HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

HelmholtzZentrum münchen
German Research Center for Environmental Health

Efficient accurate and robust statistical inference for deterministic and stochastic models of biochemical systems

Yannik Schälte

September 2021

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik

Efficient accurate and robust statistical inference for deterministic and stochastic models of biochemical systems

Yannik Schälte

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. Elisabeth Ullmann

Prüfer der Dissertation:

1. Prof. Dr.-Ing. Jan Hasenauer
2. Prof. Dr. Barbara Kaltenbacher
3. Prof. Dr. Michael Stumpf

Die Dissertation wurde am 27.09.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 09.05.2022 angenommen.

Acknowledgments

I want to thank ...

... first of all Jan Hasenauer, for giving me the opportunity to work in his group, to learn so many new things, for his great supervision, guidance, and inputs in numerous discussions.

... my present and former colleagues in the work group, in particular Anna Fiedler, Benjamin Ballnus, Carolin Loos, Daniel Weindl, Dantong Wang, Dilan Pathirana, Elba Raimúndez-Álvarez, Emad Alamoudi, Emmanuel Klinger, Erika Dudkin, Fabian Fröhlich, Felipe Reck, Jakob Vanhoefer, Leonard Schmiester, Lorenzo Contento, Manuel Huth, Marc Vaisband, Paul Jost, Paul Stapor, Philipp Städter, Polina Lakrisenko, Simon Merkt, Stephan Grein, Vanessa Nakonecnij, as well as further present and former colleagues at the ICB, in particular Akshaya Ramakrishnan, Alexander Ohnmacht, Anna Danese, Barbara Höllbacher, Benjamin Schubert, Carsten Maar, Christiane Fuchs, Dennis Rickert, Dominik Waibel, Elmar Spiegel, Emy Yue Hu, Hananeh Aliee, Hannah Busen, Ines Assum, Johanna Denkena, Karolina Worf, Katharina Schmid, Kridsadakorn Chaichoompu, Lisa Amrhein, Lisa Bast, Lisa Sikkema, Lea Schuh, Linda Krause, Malte Luecken, Maren Büttner, Maria Richter, Marius Lange, Mercè Gari, Mohammad Lotfollahi, Philipp Angerer, Ronan le Gleut, Susanne Pieschner, Svitlana Oleshko, Valerio Lupperger, Volker Bergen, and all who I forgot to list here, for lots of discussions and lots of fun.

... the Institute of Computational Biology (ICB), with its director Fabian Theis, for providing an exciting environment to work.

... Anandhi Iyappan, Claudia Reske, Marianne Antunes, Nina Fischer, Elisabeth Noheimer, Sabine Kunz, and Anna Sacher for their support in organizational matters at the ICB and the University of Bonn.

... many more people including Dennis Prangle, Scott Sisson, Michael Stumpf, and my thesis advisory committee members Barbara Kaltenbacher and Elisabeth Ullmann, for valuable inputs and discussions.

... my collaboration partners, in particular Alejandro Villaverde, Andreas Wieser, Clemens Kreutz, Frank Bergmann, Frederik Graw, Jens Timmer, Julio Banga, Jörn Starruß, Katja Radon, Lutz Brusch, Michael Hoelscher, Michael Pritsch, Wolfgang Müller, for great collaborations on e.g. CanPathPro, PEtab, FitMultiCell, and KoCo19.

... my friends and family, in particular my mum, my sister, my grandparents, and my dog.

... everyone else who I inadvertently forgot to mention above.

Abstract

Mechanistic mathematical models are powerful tools in systems biology to describe and understand biochemical systems, allowing to unravel underlying causal mechanisms. Such models depend commonly on parameters that need to be estimated from experimental data. Parameter estimation typically comprises both the identification of optimal point estimates, e.g. via optimization, and uncertainty quantification, e.g. via sampling. When the likelihood cannot be evaluated, which is often the case for complex stochastic models, likelihood-free methods such as approximate Bayesian computation (ABC) are useful. Depending on the system, data, and research question, different model types are employed, common examples including ordinary differential equation (ODE) models and multi-scale agent-based models. Increasing computational resources, growing biological knowledge, and improving experimental techniques have led to the development of increasingly complex models. While this ultimately allows for a more holistic description and understanding of biochemical systems, parameter estimation for such models becomes challenging, necessitating the development of novel methods.

In this thesis, we develop efficient, accurate, and robust parameter inference methods, specifically for optimization of large-scale ODE models, and ABC methods for Bayesian inference for stochastic models.

Firstly, we develop a hierarchical optimization approach scalable to high-dimensional ODE models that efficiently calculates optimal observable and measurement noise parameters via a problem decomposition, while using scalable adjoint sensitivity analysis to compute gradients. On a pan-cancer model with thousands of state variables and parameters, this approach improves performance by orders of magnitude and facilitates the estimation of scale, offset, and noise parameters with virtually no computational overhead and thus the integration of heterogeneous relative datasets.

Secondly, turning to ABC, we develop an efficient, self-tuned, sequential approach to perform exact inference using ABC methods in the presence of measurement noise. While ignoring noise, which is easy to do in ABC, can lead to erroneous parameter estimates, we demonstrate on various test problems that our approach provides orders of magnitude speed-up compared to established approaches, and allows to estimate noise parameters, facilitating exact inference for a wide range of problems.

Thirdly, we develop a generally applicable, adaptive distance metric for ABC that is robust to outliers in the data. While established adaptive ABC distance metrics can give erroneous parameter estimates in the presence of outliers, we demonstrate on various test problems that our approach is substantially more robust and efficient, performing well on both outlier-corrupted and outlier-free data, while making no assumptions on the type of data or model.

Fourthly, we develop an ABC distance weighting scheme accounting besides scale for informativeness of data points on parameters, by using sensitivities of inverse regression models. On various test problems, we demonstrate efficiency and robustness of our approach, in particular its ability to identify uninformative data points. Further, we demonstrate general advantages of accounting for both scale and informativeness, and tackle problems of established regression-based approaches in the presence of parameter non-identifiability.

In conclusion, the methods presented in this thesis allow for more accurate inference compared to established approaches, have better scaling properties, are more widely applicable, more robust to heterogeneous data, and computationally more efficient. We hope that these contributions will facilitate the understanding of, in particular biochemical, systems at an unprecedented level.

Zusammenfassung

Mechanistische mathematische Modelle sind in der Systembiologie ein leistungsfähiges Instrument zur Beschreibung und zum Verständnis biochemischer Systeme, mit dem sich die zugrunde liegenden kausalen Mechanismen entschlüsseln lassen. Solche Modelle hängen in der Regel von Parametern ab, welche aus experimentellen Daten geschätzt werden müssen. Die Parameterschätzung umfasst in der Regel sowohl die Ermittlung optimaler Punktschätzungen, z.B. durch Optimierung, als auch die Quantifizierung der Unsicherheit, z.B. durch Sampling. Wenn die Likelihood-Funktion nicht ausgewertet werden kann, was bei komplexen stochastischen Modellen häufig der Fall ist, sind Likelihood-freie Methoden wie approximative Bayes'sche Inferenz (engl. ABC) nützlich. Je nach System, Daten und Forschungsfrage werden verschiedene Modelltypen verwendet, wie z.B. gewöhnliche Differentialgleichungen (ODE) und agentenbasierte Modelle. Zunehmende Rechenressourcen, wachsendes biologisches Wissen und verbesserte experimentelle Techniken haben zur Entwicklung von immer komplexeren Modellen geführt. Während dies letztlich eine ganzheitlichere Beschreibung und ein besseres Verständnis biochemischer Systeme ermöglicht, wird die Parameterschätzung für solche Modelle zu einer Herausforderung, welche die Entwicklung neuer Methoden erforderlich macht.

In dieser Arbeit entwickeln wir effiziente, akkurate und robuste Methoden zur Parameterinferenz, insbesondere für die Optimierung großer ODE-Modelle, sowie ABC-Methoden für die Bayes'sche Inferenz stochastischer Modelle.

Erstens entwickeln wir einen hierarchischen Optimierungsansatz, der auf hochdimensionale ODE-Modelle skalierbar ist und optimale Beobachtungs- und Messrauschparameter über eine Problemzerlegung effizient berechnet, sowie skalierbare adjungierte Sensitivitätsanalyse zur Berechnung von Gradienten verwendet. Bei einem Krebsmodell mit Tausenden von Zustandsvariablen und Parametern verbessert dieser Ansatz die Effizienz um Größenordnungen und ermöglicht die Schätzung von Skalen-, Versatz- und Rauschparametern praktisch ohne Rechenaufwand und damit die Integration heterogener relativer Datensätze.

Zweitens entwickeln wir einen effizienten, adaptiven, sequenziellen Ansatz zur exakten Inferenz mit ABC-Methoden in Gegenwart von Messrauschen. Während das Ignorieren von Messrauschen, bei ABC leicht möglich, zu fehlerhaften Parameterschätzungen führen kann, zeigen wir an verschiedenen Testproblemen, dass unser Ansatz im Vergleich zu etablierten Ansätzen um Größenordnungen schneller ist und die Schätzung von Messrauschparametern ermöglicht, und somit exakte Inferenz für eine große Bandbreite von Problemen ermöglicht.

Drittens entwickeln wir eine allgemein anwendbare, adaptive Distanzmetrik für ABC, die robust gegenüber Ausreißern in den Daten ist. Während etablierte adaptive ABC-Distanzmetriken bei Vorhandensein von Ausreißern zu fehlerhaften Parameterschätzungen führen können, zeigen wir an verschiedenen Testproblemen, dass unser Ansatz wesentlich robuster und effizient ist und sowohl bei ausreißerbehafteten als auch ausreißerfreien Daten gut funktioniert, wobei keine Annahmen über die Art der Daten oder des Modells getroffen werden.

Viertens entwickeln wir ein ABC-Distanzgewichtungsschema, das neben der Skala auch den Informationsgehalt von Datenpunkten über die Parameter berücksichtigt, indem wir Sensitivitäten inverser Regressionsmodelle verwenden. Anhand verschiedener Testprobleme demonstrieren wir Effizienz und Robustheit unseres Ansatzes, insbesondere seine Fähigkeit, uninformative Datenpunkte zu identifizieren. Darüber hinaus demonstrieren wir allgemein die Vorteile der gleichzeitigen Berücksichtigung von Skala und Informationsgehalt, und gehen Probleme etablierter regressionsbasierter Ansätze bei Nichtidentifizierbarkeit von Parametern an.

Die in dieser Arbeit vorgestellten Methoden ermöglichen im Vergleich zu etablierten An-

sätzen genauere Inferenz, haben bessere Skalierungseigenschaften, sind breiter anwendbar, robuster gegenüber heterogenen Daten und rechnerisch effizienter. Wir hoffen, dass diese Beiträge das Verständnis von, insbesondere biochemischen, Systemen beträchtlich erweitern werden.

Contents

Chapter 1

Introduction

1.1 Research topic

Modeling

Mathematical models are in many scientific areas important tools to describe and study real-world systems, famous examples including, and far from limited to, e.g. in physics Newton's and Einstein's laws of motion [Einstein, 1905, Newton, 1833] or Bohr's atom model [Bohr, 1913], in epidemiology SIR models of disease spread [Kermack et al., 1927], in biology the Hodgkin-Huxley neuron spiking model [Hodgkin and Huxley, 1952], or in economics the Black-Scholes financial markets model [Black and Scholes, 1973]. Modeling typically comprises decision on level of abstraction, model scope and type, construction of the model, fitting it to the described system by comparison to observed data, evaluation, and, potentially, using the model to study the system and make predictions [Bender, 2012, Gershenfeld and Gershenfeld, 1999, Kitano, 2002a].

Depending on in particular purpose, knowledge, and data, there exist a variety of model types. Since the second half of the twentieth century, continuously improved computational capabilities and experimental techniques generating large quantities of data have enabled the development, simulation and analysis of increasingly complex models relying on computer simulations [Kitano, 2002b, Shiflet and Shiflet, 2014]. As by the aphorism "all models are wrong, but some are useful" attributed to the statistician George Box [Box, 1976], models only provide a simplified description and potentially approximation of a real-world system such that their limitations must be kept in mind, however they can be useful in their respective scope. Often, "Occam's razor" [Blumer et al., 1987] finds application, in the sense that, for a given purpose, among models with similar power, simple ones are preferable, as easier to analyze and computationally less demanding.

Of marked interest are quantitative mechanistic models, which explicitly simulate important causal system mechanisms, and whose equations are often based on fundamental physical or biochemical laws. Compared to purely statistical machine learning models, mechanistic models rely on hypotheses of latent causal mechanisms underlying observations, mitigating the need for calibration data, and allowing to gain insights into underlying processes between inputs and

outputs [Baker et al., 2018]. Therefore, mechanistic models particularly allow to "[measure what can be measured and] make measureable what cannot be measured", a quote attributed to Galileo Galilei.

Computational systems biology models

In the field of computational systems biology, mechanistic models are widely used to describe and understand the complex interactions in biochemical systems [Cho and Wolkenhauer, 2005, Kitano, 2002a,b]. Arguably due to the wide range of scales, systems under consideration, and research questions in the life sciences, in systems biology there exists a variety of modeling approaches, including qualitative and quantitative, deterministic and stochastic, and spatially homogeneous and resolved models [Hasenauer et al., 2015].

A popular approach, as they often provide a reasonable trade-off between quantitative insights and computational tractability, are ordinary differential equation (ODE) models [Ingalls, 2013, Klipp et al., 2005]. ODE models have been e.g. used to describe cell signaling [Bachmann et al., 2011, Fröhlich et al., 2018], cell cycle [Lloyd, 2013, Münzner et al., 2019], metabolism [Smallbone and Mendes, 2013], apoptosis [Spencer and Sorger, 2011], and diseases such as cancer [Hass et al., 2017, Raimúndez et al., 2020]. Alternative approaches when assumptions underlying ODE models, of sufficiently abundant species and spatial homogeneity, are not met, include e.g. partial differential equations (PDE) incorporating spatial dynamics [Schaff et al., 1997], chemical master equation (CME) based stochastic models describing random discrete dynamics at low copy numbers [Ingalls, 2013, Klipp et al., 2005], stochastic differential equations (SDE) for continuous stochastic dynamics [Wilkinson, 2011], or agent-based models, representing e.g. cells as interacting agents [Turner et al., 2004].

Biological and biochemical processes often interact on different spatial or temporal scales, from molecular, cellular, tissue, organ, up to organism, and potentially population, level, with ordered dynamics on the macro-scale often emerging from complex micro-scale interactions [Martins et al., 2010]. To efficiently describe the resulting macro-scale behavior, multi-scale models often couple various of the aforementioned models or model types at the different scales [Dada and Mendes, 2011, Hasenauer et al., 2015]. Multi-scale models have been e.g. used to describe heart mechanics [Hunter and Borg, 2003], cancer tissue growth [Anderson and Quaranta, 2008, Jagiella et al., 2017], whole-cell life cycle [Karr et al., 2012], and virus dynamics of e.g. HIV [Imle et al., 2019] or HCV [Durso-Cain et al., 2021].

Parameter estimation

Quantitative mechanistic models are usually subject to unknown parameters that cannot be directly measured, but need to be estimated by comparison of model outputs to observed experimental data, giving rise to an inverse problem to the forward problem of model simulation [Tarantola, 2005].

Commonly, a likelihood function relating model outputs to observed data is formulated, potentially combined with prior knowledge in a Bayesian setting [Fröhlich et al., 2019]. Parameter estimation then typically comprises both the identification of optimal point estimates, e.g. using

optimization methods, as well as uncertainty quantification, e.g. using profiles or sampling [Raue et al., 2013b]. The inverse problem gets often complicated by the computational complexity of the forward model simulation, numerical instabilities, and parameter multi-modalities, such that reliable and efficient methods and tools are of interest. If the problem is sufficiently smooth, often the case for ODE models, derivatives provide good search directions, generating interest in e.g. efficient gradient calculation methods [Fröhlich et al., 2017, Raue et al., 2013b].

Especially for complex stochastic models, the likelihood function can become intractable, in which case methods circumventing its evaluation can be used, such as approximate Bayesian computation (ABC) [Hartig et al., 2011, Price et al., 2018, Sisson et al., 2018]. In systems biology, ABC has in particular facilitated inference for complex multi-scale models [Durso-Cain et al., 2021, Jagiella et al., 2017].

1.2 Challenges

With improving computational capabilities and growing biological knowledge, increasingly complex models have been developed, which facilitate a more holistic understanding at a system level, ultimately the goal of systems biology [Kitano, 2002a]. To cope with such models, novel, scalable methods have been and are being developed, such as efficient gradient calculation methods for ODE models via adjoint sensitivity analysis [Fröhlich et al., 2017], high-performance infrastructure parallelized tools [Dutta et al., 2017, Klinger et al., 2018, Schmiester et al., 2019], or methods that exploit or adapt to the problem structure [Loos et al., 2018, Prangle, 2017], render the analysis robust to errors in the data [Maier et al., 2017], or identify informative data points [Fearnhead and Prangle, 2012]. Yet, there exist countless open questions and challenges in computational systems biology, of which we address the following in this thesis:

(i) Large-scale ODE models have the potential to provide deeper insights, their parameterization however is challenging and requires scalable methods. In addition, data used to calibrate large-scale models are often only on a relative scale, and noise levels unknown. Hierarchical optimization approaches have been developed to efficiently estimate linear observable transformations and noise parameters for ODE models [Loos et al., 2018, Weber et al., 2011]. However, these can so far in particular not be combined with adjoint sensitivity analysis [Fröhlich et al., 2017], prohibiting their application to large-scale models.

(ii) The correct and efficient assessment of measurement noise is also a problem in ABC, in which the overall approximation error introduced is often unclear, and it is unfortunately easy to neglect noise altogether, leading to erroneous parameter estimates. An interpretation of ABC as giving exact inference for a noise-corrupted model has been used to define algorithms that perform exact inference in the presence of noise [Daly et al., 2017, van der Vaart et al., 2018, Wilkinson, 2013]. However, existing approaches only consider special noise models, and do not scale to computationally demanding models and high-dimensional data.

(iii) Outliers or highly deviant data points are problematic for parameter estimation [Maier et al., 2017], however in ABC commonly used distance metrics comparing simulated and observed data are highly sensitive to outliers, yielding erroneous or uncertain estimates.

As the automatic or manual removal of outliers can be difficult especially for structured data, ABC methods that yield efficient and robust inference for both outlier-corrupted and outlier-free data are of interest.

(iv) The practical performance of ABC relies on its ability to efficiently compare informative data features via distance metrics and summary statistics. Common distance metrics adapt to the problem structure by scale normalization, but do not take informativeness of different data points into account. Approaches have been developed to construct low-dimensional summary statistic representations, e.g. based on inverse regression models [Fearnhead and Prangle, 2012], however these may lose information and substantially rely on the accuracy of employed models. Thus, robust methods that directly operate on the full data and weight them by informativeness, instead of only correcting for scale, are of interest.

1.3 Contributions of this thesis

The overall aim of this thesis could be described as developing efficient, accurate, robust computational methods that facilitate scalable parameter estimation for complex and high-dimensional problems. We tackle various problems in the two fields of optimization for ODE models, and of likelihood-free inference for stochastic models via ABC. Specifically, we address in this thesis the challenges outlined in Section 1.2 by the following contributions:

(I) *Scalable hierarchical optimization for ODE models using adjoint sensitivity analysis.* This contribution addresses challenge (i). We present an approach that combines adjoint sensitivity analysis for efficient high-dimensional gradient calculation with a hierarchical formulation that allows to efficiently handle affine observable transformations and noise parameters. On a pan-cancer model with > 4000 parameters, we demonstrate far superior performance compared to established approaches.

(II) *Exact efficient inference in ABC with noisy measurements.* This contribution addresses challenge (ii). Besides illustrating the problem of measurement noise in ABC and discussing conceptual solutions, we present an efficient, self-tuned sequential method that allows to perform exact likelihood-free inference in the presence of measurement noise, and to estimate noise parameters alongside. On various test problems, we demonstrate orders of magnitude speed-up compared to established approaches, facilitating exact inference via ABC for a wide range of application problems.

(III) *Outlier-robust adaptive distances in ABC.* This contribution addresses challenge (iii). We present a widely applicable, adaptive distance metric that is robust to outliers in the data, further allowing for online, simulation-based detection thereof. On various test problems with both outlier-free and outlier-corrupted data, we demonstrate, compared to established approaches, its substantially more robust performance.

(IV) *Informative and adaptive distances and summary statistics in ABC.* This contribution addresses challenge (iv). We present an approach defining distance weights based on sensitivities of inverse regression models to account for data informativeness. Further, we discuss advantages of the combination of scale-normalization and regression-based summary statistics, and we discuss shortcomings of regression-based summary statistics in the presence

of parameter non-identifiability. On test problems, we demonstrate superior and robust performance of the sensitivity-weighted approach over established approaches.

Thus, (I) deals with ODE models, while (II-IV) are in the field of ABC. In part, (I) and (II) deal with efficient and accurate ways of accounting for measurement noise in inference for complex models, while (III) further considers outliers that cannot be explained by common measurement noise. (IV) complements (III) by focusing on informative data, thus increasing efficiency.

Some of these contributions are part of articles that have either been already peer-reviewed and published, are currently under peer-review, or are in preparation. Therefore, parts of this thesis are based on and partly identical to the following first-author or shared first-author publications of the thesis author:

- Schmiester, L.*, **Schälte, Y.***, Fröhlich, F., Hasenauer, J. and Weindl, D. (2019). Efficient parameterization of large-scale dynamic models based on relative measurements. Bioinformatics, 36(2), pp.594-602. (*equal contribution)
- **Schälte, Y.** and Hasenauer, J. (2020). Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. Bioinformatics, 36(Supplement 1), pp.i551-i559.
- **Schälte, Y.**, Alamoudi, E. and Hasenauer, J. (2021). Robust adaptive distance functions for approximate Bayesian inference on outlier-corrupted data. bioRxiv.
- **Schälte, Y.** and Hasenauer, J. (2021). Informative and adaptive distances and summary statistics in approximate Bayesian computation. *In preparation.*

Further, results of the following publications of the thesis author are used, albeit not discussed in detail:

- **Schälte, Y.**, Stapor, P. and Hasenauer, J. (2018). Evaluation of derivative-free optimizers for parameter estimation in systems biology. IFAC-PapersOnLine, 51(19), pp.98-101.
- Schmiester, L.*, **Schälte, Y.***, Bergmann, F., Camba, T., Dudkin, E., Egert, J., Fröhlich, F., Fuhrmann, L., Hauber, A. L., Kemmer, S., Lakrisenko, P., Loos, C., Merkt, S., Müller, W., Pathirana, D., Raimúndez, E., Refisch, L., Rosenblatt, M., Stapor, P., Städter, P., Wang, D., Wieland, F.-G., Banga, J. R., Timmer, J., Villaverde, A. F., Sahle, S., Kreutz, C., Hasenauer, J. and Weindl, D. (2021). PEtab—Interoperable specification of parameter estimation problems in systems biology. PLoS computational biology, 17(1), p.e1008646. (*equal contribution)
- Städter, P.*, **Schälte, Y.***, Schmiester, L.*, Hasenauer, J. and Stapor, P. (2021). Benchmarking of numerical integration methods for ODE models of biological systems. Scientific reports, 11(1), pp.1-11. (*equal contribution)
- Fröhlich, F., Weindl, D., **Schälte, Y.**, Pathirana, D., Paszkowski, Ł., Lines, G.T., Stapor, P. and Hasenauer, J. (2021). AMICI: High-Performance Sensitivity Analysis for Large Ordinary Differential Equation Models. Bioinformatics, p.btab227.

Other publications of the thesis author's doctoral research that are not directly used in this thesis are:

- Olbrich, L.*, Castelletti, N.*, **Schälte, Y.***, Garí, M.*, Pütz, P., Bakuli, A., ..., Fuchs, C., Wölfel, R., Hasenauer, J., Hoelscher, M., and Wieser, A. (2021). A Serology Strategy for Epidemiological Studies Based on the Comparison of the Performance of Seven Different Test Systems-The Representative COVID-19 Cohort Munich. medRxiv. (*equal contribution)

- Pritsch, M., Radon, K., Bakuli, A., Le Gleut, R., Olbrich, L., Noller, G., Michelle, J., Saathoff, E., Castelletti, N., Garí, M., Pütz, P., **Schälte, Y.**, ..., Hasenauer, J., Fuchs, C., Wieser, A., and Hoelscher, M. (2021). Prevalence and risk factors of infection in the representative COVID-19 cohort Munich. International journal of environmental research and public health, 18(7), p.3572.

- Radon, K., Bakuli, A., Pütz, P., Le Gleut, R., Noller, J.M.G., Olbrich, L., Saathoff, E., Garí, M., **Schälte, Y.**, ..., Wieser, A., Hoelscher, M., Hasenauer, J., and Fuchs, C. (2021). From first to second wave: follow-up of the prospective Covid-19 cohort (KoCo19) in Munich (Germany). medRxiv.

- Syga, S., David-Rus, D., **Schälte, Y.**, Meyer-Hermann, M., Hatzikirou, H. and Deutsch, A. (2020). Inferring the effect of interventions on COVID-19 transmission networks. arXiv preprint arXiv:2012.03846.

- Durso-Cain, K., Kumberger, P., **Schälte, Y.**, Fink, T., Dahari, H., Hasenauer, J., Uprichard, S.L. and Graw, F. (2021). HCV spread kinetics reveal varying contributions of transmission modes to infection dynamics. bioRxiv.

- Vanhoefer, J., Marta, R., Pathirana, D., **Schälte, Y.** and Hasenauer, J. (2021). yaml2sbml: Human-readable and-writable specification of ODE models and their conversion to SBML. Journal of Open Source Software, 6(61), p.3215.

- Contento, L., Castelletti, N., Raimúndez, E., Le Gleut, R., **Schälte, Y.**, Stapor, P., Hinske, L. C., Hölscher, M., Wieser, A., Radon, K., Fuchs, C., Hasenauer, J. (2021). Integrative modelling of reported case numbers and seroprevalence reveals time-dependent test efficiency and infection rates. *In preparation.*

Besides these articles, the thesis author has been involved in the development and maintenance of the following computational toolboxes, which, in particular but not only, provide reusable, tested and documented implementations of various of the methods presented in this thesis, and have already been used in numerous peer-reviewed publications:

- **pyPESTO** (`https://github.com/icb-dcm/pypesto`): A Python-based parameter estimation toolbox, providing an integrated framework including in particular optimization, profiling and sampling algorithms.

- **pyABC** (`https://github.com/icb-dcm/pyabc`): A Python-based package for sequential approximate Bayesian computation on local and high-performance infrastructure.

- **PEtab** (`https://github.com/petab-dev/petab`): A systems biology parameter estimation standard based on SBML models and TSV files, alongside a Python-based library.
- **AMICI** (`https://github.com/amici-dev/amici`): A C++-based ODE simulation and sensitivity calculation package with interfaces to Python and MatLab, interfacing the highly performant ODE and DAE SUNDIALS solvers CVODES and IDAS. It allows the definition of likelihood functions and provides first and second order derivatives using forward, adjoint, and steady-state sensitivities.
- **yaml2sbml** (`https://github.com/yaml2sbml-dev/yaml2sbml`): A Python-based tool to convert ODE models specified in the easily accessible YAML format to SBML and PEtab.
- **FitMultiCell** (`https://gitlab.com/fitmulticell/fit`): A Python-based pipeline that combines pyABC with Morpheus (`https://gitlab.com/morpheus.lab/morpheus`), a tool to model and simulate multi-cellular and multi-scale systems.
- **PESTO** (`https://github.com/icb-dcm/pesto`): A MatLab-based parameter estimation toolbox, the predecessor of pyPESTO.

1.4 Outline of this thesis

The remainder of this thesis is structured as follows: In Chapter 2, we introduce terminology and provide background knowledge on the underlying inference concepts, with focus on ODE based problems, as well as ABC. Chapters 3-6 contain the main contributions of this thesis, the first one on ODE problems, and the remaining ones on ABC methods. In Chapter 3, we present the adjoint hierarchical optimization approach for large-scale ODE models with affine observable transformations. In Chapter 4, we discuss implications of the ignorance of measurement noise in ABC, and present an efficient exact method to address it. In Chapter 5, we discuss the impact of outliers in the data on common ABC distance metrics, and present a widely applicable robust method. In Chapter 6, we discuss the combined problem of data informativeness and scale on ABC analyses, and present sensitivity-weighted distance metrics. The thesis is concluded in Chapter 7 with a short summary and an outlook on potential further steps.

Chapter 2

Background

In this chapter, we introduce and provide background information on the terminology and methodology underlying and relevant throughout this thesis. We begin with the notions of process, observation and noise models in systems biology. Then, we give a general introduction of relevant statistical inference methods, covering point estimates and uncertainty quantification. Thereafter, we firstly focus on gradient calculation and optimization methods for ordinary differential equation models, and secondly introduce approximate Bayesian computation. While this chapter provides a general and rather high-level overview, the subsequent main contribution chapters may contain further specific background relevant for the respective topics.

Parts of this chapter are based on and partly identical to the following publications of the thesis author:

- **Schälte, Y.**, Stapor, P. and Hasenauer, J. (2018). Evaluation of derivative-free optimizers for parameter estimation in systems biology. IFAC-PapersOnLine, 51(19), pp.98-101.
- Schmiester, L.*, **Schälte, Y.***, Fröhlich, F., Hasenauer, J. and Weindl, D. (2019). Efficient parameterization of large-scale dynamic models based on relative measurements. Bioinformatics, 36(2), pp.594-602. (*equal contribution)
- **Schälte, Y.** and Hasenauer, J. (2020). Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. Bioinformatics, 36(Supplement 1), pp.i551-i559.
- **Schälte, Y.**, Alamoudi, E. and Hasenauer, J. (2021). Robust adaptive distance functions for approximate Bayesian inference on outlier-corrupted data. bioRxiv.
- **Schälte, Y.**, and Hasenauer, J. (2021). Informative and adaptive distances and summary statistics in ABC. *In preparation.*

2.1 Modeling of biochemical systems

In this section, we define the mathematical model including process description, observation and noise model, likelihood function and Bayesian posterior, which will be subject to the statistical

inference methods discussed in this thesis.

Throughout this thesis, we consider for simplicity of notation and argumentation, and as it represents the most common application case, continuous random variables given via probability density functions with respect to Lebesgue measure on $\mathbb{R}^d$, $d \geq 1$. Most statements however similarly hold for discrete random variables given via probability mass functions with respect to count measure, or mixed continuous-discrete variables.

2.1.1 Mathematical models of biochemical processes

Throughout this thesis, we attempt to describe the behavior of a vector of biochemical states x via a generative *process model*

$$x \sim \pi(x|\theta) \in \mathbb{R}^{n_x}, \tag{2.1}$$

which may be deterministic or stochastic, and continuous or discrete. The states are typically cell or molecule abundances, counts or concentrations, of biochemical species such as proteins, protein complexes, ligands, or RNA levels, or also e.g. spatially resolved cell structures, depending on the model purpose and resolution. Here, $\theta \in \mathbb{R}^{n_\theta}$ denotes unknown model parameters that we intend to estimate, such as reaction rate coefficients or initial abundances. The notation of, e.g., n_x, n_θ, indicates, by slight abuse of notation, the dimensions of states and parameters, and will be used similarly throughout.

2.1.2 Observation and noise model and likelihood function

Typically, not all of x is directly measurable, but only a vector of *observables*

$$y = h(x, \theta) \in \mathbb{R}^{n_y}, \tag{2.2}$$

with a possibly parameter-dependent *observation function* h, which can include both projection and recombination components.

However, even if the model perfectly reflects the underlying processes, measurements are usually noise-corrupted. Thus, we consider some possibly parameter-dependent *noise distribution* $\pi(\bar{y}|y, \theta)$ such that the actually observed data $\bar{y}_{\text{obs}} \in \mathbb{R}^{n_y}$ are a realization

$$\bar{y}_{\text{obs}} \sim \pi(\bar{y}|y, \theta). \tag{2.3}$$

All things combined, we can formulate a *likelihood*

$$\bar{y}_{\text{obs}} \sim \pi(\bar{y}|\theta) = \int \pi(\bar{y}|y, \theta)\pi(y|\theta)\, dy = \int \pi(\bar{y}|h(x, \theta), \theta)\pi(x|\theta)\, dx \tag{2.4}$$

of observing data $\bar{y}_{\text{obs}}$ given model parameterization θ, which simplifies for deterministic process models via a Dirac distribution to $\pi(\bar{y}|\theta) = \pi(\bar{y}|h(x, \theta), \theta)$.

Often, a system is measured at different time points, with technical or biological replicates, or under different experimental conditions. For ease of notation and without loss of generality, we

assume that $x, y, \bar{y}$ already incorporate such temporal, replicate, and condition multiplicity, and hold dependence on potential time, replicate, or condition-specific inputs implicit, unless a more explicit formulation is required.

2.1.3 Bayesian posterior distribution

In Bayesian inference, the likelihood (2.4) is combined with a *prior distribution* $\pi(\theta)$ on the parameters encoding information and beliefs before observing the data. Together, by Bayes' Theorem, these give the *posterior distribution*

$$\pi(\theta|\bar{y}_{\text{obs}}) = \frac{\pi(\bar{y}_{\text{obs}}|\theta)\pi(\theta)}{\pi(\bar{y}_{\text{obs}})} \propto \pi(\bar{y}_{\text{obs}}|\theta)\pi(\theta), \tag{2.5}$$

where the *evidence* $\pi(\bar{y}_{\text{obs}}) = \int \pi(\bar{y}_{\text{obs}}|\theta)\pi(\theta)\, d\theta$ is a normalization constant that is usually hard to compute, but can luckily be disregarded in common inference methods [Kramer et al., 2010].

In systems biology, prior knowledge on parameters of biochemical systems has been collected in publicly available databases [Chang et al., 2009, Wittig et al., 2012], which can be queried manually or automatically. Often, lacking more detailed information, the prior is simply a uniform distribution over a biologically plausible regime, $\theta \sim U(V)$ with $V \subset \mathbb{R}^{n_\theta}$.

2.1.4 Ordinary differential equation models

In systems biology, there exists a variety of modeling approaches for (2.1), including qualitative and quantitative, stochastic and deterministic, and spatially homogeneous and resolved models, arguably owing to the variety of system and data types and model purposes corresponding to different research questions [Hasenauer et al., 2015] (see also Chapter 1).

While we will later also use other model types as application examples, here we present in more detail ordinary differential equation (ODE) models, both exemplarily and as we will exploit their precise structure later. ODEs are a very common approach to biochemical kinetics modeling, as they allow quantitative insights while being computationally comparably tractable. Assumptions underlying ODE models are that all involved species are sufficiently abundant such that the processes can be described by deterministic concentration dynamics, and that spatial inhomogeneity is negligible.

Temporal evolution

Typically based on chemical reaction rate equations assuming mass action kinetics, or e.g. Michaelis-Menten or Hill enzyme kinetics [Wilkinson, 2009], ODE models describe the temporal evolution of a state vector $x = x(t, \theta) \in \mathbb{R}^{n_x}$ of continuous species concentrations by the parameterized initial value problem

$$\dot{x}(t, \theta) = f(x(t, \theta), \theta), \quad x(t_0, \theta) = x_0(\theta) \tag{2.6}$$

with vector field $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$, time $t \in [t_0, T] \subset \mathbb{R}$, and unknown parameters $\theta \in \mathbb{R}^{n_\theta}$. Thus, for ODE models (2.1) is a deterministic mapping.

We assume that f is Lipschitz-continuous in x, such that the Picard-Lindelöf Theorem guarantees the existence and uniqueness of a solution $x(t)$ at least in a neighborhood of the initial condition [Coddington and Levinson, 1955]. While in general a unique solution over $[t_0, T]$ may not exist, it usually does for biologically plausible inputs, and assuming the model describes the system behavior sufficiently accurately, since biological systems are subject to fundamental physical constraints such as limitation of speed.

In general, an analytical solution to (2.6) is not available, necessitating the use of numerical methods, such as Runge-Kutta, Adams-Moulton, or Backward differentiation formula (BDF) [Hindmarsh et al., 2005, Maiwald and Timmer, 2008, Mendes et al., 2009]. In a benchmark of different numerical solvers on 142 biological models, we could give guidelines on the choice of solver and hyperparameters, to ensure fast and reliable simulations [Städter et al., 2021]. Overall, we found BDF, with Newton-type solver for the non-linear problem within the implicit method, and sparse direct solver (KLU) for the inner linear problem to work robustly and efficiently. In particular, this supported previous hypotheses that biological problems often exhibit (informally) stiff dynamics, indicated by the presence of both fast and slow dynamics, in which case implicit methods are favorable [Garfinkel et al., 1977, Mendes et al., 2009].

Observation model and measurement data

As typically not all states can be observed, we define an observable function $h : x \mapsto y$ as in (2.2), here treating for notational simplicity, as in Sections 2.1.1 and 2.1.2, x as a vector of states at potentially multiple observed time points, keeping in particular time dependence implicit. While the underlying process behavior is assumed to follow a deterministic ODE, measurements are usually noise-corrupted. Most frequently, noise is assumed to follow an additive Gaussian distribution

$$\bar{y}_i = y_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

A Gaussian distribution can be loosely motivated by the Central Limit Theorem, assuming a variety of small independent noise contributions. The random variables $\{\varepsilon_i\}_{i \leq n_y}$ are commonly assumed to be independent, unless correlation e.g. between time points or conditions must be accounted for. If the noise standard deviations σ_i are unknown, they can be modeled via additional parameters, $\sigma_i = \sigma_i(\theta)$. Assuming independence, the likelihood function is given by

$$\pi(\bar{y}_{\text{obs}}|\theta) = \prod_{i \leq n_y} \frac{1}{\sqrt{2\pi\sigma_i(\theta)^2}} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{\text{obs},i} - y_i(\theta)}{\sigma_i(\theta)}\right)^2\right), \tag{2.7}$$

with $y(\theta) = h(x(\theta), \theta)$.

During the measurement process, errors can occur that cannot be described by general measurement noise, including e.g. human labeling or technical errors. Those distorted measurements, also referred to as outliers, can lead to erroneous parameter estimates if unaccounted for. There exist various methods that try to detect outliers [Ben-Gal, 2005, Hodge and Austin, 2004, Niu et al., 2011] in order to remove them from the data. However, actual outliers can be hard to

distinguish from measurement noise, especially in high-dimensional, complex data. An alternative is the use of models that are stable with respect to outliers. A common method is to use heavy-tailed distributions such as the Laplace distribution [Maier et al., 2017], given by

$$\pi(\bar{y}_{\text{obs}}|\theta) = \prod_{i \leq n_y} \frac{1}{2\sigma_i(\theta)} \exp\left(-\frac{|\bar{y}_{\text{obs},i} - y_i(\theta)|}{\sigma_i(\theta)}\right). \tag{2.8}$$

If noise levels are proportional to measurement values, a multiplicative noise model can be used instead of an additive one [Raue et al., 2013b].

2.2 Statistical inference methods

Irrespective of the exact process model used, the model $\pi(\bar{y}|\theta)$ typically depends on latent parameters θ, such as kinetic rates, observable scale factors, or noise terms, which cannot be directly measured. The goal of *statistical inference* is to, based on the model and measured data $\bar{y}_{\text{obs}}$, make statements about the latent parameters. One is usually interested in firstly good point estimates, but secondly also uncertainty quantification. Extending beyond mere parameter estimation, this then allows to evaluate the ability of the model to describe the data, to compare models, or to make quantitative predictions (e.g. Raue et al. [2013b], Villaverde et al. [2021], Wilkinson [2011]).

A fundamental distinction is between frequentist and Bayesian paradigms, following the respective notions of probability: In *frequentist inference*, the underlying model parameters are considered fixed and statements made only in terms of probabilities corresponding to long-term expected frequencies upon repeated draws of data from the same experiment, as only repeatable random events can be assigned a probability. In contrast, in *Bayesian inference*, probabilities are viewed as a more general concept, and probabilistic statements made also about parameters, updating prior beliefs on parameters by measured data. Thus, in Bayesian inference parameters are regarded as random variables, and are assigned probabilities before and after observing the data.

Frequentist inference typically deals with testing hypotheses, identifying the underlying true parameters, or giving an interval covering the true parameters with high probability upon repetition. Bayesian inference makes statements typically in terms of expectations over the full parameter posterior distribution. In practice, the choice of paradigm and method is often rather based on the problem at hand, e.g. the informativeness of the data, the existence or necessity of prior knowledge, and the computational complexity [Raue et al., 2013a].

2.2.1 Point estimates

Point estimates are supposed to give a, with respect to some criteria, single "best" guess of the underlying parameters.

Maximum likelihood estimation

A common method is *Maximum likelihood (ML) estimation (MLE)*, which aims to find the parameters that maximize the likelihood of observed data as a function of θ [Fisher, 1922], giving an optimization problem

$$\theta^{\mathrm{ML}} = \arg\max_{\theta} \pi(\bar{y}_{\mathrm{obs}}|\theta). \tag{2.9}$$

Already in order to limit the search space, the parameters are usually constrained to an appropriate domain $\theta \in V \subset \mathbb{R}^{n_\theta}$, so-called box constraints. The interpretation of MLE as yielding those parameters under which the observed data are most probably is intuitive, and its application flexible, so that MLE has become an important tool in statistical inference. The ML estimator is an example of frequentist extremum estimators. While it has generally no optimality properties for finite sample sizes, MLE possesses, like other point estimators, under certain assumptions, in particular identifiability, a number of desirable limit properties towards infinite sample size, such as consistency (convergence in probability to true parameter), efficiency (no other consistent estimator has lower mean square error), asymptotic normality ($\theta^{\mathrm{ML}} \sim \mathcal{N}(\theta^{\mathrm{true}}, \mathrm{FIM}^{-1})$ with FIM the Fisher information matrix), and functional invariance (conserved under reparameterization) [Newey and McFadden, 1994, Wilks, 1962, Zacks, 2014].

