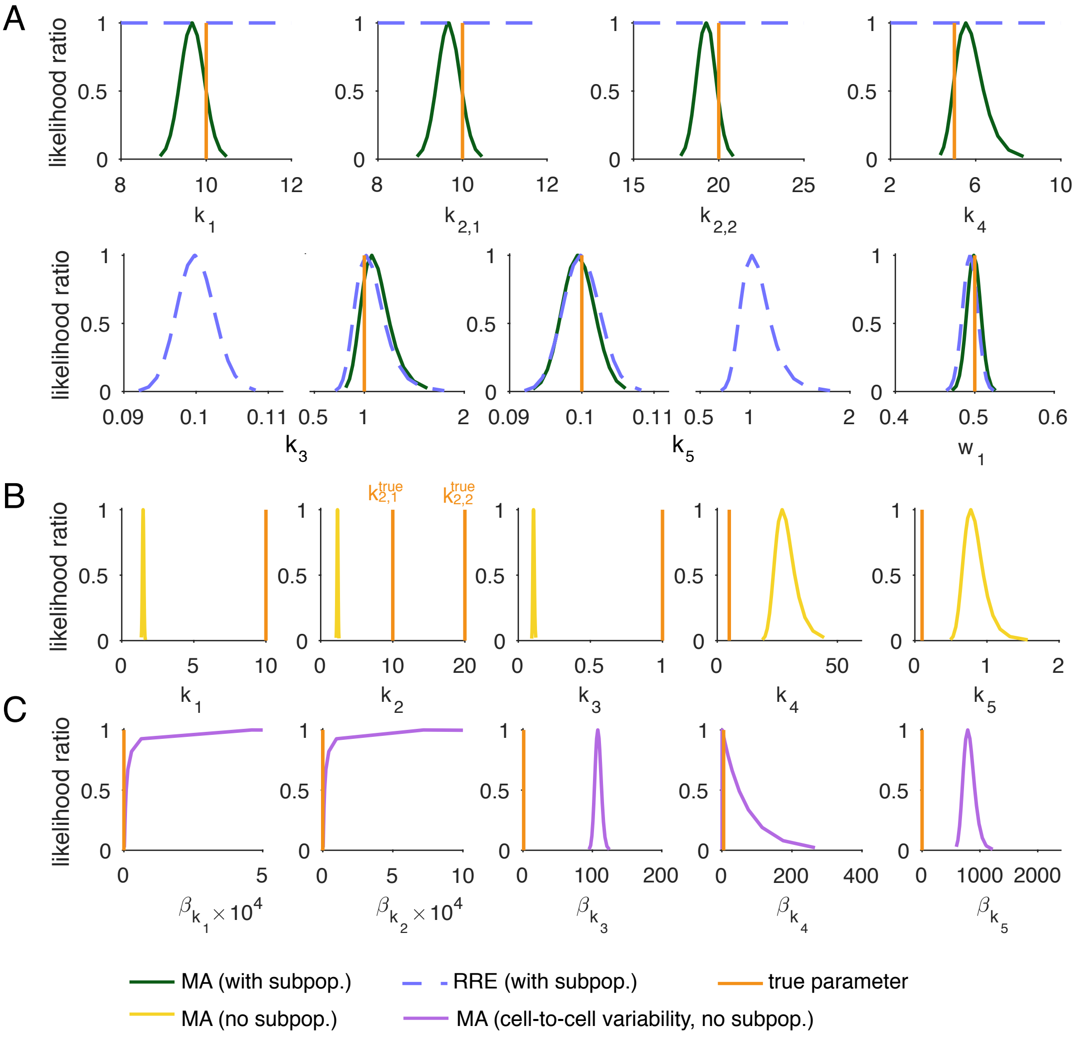

Figure 4.7: Uncertainty analysis for intrinsic noise. (A) Profile likelihoods of the parameters for the models capturing the subpopulation structure. (B) Profile likelihoods of the parameters for the model using the MA without accounting for subpopulations. (C) Profile likelihoods of the mean rate constants for the model using the MA, accounting for cell-to-cell variability of all rate constants but not for subpopulations. This corresponds to the method proposed by Zechner et al. (2012). Note that the range in x-direction differs for subplots (A)-(C). This figure is adapted from Figure S6E-G of the author's publication (Loos et al., 2018b).

mixture model (2.17), and a hierarchical model accounting for the mean and covariances, which were obtained by the MA (2.14, 2.15), both accounting for two subpopulations. For the RRE model 10 parameters for the parametrization of the variances were introduced, yielding in total $n_\theta = 17$. The model using the MA only comprised $n_\theta = 7$, since a mechanistic description of the variances was incorporated. For parameter estimation, the kinetic parameters were restricted to the interval $[10^{-3}, 10^{3}]$ and the $\log_{10}$-transformed parameters were fitted, whereas the weight w_1 was restricted to $[0, 1]$ and fitted linearly. For the RRE model, the parameters for the variance were assumed to lie within $[10^{-4}, 10^{2}]$ and also fitted in $\log_{10}$-space. Second, we studied two models that incorporate the mechanistic description of the variance by the MA, but did not consider the presence of two subpopulations (MA, no subpop.). One of these models, however, accounts for cell-to-cell variability of each parameter (MA, cell-to-cell variability, no subpop.), which corresponds to the description by Zechner et al. (2012).

The models not accounting for subpopulation structures did not fit the data at all (Figure 4.6B). Even the included variability in parameters did not improve the fits substantially. In contrast, both subpopulation models provided a good fit to the data. However, the BIC for the MA model was substantially better than for the RRE model ($\mathrm{BIC}_{\mathrm{RRE}} - \mathrm{BIC}_{\mathrm{MA}} = 79.09$).

Furthermore, we studied the uncertainty of the parameter estimates using profile likelihoods (Figure 4.7). Using the MA with subpopulations, all parameters were identifiable, indicated by a narrow profile. This was not the case for RREs, for which some parameters could not be identified from the data and showed a flat profile. For the case of no subpopulations, most of the true parameters did not lie within the estimated intervals (Figure 4.7B&C). This emphasizes the importance of taking into account subpopulation structures.

4.2.4 Accounting for correlations in multivariate measurements

Many single-cell technologies provide multivariate measurements and therefore convey information about the correlations between the observables. To incorporate this, we extended our hierarchical modeling framework to multivariate data and demonstrated its capability to reconstruct the differential protein expression of cellular subpopulations (Kharchenko et al., 2014; Sauvageau et al., 1994) using simulated data. We considered a model describing the abundance of two proteins, the expression of which is regulated by stimulus u (Figure 4.8A). The influence of u varies between cell populations and is therefore able to capture, e.g., different levels of membrane receptors. We generated multivariate data by simulating a single-cell model.

The simple model of differential protein expression considers six reactions

$$\begin{aligned}
&\mathrm{R}_1: \quad && \emptyset \to \mathrm{A}\,, \quad && \text{rate} = k_1\,,\\
&\mathrm{R}_2: \quad && \emptyset \to \mathrm{B}\,, \quad && \text{rate} = k_1\,,\\
&\mathrm{R}_3: \quad && \emptyset \to \mathrm{A}\,, \quad && \text{rate} = k_2 u\,,\\
&\mathrm{R}_4: \quad && \emptyset \to \mathrm{B}\,, \quad && \text{rate} = k_3 u\,,\\
&\mathrm{R}_5: \quad && \mathrm{A} \to \emptyset\,, \quad && \text{rate} = k_4[\mathrm{A}]\,,\\
&\mathrm{R}_6: \quad && \mathrm{B} \to \emptyset\,, \quad && \text{rate} = k_4[\mathrm{B}]\,,
\end{aligned}$$

comprising the basal expression with rate k_1, degradation with rate constant k_4 and stimulus-induced expression, depending on u, with rate constants k_2 and k_3 for protein A and B, respectively. The corresponding ODE system for the temporal evolution of $(x_1, x_2) = ([\mathrm{A}], [\mathrm{B}])$ is

$$\begin{aligned}
\dot{x}_1 &= k_1 + k_2 u - k_4 x_1\,,\\
\dot{x}_2 &= k_1 + k_3 u - k_4 x_2\,,
\end{aligned}$$

with initial conditions

$$x_1(0) = x_2(0) = \frac{k_1}{k_4}\,,$$

obtained by assuming that the system was in steady state before the stimulus was added at 0 min. Two subpopulations were assumed, one showing high expression of A while the other showed high expression of B after stimulation with u. The degradation rate constant k_4 was considered to be cell-to-cell variable,

$$k_4 \sim \log\mathcal{N}(\beta_{k_4}, \sigma_{k_4}^2)\,, \tag{4.16}$$

with median β_{k_4} and scale σ_{k_4} which were equal between the subpopulations. The measurements were exposed to log-normally distributed multiplicative measurement noise parametrized by σ_{noise}.

The hierarchical model accounted for the subpopulation variability of k_2 and k_3 and the cell-to-cell variability of k_4. This yielded the subpopulation parameters

$$\boldsymbol{\beta}_s = \begin{pmatrix} k_1 \\ k_{2,s} \\ k_{3,s} \\ \beta_{k_4} \\ \sigma_{\text{noise}} \end{pmatrix} \begin{matrix} \text{homogeneous}\,, \\ \text{subpopulation variable}\,, \\ \text{subpopulation variable}\,, \\ \text{cell-to-cell variable}\,, \\ \text{homogeneous}\,, \end{matrix}$$

$$D_{s,ij} = \begin{cases} \sigma_{k_4}^2\,, & \text{for } i = j = 4\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Using our hierarchical approach confirmed the ability of the proposed model to reproduce the data (Figure 4.8B) and to provide reliable parameter estimates (Figure 4.8C). Such multivariate data cannot be exploited by existing model-based approaches. When the temporal evolution of proteins is measured individually, the correlation information is missing and a symmetry arises in the system (Figure 4.8D). This is reflected in the multimodal profiles of the parameters $k_{2,1}$ and $k_{2,2}$, indicating a lack of practical identifiability.

Our framework exploits the correlation structures of multivariate data, which in this simulation example allowed us to conclude that each subpopulation had a high expression of only a single protein. This only becomes possible when the correlations are analyzed.

4.3 Application example: Pain sensitization

To assess the modeling framework in a real application setting, we studied pain sensitization in primary sensory neurons. These neurons are highly heterogeneous cells which are involved in pain sensitization. In this section, we studied signal transduction in the extracellular-signal regulated kinase (Erk) pathway, a signaling cascade that is involved in a range of biological processes. The specific focus of this section is pain sensitization in response to nerve growth factor (NGF) stimulation (Andres et al., 2012; Hucho and Levine, 2007; Ji et al., 2009). The neurons encounter a broad range of extracellular environments, including various extracellular scaffolds, and are highly heterogeneous. Performing single-cell microscopy experiments (Andres et al., 2010; Isensee et al., 2014), we investigated the influence of extracellular scaffolds on the response of individual subpopulations.

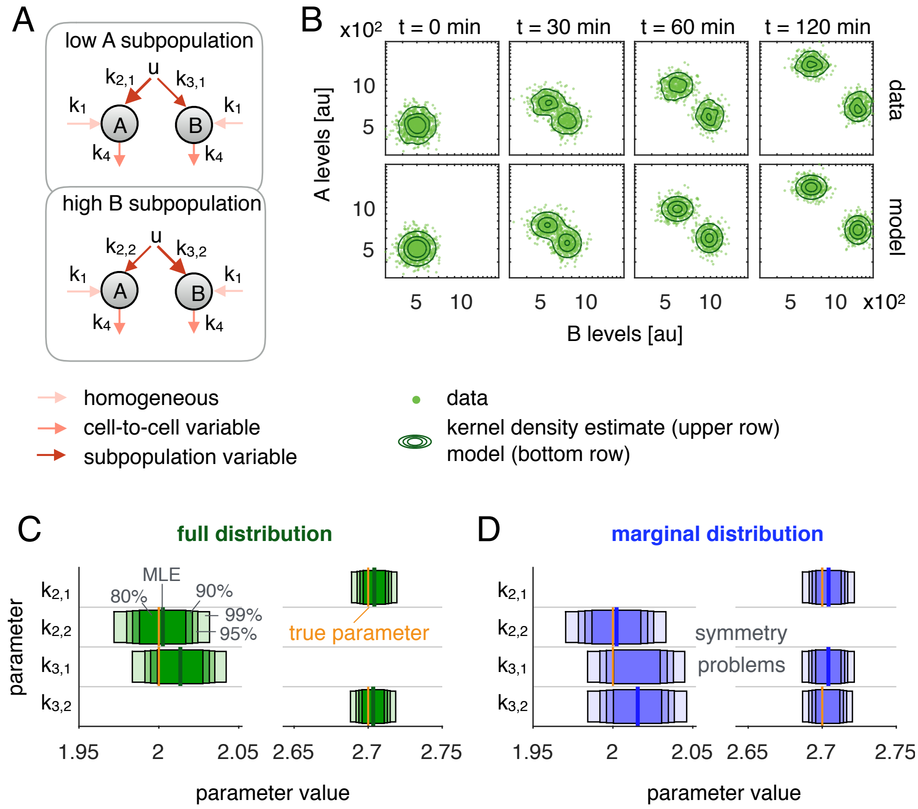

Figure 4.8: Reconstruction of differential protein expression in heterogeneous populations using multivariate data. (A) Model of differentially expressed proteins A and B. (B) Upper row: data points (1000 cells per time point) and kernel density estimation. Lower row: data points and model for the full distribution. (C,D) CIs for the parameters of the model using (C) the full distribution and (D) the marginal distributions. Horizontal bars show the CIs corresponding to the 80%, 90%, 95%, and 99% confidence levels. The vertical lines show the MLE. This figure is a modified version of Figure 4 of the author's publication (Loos et al., 2018b).

4.3.1 NGF-induced Erk signaling in primary sensory neurons

We applied the hierarchical modeling approach to investigate the influence of an extracellular scaffold on NGF-induced Erk1/2 activation in cultures of adult sensory neurons (Figure 4.9A). This was done by monitoring the rates of NGF-mediated Erk1/2 phosphorylation in dissociated cultures of the primary sensory neurons of rat dorsal root ganglia. NGF-mediated Erk1/2 signaling has been shown to play a crucial role in nociceptor sensitization in thermal and mechanical hyperalgesia (Malik-Hall et al., 2005; Zhuang et al., 2004). Primary sensory neurons form a heterogeneous population, from which, upon NGF stimulation, a subpopulation reacts with a graded Erk1/2 phosphorylation response. Previous models have attempted to approximate this by assuming the existence of responders and non-responders with differing levels of the NGF receptor TrkA (Hasenauer et al., 2014). Here, we refined this substantially by modeling the overall population using two heterogeneous subpopulations that differed in their average response. To calibrate this refined model, we collected quantitative single-cell microscopy data on NGF-induced Erk1/2 phosphorylation kinetics and dose response curves using immunofluorescence labeling of pErk1/2 alone, co-labeled with Erk1/2 and TrkA antibodies (see STAR Methods of (Loos et al., 2018b) for more details on the experimental setup).