In practice, instead of using the likelihood directly, one commonly works with the log-likelihood $\log \pi(\bar{y}_{\mathrm{obs}}|\theta)$, giving the maximization problem

$$\theta^{\mathrm{ML}} = \arg\max_{\theta} \log \pi(\bar{y}_{\mathrm{obs}}|\theta). \tag{2.10}$$

Because the logarithm is strictly increasing, this problem is equivalent to (2.9). Practically, working with the log-likelihood is more convenient, as it is firstly numerically more stable, and secondly many common probability distributions, particularly the exponential family, are only logarithmically concave, concavity (or convexity in minimization) being a desirable property in optimization [Nocedal and Wright, 2006].

While we quantify the support observed data give to model parameters via a statistical likelihood function $\pi(\bar{y}_{\mathrm{obs}}|\theta)$, occasionally inference of optimal parameters is done solely in terms of cost functions. Typical examples of such are (weighted) Euclidean or Manhattan distances, which are however, respectively, equivalent to Gaussian or Laplace log-likelihoods with known noise variables, thus those approaches can be integrated in the statistical framework presented here. E.g. for a Gaussian noise model we have

$$\log \pi(\bar{y}_{\mathrm{obs}}|\theta) = -\frac{1}{2} \sum_{i \leq n_y} \left[\log(2\pi\sigma_i(\theta)^2) + \left(\frac{\bar{y}_{\mathrm{obs},i} - y_i(\theta)}{\sigma_i(\theta)} \right)^2 \right]. \tag{2.11}$$

Maximum a-posteriori and Bayesian point estimation

The equivalent of MLE in a Bayesian framework is the *maximum a-posteriori (MAP)* estimate, defined as the maximum over the posterior distribution

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} \pi(\theta|\bar{y}_{\mathrm{obs}}) = \arg\max_{\theta} \left[\log \pi(\bar{y}_{\mathrm{obs}}|\theta) + \log \pi(\theta)\right]. \tag{2.12}$$

The optimization problems (2.10) or (2.12) can usually not be solved analytically and can possess multiple modes, necessitating the use of iterative global and local optimization methods. For these, at least for smooth problems, an efficient calculation of derivatives is often essential. We discuss common methods for ODE models in Section 2.3.

MLE with box constraints $\theta \in V \subset \mathbb{R}^{n_\theta}$ can also be interpreted as the MAP estimate with a uniform prior distribution $U(V)$, as then $\pi(\theta|\bar{y}_{\mathrm{obs}}) \propto \pi(\bar{y}_{\mathrm{obs}}|\theta)$. It should however be noted that MAP is, unlike ML, not invariant under reparameterization, which introduces a Jacobian in density transformation and usually results in a shift of the maximum. Thus, as for Bayesian inference in general, care should be taken regarding the space on which parameter priors are defined (e.g. in linear or log-space, a transformation commonly applied to e.g. reaction rate coefficients).

Other frequently reported Bayesian point estimates are the posterior mean $\mathbb{E}_{\pi(\theta|\bar{y}_{\mathrm{obs}})}[\theta]$ and median, which minimize, respectively, the expected mean square error and mean absolute error over the posterior distribution, also referred to as Bayes estimators [Jaynes, 2003]. Bayesian inference frequently involves generating a sample from the posterior distribution, in which case these are straightforward to calculate.

2.2.2 Uncertainty analysis: Local approximations and profiles

Beyond optimal point estimates, as obtained e.g. via ML or MAP estimation, one is, independent of frequentist or Bayesian paradigm, usually interested in assessing the uncertainty in such estimates, i.e. how well constrained its values are given model and observed data. In particular in the presence of non-identifiability or few data, the obtained estimates might be unreliable, and conclusions drawn not taking into account parameter uncertainty can e.g. underestimate prediction uncertainty and associated risks, and need in general not yield representative predictions [Maier et al., 2020].

In frequentist inference, uncertainty quantification typically amounts to identifying for a parameter mapping $\xi : V \subset \mathbb{R}^{n_\theta} \to \Gamma$ (e.g. identity or projection) a data-based confidence region $C(\bar{y}_{\mathrm{obs}}) \subset \xi(V)$ that contains the true parameter with high probability in the sense of long-term frequency, $\mathbb{P}_\theta[\{\bar{y} \in \mathbb{R}^{n_y} : \xi(\theta) \in C(\bar{y})\}] \geq 1 - \alpha$, to a confidence level of e.g. $1 - \alpha = 95\%$. For 1-dimensional projections, confidence intervals are common. Analogous in Bayesian statistics are credible regions or marginal credible intervals C covering e.g. 95% of the posterior distribution probability mass, $\mathbb{P}_{\pi(\theta|\bar{y}_{\mathrm{obs}})}[C] \geq 1 - \alpha$.

To select (approximate) confidence regions or intervals, there exist various strategies. For the problems considered here, they are often based on the ML estimate θ^{ML}. Given its asymptotic

normality, e.g. the FIM or Hessian of the negative log-likelihood at θ^{ML} can be used to obtain symmetric asymptotic confidence intervals via a local quadratic approximation [Venzon and Moolgavkar, 1988, Wilks, 1962]. This is computationally cheap and scalable to high-dimensional problems, where often no other methods are feasible, and can therefore be used to provide first insights [Kapfer et al., 2019, Villaverde et al., 2019], however especially for non-linear problems and few data, this approximation is often crude [Joshi et al., 2006].

Alternatives that are often more robust for small sample sizes can be derived from Wilk's Theorem of asymptotic χ^2 distribution of likelihood ratios [Raue et al., 2009, Venzon and Moolgavkar, 1988]. For 1-dimensional confidence intervals, similarly *profile likelihoods* can be used, i.e. maximum projections $\mathrm{PL}_i(\theta_i) = \max_{\hat{\theta}:\hat{\theta}_i=\theta_i} \log \pi(\bar{y}_{\mathrm{obs}}|\hat{\theta})$, $i = 1, \ldots, n_\theta$, onto single coordinates. In systems biology, these were introduced by Raue et al. [2009], as they are in particular robust in the presence of structural non-identifiabilities [Fröhlich et al., 2014]. They can be calculated using e.g. step-wise optimization or integration methods and are generally more expensive but also more accurate than local approximations [Stapor et al., 2018]. In a Bayesian context, similarly *profile posteriors* can be employed [Raue et al., 2013a].

For dynamical systems, uncertainty analysis typically also relates to structural (parameters can be theoretically uniquely determined under model and observation function given infinite data) and practical (parameters can be sufficiently constrained given the actual data) identifiability analysis. Non-identifiabilities may hinder inference approaches, their identification and model refinement thus being desirable [Villaverde et al., 2021].

2.2.3 Uncertainty analysis: Bayesian sampling

In Bayesian inference, one usually tries to take the full posterior distribution into account, thus statements are typically made in terms of expectations $\mathbb{E}_{\pi(\theta|\bar{y}_{\mathrm{obs}})}[\xi(\theta)]$ for functions $\xi : \mathbb{R}^{n_\theta} \to \mathbb{R}$. Examples are mean and variance as, respectively, measures of average behavior and variability. As the (non-normalized) posterior distribution is for most models not directly tractable, a common strategy is to draw a representative sample $\theta_i \sim \pi(\theta|\bar{y}_{\mathrm{obs}})$, $i = 1, \ldots, N$, for a population size $N \in \mathbb{N}$, from the posterior distribution and perform Monte Carlo integration

$$\mathbb{E}_{\pi(\theta|\bar{y}_{\mathrm{obs}})}[\xi(\theta)] \approx \frac{1}{N} \sum_{i \leq N} \xi(\theta_i).$$

In general, for a random variable $X \in \mathbb{R}^d$ and function ξ such that $\xi(X)$ is integrable, the Monte Carlo estimate of the expectation is given as

$$\mathbb{E}_X[\xi] \approx \hat{\xi} := \frac{1}{N} \sum_{i \leq N} \xi(x_i), \tag{2.13}$$

for i.i.d. $x_i \sim X$, $i = 1, \ldots, N$. The estimate is unbiased, i.e. $\mathbb{E}_X[\hat{\xi}] = \mathbb{E}_X[\xi]$. Further, $\hat{\xi} \to \mathbb{E}_X[\xi]$ almost surely for $N \to \infty$ by the strong law of large numbers, with error of order $N^{-1/2}$ if $\xi(X)$ has finite variance (independent of the problem dimension d) [Robert and Casella, 2004].

Sampling is computationally often more expensive than e.g. profile calculation [Ballnus et al., 2017], however usually provides the most holistic information beyond optimal points and is, up

Algorithm 2.1 Rejection sampling

repeat
  sample $x \sim g$, $u \sim U[0,1]$
  accept x if $u \leq \tilde{f}(x)/(M \cdot \tilde{g}(x))$
until acceptance
out: a sample $x \sim f$

to Monte Carlo error, exact. In the following, we shortly discuss common sampling approaches, in particular in preparation of Section 2.4 on ABC.

Rejection and importance sampling

Monte Carlo integration requires the simulation of random variables. For simple distributions, inverse transform can be used, if the cumulative distribution function can be efficiently inverted, or other transformations exploiting relations to distributions that can be efficiently sampled from [Robert and Casella, 2004]. For distributions considered here, this is scarcely possible. More common is the use of methods based on rejection or importance sampling. While rarely used in their pure form, we shortly sketch them here as they will find re-use in the discussion of ABC. Consider a random variable $X \in \mathbb{R}^d$ given via density function f (in our applications, usually $f(x) \mathrel{\widehat{=}} \pi(\theta|\bar{y}_{\text{obs}})$). Instead of directly sampling from f, a different (non-normalized) proposal distribution g can be used that is easy to sample from, with $f \ll g$, i.e. f is absolutely continuous with respect to g. Both approaches generate samples $x \sim g$ and, respectively, reject some or re-weight them in order to target f.

Proposition 2.1 (Rejection sampling)**.** *Consider non-normalized versions $\tilde{f}, \tilde{g}$ of densities f, g, such that $\tilde{f}(x) \leq M \cdot \tilde{g}(x)$ for all x for some $M \in \mathbb{R}_{>0}$. Then, we can generate a sample from f by repeatedly sampling $x \sim g$ and accepting with probability $\tilde{f}(x)/(M \cdot \tilde{g}(x))$, until a sample gets accepted.*

Here, non-normalized means that the densities are known only up to normalization, $f = C_f \cdot \tilde{f}$, $g = C_g \cdot \tilde{g}$, which is practically relevant as normalization constants can be hard to compute. Practically, this procedure can be implemented as in Algorithm 2.1.

Proof. Given $X \sim g$, $U \sim U[0,1]$, for a measurable set $\mathcal{A}$, the probability of $X \in \mathcal{A}$ conditioned on acceptance is

$$\mathbb{P}[X \in \mathcal{A} \mid U \leq \tfrac{\tilde{f}(X)}{M \cdot \tilde{g}(X)}] = \frac{\int_{\mathcal{A}} \frac{\tilde{f}(x)}{M \cdot \tilde{g}(x)} g(x)\, dx}{\int \frac{\tilde{f}(x)}{M \cdot \tilde{g}(x)} g(x)\, dx} = \int_{\mathcal{A}} f(x)\, dx = \mathbb{P}_f[\mathcal{A}],$$

as all normalization constants cancel out. For further details and discussion, see e.g. Robert and Casella [2004, Chapter 2] or Sisson et al. [2018, Chapter 4]. □

A major drawback of rejection sampling is the requirement of a normalizing constant M, which should further be chosen as a tight bound, as it is inversely proportional to the acceptance rate

$\mathbb{P}[U \leq \frac{\tilde{f}(X)}{M \cdot \tilde{g}(X)}]$. In importance sampling, samples are, instead of rejecting some, assigned a weight to give expectations over the target distribution:

Proposition 2.2 (Importance sampling)**.** *Consider densities $f \ll g$. With weights $v(\theta) := f(\theta)/g(\theta)$, for functions ξ integrable under f holds, assuming all integrals exist, for i.i.d. $x_i \sim g$, $i = 1, \ldots, N$,*

$$\mathbb{E}_f[\xi] = \mathbb{E}_g[v\xi] \approx \hat{\xi} := \frac{1}{N} \sum_{i \leq N} v(x_i)\xi(x_i), \tag{2.14}$$

where the Monte Carlo estimate is unbiased, i.e. $\mathbb{E}_g[\hat{\xi}] = \mathbb{E}_f[\xi]$, and $\hat{\xi} \to \mathbb{E}_f[\xi]$ almost surely for $N \to \infty$.

Usually, we know at least one distribution only up to normalization. In that case, we use the following

Proposition 2.3 (Self-normalized importance sampling)**.** *Consider non-normalized versions $\tilde{f}, \tilde{g}$ of densities f, g, with $f \ll g$. With weights $w(x) := \tilde{f}(x)/\tilde{g}(x) \propto v(x)$, for functions ξ integrable under f holds, for i.i.d. $x_i \sim g$, $i = 1, \ldots, N$, and self-normalized weights $W_i := w(x_i)/\sum_{j \leq N} w(x_j)$,*

$$\mathbb{E}_f[\xi] = \frac{\mathbb{E}_g[w\xi]}{\mathbb{E}_g[w]} \approx \hat{\xi} := \frac{\frac{1}{N}\sum_{i \leq N} w(x_i)\xi(x_i)}{\frac{1}{N}\sum_{i \leq N} w(x_i)} = \sum_{i \leq N} W_i \xi(x_i), \tag{2.15}$$

where $\hat{\xi} \to \mathbb{E}_f[\xi]$ almost surely for $N \to \infty$.

Proof of 2.2 and 2.3. The first identity in (2.14) is just an application of the Radon-Nikodym change of measure, informally $\int \frac{f(x)}{g(x)} \xi(x) g(x)\, dx = \int \xi(x) f(x)\, dx$, unbiasedness and asymptotic convergence follow as in (2.13). In (2.15), effectively a second Monte Carlo estimate is employed in the denominator, to approximate the unknown normalization constant. While $\hat{\xi}$ is here generally not unbiased due to the biased estimation of the inverse normalization constant, we may rewrite (2.15) as

$$\hat{\xi} = \frac{\frac{1}{N}\sum_{i \leq N} v(x_i)\xi(x_i)}{\frac{1}{N}\sum_{i \leq N} v(x_i)},$$

with v the normalized weights, where numerator and denominator converge almost surely, i.e. outside null sets, to, respectively, $\mathbb{E}_f[\xi]$ and 1 by the strong law of large numbers, thus so does their quotient, as unions of null sets are null sets. See e.g. Robert and Casella [2004, Chapter 3] or Sisson et al. [2018, Chapter 2] for further details. □

Besides being a means to sample from an intractable distribution, importance sampling can also be used as a variance reduction method, or e.g. to focus on regions of interest of ξ [Robert and Casella, 2004, Chapter 3].

Effective importance sample size

As particles in importance sampling are weighted, some contribute more to estimates than others, which can be interpreted as having a lower *effective sample size (ESS)* than a sample of the same

size directly from the target distribution. A common definition of ESS for importance sampling [Liu et al., 1998, Martino et al., 2017] can be motivated as follows: Consider a weighted linear combination

$$S_N = \frac{\sum_{i \leq N} w_i X_i}{\sum_{i \leq N} w_i}$$

of independent random variables X_i of variance $\sigma^2 > 0$. The unweighted mean $\frac{1}{\tilde{N}} \sum_{i \leq \tilde{N}} X_i$ of $\tilde{N}$ variables has variance $\sigma^2/\tilde{N}$. Equating gives

$$\frac{\sigma^2}{\tilde{N}} \stackrel{!}{=} \mathrm{Var}[S_N] = \frac{\sum_{i \leq N} w_i^2}{(\sum_{i \leq N} w_i)^2} \sigma^2 \Rightarrow \mathrm{ESS} := \tilde{N} = \frac{(\sum_{i \leq N} w_i)^2}{\sum_{i \leq N} w_i^2} \tag{2.16}$$

as a scale-invariant quantification of the ESS of the weighted sum.

Markov chain Monte Carlo sampling

Stand-alone, rejection and importance sampling suffer for higher-dimensional problems from the need of an appropriate choice for the proposal density g similar to the target density, as discrepancies result in high rejection rates or highly variable weights with a low ESS. More commonly used are methods based on Markov chains (MCMC) or sequential importance sampling (SMC) [Neal, 2011, Robert and Casella, 2004, Sisson et al., 2018].

Markov chain Monte Carlo (MCMC) methods construct a Markov chain $\{x_i\}_{i \geq 1}$ that has the target density f as its stationary distribution. Various methods exist to define the Markov transition kernel. One of the first is the Metropolis-Hastings (MH) algorithm, which, given a current state x, firstly proposes a new state $x' \sim K(x'|x)$ via some perturbation kernel K, and secondly accepts and moves to x' with acceptance probability

$$A(x'|x) = \min \left[1, \frac{f(x')}{f(x)} \frac{K(x|x')}{K(x'|x)}\right], \tag{2.17}$$

otherwise stays at x [Hastings, 1970, Metropolis et al., 1953]. This step is such that the transition probability distribution $P(x'|x) = A(x'|x)K(x'|x)$ satisfies the condition of detailed balance $P(x'|x)f(x) = P(x|x')f(x')$ necessary for stationarity of f. Provided the Markov process is ergodic, convergence of the empirical distribution to f is guaranteed [Robert and Casella, 2004]. Variations of the MH algorithm have been proposed that e.g. choose g adaptively [Haario et al., 2001], delay rejection [Haario et al., 2006], use multiple Markov chains e.g. via parallel tempering [Miasojedow et al., 2013, Vousden et al., 2016] or parallel hierarchical sampling [Rigat and Mira, 2012], sample from conditional marginals [Casella and George, 1992], or employ derivative information, e.g. via Hamiltonian Monte Carlo [Hoffman and Gelman, 2014, Neal, 2011] or the Metropolis-adjusted Langevin algorithm [Girolami and Calderhead, 2011]. MCMC-based sampling is popular e.g. for small-to-medium sized ODE models in systems biology [Ballnus, 2019, Ballnus et al., 2017]. For further information on MCMC methods see e.g. Neal [2011], Robert and Casella [2004], Wilkinson [2011], or Ballnus et al. [2017] for a review of different variants on ODE models, or Robert and Casella [2011] for a historical review of MCMC methods.

While widely used in Bayesian computing since the 1990s [Martin et al., 2020], MCMC has a number of disadvantages: While long-term convergence to the target distribution is guaranteed,

practical assessment of sample quality is challenging. Tests such as Geweke or Gelman-Rubin can be used to reject the hypothesis of converged chains [Brooks and Gelman, 1998, Geweke, 1992]. Often an initial burn-in phase, which may deviate from the long-term behavior, needs to be identified and disregarded. Thus, performing optimization prior to sampling to obtain start points in high-density regions is recommended [Ballnus et al., 2017]. Further, MCMC depends on the quality of the transition kernel, as too large, unstructured, or small average jump sizes can result in slowly mixing chains with high autocorrelation. Moreover, while e.g. parallel tempering allows to use parallel infrastructure to some degree, a chain is inherently sequential and the use of massively parallel infrastructure thus limited.

Sequential Monte Carlo sampling

In importance sampling, designing a proposal distribution g that places a large number of samples in regions of high density or interest can be difficult. *Sequential Monte Carlo* (SMC) methods overcome this problem by constructing a sequence of intermediary distributions $f_t(x)$, $t = 1, \dots, n_t$, with $f_{n_t} = f$. For every generation t, a population of particles $P_t = \{(x_i^t, w_i^t)\}_{i \geq 1}$ targeting f_t is generated from a proposal distribution g_t, with (potentially non-normalized) importance weights $w_t(x) = f_t(x)/g_t(x)$. The proposal distribution g_t is based on the previous population P_{t-1}, with a typically broad initial distribution g_1, e.g. $g_1 = f_1$, or the prior in a Bayesian setting.

While SMC as presented above only defines a general concept, there exist a variety of strategies to create concrete, efficient and robust SMC algorithms. For example, geometrically tempered intermediary distributions $f_t = f^{1/T_t} \cdot g_1^{1-1/T_t}$ with $T_t \searrow 1$, can be used to smoothly transition to the target distribution [Del Moral et al., 2006, Gelman and Meng, 1998, Neal, 1993]. In the case of Bayesian inference with data $\bar{y}_{\text{obs}} = \{\bar{y}_{\text{obs},i}\}_{i \leq n_t}$ consisting of n_t observations, e.g. via a hidden Markov model, alternatively $f_t(\theta) = \pi(\theta | \bar{y}_{\text{obs},1}, \dots, \bar{y}_{\text{obs},t})$ using partial data can be used. In that case, SMC is also referred to as particle filter [Chopin, 2002, Del Moral, 1997]. As one easily sees, the variance-minimizing proposal distribution is $g_t = f_t$, from which sampling is in general however not possible. Instead, commonly samples are be drawn from the previous weighted population and perturbed via a transition kernel [Sisson et al., 2007]. As the variability of importance weights especially in later generations can get large, resulting in a low ESS (2.16), occasional (e.g. when the ESS falls below half the population size) resampling from population P_t according to the particle weights has been suggested, to concentrate on samples with large weights [Douc and Cappé, 2005, Doucet et al., 2000]. Complementarily, rejuvenation e.g. via an MCMC kernel can be used, which moves each resampled particle, increasing particle diversity [Del Moral et al., 2006]. While usually particles from generations for $t < n_t$ are discarded in the final result P_{n_t}, there exist approaches to make use of all [Dau and Chopin, 2020]. Further, the sequential nature of SMC allows to successively adjust hyperparameters to the problem, or to e.g. use different levels of model resolution [Latz et al., 2018]. For further information on SMC methods, see e.g. Doucet et al. [2000], Del Moral et al. [2006], Sisson et al. [2018].

An essential advantage of SMC over MCMC alone is that the particles for generation t are independent, such that their generation can be parallelized, enabling the full exploitation of high-performance infrastructure, scaling to thousands of cores. Further, the population-based approach is inherently better suitable for multi-modal distributions, while for the efficient exploration of complex local shapes such as non-linear correlations, gradient-based methods, such

as Hamiltonian Monte-Carlo, may perform better. Unlike a single MCMC chain, SMC particles of a population are uncorrelated, such that a determination of burn-in phase or assessment of convergence are not required [Sisson et al., 2007]. Each generation already provides a sample from a tempered target, with the tempering error decreasing over time. However, the ESS may be low, which can result in degenerated, non-representative populations [Klinger and Hasenauer, 2017].

2.3 Parameter optimization for ODE models

In parameter inference, optimization is often a first step to identify optimal point estimates (Section 2.2.1). In this section, we give a short overview of common methods, with focus on ODE models. We consider a generic objective function to minimize, typically given as

$$\min_\theta J(\theta) \quad \text{with} \quad J(\theta) = -\log \pi(\bar{y}_{\text{obs}}|\theta) \quad \text{or} \quad J(\theta) = -\log \pi(\theta|\bar{y}_{\text{obs}}) \tag{2.18}$$

via (2.10) or (2.12) as the negative log-likelihood or negative log-posterior, the underlying process model being given via an ODE (2.6), with some observation model (2.2) and noise model (2.3) (e.g. given via normal (2.11) or Laplace noise). The formulation as a minimization instead of a maximization problem is just due to convention in the field of optimization. While optimization can be performed for any generic objective function, (2.18) has the advantage that it enables statistical interpretation and the use of previously outlined methods for uncertainty quantification.

2.3.1 Numerical optimization methods

Minimizing the objective function (2.18) presents a continuous, usually non-convex and often multi-modal optimization problem. Further, the solution of the ODE involved in every objective function evaluation (and its derivative, see Section 2.3.2) can usually only be numerically approximated, and is computationally expensive [Fröhlich et al., 2019, Klipp et al., 2005, Raue et al., 2013b]. Commonly, parameters are subjected in (2.18) to box constraints $V \subset \mathbb{R}^{n_\theta}$ of biologically plausible values, which for Bayesian inference carry the interpretation of a uniform prior (see Section 2.2).

A variety of numerical optimization approaches exist to solve (2.18) [Fröhlich et al., 2019, Nocedal and Wright, 2006]. Broadly, these can be classified as global and local, or as derivative-free and derivative-based methods, or combinations thereof. Local searches modify candidate solutions locally, in attempts of converging to the closest local optimum. While local optima may be of interest on their own, the overall aim is usually to find the global optimum. Thus, for multi-modal problems global searches are required that consider the whole search space. Derivative-based methods exploit first and possibly higher-order derivatives, in contrast to derivative-free methods.

Global optimization methods

Global optimization methods aim to efficiently explore the whole search space, often involving both global exploration, to find remote areas of interest, and local exploitation or concentration, to refine current solutions. The arguably simplest approach is multi-start local optimization, which performs many local optimizations initiated at different randomly chosen starting points [Raue et al., 2013b]. Numerous further methods have been developed, often motivated by natural, e.g. biological or physical, principles, simulating e.g. swarm behavior [Kennedy, 2011, Vaz and Vicente, 2007], evolutionary processes [Bäck, 1996, Hansen and Ostermaier, 1996], or physical annealing [Kirkpatrick et al., 1983]. Hybrid approaches such as scatter searches explicitly combine global and local approaches in an iterative process [Egea et al., 2014].

Local optimization methods

Derivative-based local searches use derivatives of J to find a good search direction and step length [Nocedal and Wright, 2006]. For example, via first and second-order Taylor approximations, assuming sufficient differentiability, one easily sees that the negative gradient $-\nabla J(\theta)$ gives the direction of steepest descent, while the Hessian $\nabla^2 J(\theta)$ allows to formulate a local quadratic approximation, containing information about local curvature, in particular its minimization giving Newton's step $s = -(\nabla^2 J(\theta))^{-1} \nabla J(\theta)$. Thus, various algorithms using first, second, or higher-order derivative information for local optimization have been developed. A basic distinction is into line-search methods, which first define a direction of descent, and then determine an adequate step length ensuring sufficient decrease [Wächter and Biegler, 2006], and trust-region methods, which instead first define a region giving a maximum step length, and in which a candidate step is subsequently proposed by minimizing a surrogate model [Coleman and Li, 1996]. Typical search directions are based on steepest descent via the gradient, [Curry, 1944], conjugate gradients [Branch et al., 1999], Newton's method using explicit second-order derivatives, or quasi-Newton methods such as (L)BFGS [Fletcher and Powell, 1963, Goldfarb, 1970, Liu and Nocedal, 1989] or SR1 [Byrd et al., 1996], which use iterative first-order approximations. See e.g. Boyd and Vandenberghe [2004], Nocedal and Wright [2006] for further information.

In contrast, derivative-free local methods do not use or approximate gradients, but rather employ heuristics to find good descent directions, including simplex-based methods [Nelder and Mead, 1965], hill climbing [De La Maza and Yuret, 1994], or linear or quadratic models [Powell, 1994, 2009]. Unlike for derivative-based methods [Nocedal and Wright, 2006], for derivative-free methods (efficient) convergence guarantees can scarcely be given [Rios and Sahinidis, 2013].

As we also observed in an own study [Schälte et al., 2018], if derivatives can be calculated efficiently and robustly, derivative-free local methods are usually inferior to derivative-based ones, especially for higher-dimensional problems, however can be advantageous for unstable derivatives or non-smooth problems (see Figure 2.1). For ODE models in systems biology, various studies have demonstrated that global optimization methods with gradient-based local searches outperform other approaches [Egea et al., 2014, Raue et al., 2013b, Schälte et al., 2018, Schmiester et al., 2021b, Villaverde et al., 2018]. Multi-start local optimization with efficient gradient calculation has been shown to perform robustly for systems biology ODE models [Raue et al., 2013b]. A particular advantage is that it can easily be parallelized, which is also possible but more intricate

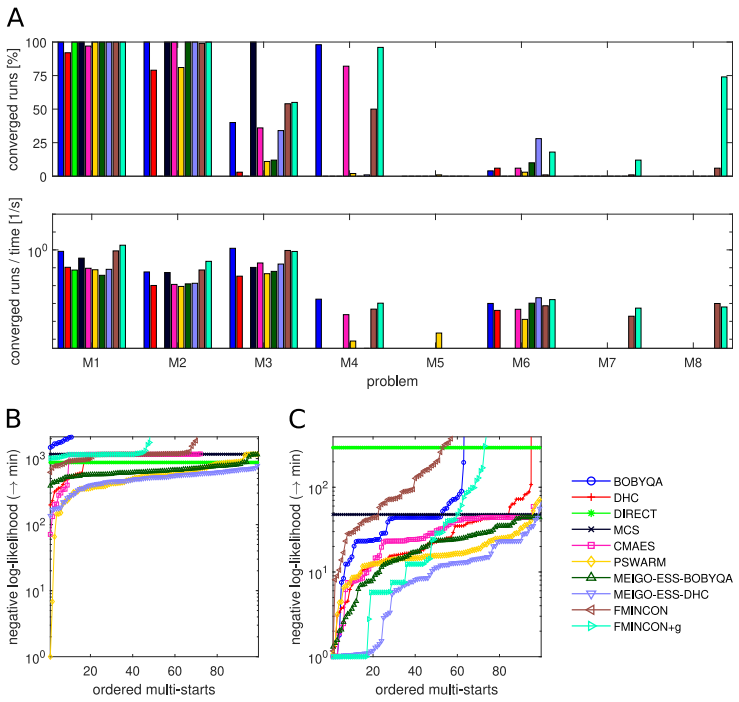

Figure 2.1: Performance of various optimization methods on eight ODE models of biochemical reaction networks. M1 ($n_\theta = 2$) describes a conversion reaction, M2 ($n_\theta = 4$) an enzyme-catalyzed reaction, M3 ($n_\theta = 5$) mRNA transfection [Leonhardt et al., 2014], M4 ($n_\theta = 7$) a discretized spatial model of Pom1 gradient formation [Hross et al., 2016], M5 ($n_\theta = 11$) a Hopf bifurcation [Ballnus et al., 2017], M6 ($n_\theta = 17$) Epo-induced JAK-STAT signaling [Swameye et al., 2003], M7 ($n_\theta = 28$) RAF-MEK-ERK signaling, and M8 ($n_\theta = 48$) histone methylation [Zheng et al., 2012], always assuming Gaussian additive noise (2.11). Considered were derivative-free local optimizers BOBYQA and DHC, derivative-free global optimizers DIRECT, MCS, CMAES, PSWARM, and MEIGO, as well as gradient-based FMINCON, with (+g) or without provided efficient gradients. A: Number of converged runs out of 100 multi-starts as a measure of reliability (top), and converged runs per time as a measure of efficiency (bottom). B, C: Waterfall plots (i.e. sorted multi-start results) for M5, which is highly multi-modal, and higher-dimensional M6. This figure is taken from the author's publication Schälte et al. [2018], see there for details and further analyses.

for e.g. hybrid approaches with iterated exploration and exploitation phases [Penas et al., 2017]. Further, optimal values obtained from single multi-starts allow for an assessment of the objective function landscape and quality of convergence e.g. via waterfall plots (Figure 2.1B+C). However, for complex problems, more efficient global exploration e.g. via scatter searches, in combination with efficient gradient-based local optimization, has been shown beneficial [Villaverde et al., 2018].

2.3.2 Gradient calculation

For a sufficiently smooth objective function J, especially in optimization (Section 2.3.1), but also e.g. for profile calculation (Section 2.2.2) and sampling (Section 2.2.3), efficiently calculated and accurate derivatives, in particular the gradient, are useful. In this section, we discuss common strategies for ODE based objectives.

Finite differences

Finite difference approximate the derivative tangent by a secant of small step size,

$$\frac{dJ}{d\theta_k}(\theta) \approx \frac{J(\theta + h_1 e_k) - J(\theta - h_2 e_k)}{h_1 + h_2},$$

with k-th unit vector e_k and $h_1, h_2 \geq 0$. Common are central ($h_1 = h_2 > 0$), forward ($h_1 > 0 = h_2$), and backward ($h_1 = 0 < h_2$) differences. While they are easy to implement for any black-box objective, their accuracy depends on the choice of step size, and they can be numerically unstable especially if the objective relies on numerical approximations. Further, the computational complexity of the gradient is linear in n_θ [Raue et al., 2013b].

Forward sensitivity analysis

Given an ODE model, we can obtain reliable gradients semi-analytically by augmenting the underlying system of equations. Consider a general sufficiently differentiable objective

$$J(\theta) = \sum_{i=1}^{n_t} J_i(y(t_i, \theta), \theta), \tag{2.19}$$

as a sum over discrete measurement time points, with observable function

$$y(t, \theta) = h(x(t, \theta), \theta),$$

with x given via an ODE (2.1). Here, we only make time dependence explicit, as we will need it later, while keeping potential multiple observables and conditions implicit. The gradient is then given as

$$\nabla J(\theta) = \sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \theta} + \frac{\partial J_i}{\partial y} \frac{dy}{d\theta} \right] = \sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \theta} + \frac{\partial J_i}{\partial y} \frac{\partial h}{\partial \theta} + \frac{\partial J_i}{\partial y} \frac{\partial h}{\partial x} \frac{dx}{d\theta} \right], \tag{2.20}$$

where we used

$$\frac{dy}{d\theta} = \frac{\partial h}{\partial \theta} + \frac{\partial h}{\partial x} \frac{dx}{d\theta}.$$

Note that here and in the following, we drop all obvious arguments for ease of readability, and indicate by ∂ partial and by d total derivatives. While all other derivatives are easily derived, the problematic part here are the state sensitivities $s^x := \frac{dx}{d\theta}$, as the state x itself is given only via the solution of an ODE.

Differentiating s^x by time, we can find via Schwarz' Theorem and ODE (2.1) an additional system of ODEs, the forward sensitivity equations

$$\dot{s}^x(t, \theta) = \frac{df}{d\theta} = \frac{\partial f}{\partial \theta} + \frac{\partial f}{\partial x} s^x, \quad s^x(t_0, \theta) = \frac{dx_0}{d\theta}(\theta)$$

These can be solved alongside (2.1) over $[t_0, t_{n_t}]$, giving an augmented system of ODEs of dimension $n_x(1 + n_\theta)$, the solution of which is the state x together with its sensitivities s^x, allowing to calculate (2.20). The errors in the numerical approximation of x and s^x can be controlled jointly, allowing an accurate gradient computation. While forward sensitivity analysis (FSA) is generally

more accurate and efficient than finite differences due to exploitation of common structure in the joint solution of the augmented system, it scales still linearly in n_θ [Hindmarsh et al., 2005, Leis and Kramer, 1988, Raue et al., 2013b].

Adjoint sensitivity analysis

Adjoint sensitivity analysis (ASA) circumvents evaluation of s^x altogether by introduction of an adjoint state $p : [t_0, t_{n_t}] \to \mathbb{R}^{n_x}$ taking the role of a Lagrange multiplier. This can be advantageous when sensitivities of few outputs have to be calculated with respect to many parameters [Cao et al., 2003, Fröhlich et al., 2017, Hindmarsh et al., 2005, Li and Petzold, 2004]. Defining $J_0 = 0$, we can write for $i = 0, \dots, n_t - 1$, by adding a zero-term,

$$\begin{aligned}\frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\frac{dx}{d\theta} &= \frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\frac{dx}{d\theta} + \int_{t_i}^{t_{i+1}} p^T\left(\frac{d\dot{x}}{d\theta} - \frac{\partial f}{\partial \theta} - \frac{\partial f}{\partial x}\frac{dx}{d\theta}\right) dt \\ &= \frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\frac{dx}{d\theta} + \left(\lim_{t \nearrow t_{i+1}} p^T\frac{dx}{d\theta} - \lim_{t \searrow t_i} p^T\frac{dx}{d\theta}\right) \\ &\quad - \int_{t_i}^{t_{i+1}} \left(\dot{p}^T + p^T\frac{\partial f}{\partial x}\right)\frac{dx}{d\theta}\, dt - \int_{t_i}^{t_{i+1}} p^T\frac{\partial f}{\partial \theta}\, dt,\end{aligned}$$

using integration by parts in the second step. Here, it becomes apparent that it is convenient to choose p such that it satisfies on $(t_i, t_{i+1}]$, $i = 0, \dots, n_t - 1$, the backward differential equation

$$\dot{p} = -\frac{\partial f}{\partial x}^T p \quad \text{with terminal condition} \quad p(t_{i+1}, \theta) = \lim_{t \searrow t_{i+1}} p(t, \theta) - \frac{\partial J_{i+1}}{\partial y}\frac{\partial h}{\partial x}, \tag{2.21}$$

with $p = 0$ for $t > t_{n_t}$. This is such that in the above expression the first integral vanishes, and due to the re-initialization of the backward equation at measurement time point t_i, the expression simplifies to

$$\begin{aligned}\frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\frac{dx}{d\theta} &= \frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\frac{dx}{d\theta} + p^T\frac{dx}{d\theta}\bigg|_{t=t_{i+1}} - \left(p^T + \frac{\partial J_i}{\partial y}\frac{\partial h}{\partial x}\right)\frac{dx}{d\theta}\bigg|_{t=t_i} - \int_{t_i}^{t_{i+1}} p^T\frac{\partial f}{\partial \theta}\, dt \\ &= p^T\frac{dx}{d\theta}\bigg|_{t=t_{i+1}} - p^T\frac{dx}{d\theta}\bigg|_{t=t_i} - \int_{t_i}^{t_{i+1}} p^T\frac{\partial f}{\partial \theta}\, dt\end{aligned}$$

Summing over $i = 0, \dots, n_t - 1$, simplifying the telescope sum, and using $p(t_{n_t}) = -\frac{\partial J_{n_t}}{\partial y}\frac{\partial h}{\partial x}$, we find for (2.20)

$$\nabla J(\theta) = \sum_{i=1}^{n_t}\left[\frac{\partial J_i}{\partial \theta} + \frac{\partial J_i}{\partial y}\frac{\partial h}{\partial \theta}\right] - p^T\frac{dx_0}{d\theta}\bigg|_{t=t_0} - \int_{t_0}^{t_{n_t}} p^T\frac{\partial f}{\partial \theta}\, dt\,. \tag{2.22}$$

This is independent of the state sensitivities, unlike FSA, except at t_0, which are usually straightforward to calculate. With (2.1) and (2.21), in ASA two ODEs of dimension n_x must be solved, besides a quadrature of dimension n_θ, which is however typically computationally less expensive. In practice, computation time for ASA was found to hardly increase with n_θ [Fröhlich et al., 2017, Özyurt and Barton, 2005]. As for FSA, for ASA dedicated solvers exist that solve the forward and adjoint ODEs jointly [Fröhlich et al., 2021, Hindmarsh et al., 2005, Serban and Hindmarsh,

2005]. A downside is however that, unlike FSA, ASA cannot be employed for least-square type optimization. Especially for large-scale sytems with many parameters, ASA has been shown to be by far the most efficient method [Fröhlich et al., 2017, Fröhlich et al., 2018, Kapfer et al., 2019].

While most computational tools trivially allow to calculate derivatives via finite differences, only few (e.g. Raue et al. [2015], Fröhlich et al. [2021]) allow FSA [Fröhlich et al., 2019]. To our knowledge, only AMICI (see Section 1.3) currently allows ASA [Fröhlich et al., 2021]. pyPESTO can be used with all three methods for gradient calculation, the latter two via AMICI.

Higher-order derivatives

Finite differences, FSA, and ASA can similarly also be employed to obtain higher-order derivatives, following similar principles. In particular the Hessian is of interest as a measure of curvature in optimization and profile calculation (see e.g. Stapor et al. [2018]).

2.4 Approximate Bayesian Computation

In the statistical inference methods presented heretofore, we have always assumed that we can evaluate the likelihood $\pi(\bar{y}_{\text{obs}}|\theta)$, and thus, in Bayesian inference, can calculate the posterior density, at least up to normalization. However, in practice this is often not possible for complex stochastic models, for numerical or computational reasons [Hasenauer, 2015, Jagiella et al., 2017, Martin et al., 2020]. In such cases, it is however often still relatively easy to simulate data $\bar{y} \sim \pi(\bar{y}|\theta)$ from the model, which amounts to implementing the model as a generative process, taking parameters and returning simulated data. Examples of such models include Markov processes, stochastic differential equations, or stochastic agent-based models (see Section 1). In systems biology, these are frequently used to describe e.g. gene expression, signal transduction, or multi-cellular processes, whenever the system dynamics cannot be adequately captured by a deterministic model [Lenive et al., 2016, Picchini, 2014, Wilkinson, 2009]. Especially to describe multi-scale and multi-cellular systems, often different modeling approaches on different spatial or temporal scales are combined, yielding highly specific high-dimensional stochastic models [Durso-Cain et al., 2021, Hasenauer, 2015, Imle et al., 2019].

Likelihood-free inference methods, i.e. methods that do not require evaluating the likelihood, have thus gained interest. For example, given an intractable target density f, pseudo-marginal methods have been developed, which use an unbiased estimator $\hat{f}_x$, i.e. $\mathbb{E}[f_x] = f(x)$, in place of $f(x)$, based on a finite number of samples [Andrieu and Roberts, 2009, Warne et al., 2020], for example particle MCMC (PMCMC) employing a particle filter for likelihood estimation in (2.17) [Andrieu et al., 2010]. While pseudo-marginal methods conceptually generate samples from the true target distribution, reliable sample-based likelihood estimators are expensive, and the Monte-Carlo error may be not unsubstantial [Martin et al., 2020]. In contrast, approximate methods only target an approximation to the target distribution. Especially, approximate Bayesian computation (ABC) [Sisson et al., 2018] has become increasingly popular in various research areas, owing to its asymptotic exactness, simplicity, and its broad applicability. In a

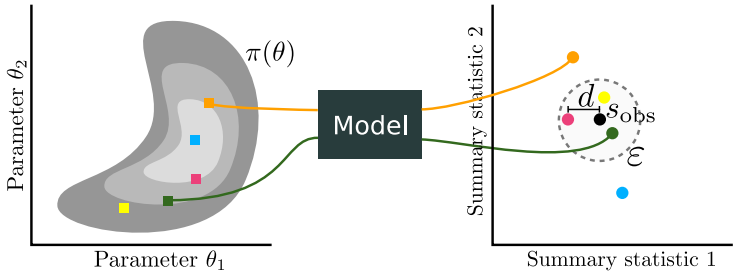

Figure 2.2: Illustration of approximate Bayesian computation (ABC). Given parameters θ sampled e.g. from the prior $\pi(\theta)$, or another proposal distribution $g(\theta) \gg \pi(\theta)$, data $\bar{y}$ are simulated via a generative black-box model, and compared to observed data $\bar{y}_{\text{obs}}$ on the level of summary statistics $s_{\text{obs}} = s(\bar{y}_{\text{obs}})$, via a distance metric d. The particle $(\theta, \bar{y})$ is accepted if their discrepancy is below an acceptance threshold, $d(s(\bar{y}), s(\bar{y}_{\text{obs}})) \leq \varepsilon$, otherwise rejected.

nutshell, in ABC the likelihood evaluation is replaced by simulating data from the model, and accepting if simulated and observed data are sufficiently close.

There are various alternative approximate likelihood-free inference methods. Indirect inference methods use an auxiliary tractable model $\pi_a(\bar{y}|\phi)$ with parameters ϕ, which can be a simplified mechanistic or also a purely data-analytic model. Inference is then based on the binding function $\phi(\theta) := \arg\max_\phi Q(\{\bar{y}_i\}_{i\leq n}|\phi)$, where Q is an estimating function, e.g. the likelihood π_a, and $\bar{y}_i \sim \pi(\bar{y}|\theta)$ are i.i.d. simulations from the original model. Bayesian synthetic likelihood (BSL) approaches use a multivariate normal approximation to the intractable (summary statistic) likelihood, with mean and covariance matrix dependent on θ and determined via simulation [Martin et al., 2020, Price et al., 2018]. Indirect inference, and in particular BSL approaches have the potential to be more efficient than ABC [Price et al., 2018], which can be interpreted as using a non-parametric likelihood approximation (see below), however rely on the adequacy of the auxiliary model.

2.4.1 The basic method

The core principle of ABC, as introduced in Pritchard et al. [1999] following ideas in Tavaré et al. [1997], consists of the following three steps (illustrated in Figure 2.2):

1. sample parameters $\theta \sim \pi(\theta)$,
2. simulate data $\bar{y} \sim \pi(\bar{y}|\theta)$,
3. accept $(\theta, \bar{y})$ if $d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon$.

Here, $d : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}_{\geq 0}$ is a distance metric measuring the proximity of simulated and observed data (see Chapter 5 for commonly used metrics), and $\varepsilon \geq 0$ is an acceptance threshold defining the permitted discrepancy. This is repeated until sufficiently many, say N, particles have been accepted, see Algorithm 2.2.

Like most ABC algorithms, this routine actually targets the joint distribution over parameters and data,

$$(\theta, \bar{y}) \sim \pi_{\text{ABC},\varepsilon}(\theta, \bar{y}|\bar{y}_{\text{obs}}) \propto I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)\pi(\theta), \tag{2.23}$$

Algorithm 2.2 ABC-Rejection algorithm

while less than N acceptances **do**
  sample parameter $\theta \sim \pi(\theta)$
  simulate data $\bar{y} \sim p(\bar{y}|\theta)$
  accept θ if $d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon$
end while
output: samples $\{\theta_i\}_{i \leq N}$

where I denotes the indicator function, such that by projection to the first component, samples from the target marginal of interest, the ABC posterior

$$\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}}) \propto \int I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)\, d\bar{y} \cdot \pi(\theta), \tag{2.24}$$

can be obtained [Sisson et al., 2018, Chapter 4]. Specifically, while in the first two steps samples $(\theta, \bar{y}) \sim \pi(\bar{y}|\theta)\pi(\theta)$ are generated, in the third step Algorithm 2.2, thus also called ABC-Rejection, implements a (here binary) rejection step as in Proposition 2.1, such that accepted particles exactly target (2.23).

2.4.2 Asymptotic behavior

Monte-Carlo estimates based on samples $\{\theta_i\}_{i \leq N}$ from (2.23) converge to integrals over (2.24), $\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}})$, for $N \to \infty$, as in Section 2.2.3. In turn, $\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}})$ serves as an approximation to the actual parameter posterior distribution $\pi(\theta|\bar{y}_{\text{obs}}) \propto \pi(\bar{y}_{\text{obs}}|\theta)\pi(\theta)$. Defining the ABC likelihood

$$\pi_{\text{ABC},\varepsilon}(\bar{y}_{\text{obs}}|\theta) \propto \int I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)\, d\bar{y}, \tag{2.25}$$

we can write $\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}}) \propto \pi_{\text{ABC},\varepsilon}(\bar{y}_{\text{obs}}|\theta)\pi(\theta)$, and thus interpret ABC as regular Bayesian analysis with an approximated likelihood [Sisson et al., 2018, Chapter 2], as an average over likelihoods of simulation values close to $\bar{y}_{\text{obs}}$.

For discrete data, the observation that the chance of simulating $\bar{y} = \bar{y}_{\text{obs}}$ is given by the probability mass function $\pi(\bar{y}_{\text{obs}}|\theta)$, accepting if $\bar{y} = \bar{y}_{\text{obs}}$ in the third step gives rise to an exact likelihood-free algorithm [Rubin, 1984], while for continuous data this is not possible. Rather, for small ε, $\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}})$ is usually assumed to approximate $\pi(\theta|\bar{y}_{\text{obs}})$, with limit

$$\pi_{\text{ABC},\varepsilon}(\theta|\bar{y}_{\text{obs}}) \to \pi(\theta|\bar{y}_{\text{obs}}) \quad \text{for} \quad \varepsilon \to 0 \tag{2.26}$$

in an appropriate sense. Informally, with $K_\varepsilon(\bar{y}) := I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon]$, if $K_\varepsilon \to \delta_{\bar{y}_{\text{obs}}}$ converges to the Dirac delta distribution at $\bar{y}_{\text{obs}}$ for $\varepsilon \to 0$, then $\pi_{\text{ABC},\varepsilon}(\bar{y}_{\text{obs}}|\theta) \to \pi(\bar{y}_{\text{obs}}|\theta)$ such that (2.26) follows. This can be formulated more rigorously for general acceptance kernels K_ε e.g. in terms of generating functions of Dirac delta sequences [Kanwal, 1998]. For the most commonly used ABC formulation, with distance metric e.g. induced via an Lp norm, we give an explicit proof of (2.26) in Section 5.3.3.

2.4.3 Summary statistics

Due to the curse of dimensionality, as simulating $\bar{y} \approx \bar{y}_{\text{obs}}$ is unlikely for high-dimensional data, ABC inference is often based on a summary statistics representation $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_s}$ of the data, with $n_s \leq n_y$, the proximity $d(s(\bar{y}), s(\bar{y}_{\text{obs}}))$ of which is evaluated in the third step above, for a distance metric $d : \mathbb{R}^{n_s} \times \mathbb{R}^{n_s} \to \mathbb{R}_{\geq 0}$. Denoting $s = s(\bar{y})$, $s_{\text{obs}} = s(\bar{y}_{\text{obs}})$, and $\pi(s|\theta) \propto \int \delta_s(s(\bar{y}))\pi(\bar{y}|\theta)\, d\bar{y}$ the intractable summary statistics likelihood, the population of accepted particles constitutes a sample from the approximate posterior distribution

$$\pi_{\text{ABC}}(\theta|s_{\text{obs}}) \propto \int I[d(s, s_{\text{obs}}) \leq \varepsilon]\pi(s|\theta)\, ds \cdot \pi(\theta).$$

Only if s is sufficient, i.e. $\pi(\theta|\bar{y}) = \pi(\theta|s(\bar{y}))$, can the original posterior $\pi(\theta|\bar{y}_{\text{obs}})$ be recovered in the approximate limit $\varepsilon \to 0$. In practice, sufficient statistics are however scarcely known, such that commonly focus is on the construction of lower-dimensional, informative summaries of the data. Semi-automatic methods to derive such statistics have been developed, e.g. based on subset selection, selecting an optimal subset from a pool of candidate summary statistics [Joyce and Marjoram, 2008, Nunes and Balding, 2010], auxiliary likelihoods, using aforementioned indirect inference methods for summary statistic construction [Drovandi et al., 2011, Gleim and Pigorsch, 2013], or regression, learning an inverse mapping from data to parameters [Borowska et al., 2021, Fearnhead and Prangle, 2012, Jiang et al., 2017]. See e.g. Blum et al. [2013] or Sisson et al. [2018, Chapter 4] for further information.

In summary, ABC methods have four levels of approximation (adapted from Sisson et al. [2018, Section 1]):

- Firstly, every model is just an approximation of a real-world system.
- Secondly, as for every sampling method, there is a Monte-Carlo error due to finite sample size.
- Thirdly, also simulations not exactly matching the data are accepted, effectively using a different likelihood (2.25).
- Fourthly, insufficient summary statistics result in a loss of information.

Typically, there is a trade-off between the last two sources of errors, with good summary statistics being both low-dimensional, and informative [Blum et al., 2013]. Conversely, ABC methods allow to consider more realistic models of real-world systems, thus reducing the first error.

In this thesis, we will for simplicity of notation usually assume that the summary statistics mapping is already incorporated in y, unless mentioned otherwise.

2.4.4 Sequential Monte Carlo sampling

The above vanilla ABC algorithm exhibits a trade-off between decreasing the acceptance threshold ε to obtain a better posterior approximation, and maintaining high acceptance rates, further it suffers from the deficiencies common to plain rejection and importance sampling algorithms.

Algorithm 2.3 ABC-SMC algorithm

for $t = 1, \ldots, n_t$ **do**
  while less than N acceptances **do**
    sample parameter $\theta \sim g_t(\theta)$
    simulate data $\bar{y} \sim p(\bar{y}|\theta)$
    accept θ if $d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon_t$
  end while
  compute weights $w_i^t = \frac{\pi(\theta_i^t)}{g_t(\theta_i^t)}$, for accepted parameters $\{\theta_i^t\}_{i \leq N}$
  normalize weights $W_i^t = w_i^t / \sum_j w_j^t$
  define g_{t+1} and ε_{t+1} based on $\{(\theta_i^t, W_i^t)\}_{i \leq N}$
end for
output: weighted samples $\{(\theta_i^{n_t}, W_i^{n_t})\}_{i \leq N}$

Instead, any of the likelihood-based approaches mentioned in Section 2.2.3 can be "ABC-fied", commonly then targeting the joint posterior $(\theta, \bar{y}) \sim \pi_{\text{ABC},\varepsilon}(\theta, \bar{y}|\bar{y}_{\text{obs}})$. For example, a Metropolis-Hastings ABC-MCMC algorithm can be formulated by replacing (2.17) by

$$A(\theta'|\theta) = \min\left[1, \frac{I[d(\bar{y}', \bar{y}_{\text{obs}}) \leq \varepsilon]}{I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon]} \frac{\pi(\theta')}{\pi(\theta)} \frac{K(\theta|\theta')}{K(\theta'|\theta)}\right],$$

with proposal distribution $K((\theta', \bar{y}')|(\theta, \bar{y})) = \pi(\bar{y}'|\theta')K(\theta'|\theta)$, free of intractable likelihood terms [Bortot et al., 2007, Marjoram et al., 2003]. See e.g. Sisson et al. [2018, Chapter 4] for an overview of ABC samplers.

In this thesis, we will focus on samplers using sequential Monte Carlo (SMC) methods (see Section 2.2.3), based on Sisson et al. [2007], Toni et al. [2009]. SMC methods are particularly amenable to combination with ABC, as the acceptance threshold ε provides a natural base for intermediate distributions, allowing for a gradual decrease of ε and thus improvement of the posterior approximation, and avoiding prior tuning of hyperparameters. In ABC-SMC, a series of particle populations $P_t = \{(\theta_i^t, \bar{y}_i^t, w_i^t)\}_{i \leq N}$ with acceptance thresholds $\varepsilon_1 > \ldots > \varepsilon_{n_t}$, $t = 1, \ldots, n_t$, are generated, constituting samples of successively better posterior approximations (Algorithm 2.3). Particles for generation t are sampled from a proposal distribution $g_t(\theta) \gg \pi(\theta)$ based on the previous generation's accepted particles P_{t-1}, e.g. via a kernel density estimate, only initially $g_1(\theta) = \pi(\theta)$. The importance weights w_t^i are the corresponding non-normalized Radon-Nikodym derivatives (Proposition 2.3)

$$w_t(\theta) = \frac{\pi(\theta)}{g_t(\theta)}, \tag{2.27}$$

see Section 2.4.5 for details on the employed rejection-importance sampling algorithm and weighting scheme.

An acceptance threshold scheme that has been shown to perform well and is also employed in most parts of this work, is to adaptively set ε_t to a quantile, e.g. the median, of the previous generation's accepted distances [Drovandi and Pettitt, 2011, Klinger and Hasenauer, 2017]. In addition, we base the initial threshold on a calibration sample of size N, which is also used to calibrate any other components that require problem-specific adjustment. See e.g. Silk et al. [2013] for alternatives. A common form of the proposal distribution, which we also employ

Algorithm 2.4 ABC-IS algorithm

while less than N acceptances **do**
  sample parameter $\theta \sim g(\theta)$
  simulate data $\bar{y} \sim p(\bar{y}|\theta)$
  accept θ if $d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon$
end while
compute weights $w_i = \frac{\pi(\theta_i)}{g(\theta_i)}$, for accepted parameters $\{\theta_i\}_{i \leq N}$
normalize weights $W_i = w_i / \sum_j w_j$
output: weighted samples $\{(\theta_i, W_i)\}_{i \leq N}$

here, is based on a multivariate normal perturbation kernel, $g_t(\theta) \propto \sum_{i=1}^{N} \mathcal{N}(\theta|\theta_i^{t-1}, \Sigma_{t-1})$ with generation-specific covariance matrix $\Sigma_{t-1} \propto \Sigma(\{(\theta_i^{t-1}, w_i^{t-1})\}_{i \leq N})$ based on the previous generation's sample covariance matrix, thus exploiting correlations, which can further be localized [Filippi et al., 2013]. For discrete parameters, similar random-walk approaches can be employed [Syga et al., 2020]. For details on the underlying ABC-SMC implementation used here see e.g. Klinger and Hasenauer [2017], Klinger et al. [2018].

Advantages of ABC-SMC include the simple possibility to adjust components such as the aforementioned acceptance threshold [Drovandi and Pettitt, 2011, Silk et al., 2013] and distance metric [Harrison and Baker, 2020, Prangle, 2017] to the problem structure, as well as the same advantages as for SMC in general, as mentioned in Section 2.2.3, in particular adjustment of proposal distribution [Filippi et al., 2013] and ease of parallelization, which is paramount for computationally expensive models. Throughout this thesis, we employ parallelization based on dynamic scheduling [Klinger et al., 2018], which uses all available computational infrastructure for sampling at near-all times, afterwards correcting for simulation time bias, and which has been shown to substantially reduce the overall wall-time compared to established static scheduling, where only exactly as many particles as required are generated. Further, especially ABC-MCMC methods can suffer from bad mixing, as the step acceptance probability is proportional to the likelihood, requiring the simulation of data close to $\bar{y}_{\text{obs}}$ [Sisson et al., 2007].

2.4.5 Rejection-importance sampling

All ABC-SMC routines used in this thesis employ a sequence of importance sampling (ABC-IS) generations. For a proposal distribution $g(\theta) \gg \pi(\theta)$ and a target population size N, we use an ABC-IS scheme as shown in Algorithm 2.4. In ABC-SMC, Algorithm 2.4 is then iterated over successively refined acceptance thresholds ε, proposals $g(\theta)$, and, possibly, distances d as well as summary statistics. This form of ABC-IS generates samples from the distribution

$$(\theta, \bar{y}) \sim G(\theta, \bar{y}) \propto I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon] \pi(\bar{y}|\theta) g(\theta). \tag{2.28}$$

Here, the $g(\theta)$ is because we sample $\theta \sim g(\theta)$, additionally we simulate data $\bar{y}$ from the likelihood and discard particle not satisfying the acceptance criterion, such that, conditioned on θ, accepted simulations are distributed as

$$\bar{y}|\theta \sim I[d(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon] \pi(\bar{y}|\theta).$$

Therefore, the importance weights, Radon-Nikodym derivatives, of the proposal distribution (2.28) against the target (2.23) are given by

$$v(\theta, \bar{y}) = \frac{\pi_{\mathrm{ABC},\varepsilon}(\theta, \bar{y}|\bar{y}_{\mathrm{obs}})}{G(\theta, \bar{y})} = \frac{C \cdot I[d(\bar{y}, \bar{y}_{\mathrm{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)\pi(\theta)}{C_G \cdot I[d(\bar{y}, \bar{y}_{\mathrm{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)g(\theta)} = \frac{C}{C_G}\frac{\pi(\theta)}{g(\theta)} = \frac{C}{C_G}w(\theta),$$

with normalization constants C, C_G such that $\iint C \cdot I[d(\bar{y}, \bar{y}_{\mathrm{obs}} \leq \varepsilon]\pi(\bar{y}|\theta)\, d\bar{y} \cdot \pi(\theta)\, d\theta = 1 = \iint C_G \cdot I[d(\bar{y}, \bar{y}_{\mathrm{obs}}) \leq \varepsilon]\pi(\bar{y}|\theta)\, d\bar{y} \cdot g(\theta)\, d\theta$. Here, $w(\theta) = \pi(\theta)/g(\theta)$ denotes the importance weight factor (2.27) related to parameter sampling.

Supposing we have i.i.d. samples $(\theta_i, \bar{y}_i) \sim G(\theta, \bar{y})$, $i = 1, \ldots, N$, if all normalization constants were known, for a function $\xi : \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_y} \to \mathbb{R}$ an unbiased and asymptotically exact importance estimate over the ABC posterior could be obtained via

$$\mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon}(\theta,\bar{y}|\bar{y}_{\mathrm{obs}})}[\xi] = \mathbb{E}_{G(\theta,\bar{y})}[v\xi] \approx \frac{1}{N}\sum_{i \leq N} v_i \xi(\theta_i, \bar{y}_i),$$

with weights $v_i = v(\theta_i, \bar{y}_i)$, see Proposition 2.2. Typically, ξ will just be a function of θ, e.g. $\xi(\theta, \bar{y}) = \theta$ yielding the mean over the ABC parameter posterior distribution. In general, C/C_G is however not known and thus v only up to normalization, such that one has to resort to the estimate

$$\mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon}(\theta,\bar{y}|\bar{y}_{\mathrm{obs}})}[\xi] = \frac{\mathbb{E}_{G(\theta,\bar{y})}[w\xi]}{\mathbb{E}_{G(\theta,\bar{y})}[w]} \approx \frac{\frac{1}{N}\sum_{i \leq N} w_i \xi(\theta_i, \bar{y}_i)}{\frac{1}{N}\sum_{i \leq N} w_i} = \sum_{i \leq N} W_i \xi(\theta_i, \bar{y}_i), \tag{2.29}$$

with self-normalized weights

$$W_i := \frac{w_i}{\sum_{j \leq N} w_j},$$

as in Algorithms 2.3 and 2.4. The approximation (2.29) is in general no longer unbiased, but asymptotically so for $N \to \infty$ by Proposition 2.3 applied to the joint sampling space over parameters and data.

In Algorithm 2.4, we use a rejection step based on the distance value, inside the importance sampling routine. Alternatively, this step could be replaced by importance sampling, effectively giving a proposal distribution $G(\theta, \bar{y}) \propto \pi(\bar{y}|\theta)\pi(\theta)$, such that the acceptance kernel becomes part of the weighting, giving weights $v(\theta, \bar{y}) \propto I[d(\bar{y}, \bar{y}_{\mathrm{obs}}) \leq \varepsilon]w(\theta)$, which would result in a number of samples generated equal to N, however in a high weight variability and low ESS (2.16), with many weights at zero. A similar interplay of rejection and importance sampling on different levels is source to many variations of ABC algorithms [Sisson et al., 2018, Chapter 4].

2.4.6 Further topics

ABC is an active research field with many developments, applications and algorithms beyond what has been and will be discussed in this thesis. For example, various strategies for MCMC and SMC algorithms as discussed in Section 2.2.3 can be transferred to an ABC context [Sisson et al., 2018], post-processing of results to account for simulation mismatch via regression can be performed [Beaumont et al., 2002], model selection can be formulated as a hierarchical problem [Toni et al., 2009], low-fidelity models can be used to determine whether costly simulations should

be performed [Prescott and Baker, 2021], the acceptance step can be re-interpreted as introducing a noise term [Wilkinson, 2013] (see also Chapter 4), or components such as population size [Klinger and Hasenauer, 2017] or distance metric [Prangle, 2017] can be automatically adjusted to the problem (see also Chapter 5). See e.g. Sisson et al. [2018] for an overview of ABC methods, history, and recent developments.

2.5 Standardization

It is important to ensure findability, accessibility and reusability of data and models, interoperability of analysis pipelines, and reproducibility of results [Wilkinson et al., 2016]. In the field of systems biology, thus several community standards have been defined [Stanford et al., 2019].

In particular, biochemical process models can be defined in the Systems Biology Markup Language (SBML) [Hucka et al., 2003], CellML [Cuellar et al., 2003], or the BioNetGen Language (BNGL, for rule-based formulations) [Harris et al., 2016]. Atop model definition, simulation experiments can be described via the Simulation Experiment Description Markup Language (SED-ML) [Waltemath et al., 2011], or similarly the Systems Biology Results Markup Language (SBRML) [Dada et al., 2010]. Various software tools, e.g. for visualization, simulation or inference, support such model standards as exchange format (e.g. Balsa-Canto and Banga [2011], Fröhlich et al. [2021], Hoops et al. [2006], Raue et al. [2015]). Furthermore, both data (e.g. Barretina et al. [2012], Eduati et al. [2017], Li et al. [2017]) and models (e.g. Li et al. [2010], Olivier and Snoep [2004]) have been made available in large online databases, which supports reusability.

Complementary to existing standards, we have in Schmiester et al. [2021a] developed the parameter estimation format PEtab (`https://petab.readthedocs.io`). It allows to describe all components required to define a parameter estimation problem, including data, observation and noise model, parameter boundaries and scales, priors, and condition-specific mappings, via human- and machine-readable TSV files, besides the model specification via SBML. It is supported by currently nine systems biology toolboxes, namely COPASI [Hoops et al., 2006], Data2Dynamics [Raue et al., 2015], dMod [Kaschek et al., 2019], parPE [Schmiester et al., 2019], MEIGO [Egea et al., 2014], AMICI [Fröhlich et al., 2021], pyPESTO [Schälte et al., 2021b], and pyABC [Klinger et al., 2018], as well as more recently SBML2Julia [Lang et al., 2020]. Initially focused on single-model ODE problems, extensions to encompass e.g. model selection problems and spatial multi-cellular or multi-scale models [Starruß et al., 2014] are underway (not published yet). Further, we have in Vanhoefer et al. [2021] developed a format for manual and programmatic ODE model construction in YAML with export to executable SBML models, and PEtab parameter inference problems.

Chapter 3

Scalable hierarchical optimization using adjoint sensitivity analysis

Mechanistic modeling of biochemical processes using ordinary differential equation models facilitates the quantitative understanding of biological processes. Usually, a model possesses unknown parameters that need to be estimated from experimental data. In order to get a more holistic understanding of processes and reaction networks, comprehensive models comprising hundreds to thousands of species and parameters have been developed. Parameter estimation for such large-scale models is challenging, in particular as data are often only available on a relative scale and noise levels are unknown. Thus, scaling parameters and noise parameters must be estimated alongside model parameters influencing the process dynamics.

In this chapter, we present a novel adjoint hierarchical optimization method combining the efficient analytical calculation of optimal scaling, offset, and noise model parameters via a hierarchical formulation with the scalable, efficient evaluation of objective function gradients using adjoint sensitivity analysis. We evaluate the novel approach on a pan-cancer model with > 1000 states and > 4000 parameters [Fröhlich et al., 2018], demonstrating that it provides a substantial improvement of optimizer performance, and facilitates the estimation of scale, offset and noise parameters with virtually no computational overhead.

This chapter is based on and partly identical to the following shared first-author publication of the thesis author:

- Schmiester, L.*, **Schälte, Y.***, Fröhlich, F., Hasenauer, J. and Weindl, D. (2019). Efficient parameterization of large-scale dynamic models based on relative measurements. Bioinformatics, 36(2), pp.594-602. (*equal contribution)

In this publication, the thesis author developed the hierarchical optimization approach with efficient gradients based on adjoint sensitivities and analytical solutions for optimal affine observable transformations and noise parameters, which we focus on here. The implementation was performed jointly with, and the evaluation on a high-dimensional application problem mainly by, Leonard Schmiester and Daniel Weindl.

3.1 Introduction

Ordinary differential equation (ODE) models are widely used in systems biology to mechanistically describe and understand the temporal evolution of biochemical reaction networks (Section 2.1.4). Growing biological knowledge, efficient simulation routines, and increasing, in particular parallel, computational power have facilitated the development of larger and more comprehensive models, with up to thousands of state variables and parameters [Bouhaddou et al., 2018, Fröhlich et al., 2018]. Such large-scale models give a more detailed and holistic understanding at a system level, which decoupled or simplified models cannot provide [Kitano, 2002a], e.g. allowing to capture pathway cross-talk, or to understand and predict drug effects and combinations [Fröhlich et al., 2018, Hass et al., 2017, Korkut et al., 2015]. However, with increasing numbers of state variables and parameters, parameterization of large-scale models gets challenging [Kapfer et al., 2019].

Commonly, data-driven parameter estimation is based on the quantitative comparison of simulated and measured observables via a noise model (Section 2.1.2). Large-scale transcriptomics, proteomics, and pharmacological data sets have been acquired and made available in public databases (e.g. Barretina et al. [2012], Eduati et al. [2017], Li et al. [2017]). However, such datasets often contain only measurement values relative to the absolute concentrations of the species under consideration, often requiring an unknown affine transformation, via scaling factors and offsets, to enable the quantitative comparison of model simulations and data [Degasperi et al., 2017, Raue et al., 2013b, Weber et al., 2011]. For example, frequently employed measurement techniques such as Western blotting [Renart et al., 1979] or flow cytometry [Herzenberg et al., 2006] only provide relative information about absolute molecular quantities, in addition measurements in large-scale databases are often normalized and thus also only relative. An alternative approach occasionally employed is the comparison of relative changes only, termed data-driven normalization by Degasperi et al. [2017], by dividing simulations and data in the same way by a control. While this does not increase the problem dimension, the noise distribution is non-trivial and residuals are not uncorrelated, which can result in incorrect confidence intervals (see e.g. Thomaseth and Radde [2016] and the supplementary information of Loos et al. [2018]). In addition to the problem of relative data, measurements are usually noise-corrupted. When no reliable estimates of noise levels are available, e.g. via experimental replicates, corresponding noise parameters can also be learned simultaneously with the model dynamics. Thus, frequently parameters describing observable transformations, and parameters quantifying noise levels must be learned alongside parameters governing the dynamics of the modeled states (2.1). This increases the parameter dimension (e.g. by a factor of two in Bachmann et al. [2011]), and it was shown that even the presence of only a few scaling parameters can substantially harm optimizer performance [Degasperi et al., 2017].

To facilitate inference for large-scale ODE models, scalable methods have been developed, in particular exploiting parallelization [Fröhlich et al., 2018, Penas et al., 2015], and adjoint sensitivity analysis (ASA) for efficient gradient computation (Section 2.3.2), permitting efficient local optimization (Section 2.3.1).

To improve inference in the special case of data with unknown linear transformations, and assuming additive normal noise with known standard deviations, Weber et al. [2011] developed a hierarchical optimization approach that, instead of estimating parameters governing model dy-

namics and scaling parameters defining the observable transformation jointly, exploits the fact that for given dynamics parameters and model simulations, the optimal scaling parameters can be computed analytically, improving optimizer convergence for the reduced problem of estimating the remaining parameters. This approach was generalized by Loos et al. [2018], who provided analytical solutions for both scaling factors and noise parameters, for different noise distributions, in particular additive normal and Laplace noise. However, their approach only considers scaling factors, but not offsets, which are often required to e.g. describe background noise, or when normalization such as mean subtraction was applied. Further, their approach is not compatible with ASA, but only forward sensitivity analysis (FSA), which is computationally prohibitive for large-scale models [Fröhlich et al., 2018].

In this chapter, we firstly take a more general perspective on hierarchical optimization, providing insights into how splitting up a problem hierarchically can be beneficial. Then, for ODE models, we demonstrate how a general hierarchical problem can be integrated with FSA and, in particular, ASA, permitting efficient gradient-based optimization for high-dimensional problems. Complementarily, we provide analytical optimal solutions for the specific case of affine observable transformations and unknown additive normal noise levels, i.e. for scaling factors, offsets, and noise parameters, yielding a highly efficient solution of the corresponding hierarchical inner sub-problem. We evaluate the novel hierarchical approach on a large-scale pan-cancer signaling model. Compared to standard optimization, we demonstrate substantially improved performance in terms of parameter estimates and fits. In particular, the hierarchical approach provides an unbiased way to weight heterogeneous data sets, and estimate observable and noise parameters with negligible computational overhead.

3.2 Hierarchical optimization with adjoint sensitivity analysis

In this chapter, we employ the notation of ODE constrained optimization problem from Section 2.1.4, with modifications to encompass the hierarchical approach presented here, in particular splitting the parameter vector hierarchically into outer and inner parameters. While in Section 3.2.1 we consider a generic objective function, it is thereafter based on the inference problem (2.18) as a negative log-likelihood, with an underlying ODE model and specific observation function and noise model.

3.2.1 A general hierarchical optimization problem

Splitting the parameter vector $\theta = (\psi, \eta)$ into *outer* parameters ψ and *inner* parameters η, we consider for an objective function $J : V \subset \mathbb{R}^{n_\psi} \times \mathbb{R}^{n_\eta} \to \mathbb{R}$, $(\psi, \eta) \mapsto J(\psi, \eta)$ the minimization problem

$$\inf_{\psi,\eta} J(\psi, \eta), \tag{3.1}$$

over a suitable domain V. Instead of solving (3.1) jointly, we decompose the problem hierarchically as

$$\inf_{\psi} \hat{J}(\psi) \quad \text{where} \quad \hat{J}(\psi) := \inf_{\eta} J(\psi, \eta), \tag{3.2}$$

i.e. into an outer problem of minimizing over ψ the inner problem of minimizing over η conditioned on ψ. Problems (3.1) and (3.2) are clearly equivalent.

Assuming overall infimum and all conditional infima to be finite and assumed, which will be the case in our applications, we can further define

$$\hat{\eta}(\psi) := \arg\min_{\eta} J(\psi, \eta) \quad \text{such that} \quad \hat{J}(\psi) = J(\psi, \hat{\eta}(\psi)). \tag{3.3}$$

Note that in general $\hat{\eta}(\psi)$ need not be uniquely defined.

In practice, a hierarchical decomposition can be in particular beneficial if the inner problem can be efficiently solved, i.e. the conditional minimum given ψ efficiently calculated, as then the outer problem is of lower dimension and potentially simpler than the joint problem. When iterative numerical optimization as outlined in Section 2.3.1 is used to solve the joint or outer, and potentially the inner, problem, practically the hierarchical problem translates to an iterative two-step procedure, where in an outer loop parameters ψ are proposed, conditional on which the inner problem is solved for η, either numerically in an inner loop, or analytically, yielding the corresponding function value $\hat{J}(\psi)$ as well as potentially derivative information, such as the gradient $\nabla\hat{J}(\psi)$, to be used in the outer loop to guide exploration for the next step.

It is of interest whether a numerical optimizer will converge to the same local minimum for the joint and the hierarchical problem. This can in general not be answered, as it depends on the properties of both J and the employed optimizer. Effectively, the objective function becomes a different one. It is straightforward to construct problems with multi-modal inner problems, in which local optimizers started from the same point (projected onto ψ for the hierarchical problem) will converge to different local minima in both approaches, because the inner problem may focus on a different valley of attraction. However, in general the reduction of dimension of the hierarchical outer problem compared to the original joint one, as well as the objective function landscape simplification as a profile projection can be expected to yield improvements in optimizer convergence, as in practice e.g. confirmed by Loos et al. [2018], Weber et al. [2011]. In particular, there is a surjection of local minima from the original problem to the hierarchical outer one, which becomes a bijection if the inner problem is uni-modal with unique minimum.

To solve the joint and the hierarchical outer problem, and potentially the hierarchical inner problem if an analytical solution is not available, derivatives are highly useful to guide numerical local searches. As we will thus later be interested in calculating derivatives for hierarchical problems, the following observation is central:

Theorem 3.1. *Assume that J and $\hat{\eta}$ are k times continuously differentiable, and the conditional minimum satisfies*

$$\partial_\eta J(\psi, \hat{\eta}(\psi)) = 0 \tag{3.4}$$

for all ψ. Then

$$\nabla^\kappa \hat{J}(\psi) = \partial_\psi^\kappa J(\psi, \hat{\eta}(\psi))$$

for all $\kappa \leq k$.

Proof. It is by chain rule

$$\nabla\hat{J}(\psi) = \partial_\psi J(\psi, \hat{\eta}(\psi)) + \partial_\eta J(\psi, \hat{\eta}(\psi))\nabla_\psi\hat{\eta}(\psi) = \partial_\psi J(\psi, \hat{\eta}(\psi)),$$

due to (3.4). The general statement for higher-order derivatives follows by induction and Schwarz' Theorem. □

This implies that e.g. gradient and Hessian of $\hat{J}$ are just the projections of the corresponding derivatives of J to the first component, such that the objective function landscape relevant for e.g. optimization of the hierarchical outer problem, such as convexity, is based on and simplified from the original joint one. Note that (3.4) is satisfied in particular if $\hat{\eta}(\psi) \in V \backslash \partial V$ is assumed in the interior of V, by necessary condition of a local minimum.

In our applications, $\hat{\eta}$ will be constructively given. However, in general note that assuming e.g. that J is twice continuously differentiable and convex in the second component, i.e. $\partial_\eta^2 J(\psi, \eta) > 0$ for all (ψ, η), and that the conditional minimum $\min_\eta J(\psi, \eta)$ is assumed in the interior $V \backslash \partial V$ of V for all ψ, the Implicit Function Theorem gives existence and (global) uniqueness of a continuously differentiable function $\hat{\eta}(\psi)$ such that $\partial_\eta J(\psi, \eta) = 0 \Leftrightarrow \eta = \hat{\eta}(\psi)$, with $\hat{J}(\psi) = J(\psi, \hat{\eta}(\psi))$.

3.2.2 ODE constrained hierarchical problems with efficient inner problem

In this section, we formulate a generic hierarchical problem based on a biochemical ODE model as introduced in Section 2.1.4. Splitting the parameter vector $\theta = (\psi, \eta)$, we assume that we can decompose the observation function (2.2) as

$$y(t, \psi, \eta) = h(\tilde{h}(x(t, \psi), \psi), \eta),$$

where

$$\tilde{y}(t, \psi) = \tilde{h}(x(t, \psi), \psi) \tag{3.5}$$

denotes the mapping from states x to raw, unscaled, observables $\tilde{y}$, which need to be transformed by h in order to allow for a comparison to the observed data $\bar{y}_{\text{obs}}$. Here, states x and unscaled observables $\tilde{y}$ depend on outer parameters ψ instead of the full vector θ, while h directly depends only on inner parameters η. Similar to (2.19), we then consider a generic objective function

$$J(\psi, \eta) = \sum_{i=1}^{n_t} J_i(\tilde{y}(t_i, \psi), \psi, \eta),$$

where in particular noise parameters could be included in η instead of ψ. Note that often, $\tilde{h}$ and J_j will not directly depend on ψ, however the here used formulation allows for greater flexibility, permitting to not treat all parameters not directly involved in the simulation of x as inner parameters. Further, we now regard J_i as a function of the unscaled observables $\tilde{y}$, such that, besides the noise model, also the observable transformation is considered as a part of J_i here, for ease of notation.

Assuming all conditional minima of the hierarchical problem $\min_\psi \min_\eta J(\psi, \eta)$ to exist, we can, as in (3.3), further define

$$\hat{\eta}(\psi) = \arg\min_\eta J(\psi, \eta)$$

and

$$\hat{J}(\psi) = \min_{\eta} J(\psi, \eta) = J(\psi, \hat{\eta}(\psi)). \tag{3.6}$$

3.2.3 Relative measurements

To provide context prior to further discussion, let us introduce the specific hierarchical problem of interest in our applications. When measured data are only available on a relative scale, often an affine transformation can be used to map model outputs to data. While this is usually implicitly incorporated in the observation function h and parameters θ, here we use an explicit formulation, in order to decompose the inference problem later. With scaling factors s and offsets b, we assume that observed data $\bar{y}_{\text{obs}} = \{\bar{y}_{\text{obs},i}\}_{i \leq n_y}$ can be compared to rescaled simulated observables

$$y_i = s_i \cdot \tilde{y}_i + b_i,$$

where $\tilde{y}$ denotes unscaled observables (3.5) derived from x. Assuming additive normal noise of variance σ_i, the negative log-likelihood is then explicitly given as

$$J(\psi, \eta) = \frac{1}{2} \sum_{i \leq n_y} \left[\log(2\pi\sigma_i^2) + \left(\frac{\bar{y}_{\text{obs},i} - (s_i \cdot \tilde{y}_i + b_i)}{\sigma_i} \right)^2 \right], \tag{3.7}$$

where $\tilde{y}_i = \tilde{y}_i(\psi)$ and $s_i, b_i, \sigma_i = s_i(\eta), b_i(\eta), \sigma_i(\eta)$ are, if they are unknown, incorporated via projection in additional parameters η, besides the parameters ψ governing the model dynamics, so that the overall parameter vector is $\theta = (\psi, \eta)$.

Commonly, experimental data are available for various observables, at various discrete time points, and potentially for various experimental conditions or replicate measurements. Often, observable transformations and noise parameters are assumed to coincide for multiple such data points, e.g. when the same normalization is applied to all data points within an experiment. In that case, only a single parameter is to be incorporated in η, with projections coinciding across the corresponding data points, e.g. $s_{i_1} = s_{i_2} = \eta_i$ for some i, for data points i_1, i_2 sharing the same scaling factor.

In the standard approach, dynamics parameters, observable transformation parameters, and noise parameters are optimized simultaneously, e.g. via multi-start local optimization, see Section 2.3.1. However, we can exploit the specific structure of the problem by considering it hierarchically. The time-critical part in calculating the objective function is usually the numerical simulation of x. As observable transformation parameters and noise parameters do not affect x, they can be efficiently optimized conditional on a single simulation of x for given outer parameters ψ. The same applies to the generic hierarchical problem in Section 3.2.2. Given their simple functional form, in some scenarios even analytic formulas for the inner problem are available, see Section 3.2.5.

3.2.4 Derivatives for ODE problems with exactly solved inner problem

In this section, we are concerned with efficient methods of calculating gradients for the outer parameters of an ODE constrained hierarchical problem, which is crucial for efficient and reliable parameter optimization (see Section 2.3). We assume J and $\hat{\eta}$ as defined in Section 3.2.2 to be

Algorithm 3.1 Hierarchical optimization with exactly solved inner problem, using forward sensitivity analysis

1. compute $\tilde{y}$, $\frac{d\tilde{y}}{d\psi}$ via the augmented ODE system using FSA, given ψ
2. compute $\hat{\eta}(\psi)$, given ψ, $\tilde{y}$
3. compute $\hat{J}(\psi)$, $\nabla\hat{J}(\psi)$ via (3.6) and (3.9), given ψ, $\tilde{y}$, $\hat{\eta}(\psi)$, $\frac{d\tilde{y}}{d\psi}$

sufficiently differentiable.

Easily possible is the use of finite differences. However, as these are neither robust nor efficient, in the following we present routines using forward (FSA) and adjoint sensitivity analysis (ASA) (Section 2.3.2).

Forward sensitivity analysis

Similar to (2.20), we can write the gradient of (3.6) as

$$\nabla\hat{J}(\psi) = \frac{\partial J}{\partial\psi} + \frac{\partial J}{\partial\eta}\frac{d\hat{\eta}}{d\psi} = \sum_{i=1}^{n_t}\left[\frac{\partial J_i}{\partial\psi} + \frac{\partial J_i}{\partial\tilde{y}}\frac{d\tilde{y}}{d\psi} + \frac{\partial J_i}{\partial\eta}\frac{d\hat{\eta}}{d\psi}\right], \tag{3.8}$$

dropping all obvious arguments for ease of notation. Assuming that the inner problem was solved such that $\partial_\eta J(\psi, \hat{\eta}(\psi)) = 0$, which we refer to as solving the inner problem exactly, (3.8) simplifies by Theorem 3.1 to

$$\nabla\hat{J}(\psi) = \sum_{i=1}^{n_t}\left[\frac{\partial J_i}{\partial\psi} + \frac{\partial J_i}{\partial\tilde{y}}\frac{d\tilde{y}}{d\psi}\right]. \tag{3.9}$$

Here, $\frac{d\tilde{y}}{d\psi} = \frac{\partial\tilde{h}}{\partial\psi} + \frac{\partial\tilde{h}}{\partial x}\frac{dx}{d\psi}$ depends on the state sensitivities, as in (2.20). In practice, we can thus, in an optimizer iteration, for a given ψ, calculate $\hat{J}(\psi)$ and $\nabla\hat{J}(\psi)$ using FSA as outlined in Algorithm 3.1: First, only the unscaled observables $\tilde{y}$ and their sensitivities $\frac{d\tilde{y}}{d\psi}$ are computed, involving a simulation of the model states x. Then, based on $\tilde{y}$, the conditionally optimal inner parameters $\hat{\eta}(\psi)$ are calculated. In the last step, all parts are combined to calculate the transformed observables and thence the objective function value $\hat{J}(\psi)$, as well as its derivative $\nabla\hat{J}(\psi)$. This is essentially what Loos et al. [2018] did for a specific inner problem and specific noise models, thus here recovered in a more general context.

Adjoint sensitivity analysis

For large-scale models, the computational cost of FSA is prohibitive, such that ASA would be preferable for the hierarchical outer problem. A problem however is that the terminal conditions of the adjoint state (2.21), updated at each measurement time point, depend on the data $\bar{y}_{\text{obs}}$ and the transformed simulated observables y via the objective function values J_i. In turn, y depends, unlike $\tilde{y}$, on the inner parameters η, as does J_i directly via noise parameters. Thus, we cannot first simulate the ODE and then compute optimal inner parameters and assemble the objective function gradient, as done with FSA.