We employed the model proposed by Hasenauer et al. (2014), which comprises the reactions

$$\begin{aligned}
&\mathrm{R}_1: && \mathrm{TrkA} + \mathrm{NGF} \to \mathrm{TrkA{:}NGF}\,, && \text{rate} = k_1[\mathrm{TrkA}][\mathrm{NGF}]\,,\\
&\mathrm{R}_2: && \mathrm{TrkA{:}NGF} \to \mathrm{TrkA} + \mathrm{NGF}\,, && \text{rate} = k_2[\mathrm{TrkA{:}NGF}]\,,\\
&\mathrm{R}_3: && \mathrm{Erk} \to \mathrm{pErk}\,, && \text{rate} = k_3[\mathrm{TrkA{:}NGF}][\mathrm{Erk}]\,,\\
&\mathrm{R}_4: && \mathrm{Erk} \to \mathrm{pErk}\,, && \text{rate} = k_4[\mathrm{Erk}]\,,\\
&\mathrm{R}_5: && \mathrm{pErk} \to \mathrm{Erk}\,, && \text{rate} = k_5[\mathrm{pErk}]\,.
\end{aligned}$$

Conservation of mass yields

$$\begin{aligned}
[\mathrm{TrkA}] + [\mathrm{TrkA{:}NGF}] &= [\mathrm{TrkA}]_0\,,\\
[\mathrm{NGF}] + [\mathrm{TrkA{:}NGF}] &= [\mathrm{NGF}]_0\,,\\
[\mathrm{Erk}] + [\mathrm{pErk}] &= [\mathrm{Erk}]_0\,.
\end{aligned}$$

To eliminate structurally non-identifiable parameters, the model was reparametrized to

$$\begin{aligned}
\dot{x}_1 &= k_1[\mathrm{NGF}]_0(k_3[\mathrm{TrkA}]_0 - x_1) - k_2x_1\,, && x_1(0) = 0\,, \\
\dot{x}_2 &= (x_1 + k_4)(s_P[\mathrm{Erk}]_0 - x_2) - k_5x_2\,, && x_2(0) = \frac{k_4 s_P[\mathrm{Erk}]_0}{(k_4+k_5)}\,,
\end{aligned} \tag{4.17}$$

with $x_1 = k_3[\mathrm{TrkA{:}NGF}]$ and $x_2 = s_P[\mathrm{pErk}]$. The observables for the considered experimental conditions are

$$\mathbf{y}_e = \begin{cases} s_{P,e}[\mathrm{pErk}] + o_{P,e}\,, & e = 1,2\,,\\[2ex] (s_{P,e}[\mathrm{pErk}] + o_{P,e}, s_T[\mathrm{TrkA}]_0 + o_T)^T\,, & e = 3\,,\\[2ex] (s_{P,e}[\mathrm{pErk}] + o_{P,e}, s_E[\mathrm{Erk}]_0 + o_E)^T\,, & e = 4\,, \end{cases}$$

to compare the subpopulations on poly-D-lysine (PDL). This comprises pErk1/2 kinetics ($e = 1$), pErk1/2 dose response ($e = 2$), pErk/TrkA dose response ($e = 3$), and pErk/Erk dose response ($e = 4$). Furthermore, the observables to study the effects of the extracellular scaffolds PDL and collagen type I (Col I) on the neurons (PDL: $e = 1, 3, 5, 7$, Col I: $e = 2, 4, 6, 8$) are:

$$\mathbf{y}_e = \begin{cases} s_{P,e}[\text{pErk}] + o_{P,e}\,, & e = 1,2,3,4\,, \\ (s_{P,e}[\text{pErk}] + o_{P,e}, s_T[\text{TrkA}]_0 + o_T)^T\,, & e = 5,6\,, \\ (s_{P,e}[\text{pErk}] + o_{P,e}, s_E[\text{Erk}]_0 + o_E)^T\,, & e = 7,8\,. \end{cases}$$

This comprises pErk1/2 kinetics ($e = 1, 2$), pErk1/2 dose response ($e = 3, 4$), pErk/TrkA dose response ($e = 5, 6$), and pErk/Erk dose response ($e = 7, 8$).

The pErk1/2, TrkA and Erk1/2 levels could only be measured up to some scaling constants denoted by s_P, s_T and s_E, respectively, and with some offsets denoted by o_P, o_T and o_E. Each observable was assumed to be subject to multiplicative log-normally distributed measurement noise parametrized by $\sigma_{P,e,\text{noise}}, \sigma_{T,\text{noise}}$ and $\sigma_{E,\text{noise}}$. For the comparison of the extracellular scaffold, the same scaling, offset, and measurement noise parameters were used for PDL and Col I. For each subpopulation, we used the sigma-point approximation accounting for cell-to-cell variability in cellular TrkA activity and Erk1/2 levels. The covariance between TrkA activity and relative Erk1/2 expression was parametrized, accounting for correlations, with the matrix logarithm parametrization $M(\sigma_T, \sigma_E, \sigma_{TE}) \in \mathbb{R}^{2\times 2}$. All other entries of $\mathbf{D}_s$ were assumed to be 0.

4.3.2 Differences between subpopulations

Differences between the responses of responders and non-responders are likely caused by the expression of the receptor corresponding to the stimulus. In case of NGF the activation of the TrkA receptor leads classically to Erk1/2 sensitization signaling. Thus we first validated our modeling approach by predicting causal differences between subpopulations and its accordance with described differences in TrkA expression. We used experimental kinetic and dose response data from sensory neurons cultured on the adherence substrate PDL.

We accounted for all possible combinations of subpopulation variability of k_1, k_2, k_4, k_5, $k_3[\text{TrkA}]_0$, and $s_{P,e}[\text{Erk}]_0$. This yielded in total $2^6 = 64$ models that were tested, ranging from $n_\theta = 26$ parameters, for the model assuming no subpopulations at all, to $n_\theta = 33$ parameters, assuming that the subpopulations differ in all parameters. To take into account all hierarchical models, we considered the BIC weights (2.39). The BIC-based ranking scheme suggested that cellular TrkA activity ($k_3[\text{TrkA}]_0$) made the greatest contribution (Figure 4.9B). This was indicated by a high BIC weight and the substantially better mean rank of the models using differences in cellular TrkA activity compared with those using other differences. Using a model including only additional subpopulation variability of TrkA expression levels produced an excellent fit to the experimental data (Figures 4.9C). This difference was also confirmed experimentally in the cultures (Figure 4.9D).

The potential differences which follow cellular TrkA activity are the relative Erk1/2 expression levels ($s[\text{Erk}]_0$) and the dephosphorylation rate constant (k_5). However, our experimental data showed no statistically significant difference in total Erk1/2 levels between responders and non-responders (Figure 4.9E). To assess the relevance of the dephosphorylation rate and thus the corresponding phosphatase activity, we performed experiments in which we monitored the pErk1/2 decline dynamics after inhibiting the mitogen-activated protein kinase (Mek) that phosphorylates Erk1/2. If the phosphatase activity varies, we would expect to observe different equilibration dynamics. To validate whether the two subpopulations differ in their dephosphorylation/phosphotase activity (parametrized by k_5), we inhibited cells with the Mek-inhibitor U0126 (10 μM). NGF binds to the TrkA+ subpopulation and activates pErk1/2 signaling, whereas GDNF binds to the Ret receptor on the opposing subpopulation (TrkA-) and yields pErk1/2 signaling in this neuronal subgroup. Cells were pre-stimulated for 1 h with the combined stimuli NGF (20 ng/ml) and GDNF (100 ng/ml) to obtain responses in both subpopulations. We measured pErk1/2 levels to obtain the dynamics of the dephosphorylation as well as TrkA levels to distinguish the two subpopulations. Cells were considered to belong to the TrkA+ subpopulation if their intensity was above 670 and to the TrkA- subpopulation if their intensity was below 630. The measurements were taken every 3 minutes between 0 and 37 min for four replicates.

To obtain the de-phosphorylation rate constant k_5, we normalized the values of pErk1/2 to 1 at $t =$0 min and 0 at $t_{\text{max}} = 37$ min. We fitted an exponential decay,

$$E(t) = E_c \exp(-k_5 t) + E_o \,,$$

to the scaled data of the four replicates. The scaling E_c and offset E_o could be determined from the boundary conditions

$$E(0) = 1 \quad \text{and} \quad E(t_{\text{max}}) = 0 \,.$$

This yielded four values for the de-phosphorylation in the TrkA+ subpopulation and in the TrkA- subpopulation. A two-sample t-test with Welch's correction gave a p-value of 0.62, indicating that the dephosphorylation rates in the two subpopulations were not significantly different (Figures 4.9F).

We compared the results of model selection by BIC and log pointwise posterior density. This was done for the models accounting for no or one difference between the subpopulations. We considered this reduced set of models for the comparison, as the sampling for the calculation of the log pointwise predictive density and the calculation of the Bayes factors took on average 780 CPU hours per model. The BIC values, the log pointwise posterior density and the Bayes factors strongly prefer the model accounting for differing TrkA levels over all other models ($\Delta\text{BIC} > 7\cdot 10^3$). We found that the log pointwise posterior density is sensitive to the splitting of the data set, with smaller test and training data sets preferring less complex models. The results in Figure 4.9G are shown for splitting the data set in two parts, which gave a rank correlation of $r = 0.61$. The Bayes factors even yielded a rank correlation of $r = 1$, indicating that the Bayes factors are indeed well approximated by the BIC for these models.

The final model accounted for subpopulation differences in cellular TrkA activity and also took into account differences in the variance of TrkA activity between the subpopulations. The fits for parts of the data are visualized in Figure 4.9C. Using the final calibrated model, we predicted the relation between pErk1/2 levels at 0 and 120 min by drawing parameters from the inferred single-cell parameter distribution and simulating the ODE model (Figure 4.9H).

In summary, this analysis of subpopulation structures demonstrates that the hierarchical approach using experimental data provided an appropriate ranking of differences which could be demonstrated experimentally and is in line with literature (reviewed in, e.g., (Mantyh et al., 2011)).

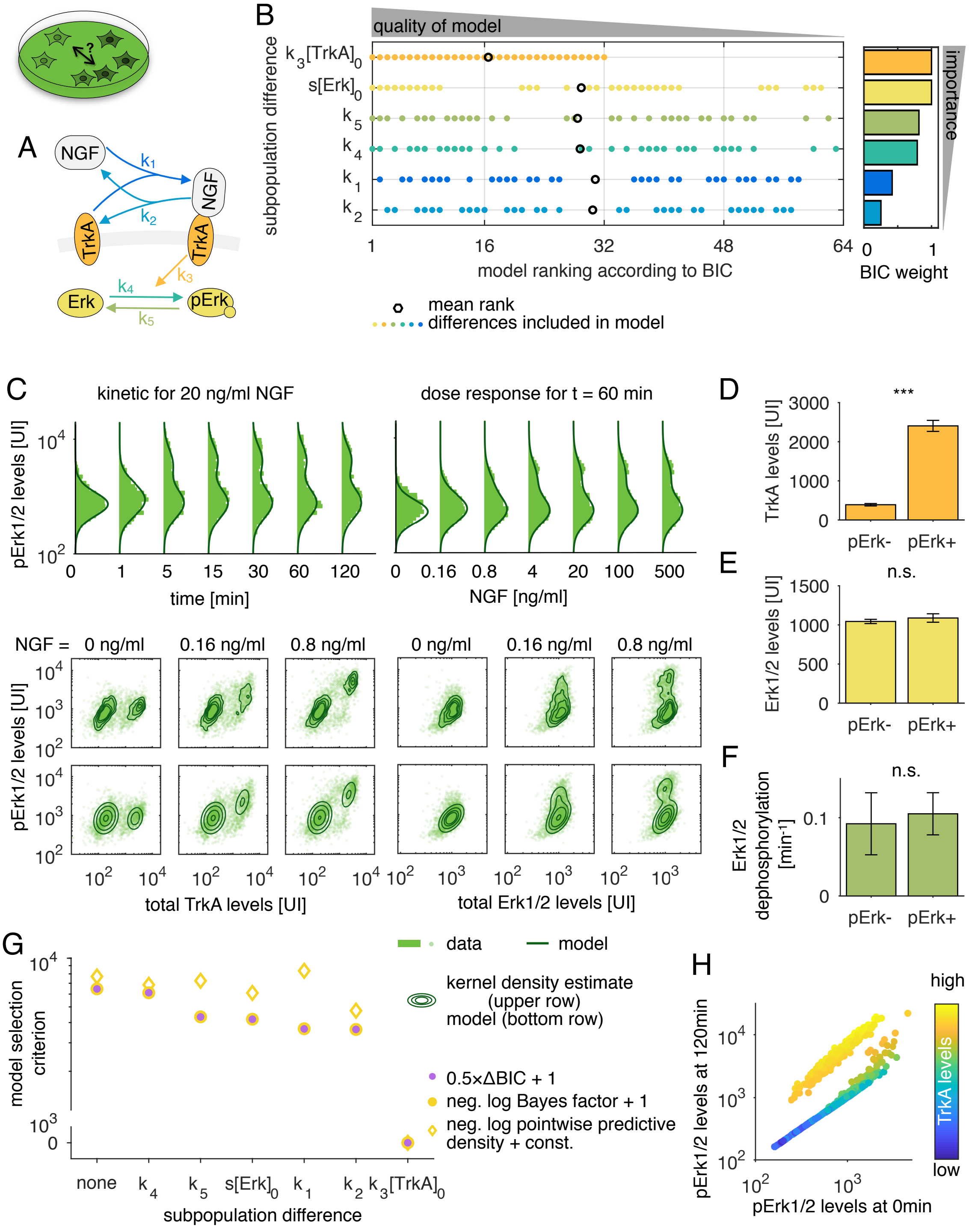

Figure 4.9: Sources of heterogeneity between subpopulations in primary sensory neurons. Figure caption on next page.