Algorithm 3.2 Hierarchical optimization with exactly solved inner problem, using adjoint sensitivity analysis

1. compute $\tilde{y}$ via the state ODE (2.1), given ψ
2. compute $\hat{\eta}(\psi)$, given ψ, $\tilde{y}$
3. compute adjoint state p via the adjoint ODE (2.21), given $\tilde{y}$, $\hat{\eta}(\psi)$
4. compute $\hat{J}(\psi)$, $\nabla\hat{J}(\psi)$ via (3.6) and (2.22), given ψ, $\tilde{y}$, $\hat{\eta}(\psi)$, p

Yet, the combination of hierarchical optimization with ASA is possible. The core insight is that, as long as condition (3.4) is fulfilled, i.e. the inner problem solved exactly, can the inner parameters be treated, once known, as constants in the calculation of $\nabla\hat{J}(\psi) = \partial_\psi J(\psi, \hat{\eta}(\psi))$ via Theorem 3.1 as the partial derivative of J with respect to ψ. More concretely, (3.9) does not involve derivatives with respect to η and thus has virtually the same form as (2.20), with h replaced by $\tilde{h}$ and θ by ψ. Explicitly, the adjoint state (2.21) here takes the form

$$\dot{p} = -\frac{\partial f}{\partial x}^T p \quad \text{with terminal condition} \quad p(t_{i+1}, \psi) = \lim_{t \searrow t_{i+1}} p(t, \psi) - \frac{\partial J_{i+1}}{\partial \tilde{y}} \frac{\partial \tilde{h}}{\partial x}$$

on $(t_i, t_{i+1}]$, for $i = 0, \ldots, n_t - 1$, with p implicitly dependent on $\hat{\eta}(\psi)$ via J_{i+1}, such that the objective gradient can be written as

$$\nabla\hat{J}(\psi) = \sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \psi} + \frac{\partial J_i}{\partial \tilde{y}} \frac{\partial \tilde{h}}{\partial \psi} \right] - p^T \frac{dx_0}{d\psi}\bigg|_{t=t_0} - \int_{t_0}^{t_{n_t}} p^T \frac{\partial f}{\partial \psi}\, dt\,.$$

Thus, we can proceed as outlined in Algorithm 3.2: First, given outer parameters ψ, we simulate $\tilde{y}$, however without sensitivities. Having then derived $\hat{\eta}(\psi)$, we can simulate the adjoint state p as in (2.21) with h replaced by $\tilde{h}$, and can thereby calculate $\nabla\hat{J}(\psi)$ using ASA. The procedure is also visualized in Figure 3.1 for the specific case of an affine observable transformation.

To implement Algorithm 3.2, we can represent $\hat{\eta}(\psi)$ as fixed inputs, which common ODE simulation and sensitivity calculation tools such as AMICI already support besides parameters, in order to accommodate for e.g. condition-specific variables defined at run-time. While Algorithm 3.2 solves the forward and adjoint equations exactly once and thus has no computational overhead over the standard joint approach, a simpler implementation with some computational overhead is possible by in the third step simply re-simulating the forward equations besides the adjoint ones, with adjusted inputs $\hat{\eta}$. Similarly, in Algorithm 3.1 in the first step only $\tilde{y}$ can be simulated, and in the third step re-simulated alongside the FSA equations, with $\hat{\eta}(\psi)$ encoded as fixed inputs, yielding $\hat{J}(\psi)$ and $\nabla\hat{J}(\psi)$. This gives for both FSA and ASA, at slight computational overhead, easily implementable algorithms, without the need to manually assemble results, especially the gradient.

3.2.5 Analytical formulas for optimal scalings, offsets, and normal noise variances

While we discussed in the previous section the efficient gradient calculation for the outer problem, in this section we derive an explicit analytical solution to the inner problem in specific situations, thus yielding a highly efficient solution. We consider the setting of Section 3.2.3 of unknown

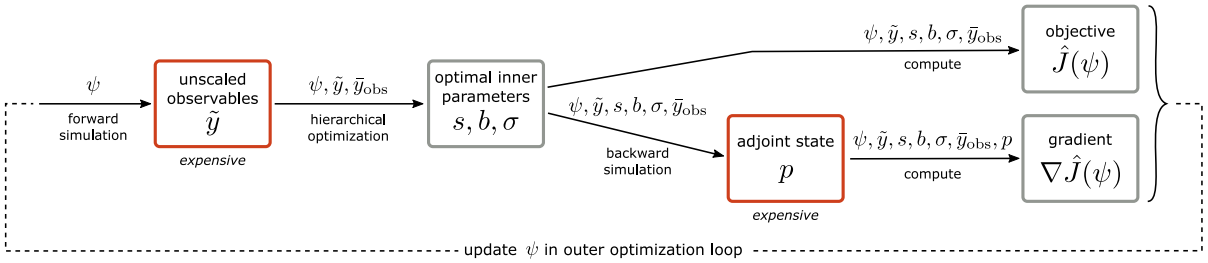

Figure 3.1: Illustration of the hierarchical optimization scheme using adjoint sensitivities. In the outer loop, ψ is updated by the employed iterative gradient-based optimization method. When a new value of ψ is proposed, an inner loop is entered, in which unscaled observables $\tilde{y}$ are simulated, and optimal inner parameters $\eta = (s, b, \sigma)$ conditional on ψ computed. Then, the adjoint state p is simulated, and objective function value $\hat{J}(\psi)$ and gradient $\nabla \hat{J}(\psi)$ are returned. Here, the solution of the inner problem is shown in detail. The red boxes involve the simulation of ODEs and are thus computationally more expensive. If the gradient is not required in an optimizer iteration, the adjoint and gradient steps can be omitted. This figure is adapted from the author's publication Schmiester et al. [2019].

scaling factors, offsets, and noise parameters for additive normal noise.

As already outlined, often observable and noise parameters will be shared across data points. Denote index sets $I^s_{i_s}, I^b_{i_b}, I^\sigma_{i_\sigma}$, for $i_s = 1, \dots, n_s$, $i_b = 1, \dots, n_b$, $i_\sigma = 1, \dots, n_\sigma$, for data points sharing, respectively, scaling factors, offsets, and noise parameters, i.e. e.g. $s_i = s_{i_s}$ for all $i \in I^s_{i_s}$. We will make the following assumptions, necessary for the validity of the here derived analytical solutions. We assume firstly that $\{I^s_{i_s}\}_{i_s} = \{I^b_{i_b}\}_{i_b}$, i.e. scaling factors and offsets are shared across the same data points, and secondly that for every i_s, there is an i_σ with $I^s_{i_s} \subset I^\sigma_{i_\sigma}$, i.e. data points that share a scaling parameter also share the noise parameter. For a subset on which σ is externally provided, e.g. calculated from experimental replicates, the second assumption can be dropped, as can the first assumption when either scalings or offsets do not exist or are externally provided. From an application point of view, both assumptions constitute usually no restriction, while there are counter-examples. In particular, scaling factors and offsets are typically shared among the same data points, as required by the first assumption, while the second assumption states that noise parameters are assigned on the same level as observable transformations, or more coarsely.

Scaling factors and offsets

Consider single scaling factor and offset parameters, i.e. objective (3.7) restricted to a subset of indices I such that $s_i \equiv s$, $b_i \equiv b$, and, by the assumptions, $\sigma_i \equiv \sigma$, or noise parameters are pre-defined. Bare constraints on the inner parameters, a necessary condition for a local optimum is $\partial_{s,b} J = 0$, giving

$$0 = \partial_s J = -\sum_i \frac{\bar{y}_{\text{obs},i} - (s\tilde{y}_i + b)}{\sigma_i^2} \tilde{y}_i \Rightarrow s = \left(\sum_i \frac{\tilde{y}_i^2}{\sigma_i^2} \right)^{-1} \left(\sum_i \frac{(\bar{y}_{\text{obs},i} - b)\tilde{y}_i}{\sigma_i^2} \right) \tag{3.10}$$

and

$$0 = \partial_b J = -\sum_i \frac{\bar{y}_{\text{obs},i} - (s\tilde{y}_i + b)}{\sigma_i^2} \Rightarrow b = \left(\sum_i \frac{1}{\sigma_i^2} \right)^{-1} \left(\sum_i \frac{\bar{y}_{\text{obs},i} - s\tilde{y}_i}{\sigma_i^2} \right). \tag{3.11}$$

Note that if noise parameters are estimated, by assumption $\sigma_i \equiv \sigma$, such that they drop out. If either scaling or offset is fixed, and in that case not necessarily all to the same value, we are done. In particular, for $b = 0$, (3.10) recovers the result by Weber et al. [2011]. If both scaling factors and offsets are to be optimized, inserting (3.10) into (3.11) gives

$$\left[1 - \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1} \left(\frac{\sum_i \frac{\tilde{y}_i}{\sigma_i^2} \sum_j \frac{\tilde{y}_j}{\sigma_j^2}}{\sum_i \frac{\tilde{y}_i^2}{\sigma_i^2}}\right)\right] b = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1} \left(\sum_i \frac{\bar{y}_{\text{obs},i}}{\sigma_i^2} - \frac{\sum_i \frac{\bar{y}_{\text{obs},i}\tilde{y}_i}{\sigma_i^2} \sum_j \frac{\tilde{y}_j}{\sigma_j^2}}{\sum_i \frac{\tilde{y}_i^2}{\sigma_i^2}}\right).$$

This defines b if the left-hand factor is non-zero. Otherwise, the simulations are not diverse enough to identify both s and b, in which case we can simply fix $b = 0$. In either case, s is consequently determined via (3.10).

Inspection of the second-order derivatives

$$\partial_s^2 J = \sum_i \frac{\tilde{y}_i^2}{\sigma_i^2} \geq 0 \quad \text{and} \quad \partial_b^2 J = \sum_i \frac{1}{\sigma_i^2} > 0$$

shows by sufficiency that the found points are indeed local minima, global due to convexity of the inner problems.

Noise parameters

All scaling factors s_i and offsets b_i calculated, similarly consider a single noise parameter, i.e. objective (3.7) restricted to a subset of indices I such that $\sigma_i \equiv \sigma$. The necessary condition $\partial_{\sigma^2} J$ for a local optimum gives

$$0 = \partial_{\sigma^2} J = \frac{1}{2} \sum_i \left[\frac{1}{\sigma^2} - \frac{(\bar{y}_{\text{obs},i} - (s_i\tilde{y}_i + b_i))^2}{(\sigma^2)^2}\right] \Rightarrow \sigma^2 = \left(\sum_i 1\right)^{-1} \left(\sum_i (\bar{y}_{\text{obs},i} - (s_i\tilde{y}_i + b_i))^2\right). \tag{3.12}$$

Inspection of the second derivative at the optimal point gives

$$\partial_{\sigma^2}^2 J = \frac{1}{2} \sum_i \frac{2(\bar{y}_{\text{obs},i} - (s_i\tilde{y}_i + b_i))^2 - \sigma^2}{(\sigma^2)^3} = \frac{1}{2} \frac{\sum_i 1}{(\sigma^2)^2} > 0,$$

i.e. we have a local minimum. An exception is if the optimal $\sigma^2 = 0$, i.e. the transformed simulated and observed data exactly match, which is however unlikely to occur in practice. In particular, for $b = 0$ and estimated noise parameters, (3.10) and (3.12) recover a result by Loos et al. [2018].

As no constraints were imposed on the range of values the scaling factors, offsets, and noise parameters can take, the inner problem was solved exactly by necessary conditions of a local minimum, such that the requirements of Section 3.2.4 for efficient gradient calculation using FSA or ASA apply. Further, the inner problem is uni-modal, such that there is a bijection of local minima of the joint and the outer hierarchical problem.

3.2.6 Derivatives for ODE problems with inexactly solved inner problem

So far, we have only considered ODE constrained hierarchical problems such that the inner problem satisfies condition (3.4), i.e. $\partial_\eta J(\psi, \hat{\eta}(\psi)) = 0$, in which case (3.8) simplifies to (3.9), and the derivative calculation approaches using in particular ASA as presented in Section 3.2.4 are applicable, as the inner parameters can be regarded as fixed when calculating the partial derivative of the outer problem. However, it can occur that (3.4) is not satisfied. In particular, constraints may be imposed on the range of values the inner parameters η can take, e.g. restrictions to numerically feasible values, or to biologically plausible regimes, similar to box constraints commonly applied to ψ. We refer to $\partial_\eta J(\psi, \hat{\eta}(\psi)) \neq 0$ for brevity as solving the inner problem inexactly, while it may obviously still be the exact solution under constraints. It is desirable to be able to efficiently calculate accurate derivatives for the outer problem in this case too. In this section, we present approaches of doing so. Note that this is rather a conceptual excursion, going beyond what is discussed and applied in Schmiester et al. [2019].

The problem in case of an inexactly solved inner problem is that in (3.8), the terms $\frac{d\hat{\eta}}{d\psi}$, with $\hat{\eta}(\psi)$ the solution of the inner problem subject to constraints, also depend on the observable sensitivities $\frac{d\tilde{y}}{d\psi}$, and thus on the state sensitivities. To make the dependence explicit, define

$$\hat{\eta}(\psi) = G(\{\tilde{y}_k\}_{k=1,\dots,n_t}, \psi).$$

Then,

$$\frac{d\hat{\eta}}{d\psi} = \frac{\partial G}{\partial \psi} + \sum_{k=1}^{n_t} \frac{\partial G}{\partial \tilde{y}_k} \left(\frac{\partial \tilde{h}_k}{\partial \psi} + \frac{\partial \tilde{h}_k}{\partial x_k} \frac{dx_k}{d\psi} \right),$$

making the time dependence $\tilde{h}_k = \tilde{h}(x(t_k, \psi), \psi)$, $x_k = x(t_k, \psi)$ explicit, such that e.g. $\frac{dx_k}{d\psi} = \frac{dx(t_k,\psi)}{d\psi}$ denotes the derivative of states evaluated at time t_k. Thus, from (3.8),

$$\nabla \hat{J}(\psi) = \sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \psi} + \frac{\partial J_i}{\partial \tilde{y}_i} \frac{\partial \tilde{h}_i}{\partial \psi} + \frac{\partial J_i}{\partial \tilde{y}_i} \frac{\partial \tilde{h}_i}{\partial x_i} \frac{dx_i}{d\psi} + \frac{\partial J_i}{\partial \eta} \frac{\partial G}{\partial \psi} + \sum_{k=1}^{n_t} \frac{\partial J_i}{\partial \eta} \frac{\partial G}{\partial \tilde{y}_k} \left(\frac{\partial \tilde{h}_k}{\partial \psi} + \frac{\partial \tilde{h}_k}{\partial x_k} \frac{dx_k}{d\psi} \right) \right]. \tag{3.13}$$

If we use FSA, i.e. explicitly simulate $\frac{dx}{d\psi}$, then we can simply calculate (3.13) without further ado, provided all derivatives involving G are, analytically or numerically, available. However, the application of ASA is prohibited as additional terms involve $\frac{dx}{d\psi}$, which would thus not drop out. Namely, terms involving state sensitivities $\frac{dx_i}{d\psi}$ are, regrouped by time dependence,

$$\sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \tilde{y}_i} + \sum_{k=1}^{n_t} \frac{\partial J_k}{\partial \eta} \frac{\partial G}{\partial \tilde{y}_i} \right] \frac{\partial \tilde{h}_i}{\partial x_i} \frac{dx_i}{d\psi}.$$

We see by comparison with the derivation of the adjoint state (2.21) in Section 2.3.2 that we can define a new backward equation with same law $\dot{p} = -\frac{\partial f}{\partial x}^T p$ but terminal condition

$$p(t_{i+1}, \psi) = \lim_{t \searrow t_{i+1}} p(t, \psi) - \left[\frac{\partial J_{i+1}}{\partial \psi} + \sum_{k=1}^{n_t} \frac{\partial J_k}{\partial \eta} \frac{\partial G}{\partial \tilde{y}_{i+1}} \right] \frac{\partial \tilde{h}_{i+1}}{\partial x_{i+1}},$$

Table 3.1: Datasets used for parameter estimation. The number of inner parameters of certain types is indicated, followed, in parentheses, by the number of parameters that were computed analytically in the hierarchical setting. The noise parameter for Dataset 1 was set to one if the dataset was considered individually.

	Dataset 1 (CCLE)	Dataset 2 (MCLP)
# datapoints	5281	1799
# cell-lines	96	54
# observables	1	48
# scalings	96(96)	0
# offsets	0	102(48)
# noise parameters	1(1)	48(48)

such that the gradient can be written as

$$\nabla \hat{J}(\psi) = \sum_{i=1}^{n_t} \left[\frac{\partial J_i}{\partial \psi} + \frac{\partial J_i}{\partial \tilde{y}_i} \frac{\partial \tilde{h}_i}{\partial \psi} + \frac{\partial J_i}{\partial \eta} \frac{\partial G}{\partial \psi} + \sum_{k=1}^{n_t} \frac{\partial J_i}{\partial \eta} \frac{\partial G}{\partial \tilde{y}_k} \frac{\partial \tilde{h}_k}{\partial \psi} \right] - p^T \frac{dx_0}{d\psi}\bigg|_{t=t_0} - \int_{t_0}^{t_{n_t}} p^T \frac{\partial f}{\partial \psi}\, dt\,.$$

With this modified adjoint problem, we can thus employ ASA to efficiently calculate the gradient of the outer problem for large-scale problems also for an inexactly solved inner problem. While this provides a conceptual solution, it has currently however not been implemented anywhere.

3.2.7 Implementation

Together with Daniel Weindl and Leonard Schmiester, the hierarchical optimization approach using ASA from Sections 3.2.4 and 3.2.5 was implemented as part of the C++ package parPE (`https://github.com/icb-dcm/parpe`), which provides functionality to solve large-scale parameter optimization problems on high performance infrastructure using massive parallelization. For ease of implementation, we re-simulated the forward trajectory during the simulation of the adjoint state, as discussed in Section 3.2.4. Further, together with previous implementations by Carolin Loos, we also provided implementations of hierarchical optimization using FSA and ASA for the MatLab toolbox PESTO, and the Python package pyPESTO. A snapshot of all code and data underlying the study in this chapter is on Zenodo (`https://doi.org/10.5281/zenodo.3254429` and `https://doi.org/10.5281/zenodo.3254441`).

3.3 Application to a large-scale pan-cancer pathway model

We applied the hierarchical optimization approach presented in Section 3.2, with analytically solved inner problem based on affine observable transformations and normal noise model, to a large-scale pan-cancer pathway model developed by Fröhlich et al. [2018]. The model comprises 1396 biochemical species, mostly proteins and protein complexes, and 4232 unknown parameters, and can be individualized to specific cancer cell-lines using genetic profiles and gene expression data. As the evaluation was performed mainly by Leonard Schmiester and Daniel Weindl, we here only report main findings, and refer to the publication [Schmiester et al., 2019] for details.

For model individualization and calibration, two datasets were considered (Table 3.1). Dataset 1, the dataset considered in Fröhlich et al. [2018], consists of viability measurements for 96 cancer cell-lines in response to 7 drugs at 8 concentrations, available in the Cancer Cell Line Encyclopedia (CCLE) [Barretina et al., 2012] as relative measurements. To account for this, Fröhlich et al. [2018] simulated both condition and control and divided the simulations [Degasperi et al., 2017]. However, this approach is not applicable when multiple observables need to be considered, e.g. additional data types incorporated, or more complex normalization applied, and gives correlated noise models (see the introduction). Therefore, we transformed the model outputs by cell-line specific scaling factors $s_{\text{cell-line}_{i_s}}$, giving the observation function

$$y_{\text{viability}_i} = s_{\text{cell-line}_{i_s}} \cdot \tilde{y}_{\text{viability}_i},$$

with i iterating over data points belonging to cell-line i_s, and $\tilde{y}_{\text{viability}_i}$ a weighted sum of species assumed to determine cell viability. Normal measurement noise with common variance, $\bar{y}_{\text{viability}_i} \sim \mathcal{N}(y_{\text{viability}_i}, \sigma^2_{\text{viability}})$, was assumed.

The viability dataset was complemented by molecular measurements from the MD Anderson Cell Lines Project (MCLP) [Li et al., 2017], Dataset 2, which contains reverse phase protein array (phospho-)proteomic measurements for various cancer cell-lines. 32 proteins and 16 phosphoproteins were identified that were also covered by the model, and in total 54 cell-lines overlapped with the 95 cell-lines from Dataset 1. In the MCLP database, measurements are normalized across cell-lines and proteins by subtracting the respective median from the $\log_2$-transformed measured values. To account for this, we included one cell-line specific offset $b_{\text{cell-line}_{j_b}}$, and one protein-specific offset $b_{\text{protein}_{i_b}}$, giving the observation function

$$y_{\text{protein}_{i_b},\text{cell-line}_{j_b}} = \log_2(z_{\text{protein}_{i_b},\text{cell-line}_{j_b}}) + b_{\text{cell-line}_{j_b}} + b_{\text{protein}_{i_b}},$$

with $z_{\text{protein}_{i_b},\text{cell-line}_{j_b}}$ the simulated absolute protein levels. Normal measurement noise with protein-specific variance, $\bar{y}_{\text{protein}_{i_b},\text{cell-line}_{j_b}} \sim \mathcal{N}(y_{\text{protein}_{i_b},\text{cell-line}_{j_b}}, \sigma^2_{\text{protein}_{i_b}})$, was assumed.

The integration of viability and molecular measurements considers information on two levels, which can improve model reliability, however requires a substantial number of observation and noise parameters (Table 3.1).

3.3.1 Adjoint sensitivity analysis facilitates gradient calculation on large-scale problems

To compare FSA and ASA on this large-scale problem, we calculated gradients using both approaches for a subset of parameters ranging from one to all dynamic parameters. As computation times for FSA on the full problem were prohibitively large, we used a random subset of simulation conditions. This showed that the computation time for FSA scaled linearly with the number of parameters, while it was constant in the number of parameters for ASA, leading to a roughly 2700-fold speed-up when considering all parameters (supplementary information of Schmiester et al. [2019], Figure S1A). From this and based on the number of gradient evaluations performed using the adjoint hierarchical approach, we estimated the CPU time for a single local optimization using the forward hierarchical approach [Loos et al., 2018] to be on the order of 100-1000 years, roughly three orders of magnitude more than the adjoint hierarchical approach (supplementary

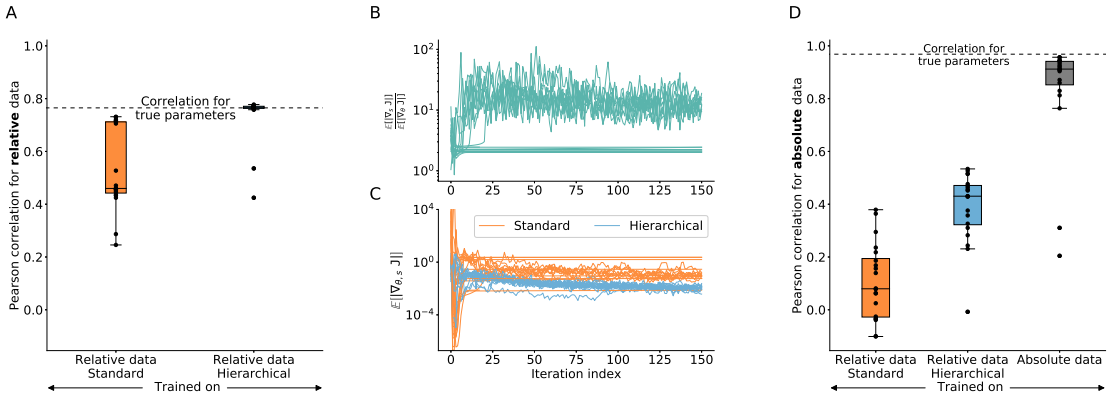

Figure 3.2: Convergence of standard and hierarchical optimization. Parameter estimation results using a simulated version of Dataset 1 from Table 3.1 with the Ipopt optimizer. For each setting, 20 multi-start optimizer runs were performed. A: Pearson correlations of relative training data and corresponding model simulation after training on relative data, using standard and hierarchical optimization. The dashed line indicates the correlation for the true parameters used to generate the noise-corrupted training data. B: Ratio of the average gradient contribution for scaling parameters over dynamic parameters along optimizer run trajectories, using standard optimization. C: Average gradient for standard and hierarchical optimization. Only the parameters that were optimized numerically were taken into account. D: Pearson correlations of absolute data and corresponding model simulation after training on (left and middle) relative data and (right) absolute data. This figure is taken from the author's publication Schmiester et al. [2019].

information of Schmiester et al. [2019], Figure S1B).

3.3.2 Evaluation using simulated data

To analyze the effect of observable parameters on optimizer performance for large-scale problems, we trained the model on noise-corrupted simulated relative data from Dataset 1, using parameters from previous optimization runs. We performed optimization using both the standard approach, and the adjoint hierarchical approach, with analytical conditionally optimal solutions for the scaling factors. For the joint and outer problems, we used multi-start local optimization, with the local gradient-based interior point optimizer Ipopt [Wächter and Biegler, 2006]. Compared to the standard approach, the hierarchical approach achieved substantially better Pearson correlations between data and optimal simulation, for almost all of 20 local starts, in most cases close to correlations between data and simulations using the underlying ground-truth parameters, indicating a good quality of fit (Figure 3.2A). Compared to the simulation of the ODE, the computation time needed to solve the inner problem was roughly five orders of magnitude faster, and thus negligible.

One reason for this substantial performance improvement may be the dimension reduction of the outer problem. However, the inner parameters represented only roughly 2% of all parameters. To assess this, we evaluated the average absolute objective function gradients with respect to scaling ($E[|\nabla_s J|]$) and dynamic ($E[|\nabla_\psi J|]$) parameters (Figure 3.2B+C). Their ratio revealed that the joint objective function was in many cases roughly 10 times more sensitive to scaling parameters than to dynamic ones, throughout the entire optimizer trajectory. Thus, removing the scaling factors via the hierarchical approach improved the conditioning of the optimization

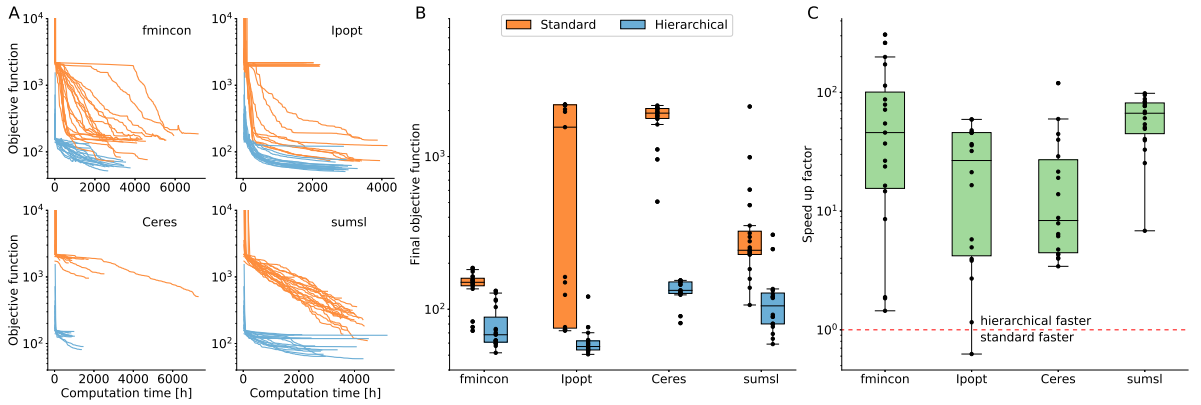

Figure 3.3: Computational efficiency of standard and hierarchical optimization for multiple optimization algorithms. A: Optimizer trajectories for the four local gradient-based optimizers fmincon, Ipopt, Ceres and sumsl using standard and hierarchical optimization. Dataset 1 was used. Fmincon runs were performed on different systems and using a different implementation than the other optimizers, so that absolute computation times are not comparable. B: Final objective function values obtained after 150 iterations by the different optimizers using standard and hierarchical optimization. C: Speed-up of hierarchical compared to standard optimization, defined by the computation time a hierarchical run required to find the final objective function value of the corresponding standard run, for every local optimization (or vice versa if standard optimization found a better final value). The dashed red line indicates equal speed of standard and hierarchical approach. This figure is taken from the author's publication Schmiester et al. [2019].

problem, which has an impact on convergence rate [Boyd and Vandenberghe, 2004, Chapter 9.3].

To investigate the information loss from the use of relative data, we further performed optimization based on simulated absolute data, using the same underlying ground-truth parameters. For comparison, we predicted absolute values from the parameters inferred from relative data. The substantially lower Pearson correlation coefficients far from one indicate that there was indeed a loss of information due to the normalization process (Figure 3.2D). However, also here the hierarchical approach outperformed the standard joint approach, indicating that it was able to extract more information from the data.

3.3.3 All tested optimizers benefit from hierarchical optimization

To provide a thorough comparison of standard and hierarchical optimization, we performed optimization based on the measured viability data, Dataset 1, using both approaches and four different gradient-based local optimizers that employ different update schemes, e.g. based on line searches or trust-region methods. For all four, the hierarchical approach gave consistently better objective function values after the same computation time, with a generally lower variability between starts (Figure 3.3A+B). Further, for most starts using the hierarchical approach, we observed a considerable decrease of objective function value in the first few iterations, compared to a more gradual decrease for the standard approach. To quantify the computation time benefit, we determined the computation time required by the hierarchical approach to achieve the final function value obtained using the standard approach (Figure 3.3C). Except for a single start, the hierarchical approach was consistently more efficient, with median speed-ups between one and two orders of magnitude. Given that a single local optimization run in the standard approach

required up to several thousand hours of computation time, the efficiency improvement achieved using hierarchical optimization, obtaining better results in tens to hundreds of hours, is crucial.

3.3.4 Hierarchical optimization enables integration of heterogeneous data sets

As viability measurements (Dataset 1) provide limited information on molecular mechanisms, we complemented them using (phospho-)protein measurements (Dataset 2). We estimated noise parameters for both datasets to ensure an unbiased weighting. Further, Dataset 2 introduced cell-line and protein specific offsets. When using the hierarchical approach on both datasets, we treated as inner parameters (i) the cell-line specific scaling factors of the viability observables, (ii) the protein specific offsets of the log-transformed protein observables, and (iii) the noise parameters of both datasets, resulting in a total of 193 inner parameters. The cell-line specific offsets were treated as outer parameters, as the analytical solution from Section 3.2.5 requires nested sets of parameters sharing offsets and noise parameters, and further does not accommodate the estimation of multiple offsets.

We performed optimization using the standard and hierarchical approach, using Ipopt for the outer problem. Again, the hierarchical approach yielded substantially better objective function values and was computationally substantially more efficient (Schmiester et al. [2019], Figure 4A). While the standard approach yielded values around $J \approx 10^4$, for the hierarchical one the values can be grouped into two categories. The first group yielded objective function values similar to the standard approach, while the second group yielded values $J < 3 \times 10^3$. The simulations obtained from the standard approach and the first group were able to fit the viability measurements, but failed to describe the protein measurements (Schmiester et al. [2019], Figure 4B). In contrast, the second hierarchical group fitted both viability and protein measurements well. Consequently, only the hierarchical approach was able to integrate both datasets, finding parameters that provided an overall better description of the data.

3.4 Discussion

While large-scale mechanistic models have the potential to lead to a more holistic and detailed understanding of biochemical processes, their parameterization is challenging, requiring efficient, scalable methods. To calibrate such models, large, heterogeneous datasets are required, which often provide only relative measurements. In this chapter, we combined the concept of hierarchical optimization to efficiently handle observable transformations and noise parameters [Loos et al., 2018, Weber et al., 2011] with adjoint sensitivity analysis to efficiently calculate objective function gradients [Fröhlich et al., 2017]. Further, we derived analytical formulas for optimal scaling factors, offsets and noise parameters under the assumption of additive normal noise, extending the work of Loos et al. [2018] to affine transformations, allowing for a highly efficient solution of the hierarchical inner problem. In addition, we presented a more generic view on hierarchical optimization, independent of noise and observation model, generalizing previous considerations, and permitting the use of forward and adjoint sensitivity analysis also for an inner problem subjected to constraints.

We evaluated the adjoint hierarchical approach on a large-scale pan-cancer pathway model by Fröhlich et al. [2018], on which adjoint sensitivity analysis enabled inference in the first place. The comparison of hierarchical to standard optimization revealed substantially improved optimizer convergence using the hierarchical approach in all considered scenarios. We obtained speed-ups of one to two orders of magnitude, irrespective of the used local optimizer. Given that the overall CPU time is on the order of thousands of hours, this is a substantial improvement. On synthetic data, for the hierarchical approach correlations of the data with simulations for inferred parameters were close to correlations with simulations for the underlying ground-truth parameters. Further, only the hierarchical approach was able to integrate two heterogeneous datasets of viability and protein measurements, by effectively weighting them appropriately via conditionally optimal noise levels. The inferred noise parameters provide estimates for the scale of measurement noise when no or few experimental replicates are available, which is the case in many large-scale databases (e.g. CCLE and MCLP). While previous studies had already demonstrated a detrimental impact of observable rescalings for relative data on optimizer convergence [Degasperi et al., 2017], we identified comparably large gradients with respect to scalings as a possible explanation, resulting in numerically stiff gradients for the joint problem. Our hierarchical approach fully circumvents this problem, allowing for the introduction of observable transformations and noise parameters at virtually no computational overhead, and without increasing the dimension of the outer optimization problem.

Here, we only considered additive normal measurement noise, which is not applicable to all problems. One alternative is Laplace noise, the assumption of which as noise model is more robust in the presence of outliers in the data (Maier et al. [2017], see also Chapter 5 in the context of approximate Bayesian computation). Restricted to scaling factors, noise parameters, and forward sensitivity analysis, Laplace noise was considered already in Loos et al. [2018]. Further, in many applications log-transformed observables are considered (Raue et al. [2013b], see also the supplementary information of Loos et al. [2018]). Deriving analytical solutions there for an inner problem consisting of affine observable transformations and noise parameters would allow for highly efficient hierarchical optimization for a greater variety of applications. Further, already the here considered application problem defined multiple offsets per observable. Extending the inner problem to all offsets may improve performance of the hierarchical approach further here.

It may not always be possible or straightforward to derive an analytical solution of the inner problem. In that case, the inner problem can be solved numerically, which further has the advantage of a straightforward implementation. As discussed in Section 3.2.1, in that case an investigation of the inner problem structure, e.g. regarding convexity, is beneficial, to guide e.g. the number of local starts to perform. In tests, we observed for small-scale models the computation time spent in the numerical solution of an inner problem to be not unsubstantial, which may however be different for large-scale problems, and more efficient implementations. A systematic benchmark of such approaches would be required. Moreover, as efficient and scalable numerical ODE solvers as employed in AMICI are implemented in C++, a re-implementation of the inner problem, solved both analytically and numerically, in C++, and an integrated formulation avoiding the repeated simulation of the forward problem as employed for the application example in this chapter, could improve efficiency further.

When numerical difficulties occur due to extreme inner parameters, or e.g. boundaries on the range of values they can take, imposing constraints on the inner problem may be required. While

we have, in Section 3.2.6, given an outlook on how such a problem could be efficiently solved using both forward and adjoint sensitivity analysis, the implementation of such a routine would be more intricate. An evaluation of this approach, and a comparison to e.g. just considering the partial derivatives as in Section 3.2.4, may be of interest.

Recently, optimization using mini-batch stochastic gradient descent, which uses in each generation only a subset of the full dataset to update the parameters, has been demonstrated to improve performance for inference on mechanistic ODE models with many experimental conditions [Stapor et al., 2019]. A combination of mini-batch stochastic and hierarchical optimization would require a more careful batch selection or consideration of conditionally optimal inner parameters, but could provide further improvements for large datasets.

Similar hierarchical approaches have more recently also been used e.g. for marginalized sampling (not published yet), and for categorical data via an optimal quantitative surrogate data based approach [Schmiester et al., 2020, 2021b]. This shows that the general concept thus has promising further application fields. The possibility of a combination with adjoint sensitivity analysis for scalable gradient evaluation in those cases however remains to be investigated.

In conclusion, we developed a scalable hierarchical framework to efficiently estimate observable transformations and noise parameters, and using adjoint sensitivity analysis to efficiently calculate objective function gradients. We expect this approach, as well as the general concept, to improve inference in many application problems.

Chapter 4

Exact efficient inference in ABC with noisy measurements

While increasingly popular for likelihood-free inference, the approximation error introduced in approximate Bayesian computation (ABC) is often not apparent. In this chapter, we demonstrate exemplarily how ABC yields erroneous parameter estimates when measurement noise present in the data is neglected, which often happens in practice. Then, we discuss ways of addressing measurement noise in the analysis. In particular, we present a novel exact, efficient, adaptive, sequential importance sampling based algorithm, applicable to general model types and noise models. We demonstrate how, compared to established approaches, we obtain orders of magnitude speed-up, thus facilitating exact inference via ABC for a wide range of application problems for the first time.

This chapter is based on and partly identical to the following publication, in which all methods were developed and all analyses performed by the thesis author:

- **Schälte, Y.** and Hasenauer, J. (2020). Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. Bioinformatics, 36 (Supplement 1), pp. i551-i559.

4.1 Introduction

Measurement data are usually corrupted by noise due to unavoidable inaccuracies in the measurement process. In likelihood-based inference, it has been widely adopted to include a model of measurement noise in the likelihood function [Raue et al., 2013b]. Especially for deterministic process models, this allows to define a non-degenerate likelihood in the first place. In contrast, in likelihood-free inference, in particular ABC, it is easy to ignore noise altogether, due to the non-necessity of even formulating a likelihood that relates process states to measured values. Further, the various inherent approximation levels of ABC often do not allow to pinpoint error sources from the result (see the discussion in Section 2.4.3). In the past, noise has often been disregarded in ABC analyses [Eriksson et al., 2019, Imle et al., 2019, Jagiella et al., 2017, Lenive et al., 2016,

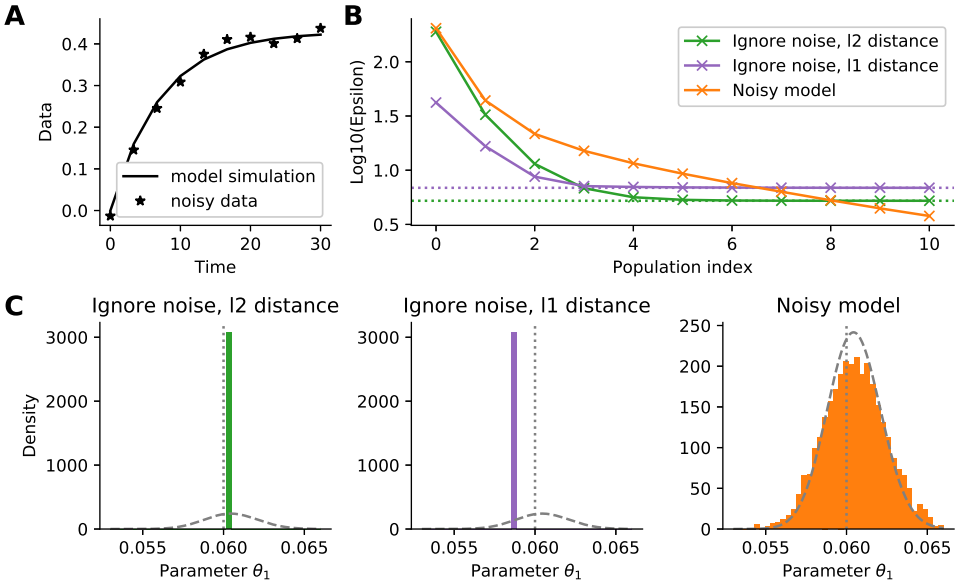

Figure 4.1: Illustrative conversion reaction ODE example. (A) The employed data. (B) Acceptance thresholds over sequential ABC-SMC iterations for 3 estimation methods: Using a non-noisy model and an L2 or L1 distance in the ABC acceptance step, and using a noisy model (and an L2 distance). The minimum obtainable values for the respective distances are indicated by dashed lines. (C) Histograms of the corresponding ABC posterior approximations after the last iterations. The true posteriors are indicated by dashed, the true parameter values by dotted lines. This figure is taken from the author's publication Schälte and Hasenauer [2020].

Toni et al., 2009]. Asymptotic exactness, i.e. recovery of the true posterior distribution, of ABC is however guaranteed only if the data-generation process is perfectly reproduced. Omitting the measurement noise can lead to substantially wrong parameter estimates.

The problem is illustrated in Figure 4.1 on an ODE model of a conversion reaction, with one unknown parameter θ_1. Synthetic data $\bar{y}_{\text{obs}}$ were generated by adding normal noise to the model simulation $y(\theta_{\text{true}})$ (Figure 4.1A). Three different ABC-SMC analyses were performed: Using the noise-free model y together with an L1 (I) or L2 (II) distance, and adding a corresponding normal random variable to the model output to account for noise (III). Usually, in ABC we would hope to decrease the acceptance threshold ε asymptotically to 0 for $t \to \infty$. For (I) and (II), this was however not possible (Figure 4.1B): The thresholds converged to different positive values. This is reflected also in the inferred posterior distributions (Figure 4.1C), where (I) and (II) converged to different point estimates. The parameters corresponding to these point estimates are maximum likelihood estimators under the assumption of normal (I) or Laplace (II) noise (see the supplementary information of Schälte and Hasenauer [2020, Section 1.5] for details), which is far from the result one would aspire in a Bayesian analysis. In contrast, (III) gives a good approximation of the true posterior. In the supplementary information of Schälte and Hasenauer [2020], we illustrate consequences of neglecting measurement noise on further models, including stochastic ones.

In practice, errors in ABC parameter estimates due to model error and in particular the omission of measurement noise can be hard to detect, as usually the model is not deterministic, and the inference computationally so expensive that convergence cannot be reached, such that e.g. in the above analysis not a point estimate, but a highly concentrated distribution indistinguishable

from a posterior for highly informative data would be the result, making it important to correctly account for measurement noise.