Figure 4.9: Sources of heterogeneity between subpopulations in primary sensory neurons. (A) Pathway model of NGF-induced Erk signaling. (B) Ranking according to the BIC values for the 64 hierarchical models, in which the colored dots indicate those parameters that are assumed to differ between the subpopulations. The importance of the differences is ranked according to the BIC weights. The black circles indicate the mean rank of the models including the corresponding difference. (C) Data and fit for measurements of pErk1/2 levels (approximately 1400 cells per time point and 4300 cells per dosage) and multivariate measurements of pErk/TrkA and pErk/Erk levels (approximately 3000 cells per dosage) measured for 60 min under NGF stimulation with indicated concentrations. The measured values are in arbitrary units of intensity. For the multivariate data, the contour lines of the kernel density estimation of the data and the level sets of the density of the hierarchical model are shown. Mean and standard deviation of (D) TrkA levels ($n_r = 4$ replicates) (E) Erk1/2 levels ($n_r = 4$) and (F) Erk1/2 dephosphorylation ($n_r = 4$) of non-responsive (pErk-) and responsive (pErk+) sensory neurons after NGF stimulation with varying concentrations (as indicated in (C) for 60 min). (G) Comparison of models, which account for one or no difference between subpopulations, using BIC, log pointwise predictive densities and Bayes factors. (H) Predicted single-cell trajectories for the optimal parameter values, showing the relation between pErk1/2 levels in steady state (0 min) and after stimulation with NGF (120 min). The color of the cells indicates the TrkA level, which is assumed to be constant over time. This figure is a modified version of Figures 5 and S4B&E of the author's publication (Loos et al., 2018b).

4.3.3 Differences mediated by extracellular scaffolds

Even though matrix molecules have been investigated for their impact on signaling pathways underlying neurite outgrowth (Chen et al., 2007; Myers et al., 2011), much less is known about the role of cell scaffolds in sensitization and thus sensitization signaling of nociceptive neurons. To approach this, we compared the modification of the well-described NGF initiated sensitization signaling pathway by the two example scaffolds, Col I, a classical extracellular matrix protein that forms receptor-matrix interactions, and by PDL, a neutral scaffolding that promotes cell adherence by electrostatic interaction. We determined the kinetics and dose response curves of NGF-induced Erk1/2 phosphorylation in sensory neurons cultured overnight on Col I or PDL. We found that the mean Erk1/2 activation was approximately 17 % higher in Col I compared to PDL after NGF treatment (Figure 4.10A for pErk1/2 dose responses). In addition to showing increased NGF-induced Erk1/2 activation, the number of cells was observed to be 1.5 times lower in the collagen cultures than in the PDL cultures. These observations raised questions about the source of the measured increase in mean NGF-mediated Erk1/2 activation. We considered two hypotheses: (i) the increase results from a biological action of the different scaffolds onto the neurons; and (ii) the increase reflects a shift of the subpopulation sizes arising from a nonrandom loss of parts of the high-responder subpopulation due to reduced cell adherence in the collagen cultures.

To unravel the causal differences between the primary sensory neurons cultured on PDL and on Col I, we used the model which assumes subpopulation differences in TrkA levels. The model for each adherence substrate accounted for the cell-to-cell variability of Erk1/2 and the inter- and intra-subpopulation variability of cellular TrkA activity. The differences between the extracellular scaffolds were parametrized as

$$\kappa_{k_1}, \kappa_{k_2}, \kappa_{k_4}, \kappa_{k_5}, \kappa_{\beta_{k_3[\mathrm{TrkA}]_0}}, \kappa_{\beta_{c[\mathrm{Erk}]_0}}, \kappa_w \,,$$

and the parameters were related by

$$k_{1,\mathrm{Col\,I}} = k_{1,\mathrm{PDL}} 10^{\kappa_{k_1}} \,.$$

Accounting for these seven potential differences, we defined 128 hierarchical models. Each model was fitted to the data with multi-start local optimization using at least 20 starts. We sorted the models with respect to their BIC value. The BIC weights for the differences were computed by summing over the BIC weights (2.39) of the models accounting for the corresponding differences. The model ranked first by the BIC (Figures 4.10B) gave a good fit to the data and suggested differences not only in cellular TrkA activity ($k_3[\mathrm{TrkA}]_0$) but also in Erk1/2 expression ($s[\mathrm{Erk}]_0$), and Erk1/2 dephoshorylation (k_5) (Figures 4.10C&D). These differences were assumed to explain the higher response on Col I, and therefore supported hypothesis (i). The model that assumed no difference between the extracellular scaffolds (rank 128) or changes only in the relative size of the subpopulations (rank 127) performed worst, indicating that hypothesis (ii) failed to explain the data. These results confirmed the model-based analysis and suggested an impact of the classical extracellular matrix protein Col I on protein expression. Indeed, the differences in relative TrkA and Erk1/2 expression levels could be confirmed (Figures 4.10E&F).

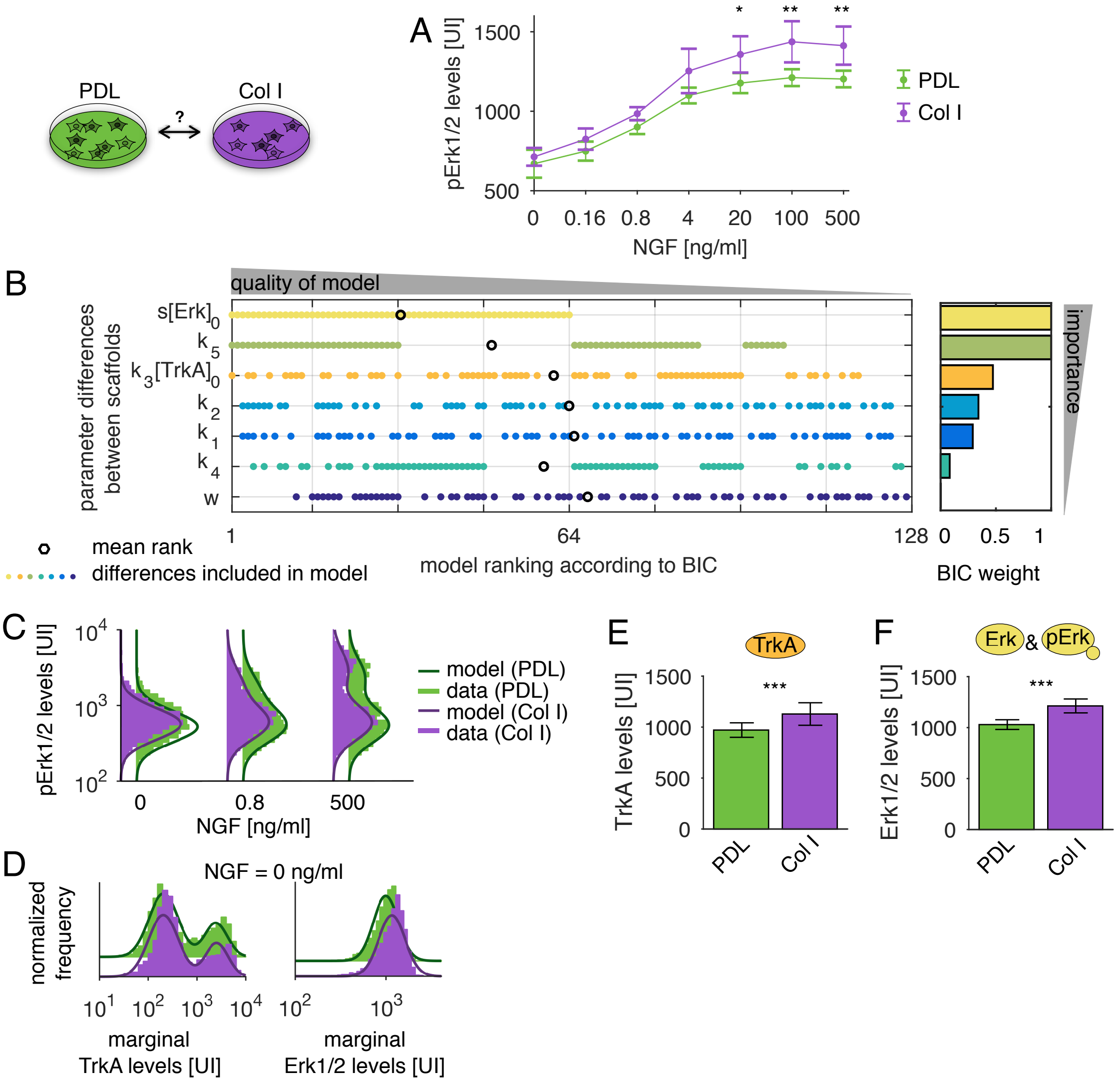

Figure 4.10: Differences in NGF-induced Erk1/2 phosphorylation mediated by different extracellular scaffolds. Primary sensory neurons were provided with the two different scaffolds PDL and Col I in an overnight culture. (A) Sensory neurons grown on the Col I substrate showed a significantly higher mean phospho-Erk1/2 response to indicated doses of NGF after 1 h of stimulation. Means and standard deviations of four replicates are shown. (B) BIC-based ranking for the potential differences between culture conditions. The colored dots indicate which parameters are assumed to differ between the extracellular scaffolds. (C) Experimental data and fit for measurements of pErk1/2 distributions from Col I (approximately 2300 cells per dosage) and PDL (approximately 4300 cells per dosage) cultured neurons after treatment with indicated NGF concentrations for 1 h. (D) Marginal levels for TrkA and Erk1/2, which were assumed to be constant over varying doses and time (approximately 2000 cells in Col I and 2900 in PDL). Mean and standard deviation of (E) TrkA and (F) Erk1/2 levels of NGF dose response curve data, which showed significant elevations in Col I treated neurons. For this calculation, 24 samples were used (4 replicates for 6 doses). This figure is adapted from Figure 6 of the author's publication (Loos et al., 2018b).

4.4 Robust calibration of hierarchical population models

In Section 3.1, we demonstrated the importance of incorporating robust distributions when studying population average data. For single-cell snapshot data, the probability of outlier in the data is even higher due to the high number of data points (Pyne et al., 2009). However, the results obtained in Section 3.1 cannot directly be transferred to population models due to specificities of the different data and model types.

In this section, we provide the likelihood functions for several distribution assumptions which can be incorporated into the hierarchical population modeling framework. In particular, we investigate the choice of distribution ϕ in (4.9). For each distribution, we derive the function g_φ defined in (4.3), which maps the mean and covariances of the species to the distribution parameters $\boldsymbol{\varphi}$. We assess the influence of the distribution assumption for simulated data of the models studied in Section 4.2 and the experimental data studied in Section 4.3.2.

4.4.1 Distribution assumptions

The **multivariate normal distribution** (4.4, 4.12) and its incorporation in the hierarchical population modeling framework were introduced in Section 4.1.1. This distribution has the parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n_y \times n_y}$ and was used for the analyses in previous Sections 4.2 and 4.3.

For the incorporation of the **multivariate skew normal**, we followed the definition of Pyne et al. (2009). The distribution has the parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$, with location $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$, skew parameters $\boldsymbol{\delta} \in \mathbb{R}^{n_y}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n_y \times n_y}$. The probability density function is

$$\phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = 2\phi_{n_y}(\bar{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}(\bar{\mathbf{y}} - \boldsymbol{\mu}))\,,$$

with $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\delta}\boldsymbol{\delta}^T$ and $\boldsymbol{\alpha} = \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}/(1 - \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\delta})^{\frac{1}{2}}$, ϕ_{n_y} denoting the multivariate normal density with n_y dimensions, and Φ denoting the cumulative distribution function of a univariate standard normal distribution. The log-density function is given by

$$\log \phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = \log(2) + \log(\phi_{n_y}(\bar{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\Omega})) + \log(\Phi(\boldsymbol{\alpha}(\bar{\mathbf{y}} - \boldsymbol{\mu})))\,.$$

Assuming that the distribution parameters depend on parameter vector $\boldsymbol{\theta}$, the gradient is given by

$$\begin{aligned}
\frac{\partial \log \phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial \theta_i} &= \frac{\partial \phi_{n_y}(\bar{\mathbf{y}}|\boldsymbol{\mu}(\boldsymbol{\theta}),\boldsymbol{\Omega}(\boldsymbol{\theta}))}{\partial \theta_i} \frac{1}{\phi_{n_y}(\bar{\mathbf{y}}|\boldsymbol{\mu}(\boldsymbol{\theta}),\boldsymbol{\Omega}(\boldsymbol{\theta}))} + \\
&\quad \frac{1}{\Phi(\boldsymbol{\alpha}(\boldsymbol{\theta})(\bar{\mathbf{y}}\boldsymbol{\mu}(\boldsymbol{\theta})))} \phi_1(\boldsymbol{\alpha}(\boldsymbol{\theta})(\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))) \cdot \\
&\quad \left(-\boldsymbol{\alpha}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} + (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})) \frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \theta_i} \right),
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \left[\boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \underbrace{\left(1 - \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{1}{2}}}_{:=a(\boldsymbol{\theta})} \right] \\
&= \left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i} \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} + \boldsymbol{\delta}(\boldsymbol{\theta})^T \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i} \right) a(\boldsymbol{\theta}) + \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \frac{\partial a(\boldsymbol{\theta})}{\partial \theta_i}, \\
\frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i} &= -\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})}{\partial \theta_i} \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \\
&= -\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \sqrt{\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T} \right)^{-1} \left(\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i} + \frac{1}{2} \left(\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T\right)^{-\frac{1}{2}} \cdot \right. \\
&\quad \left. \left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i} \boldsymbol{\delta}^T + \boldsymbol{\delta}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i} \right) \right) \cdot \left(\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \sqrt{\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T} \right)^{-1}, \\
\frac{\partial a(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{1}{2} \left(1 - \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{3}{2}} \frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i} \\
&= \frac{1}{2} \left(1 - \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{3}{2}} \left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i} \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \boldsymbol{\delta}(\boldsymbol{\theta}) + \right. \\
&\quad \left. \boldsymbol{\delta}(\boldsymbol{\theta})^T \left(\frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i} \boldsymbol{\delta}(\boldsymbol{\theta}) + \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i} \right) \right).
\end{aligned}$$

The derivative of the multivariate normal density is given in (4.12). Following Pyne et al. (2009) and Sahu et al. (2003), the mean and covariance matrix of the multivariate skew normal distribution are given by

$$\begin{aligned}
\mathbf{m} &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \boldsymbol{\delta}, \\
\mathbf{C} &= \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right) \boldsymbol{\delta}\boldsymbol{\delta}^T.
\end{aligned}$$

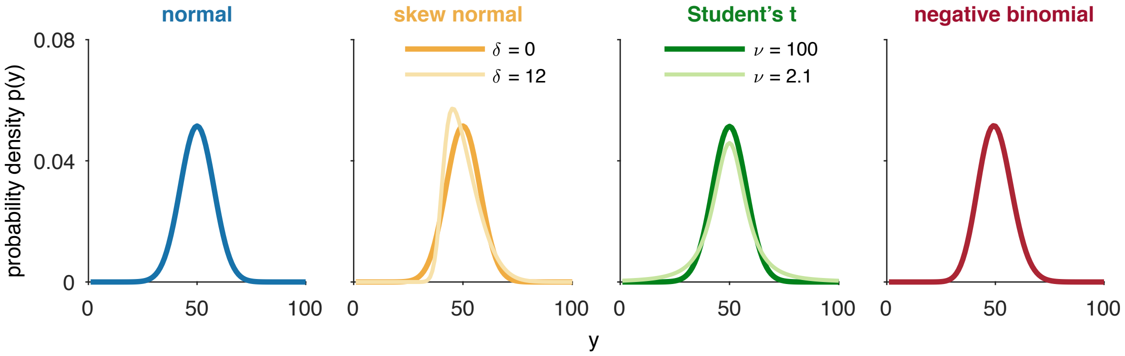

Figure 4.11: Robust distributions for the hierarchical population model. The visualized distributions (normal, skew normal, Student's t and negative binomial) all have mean $\mathbf{m} = 50$ and variance $\mathbf{C} = 60$.