In this chapter, we discuss ways of addressing measurement noise: Either the model output can be randomized, or the ABC acceptance step modified, in accordance with the noise model. The latter method builds on the insight by Wilkinson [2013] that ABC gives exact inference, however from the wrong model, namely an additional implicit uniform noise model induced by the acceptance step. Introduced by Wilkinson [2013] for ABC-Rejection and ABC-MCMC and used in van der Vaart et al. [2018] in the case of replicate measurements, the approach was extended by Daly et al. [2017] to ABC-SMC, presenting two algorithms to perform exact inference given additive normal measurement noise, and relying on certain tuning parameters. Here, we extend the existing ideas by presenting an exact algorithm based on the ABC-SMC formulation of Toni et al. [2009], applicable to various model types and noise models. We present robust approaches to automatically select certain hyperparameters, like initialization and step size selection, and include ideas from rejection control importance sampling [Liu et al., 1998, Sisson and Fan, 2018] to increase sampling efficiency. We evaluate and compare the presented methods on various models including ordinary and stochastic differential equations, discrete Markov jump processes, and agent-based models, and noise models including normal, Laplace, and Poisson noise.

4.2 Background: Model error in ABC

In this chapter, we employ the notation of general parameter inference problem from Section 2.1, and of ABC-SMC methods from Section 2.4. In particular, let θ denote parameters, $\bar{y}_{\text{obs}}$ the observed data, d the ABC distance metric, ε the acceptance threshold, and N the number of samples constituting a population.

In ABC, instead of the full likelihood $\pi(\bar{y}|\theta)$ (2.4), often the model likelihood $\pi(y|\theta)$ (2.2) is considered in the formulation of the ABC algorithms in Section 2.4, i.e. the simulation and acceptance steps from Section 2.4.1 are defined as

2.' simulate data $y \sim \pi(y|\theta)$,

3.' accept (θ, y) if $d(y, \bar{y}_{\text{obs}}) \leq \varepsilon$.

This effectively assumes perfect measurements, which is practically rather uncommon. In the following, we assume that the observed data $\bar{y}_{\text{obs}}$ are noise-corrupted and thus a realization of a distribution $\pi(\bar{y}|\theta)$, resulting from the noise-free model likelihood $\pi(y|\theta)$ via a parameterized noise model $\pi(\bar{y}|y, \theta)$ (2.3), so that we can write $\pi(\bar{y}|\theta) = \int \pi(\bar{y}|y, \theta)\pi(y|\theta)\, dy$. This is in line with the notation of Section 2.1. We assume that we can simulate data $y \sim \pi(y|\theta)$ from the model likelihood, while evaluation of the model likelihood is not possible. The noise model $\pi(\bar{y}|y, \theta)$ is usually simple, e.g. a normal or a Laplace distribution, thus we may assume that we can evaluate it. More generally, we thus consider an intractable likelihood that we can decompose into a first, intractable, component, the model likelihood, and a terminal, tractable, component, the noise model.

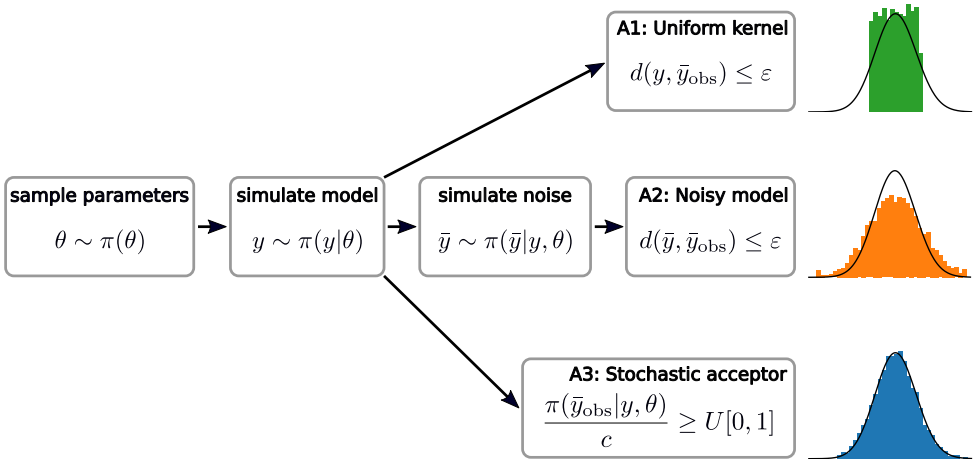

Figure 4.2: The different conceivable ways of accounting for noise in ABC. (A1): Using a uniform kernel with $\varepsilon \gg 0$. The posterior is visibly off. (A2): Adding random measurement noise to the model output. The posterior still has a slightly higher variance. (A3): Modifying the acceptance kernel. The posterior matches the true distribution accurately. This figure is adapted from the author's publication Schälte and Hasenauer [2020].

4.2.1 Approaches to account for noise

ABC inference with the noise-free model $\pi(y|\theta)$ but noise-corrupted data $\bar{y}_{\text{obs}}$ implies that inference for the wrong model is performed. To tackle this discrepancy, there are in principle three different approaches, visualized in Figure 4.2.

Using an appropriate uniform kernel

The first approach (A1) is to simply use a uniform acceptance kernel based on an appropriately defined distance metric d, and an acceptance threshold ε chosen comparably large, such that the resulting acceptance kernel is similar to the underlying noise distribution, e.g. regarding variance (e.g. in Daly et al. [2017], Toni et al. [2009]). An advantage of this approach is its computational efficiency, since acceptance is deterministic. Further, it is straightforward to apply, as it only uses standard methods available in most ABC implementations. However, a major problem is that this approach effectively assumes a uniform noise distribution (see Theorem 4.1), which is in practice rarely appropriate. In addition, the exact choice of ε is ambiguous. Here, we only mention this approach for completeness, and focus on (asymptotically) exact methods.

Randomizing the model output

The second approach (A2) is to account for measurement noise in the forward model, by explicitly simulating it on top of the model simulation, $y \to \bar{y} \sim \pi(\bar{y}|y,\theta)$, and thus using $\pi(\bar{y}|\theta)$ as generative model in second step above (e.g. in Toni and Stumpf [2010]). An advantage of this approach is that its application is again straightforward, only requiring a basic ABC implementation. In addition, if the noise model depends on unknown parameters, these can in principle be included in the overall parameter vector θ and estimated alongside. Further, this

approach is in particular applicable to fully black-box models, where noise cannot be separated. A major problem is however that the randomness in the noise simulation can make the fit to observed data challenging. Further, the comparison of simulated and observed data still requires a uniform acceptance kernel with threshold $\varepsilon > 0$. A2 is asymptotically exact for $\varepsilon \to 0$ (see e.g. Theorem 5.1), however in practice a small approximation error remains and can be hard to quantify.

Modifying the acceptance step

The third approach (A3) simulates noise-free data $y \sim \pi(y|\theta)$, but modifies the acceptance step: Based on the insight by Wilkinson [2013] that "ABC gives exact inference, but for the wrong model", we modify the acceptance step to

3." accept (θ, y) with probability $\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c}$.

Here, $c \geq \max_{y,\theta} \pi(\bar{y}_{\text{obs}}|y,\theta)$ is a normalization constant such that the quotient is in the interval $[0,1]$. This step can be implemented by sampling $u \sim U[0,1]$ and accepting if $\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c} \geq u$.

The following theorem shows that this indeed gives exact inference from the true posterior:

Theorem 4.1 (Exact noisy ABC)**.** *Consider prior $\pi(\theta)$, model likelihood $\pi(y|\theta)$, and noise model $\pi(\bar{y}|y,\theta)$, and assume $\bar{y}_{obs} \sim \pi(\bar{y}|\theta) = \int \pi(\bar{y}|y,\theta)\pi(y|\theta)\, dy$. Then the above ABC algorithm, i.e. sampling $\theta \sim \pi(\theta)$, simulating $y \sim \pi(y|\theta)$, and accepting with probability $\pi(\bar{y}_{obs}|y,\theta)/c$ with $c \geq \sup_{\theta,y} \pi(\bar{y}_{obs}|y,\theta)$ targets the correct posterior distribution $\pi(\theta|\bar{y}_{obs}) \propto \pi(\bar{y}_{obs}|\theta)\pi(\theta)$.*

Proof. Adapted from Wilkinson [2013]. Denote by A the event that parameter θ is accepted. Then for arbitrary $c > 0$ the distribution of accepted parameters is given by

$$\pi(\theta|A) = \frac{\pi(A|\theta) \cdot \pi(\theta)}{\pi(A)} \propto \int \min\left[\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c}, 1\right] \pi(y|\theta)\, dy \cdot \pi(\theta)$$

by definition of the acceptance criterion.

Moreover, since we assume the data to be generated under $\bar{y}_{\text{obs}} \sim \pi(\bar{y}|\theta)$, the true posterior distribution is given by

$$\pi(\theta|\bar{y}_{\text{obs}}) = \frac{\pi(\bar{y}_{\text{obs}}|\theta)\pi(\theta)}{\pi(\bar{y}_{\text{obs}})} \propto \int \pi(\bar{y}_{\text{obs}}|y,\theta)\pi(y|\theta)\, dy \cdot \pi(\theta).$$

Thus, if $c \geq \sup_{\theta,y} \pi(\bar{y}_{\text{obs}}|y,\theta)$, we find $\pi(\theta|A) = \pi(\bar{y}_{\text{obs}}|\theta)$, i.e. accepted samples constitute a sample from the true posterior distribution. □

This is a reformulation of the main result in Wilkinson [2013], extending it to arbitrary noise models, and allowing the noise model to be parameter-dependent, which is crucial in practice, as noise model parameters such as variances are often unknown [Raue et al., 2013b]. Including these as parameters allows to estimate them alongside the dynamical model parameters. An

intuitive explanation for why e.g. variance parameters are not estimated arbitrarily large in the above formulation is that, while allowing to accept simulations that do not tightly match the observed data, large variances are implicitly punished by a flatter density, leading to smaller acceptance rates.

It is thus the non-degenerate noise model that allows us to perform likelihood-free inference in an exact manner (i.e., up to Monte Carlo error), although we still cannot evaluate the full likelihood in general. Only for deterministic models do we thus evaluate the full likelihood, such that this approach is equivalent to likelihood-based sampling techniques as described in Section 2.2.3. Wilkinson [2013] integrated this idea in an ABC-rejection and an ABC-MCMC algorithm, and Daly et al. [2017] introduced two sequential implementations for Gaussian noise. Building on both works, we present here a self-tuned sequential, efficient algorithm applicable to a broad spectrum of noise models.

4.3 Towards an efficient exact sampler

To increase sampling efficiency, we want to integrate the exact sampler A3 in an SMC scheme, for which we need to replace the gradual decrease of the acceptance threshold ε. Motivated by MCMC parallel tempering [Earl and Deem, 2005], the basic idea of our approach is to temper the acceptance kernel to mediate from prior to posterior. In this context, see also the discussion on variations of likelihood-based SMC samplers in Section 2.2.3. We introduce temperatures $T_1 >, \ldots, T_{n_t} = 1$, and, for a given temperature $T = T_t$, modify the above third acceptance step to

3.” accept (θ, y) with probability $\left(\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c}\right)^{1/T}$.

This yields samples from the distribution

$$\pi_{\text{ABC},t}(\theta|\bar{y}_{\text{obs}}) \propto \int \pi(\bar{y}_{\text{obs}}|y,\theta)^{1/T}\pi(y|\theta)\pi(\theta)\,dy, \tag{4.1}$$

such that $T_{n_t} = 1$ gives samples from the target posterior distribution. As usual and described in Section 2.4.4, sequential importance sampling is performed, with a proposal distribution $g_t(\theta)$ at generation t, and the accepted particles subsequently weighted by $w_t(\theta) = \pi(\theta)/g_t(\theta)$. This tempering scheme is applicable to any noise model, with the exception of uniform noise, for which the acceptance step would remain unchanged in t, but this noise model can be effectively dealt with already by standard ABC.

In the following, we propose approaches to select the normalization constant c, the temperature schedule, and the initial temperature.

4.3.1 Normalization choice and correction

A problem persistent in the approaches by Wilkinson [2013] and Daly et al. [2017] is the choice of normalization constant c. If c is too small such that higher values occur during sampling,

the effective noise distribution sampled from is not $\pi(\bar{y}_{\text{obs}}|y,\theta)$, but $\min\left[\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c},1\right]$, i.e. it is flattened out to a uniform distribution at values of high probability. In particular as $c \to 0$, the effective noise distribution becomes one similar to a uniform distribution of increasing variance. For a more illustrative discussion of this see Daly et al. [2017].

As discussed in Section 2.4.5 in the context of standard ABC, the rejection step could be replaced by non-normalized importance weights $\pi(\bar{y}_{\text{obs}}|y,\theta)w(\theta)$, not requiring a choice of c altogether, which would however result in highly variable weights and consequently a low ESS (2.16), thus we want to employ a rejection mechanism here.

A trivial choice for c is the highest mode of the noise distribution, which is for common noise models assumed at $y = \bar{y}_{\text{obs}}$. In practice, it is however often unlikely or impossible for the noise-free model to exactly replicate the noise-corrupted measured data, such that unnecessarily low acceptance rates are the consequence. Of interest is thus the point

$$\bar{c} = \max_{y,\theta \text{ realizable}} \pi(\bar{y}_{\text{obs}}|y,\theta) \tag{4.2}$$

that is realizable under the model. For deterministic models, this point is the maximum likelihood value and can be computed e.g. by optimization. For stochastic models it is generally unknown. Due to the problem of finding good values for c, Daly et al. [2017] disregard the ABC-SMC sampler based on Toni and Stumpf [2010] that we also employ here, in favor of a sampler based on Del Moral et al. [2012], although the former had shown superior accuracy. We can however solve the problems of too low acceptance rates or alternatively the decapitation of the noise model by correcting for the discrepancy in acceptance distribution. Based on ideas from rejection control importance sampling (RCIS, Sisson et al. [2018, Chapter 4]), given a proposal distribution $g(\theta)$, a temperature $T \geq 1$, and any $c > 0$, we reformulate the above third acceptance step as

3.” accept with probability $\min\left[\left(\frac{\pi(D|y,\theta)}{c}\right)^{1/T},1\right]$,

and modify the importance weights to

$$w(\theta,y) = \frac{\pi(\bar{y}_{\text{obs}}|y,\theta)^{1/T}}{\min\left[\left(\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c}\right)^{1/T},1\right]} \cdot \frac{\pi(\theta)}{g(\theta)}, \tag{4.3}$$

and formulate the following

Theorem 4.2 (Importance-weighted acceptance)**.** *With the notation of Theorem 4.1, consider a proposal distribution $g(\theta) \gg \pi(\theta)$, a temperature $T \geq 1$, and let $c > 0$ arbitrary. For an ABC algorithm that samples $\theta \sim g(\theta)$, simulates $y \sim \pi(y|\theta)$, and accepts with probability $\min\left[(\pi(\bar{y}_{obs}|y,\theta)/c)^{1/T},1\right]$, the weighted samples $(\theta,y,w(\theta,y))$, with non-normalized weights w as in* (4.3)*, target the distribution* (4.1)*.*

In particular, for $T = 1$, we thus sample from the target posterior distribution.

Proof. As in Section 2.4.5, we can interpret the above ABC routine as generating tuples (θ,y) in joint space over parameters and, here, noise-free data. Similar to the proof of Theorem 4.1,

with acceptance event A, the distribution of accepted particles is

$$\pi(\theta, y|A) \propto \min\left[\left(\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c}\right)^{1/T}, 1\right]\pi(y|\theta)g(\theta),$$

while the joint target distribution is

$$\pi_{\text{ABC}}(\theta, y|\bar{y}_{\text{obs}}) \propto \pi(\bar{y}_{\text{obs}}|y,\theta)^{1/T}\pi(y|\theta)\pi(\theta).$$

Thus, we need to, by Proposition 2.3, apply exactly the importance weights (4.3), where $\pi(y|\theta)$ cancels out as usual. Samples for θ alone are obtained by marginalization. □

This means that we can, for arbitrary c, correct for accepting from the decapitated noise distribution by modifying the acceptance weights accordingly. While a smaller c leads to higher acceptance rates, it also leads to an increase in the Monte Carlo error by weight degeneration and consequently a low ESS, therefore it must be chosen carefully. In a sequential approach, we can iteratively update c by taking into account previously observed values, while the samples gradually uncover higher-density regions. In this chapter, we set it by default to the maximum of the values found in previous generations. When acceptance rates are too low, we set it to $\eta^{-T}c$, where c is the maximum found value, and $\eta \geq 1$ increases acceptance rates by roughly up to that factor. Other schemes, e.g. based on quantiles of observed values, are possible, and would require further studies. Before the first generation, we draw a calibration sample from the prior to select the initial normalization.

4.3.2 Temperature update schemes

In order to increase robustness and automation, it remains to select the temperature sequence $T_1 > \ldots > T_{n_t}$ appropriately, balancing information gain per generation, and acceptance rate. In general, the overall required number of simulations depends on both the number of intermediate populations, and the difficulty of jumping between subsequent distributions. In the following, we propose two schemes.

Acceptance rate scheme

The idea underlying this scheme is to match a specified target acceptance rate, i.e. to choose $T_t = T$ such that the expected acceptance rate

$$\gamma = \int\left(\int \min\left[\left(\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c_t}\right)^{1/T}, 1\right]\pi(y|\theta)\,dy\right)g_t(\theta)\,d\theta$$

matches a specified target rate γ_{target}, where c_t is the previously specified normalization constant for generation t. As we have no samples from $g_t(\theta)$ prior to generation t, we employ an importance

sampling estimate based on the previous proposal distribution g_{t-1} as

$$\gamma = \int u_t(\theta) \left(\int \min \left[\left(\frac{\pi(\bar{y}_{\text{obs}}|y,\theta)}{c_t} \right)^{1/T}, 1 \right] \pi(y|\theta)\, dy \right) g_{t-1}(\theta)\, d\theta$$
$$\approx \frac{1}{M_{t-1}} \sum_{i=1}^{M_{t-1}} u_t(\theta_i^{t-1}) \min \left[\left(\frac{\pi(\bar{y}_{\text{obs}}|y_i^{t-1}, \theta_i^{t-1})}{c_t} \right)^{1/T}, 1 \right], \quad (4.4)$$

with Radon-Nikodym derivatives $u_t(\theta) = g_t(\theta)/g_{t-1}(\theta)$. Here, in the second line we approximate the outer integral via a Monte Carlo sample, using all $M_{t-1} \geq N$ particles $\{(\theta_i^{t-1}, y_i^{t-1})\}_{i \leq M_{t-1}}$ sampled in generation $t-1$, including rejected ones in order to be representative of the sampling process and avoid a bias to accepted particles. The inner integral is approximated by the corresponding single value. The Radon-Nikodym derivatives can be omitted if subsequent proposal distributions are sufficiently close. In practice, it may for computational efficiency further suffice to restrict to a random subset of all generated particles.

Matching $\gamma \approx \gamma_{\text{target}}$ is a one-dimensional bounded optimization problem that can be efficiently solved. We employed a bisectional search on the logarithm of the inverse temperature. Compared to the overall run time of ABC analyses, we found this optimization to be temporally negligible. Although this scheme only provides a rough estimate of the expected acceptance rate, it proved sufficient for our purpose.

Assuming for $t \to \infty$ convergence $c_t \to c_\infty$ and $g_t \to g_\infty$ similar in shape to the true posterior distribution, a temperature T proposed via the above acceptance rate scheme with fixed γ will converge to a value $T_\infty > 1$ in general. Therefore, the above scheme needs to be backed up by an additional scheme, e.g. the following one, that ensures $T \to 1$, at the cost of then lower acceptance rates.

Exponential decay scheme

In standard likelihood-based parallel tempering MCMC, empirically a geometric progression, i.e. a scheme with constant ratios between adjacent temperatures, has shown to yield similar probabilities for swaps [Predescu et al., 2004, Sugita et al., 2000]. Since a similar approach was recently successfully applied in an ABC-SMC setting for normal noise [Daly et al., 2017], we used a geometric progression here as well, by specifying a ratio $\alpha \in (0, 1)$ such that $T_t = \alpha T_{t-1}$.

The temperature actually used in generation t was then set to the minimum of the temperatures proposed by the acceptance rate scheme, which ensures that the new criterion is sufficiently different, and the exponential decay scheme, which ensures an eventual decrease to $T = 1$.

Initial temperature

The initial temperature T_1 should be low enough to avoid simply sampling from the prior without gaining information, but high enough for reasonable acceptance rates. It is a crucial tuning parameter, which e.g. in Daly et al. [2017] was simply set to a constant value. Here, we propose to employ the above acceptance rate scheme with $g_0(\theta) = g_1(\theta) = \pi(\theta)$, using a calibration sample

$\{(\theta_i^0, y_i^0)\}_{i \leq N}$ from the prior, the same also used to define the initial normalization constant c.

We denote by ASSA in the following the here proposed exact ABC-SMC algorithm with an adaptive sequential stochastic acceptor, i.e. with normalization constant c sequentially updated with weight correction as described in Section 4.3.1, and temperature T selected as the minimum of the two schemes described in Section 4.3.2, with target acceptance rate $\gamma_{\text{target}} = 0.3$ and decay factor $\alpha = 0.5$.

4.3.3 Implementation

We implemented all methods presented in this chapter in pyABC. We put emphasis on a modular and easy-to-use, well-documented implementation. To ensure numerical stability, operations such as density calculation and optimization of T were performed in log-space. The code underlying the study in this chapter can be found on GitHub (`https://github.com/yannikschaelte/Study-ABC-Noise`), a snapshot of code and data is on Zenodo (`https://doi.org/10.5281/zenodo.3631120`). For details on the hardware used for the study, see the supplementary information of Schälte and Hasenauer [2020], Section 5.

As proposal distribution, we used a multivariate normal kernel with adaptive covariance matrix based on the previous generation's weighted accepted parameters, with bandwidth according to Silverman's rule of thumb, and used the median of the previous generation's accepted distances as the new acceptance threshold ε when using a uniform acceptance kernel [Klinger and Hasenauer, 2017]. Unless mentioned otherwise, we used a population size of $N = 1000$. In general, the sufficiency of a population size can be checked by comparing results for different population sizes. While this was for computational reasons not explicitly done here, preliminary tests and previous studies for the considered problems suggest that the here used population sizes were sufficient (see e.g. Jagiella et al. [2017]).

4.4 Application to test problems

To verify the exactness of ASSA, and to study its performance and compare it to alternative approaches, we considered six test problems covering various process model types, including ordinary differential equations (ODE), stochastic differential equations (SDE), Markov jump processes (MJP), and agent-based models (ABM), as well as various noise model types, including normal, Laplace, and Poisson noise. The core model features are described in Table 4.1.

4.4.1 Test problems

Models M1 and M2 are ODE models of a conversion reaction $A \underset{\theta_2}{\overset{\theta_1}{\rightleftharpoons}} B$, a common component in biochemical reaction networks, the analytical solution to which is given by

$$\begin{pmatrix} A \\ B \end{pmatrix}(t, \theta) = \frac{1}{\theta_1 + \theta_2} \left[\begin{pmatrix} \theta_2 & \theta_2 \\ \theta_1 & \theta_1 \end{pmatrix} - \begin{pmatrix} -\theta_1 & \theta_2 \\ \theta_1 & -\theta_2 \end{pmatrix} \exp(-(\theta_1 + \theta_2)t) \right] \begin{pmatrix} A_0 \\ B_0 \end{pmatrix}.$$

Table 4.1: Properties of test models used in Chapter 4: Identifier, description, process model type, noise model type, number of parameters n_θ and data n_y.

Id	Description	Type	Noise	n_θ	n_y
M1	Conversion reaction	ODE	Normal	2	10
M2	Conversion reaction	ODE	Laplace	2	10
M3	Hodgkin-Huxley neurons	SDE	Normal	2	100
M4	Gene expression	MJP	Poisson	2	10
M5	STAT5 dimerization	ODE	Normal	11	48
M6	Tumor spheroid growth	ABM	Normal	7	30

We assumed known initial conditions $(A_0, B_0) = (1, 0)$ and only species A to be measured. Both reaction rate constants were estimated, with ground truth values $(\theta_1, \theta_2) = (0.06, 0.08)$. Unless mentioned otherwise, we assumed $n_y = 10$ equidistant measurements in the interval $[0, 30]$. In M1, the measured data were assumed to be corrupted by independent additive normal noise (2.7) with standard deviation $\sigma = 0.02$. While this is a commonly used noise assumption, for outlier-corrupted data a Laplace distribution has shown preferable [Maier et al., 2017]. Thus, in M2, we instead assumed independent additive Laplace noise (2.8) with scale parameter $\sigma = 0.02$. We assumed uniform priors on $[0, 0.4]$ for both parameters.

Model M3 is an SDE model of intrinsic ion channel noise in Hodgkin-Huxley neurons, based on Goldwyn et al. [2011]. We used a Fortran 95 implementation made available on ModelDB (`https://senselab.med.yale.edu/ModelDB`). We assumed only the fraction of open K channels to be measured at $n_y = 100$ equidistant time steps over a time frame covering roughly 9 oscillations, and employed an independent additive normal noise model with $\sigma = 0.05$ accounting for inaccuracies in the measurement process. We estimated two parameters dc and membrane_dim, describing the input current, and the square root of the membrane area. We assumed uniform priors dc $\in [2, 30]$ and membrane_dim $\in [1, 12]$, with ground truth values $(\text{dc}, \text{membrane\_dim}) = (20, 10)$.

Model M4 describes mRNA synthesis and decay, $\emptyset \xrightarrow{p_1} \text{mRNA} \xrightarrow{p_2} \emptyset$. To capture the intrinsic stochasticity of this process at low copy numbers, we used an MJP model based on the chemical master equation (CME), with simulations via Gillespie's direct algorithm [Gillespie, 1977], using an own implementation. We assumed mRNA counts to be measured at $n_y = 10$ equidistant time points over the interval $[0, 90]$. We employed a Poisson noise model, which is frequently used for regression of count data [Coxe et al., 2009]. We estimated both transcription rate constant p_1 and decay rate constant p_2, assuming uniform priors $p_1 \in [0, 30]$, $p_2 \in [0, 0.2]$, with ground truth values $(p_1, p_2) = (10, 0.1)$.

Model M5 is an ODE model developed by [Boehm et al., 2014] describing the homo- and heterodimerization of the transcription factors STAT5A and STAT5B, for which three types of data with 16 measurements each are available. In the original publication, independent additive normal measurement noise was assumed, which we also did here, and optimal parameter point estimates were obtained via optimization. We estimated 11 log-scaled model parameters, including three standard deviations of the noise model, one for each data type.

Model M6 is a multi-scale ABM describing the growth of a tumor spheroid, projected onto a two-dimensional plane, as described in Jagiella et al. [2017]. We used an implementation in C++ (`https://github.com/icb-dcm/tumor2d`). Single cells were modeled as stochastically interacting agents, coupled to the dynamics of extracellular substances modeled via partial differential

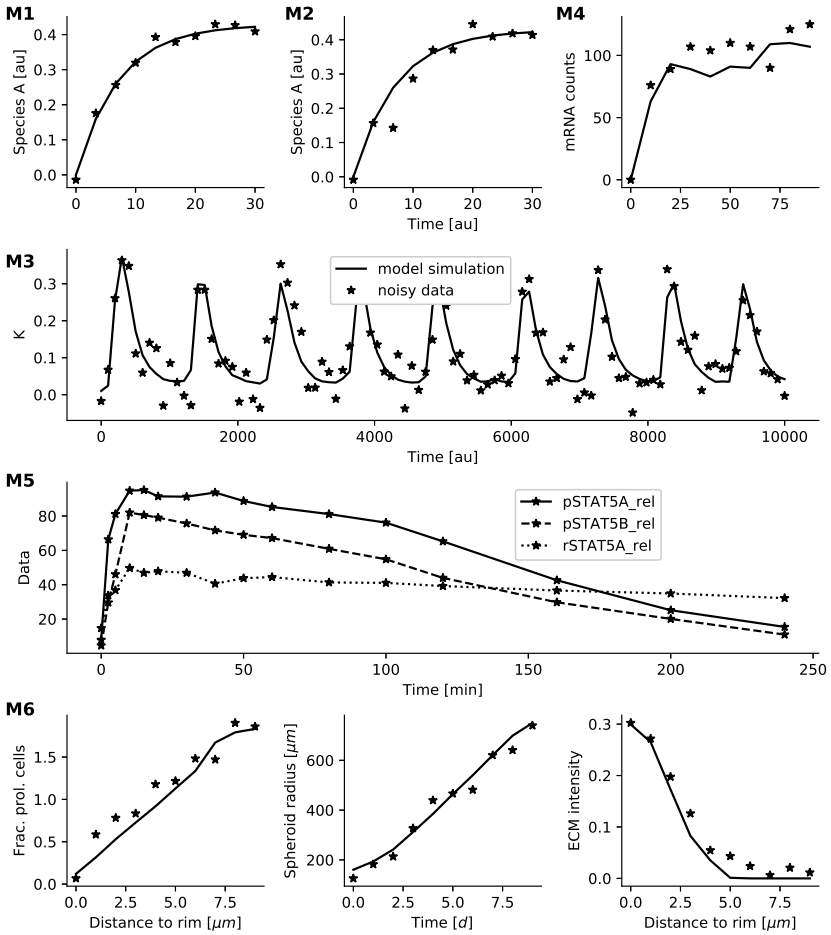

Figure 4.3: Selection of data sets employed for inference on models M1-6. For all but M5, ground truth simulations and noisy data created by adding noise to the simulations are shown. As for model M5 no ground truth exists, only the experimentally observed values of the three data types are shown and connected by lines. This figure is taken from the supplementary information of the author's publication Schälte and Hasenauer [2020].

equations (PDE). The model describes three observables: The spheroid radius over time, and the extra-cellular matrix (ECM) density and the fraction of proliferating cells, at different distances from the rim, observed at a single time point. For the three data types, we assumed independent additive normal noise of different variances. The model has seven unknown parameters, which were estimated on log-scale, with uniform priors. Similar ABMs describing biological processes such as pathogen spread or tissue growth have recently been frequently analyzed using ABC methods (e.g. Durso-Cain et al. [2021], Imle et al. [2019]).

For models M1-4 and M6, we created artificial data by simulating the noise-free model under the ground truth parameters, and adding noise to the simulations according to the respective noise distributions. Model M5 is based on real data without known ground truth but reported literature values. Further details on the models can be found in the supplementary information of Schälte and Hasenauer [2020], exemplary simulated data are shown in Figure 4.3.

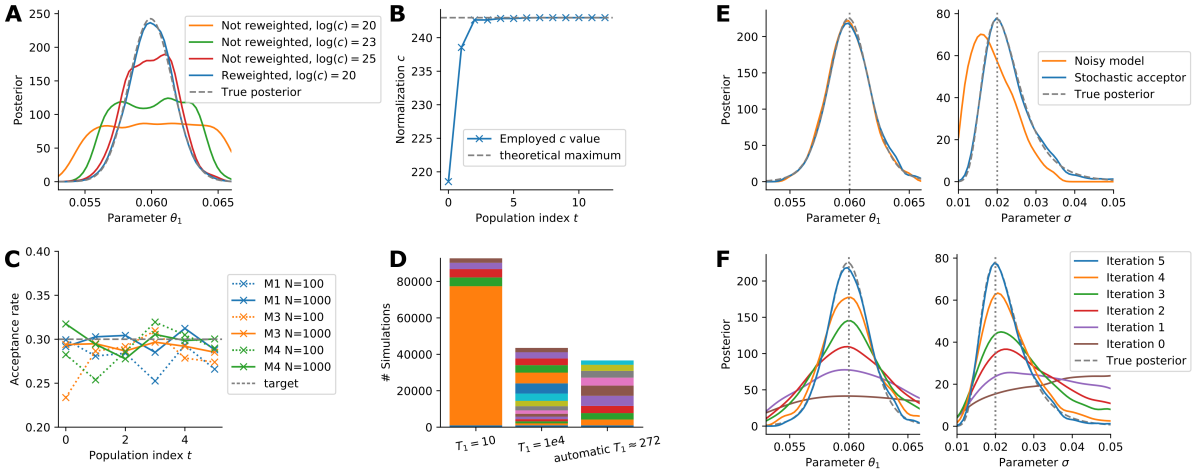

Figure 4.4: Evaluation of the properties of the components of the proposed sampling scheme. (A) Kernel density estimates (KDE) of four posterior estimates generated for model M1 with one unknown parameter and 10 data points, with different normalization constants c, only once correcting for the acceptance bias by re-weighting. (B) Normalization constant c over the iterations for inference for model M1 with 100 data points. (C) Acceptance rate in sampling runs for models M1, M3 and M4 for population sizes $N = 100$ and $N = 1000$, over six generations, using only the acceptance rate criterion was used for temperature scheduling. Note that the population index t is offset by 1 in (B) and (C), starting with 0 instead of 1. (D) Total number of simulations in runs for model M1 for different initial temperatures T_1. The temperature updates were performed using only the exponential decay scheme. The colors indicate individual iterations, starting at the bottom with the simulations spent in the calibration iteration, and then from $t = 1$ upwards. (E,F) KDE of simulations obtained for model M1 with one unknown dynamic parameter, and also estimating the normal noise variance. (E) Comparison of a run using the stochastic acceptor, and a run using a noisy model output. (F) Posterior estimates over the sequential iterations using the stochastic acceptor. This figure is taken from the author's publication Schälte and Hasenauer [2020].

4.4.2 Re-weighting reliably corrects for normalization bias

To verify that the RCIS re-weighting derived in Section 4.3.1 allows for exact Monte Carlo inference regardless of the normalization constant c, we compared, on model M1 with a single estimated parameter, the posterior approximations obtained for different fixed values of c to the true posterior (Figure 4.4A). The theoretical maximum value of the likelihood of the measured data $\bar{y}_{\text{obs}}$ being $\bar{c} = \max_\theta \pi(\bar{y}_{\text{obs}}|y(\theta), \theta) \approx 25.5$, as obtained via multi-start local optimization, we considered values $c \in \{20, 23, 25\}$. As already pointed out by Daly et al. [2017], employing such a too low normalization without weight correction resulted in a flattening out of the posterior distribution at high-density values. The difference to the true posterior became smaller for larger c. However, if we corrected for the bias induced by the too low normalization by re-weighting as in Theorem 4.2, we obtained, even for the lowest $c = 20$, a posterior that matched the true distribution well.

When using the proposed strategy of automatically updating c after each generation to the hitherto highest observed value, we observed for all test models that the value of c converged over time, larger jumps occurring solely in the first few generations. Exemplarily, for model M1 with 2 parameters and 100 data points, c converged to the known theoretical minimum upper bound $\bar{c}$ (4.2) (Figure 4.4B).

4.4.3 Acceptance rate prediction is sufficiently reliable

To analyse the performance of the acceptance rate prediction introduced in Section 4.3.2 as a means to choose adequate temperature steps T, we ran six generations for models M1-3, for population sizes $N = 100$ and $N = 1000$, with a target acceptance rate of $\gamma_{\text{target}} = 0.3$ (Figure 4.4C). Already for $t = 1$, the fit was decent with values fluctuating around the target value, thus permitting to find appropriate initial temperatures. As expected, for a higher population size, the fluctuations decreased. Overall, approximation (4.4) appears to predict acceptance rates sufficiently well for the purpose of temperature selection.

To demonstrate the importance of a proper selection of the initial temperature, we fixed it for model M1 to three different values, $T_1 = 1e1$, $T_1 = 1e4$, and $T_1 \approx 272$, the latter being the temperature proposed by the acceptance rate scheme. For the following generations we here employed only the exponential decay scheme (Figure 4.4C). Too low an initial temperature resulted in many simulations being necessary in the initial generation to bridge from $\pi(\theta)$ to $\int \pi(\bar{y}_{\text{obs}}|y,\theta)^{1/T_1}\pi(y|\theta)\,dy \cdot \pi(\theta)$, sometimes even more than for the entire analysis using the automatic acceptance rate criterion based initial temperature. Conversely, too high an initial temperature yielded little information gain and thus a waste of computation time in the first few generations.

In most cases, the temperature proposed via the acceptance rate scheme for a fixed γ_{target} did not converge to $T = 1$, at which we have exact inference, which was to be expected, as explained in Section 4.3.2, and necessitated the presence of a secondary scheme to ensure $T \to 1$, e.g. the exponential decay scheme thus used subsequently. We observed that the acceptance rate criterion reliably proposed good initial temperatures, and allowed for considerable temperature jumps of up to orders of magnitude in the first few generations, which accelerated convergence substantially, while in later generations the exponential decay scheme took over, with decreasing acceptance rates.

4.4.4 Noise parameters can be estimated

In Theorem 4.1, we allowed the noise model $\pi(\bar{y}|y,\theta)$ to be parameter-dependent. To verify this, we considered model M1, estimating the standard deviation σ of the additive normal noise model alongside rate constant θ_1 (Figure 4.4E). The posterior distributions for both parameters obtained via ASSA indeed match the ground truth well. Conceptually, also employing approach A2, i.e. explicitly simulating the noise as part of the model, and using a standard distance and letting $\varepsilon \to 0$, should allow to estimate noise parameters in the approximate limit. Remarkably however, even for this simple model and after $6e6$ simulations, compared to $1e5$ for ASSA, the quality of the estimated posterior distribution for σ was considerably worse, while θ_1 was well estimated. This confirms Theorem 4.1 and demonstrates that inference of noise parameters is practically possible.

Exemplarily considering the same model, the gradual sequential improvement of the posterior approximation obtained via ASSA over the generations (Figure 4.4F) indicates that the combined temperature update scheme suggests steps with an appropriate, even information gain per generation.

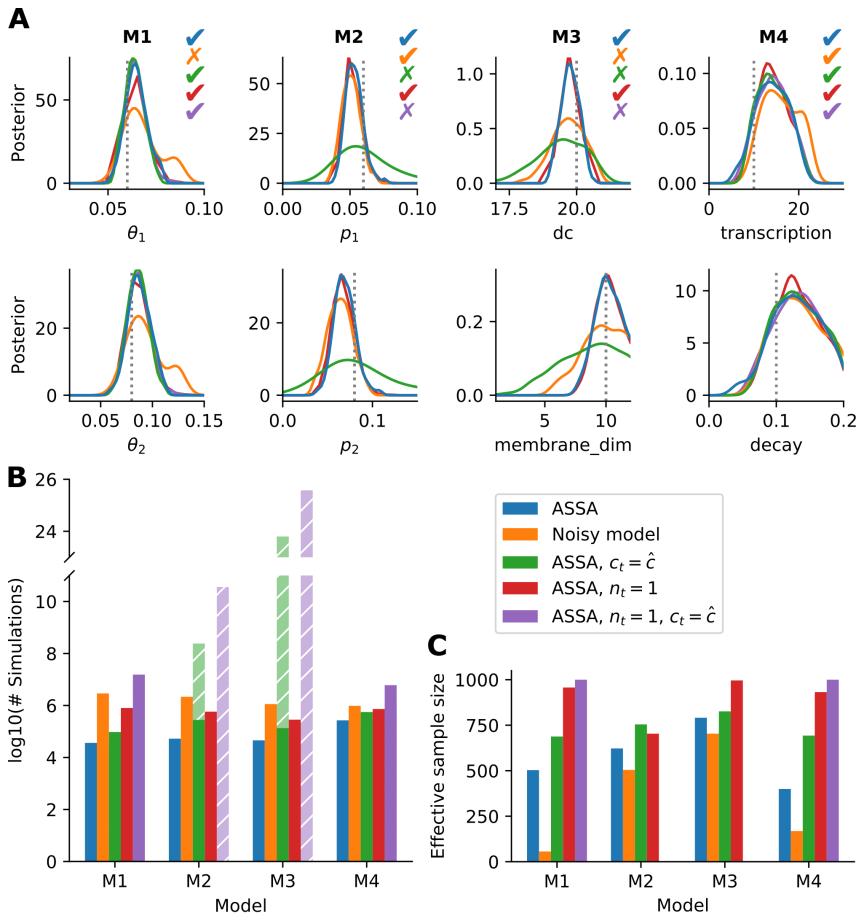

Figure 4.5: Method comparison for models M1-4. (A) Posterior marginals obtained using the five different samplers. For runs that had to be stopped due to an acceptance rate below 1e-3, the last finished population is shown. True parameters are indicated by dotted lines. The colored check marks indicate the methods reaching $T = 1$, or for the noisy model an adequate approximation, by visual inspection. (B) Total number of simulations over all iterations. For samplers that had to be stopped early, in addition to the number of samples until stopping, estimates for an analysis that would reach $T = 1$ are shown hatched, based on the approximation given in the text. (C) Effective sample sizes for the final posterior estimate obtained by each algorithm in its last iteration. This figure is taken from the author's publication Schälte and Hasenauer [2020].

4.4.5 Applicable to various process and noise model types

To evaluate the performance of the proposed method ASSA on different types of process models $\pi(y|\theta)$ and noise models $\pi(\bar{y}|y,\theta)$, we compared it to alternative approaches on models M1-4. Firstly, we employed ASSA with normalization constant not dynamically adapted but set to $c = \hat{c} := \max_{y,\theta} \pi(\bar{y}_{\text{obs}}|y,\theta)$, over the full data space, ignoring whether the model is able to simulate such values, as was done by Wilkinson [2013]. For all here considered noise models, this corresponds to $y = \bar{y}_{\text{obs}}$ (see the supplementary information of Schälte and Hasenauer [2020, Section 4]). Secondly, we employed ASSA with only one iteration, $n_t = 1$, giving exact ABC-Rejection. Thirdly, we considered ASSA with both $c = \hat{c}$ and $n_t = 1$. As last alternative, we considered the noisy model approach A2 in a sequential form, using an L2 distance. Runs were stopped when exact inference with $T = 1$ was reached, which was always the case for ASSA, or the acceptance rate fell below 10^{-3}.

For the deterministic models M1 and M2, we could confirm that the distributions inferred by ASSA closely matched the true posteriors. While such an analysis is not that easily possible for the stochastic models M3 and M4, the posterior approximations obtained using ASSA for all four models are centered around the true parameters to a degree that seems reasonable given model and data (Figure 4.5A). When other samplers of type A3 reached $T = 1$, the posterior approximations were similar in shape. The posterior approximations obtained by the noisy model approach A2 had for three models a larger variance than those obtained by ASSA, indicating that this approach is less efficient.