This yields for $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta}), \boldsymbol{\delta}(\boldsymbol{\theta}))$ the relation

$$\begin{aligned}\boldsymbol{\mu}_s(\boldsymbol{\theta}) &= \mathbf{m}_s^y(\boldsymbol{\theta}) - \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}(\boldsymbol{\theta})\,,\\ \boldsymbol{\Sigma}_s(\boldsymbol{\theta}) &= \mathbf{C}_s^y(\boldsymbol{\theta}) - \left(1 - \frac{2}{\pi}\right)\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T + \boldsymbol{\Gamma}(\boldsymbol{\theta})\,,\end{aligned}$$

with measurement noise matrix $\boldsymbol{\Gamma}$ (4.7). The derivatives are given by

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{m}_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \sqrt{\frac{2}{\pi}}\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\,,\\ \frac{\partial \boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{C}_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \left(1 - \frac{2}{\pi}\right)\left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\boldsymbol{\delta}(\boldsymbol{\theta})^T + \boldsymbol{\delta}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i}\right) + \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\theta})}{\partial \theta_i}\,.\end{aligned}$$

The entries of the skew parameter vector $\boldsymbol{\delta}$ are allowed to be different and are not linked to the simulated moments of the system $\mathbf{m}_s^y$ and $\mathbf{C}_s^y$. However, the entries are restricted in a way that $\boldsymbol{\Sigma}_s$ needs to be positive definite.

The **multivariate Student's t** distribution has distribution parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with location $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$, shape matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n_y \times n_y}$ and degree of freedom $\nu \in \mathbb{R}_+$. The probability density function reads

$$\begin{aligned}\phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}) &= \frac{\Gamma\left(\frac{\nu+n_y}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{n_y}{2}}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{1}{\nu}\mathbf{Z}\right)^{\frac{\nu+n_y}{2}}}\,,\\ \text{with}\quad \mathbf{Z} &= (\bar{\mathbf{y}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})\,,\end{aligned}$$

with log-density function

$$\begin{aligned}\log\phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = &\ \log\Gamma\left(\frac{\nu+n_y}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \log\left(|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\right) - \\ &\frac{n_y}{2}\log(\pi\nu) - \frac{\nu+n_y}{2}\log\left(1+\frac{1}{\nu}\mathbf{Z}\right).\end{aligned}$$

The gradient is given by

$$\begin{aligned}\frac{\partial\log\phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial\theta_i} &= \frac{1}{2}\Bigg(\Bigg(\Psi\left(\frac{\nu(\boldsymbol{\theta})+n_y}{2}\right) - \Psi\left(\frac{\nu(\boldsymbol{\theta})}{2}\right) - \frac{n_y}{\nu(\boldsymbol{\theta})} + \\ &\frac{\mathbf{Z}(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+n_y) - \nu(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta}))\log\left(1+\frac{1}{\nu(\boldsymbol{\theta})}\mathbf{Z}(\boldsymbol{\theta})\right)}{\nu(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta}))}\Bigg)\frac{\partial\nu(\boldsymbol{\theta})}{\partial\theta_i} \\ &- \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_i}\right) - \frac{\nu(\boldsymbol{\theta})+n_y}{\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta})}\frac{\partial\mathbf{Z}(\boldsymbol{\theta})}{\partial\theta_i}\Bigg),\end{aligned}$$

with

$$\begin{aligned}\frac{\partial\mathbf{Z}(\boldsymbol{\theta})}{\partial\theta_i} = &\ (\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i}\right)^T + \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i}\right)^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta})) + \\ &(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}}{\partial\theta_i}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta})),\end{aligned}$$

and digamma function Ψ as in (3.7). For $\nu > 2$, the mean and covariance matrix of the multivariate skew normal distribution are given by

$$\begin{aligned}\mathbf{m} &= \boldsymbol{\mu}, \\ \mathbf{C} &= \frac{\nu}{\nu-2}\boldsymbol{\Sigma}.\end{aligned}$$

This yields $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta}), \nu(\boldsymbol{\theta}))$ with

$$\begin{aligned}\boldsymbol{\mu}_s(\boldsymbol{\theta}) &= \mathbf{m}_s^y(\boldsymbol{\theta}), \\ \boldsymbol{\Sigma}_s(\boldsymbol{\theta}) &= \frac{\nu(\boldsymbol{\theta})-2}{\nu(\boldsymbol{\theta})}\left(\mathbf{C}_s^y(\boldsymbol{\theta}) + \boldsymbol{\Gamma}(\boldsymbol{\theta})\right),\end{aligned}$$

and derivatives

$$\begin{aligned}\frac{\partial\boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial\theta_i} &= \frac{\partial\mathbf{m}_s^y(\boldsymbol{\theta})}{\partial\theta_i}, \\ \frac{\partial\boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial\theta_i} &= \frac{\nu(\boldsymbol{\theta})-2}{\nu(\boldsymbol{\theta})}\left(\frac{\partial\mathbf{C}_s^y(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\theta})}{\partial\theta_i}\right) + \left(\mathbf{C}_s^y(\boldsymbol{\theta}) + \boldsymbol{\Gamma}(\boldsymbol{\theta})\right)\frac{2}{\nu(\boldsymbol{\theta})^2}\frac{\partial\nu(\boldsymbol{\theta})}{\partial\theta_i}.\end{aligned}$$

As for the skewness parameter in the case of the skew normal distribution, the degree of freedom ν is not linked to the simulated moments of the system.

A further distribution assumption which is often employed in the analysis of single-cell data is the **negative binomial** distribution (Grün et al., 2014). For a two stage model of gene expression, the protein number follows a negative binomial distribution, if the ratio of mRNA degradation to protein degradation is high (Shahrezaei and Swain, 2008). This distribution has the parameters $\boldsymbol{\varphi} = (\tau, \rho)$ with $\tau > 0$ and $\rho \in [0, 1]$. In contrast to the other distributions in this section, this distribution is only defined for the one-dimensional case. However, potentially extensions to higher dimensions could be employed (Shi and Valdez, 2014). The probability density function reads

$$\phi(\bar{y}|\boldsymbol{\varphi}) = \binom{\bar{y} + \tau - 1}{\bar{y}} (1-\rho)^{\bar{y}} \rho^{\tau} \,,$$

and the log-density function

$$\log \phi(\bar{y}|\boldsymbol{\varphi}) = \ \log(\Gamma(\bar{y} + \tau)) - \log(\Gamma(\bar{y} + 1)) - \log(\Gamma(\tau)) + \bar{y} \log(1-\rho) + \tau \log(\rho) \,.$$

The derivative of the log-density function is

$$\begin{aligned} \frac{\partial \log \phi(\bar{y}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial \theta_i} = &\left(\Psi(\bar{y} + \tau(\boldsymbol{\theta})) - \Psi(\tau(\boldsymbol{\theta})) + \log(\rho(\boldsymbol{\theta})) \right) \frac{\partial \tau(\boldsymbol{\theta})}{\partial \theta_i} + \\ &\left(\frac{\bar{y}}{1 - \rho(\boldsymbol{\theta})} + \frac{\tau(\boldsymbol{\theta})}{\rho} \right) \frac{\partial \rho(\boldsymbol{\theta})}{\partial \theta_i} \,. \end{aligned}$$

The mean and variance of the negative binomial distribution are

$$\begin{aligned} m &= \frac{(1-\rho)\tau}{\rho} \,, \\ C &= \frac{(1-\rho)\tau}{\rho^2} \,. \end{aligned}$$

Thus, the distribution parameters $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\rho_s(\boldsymbol{\theta}), \tau_s(\boldsymbol{\theta}))$ are mapped to the moments and measurement noise via

$$\rho_s(\boldsymbol{\theta}) = \frac{m_s^y(\boldsymbol{\theta})}{C_s^y(\boldsymbol{\theta}) + \sigma_{\text{noise}}(\boldsymbol{\theta})} \,, \tag{4.18}$$

$$\tau_s(\boldsymbol{\theta}) = \frac{m_s^y(\boldsymbol{\theta})^2}{C_s^y(\boldsymbol{\theta}) + \sigma_{\text{noise}}(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta})} \,. \tag{4.19}$$

If no measurement noise is taken into account, it is required that $C > m$ such that $\tau > 0$. The derivatives of the distribution parameters are

$$\begin{aligned}
\frac{\partial \rho_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{1}{C_s^y(\boldsymbol{\theta}) + \sigma_{\text{noise}}(\boldsymbol{\theta})} \frac{\partial m_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \frac{m_s^y(\boldsymbol{\theta})}{C_s^y(\boldsymbol{\theta}) + \sigma_{\text{noise}}(\boldsymbol{\theta})} \left(\frac{\partial C_s^y(\boldsymbol{\theta})}{\partial \theta_i} + \frac{\partial \sigma_{\text{noise}}(\boldsymbol{\theta})}{\partial \theta_i} \right) \\
\frac{\partial \tau_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{m_s^y(\boldsymbol{\theta})(2C_s^y(\boldsymbol{\theta}) + 2\sigma_{\text{noise}}(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))}{(C_s^y(\boldsymbol{\theta}) + \sigma_{\text{noise}}(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))^2} \frac{\partial m_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \\
&\quad \frac{m_s^y(\boldsymbol{\theta})^2}{(C_s^y(\boldsymbol{\theta})\sigma_{\text{noise}}(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))^2} \left(\frac{\partial C_s^y(\boldsymbol{\theta})}{\partial \theta_i} + \frac{\partial \sigma_{\text{noise}}(\boldsymbol{\theta})}{\partial \theta_i} \right).
\end{aligned}$$

4.4.2 Evaluation of influence of distribution assumptions for simulated data

We simulated univariate data for three different models: (i) conversion process (Figure 4.4A, 4.12A); (ii) two stage gene expression (Figure 4.6A, 4.12B); and (iii) birth-death process (Figure 4.8A, 4.12C). The models (ii) and (iii) are commonly used for the description of gene expression. For each model, we chose three parameter vectors, three numbers of time points and four numbers of cells per time point (50, 100, 500, 1000). This yielded 108 data sets which were simulated using the SSA. The differences in the measurements for individual cells arose solely due to intrinsic noise and no additional measurement noise was added to the data.

To assess the distribution assumption in the hierarchical population model, we generated data for different outlier scenarios (Figure 4.12D):

(i) *no outliers*: no outliers were included in the data.

(ii) *zeros*: the measured concentration at a certain time point t_k is zero, e.g., due to a missing label or entry. Consequently, we measured $\bar{y}_j = 0$.

(iii) *doublets*: two cells were wrongly measured as one cell. To simulate this, the measured value of a random cell was doubled.

(iv) *uniform*: cells were randomly chosen and their measured value assigned to the rounded value of a uniformly distributed value on an interval which is 1.5 times broader than the range $I_{\text{no outlier}}$ of the measurements without outliers.

We assumed 2%, 5% and 10% of the cells to be outliers for scenarios (ii), (iii), and (iv), respectively. We calibrated the hierarchical population models based on all data sets for the different distribution assumptions with 30 multi-starts. For this, we assumed the true underlying source of subpopulation variability to be known and did not allow for measurement noise.

For the model of the birth-death process (Figure 4.12C), the first time point follows a Poisson distribution, for which the mean equals the variance. The MA provides that mean and variance equal the ratio of stimulus-independent synthesis to degradation. Linking the moments for the first time point to the distribution parameters of the negative binomial distribution is thus not possible, since it hurts the restriction that the variance needs to be higher than the mean and results in undefined τ (4.19). The density cannot be evaluated for the first time point without the assumption of additional measurement noise which would increase the variance. Thus, the negative binomial distribution could not be applied to study this model.

In the case of outlier-free data, the best distribution assumption according to the BIC differs between the studied models. For the conversion process and birth-death process, the normal distribution seemed most appropriate, while for the two stage gene expression model the Student's t and the negative binomial distributions were chosen. It is to be expected that the negative binomial distribution fits well for this model, since this process follows a negative binomial distribution in steady state if the ratio between mRNA and protein degradation rate is high (Shahrezaei and Swain, 2008). As soon as outliers were introduced to the data, the Student's t distribution provided most often the best BIC.

The MSE obtained with the Student's t distribution was similarly low as the best obtained MSE and did not change a lot for the outlier-corrupted data sets. However, the MSE obtained for other distributions increased substantially in the presence of outliers.

For the here considered models and data sets, the computation time, convergence and, thus, performance were not influenced by the presence of outliers. For the conversion process, the negative binomial distribution needed the lowest computation time, directly followed by the normal distribution. For the other models, the normal distribution required the lowest computation time. This might be due to the lower dimension of the optimization problem since no additional parameters were estimated from the data such as it is the case for the skew normal and the Student's t distributions. In terms of converged starts, we observed some differences between the considered models. For the conversion process the negative binomial distribution, and for the remaining models the skew normal distribution, provided the lowest number of converged starts, while the normal and Student's t distributions did not seem to suffer from convergence problems. The influence of the outliers is also visualized for example data sets of the outlier scenarios where the normal, skew normal and negative binomial distributions were clearly deviated by the outliers (Figure 4.12E). In this simulation study, the degrees of freedom of the models were limited since no measurement noise was incorporated in the population models.

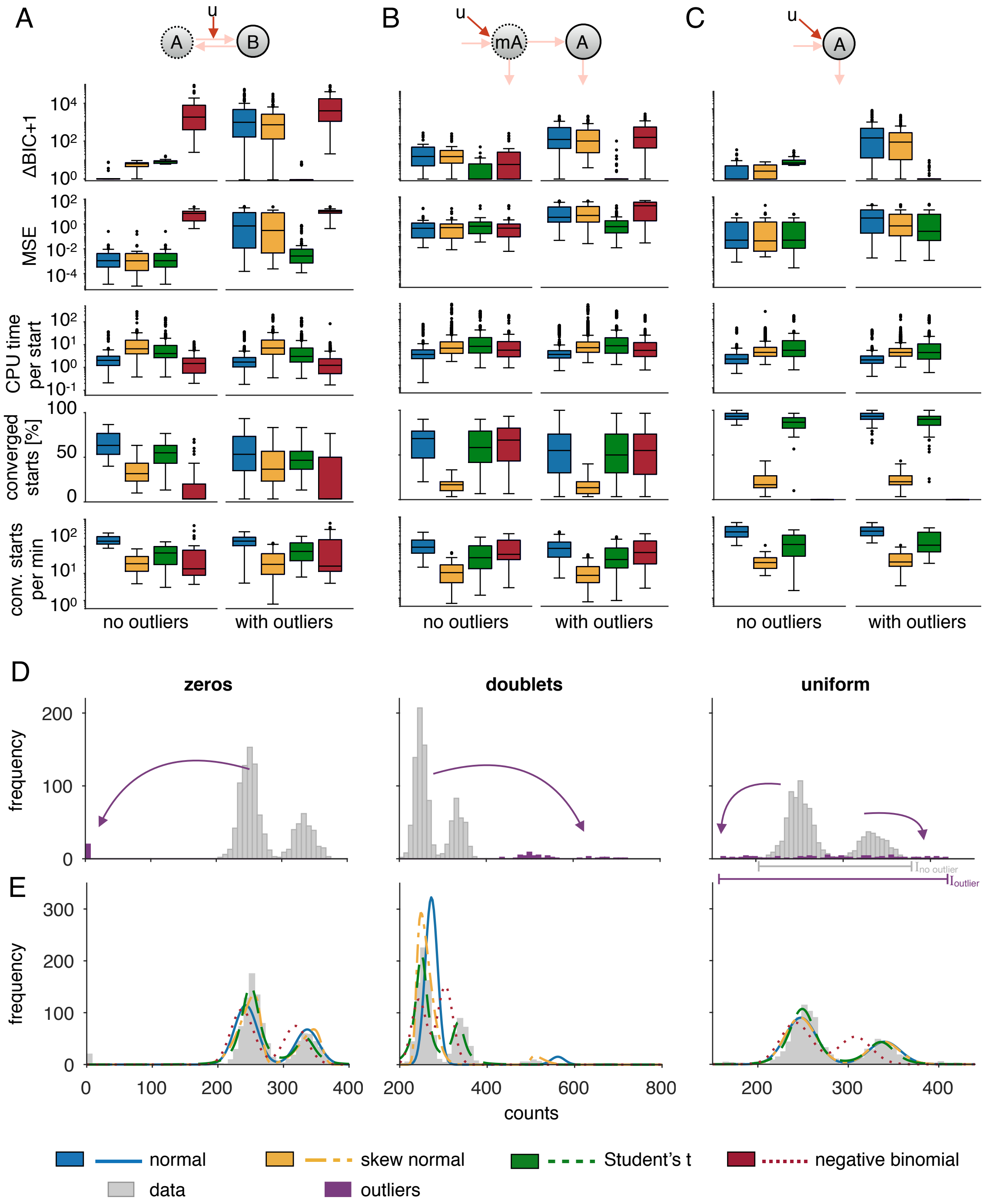

Figure 4.12: Robust distributions within the hierarchical population modeling framework. (A-C) Comparison of BIC values, MSE, CPU time per optimization start, number of converged starts and performance for the distribution assumptions for the models of a (A) conversion process, (B) two stage gene expression and (C) birth-death process. We considered outlier-free and outlier-corrupted data. Each boxplot showing the CPU time per optimization start has 1080 points (36 data sets and 30 multi-starts) and the other boxplots comprise 36 points. (D) Outlier scenarios for single-cell snapshot data illustrated for example data sets of a conversion process with (E) corresponding fits obtained by different distribution assumptions.

4.4.3 NGF-induced Erk signaling

To test the distributions in a real application setting, we reanalyzed the data and the hierarchical population model introduced in Section 4.3.2. For these data, we cannot apply the negative binomial distribution, since the distribution requires integer valued measurements and is only defined for univariate measurements. In this section, we used log-transformed simulations and data. Calibrating the hierarchical population models using the normal (as in Section 4.3.2), the skew normal and the Student's t distributions, we found that for the univariate data of the pErk1/2 kinetics, the model fits cannot visually be distinguished (Figure 4.13A). However, for the bivariate data, the skew normal and the Student's t distributions fit the data better (Figure 4.13B). We cannot assess the MSE since the true parameters are not known. Visualizing the likelihood waterfall plots (Figure 4.13C) and analyzing the performance of the optimizations (Figure 4.13D), we found that the Student's t distribution substantially outperformed the other distributions. Interestingly, the skew normal distribution which showed bad performance for the simulation study has here a comparable performance as the normal distribution. The skew normal distribution provided the best likelihood value. However, it also has the highest number of parameters since for the bivariate measurements each dimension is allowed to have different skewness parameters.

4.5 Summary and discussion

Elucidating the causes of cellular heterogeneity is a challenging task in systems biology and requires appropriate mechanistic models for the use with single-cell data. In this chapter, we proposed a hierarchical modeling framework that for the first time allowed different levels of heterogeneity to be investigated, including subpopulation structures and cell-to-cell variability within subpopulations. Beyond cell-to-cell variability, the method accounts for measurement noise and is able to deconvolute these sources.

This modeling approach unifies available mechanistic modeling and inference frameworks (Hasenauer et al., 2014; Zechner et al., 2012), complements available statistical methods and exploits efficient simulation methods for cellular subpopulations. The proposed method facilitates the integration and simultaneous analysis of multiple data sets, without requiring complex pre-processing of the data (Lee et al., 2011).

Differences between cell types can be analyzed and modeled in the same manner as differences between cellular subpopulations. The method is also able to handle more than two subpopulations, and the number of subpopulations can even be inferred using a data-driven

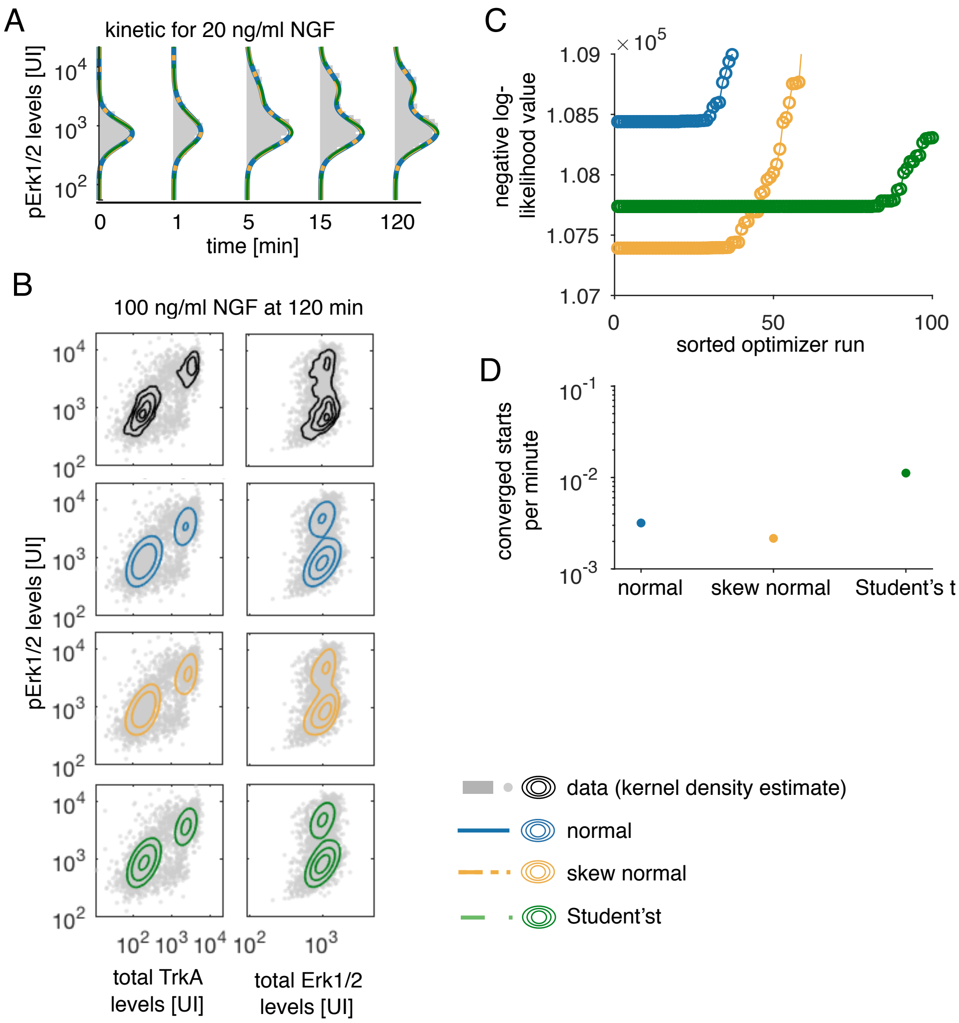

Figure 4.13: Robust distributions for NGF-induced Erk signaling. (A,B) Data and model fits for (A) univariate measurements of pErk1/2 levels and (B) bivariate measurements for pErk/TrkA and pErk/Erk levels. (C) Likelihood waterfall plot for the three different distribution assumptions. The best 80 values are shown and in total 500 multi-starts were performed. (D) The performance of the optimization measured as number of converged starts per minute.

approach. Procedures such as a forward-backward algorithm (Section 2.4.2) or reversible jump Markov Chain Monte Carlo (Green, 1995) could be implemented to simultaneously perform parameter estimation and model selection.

In this thesis, we incorporated various distributions to model the cell population. We found that the normal distribution assumption was appropriate when not many outliers are to be expected. The negative binomial distribution did not provide appropriate results for all scenarios, is highly restricted by the relation between mean and variance and can only describe univariate data. A trivial extension to allow for multivariate measurements would be the product distribution, which neglects correlations. Multivariate extensions which account for correlations could be incorporated (Shi and Valdez, 2014). If the data is outlier-corrupted, robust alternatives such as the Student's t distribution might be more reasonable. This distribution also provides reliable results when the data is outlier-free and could be considered as a default distribution assumption. If more information is available about the precise type of outliers, e.g., that they arise due to dropout events in single-cell RNA-seq data, computational methods can be adapted accordingly (Eraslan et al., 2019; Pierson and Yau, 2015). While the Student's t distribution suffered from problems of over-fitting in the case of population-average data (Chapter 3), the number of measurements in single-cell data sets is usually much higher and we do not expect to face the same problems as observed in Section 3.1.2. To allow for different degrees of freedom in multivariate measurements, a t copula could be employed (Luo and Shevchenko, 2010). Also, a skewed version of the Student's t distribution as, e.g., used by Pyne et al. (2009) could be incorporated.

The inference of mechanistic models from single-cell data relies on statistical models for the measurement and sampling process. In many modeling studies using single-cell data, no distinction is made between cells from different batches, obscuring cell-to-cell variability and differences between experimental batches (Hicks et al., 2017). We observed that model selection is often biased towards complex models. To circumvent this issue, we used a ranking of potential differences rather than a precise measure of statistical significance. However, this problem will need to be addressed, as the use of single-cell data is increasingly common.

In summary, we proposed the use of hierarchical population models as a novel tool to study heterogeneity in multivariate single-cell data and evaluated their performance. Our framework is the first to account for multiple levels of heterogeneity simultaneously. Our results on simulation and application examples suggested that this method can be used to obtain a more holistic understanding of cellular heterogeneity.

Chapter 5

Summary and conclusion

Each step in the building, calibration and comparison of mathematical models for studying biological processes faces different challenges depending on the biological questions and data types considered. In this thesis, we covered all these steps and addressed the research questions posed in Section 1.2. These are the methodological questions of robust and efficient modeling and model calibration for population averages and single-cells as well as the biological questions of histone methylation and pain sensitization.

In Chapter 3 we addressed the problems of outlier-occurrence and large number of parameters for ODE models which are calibrated to population average data. We used distributions with heavier tails than the Gaussian distribution and derived the gradients and Hessians of the corresponding likelihood functions to enable an efficient optimization. The alternative distributions allowed for robust estimation results for our considered outlier scenarios. This provided the first comprehensive evaluation of alternative distributions in the biosciences, where specific experimental errors and outliers occur. The Cauchy and Student's t distributions faced problems with over-fitting when the sample size is too small. However, we found that the Laplace distribution provides a good trade-off between robustness and computational complexity. For the Gaussian and Laplace distributions, we were able to employ a hierarchical scheme for optimization, where we calculate the optimal scaling and noise parameters, which do not contribute to the dynamics of the system, analytically. The hierarchical approach to optimization yielded a substantial increase in optimizer performance, measured in terms of the number of converged starts as well as computation time required. Furthermore, the approach can also be used to calculate profile likelihoods of the parameters. We evaluated the proposed methods for real experimental data of histone modifications. We built models which represent different hypothesis about the mechanisms of histone H3 methylation and calibrated these models employing Laplace noise and hierarchical optimization. We found that a model assuming that the histone tails are only methylated up to a predefined final state seems more appropriate.

In Chapter 4 we provided a framework for mechanistically studying heterogeneous subpopulations based on single-cell snapshot data. For this, we proposed the hierarchical population model which unifies existing modeling techniques and facilitates the inclusion of different types of heterogeneity. This includes cellular variability between and within subpopulations. Similarly to Section 3.1, we also provided the equations to incorporate alternative distribution assumptions to the Gaussian distribution and assessed their influence on optimization results and performance. Since the individual data types are rather specific, the results obtained in Chapter 3 could not directly be transferred to the results on single-cell data. Indeed, we found that for our examples, the Student's t distribution did not suffer from the problems of over-fitting which occurred for population average data. We also evaluated the methods proposed in Chapter 4 for real experimental data, in particular, for data of NGF-induced Erk signaling in primary sensory neurons. Performing model selection for a large number of models, we found that extracellular scaffolds have an impact on intracellular signaling but do not change the subpopulation composition.