4.4.6 Orders of magnitude speed-up over established methods

In parameter inference, especially for dynamical models considered in systems biology and in ABC, model simulations are typically the time-critical part (see e.g. Jagiella et al. [2017]). Thus, here we considered the required total number of model simulations as a measure of efficiency (Figure 4.5B). For all models, ASSA required the least simulations. For model M4, the advantage over the other methods was the smallest with a factor of 2, which may be explained by high stochasticity of the process model itself, not unlikely to replicate the noisy data. The other models have a higher signal-to-noise ratio, such that model simulations close to the noisy data were less likely. For models M2 and M3, ASSA with $c = \hat{c}$, and the version with $n_t = 1$, did not even reach $T = 1$ under the permitted computational budget, resulting in a higher posterior variance due to an overestimation of the noise variance, or no results being obtained at all for $n_t = 1$ (see Figure 4.5A).

For our analysis, we provided all methods with a computational budget that was by far sufficient for ASSA. Yet, for some runs with $c = \hat{c}$, this budget was still insufficient. To compare the expected computational budget required for exact sampling, we made use of the roughly inverse proportional dependence of the acceptance rate on the normalization factor c (4.4). With s the number of simulations required by ASSA in the last iteration, we get the estimate $\hat{s} = \frac{\hat{c}}{c}s$ for the required number of simulations for samplers using $c = \hat{c}$ in the last iteration. The estimates indicate that the sequential samplers with $c = \hat{c}$ would, respectively on model M2 and M3, require at least 4 and 20 orders of magnitude more simulations than those with self-tuned c (Figure 4.5B). Even with massive parallelization, exact inference with $T = 1$ would thus not be possible without the proposed self-tuned scheme for c.

On all models M1-4, ASSA required, respectively, about 22, 11, 6, and 2 times less simulations than with $n_t = 1$, indicating that the ABC-SMC scheme is superior to ABC-Rejection, as the temperature selection scheme allows to efficiently bridge from prior to posterior.

To assess the influence of the number of data points, we considered models M1-3 with between 10 and 1000 data points. The value of $\hat{c}$ grew exponentially with the number of data points, while this was not the case for the highest c possible under the model, $\bar{c}$ (4.2), the limiting value for ASSA. Indeed, we found for models M1-3 that for ASSA the number of required simulations increased only moderately. The ratio $\hat{c}/c$ with c the value used by ASSA in the last iteration increased e.g. for model M1 from about 10 for 10 data points to more than 1e200 for 1000 data points. This indicates that ASSA scales well with the number of data points, while approaches with too large normalization c quickly become computationally infeasible. See the supplementary information of Schälte and Hasenauer [2020], Section 8, for further details.

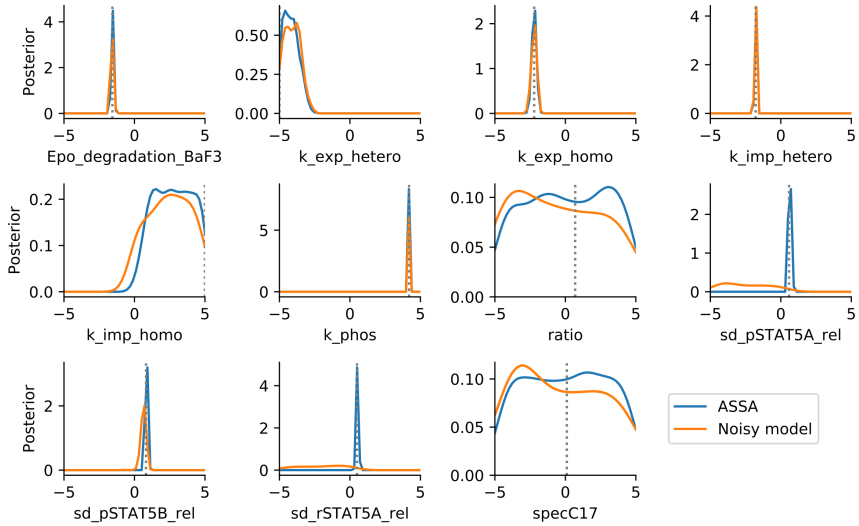

Figure 4.6: Posterior marginals for model M5, using ASSA and the noisy model sampler. The literature MAP values are indicated by dotted lines. The noise parameters have prefix "sd_". This figure is taken from the author's publication Schälte and Hasenauer [2020].

While the number of required model simulations provides information about efficiency, it does not account for stability of the results obtained via the different methods. This we assessed here via the effective sample size (ESS) (2.16). As we inflate the weights by (4.3) if the criterion exceeds 1, it is to be expected that ASSA has a lower ESS than the alternative stochastic acceptors, which was indeed the case on the test models (Figure 4.5C). However, for all models the ESS was still reasonably high, indicating that the populations were not degenerated, and that ASSA is substantially more efficient than other approaches in terms of the ratio of ESS and required number of samples. E.g. on model M3 an increase of the population size by 25% may be expected to yield an ESS of more than 1000 for ASSA. Compared to the orders of magnitude differences between sample numbers, this renders ASSA highly efficient (see also the supplementary information of Schälte and Hasenauer [2020], Figure S13).

4.4.7 Scales to challenging application problems

To assess the performance of the proposed approach on more realistic application examples, we considered models M5 and M6. Here, since the acceptance rates when updating the normalization constant c always to the highest so far observed value were too low, we multiplied c by a factor η of, respectively, 20 and 10, once the acceptance rate fell below 0.1, increasing acceptance rates at the cost of higher weight variability, as introduced in Section 4.3.1. As the other exact approaches would have no chance of success here due to too high sample numbers, as already seen on M1-4, here we only compared ASSA and a sequential version of the noisy model approach A2.

For model M5, we used a population size of $N = 1e4$ to guarantee stability of the results. The posterior marginals (Figure 4.6) indicate that seven parameters can be accurately estimated, two parameters can be constrained, and two parameters are non-identifiable. The posterior distributions recovered the reported literature values [Boehm et al., 2014], which had been obtained by optimization, and the parameter samples provide an accurate description of the experimental data (supplementary information of Schälte and Hasenauer [2020], Figure S14). Importantly,

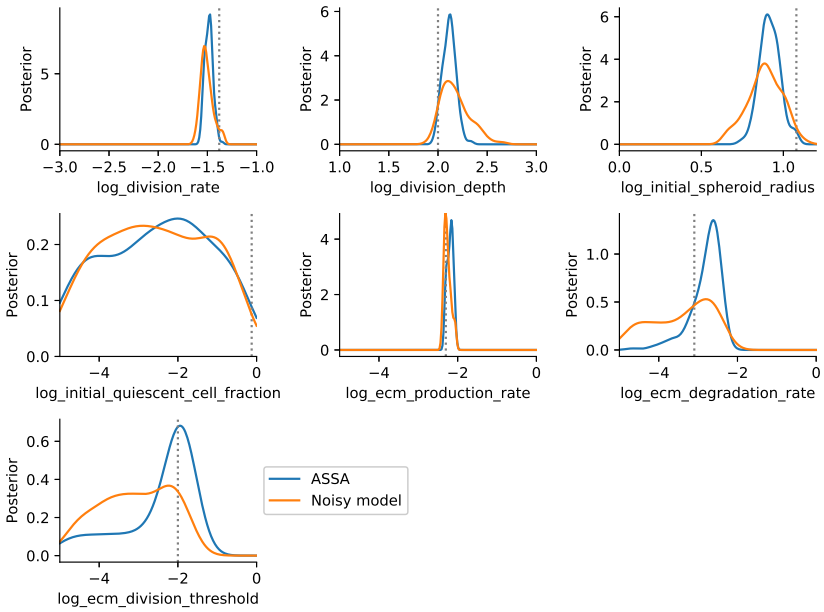

Figure 4.7: Posterior marginals for model M6, using ASSA and the noisy model sampler. Ground truth parameter values are indicated by dotted lines. This figure is taken from the supplementary information of the author's publication Schälte and Hasenauer [2020].

in addition to kinetic parameters, ASSA was capable of identifying the standard deviations of the normal noise models on all three parameters with narrow confidence intervals. In contrast, the application of the noisy model approach A2 was in particular unable to estimate two out of three noise parameters at all. The ESS for the stochastic acceptor was 1011, using in total 8e6 simulations, for the noisy model only 83, using in total 12e6 simulations.

For model M6, we used a population size of only $N = 5e2$ due to the high simulation time of the model. Again, the comparison of the posterior marginals with the noisy model approach A2, which was given a similar computational budget as required by ASSA, revealed that ASSA extracted overall more information (Figure 4.7) expressed by narrower confidence intervals. The parameter samples described the data accurately under the assumed noise model (supplementary information of Schälte and Hasenauer [2020], Figure 16). Given the data, the parameters related to initial conditions cannot be inferred well, but for the others the reference values are accurately matched. The numbers of simulations were, respectively for ASSA and the noisy model approach A2, 4.0e5 and 3.2e5, with an ESS of 222 and 153.

4.5 Discussion

While it allows to perform parameter inference on a wide range of models, we saw in this chapter that ABC can easily give erroneous and misleading parameter estimates when measurement noise is not properly accounted for. We discussed ways of dealing with noise in ABC and in particular presented a novel adaptive, sequential importance sampling based algorithm (ASSA), broadly applicable to various process models and noise models. On several test problems, we demonstrated that the proposed algorithm is accurate and up to orders of magnitude more ef-

ficient than existing approaches. We achieved this efficiency gain by firstly learning a required normalization constant in a self-tuned manner, correcting for potential acceptance bias by importance re-weighting. Secondly, we devised an adaptive tempering scheme to transition from prior to posterior with sensible information gains per generation, in particular permitting by acceptance rate prediction to find good initial temperatures, usually a critical tuning parameter. Consequently, we were able to perform exact likelihood-free inference on models on which this was hitherto not feasible.

Our approach adapts to the problem structure by learning appropriate values for critical tuning parameters on the fly, making it both stable and applicable to diverse problems, and avoiding manual tuning, which could otherwise be time-intensive. This is a key difference to related approaches, in which parameters were manually adjusted to the individual problems. Furthermore, our approach permits estimating noise parameters such as variances, which are in practice often unknown, facilitating integrated analysis workflows [Hross et al., 2018].

As ABC has in the past been frequently applied without a proper noise formulation, the question may be raised whether this may have lead to erroneous parameter estimates, in particular concerning uncertainties. Unfortunately, this is difficult to answer. However, as we have here been able to apply exact likelihood-free inference even to computationally demanding complex inference problems, we expect our algorithm to be broadly applicable, thus improving the reliability of parameter estimates for a broad spectrum of applications.

Still, the presented method could be improved in various regards. One hyperparameter that to some degree required manual selection on more complex application problems was the multiplicative factor η to the normalization constant c, which induces a trade-off between acceptance rate and weight heterogeneity and thus between rejection and importance sampling. It could potentially be automatically subjected to limits on acceptance rates or based on quantiles of observed values. For models with estimated small noise terms and high-dimensional data, acceptance rates of the presented approach can conceptually become prohibitively low, necessitating further robust concepts. Another possible path of future research concerns the tempering steps. The overall required number of simulations depends both on the number of intermediate populations, and the number of samples required in each. The exponential decay scheme used here was rather loosely motivated by analogies to MCMC parallel tempering, and could be improved, similar to dynamically adjusted temperature steps in likelihood-based MCMC [Predescu et al., 2004, Vousden et al., 2016], or likelihood-based SMC concepts [Del Moral et al., 2006, Gelman and Meng, 1998]. Further, for likelihood-based SMC, there exist approaches that try to keep the effective sample size constant [Latz et al., 2018]. Investigating similar concepts in a likelihood-free context could be of interest. Another interesting use of tempering is for thermodynamic integration, allowing to compute Bayes factors. The presented sequential method could potentially be extended to allow to perform model selection, as an alternative to e.g. Toni and Stumpf [2010].

In conclusion, we demonstrated the importance and the benefits of using proper noise models in ABC. The proposed algorithm enables exact inference on a broad range of problems for the first time. Having further provided an implementation in pyABC facilitating massive parallelization, we expect this approach to be useful in many applications.

Chapter 5

Robust adaptive distances in ABC

Complications for inference in general, and approximate Bayesian computation (ABC) in particular, can arise from the presence of outliers in the data. In this chapter, we demonstrate how established ABC distance metrics are highly sensitive to outliers, resulting in erroneous or highly uncertain parameter estimates. Then, we present a self-tuned outlier-robust distance metric, based on a popular adaptive distance weighting concept enabling efficient inference for data varying on different scales, complemented by a simulation-based online outlier-detection and down-weighting scheme. We evaluate and compare the presented methods on various test models covering different model types, problem features, and outlier scenarios. Our evaluation demonstrates substantial improvements on outlier-corrupted data, while giving at least comparable performance on outlier-free data. Note that while some problems related to outliers could be solved by robust noise models via an exact approach as in the previous chapter (discussed below), in this chapter, we employ a black-box model perspective on ABC, as is commonly done in most research, not permitting such a model decomposition.

This chapter is based on and partly identical to the following publication, in which all methods were developed and all analyses performed by the thesis author:

- **Schälte, Y.**, Alamoudi, E. and Hasenauer, J. (2021). Robust adaptive distance functions for approximate Bayesian inference on outlier-corrupted data. bioRxiv.

5.1 Introduction

In ABC, simulated data are evaluated based on their proximity to observed data via summary statistics and distance metrics. As distance metric, often a simple Minkowski distance is used, alternative approaches include e.g. Kullback-Leibler divergence [Jiang, 2018] or Wasserstein distances [Bernton et al., 2017] avoiding the use of summary statistics altogether. A popular distance metric introduced by Prangle [2017] exploits the structure of ABC-SMC algorithms by using a weighted Euclidean distance that adjusts to the problem structure by iteratively updating model output, or summary statistic, weights to normalize contributions at different scales, thus enabling the automatic integration of heterogeneous data.

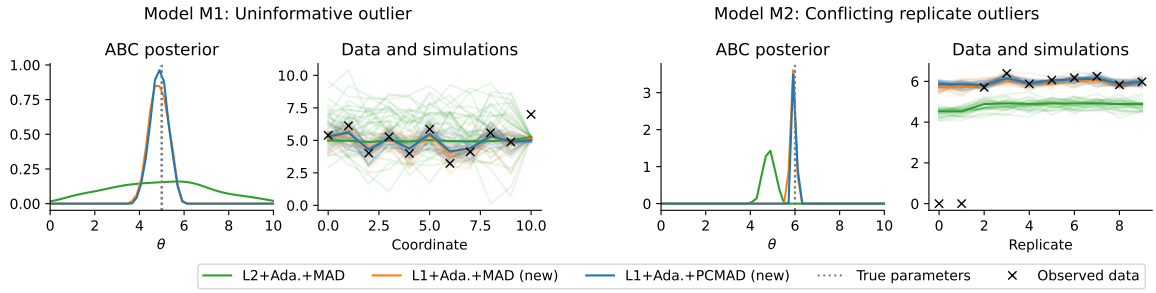

Figure 5.1: Illustration of the impact of outliers on ABC using two test models and three distance metrics, namely the established L2+Ada.+MAD distance introduced in Prangle [2017], and two novel ones, L1+Ada.MAD and L1+Ada.+PCMAD. Left: A model with 10 $\mathcal{N}(\theta, 1)$ distributed data points, and an uninformative $\mathcal{N}(5, 0.1^2)$ distributed one, whose observation however is an outlier. Right: A model with 10 $\mathcal{N}(\theta, 0.2^2)$ distributed data points, the first two of which however are outliers. For both test problems, on the respective left the obtained ABC posterior approximation is shown, and on the right the underlying outlier-corrupted data, together with, for all three distances, light-colored lines of 30 exemplary accepted simulations from the last ABC-SMC generation, and the respective sample means as darker-colored lines. It can be seen how the established distance yields highly uncertain (left) or biased (right) estimates, while the novel methods give far more accurate estimates of the underlying true parameter. The distance metrics and models shown here are properly introduced in Sections 5.3 and 5.4. This figure is taken from the author's publication Schälte et al. [2021a].

Errors can occur in the data generation process of individual measurements, which result in outliers in the data [Ghosh and Vogt, 2012, Motulsky and Christopoulos, 2003]. By an outlier, we informally denote a data point corrupted by errors that are not expected under the given experimental setup. Outliers typically result in large deviations from the expected behavior, which can hardly be explained by the model, already accounting for common measurement noise. Reasons for outliers exist plenty, e.g. technical limitations, external perturbations, or human errors such as missing or incorrect labels [Maier et al., 2017].

In parameter inference, outliers can be problematic, because they can result in erroneous or highly uncertain parameter estimates. Consequently, methods have been developed that aim to detect and remove outliers from the data, prior to and often independent of the used inference method [Ben-Gal, 2005, Hodge and Austin, 2004, Niu et al., 2011]. However, for noise-corrupted high-dimensional or highly structured data with few replicates, which are common in systems biology and also as applications of ABC [Durso-Cain et al., 2021, Jagiella et al., 2017], such methods may be unreliable, and the complete removal of points that are not actually outliers may bias the analysis or increase uncertainty [Motulsky and Christopoulos, 2003]. To circumvent this, estimators that are robust to outliers have been developed, using heavy-tailed distributions [Berger et al., 1994, Fernández and Steel, 1999, Huber et al., 1964, Tarantola, 2005] or pseudo-likelihoods with robust loss functions or divergences [Basu et al., 1998, Chérief-Abdellatif and Alquier, 2020, Jewson et al., 2018]. For ODE models of biochemical systems, the use of heavy-tailed likelihood functions such as Laplace, Huber, or Student's t, rather than the commonly used normal distribution, has proven considerably more robust against outliers, without their explicit removal [Maier et al., 2017].

In ABC, most such methods are however not applicable, as they require a tractable likelihood function. In this chapter, we propose to make ABC robust in the presence of outliers by a novel robust adaptive distance metric, based on the scale-normalizing adaptive distance concept

by Prangle [2017]. It can easily be shown that the original algorithm formulation is sensitive to outliers and can lead to biased or uncertain estimates (see Figure 5.1 for an illustration of possible problems on simple test models). We suggest the use of an outlier-robust norm, and additionally present a routine for simulation-based online outlier detection and down-weighting.

There already exist some approaches aiming at robustness in ABC inference, not necessarily tailored to outliers, but more generally model misspecification. Notably, these include Hellinger or Cramér-von-Mises distances as robust distances between probability distributions [Frazier, 2020], operating on high-dimensional data circumventing the definition of summary statistics, and approaches using distances based on M-estimating functions [Ruli et al., 2020] and γ-divergence estimators [Fujisawa et al., 2021]. These are however designed for data with many i.i.d. replicates, which are not commonly available e.g. in systems biology applications, where data are usually highly structured with single or few replicates, and potentially high-dimensional. Another approach is presented in Frazier et al. [2020], who suggest to address model misspecification by parameterized adjustments of either summary statistics or distance metric weights. This approach is similar to ours in allowing the detection and down-weighting of inconsistent data points, and could in principle be combined with a scale adaptation scheme as done here. However, it augments the parameter vector by the number of summary statistics, thus appearing to require an appropriate, sufficiently low-dimensional, choice thereof. Lastly, a reformulation of the acceptance step as in Chapter 4 would straightforwardly allow for the use of heavy-tailed noise models similarly to Maier et al. [2017]. Yet, as in ABC the model must usually be regarded as a black box, such an approach is not generally applicable and not pursued here.

Our approach makes no assumptions on the structure or dimension of data and is easy to adopt, allows for the interpretation of generated weights, and yields efficient inference by combination of robust measures with scale adaptation. We evaluate and compare the presented methods on several test problems covering various model types, problem features, and outlier scenarios, including a realistic application example of an agent-based model of tumor growth, on both outlier-free and outlier-corrupted data.

5.2 Background: Adaptive distances

In this chapter, we employ the notation of general parameter inference problem from Section 2.1, and of ABC-SMC methods from Section 2.4. For simplicity of notation and as we require, unlike in Chapter 4, no separation of process and noise model, we assume $\bar{y} = y$, i.e. a potential noise model is already incorporated in y. We refer to y as model output, which then includes potentially applied summary statistics. In particular, let θ denote parameters, y_{obs} the observed data, $\pi(\theta|y_{\text{obs}}) \propto \pi(y_{\text{obs}}|\theta)\pi(\theta)$ the posterior distribution of interest, d the ABC distance metric, ε the acceptance threshold, N the number of samples constituting a population, and $P_t = \{(\theta_i^t, y_i^t, w_i^t)\}_{i \leq N}$ the population of accepted particles in generation t.

A common choice of distance metric d is, induced by the respective Lp norm, a weighted Minkowski distance

$$d(y, y_{\text{obs}}) = \|r \cdot (y - y_{\text{obs}})\|_p = \left(\sum_{i_y=1}^{n_y} \left| r_{i_y}(y_{i_y} - y_{\text{obs},i_y}) \right|^p \right)^{1/p}, \tag{5.1}$$

with $p \geq 1$, as a generalization of, most commonly, Euclidean (L2 norm, $p = 2$), Manhattan (L1 norm, $p = 1$), or Chebyshev (maximum or infinity norm, $p \to \infty$) distance, where the summation is over the model output, or summary statistic, coordinates i_y, and the r_{i_y} are model output specific weights. Frequently, simply unit weights $r = 1$ are used (e.g. Borowska et al. [2021], Fearnhead and Prangle [2012], Jiang et al. [2017], Toni and Stumpf [2010], Warne et al. [2018]). It has been argued that, empirically, among similar distance metrics, the exact form does not matter [McKinley et al., 2009, Owen et al., 2015]. Also theoretically, for $\varepsilon \to 0$ the approximation converges to the true posterior independent of the exact distance used (see Theorem 5.1), provided the data generation model is specified correctly (see Chapter 4 for a discussion of the impact of distance metric in the case of model misspecification).

As demonstrated in particular in Prangle [2017], adjusting the distance metric to the problem structure can however improve parameter estimates under a limited computational budget. The precise problem Prangle [2017] tackle is that data or summary statistics can be and vary on different scales. Unweighted, highly variable statistics dominate the acceptance decision, although the scale of a statistic is generally not informative of its relevance. This can be corrected for by the choice of weights r_{i_y} in (5.1). A common choice is as inversely proportional to measures of variability of the respective statistics, i.e.

$$r_{i_y} = 1/\sigma_{i_y} \tag{5.2}$$

with σ_{i_y} e.g. given via empirical standard deviations, $\sigma_{i_y} = \mathrm{Var}(\{y_{i,i_y}\}_{i\geq 1})^{1/2}$, where $\{y_i\}_{i\geq 1}$ are a calibration sample. An alternative measure suggested by Csilléry et al. [2012] for being more robust to extreme sample values, and used in Prangle [2017] and also throughout this chapter, is the median absolute deviation (MAD) to the sample median,

$$\sigma_{i_y} = \mathrm{MAD}_{i_y} := \mathrm{median}(\{|y_{i,i_y} - \mathrm{median}(\{y_{i',i_y}\}_{i'\geq 1})|\}_{i\geq 1}). \tag{5.3}$$

It is straightforward to use weights based on a calibration sample, e.g. from the prior distribution, prior to the actual ABC analysis [Beaumont et al., 2002, Csilléry et al., 2012, Fearnhead and Prangle, 2012]. However, while applicable to ABC-Rejection, Prangle [2017] demonstrate that in an ABC-SMC framework, the distribution of model outputs in later generations can differ considerably from one obtained in pre-calibration. Their relative variability may change as the proposal distribution focuses on high-density regions in parameter space, such that weights obtained in calibration may no longer yield similar contributions of model outputs or summary statistics. Thus, they suggest a sequentially updated generation-specific distance metric d_t, by updating the weights $r_j = r_j^t$ in distance (5.1) anew for each generation. The calibration sample for generation t consists of all samples created in generation $t-1$, including not only accepted ones but also rejected ones, which makes the model output distribution representative of the sampling process and allows for greater flexibility of adaptation. This serves as an approximation to the expected model output scale distribution under the proposal distribution g_t of generation t, without additional simulation cost. To ensure nested acceptance regions, Prangle [2017] then define the acceptance criterion for generation t as

3. accept (θ, y) if $d_{t'}(y, y_{\mathrm{obs}}) \leq \varepsilon_{t'}$ for all $t' \leq t$.

When using an adaptive scheme for ε_t, e.g. based on quantiles of distance values of accepted

particles as introduced in Section 2.4, it must be based on distance values recalculated with the new distance metric, $\{d_t(y_i^{t-1}, y_{\text{obs}})\}_{i \leq N}$.

While the above algorithm bases the weights for generation t on samples from generation $t-1$ with a different proposal distribution, Prangle [2017] further suggest a second algorithm that bases the weights on the current generation, by delaying the acceptance criterion definition. As this algorithm was found to not substantially improve results over the above-introduced one, while being not as integrable into existing ABC-SMC frameworks, we did not pursue this second formulation further here.

5.3 Robust adaptive distances for outlier-corrupted data

In this section, we describe distance metrics that build upon the adaptive distance metrics introduced in Section 5.2, but are robust to outliers in the data. We assume the data $y_{\text{obs}} \in \mathbb{R}^{n_y}$ to be a realization of the model $\pi(y|\theta)$, except for single outliers that are generated by another confounding mechanism.

5.3.1 Outlier-robust adaptive distances

While using MAD as a robust measure of sample variability for weight definition, Prangle [2017] base the overall distance on an L2 norm ($p = 2$ in (5.1)). Squared residuals emphasize large errors, which may be desirable, however makes the analysis highly sensitive to outliers, which typically result in large residuals that can dominate the distance and reduce the relative importance of other model outputs, leading to an overall worse performance.

In regression analysis, robust approaches have been developed that are more robust to outliers than standard least squares. The arguably most common alternative are absolute deviations, corresponding to an L1 norm ($p = 1$ in (5.1)). For ODE models, it was shown in Maier et al. [2017] that replacing the most common assumption of a normal noise model, corresponding to a weighted L2 distance (see Section 2.2.1), by heavy-tailed distributions such as Laplace, corresponding to an L1 distance, Huber, Cauchy or Student's t, makes the analysis considerably more robust to outliers, while performing roughly comparably on outlier-free data. In ABC, if noise model and dynamics description can be decoupled, similar approaches can be employed, either by simulating e.g. Laplace measurement noise, or using an appropriate acceptance kernel [Schälte and Hasenauer, 2020]. However, in general, in ABC the model must be regarded as a black box that does not allow to do so. In that case, robustification and model error correction must happen on the level of the distance metric, which quantifies the relative impact of data points.

Here, based on the above motivation, we propose to replace the weighted L2 norm in the adaptive distance formulation by a weighted L1 norm, yielding a distance metric both robust to outliers and adaptive to the problem structure. In the following, we shortly motivate why this change of norm renders the analysis more robust to outliers.

Consider the distribution of a single model output, or summary statistic, given via a random

variable $Y \in \mathbb{R}$, with variance $\mathbb{E}[Y^2] = \sigma^2 < \infty$, and observed value y_{obs}. We are interested in the variability of the corresponding distance component in (5.1) around its average value, as similar levels of variability of different model outputs impact the acceptance decision similarly. For the variance of an unweighted L1 distance component $|Y - y_{\text{obs}}|$ holds

$$\begin{aligned} 0 \leq \text{Var}[|Y - y_{\text{obs}}|] &= \mathbb{E}[(Y - y_{\text{obs}})^2] - \mathbb{E}[|Y - y_{\text{obs}}|]^2 \\ &= \text{Var}[Y] + \mathbb{E}[Y - y_{\text{obs}}]^2 - \mathbb{E}[|Y - y_{\text{obs}}|]^2 \leq \text{Var}[Y] = \sigma^2. \end{aligned}$$

Moreover, $\text{Var}[|Y - y_{\text{obs}}|] \to \sigma^2$ in the limit $|y_{\text{obs}}| \to \infty$ of extreme outliers, with smaller variance for observed values closer to $\mathbb{E}[Y]$. Thus, for simplicity assuming that in (5.1) normalization is by the distribution standard deviation σ, it follows for the variance of the corresponding component of a weighted L1 distance $0 \leq \text{Var}[|Y - y_{\text{obs}}|/\sigma] \leq 1$. Note that the absolute value of the distance component will be large for large outliers, however it only acts as an offset inconsequential to adaptive acceptance threshold schemes, resulting in an offset to automatically determined acceptance thresholds ε, whilst the variance is constrained.

For L2 distances, a similar boundedness does not apply. For example, if $Y \sim \mathcal{N}(0, \sigma^2)$, it is $\text{Var}[(Y - y_{\text{obs}})^2] = 2\sigma^4(1 + 2y_{\text{obs}}^2/\sigma^2)$, and therefore for the weighted distance component $\text{Var}[((Y - y_{\text{obs}})/\sigma)^2] = 2(1 + 2y_{\text{obs}}^2/\sigma^2)$, diverging for $|y_{\text{obs}}| \to \infty$, as one easily sees from properties of non-central χ^2 distributions.

5.3.2 Online outlier detection and down-weighting via bias correction

Even when using an outlier-insensitive distance such as L1, large outliers still influence the analysis, rendering it desirable to detect and down-weight them. Here, we propose an online simulation-based approach to do so in ABC-SMC by complementing MAD (5.3), as a measure of sample variability, by a measure of deviation from the observed value, such as the median absolute deviation to observation (MADO) for model output i_y,

$$\text{MADO}_{i_y} := \text{median}(\{|y_{i,i_y} - y_{\text{obs},i,i_y}|\}_{i \geq 1}).$$

To account for both in-sample variability and deviation to observed value, we propose to then define the weight in (5.1) via the combined median absolute deviation (CMAD) as

$$r_{i_y} = 1/\,\text{CMAD}_{i_y} := 1/(\text{MAD}_{i_y} + \text{MADO}_{i_y}). \tag{5.4}$$

If the model describes a data point well, one would expect the simulations to be close to the observed data, roughly on the same order as the in-sample variation of simulations, such that the additional term does not substantially hinder the purpose of variance normalization. Conversely, outliers with large deviations from commonly sampled values are down-weighted. Note that here we only down-weight data points while not removing them entirely as long as $r_j \neq 0$, which conceptually permits the recovery from badly assigned weights at least in the approximate limit as $t \to \infty$, i.e. no information is lost.

Especially in early iterations and for uninformative priors, common model simulations may deviate from observed data substantially, as may the relative variability of different model simulations, compared to later generations. In order to not "punish" model outputs that happen to deviate

more from the observed data initially than others, we propose to apply the deviation correction only if no more than a small fraction of the data points exhibit a substantial deviation. Concretely, here we only used CMAD for weighting in a generation if

$$\#\{i_y : \mathrm{MADO}_{i_y} > 2 \cdot \mathrm{MAD}_{i_y}\}/n_y \leq 1/3,$$

resorting to MAD otherwise. These hyperparameters are clearly heuristics and may require tuning on some problems, however e.g. more than one outlier in three data points should in practice hardly occur, while only counting as outliers points with sufficiently high deviation compared to the in-sample variability focuses on large deviations that could otherwise considerably impact the analysis, while small outliers are less problematic (see also Section 5.3.1). We denote this weighting scheme, which perhaps uses CMAD and otherwise only MAD, as PCMAD.

As outliers usually get apparent only in later generations, where simulations resemble the observed data more and fluctuate less, the application of such outlier correction methods only makes sense in an adaptive framework, as done here.

Instead of the robust measures MAD and MADO employed here, in principle also other measures of variability and deviation can be used, such as the common standard deviation and bias. We can then statistically interpret the above introduce weighting scheme as, informally, replacing the in-sample standard deviation by the root mean square error $\mathbb{E}[(y - y_{\mathrm{obs}})^2]^{1/2} = (\mathrm{Var}[y] + \mathrm{Bias}[y, y_{\mathrm{obs}}]^2)^{1/2}$, or robust alternatives thereof, regarding the simulations as predictors of the observed value.

5.3.3 Convergence

Under certain assumptions on the adaptive distance weights in the ABC-SMC formulation above, convergence of the approximate posterior to the true one, $\pi_{\mathrm{ABC},\varepsilon_t}(\theta|y_{\mathrm{obs}}) \to \pi(\theta|y_{\mathrm{obs}})$ in an appropriate sense for $t \to \infty$ and $\varepsilon_t \to 0$, holds, as for fixed-distance ABC in general under mild assumptions, as made explicit by the following

Theorem 5.1. *Consider prior $\pi(\theta)$ and likelihood $\pi(y|\theta)$ densities, with posterior $\pi(\theta|y_{obs}) \propto \pi(y_{obs}|\theta)\pi(\theta)$, and measurable distance metrics d_t. Assume that for the acceptance regions $A_t = \{y : d_t(y, y_{obs}) \leq \varepsilon_t\}$ holds $\lim_{t\to\infty} |A_t| = 0$, where $|\cdot|$ denotes the Lebesgue measure, and that the A_t have bounded eccentricity at y_{obs} (also referred to as shrinking regularly to y_{obs}), i.e. there is a constant $c > 0$ such that for every A_t there is a ball B_t with $y_{obs} \in B_t$, $A_t \subset B_t$, and $|A_t| \geq c\,|B_t|$. Then, for functions $\xi : \mathbb{R}^{n_\theta} \to \mathbb{R}$ with $\mathbb{E}_{\pi(\theta)}[|\xi|] < \infty$ holds*

$$\mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon_t(\theta|y_{obs})}}[\xi] \xrightarrow{t\to\infty} \mathbb{E}_{\pi(\theta|y_{obs})}[\xi] \quad \textit{for almost-all } y_{obs}. \tag{5.5}$$

Recall that in this chapter, unlike in Section 2.4, we assume for notational simplification that $y = \bar{y}$ incorporates measurement noise, and thus accurately describes the underlying data generation process $y_{\mathrm{obs}} \sim \pi(y|\theta)$, and potentially incorporates summary statistics s.

Proof. Adapted from Barber et al. [2015] to adaptive distances, and from Prangle [2017] to a

more rigorous formulation. Define

$$\phi_\xi(y) = \int \xi(\theta)\pi(y|\theta)\pi(\theta)\,d\theta \quad \text{and} \quad \phi_\xi^t(y_{\text{obs}}) = \frac{1}{|A_t|}\int_{A_t} \phi_\xi(y)\,dy\,. \tag{5.6}$$

Note that ϕ_ξ is integrable, as $\mathbb{E}_{\pi(\theta)}[\xi] = \mathbb{E}_{\pi(\theta,y)}[\xi]$, where $\pi(\theta, y) = \pi(y|\theta)\pi(\theta)$ denotes the joint distribution of θ and y, and using Fubini's Theorem. Thus, we have, with the same notation for $\xi \equiv 1$,

$$\mathbb{E}_{\pi(\theta|y_{\text{obs}})}[\xi] = \frac{\phi_\xi(y_{\text{obs}})}{\phi_1(y_{\text{obs}})} \quad \text{and} \quad \mathbb{E}_{\pi_{\text{ABC},\varepsilon_t}(\theta|y_{\text{obs}})}[\xi] = \frac{\phi_\xi^t(y_{\text{obs}})}{\phi_1^t(y_{\text{obs}})}.$$

Given integrability of $\phi_1(y) = \pi(y)$ and $\phi_\xi(y)$, $|A_t| \to 0$, and bounded eccentricity of the A_t, the Lebesgue Differentiation Theorem [Stein and Shakarchi, 2005, Corollaries 1.6 and 1.7] gives $\phi_1^t(y_{\text{obs}}) \to \phi_1(y_{\text{obs}})$ and $\phi_\xi^t(y_{\text{obs}}) \to \phi_\xi(y_{\text{obs}})$ for $t \to \infty$, for almost-all y_{obs}, such that (5.5) follows. □

Note that the assumption that $\mathbb{E}_{\pi(\theta)}[|\xi|] < \infty$, i.e. ξ is integrable under the prior, is not overly restrictive, as we are interested in integrability under the posterior, but could be relaxed to local integrability of ϕ_ξ in an environment of y_{obs} for (5.6). If all functions all continuous, convergence in (5.5) holds point-wise, as one easily sees, see e.g. the proof of Stein and Shakarchi [2005, Theorem 1.3]. As y_{obs} is in practice an observed sample, the statement "for almost-all y_{obs}" is no restriction. See e.g. Barber et al. [2015] for further discussion, in particular of convergence speed. While we have here assumed a Lebesgue density $\pi(y, \theta)$, a similar argumentation also holds for discrete parameters θ, which only affects the form of ϕ_ξ. For discrete data y, the assumptions on the A_t imply simulations $y = y_{\text{obs}}$ eventually, in which case the statement (5.5) directly follows.

In our case, convergence of the acceptance regions to y_{obs}, $|A_t| \to 0$, does not necessarily hold in common ABC-SMC implementations, e.g. when ε is quantile-based. Bounded eccentricity holds if, as on easily sees, $r_j^t \neq 0$ for all weights, and the maximum weights ratio

$$\sup_t \frac{\max_{i_y \leq n_y} r_{i_y}^t}{\min_{i_y \leq n_y} r_{i_y}^t} < \infty$$

is bounded. Practically, this can be achieved by constraining the relative weight range to a compact interval in $(0, \infty)$, or e.g. by updating the distance metric only a finite number of times. In the results presented here, we did not employ any such restrictions, as we focused on the ability of the methods to retrieve information under a limited simulation budget. Yet, reliable strategies to set such constraints may be practically of relevance and may require further investigation.

5.3.4 Implementation

We implemented all methods presented in this chapter in pyABC, putting emphasis on a modular and easy-to-use, well-documented implementation. The code underlying the study in this chapter can be found on GitHub at `https://github.com/yannikschaelte/study_abc_rad`, a snapshot of code and data is on Zenodo at `http://doi.org/10.5281/zenodo.5136475`. The implemented methods further e.g. allow to impose restrictions on how and when to update weights, and allow the combination with automatic and adaptive summary statistic calculation schemes as

Table 5.1: Properties of test models used in Chapter 5: Identifier, short description, number of parameters n_θ and data points n_y, population size N and maximum number of model simulation after which an analysis was terminated.

ID	Description	n_θ	n_y	N	Max. sim.
M1	Informative and uninformative Gaussian variables	1	11	1000	100000
M2	Independent Gaussian replicates	1	10	1000	100000
M3	Conversion reaction ODE model	2	10	1000	100000
M4	g-and-k distribution order statistics	4	7	1000	100000
M5	Lotka-Volterra Markov jump process model	3	32	200	50000
M6	Tumor spheroid growth agent-based model	7	150	500	150000

in Fearnhead and Prangle [2012], see also Chapter 6. All simulations were performed on the GCS Supercomputer JUWELS at Jülich Supercomputing Center (JSC), using up to 384 cores in parallel.

If not stated otherwise, we defined the acceptance criterion in generation t as $d_{t'}(s, s_{\text{obs}}) \leq \varepsilon_{t'}$ for all $t' \leq t$, including previous acceptance criteria, to ensure nested acceptance regions. The distance weights were calculated based on all samples generated in the previous generation, including accepted and rejected ones. As transition kernel, we used a multivariate normal distribution with covariance kernel proportional to the previous generation's weighted sample covariance matrix by Silverman's rule of thumb [Klinger et al., 2018]. Acceptance thresholds were automatically selected as the median of the updated distance values $\{d_t(y_i^{t-1}, y_{\text{obs}})\}_{i \leq N}$ of the previous generation's accepted particles.

5.4 Application to test problems

We evaluated the proposed methods on five test problems M1-5 covering different problem features, model types, and outlier scenarios, and a more realistic application problem M6, on both outlier-free and outlier-corrupted data, and compared them to established calibrated and adaptive distance metrics as introduced in Prangle [2017].

5.4.1 Test problems

An overview of the core problem features is given in Table 5.1. Details on all test problems can be found in the supplementary information of Schälte et al. [2021a, Chapter 1].

M1 consists of 10 observables $y_j \sim \mathcal{N}(\theta, 1)$, $j = 1, \dots, 10$, for a parameter θ with prior $\theta \sim U[0, 10]$ and true value $\theta = 5$, and one uninformative variable of low variance $y_{11} \sim \mathcal{N}(5, 0.1^2)$. As outlier, we set $y_{11} = 7$.

M2 consists of 10 observables $y_j \sim \mathcal{N}(\theta, 0.2^2)$, $j = 1, \dots, 10$, with prior $\theta \sim U[0, 10]$ and true value $\theta = 6$. While all observables are informative of the parameters, we set two to $y_j = 0$ to simulate conflicting information in the data due to outliers.

M3 is an ODE model of a conversion reaction $A \overset{\theta_1}{\to} B$, $B \overset{\theta_2}{\to} A$, assuming species B to be observed

at 10 evenly spaced time-points. We assumed additive $\mathcal{N}(0, 0.02^2)$ distributed noise. True parameters are $(\log\theta_1, \log\theta_2) = (-1.5, -1.5)$, estimated on log-scale, with prior $U[-3.5, 1]^{\otimes 2}$. Initial conditions $(A_0, B_0) = (1, 0)$ were assumed known. As outlier, we randomly set one data point to zero, which could in practice e.g. occur as a wrongly assigned missing value. This model is similar to a test model used in Chapter 4 and identical to the first test model in Maier et al. [2017].

M4 and M5 are common ABC test problems, adopted directly from the application examples in Prangle [2017]. M4 is based on the g-and-k distribution, which is defined via its quantile function

$$q(x) = A + B\left(1 + c\frac{1-\exp(-gz(x))}{1+\exp(-gz(x))}\right)(1+z(x)^2)^k z(x),$$

where $z(x)$ is the quantile function of a standard normal distribution. It does not have a closed-form likelihood function, but can be sampled from via $z \sim \mathcal{N}(0,1)$. As commonly done, we fixed $c = 0.8$ and estimated the remaining parameters. We used as summary statistics seven order statistics at indices $(1250, 2500, ..., 8750)$ out of $10,000$ independent samples. We used independent $U[0, 10]$ priors on the model's four parameters A, B, g, k, and ground truth values $(A, B, g, k) = (3, 1, 1.5, 0.5)$. As outlier, we considered randomly setting one observable to zero. M5 is a Markov jump process model of a Lotka-Volterra predator-prey population model. The underlying simplified population dynamics of prey x_1 and predators x_2 are given via

$$\begin{aligned}
(x_1, x_2) &\xrightarrow{\theta_1} (x_1 + 1, x_2) && \text{(prey growth)},\\
(x_1, x_2) &\xrightarrow{\theta_2} (x_1 - 1, x_2 + 1) && \text{(predation)},\\
(x_1, x_2) &\xrightarrow{\theta_3} (x_1, x_2 - 1) && \text{(predator death)}.
\end{aligned}$$

Simulations were performed via Gillespie's direct algorithm [Gillespie, 1977], using the author's Python package ssa (`https://pypi.org/project/ssa`). We assumed both species to be observed under additive $\mathcal{N}(0, \exp(2.3))$ distributed noise, at 16 evenly spaced time-points over a span of roughly four periods. The model has three conversion rate parameters, which were estimated on log-scale, with wide independent $U[-6, 2]$ priors and ground-truth values $(\theta_1, \theta_2, \theta_3) = (1, 0.005, 0.6)$. As outliers, we considered multiplying 6 observables at 3 random time-points by a factor of 10, which can in practice occur e.g. due to a wrong exponent. M5 exhibits highly variable dynamics, with different average behavior under the wide prior compared to the posterior, and may thus be regarded as a more challenging inference problem.