We discussed ideas for possible extensions of all proposed methods in the corresponding Sections 3.5 and 4.5. These ideas comprised, e.g., the use of adjoint sensitivities within the hierarchical approach for optimization to enable the efficient calibration of even larger models. Furthermore, methods for simultaneously estimating the parameters and performing model selection could be employed, and other promising alternative distributions could be incorporated in the frameworks. Further work could also be directed towards enabling the fitting of mechanistic models with a large number of parameters for single-cell data. For this, approaches from Chapter 3 for improving efficiency of optimization could be employed for studying single-cell data, e.g., analytically calculating optimal parameter values for parameters not required for model simulation. Furthermore, adjoint sensitivity analysis could not only be used for the calibration of ODE models, but also to facilitate a more efficient calibration of population models. Beyond population average and single-cell snapshot data, single-cell time-lapse data are commonly studied but were not considered in this thesis. These data provide temporal information about individual cells and the concepts proposed in this thesis could also be extended to include this information.

To conclude, this thesis proposed methods for the mathematical modeling and model calibration for different types of biological data. We showed the importance and capabilities of the methods by applying the methods to real biological problems and gaining new insights into the mechanisms of histone methylation and pain sensitization. This demonstrated that the methods and their application to study a broad range of biological questions will enable progress towards the aim of systems biology, namely, getting a holistic, mechanistic understanding of biological systems.

Appendix

A. Hessian matrices

We provide the Hessian matrices for the Gaussian, Laplace, Cauchy, Student's t and Huber distributions which were introduced in Section 3.1.

The Hessian matrix for the **Gaussian distribution** (3.2) for $l, m = 1, \ldots, n_\theta$ is given by

$$\begin{aligned}\frac{\partial \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_l \partial \theta_m} = -\frac{1}{2}\sum_j \Bigg[& -\frac{1}{\sigma_j^4(\boldsymbol{\theta})}\left(1-2\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}\right)\frac{\partial\sigma_j^2(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j^2(\boldsymbol{\theta})}{\partial\theta_m} \\ & + \frac{1}{\sigma_j^2(\boldsymbol{\theta})}\left(1-\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}\right)\frac{\partial^2\sigma_j^2(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\ & + 2\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^4(\boldsymbol{\theta})}\left(\frac{\partial\sigma_j^2(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial\sigma_j^2(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\right) \\ & + 2\frac{1}{\sigma_j^2(\boldsymbol{\theta})}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} - 2\frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})}\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m}\Bigg].\end{aligned}$$

For optimization, we approximated the Hessian by neglecting terms which depend on the second-order derivative of the outputs with respect to the parameters. This is based on the assumption that for good fits $\bar{y}_j - y_j(\boldsymbol{\theta})$ is small and, thus, also the influence of the second-order sensitivities is small.

The Hessian matrix for the **Laplace distribution** (3.3) is given by

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \theta_l \theta_m} = \sum_j \Bigg[& \left(-\frac{1}{\sigma_j(\boldsymbol{\theta})} + \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j^2(\boldsymbol{\theta})}\right)\frac{\partial^2\sigma_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\ & + \left(\frac{1}{\sigma_j^2(\boldsymbol{\theta})} - \frac{2|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j^3(\boldsymbol{\theta})}\right)\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m} \\ & - \frac{\mathrm{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\sigma_j^2(\boldsymbol{\theta})}\left(\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\right) \\ & + \frac{\mathrm{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\sigma_j(\boldsymbol{\theta})}\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m}\Bigg].\end{aligned}$$

Note that, in contrast to Gaussian noise, the term including the second-order sensitivities has an influence on the Hessian for Laplace noise even for small deviations of the measurement and observable.

The Hessian matrix for the **Huber distribution** (3.4) is given by

$$
\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} = \sum_j \; & \chi_{\{|\mathrm{res}_j(\boldsymbol{\theta})| \leq \tau(\boldsymbol{\theta})\}}(j) \cdot \Bigg[\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \frac{\bar{y}_j - y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} \\
& + \frac{\partial^2 \sigma_j(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \left(-\frac{1}{\sigma_j(\boldsymbol{\theta})} + \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^3(\boldsymbol{\theta})} \right) \\
& + \frac{\partial^2 \tau(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \left(\frac{\frac{2}{\tau^2(\boldsymbol{\theta})} e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\sqrt{2\pi}\mathrm{erf}\left(\frac{\tau(\boldsymbol{\theta})}{\sqrt{2}}\right) + \frac{2}{\tau(\boldsymbol{\theta})} e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}} \right) \\
& - \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} \frac{1}{\sigma_j^2(\boldsymbol{\theta})} \\
& - \left(\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_l} \right) \frac{2\left(\bar{y}_j - y_j(\boldsymbol{\theta})\right)}{\sigma_j^3(\boldsymbol{\theta})} \\
& + \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_m} \left(\frac{1}{\sigma_j^2(\boldsymbol{\theta})} - 3 \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j^4(\boldsymbol{\theta})} \right) \\
& + \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_m} \Bigg(\frac{2\sigma_j^2(\boldsymbol{\theta}) c_{\mathrm{Huber},j}(\boldsymbol{\theta})^2 e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\tau(\boldsymbol{\theta})} \cdot \\
& \left(\frac{2e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\tau^3(\boldsymbol{\theta})} - \frac{1 + \frac{2}{\tau^2(\boldsymbol{\theta})}}{\sigma_j(\boldsymbol{\theta}) c_{\mathrm{Huber},j}(\boldsymbol{\theta})} \right) \Bigg) \Bigg] \\
+ \; & \chi_{\{|\mathrm{res}_j(\boldsymbol{\theta})| > \tau(\boldsymbol{\theta})\}}(j) \cdot \Bigg[\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \frac{\tau(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} \mathrm{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta})) \\
& + \frac{\partial^2 \sigma_j(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \left(-\frac{1}{\sigma_j(\boldsymbol{\theta})} + \frac{\tau(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} |\bar{y}_j - y_j(\boldsymbol{\theta})| \right) \\
& + \frac{\partial^2 \tau(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_m} \left(\frac{\frac{2}{\tau^2(\boldsymbol{\theta})} e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\sqrt{2\pi}\mathrm{erf}\left(\frac{\tau(\boldsymbol{\theta})}{\sqrt{2}}\right) + \frac{2}{\tau(\boldsymbol{\theta})} e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}} - \frac{|(\bar{y}_j - y_j(\boldsymbol{\theta}))|}{\sigma_j(\boldsymbol{\theta})} - \tau(\boldsymbol{\theta}) \right) \\
& - \left(\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_l} \right) \frac{\mathrm{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta}))\tau(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})} \\
& + \left(\frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_m} \frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l} \right) \frac{\mathrm{sgn}(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\sigma_j(\boldsymbol{\theta})} \\
& + \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_m} \left(\frac{1}{\sigma_j^2(\boldsymbol{\theta})} - 2 \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})| \tau(\boldsymbol{\theta})}{\sigma_j^3(\boldsymbol{\theta})} \right) \\
& + \left(\frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_m} \frac{\partial \sigma_j(\boldsymbol{\theta})}{\partial\theta_l} \right) \frac{|\bar{y}_j - y_j(\boldsymbol{\theta})|}{\sigma_j^2(\boldsymbol{\theta})} \\
& + \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_l} \frac{\partial \tau(\boldsymbol{\theta})}{\partial\theta_m} \Bigg(\frac{2\sigma_j^2(\boldsymbol{\theta}) c_{\mathrm{Huber},j}(\boldsymbol{\theta})^2 e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\tau(\boldsymbol{\theta})} \cdot \\
& \left(\frac{2e^{-\frac{1}{2}\tau^2(\boldsymbol{\theta})}}{\tau^3(\boldsymbol{\theta})} - \frac{1 + \frac{2}{\tau^2(\boldsymbol{\theta})}}{\sigma_j(\boldsymbol{\theta}) c_{\mathrm{Huber},j}(\boldsymbol{\theta})} \right) - 1 \Bigg) \Bigg] .
\end{aligned}
$$

For residuals for which the absolute value is greater than τ, the Hessian depends on second-order sensitivities, even for small deviations of measurement and observable.

The Hessian matrix for the **Cauchy distribution** (3.5) is given by

$$
\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial\theta_l \theta_m} &= \sum_j \Bigg[\bigg(\frac{1}{\sigma_j(\boldsymbol{\theta})} - 2\frac{\sigma_j(\boldsymbol{\theta})}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2}\bigg)\frac{\partial^2 \sigma_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\
&+ \bigg[\frac{4\sigma_j(\boldsymbol{\theta})^2}{((\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2)^2} - \frac{1}{\sigma_j(\boldsymbol{\theta})^2} - \frac{2}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2}\bigg]\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m} \\
&- 4\frac{\sigma_j(\boldsymbol{\theta})(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\Big((\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j^2(\boldsymbol{\theta})\Big)^2}\left(\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} + \frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\right) \\
&+ \frac{2}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2}\left(\frac{2(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2} - 1\right)\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m} \\
&+ 2\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))}{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2 + \sigma_j(\boldsymbol{\theta})^2}\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m}\Bigg].
\end{aligned}
$$

Assuming that the deviation between measurement and observable is small, we can again neglect the second-order sensitivities. This provides an approximation which only depends on the first-order sensitivities.

The Hessian matrix for the **Student's t distribution** (3.6) is given by

$$
\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} =& \sum_j \Bigg[\frac{1}{2}\Bigg(\Psi\left(\frac{\nu_j(\boldsymbol{\theta})+1}{2}\right) - \Psi\left(\frac{\nu_j(\boldsymbol{\theta})}{2}\right) - \log\left(1 + \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\nu_j(\boldsymbol{\theta})\sigma_j(\boldsymbol{\theta})^2}\right) \\
& - \frac{1}{\nu_j(\boldsymbol{\theta})} + \frac{\nu_j(\boldsymbol{\theta})+1}{\nu_j^2(\boldsymbol{\theta})\sigma_j^2(\boldsymbol{\theta})}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2}}\Bigg)\frac{\partial^2\nu_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\
& - \left(\frac{1}{\sigma_j(\boldsymbol{\theta})} - \frac{\nu_j(\boldsymbol{\theta})+1}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2}}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\nu_j(\boldsymbol{\theta})\sigma_j^3(\boldsymbol{\theta})}\right)\frac{\partial^2\sigma_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\
& + \frac{\nu_j(\boldsymbol{\theta})+1}{1 + \frac{1}{\nu_j(\boldsymbol{\theta})}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2}}\frac{1}{\nu_j(\boldsymbol{\theta})}\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))}{\sigma_j^2(\boldsymbol{\theta})}\frac{\partial^2 y_j(\boldsymbol{\theta})}{\partial\theta_l\partial\theta_m} \\
& + \frac{1}{2}\Bigg(\frac{1}{2}\Psi_1\left(\frac{\nu_j(\boldsymbol{\theta})+1}{2}\right) - \frac{1}{2}\Psi_1\left(\frac{\nu_j(\boldsymbol{\theta})}{2}\right) + \frac{1}{\nu_j^2(\boldsymbol{\theta})} \\
& \quad + \frac{1}{\nu_j(\boldsymbol{\theta}) + \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2}}\left(\frac{\frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2} - 1}{\nu_j(\boldsymbol{\theta}) + \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\sigma_j(\boldsymbol{\theta})^2}} - \frac{1}{\nu_j(\boldsymbol{\theta})}\right) \\
& \quad \frac{(\bar{y}_j - y_j(\boldsymbol{\theta}))^2}{\nu_j(\boldsymbol{\theta})\sigma_j^2(\boldsymbol{\theta})}\Bigg)\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_m}
\end{aligned}
$$

$$
\begin{aligned}
&+\Bigg(\frac{1}{\sigma_j^2(\boldsymbol{\theta})}+\frac{\nu_j(\boldsymbol{\theta})+1}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^4(\boldsymbol{\theta})}\\
&\quad+\Bigg(2\frac{\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}-\frac{3}{\sigma_j(\boldsymbol{\theta})}\Bigg)\Bigg)\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m}\\
&+\frac{\nu_j(\boldsymbol{\theta})+1}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}\frac{1}{\sigma_j^2(\boldsymbol{\theta})}\Bigg(2\frac{\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}-1\Bigg)\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m}\\
&+2\frac{\nu_j(\boldsymbol{\theta})+1}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}\Bigg(\frac{\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}{\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}}-1\Bigg)\cdot\\
&\quad\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))}{\sigma_j^3(\boldsymbol{\theta})}\Bigg(\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m}+\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\Bigg)\\
&+\frac{\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}-1}{(\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})})^2}\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^3(\boldsymbol{\theta})}\Bigg(\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_m}+\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial\sigma_j(\boldsymbol{\theta})}{\partial\theta_l}\Bigg)\\
&+\frac{\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}-1}{\left(\nu_j(\boldsymbol{\theta})+\frac{(\bar{y}_j-y_j(\boldsymbol{\theta}))^2}{\sigma_j^2(\boldsymbol{\theta})}\right)^2}\frac{\bar{y}_j-y_j(\boldsymbol{\theta})}{\sigma_j^2(\boldsymbol{\theta})}\Bigg(\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_l}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_m}+\frac{\partial\nu_j(\boldsymbol{\theta})}{\partial\theta_m}\frac{\partial y_j(\boldsymbol{\theta})}{\partial\theta_l}\Bigg)\Bigg],
\end{aligned}
$$

where Ψ_1 is the trigamma function, the derivative of the digamma function Ψ as in (3.7). Assuming that the deviation between measurement and observable is small, we can again neglect the second-order sensitivities. This provides an approximation which only depends on the first-order sensitivities.