M6, a more realistic application example, is the agent-based model of tumor spheroid growth as described in Chapter 4. For the three data types, here considering with finer resolution in total $n_y = 150$ data points, we again assumed independent additive normal noise of different variances. As outliers, we considered interchanging in total 20 data points in the observables' dynamic regimes.

5.4.2 Experimental setup

To systematically assess the performance of the presented distance metrics, we ran M1-5 for each distance metric 20 times on outlier-free data, each for a separate data set randomly sampled from the model under ground truth parameters θ_{GT}, and additionally 20 times on outlier-corrupted

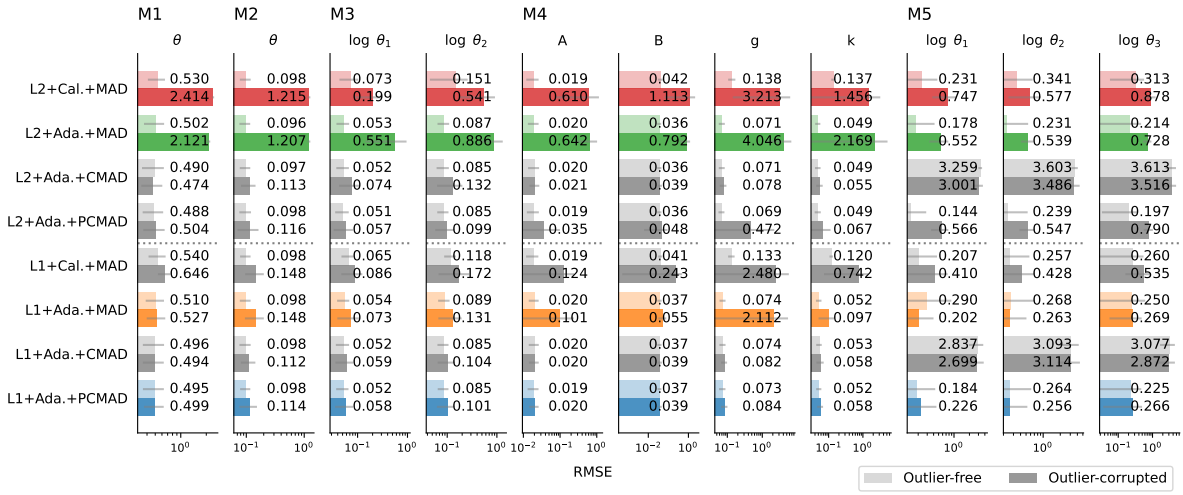

Figure 5.2: Mean RMSE for the parameters of models M1-5 (columns) obtained for 8 distance metrics (rows), using L2 or L1 norms, calibrated only in the first ("Cal.") or every ("Ada.") generation, and using MAD, CMAD, and PCMAD for distance weight calculation. Each RMSE is averaged over 20 data sets, grey lines indicate standard deviations. For each distance, the upper, lighter bar is based on outlier-free data, while the lower, darker bar is based on outlier-corrupted data. Distances of interest are colored, alternative distance combinations are shown in grey for reference. This figure is taken from the author's publication Schälte et al. [2021a].

data sets derived from the outlier-free ones. Each run was performed with a fixed population size N and given a budget of model evaluations, evaluated after each generation (see Table 5.1). The evaluation budgets were chosen rather low, in order to assess the ability of the distance metrics to extract information under a limited budget.

The performance of a distance metric was evaluated by its ability to yield accurate point estimates with low uncertainties. Therefore, we used, as a metric combining both aims, the root mean square error (RMSE) of the weighted posterior samples from the last ABC-SMC generation with respect to the ground truth parameters, i.e. the square root of $\mathbb{E}[(\theta - \theta_{\mathrm{GT}})^2] = \mathrm{Var}[\theta] + \mathrm{Bias}[\theta, \theta_{\mathrm{GT}}]^2$ under the obtained posterior approximation. While the actual posterior mean may not be unbiased for a given data set, one may expect it to be on average.

As distance metrics, we considered L2 and L1 distances ("L2" or "L1"), only pre-calibrated in the first generation or adapted in each generation ("Cal." or "Ada."), and using MAD, CMAD, or PCMAD for distance weight calculation. In the following, e.g. L2+Ada.+MAD denotes an L2 distance with adaptive MAD-based weights. While we also explore other combinations to assess the contribution of the various hyperparameters separately, we focus the analysis on the established pre-calibrated L2+Cal.+MAD, the established adaptive L2+Ada.+MAD introduced by Prangle [2017], and the here newly introduced robust L1+Ada.+MAD and L1+Ada.+PCMAD.

5.4.3 Results

L1 more robust than L2

Comparing L2- and L1-based distances (Figure 5.2, upper half vs. lower half) on outlier-free data (Figure 5.2, light bars), e.g. for L2+Ada.+MAD vs. L1+Ada.+MAD, shows no substantial differences for most models and parameters, indicating that L1 works similarly well on outlier-free data.

On outlier-corrupted data (Figure 5.2, dark bars), using an L2 distance (L2+Ada.+MAD) yielded substantially higher RMSE values compared to outlier-free data. Using an L1 distance (L1+Ada.+MAD) drastically reduced the RMSE in all cases by up to orders of magnitude, in many cases to only slightly higher levels than for outlier-free data.

Comparing pre-calibrated L2+Cal.+MAD and adaptive L2+Ada.+MAD (as well as the L1 variants) on outlier-free data confirms the finding in Prangle [2017] that the adaptive weighting scheme outperforms pre-calibrated weighting. However, e.g. the comparison of L2+Cal.+MAD and L2+Ada.+MAD on outlier-corrupted data shows that there, in some cases, an adaptive weighting scheme with L2 distances can give worse results than pre-calibrated weights, arguably due to decreasing in-sample variability with large impact on the distance metric, as outlined in Section 5.3.1. For the corresponding L1 distances, in the majority of cases, an adaptive distance metric did improve results over pre-calibration.

Outlier correction further improves estimates

Applying active outlier detection and down-weighting by PCMAD on top of using an L1 distance (L1+Ada.+PCMAD) further reduced RMSE values over MAD (L1+Ada.+MAD) on M1-4, and substantially so e.g. for the A and g parameters of M4, giving a total reduction of up to nearly 50 times compared to the established L2+Ada.+MAD. On M1-4, the use of CMAD throughout (L1+Ada.+CMAD) and not only if outliers were detected in less than one in three data points (PCMAD), yielded low RMSE values too. However, this was not the case for M5, where L1+Ada.+CMAD gave large RMSE values on both outlier-free and outlier-corrupted data. This is likely due to the fact that M5 exhibits highly flexible dynamics. In that case, blindly applying bias correction may wrongly down-weight data points that just happen to not have converged yet, while PCMAD reliably only did so in case of strong indication of few outliers.

Overall, in many cases L1+Ada.+MAD already gave substantially better results than the reference method L2+Ada.+MAD, with further improvements by active outlier detection, such as here L1+Ada.+PCMAD, which consistently yielded small RMSE values for all parameters of all test problems.

Nested acceptance regions improve performance slightly

We also assessed the effect on robustness of nested acceptance regions via checking previous acceptance criteria, $d_{t'}(s, s_{\text{obs}}) \leq \varepsilon_{t'}$, $t' \leq t$, which was employed in Prangle [2017] without

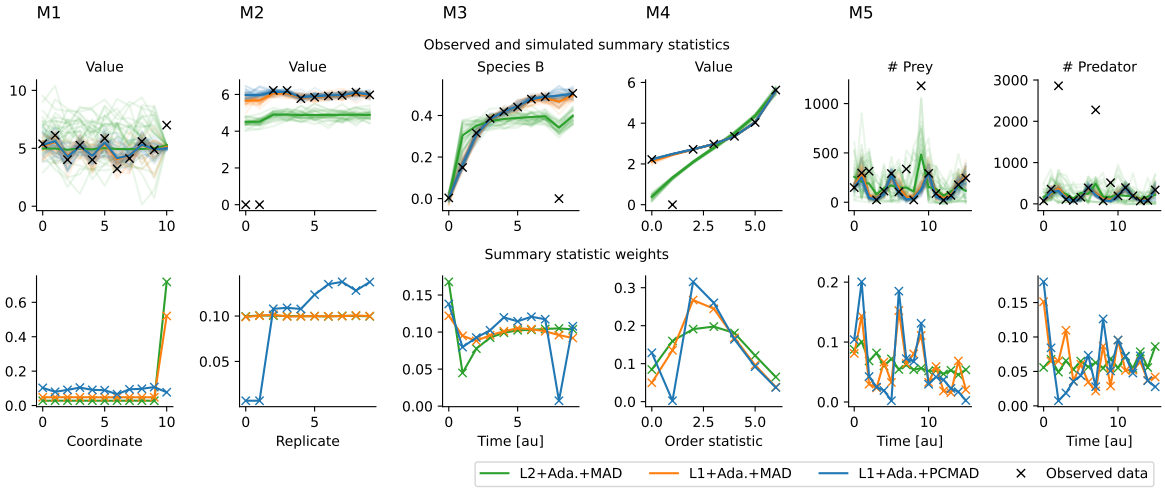

Figure 5.3: Fits and weights for models M1-5 on outlier-corrupted data, for three distances. Top: Observed data (black) and, for each distance, 30 accepted simulated data sets (lighter lines) as well as the sample means (darker lines) from the last ABC-SMC generation. Note that these are accepted simulations, not predictions; for $\varepsilon \to 0$, the accepted simulations should exactly match the observed (non-outlier) data. Bottom: The corresponding weights r_j assigned to each model output or summary statistic by the three shown distance metrics in the last generation, normalized to sum 1. For each problem, one exemplary run out of the 20 runs on outlier-corrupted data is shown. This figure is taken from the author's publication Schälte et al. [2021a].

comparison. Thus, we repeated the same ABC analyses as shown above, but only taking into account the current generation's acceptance criterion, i.e. accepting if $d_t(s, s_{\text{obs}}) \leq \varepsilon_t$ (for details see the supplementary information of Schälte et al. [2021a, Section 2]). Overall, we found no major or structural differences between using or discarding previous acceptance criteria, especially for L1 distances, on both outlier-free and outlier-corrupted data, with slightly better results when using them. An exception, where the use of previous acceptance criteria performed substantially better, was M5, which especially for L2 showed a larger RMSE for the adaptive distance compared to the pre-calibrated distance when only checking the current criterion. A reason for this may be, again, the highly flexible model dynamics, for which a more conservative acceptance criterion may be more robust. Both using and discarding previous acceptance criteria may have their advantages – using them may be more conservative and thus preferable on highly variable models, but may be unnecessary for L1 distances, further not using them gives more flexibility and may allow to escape from bad initial choices.

Robust distances yield better fits of the data

The advantage of the robust adaptive distances L1+Ada.+MAD and L1+Ada.+PCMAD over L2+Ada.+MAD became apparent not only in terms of parameter estimates, but also of fits of simulated data to the observed data (shown in Figure 5.3 exemplarily for selected data sets). For M1, the variability of simulated data for L2+Ada.+MAD was considerably higher than for the L1 distances. For M2, L2+Ada.+MAD balanced simulations visually between the eight accurate measurements and the two outliers, fitting no point well, while L1+Ada.+MAD was considerably less biased, only slightly more than L1+Ada.+PCMAD. This held similarly for M3 and M4, where the analysis for L2+Ada.+MAD converged to curves visibly different from the

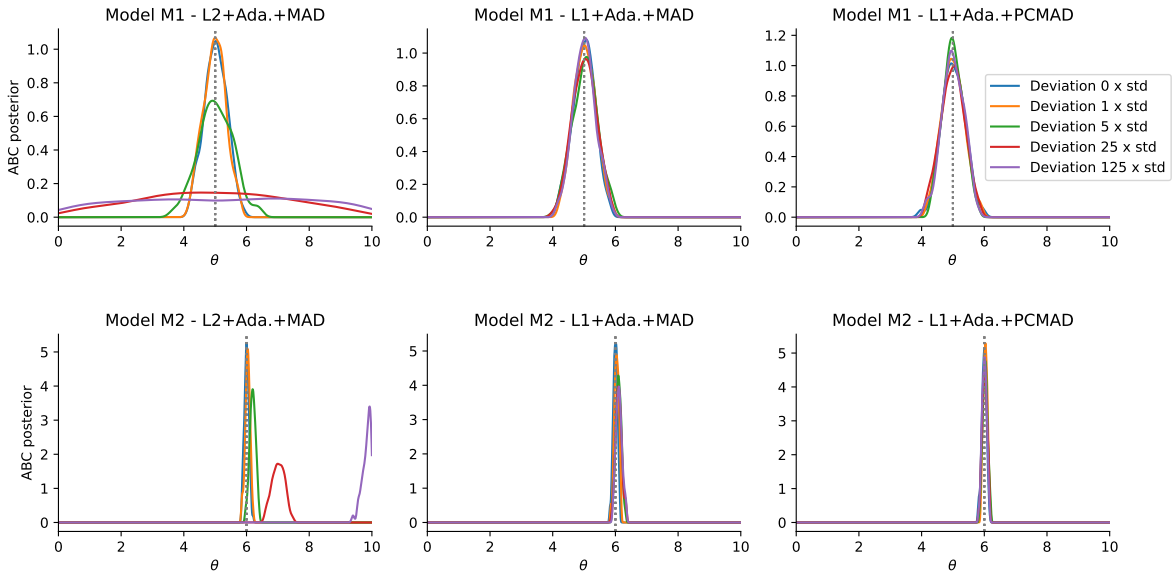

Figure 5.4: Impact of scale of deviation on the analysis. Top: Problem M1. Bottom: Problem M2. Left to right: Adaptive L2 norm with MAD weights, adaptive L1 norm with MAD weights, and adaptive L1 norm with PCMAD weights. Shown are five outlier scales with perfect data at the distribution means, except for the outlier data points which were given deviations of 0, 1, 5, 25, 125 times the standard deviation of the respective model outputs. True parameters ($\theta = 5$ for M1, $\theta = 6$ for M2) are indicated by grey dotted lines. This figure is taken from the supplementary information of the author's publication Schälte et al. [2021a].

observed data, corresponding to different optimal parameter vectors. For M5, extremes were less well captured. Overall, L1+Ada.+PCMAD provided the best fits in all these cases.

Also in the weights the respective distances assigned to data points in the last ABC-SMC generation (Figure 5.3 bottom) was this reflected. For M1-4, PCMAD reliably detected outliers and assigned them low weights, reducing their impact on the analysis. Typically, bias correction was applied only after a few generations, due to overall high initial variability, reliably converging to the actual outliers in the later generations in most cases. For M5, PCMAD did often not identify all or any outliers. Likely, the flexibility of the model does not allow to reliably detect outliers, as too many data points exhibit deviations from observed values. Thus, in this case PCMAD does not give an advantage over MAD alone, resulting in similar RMSE values, while blindly applying CMAD is detrimental, as observed before.

Robust to outliers at large scale

In order to analyze the impact of the scale of outliers on the presented methods, we considered problems M1+2 with outliers at different multiples of the corresponding model outputs' standard deviations (Figure 5.4). This revealed that the performance of L2+Ada.+MAD worsened substantially with increasing outlier scale, while the L1-based distances were considerably less affected, with slight improvements for PCMAD over MAD. This is in line with the discussion in Section 5.3.1.

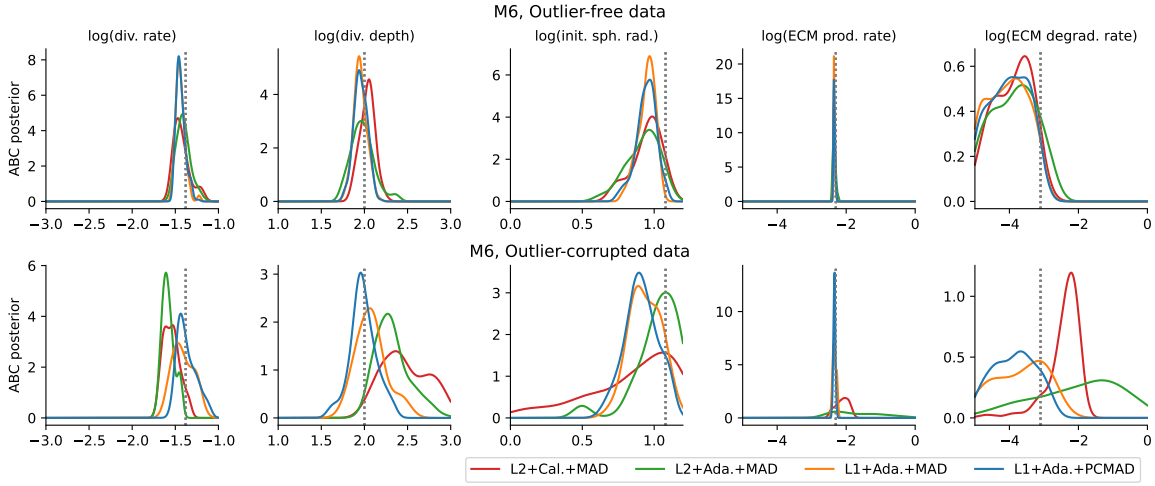

Figure 5.5: Posterior marginals for the 5 out of the 7 model parameters of model M6 showing interesting dynamics. Top: Without outliers. Bottom: With outliers. The x-axis boundaries are the uniform prior boundaries. The parameter values used to simulate the observed data are indicated by grey dotted lines. This figure is taken from the author's publication Schälte et al. [2021a].

Applicable to complex application example

Due to its computational complexity with model simulation times on the order of seconds, for the tumor spheroid growth application problem M6 we only considered single outlier-free and outlier-corrupted data sets. Here, we compared the pre-calibrated distance L2+Cal.+MAD, the adaptive L2 distance L2+Ada.+MAD, and the newly introduced adaptive L1 distances L1+Ada.+MAD and L1+Ada.+PCMAD.

Remarkably, the L1 distances performed superior to L2 on M6 also on outlier-free data, as the parameter estimates show (Figure 5.5 top). In particular L2+Ada.+MAD yielded substantially wider uncertainty estimates, also compared to L2+Cal.+MAD. This is also reflected by more scattered model simulations (Figure 5.6 top). Arguably, this is because for high-dimensional data the chance of deviant data points is high, even if the model perfectly describes the underlying data generation process. In that case, L2+Ada.+MAD may increasingly focus the analysis on points with large deviations, even if there are no systematic outliers in the data.

On outlier-corrupted data, both L2+Cal.+MAD and L2+Ada.+MAD performed badly, with highly uncertain or unreasonable parameter estimates (Figure 5.5 bottom), and simulations that visually do not match the observed data (Figure 5.6 bottom). In contrast, L1+Ada.+MAD and L1+Ada.+PCMAD performed considerably better, with parameter estimates closer to those obtained on outlier-free data, although generally with higher uncertainties, as expected given the high fraction of outliers. Overall, L1+Ada.+PCMAD yielded the lowest parameter uncertainties. Further, the outlier detection reliably corrected for virtually all outliers.

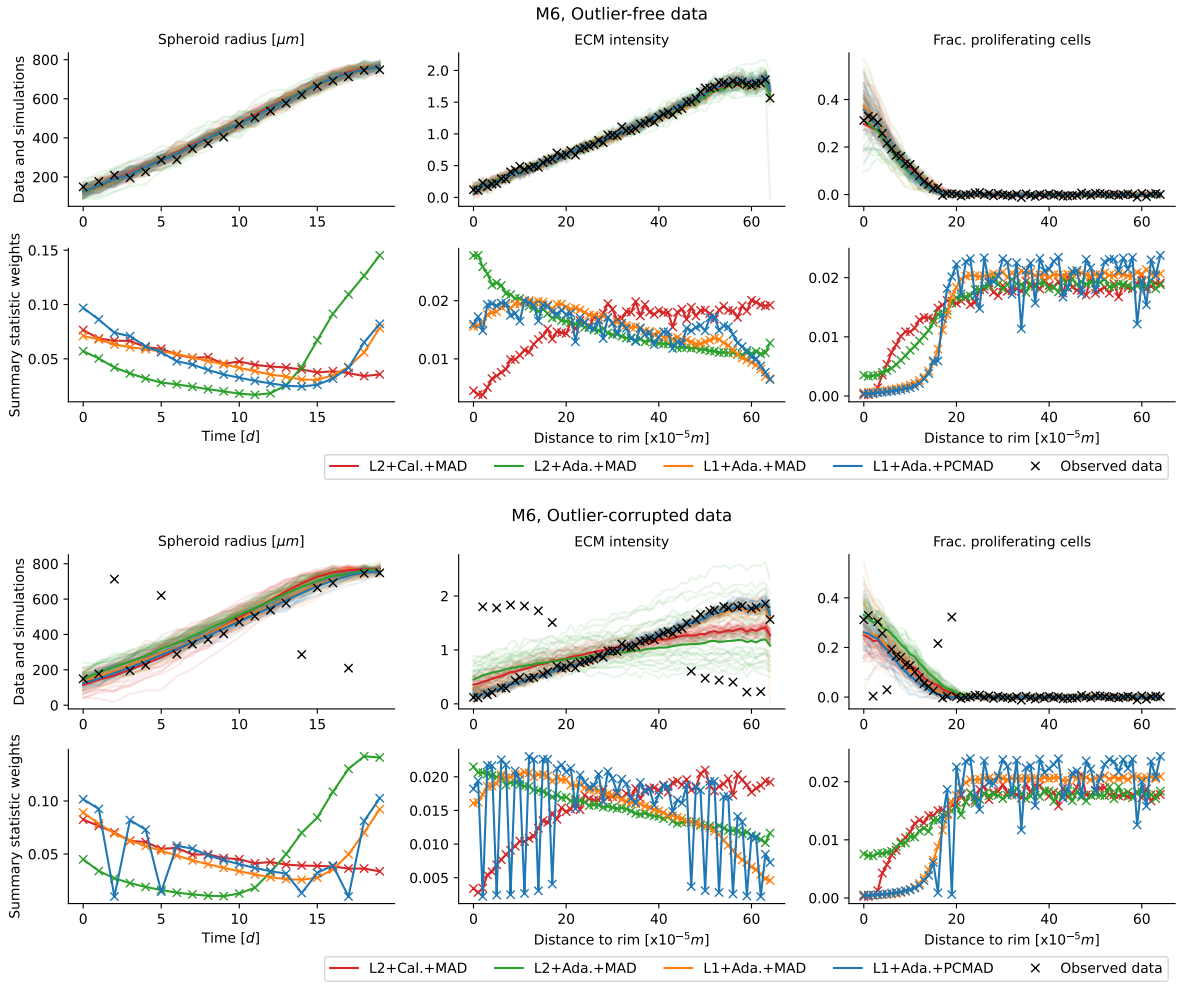

Figure 5.6: Fits and weights for four distance metrics on problem M6 on outlier-free (top) and outlier-corrupted (bottom) data. The respective upper rows show the observed data (black), and, for each distance, 30 accepted simulated data sets (light lines) as well as the sample means (darker lines) from the last ABC-SMC generation. Note that these are accepted simulations, not predictions; for $\varepsilon \to 0$, the accepted simulations should exactly match the observed (non-outlier) data. The respective lower rows show the corresponding weights assigned to each summary statistic by the four distance metrics in the last generation, normalized to sum 1. This figure is taken from the author's publication Schälte et al. [2021a].

5.5 Discussion

In this chapter, we demonstrated how established ABC distance metrics may be sensitive to outliers in the data, resulting in erroneous parameter estimates. As an extension of a popular distance metric with iteratively updated weights correcting for different data scales, we introduced a widely applicable robust adaptive distance metric based on an L1 norm, and further introduced a weighting scheme that allows for online simulation-based detection and down-weighting of outliers.

We evaluated and compared the novel methods on six test problems. Firstly, this demonstrated that the use of an outlier-robust distance metric such as L1 considerably improves performance in the presence of outliers, while performing similarly well on outlier-free data. Secondly, the outlier

detection and correction generally further improved results, as outliers were correctly identified and sufficiently down-weighted, allowing the analysis to extract information from the relevant data more efficiently. This indicates that actively identifying outliers, whether by this method or other ones as outlined in the introduction, may be of advantage. Also against extreme outliers, which severely affected reference methods, did the presented methods prove robust. As especially in high-dimensional complex data sets, e.g. based on imaging techniques commonly used in ABC applications, outliers or highly deviant data points may exist, we recommend, given our results, the consistent use of robust distance metrics in practical applications, unless there are reasons to not do so, e.g. if sensitivity to large deviations is desirable. Further, also the here introduced outlier correction method appears generally applicable, if only for detection, in order to inform early-on about potential problems in data or model.

One possible path of future research would be distances that combine the sensitivity of L2 to small deviations with the robustness of L1 to large outliers, such as a Huber distance, which however relies on the choice of hyperparameters. Also a comparison to alternative robust distance metrics as mentioned in the introduction would be of interest, wherever multiple approaches are applicable, however is beyond the scope of this chapter. Model selection could be employed to determine the most appropriate distance metric.

A limitation of our study is that we only considered relatively obvious synthetic outliers, which could easily be removed by manual inspection. This was for testing purposes, and as small outliers are less consequential and hard to detect by any method. A study in that regard and e.g. using real high-dimensional, e.g. imaging, data would be of interest.

As already stated in Prangle [2017], Mahalanobis type distances, i.e. the use of coordinate transformations, may be profitable especially for correlated data, however are numerically less robust and harder to interpret. Further, as in Prangle [2017], here we used median absolute deviations for weight calculation. Especially for discrete data, these may however yield zero weights. The handling of such deteriorated weights may in general be of relevance.

To ensure convergence, theoretically distance weights need to be either frozen after a number of generations, or their eccentricity bounded. It may be practically relevant to devise reliable strategies of doing so while allowing for sufficient weight adaptation.

A further promising research path becomes apparent in the last application example (Figure 5.6): One data type, the fraction of proliferating cells, is constant over most of the observed data points. While it is not clear whether this holds for all parameters, this raises the suspicion that these summary statistics are hardly informative of the parameters. Yet, they are assigned high weights due to their low variability. Thus, measuring the "informativeness" of summary statistics in combination with scale adaptation and outlier correction might considerably improve the analysis in such cases. This problem we address in the subsequent chapter.

In conclusion, we presented a novel robust adaptive distance metric with outlier down-weighting, and demonstrated its broad applicability. The methods are easy to adopt and have been implemented in pyABC, facilitating the straightforward use in application projects.

Chapter 6

Informative and adaptive distances and summary statistics in ABC

The practical performance of approximate Bayesian computation (ABC) relies on its ability to efficiently compare relevant features in simulated and observed data via summary statistics and distance metrics. Separately, semi-automatic schemes have been devised to firstly define low-dimensional, informative summary statistics by regression of parameters on simulated data, and to secondly adapt distance metrics to the data by scale normalization. In this chapter, which may be regarded as an extension of Chapter 5, we combine both ideas in two ways. Firstly, we demonstrate synergies of both approaches, by scale normalization for regression-based summary statistics. Secondly, we employ regression models not to transform data, but to inform additional sensitivity weights accounting for the information content of data on parameters. Moreover, we discuss problems arising for regression models under non-identifiability, and present a solution using regression target augmentation. On a dedicated test problem as well as on a realistic systems-biological application problem, we demonstrate substantially superior performance of in particular the sensitivity-weighted approach, as well as robustness and wide applicability on a collection of further test problems. Further, we demonstrate the applicability of the sensitivity-weighted distances to outlier-corrupted data.

This chapter is based on and partly identical to the following publication, in which all methods were developed and all analyses performed by the thesis author:

- **Schälte, Y.** and Hasenauer, J. (2021). Informative and adaptive distances and summary statistics in approximate Bayesian computation. *In preparation.*

6.1 Introduction

Accuracy and efficiency of ABC rely on its ability to extract information from simulated and observed data, via distance metrics and summary statistics. The adaptive distance formulation from the previous chapter, based on Prangle [2017], normalizes model outputs, or summary statistic representations thereof, by the respective scales they vary on, thus ensuring similar

contributions to the acceptance decision. However, an implicit assumption is that all model outputs are similarly informative of the parameters. If some model outputs are less or not informative, scale normalization can still be a reasonable, as widely applicable to heterogeneous data, compromise, but it would be desirable to either only consider informative statistics, or account for informativeness in the weighting scheme. For example, consider the tumor growth model from the previous section, where a substantial proportion of the data profiles appear to be essentially flat, raising the suspicion that the corresponding data points are uncorrelated with parameters and thus uninformative. Yet, due to their low variability, these data points were assigned large scale-normalizing weights (Figure 5.6).

Informative, low-dimensional summary statistics can improve performance of ABC in particular for high-dimensional, noisy data, due to the curse of dimensionality [Blum et al., 2013]. Various methods to robustly and semi-automatically construct such statistics have been developed, e.g. based on subset selection [Joyce and Marjoram, 2008, Nunes and Balding, 2010], or indirect inference [Drovandi et al., 2011, Gleim and Pigorsch, 2013], or regression [Borowska et al., 2021, Fearnhead and Prangle, 2012, Jiang et al., 2017], see Section 2.4.3. A particular approach, called "semi-automatic ABC" [Fearnhead and Prangle, 2012], uses as summary statistics the outputs of a linear regression (LR) model of parameters on data, over an adequate domain, thus an inverse model to the forward mechanistic process model. While such regression models can be heuristically motivated as summarizing the information content in the data on parameters in a single value per parameter, Fearnhead and Prangle [2012] further argue that using the posterior mean as summary statistic, of which the regression model serves as an approximation, conserves the true posterior mean in the approximate limit of vanishing acceptance threshold. To capture non-linear relationships, more sophisticated regression models have been employed, such as Gaussian processes (GP) as in Borowska et al. [2021], or neural networks (NN) as in Jiang et al. [2017].

To evaluate proximity of regression-based summary statistics, e.g. Euclidean distances have been used, or weighted Euclidean distances using weights based on calibration samples [Fearnhead and Prangle, 2012]. However, here essentially the same problems apply that motivated the use of adaptively weighted distances, shifted from the level of data to the level of parameters, or regression approximations thereof, used as summary statistics. In fact, as outlined above, to regression-based summary statistics the approach by Prangle [2017] is particularly applicable, as all statistics may be assumed to be informative. Another problem with regression-based statistics is that an inverse mapping may not always exist, for example when multiple parameters describe the observed data similarly well, corresponding to global parameter non-identifiability. In that case, an inverse model cannot reconcile all data.

In this chapter, we present two approaches that combine the concepts of adaptive distances and regression of parameters on data. Firstly, we discuss integrating statistics learning in an ABC-SMC framework with adaptive distances, and demonstrate performance improvements by scale-normalizing the derived statistics. Secondly, the focus of this chapter, we employ regression models not to transform data, but in order to inform additional sensitivity weights accounting for the information content of data on parameters. This approach may be regarded as an extension of the approaches of Prangle [2017] and the previous section, and as an alternative to an approach presented in Harrison and Baker [2020], who employ weights maximizing the difference of prior and posterior approximation, using a slightly different ABC-SMC formulation and notion of informativeness than used here (see discussion). Furthermore, we discuss the problem of

non-existence of an inverse mapping for regression models due to non-identifiability, and present a solution using augmented regression targets. On a dedicated test problem exhibiting multiple problematic features such as partly uninformative data, heterogeneous data and parameter scales, and parameter multi-modalities, we demonstrate how both scale-normalizing adaptive distances based on Prangle [2017] and Chapter 5, and regression-based summary statistics similar to Fearnhead and Prangle [2012] fail to efficiently approximate the true posterior distribution well, and demonstrate substantially improved performance of the newly introduced approaches. Similarly, on a more realistic and complex application example, we demonstrate their good performance, and demonstrate the applicability of sensitivity-weighted distance functions to outlier-corrupted data. On various common ABC test problems, we evaluate robustness.

6.2 Background: Regression-based summary statistics

In this chapter, we employ the notation of general parameter inference problem from Section 2.1, and of ABC-SMC methods from Section 2.4. For simplicity of notation and as we require no such distinction here, we assume that measurement noise is already incorporated in y, i.e. $\pi(\bar{y}|\theta) = \pi(y|\theta)$. In particular, let θ denote the parameters, y_{obs} the observed data, $\pi(\theta|y_{\text{obs}}) \propto \pi(y_{\text{obs}}|\theta)\pi(\theta)$ the posterior distribution of interest, d or d_t the ABC distance metric, ε the acceptance threshold, and $P_t = \{(\theta_i^t, y_i^t, w_i^t)\}_{i \leq N}$ the population of accepted particles in generation t.

As raw data generated by, especially spatial, models are often high-dimensional and noisy, in ABC deriving low-dimensional informative summary statistics is often an essential task. The "semi-automatic ABC" approach by Fearnhead and Prangle [2012] uses as summary statistics the outputs of a linear regression (LR) model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$, predicting parameters from simulated data. While leaving room for customization, the method works roughly as follows:

1. In an ABC pilot run, determine a posterior high-density region H.
2. Generate a population $P = \{(\theta_i, y_i)\}_{i \leq \tilde{N}} \sim \pi(y|\theta)I[\theta \in H]$, for some $\tilde{N} \in \mathbb{N}$.
3. Train a regressor model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$, $y \mapsto \theta$, on P.
4. Run the actual ABC analysis using s as summary statistics.

Step 1 uses no or manually derived summary statistics and has the purpose of defining a good training region for the regression model and can be skipped if the prior is informative [Fearnhead and Prangle, 2012]. In Borowska et al. [2021], $H = [0.5\tilde{\theta}, 2\tilde{\theta}]$ around a literature value $\tilde{\theta}$, based on manual prior experimentation, is used, which is generally not applicable as reliable literature values are usually not available. In Jiang et al. [2017], the prior is used, in one case constrained to an identifiable region, whereby step 1 is omitted. In step 4, the distance is then not in terms of the full data or model outputs y, but their summary statistics representation, acceptance defined as $d(s(y), s(y_{\text{obs}})) \leq \varepsilon$.

In step 3, Fearnhead and Prangle [2012] employed a simple LR model on potentially augmented data (specifically, on some problems taking also higher-order moments $y^1, \ldots, y^4$ into account as regressors), which they found to work sufficiently well. Jiang et al. [2017] and Borowska et al.

[2021] respectively used neural networks (NN) and Gaussian processes (GP) instead, aiming at a more accurate description of non-linear relationships, and further automating the generation of summary statistics. Arguably, the observed sufficiency of the linear model in the "semi-automatic" approach may be due to the substantial time spent in the pilot run, identifying a high-density region, in which a linear model may be sufficient, while e.g. Jiang et al. [2017] observed a clearly better posterior approximation using NN compared to LR. Also Borowska et al. [2021] reported better model predictions of GP compared to LR and LASSO.

A theoretical justification of regression-based summary statistics is that the regression model serves as an approximation to the posterior mean, e.g. for a linear model

$$s(y) = A \cdot y + b \approx \mathbb{E}_{\pi(\theta|y)}[\theta],$$

using which as summary statistic ensures that the ABC posterior approximation recovers the actual posterior mean as $\varepsilon \to 0$, the optimal Bayes estimator of θ under mean square error, see Blum et al. [2013], Fearnhead and Prangle [2012], Jiang et al. [2017], or Theorem 6.1 in this chapter, for details.

6.3 Adaptive and informative regression-based distances and summary statistics

In this section, we describe the novel methods presented in this chapter. Firstly, we describe a framework for integrated regression model training with adaptive distance functions, secondly we introduce sensitivity-weighted distance functions, and thirdly provide underlying theory on the use of expectations of parameters or transformations thereof as summary statistics.

6.3.1 Integrating summary statistics learning and adaptive distances

In previous applications, the regression approach of Section 6.2 has been used together with uniform or on a previous run pre-calibrated distance weights [Borowska et al., 2021, Fearnhead and Prangle, 2012, Jiang et al., 2017]. However, to the regression model outputs, as approximations of parameter means, the same problems apply that motivated the scale-normalized adaptive distance approach in Prangle [2017] and Section 5.2, as parameters varying on larger scales can dominate the analysis without scale adjustment, with potentially changing levels of variability over ABC-SMC generations. In fact, all summary statistics being similarly informative, the scale-normalized adaptive distance approach is particularly applicable.

Therefore, we propose to combine the regression-based summary statistics approach from Section 6.2 with the distance adaptation approach from Section 5.2, or its robust formulation from Section 5.3. In principle, the regression model can be pre-trained as previously done. Here, we however suggest to increase the level of automation and avoid the necessity of a pilot run by integrating the regression model training into the actual ABC-SMC run (see Algorithm 6.1): We begin by using the full model outputs as summary statistics (or some alternatively derived summary statistics representation thereof), with an adaptively scale-normalized distance. Then, in a generation $t_{\text{train}} \geq 1$, the regression model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$ is trained based on all particles

$\{(\theta_i^{t_{\text{train}}-1}, y_i^{t_{\text{train}}-1}, w_i^{t_{\text{train}}-1})\}_{i\leq\tilde{N}}$, $\tilde{N} \geq N$, generated in the previous generation. From $t \geq t_{\text{train}}$ onward, then the regression model outputs $s(y)$ are used as summary statistics, further using an adaptively scale-normalized distance metric with weights adjusted in each generation. The training samples include, as for adaptive distance weight calculation, also rejected samples. While this firstly increases the sample size used to train the regression model to typically a multiple of the population size N depending on the acceptance rate, giving a more robust model, it secondly gives a representative sample from the joint distribution of data and parameters, with parameters restricted to a high-density region, but not confined to $y \approx y_{\text{obs}}$, as for the distribution of accepted particles. As regression model, any model can be used, particularly including previously employed LR, NN, or GP. t_{train} may be pre-defined, or e.g. depend on the number of simulations performed, or other criteria.

The delay of regression model training until after a few generations serves to focus on a high-density region, similar to Fearnhead and Prangle [2012], such that simpler models suffice to describe the relationship of data and parameters sufficiently well. Potentially, the model could be retrained in later generations, or even in each generation. While Fearnhead and Prangle [2012] update the prior in step 4 based on the range of values obtained in the pilot run, we consider the prior as part of the problem formulation and thus do not update it. However, in generations $t \geq t_{\text{train}}$ the proposal distributions g_t will suggest most of the values within the range of values covered by $P_{t_{\text{train}}} - 1$, on which the model was trained, such that an explicit prior range update should not be necessary.

While this approach gives a high level of automation, it should be noted that model training may not always be possible to automatize, especially for more complex models such as NNs depending on hyperparameters, in which case it may also be regarded as two workflows with manual intervention at time t_{train}.

6.3.2 Regression-based sensitivity weights

The scale normalization of the adaptive distance methods from Prangle [2017] may be, operating on the full data, not ideal if data are not similarly informative. For example, there may be uninformative data points that hardly correlate with parameter values, or there may be many data points informative of one parameter, but only few informative of another parameter, such that the analysis constrains the first parameter more, if all data points are weighted equally. The adaptive distance method can even give worse performed than uniformly weighted distances, e.g. when uninformative data points underlying only background noise of small variance, which would usually hardly impact the analysis, are inflated. The regression model approach from Section 6.2 is one solution to obtain informative summary statistics, however it performs a complex transformation of the model outputs, which may hinder interpretability, and perform badly if the regression model is not accurate enough. In this section, we present an alternative approach, namely to, instead of replacing the summary statistics by the regression model outputs, only use the regression model outputs and inferred relationships to inform additional weights.

The idea is that we want to weight each data point by how "informative" it is of the underlying parameters. Informativeness we quantify via the sensitivity of how much the posterior expectations of parameters, or transformations thereof, given observed data, would vary under perturbations of the observed data y_{obs}. As in Section 6.2, we use a regression model to describe the inverse

Algorithm 6.1 ABC-SMC algorithm with regression-based summary statistics or sensitivity-weighted distances

for $t = 1, \ldots, t_{\text{train}} - 1$ **do**
  while less than N acceptances **do**
    sample parameter $\theta \sim g_t(\theta)$
    simulate data $\bar{y} \sim p(\bar{y}|\theta)$
    accept θ if $d_t(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon_t$, where d_t denotes a distance with adaptive scale weights r
  end while
  compute weights $w_i^t = \frac{\pi(\theta_i^t)}{g_t(\theta_i^t)}$, for accepted parameters $\{\theta_i^t\}_{i \leq N}$
  normalize weights $W_i^t = w_i^t / \sum_j w_j^t$
  define g_{t+1} and ε_{t+1} based on $\{(\theta_i^t, \bar{y}_i^t, W_i^t)\}_{i \leq N}$, as well as d_{t+1} based on all generated particles, if adaptive
end for
learn regression model s based on all generated particles $\{(\theta_i^{t_{\text{train}}-1}, \bar{y}_i^{t_{\text{train}}-1}, w_i^{t_{\text{train}}-1})\}_{i \geq 1}$
if using s to weight model outputs, define sensitivity weights $q_1, \ldots, q_{n_y}$
for $t = t_{\text{train}}, \ldots, n_t$ **do**
  while less than N acceptances **do**
    sample parameter $\theta \sim g_t(\theta)$
    simulate data $\bar{y} \sim p(\bar{y}|\theta)$
    accept θ if $d_t(s(\bar{y}), s(\bar{y}_{\text{obs}})) \leq \varepsilon_t$ if using the regression model to construct summary statistics, or $d_t(\bar{y}, \bar{y}_{\text{obs}}) \leq \varepsilon_t$ if using s only to define additional weights for d_t, in which case d_t denotes a distance using both scale weights r and sensitivity weights q
  end while
  compute weights $w_i^t = \frac{\pi(\theta_i^t)}{g_t(\theta_i^t)}$, for accepted parameters $\{\theta_i^t\}_{i \leq N}$
  normalize weights $W_i^t = w_i^t / \sum_j w_j^t$
  define g_{t+1} and ε_{t+1} based on $\{(\theta_i^t, \bar{y}_i^t, W_i^t)\}_{i \leq N}$, as well as d_{t+1} based on all generated particles, if adaptive
end for
output: weighted samples $\{(\theta_i^{n_t}, W_i^{n_t})\}_{i \leq N}$

mapping from data or model outputs to parameters.