B. Standard model for histone methylation

Here, we provide the equations for the absolute abundance of histone modifications of generation g corresponding to (3.25):

$$\begin{pmatrix} \dot{\tilde{x}}_{g,00} \\ \dot{\tilde{x}}_{g,01} \\ \dot{\tilde{x}}_{g,02} \\ \dot{\tilde{x}}_{g,03} \\ \dot{\tilde{x}}_{g,10} \\ \dot{\tilde{x}}_{g,11} \\ \dot{\tilde{x}}_{g,12} \\ \dot{\tilde{x}}_{g,13} \\ \dot{\tilde{x}}_{g,20} \\ \dot{\tilde{x}}_{g,21} \\ \dot{\tilde{x}}_{g,22} \\ \dot{\tilde{x}}_{g,23} \\ \dot{\tilde{x}}_{g,30} \\ \dot{\tilde{x}}_{g,31} \\ \dot{\tilde{x}}_{g,32} \end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0
\end{pmatrix}^T \begin{pmatrix}
c_g(t)\, N(t) \\
k_{00\to01}\, \tilde{x}_{g,00} \\
k_{00\to10}\, \tilde{x}_{g,00} \\
k_{01\to02}\, \tilde{x}_{g,01} \\
k_{01\to11}\, \tilde{x}_{g,01} \\
k_{02\to03}\, \tilde{x}_{g,02} \\
k_{02\to12}\, \tilde{x}_{g,02} \\
k_{03\to13}\, \tilde{x}_{g,03} \\
k_{10\to11}\, \tilde{x}_{g,10} \\
k_{10\to20}\, \tilde{x}_{g,10} \\
k_{11\to12}\, \tilde{x}_{g,11} \\
k_{11\to21}\, \tilde{x}_{g,11} \\
k_{12\to13}\, \tilde{x}_{g,12} \\
k_{12\to22}\, \tilde{x}_{g,12} \\
k_{13\to23}\, \tilde{x}_{g,13} \\
k_{20\to21}\, \tilde{x}_{g,20} \\
k_{20\to30}\, \tilde{x}_{g,20} \\
k_{21\to22}\, \tilde{x}_{g,21} \\
k_{21\to31}\, \tilde{x}_{g,21} \\
k_{22\to23}\, \tilde{x}_{g,22} \\
k_{22\to32}\, \tilde{x}_{g,22} \\
k_{30\to31}\, \tilde{x}_{g,30} \\
k_{31\to32}\, \tilde{x}_{g,31} \\
d_{K27,1}\, \tilde{x}_{g,10} \\
d_{K27,1}\, \tilde{x}_{g,11} \\
d_{K27,1}\, \tilde{x}_{g,12} \\
d_{K27,1}\, \tilde{x}_{g,13} \\
d_{K27,2}\, \tilde{x}_{g,20} \\
d_{K27,2}\, \tilde{x}_{g,21} \\
d_{K27,2}\, \tilde{x}_{g,22} \\
d_{K27,2}\, \tilde{x}_{g,23} \\
d_{K27,3}\, \tilde{x}_{g,30} \\
d_{K27,3}\, \tilde{x}_{g,31} \\
d_{K27,3}\, \tilde{x}_{g,32} \\
d_{K36,1}\, \tilde{x}_{g,01} \\
d_{K36,1}\, \tilde{x}_{g,11} \\
d_{K36,1}\, \tilde{x}_{g,12} \\
d_{K36,1}\, \tilde{x}_{g,13} \\
d_{K36,2}\, \tilde{x}_{g,02} \\
d_{K36,2}\, \tilde{x}_{g,12} \\
d_{K36,2}\, \tilde{x}_{g,22} \\
d_{K36,2}\, \tilde{x}_{g,32} \\
d_{K36,3}\, \tilde{x}_{g,03} \\
d_{K36,3}\, \tilde{x}_{g,13} \\
d_{K36,3}\, \tilde{x}_{g,23}
\end{pmatrix},$$

and initial conditions (3.27).

Bibliography

H. D.-G. Acquah. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *J. Dev. Agric. Econ.*, 2(1):001–006, 2010.

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

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

S. J. Altschuler and L. F. Wu. Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559–563, 2010.

D. F. Anderson and D. J. Higham. Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *Multiscale Model. Simul.*, 10(1):146–179, 2012.

C. Andres, S. Meyer, O. A. Dina, J. D. Levine, and T. Hucho. Quantitative automated microscopy (QuAM) elucidates growth factor specific signalling in pain sensitization. *Mol. Pain*, 6(98):1–16, 2010.

C. Andres, J. Hasenauer, F. Allgöwer, and T. Hucho. Threshold-free population analysis identifies larger DRG neurons to respond stronger to NGF stimulation. *PLoS One*, 7 (3):e34257, 2012.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

J. Bachmann, A. Raue, M. Schilling, M. E. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. D. Lehmann, J. Timmer, and U. Klingmüller. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, 7(1):516, 2011.

R. Baker, J.-M. Peña, J. Jayamohan, and A. Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.*, 14 (20170660), 2018.

G. Balázsi, A. van Oudenaarden, and J. J. Collins. Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6):910–925, 2011.

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

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Inform. Process. Lett.*, 24(6):377–380, 1987.

B. Bodenmiller, E. R. Zunder, R. Finck, T. J. Chen, E. S. Savig, R. V. Bruggner, E. F. Simonds, S. C. Bendall, K. Sachs, P. O. Krutzik, and G. P. Nolan. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.*, 30(9):858–867, 2012.

R. Boiger, J. Hasenauer, S. Hross, and B. Kaltenbacher. Integration based profile likelihood calculation for PDE constrained parameter estimation problems. *Inverse Prob.*, 32(12): 125009, 2016.

M. Bouhaddou, A. M. Barrette, A. D. Stern, R. J. Koch, M. S. DiStefano, E. A. Riesel, L. C. Santos, A. L. Tan, A. E. Mertz, and M. R. Birtwistle. A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLoS Comput. Biol.*, 14(3):e1005985, 2018.

V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer, and D. H. Busch. Disparate individual fates compose robust CD8+ T cell immunity. *Science*, 340(6132):630–635, 2013.

F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, 33(2):155–160, 2015.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: A practical information-theoretic approach.* Springer, New York, NY, 2nd edition, 2002.

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

Z.-L. Chen, W.-M. Yu, and S. Strickland. Peripheral regeneration. *Annu. Rev. Neurosci.*, 30:209–233, 2007.

K.-H. Cho and O. Wolkenhauer. Systems biology: Discovering the dynamic behavior of biochemical networks. *Biosyst. Rev.*, 1(1):9–17, 2005.

T. Choi and J. Rousseau. A note on Bayes factor consistency in partial linear models. *J. Stat. Plan. Infer.*, 166:158–170, 2015.

F. Crauste, J. Mafille, L. Boucinha, S. Djebali, O. Gandrillon, J. Marvel, and C. Arpin. Identification of nascent memory CD8 T cells and modeling of their ontogeny. *Cell Syst.*, 4(3):306–317, 2017.

H. M. Davey and D. B. Kell. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. *Microbiol. Rev.*, 60(4):641–696, 1996.

A. Degasperi, D. Fey, and B. N. Kholodenko. Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *npj Syst. Biol. Appl.*, 3(1):20, 2017.

P. D. Dixit, E. Lyashenko, M. Niepel, and D. Vitkup. Maximum entropy framework for inference of cell population heterogeneity in signaling networks. *bioRxiv*, 137513, 2019.

S. Ebinger, E. Z. Özdemir, C. Ziegenhain, S. Tiedt, C. C. Alves, M. Grunert, M. Dworzak, C. Lutz, V. A. Turati, T. Enver, H.-P. Horny, K. Sotlar, S. Parekh, K. Spiekermann, W. Hiddemann, A. Schepers, B. Polzer, S. Kirsch, M. Hoffmann, B. Knapp, J. Hasenauer, H. Pfeifer, R. Panzer-Grümayer, W. Enard, O. Gires, and I. Jeremias. Characterization of rare, dormant, and therapy-resistant cells in acute lymphoblastic leukemia. *Cancer Cell*, 30(6):849–862, 2016.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, 180(2):498–515, 2006.

G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, 10(1):390, 2019.

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

K. J. Ferrari, A. Scelfo, S. Jammula, A. Cuomo, I. Barozzi, A. Stützer, W. Fischle, T. Bonaldi, and D. Pasini. Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Mol. Cell*, 53(1):49–62, 2014.

A. Fiedler, S. Raeth, F. J. Theis, A. Hausser, and J. Hasenauer. Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints. *BMC Syst. Biol.*, 10(80), 2016.

S. Filippi, C. P. Barnes, J. Cornebise, and M. P. Stumpf. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol.*, 12(1):87–107, 2013.

S. Filippi, C. P. Barnes, P. D. W. Kirk, T. Kudo, K. Kunida, S. S. McMahon, T. Tsuchiya, T. Wada, S. Kuroda, and M. P. Stumpf. Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. *Cell Rep.*, 15(11):2524–2535, 2016.

D. S. Fischer, A. K. Fiedler, E. M. Kernfeld, R. M. J. Genga, A. Bastidas-Ponce, M. Bakhti, H. Lickert, J. Hasenauer, R. Maehr, and F. J. Theis. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.*, 37(4):461–468, 2019.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London, Ser. A*, 222:309–368, 1922.

R. Fletcher and M. J. Powell. A rapidly convergent descent method for minimization. *Comp. J.*, 6(2):163–168, 1963.

Z. R. Fox and B. Munsky. The finite state projection based fisher information matrix approach to estimate information and optimize single-cell experiments. *PLoS Comput. Biol.*, 15(1):1–23, 2019.

A. P. Frei, F.-A. Bava, E. R. Zunder, E. W. Y. Hsieh, S.-Y. Chen, G. P. Nolan, and P. F. Gherardini. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods*, 13(3):269–275, 2016.

F. Fröhlich, P. Thomas, A. Kazeroonian, F. J. Theis, R. Grima, and J. Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.*, 12(7):e1005030, 2016.

F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, 13(1):e1005331, 2017.

F. Fröhlich, T. Kessler, D. Weindl, A. Shadrin, L. Schmiester, H. Hache, A. Muradyan, M. Schütte, J.-H. Lim, M. Heinig, F. J. Theis, H. Lehrach, C. Wierling, B. Lange, and J. Hasenauer. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Syst.*, 7(6):567–579.e6, 2018.

F. Fröhlich, C. Loos, and J. Hasenauer. Scalable inference of ordinary differential equation models of biochemical processes. In G. Sanguinetti and V. A. Huynh-Thu, editors, *Gene Regulatory Networks: Methods and Protocols*, volume 1883 of *Methods in Molecular Biology*, chapter 16, pages 385–422. Humana Press, 1 edition, 2019.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Stat. Comput.*, 24(6):997–1016, 2014.

C. Gerlach, J. C. Rohr, L. Perié, N. van Rooij, J. W. J. van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, R. J. De Boer, and T. N. M. Schumacher. Heterogeneous differentiation patterns of individual CD8+ T cells. *Science*, 340(6132): 635–639, 2013.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger, editors, *Bayesian Statistics*, pages 169–193. University Press, Oxford, UK, 1992.

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

C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11:417–422, 2014.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188 (1):404–425, 1992.

D. T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306, 2000.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reaction systems. *J. Chem. Phys.*, 115:1716–1733, 2001.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Math. Comp.*, 24(109):23–26, 1970.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

D. Grün, L. Kester, and A. van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, 2014.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

J. Hasenauer, S. Waldherr, M. Doszczak, P. Scheurich, and F. Allgöwer. Density-based modeling and identification of biochemical networks in cell populations. In M. Kothare, M. Tade, O. Moses, A. Wouwer, and I. Smets, editors, *Proc. of the 9th Int. Symp. on Dynamics and Control of Process Syst.*, volume 9, pages 320–325, Leuven, Belgium, 2010.

J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(125), 2011a.

J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer. Analysis of heterogeneous cell populations: a density-based modeling and identification framework. *J. Process Control*, 21(10):1417–1425, 2011b.

J. Hasenauer, C. Hasenauer, T. Hucho, and F. J. Theis. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *PLoS Comput. Biol.*, 10(7):e1003686, 2014.

H. Hass, K. Masson, S. Wohlgemuth, V. Paragas, J. E. Allen, M. Sevecka, E. Pace, J. Timmer, J. Stelling, G. MacBeath, B. Schoeberl, and A. Raue. Predicting ligand-dependent tumors from multi-dimensional signaling features. *npj Syst. Biol. Appl.*, 3 (1):27, 2017.

H. Hass, C. Loos, E. Raimúndez-Álvarez, J. Timmer, J. Hasenauer, and C. Kreutz. Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, page btz020, 2019.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 51(1):97–109, 1970.

D. M. Hawkins. *Identification of Outliers*, volume 11. Springer Netherlands, 1980.

M. A. Henson. Dynamic modeling of microbial cell populations. *Curr. Opin. Biotechnol.*, 14(5):460–467, 2003.

L. A. Herzenberg, J. Tung, W. A. Moore, L. A. Herzenberg, and D. R. Parks. Interpreting flow cytometry data: A guide for the perplexed. *Nat. Immunol.*, 7(7):681–685, 2006.

S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 2017.

A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM T. Math. Software.*, 31(3):363–396, 2005.

V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

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

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Stat.*, 35(1):73–101, 1964.

T. Hucho and J. D. Levine. Signaling pathways in sensitization: toward a nociceptor cell biology. *Neuron*, 55(3):365–376, 2007.

S. Hug, M. Schwarzfischer, J. Hasenauer, C. Marr, and F. J. Theis. An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule. *Stat. Comput.*, 26(3):663–677, 2016.

C. Hurvich and C.-L. Tsia. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

J. Isensee, M. Diskar, S. Waldherr, R. Buschow, J. Hasenauer, A. Prinz, F. Allgöwer, F. W. Herberg, and T. Hucho. Pain modulators regulate the dynamics of PKA-RII phosphorylation in subgroups of sensory neurons. *J. Cell Sci.*, 127:216–229, 2014.

J. Isensee, M. Kaufholz, M. J. Knape, J. Hasenauer, H. Hammerich, H. Gonczarowska-Jorge, R. P. Zahedi, F. Schwede, F. W. Herberg, and T. Hucho. PKA-RII subunit phosphorylation precedes activation by cAMP and regulates activity termination. *J. Cell. Biol.*, 2018.

H. Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, 3rd edition, 1961.

R.-R. Ji, R. W. Gereau, M. Malcangio, and G. R. Strichartz. MAP kinase and pain. *Brain Res. Rev.*, 60(1):135–148, 2009.

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

B. Kaltenbacher and J. Offtermatt. A refinement and coarsening indicator algorithm for finding sparse solutions of inverse problems. *Inverse Probl. Imag.*, 5(2):391–406, 2011.

R. E. Kass and A. E. Raftery. Bayes factors. *J. Am. Stat. Assoc.*, 90(430):773–795, 1995.

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

P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11(7):740–742, 2014.

K. H. Kim and C. W. Roberts. Targeting EZH2 in cancer. *Nat. Med.*, 22(2):128, 2016.

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

A. Klimovskaia, S. Ganscha, and M. Claassen. Sparse regression based structure learning of stochastic reaction networks from single cell snapshot time series. *PLoS Comput. Biol.*, 12(12):e1005234, 2016.

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

A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell RNA sequencing. *Mol. Cell*, 58(4):610–620, 2015.

M. Komorowski, B. Finkenstädt, C. V. Harper, and D. A. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1):343, 2009.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

L. Kuepfer, M. Peter, U. Sauer, and J. Stelling. Ensemble modeling for analysis of cell signaling dynamics. *Nat. Biotechnol.*, 25(9):1001, 2007.

J. Kuha. AIC and BIC: Comparisons of assumptions and performance. *Sociol. Method. Res.*, 33(2):188–229, 2004.

A. Küper, R. Dürr, and S. Waldherr. Dynamic density estimation in heterogeneous cell populations. *IEEE Control Syst. Lett.*, 3(2):242–247, 2019.

A. Kuzmichev, K. Nishioka, H. Erdjument-Bromage, P. Tempst, and D. Reinberg. Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of zeste protein. *Genes Dev.*, 16(22):2893–2905, 2002.

M. K. Lacki and B. Miasojedow. State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Stat. Comput.*, 26 (5):951–964, 2015.

K. Lane, D. Van Valen, M. M. DeFelice, D. N. Macklin, T. Kudo, A. Jaimovich, A. Carr, T. Meyer, D. Pe'er, S. C. Boutet, and M. W. Covert. Measuring signaling and RNA-seq in the same cell links gene expression to dynamic patterns of NF-κB activation. *Cell Syst.*, 4(4):458–469.e5, 2017.

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

N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55(2):195–207, 2006.

C. H. Lee, K. H. Kim, and P. Kim. A moment closure method for stochastic reaction networks. *J. Chem. Phys.*, 130(13):134107, 2009.

G. Lee, W. Finn, and C. Scott. Statistical file matching of flow cytometry data. *J. Biomed. Inform.*, 44(4):663–676, 2011.

C. Lester, C. A. Yates, M. B. Giles, and R. E. Baker. An adaptive multi-level simulation algorithm for stochastic biological systems. *J. Chem. Phys.*, 142(2):01B612_1, 2015.

J.-R. Lin, M. Fallahi-Sichani, and P. K. Sorger. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun.*, 6:8390, 2015.

C. Loos, C. Marr, F. J. Theis, and J. Hasenauer. *Computational Methods in Systems Biology*, volume 9308 of *Lecture Notes in Computer Science*, chapter Approximate Bayesian Computation for stochastic single-cell time-lapse data using multivariate test statistics, pages 52–63. Springer International Publishing, 2015.

C. Loos, A. Fiedler, and J. Hasenauer. Parameter estimation for reaction rate equation constrained mixture models. In E. Bartocci, P. Lio, and N. Paoletti, editors, *Proc. 13th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 186–200. Springer International Publishing, 2016.

C. Loos, S. Krause, and J. Hasenauer. Hierarchical optimization for the efficient parametrization of ODE models. *Bioinformatics*, 34(24):4266–4273, 2018a.

C. Loos, K. Moeller, F. Fröhlich, T. Hucho, and J. Hasenauer. A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Syst.*, 6(5):593–603, 2018b.

X.-K. Lun, V. R. Zanotelli, J. D. Wade, D. Schapiro, M. Tognetti, N. Dobberstein, and B. Bodenmiller. Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry. *Nat. Biotechnol.*, 35(2):164–172, 2017.

X. Luo and P. V. Shevchenko. The t copula with multiple parameters of degrees of freedom: bivariate characteristics and application to risk management. *Quant. Finance*, 10(9): 1039–1054, 2010.

A. Lyons and C. Parish. Determination of lymphocyte division by flow cytometry. *J. Immunol. Methods.*, 171(1):131–137, 1994.

C. Maier. Parameter estimation for outlier corrupted data. Master's thesis, Technische Universität München, Mathematische Fakultät, 2016.

C. Maier, C. Loos, and J. Hasenauer. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725, 2017.

M. Malik-Hall, O. A. Dina, and J. D. Levine. Primary afferent nociceptor mechanisms mediating NGF-induced mechanical hyperalgesia. *Eur. J. Neurosci.*, 21(12):3387–3394, 2005.

J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, 9(34), 2011.

P. W. Mantyh, M. Koltzenburg, L. M. Mendell, L. Tive, and D. L. Shelton. Antagonism of nerve growth factor-TrkA signaling and the relief of pain. *Anesthesiology*, 115(1): 189–204, 2011.

N. V. Mantzaris, P. Daoutidis, and F. Srienc. Numerical solution of multi-variable cell population balance models: I. finite difference methods. *Comp. Chem. Eng.*, 25(11-12): 1411–1440, 2001.

W. Q. Meeker and L. A. Escobar. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am. Stat.*, 49(1):48–53, 1995.

X. L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.*, 6(4):831–860, 1996.

R. Merkle, B. Steiert, F. Salopiata, S. Depner, A. Raue, N. Iwamoto, M. Schelker, H. Hass, M. Wäsch, M. E. Böhm, O. Mücke, D. B. Lipka, C. Plass, W. D. Lehmann, C. Kreutz, J. Timmer, M. Schilling, and U. Klingmüller. Identification of cell type-specific differences in erythropoietin receptor signaling in primary erythroid and lung cancer cells. *PLoS Comput. Biol.*, 12(8):e1005049, 2016.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

F. Michor and K. Polyak. The origins and implications of intratumor heterogeneity. *Cancer Prev. Res.*, pages 1940–6207, 2010.

T. Miyashiro and M. Goulian. Single-cell analysis of gene expression by fluorescence microscopy. *Methods Enzymol.*, 423:458–475., 2007.

E. J. Molinelli, A. Korkut, W. Wang, M. L. Miller, N. P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D. B. Solit, C. A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina, and C. Sander. Perturbation biology: Inferring signaling networks in cellular systems. *PLoS Comput. Biol.*, 9(12):e1003290, 2013.

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

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, 2006.

B. Munsky and M. Khammash. Identification from stochastic cell-to-cell variation: a genetic switch case study. *IET Syst. Biol.*, 4(6):356–366, 2010.

K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

J. P. Myers, M. Santiago-Medina, and T. M. Gomez. Regulation of axonal outgrowth and pathfinding by integrin-ECM interactions. *Dev. Neurobiol.*, 71(11):901–923, 2011.

R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Stat. Comput.*, 6(4):353–366, 1996.

E. W. Nester and B. A. Stocker. Biosynthetic latency in early stages of deoxyribonucleic acidtransformation in Bacillus subtilis. *J. Bacteriol.*, 86:785–796, 1963.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in statistics*, pages 73–108. Springer, 1992.

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

J. Nocedal and S. Wright. *Numerical Optimization.* Springer Science & Business Media, 2006.

S. H. Orkin and K. Hochedlinger. Chromatin connections to pluripotency and cellular reprogramming. *Cell*, 145(6):835–850, 2011.

Y.-i. Ozaki, S. Uda, T. H. Saito, J. Chung, H. Kubota, and S. Kuroda. A quantitative image cytometry technique for time series or population analyses of signaling networks. *PLoS One*, 5(4):e9955, 2010.

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

E. Pierson and C. Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16(1):241, 2015.

M. A. Pinto, C. D. Immanuel, and F. J. Doyle III. A feasible solution technique for higher-dimensional population balance models. *Comp. Chem. Eng.*, 31(10):1242–1256, 2007.

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

S. Pyne, X. Hu, K. Wang, E. Rossin, T. Lin, L. Maier, C. Baecher-Allan, G. McLachlan, P. Tamayo, D. Hafler, P. De Jager, and J. Mesirov. Automated high-dimensional flow cytometric data analysis. *Proc. Natl. Acad. Sci. USA*, 106(21):8519–8124, 2009.

Y. Qiu, T. Hu, B. Liang, and H. Cui. Robust estimation of parameters in nonlinear ordinary differential equation models. *J. Syst. Sci. Complex.*, 29(1):41–60, 2016.

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

A. Raue, 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. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One*, 8(9):e74335, 2013.

A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. J. Theis, U. Klingmüller, B. Schöberl, and J.Timmer. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.

A. Regev, S. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. K. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, J. Lundeberg, P. Majumder, J. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Philipakis, C. P. Ponting, S. R. Quake, W. Reik, O. Rozenblatt-Rosen, J. R. Sanes, R. Satija, T. Shumacher, A. K. Shalek, E. Shapiro, P. Sharma, J. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, A. van Oudenaarden, A. Wagner, F. M. Watt, J. S. Weissman, B. Wold, R. J. Xavier, N. Yosef, and H. C. A. M. Participants. The Human Cell Atlas. *bioRxiv*, 121202, 2017.

W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432, 2007.

J. Renart, J. Reiser, and G. R. Stark. Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. USA*, 76(7):3116–3120, 1979.

F. Rigat and A. Mira. Parallel hierarchical sampling: a general-purpose class of multiple-chains MCMC algorithms. *Comp. Stat. Data Anal.*, 56(6):1450–1467, 2012.

H. Risken. *The Fokker-Planck equation: Methods of solution and applications.* Springer, Berlin / Heidelberg, 2nd edition, 1996.

P. Rumschinski, S. Borchers, S. Bosio, R. Weismantel, and R. Findeisen. Set-based dynamical parameter estimation and model invalidation for biochemical reaction networks. *BMC Syst. Biol.*, 4(69), 2010.

S. K. Sahu, D. K. Dey, and M. D. Branco. A new class of multivariate skew distributions with applications to Bayesian regression models. *Can. J. Stat.*, 31(2):129–150, 2003.

G. Sauvageau, P. M. Lansdorp, C. J. Eaves, D. E. Hogge, W. H. Dragowska, D. S. Reid, C. Largman, H. J. Lawrence, and R. K. Humphries. Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells. *Proc. Natl. Acad. Sci. USA*, 91(25):12223–12227, 1994.

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

D. Schnoerr, G. Sanguinetti, and R. Grima. Validity conditions for moment closure approximations in stochastic chemical kinetics. *J. Chem. Phys.*, 141(8):084103, 2014.

B. Schöberl, E. A. Pace, J. B. Fitzgerald, B. D. Harms, L. Xu, L. Nie, B. Linggi, A. Kalra, V. Paragas, R. Bukhalid, V. Grantcharova, N. Kohli, K. A. West, M. Leszczyniecka, M. J. Feldhaus, A. J. Kudla, and U. B. Nielsen. Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor–PI3K axis. *Sci. Signal.*, 2 (77):ra31, 2009.

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

B. Sengupta, K. J. Friston, and W. D. Penny. Efficient gradient computation for dynamical models. *NeuroImage*, 98:521–527, 2014.

R. Serban and A. C. Hindmarsh. CVODES: The sensitivity-enabled ODE solver in SUNDIALS. In *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 257–269. ASME, 2005.

V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. USA*, 105(45):17256–17261, 2008.

P. Shi and E. A. Valdez. Multivariate negative binomial models for insurance claim counts. *Insur. Math. Econ.*, 55:18–29, 2014.

R. Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, pages 147–164, 1980.

R. Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

D. Silk. *Unscented approaches to inference and design for systems and synthetic biology.* PhD thesis, Imperial College London, 2013.

J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Anal.*, 1(4): 833–359, 2006.

S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245):428–433, 2009.

P. Stapor, F. Fröhlich, and J. Hasenauer. Optimization and profile calculation of ODE models using second order adjoint sensitivity analysis. *Bioinformatics*, 34(13):i151–i159, 2018a.

P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich, and J. Hasenauer. PESTO: Parameter EStimation TOolbox. *Bioinformatics*, 34(4):705–707, 2018b.

B. Steiert, J. Timmer, and C. Kreutz. L1 regularization facilitates detection of cell type-specific parameters in dynamical systems. *Bioinformatics*, 32(17):i718–i726, 2016.

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

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

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

S. Tay, J. J. Hughey, T. K. Lee, T. Lipniacki, S. R. Quake, and M. W. Covert. Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature*, 466:267–271, 2010.

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

C. Thomaseth and N. Radde. Normalization of Western blot data affects the statistics of estimators. *IFAC-PapersOnLine*, 49(26):56–62, 2016.

R. van der Merwe. *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models.* PhD thesis, Oregon Health & Science University, 2004.

N. G. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland, Amsterdam, 3rd edition, 2007.

A. F. Villaverde, F. Froehlich, D. Weindl, J. Hasenauer, and J. R. Banga. Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics*, bty736, 2018.

H. von Foerster. Some remarks on changing populations. In J. F. Stohlman, editor, *The kinetics of cellular proliferation*, pages 382–407. Grune and Stratton, New York, 1959.

E. J. Wagner and P. B. Carpenter. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.*, 13(2):115, 2012.

S. Waldherr. Estimation methods for heterogeneous cell population models in systems biology. *J. R. Soc. Interface*, 15(147):20180530, 2018.

M. Wang and X. Sun. Bayes factor consistency for nested linear models with a growing number of parameters. *J. Stat. Plan. Infer.*, 147:95–105, 2014.

L. Wassermann. Bayesian model selection and model averaging. *J. Math. Psychol.*, 44(1): 92–107, 2000.

P. Weber, J. Hasenauer, F. Allgöwer, and N. Radde. Parameter estimation and identifiability of biological networks using relative data. In S. Bittanti, A. Cenedese, and S. Zampieri, editors, *Proc. of the 18th IFAC World Congress*, volume 18, pages 11648–11653, Milano, Italy, 2011.

D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinf.*, 8(2):109–116, 2007.

D. J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.*, 10(2):122–133, 2009.

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

J. Yao, A. Pilko, and R. Wollman. Distinct cellular states determine calcium signaling response. *Mol. Syst. Biol.*, 12(12):894, 2016.

C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. USA*, 109(21):8340–8345, 2012.

Y. Zheng, S. M. M. Sweet, R. Popovic, E. Martinez-Garcia, J. D. Tipton, P. M. Thomas, J. D. Licht, and N. L. Kelleher. Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. *Proc. Natl. Acad. Sci. USA*, 109(34):13549–13554, 2012.

Z.-Y. Zhuang, H. Xu, D. E. Clapham, and R.-R. Ji. Phosphatidylinositol 3-kinase activates ERK in primary sensory neurons and mediates inflammatory heat hyperalgesia through TRPV1 sensitization. *J. Neurosci.*, 24(38):8300–8309, 2004.