More specifically, before generation t_{train}, as in Section 6.3.1, we learn a regression model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$ on the samples $\{(\theta_i^{t_{\text{train}}-1}, y_i^{t_{\text{train}}-1}, w_i^{t_{\text{train}}-1})\}_{i \leq \tilde{N}}$ generated in the previous generation. As regression model inputs, we use normalized simulations $y/\sigma_{t_{\text{train}}}$, with σ the measure of scale used in the adaptive distance formulation, e.g. based on the MAD (5.3). Further, we z-score normalize regression model targets θ, in order to make the model independent of parameter scale. Then, we calculate the sensitivity matrix

$$S = \nabla_y s(y_{\text{obs}}) \in \mathbb{R}^{n_y \times n_\theta} \tag{6.1}$$

at the observed data. To robustly approximate derivatives, we employ finite differences (Section 2.3.2) with automatic step size control, by testing multiple step sizes and picking the one minimizing the difference to adjacent ones. Then, we define the *sensitivity weight* of model output coordinate i_y as

$$q_{i_y} = \sum_{i_\theta=1}^{n_\theta} \frac{\left|S_{i_y i_\theta}\right|}{\sum_{j_y=1}^{n_y} \left|S_{j_y i_\theta}\right|}, \tag{6.2}$$

i.e. as the sum over the absolute sensitivities of all parameters with respect to model output coordinate i_y, normalized to one per parameter to level the impact of each parameter. This normalization can be omitted, but may yield more conservative and robust weights, accounting for the fact that the regression model may describe underlying relationships imperfectly, by assigning more evenly distributed weights when all sensitivities of some parameters are small.

The final weight used in the distance function (5.1) is then given as the product of scale weight (5.2) and sensitivity weight (6.2) as

$$r_{i_y} = q_{i_y}/\sigma_{i_y}, \tag{6.3}$$

with here σ_{i_y} e.g. given via the MAD (5.3), or, also taking into account bias via PCMAD, via the CMAD (5.4). This separate treatment of scale and sensitivity weights allows to e.g. include the error correction via CMAD (5.4) of Section 5.3 in the scale correction, but not in the normalized data used for regression model inputs, which would lead to an inverse rescaling of sensitivities, and thus simultaneously account for information content, and correct for outliers. Both scale and sensitivity weights can be updated independently of each other. Here, we update the sensitivity weights only once in generation t_{train} (see Algorithm 6.1), as training especially of complex regression models may be relatively expensive compared to parallelized model simulations, while the scale weights are updated in each generation. Conceptually, the regression model could however be updated in each generation as well, in particular when using computationally simple regression models such as LR, or computationally expensive mechanistic process models.

An advantage of this approach compared to employing the regression models to define summary statistics is that the data are not transformed themselves but only re-weighted. Thus, it is less intrusive, and may be expected to be more robust against regression model misspecification. Further, the obtained weights can be easily interpreted. Indeed, as long as all weights r_{i_y} are non-zero, the original posterior $\pi(\theta|y_{\text{obs}})$ can be conceptually recovered for $\varepsilon \to 0$ by Theorem 5.1, i.e. no information is lost, while practically convergence will clearly depend on the assigned weights. Contrarily, information may be lost when insufficient summary statistics are employed. Additionally, the weights ratio could be easily bounded to limit the impact of the regression model, and ensure convergence in Theorem 5.1. Here, we did not apply a ratio bound, however we set weights numerically close to zero to a tenth of the minimum non-zero weight, to ensure that all coordinates contribute to the distance value.

6.3.3 Optimal summary statistics to recover distribution features

A problem with regression models of parameters on model simulations is that such a function need not exist. For example, consider a model with $y \sim \mathcal{N}(\theta^2, 0.1^2)$, with prior $\theta \sim U[-1, 1]$, and observed data $y_{\text{obs}} = 0.7$. The problem is symmetric in θ, such that the posterior mean is given as $\mathbb{E}_{\pi(\theta|y)}[\theta] = 0$. Using the posterior mean as summary statistic as in Blum et al. [2013], Fearnhead and Prangle [2012] would clearly recover the correct mean of the true posterior, however fail to adequately describe the, for $y_{\text{obs}} \neq 0$ bimodal, posterior shape, such as its variance or higher-order moments, which is usually of interest in Bayesian inference. Similarly, as the inverse mapping $y \mapsto \theta$ does not exist globally, given representative samples $\{\theta_i, y_i\}_{i \leq N} \sim \pi(y|\theta)\pi(\theta)$, a regression model $s : y \mapsto \theta$ cannot extract a meaningful relationship.

One approach of addressing the problem of lack of invertibility is to localize the regression

model to a domain on which it is invertible, e.g. to $\theta > 0$ in the above example. However, data features learned on a subset of samples need not be informative elsewhere. Further, we empirically found approaches to identify clusters in parameter-space, conceptually allowing to identify multi-modalities if the data are sufficiently informative, e.g. based on Gaussian mixture models with model selection over the number of components via AIC [Schwarz, 1978], to be not robust enough, thus did not pursue this approach further.

An alternative is to, instead of only learning a mapping $s : y \mapsto \theta$ which may not exist, also consider transformations $\lambda(\theta)$ of the parameters, e.g. higher-order moments $s : y \mapsto \lambda(\theta) = (\theta^1, \dots, \theta^k)$, which may be better described as functions of the data, or identifiable in the first place. For the above example, it suffices to consider θ^2 as regression target, giving a linear mapping and breaking the symmetry. While the use of parameter transformations as regression model targets is heuristically reasonable, their use can be justified by considerations of the corresponding posterior approximations. As pointed out by Jiang et al. [2017], employing as ABC summary statistics posterior expectations of transformations of the parameters, $s(y) = \mathbb{E}_{\pi(\theta|y)}[\lambda(\theta)]$, allows to recover the corresponding posterior expectations for $\varepsilon \to 0$ in the population of accepted particles. This is made explicit by the following

Theorem 6.1. *Denote the joint distribution of parameters and data $\Theta, Y \sim \pi(\theta, y)$, with prior marginal $\pi(\theta) = \int \pi(\theta, y)\, dy$, likelihood $\pi(y|\theta) = \pi(\theta, y)/\pi(\theta)$, and resulting posterior $\pi(\theta|y) = \pi(\theta, y)/\pi(y) = \pi(y|\theta)\pi(\theta)/\pi(y)$. Given a parameter transformation $\lambda : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_\lambda}$ such that $\mathbb{E}_{\pi(\theta)}[|\lambda(\theta)|] < \infty$, define summary statistics as the conditional expectation*

$$s(y) := \mathbb{E}[\lambda(\Theta)|Y = y] = \int \lambda(\theta)\pi(\theta|y)\, d\theta\,.$$

Given observed data y_{obs}, acceptance threshold ε, and assuming the distance metric $d(s(y), s(y_{obs})) = \|s(y) - s(y_{obs})\|$ is norm-induced as in (6.2), denote the ABC posterior distribution

$$\pi_{\mathrm{ABC},\varepsilon}(\theta|s(y_{obs})) \propto \int I[\|s(y) - s(y_{obs})\| \leq \varepsilon]\pi(y|\theta)\, dy \cdot \pi(\theta).$$

Then, it holds

$$\left\|\mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon}}[\lambda(\Theta)|s(y_{obs})] - s(y_{obs})\right\| \leq \varepsilon, \tag{6.4}$$

and therefore

$$\lim_{\varepsilon \to 0} \mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon}}[\lambda(\Theta)|s(y_{obs})] = \mathbb{E}[\lambda(\Theta)|Y = y_{obs}]. \tag{6.5}$$

Proof. Based on Fearnhead and Prangle [2012] and Jiang et al. [2017], a simple extension of the argumentation in the latter. Note that $s(y)$ is almost surely finite due to $\mathbb{E}[|\lambda(\theta)|] < \infty$ and Fubini's Theorem. As for the induced σ-algebras holds $\sigma(s(Y)) \subset \sigma(Y)$, $s(Y)$ is also a version of the conditional expectation $\mathbb{E}[\lambda(\Theta)|s(Y)]$, since

$$s(Y) = \mathbb{E}[s(Y)|s(Y)] = \mathbb{E}[\mathbb{E}[\lambda(\Theta)|Y]|s(Y)] = \mathbb{E}[\Theta|s(Y)]$$

by, respectively, measurability, definition, and tower property. Thus, denoting the acceptance region

$$A = \{\|s(Y) - s(y_{\mathrm{obs}})\| \leq \varepsilon\} \in \sigma(s(Y)),$$

with $\mathbb{E}[\lambda(\Theta)|A] = \mathbb{E}[\lambda(\Theta)\mathbb{1}_A]/\mathbb{E}[\mathbb{1}_A]$, we have

$$\mathbb{E}_{\mathrm{ABC},\varepsilon}[\lambda(\Theta)|s(y_{\mathrm{obs}})] = \mathbb{E}[\lambda(\Theta)|A] = \mathbb{E}[s(Y)|A],$$

such that by Jensen's inequality, given convexity of the norm,

$$\left\|\mathbb{E}_{\pi_{\mathrm{ABC},\varepsilon}}[\lambda(\Theta)|y_{\mathrm{obs}}] - s(y_{\mathrm{obs}})\right\| = \|\mathbb{E}[s(Y) - s(y_{\mathrm{obs}})|A]\| \leq \mathbb{E}[\|s(Y) - s(y_{\mathrm{obs}})\| \,|A] \leq \varepsilon.$$

(6.5) then follows directly from (6.4) by definition of $s(y_{\mathrm{obs}})$. □

Therefore, e.g. for $\lambda(\theta) = (\theta^1, \ldots, \theta^k)$, the corresponding first k moments of the true posterior distribution are recovered by an ABC analysis in the approximate limit $\varepsilon \to 0$, when taking the posterior expectation $s(y) = \mathbb{E}[\lambda(\Theta)|Y = y]$ as summary statistic. In the limit $k \to \infty$ for $\varepsilon \to 0$ and e.g. assuming the moment-generating functions exist, the approximate posterior would thus converge to the true posterior.

Obviously, the conditional posterior expectations $\mathbb{E}[\lambda(\Theta)|Y]$ are hardly available for summary statistics definition. However, we may interpret the above regression models as approximations thereof. Using higher-order moments $\lambda(\theta) = (\theta^1, \ldots, \theta^k)$ as regression targets and given a sufficiently accurate description of the underlying expectation by the regression model, we may thus expect to approximately recover the corresponding posterior moments, or more heuristically, to break symmetry and obtain a meaningful regression model $y \mapsto \lambda(\theta)$. Thus, we propose to use $\lambda(\theta)$ as regression model targets, both for summary statistics construction (Section 6.3.1), and to define sensitivity weights (Section 6.3.2).

6.3.4 Implementation

We implemented all methods presented in this chapter in pyABC, in an extensible, modular manner. In particular, we interfaced scikit-learn [Pedregosa et al., 2011], providing regression models including in particular LR, LASSO, NN, and GP. The code underlying the study in this chapter can be found on GitHub at `https://github.com/yannikschaelte/study_abc_slad`, a snapshot of code and data is on Zenodo at `http://doi.org/10.5281/zenodo.5522919`. All simulations were performed on the GCS Supercomputer JUWELS at Jülich Supercomputing Center (JSC), using up to 384 cores in parallel.

6.4 Application to test problems

6.4.1 Considered distances and summary statistics

As ABC distance metric to compare model outputs or summary statistics, we considered, given its robust performance in Chapter 5, an L1 norm, with MAD (5.3) based weights when employing scale-normalization ("Ada.", using the same notation as in Chapter 5), as well as occasionally outlier-correcting PDMAD (5.4). As acceptance criterion in generation t, we only considered d_t, but not previous distances. This was for ease of implementation and as for an L1 norm

no substantial differences were observed in Chapter 5, however note that considering previous acceptance criteria, giving nested acceptance regions, may improve performance further on some problems.

As regressor models, we considered linear regression (LR), and, as a slightly more complex model, a neural network (NN), using implementations in scikit-learn [Pedregosa et al., 2011]. Unless stated otherwise, we trained the regression model as soon as 40% of the total simulation budget had been reached. For comparison, we also considered training the regression model before the initial generation, $t_{\text{train}} = 1$, based on samples from the prior ("Init"). NN models were considered with a single hidden layer of dimension $[(n_y + n_\theta)/2]$ with ReLU activation function [Nair and Hinton, 2010], using ADAM stochastic gradient descent for optimization [Kingma and Ba, 2015], and early stopping to avoid overfitting, using a validation set of 10% of the training data. Both regression models used here were computationally comparably efficient compared to the full ABC-SMC analysis, with run-times on the order of milliseconds (LR) and seconds (NN).

When employing parameter augmentation (Section 6.3.3), we used the first four moments, $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ ("P4"). We considered both regression for summary statistics construction ("Stat", Section 6.3.1) and sensitivity weight construction ("Sensi", Section 6.3.2).

While data augmentation, e.g. including higher-order moments $y^2, \ldots, y^4$, as regression model inputs, can further improve performance, and was employed on selected problems in e.g. Blum et al. [2013], Fearnhead and Prangle [2012], here we did not do so but rather considered the regression model inputs as pre-defined, as their necessity is problem specific, and complex methods such as NN may be assumed capable of inferring the corresponding transformations themselves (unlike for regression target transformations).

For example, L1+Ada.+MAD+StatNN denotes an analysis using consistently, i.e. both before and after regression model training, an adaptive distance with MAD scale-normalizing weights, and using a neural network to construct summary statistics as soon as 40% of the total simulation budget are reached, using $\lambda(\theta) = \theta$ as regression targets, L1+Ada.+MAD+SensiLR+P4 uses an adaptive distance function with MAD scale-normalizing weights, and uses an LR model to define further sensitivity weights, using $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ as regression targets, and L1+LR uses an LR model for summary statistics construction, but does not normalize distance weights.

6.4.2 Performance on dedicated demonstration problem

To illustrate the different problems discussed and addressed in this chapter, we constructed a comparably simple test problem combining different features that can be problematic for established methods. The model possesses four parameters, $\theta = (\theta_1, \ldots, \theta_4)$, with both informative and uninformative model outputs:

- $y_1 \sim \mathcal{N}(\theta_1, 0.1^2)$ is informative of θ_1, with a relatively wide corresponding prior $\theta_1 \sim U[-7, 7]$,
- $y_2 \sim \mathcal{N}(\theta_2, 100^2)$ is informative of θ_2, with corresponding prior $\theta_2 \sim U[-700, 700]$,
- $y_3 \sim \mathcal{N}(\theta_3, 4 \cdot 100^2)^{\otimes 4}$ is a four-dimensional vector informative of θ_3, with corresponding prior $\theta_3 \sim U[-700, 700]$,

- $y_4 \sim \mathcal{N}(\theta_4^2, 0.1^2)$ is informative of θ_4, with corresponding symmetric prior $\theta_4 \sim U[-1, 1]$, however is quadratic in the parameter, resulting in a bimodal posterior distribution for $y_{\text{obs},4} \neq 0$,
- $y_5 \sim \mathcal{N}(0, 10)^{\otimes 10}$ is an uninformative 10-dimensional vector.

The model dynamics are purposely simple, such that the inverse mapping can be captured comparably easily by regression models. The problem possesses the following potentially problematic features:

- A substantial part of the data, y_5, is uninformative, such that approaches not accounting for informativeness of data may converge slower.
- Both data and parameters are on different scales, such that approaches comparing data, or, via regression-based summary statistics, parameters, without normalization may be biased towards large-scale variables. Further, e.g. the prior of θ_1 is relatively wide, such that pre-calibrated weighting is sub-optimal, as discussed in Prangle [2017].
- y_4 is quadratic in θ_4 and symmetric over the prior, such that the inverse first-order regression approaches as in Fearnhead and Prangle [2012] cannot capture a meaningful relationship.
- While the distributions of y_3 and y_4 are such that the posterior distributions of θ_3, θ_4 are identical, in solely scale-normalized approaches such as Prangle [2017] the impact of y_4 on the distance value is roughly four times as high as that of y_3, leading to potentially uneven convergence.

We studied the demonstration problem with synthetic data $y_{\text{obs},1}, y_{\text{obs},2}, y_{\text{obs},3}, y_{\text{obs},5} \equiv 0$, $y_{\text{obs},4} = 0.7$, using a population size of $N = 4e3$ with a total budget per run of $1e6$ simulations. Obtained posterior approximations using selected distance functions and summary statistics are shown in Figure 6.1.

Solely scale-normalized distances on the full model outputs without informativeness assessment converge slowly

The MAD scale-normalized adaptive distance L1+Ada.+MAD captured all posterior modes and shapes, in particular the bimodality of θ_4, correctly, however with large variances, because the uninformative model outputs y_5 were considered on the same scale as the informative model outputs $y_1, \ldots, y_4$, leading to an overall slower convergence (Figure 6.1 bottom). Further, while the true marginal posteriors of θ_2 and θ_3 coincide, L1+Ada.+MAD assigned a substantially wider variance to θ_2, as only a single model output, y_2, is informative of it, while four are of θ_3, all on the same normalized scale.

Non-scale-normalized distances converge unevenly

The analyses L1+StatLR and L1+StatNN not employing scale normalization described the marginal posterior distributions of θ_2 and θ_3, which are on the same scale, well, however had a

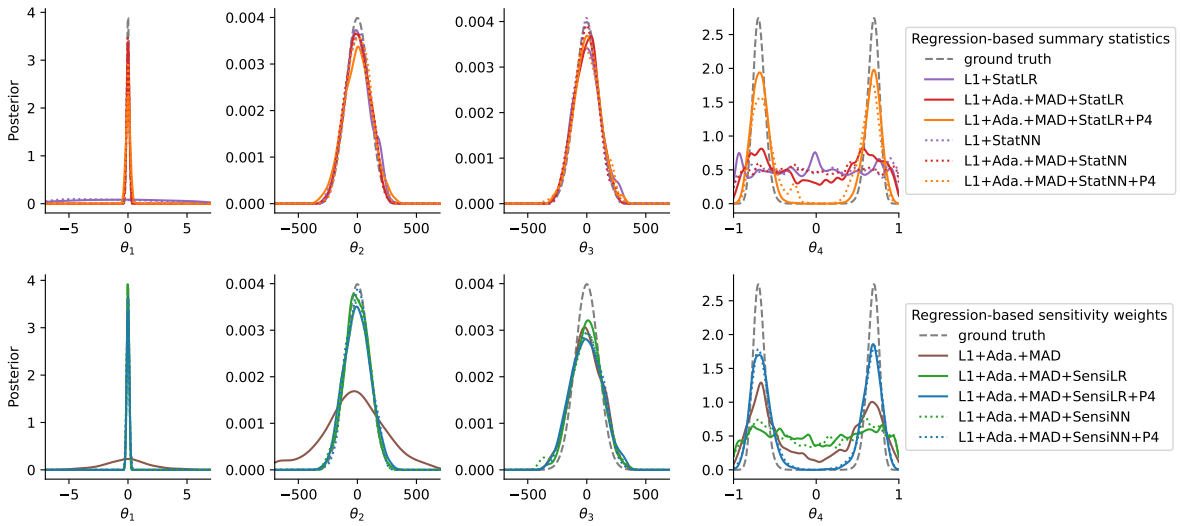

Figure 6.1: ABC marginal posterior approximations obtained for analyses using regression-based summary statistics (top, "Stat") and sensitivity weights (bottom, "Sensi") on the demonstration problem, using an underlying L1 norm, uniformly weighted, and MAD scale-normalized distance weights ("Ada.+MAD"), using a linear regression model ("LR") or a neural network model ("NN"), and optionally transformed parameters $\theta^1, \ldots, \theta^4$ as regression targets ("P4").

substantially wider variance for θ_1 (Figure 6.1 top). This is because in these analyses, regression-based parameter estimates were used as summary statistics, but θ_1 varies on a smaller scale. In comparison, all analyses employing scale normalization of model outputs or regression-based summary statistics described θ_1, θ_2, and θ_3 roughly or almost similarly well, with the exception of L1+Ada.+MAD, as outlined above.

Regression models not accounting for non-identifiability cannot approximate full posterior

All analyses that employed regression models to either construct summary statistics or define sensitivity weights, but only used the non-augmented parameters $\lambda(\theta) = \theta$ as regression targets, failed to describe the bimodal posterior distribution of θ_4 well. This is because a global mapping $y_4 \mapsto \theta_4$ does not exist, such that regression models cannot detect a meaningful relationship, corresponding to $\mathbb{E}_{\pi(\theta_4|y_4)}[\theta_4] = 0$ for all y_4. While this posterior mean is clearly captured, no further distribution information is. In comparison, analyses considering higher-order moments $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ as regression targets ("P4") captured the bimodality, as for this problem a linear mapping $\theta_4^2 \sim y_4$ exists, or a quadratic one $\theta_4^4 \sim y_4^2$.

Novel approaches fit all parameters well

The analyses L1+Ada.+MAD+{Stat{LR/NN}/Sensi{LR/NN}}+P4 combining all methods introduced in this chapter, i.e. scale normalization, regression models for summary statistics construction or sensitivity weight definition, and regression target augmentation, provided the overall best description of all posterior marginals, with roughly homogeneously small variances. For this

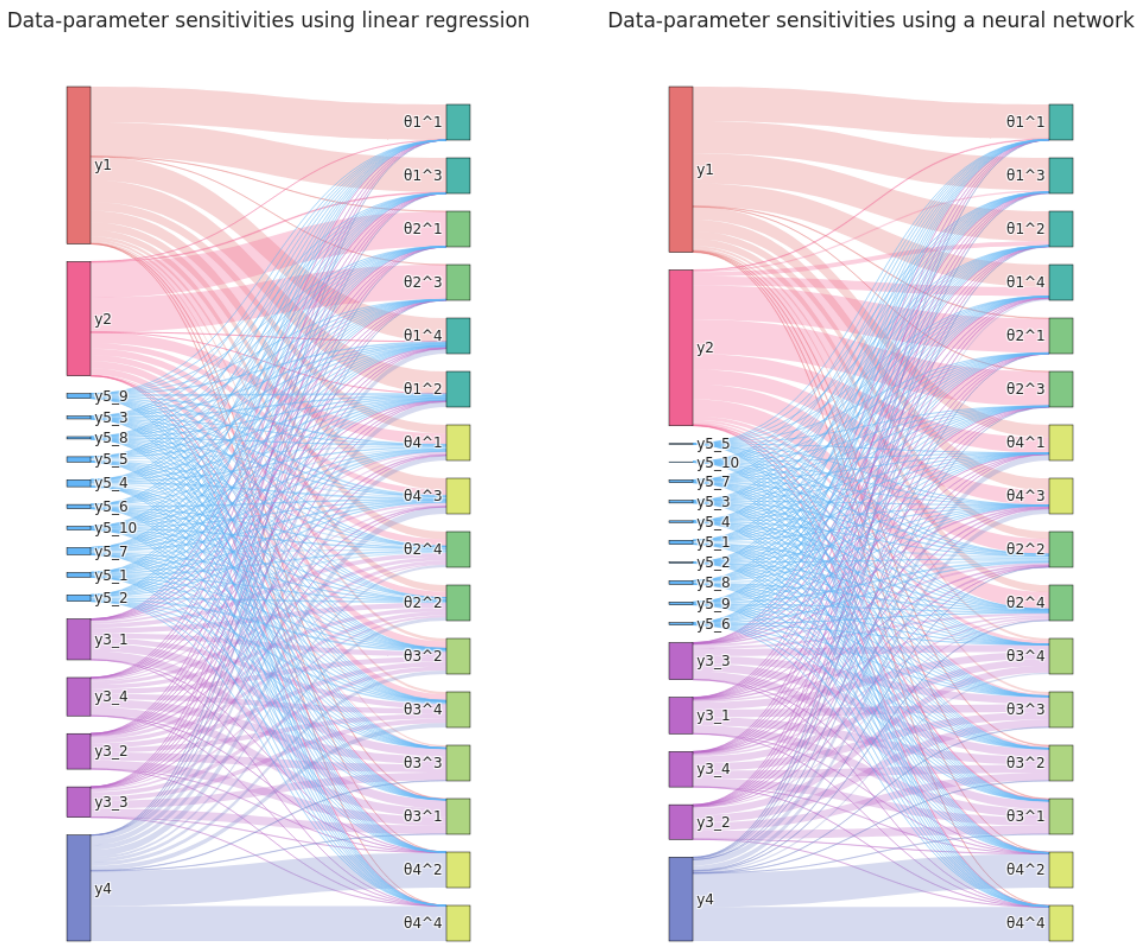

Figure 6.2: Exemplary illustration of normalized absolute data-parameter sensitivities calculated for the demonstration problem, using a linear regression model (left) and a neural network model (right), for all regression targets $\theta^1, \ldots, \theta^4$ with $\theta = (\theta_1, \ldots, \theta_4)$ (respectively on the right), with respect to all data coordinates y (respectively on the left). The absolute sensitivity matrix $|S|$ (6.1) was normalized per parameter, as in (6.2). The widths of lines connecting data and parameters, and corresponding endpoints, are proportional to their respective absolute values. In particular, the heights of the respective left endpoints are proportional to the sensitivity weights q_{i_y} (6.2). Data types, e.g. $y_{5,1}, \ldots, y_{5,10}$, and parameters with their exponents, e.g. $\theta_1^1, \ldots, \theta_1^4$, are grouped by colors.

model and considered inference scenario, no advantages of NN could be observed over LR, which already provided good posterior approximations.

For L1+Ada.+MAD+Sensi{LR/NN}, parameter θ_3 was slightly less well inferred compared to L1+Ada.+MAD+Stat{LR/NN}, inverting the situation for L1+Ada.MAD, to a smaller extent. This can be attributed to the fact that the latter approaches employ a one-dimensional interpolation of the four-dimensional vector y_3, and thus e.g. an approximation of the statistic $\frac{1}{4}\sum_{i=1}^{4} y_{3,i}$ which is sufficient of θ_3, while approaches as the former ones, not employing a lower-dimensional transformation, but only assigning (roughly equal) weights, are more subject to random noise in the single simulations. This illustrates that when low-dimensional sufficient statistics firstly

Table 6.1: Properties of test models used in Chapter 6: Identifier, short description, number of parameters n_θ and data points n_y, population size N and maximum number of model simulation after which an analysis was terminated.

ID	Description	n_θ	n_y	N	Max. sim.
T1	Conversion reaction ODE model	2	10	1000	250000
T2	One informative and one uninformative variable	1	2	1000	25000
T3	g-and-k distribution order statistics, small	4	7	1000	250000
T4	Lotka-Volterra Markov jump process model, small	3	32	500	125000
T5	g-and-k distribution order statistics, large	4	100	1000	250000
T6	Lotka-Volterra Markov jump process model, large	3	200	500	125000

exist and can secondly be easily and accurately captured by regression models, employing explicit dimension reduction can be superior to mere re-weighting.

Sensitivity weights permit further insights

The normalized absolute sensitivities (6.2) of parameters, or transformations thereof, with respect to model outputs obtained using LR and NN models are visualized for the demonstration problem for exemplary analyses in Figure 6.2. Overall, both regression models captured correlations of model outputs and parameters well, and assigned large, albeit not completely homogeneous, sensitivity weights to $y_1, \ldots, y_4$ and lower ones to y_5, with e.g. the weights for y_1 and y_2 roughly equal, or at least comparable, to the sum of the weights for $y_{3,1}, \ldots, y_{3,4}$. The description provided by the NN model was overall slightly better, assigning lower weights to y_5, and capturing e.g. the mappings $\theta_1^2 \sim y_1$ and $\theta_1^4 \sim y_1$ slightly better, which the linear model cannot accurately emulate without data augmentation (e.g. using also $y_1^2, \ldots$ as regression model inputs). As seen above, nonetheless the linear model sufficed for this problem to provide a good posterior approximation. Sensitivities of θ_4^1 and θ_4^3 were, as expected, comparably small with respect to all variables.

The weight assigned to y_4 was only roughly half the ones assigned to y_1, y_2 and y_3, because θ_4^1 and θ_4^3 could not be accurately described. Correspondingly, θ_4 had a comparably wider variance under sensitivity-weighted analyses. This could be improved by not employing parameter-wise normalization in (6.2), which however makes the analyses less robust to regression model misspecification (not shown here).

An analysis as performed here may in general serve to evaluate the plausibility of the used regression model, and allow to obtain insights into underlying relationships of parameters and model outputs, or elicit e.g. uninformative model outputs.

6.4.3 Evaluation on established test problems

To evaluate robustness and general performance of the proposed methods, we next considered six test problems T1-6, not tailored to the challenges discussed in Section 6.4.2. Core model properties as well as employed ABC-SMC population sizes N and total budgets of numbers of simulations are given in Table 6.1.

T1, T3, and T4 are problems M3, M4, and M5 from Chapter 5, respectively an ODE model of a conversion reaction, and, based on application examples in Prangle [2017], g-and-k distribution samples, and a Markov jump process model of a Lotka-Volterra model. T2 consists of two observables, thereof $y_1 \sim \mathcal{N}(\theta, 0.1^2)$ informative and $y_2 \sim \mathcal{N}(0, 1^2)$ uninformative, with wide prior $\theta \sim \mathcal{N}(0, 100^2)$, also taken from Prangle [2017]. T5 and T6 are variations of T3 and T4 with higher-dimensional data, based on application examples in Fearnhead and Prangle [2012]. T5 employs 100 order statistics out of 10,000 samples from a g-and-k distribution, with same $U[0, 10]$ priors on the four parameters A, B, g, k, considering ground truth values $(A, B, g, k) = (3, 1, 2, 0.5)$. T6 employs noise-free observations of predators and prey at 200 evenly spaced time-points over the time interval $[0, 20]$, estimating the three model parameters on linear scale, considering narrower independent priors $\theta_1 \sim U[0, 2]$, $\theta_2 \sim U[0, 0.1]$, $\theta_3 \sim U[0, 1]$, and ground truth values $(\theta_1, \theta_2, \theta_3) = (0.5, 0.0025, 0.3)$.

We considered LR and NN regression models for both summary statistics construction and sensitivity weight calculation, with and without scale normalization, and employing the regression model either in a generation after 40% of the total simulation budget, or based on a sample of size N from the prior before the initial generation ("Init"). As reference method, we considered scale-normalized L1+Ada.+MAD, which had performed well on problems T1+T3+T4 (Section 5).

We ran 10 repetitions of each inference setting on problems T1-6, using different model simulations under ground truth parameters as observed data. As measure of quality of fit, we report root mean square errors (RMSE) of the weighted posterior samples from the last ABC-SMC generation, with respect to ground truth parameters (note all problem considered here are unimodal). The results are visualized in Figure 6.3.

Delay of regression model training advantageous on complex models

For the considered LR and NN models, regression model training on prior samples ("Init") gave for most problems substantially worse results than when trained after multiple generations, after 40% of the total simulation budget. One reason for this may be that only N prior samples were used for training, compared to potentially more samples, including rejected ones, in later generations. However, also when using only N training samples in the later-trained approach (not shown here), results were better than in analyses training the regression model based on the prior. Another reason may be that after multiple generations the bulk of samples is restricted to a high-density posterior region, in which a simpler model, even a linear one, suffices to describe the underlying relationships sufficiently well. Thus, this justifies empirically the approach by Fearnhead and Prangle [2012] of using a pilot run to constrain parameters, and who found linear models to be sufficient. Jiang et al. [2017], who based their regression model on the prior, used firstly more complex NN models with three hidden layers of dimension 100, and secondly up to a million training samples, giving training times on the order of minutes, intending to capture the posterior mean accurately on the whole domain.

An exception was problem T2, on which an earlier use of regression models consistently improved performance. This can be explained by the clear linear relationships of parameters and data in this model, such that accurate regression models can be easily learned at any point in the analysis and thereafter be beneficial. Further, by design the scale-normalized distances assign a small

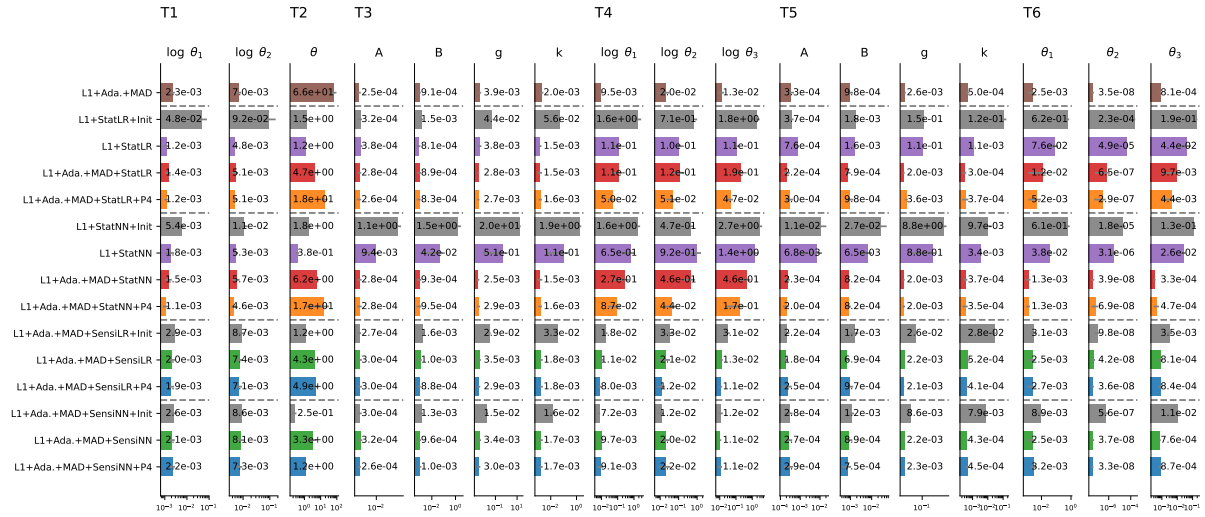

Figure 6.3: Median RMSE for the parameters of models T1-6 (columns) obtained for 15 inference methods (rows), using an L1 distance, either uniformly weighted if unspecified, or with adaptive MAD scale normalization ("Ada.+MAD"). As regression models we considered linear regression ("LR") and neural networks ("NN"), both to define summary statistics ("Stat") and sensitivity weights ("Sensi"). Some inference settings further used parameter transformations $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ as regression targets ("P4"). In some settings, the regression method was trained before the initial generation based on the prior ("Init"), otherwise after 40% of the total simulation budget if unspecified. The first row contains solely scale-normalized L1+Ada.+MAD as a reference method, followed by two blocks of three rows using regression methods for summary statistics construction, using firstly LR and secondly NN, and then by two blocks of four using sensitivity weights, using firstly LR and secondly NN. Reported values are median RMSE values over 10 replicates, with grey lines indicating MAD values. Note that here we report median RMSE values in order to depict typical behavior of the algorithms. However it should be noted that some settings yielded occasionally substantially larger large RMSE values, especially for "Stat", arguably due to regression model misspecification (not shown).

weight to the informative statistic y_1 in early generations due to the wide prior, such that the analysis progresses slowly until the uninformativeness of y_2 is recognized. This indicates that criteria regarding at what point to learn and use a regression model for a given problem may be of interest.

Scale-normalized distances improve performance for regression-based summary statistics

As the comparison of L1+Stat{LR/NN} and L1+Ada.+MAD+Stat{LR/NN} shows, for many problems the use of scale-normalized distances improved performance, particularly for T5 and T6, while it was roughly similar for T1. An exception was again T2, where in fact a plain L1 distance would be preferable over L1+Ada.+MAD at least in the first generations, for reasons outlined above. However, the behavior in such cases is completely dependent on the relative scales of informative and uninformative data points, such that scale normalization is generally advisable, additionally taking into account informativeness, as shown here, unless a manual quantification of the impact data points should have is possible and preferable.

Sensitivity-weighted distances perform highly robustly

The analyses L1+Ada.+MAD+Sensi{LR/NN}(+P4) using regression models to define sensitivity weights performed reliably, with RMSE values generally not far higher, but in some cases consistently lower, than ones obtained by L1+Ada.+MAD. This indicates that, while the sensitivity weighting could in those cases not improve performance, as sole scale normalization performed efficiently enough already, the approach is highly robust. In some cases, specifically T2, which had one clearly uninformative statistic, and arguably T5, which is a high-dimensional collection of order statistics, did the sensitivity weighting improve performance. In other cases, specifically T1, T3, T4, and T6, RMSE values for some parameters decreased, but slightly increased for other parameters, indicating that the weighting scheme re-prioritized data points, but there were no overall uninformative ones to identify.

Regression-based summary statistics can be superior but also more dependent on regression model accuracy

In various cases, e.g. when trained in the initial generation, and consistently for T4, as well as using LR on T6, did regression-based summary statistics perform inferior to both L1+Ada.+MAD and regression-based sensitivity weights. Arguably, in those cases the regression model was not accurate enough to permit its outputs to serve as low-dimensional summaries, while the sensitivity-based approach was more robust. However, in some cases, specifically for T1, and two parameters of T6 using a NN, were RMSE values obtained using regression-based summary statistics smaller than both with L1+Ada.+MAD and regression-based sensitivity weights. In both cases, the data consist of repeated measurements along trajectories of species, with reaction rates governed by the parameters. This indicates that if the lower-dimensional summary statistics representation is accurate and informative of the parameters, then its use can be beneficial.

No clear preference for regression models or parameter augmentation on considered problems

Overall, we found, for both regression based summary statistics and sensitivity weights, no clear indication of preference for LR or NN, with LR performing more robustly in many cases, but NN clearly preferable in some. A regression model selection scheme reliable in the considered setting of relatively small sample sizes might thus be beneficial. Further, the use of augmented parameters as regression targets did not substantially worsen, but did also not notably improve performance for any test problem (note that none of them is multi-modal), however performed inferior e.g. on T2, which has a clear linear mapping, such that the consideration of higher-order moments may have complicated the inference. This indicates that the use of augmented parameters as regression targets is robust, but if further information is available, a restriction to e.g. first or second order may be beneficial.

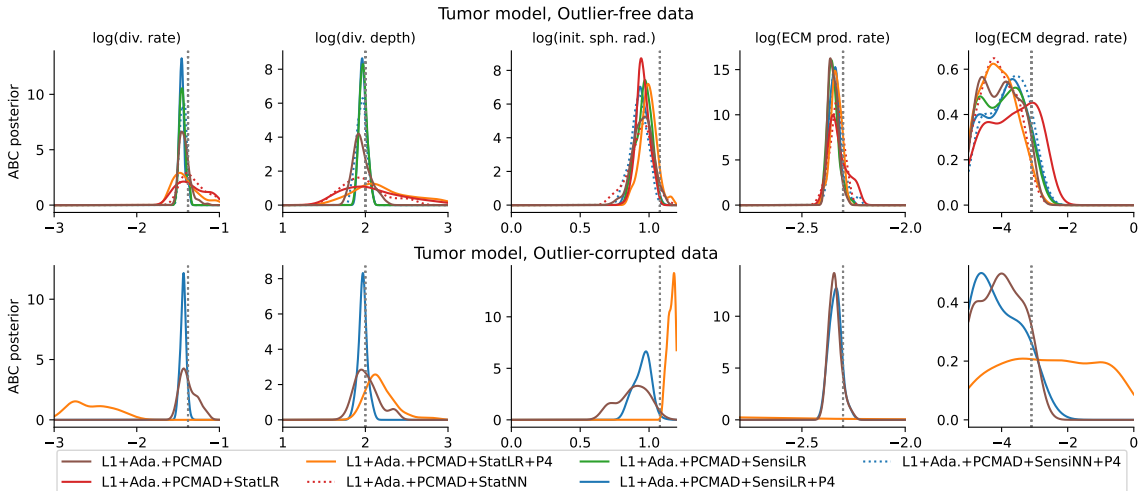

Figure 6.4: Posterior marginals for the 5 out of the 7 model parameters of the tumor problem showing interesting dynamics. Top: Without outliers. Bottom: With outliers. The parameter values used to simulate the observed data are indicated by grey dotted lines. With the exception of the ECM production rate, which is zoomed in for visibility, the plot boundaries coincide with the employed uniform prior range.

6.4.4 Performance on application example

Next, we considered the agent-based model of tumor spheroid growth (model M6 from Section 5), constituting a more realistic application example, and considering both outlier-free and outlier-corrupted data. We employed the same simulated data as in the previous chapter, using a population size of $N = 500$, with a computational budget of 150,000 simulations per analysis.

Given its computational complexity, on this problem we only considered selected approaches: Using as reference method L1+Ada.+PCMAD from the previous chapter, we employed, given the robust performance of linear models before, L1+Ada.+PCMAD+SensiLR(+P4) employing sensitivity weights, L1+Ada.+PCMAD+StatLR(+P4) employing summary statistics, both with and without augmented parameters $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ as regression targets, and further L1+Ada.+PCMAD+StatNN and L1+Ada.+PCMAD+SensiNN+P4 using neural networks. Note that here we used PCMAD (5.4) instead of MAD (5.3), in order to be able to identify outliers.

Sensitivity weights improve performance and identify informative model outputs

The use of regression models for sensitivity weight definition improved performance on the tumor model with outlier-free data over L1+Ada.+PCMAD, giving substantially lower variances for the division rate and depth parameters while giving similar results on the remaining parameters (Figure 6.4 top), and accepted simulations closely matching the observed data (Figure 6.5 top, simulations). No differences could be observed between using only the parameters themselves or higher-order moments ("P4") as regression targets. Overall, here the analyses using linear regression model gave slightly lower variances than the one based on a neural network.

On this problem, regression-based summary statistics worsened performance considerably compared to L1+Ada.+PCMAD in the considered settings, indicating that the employed regression

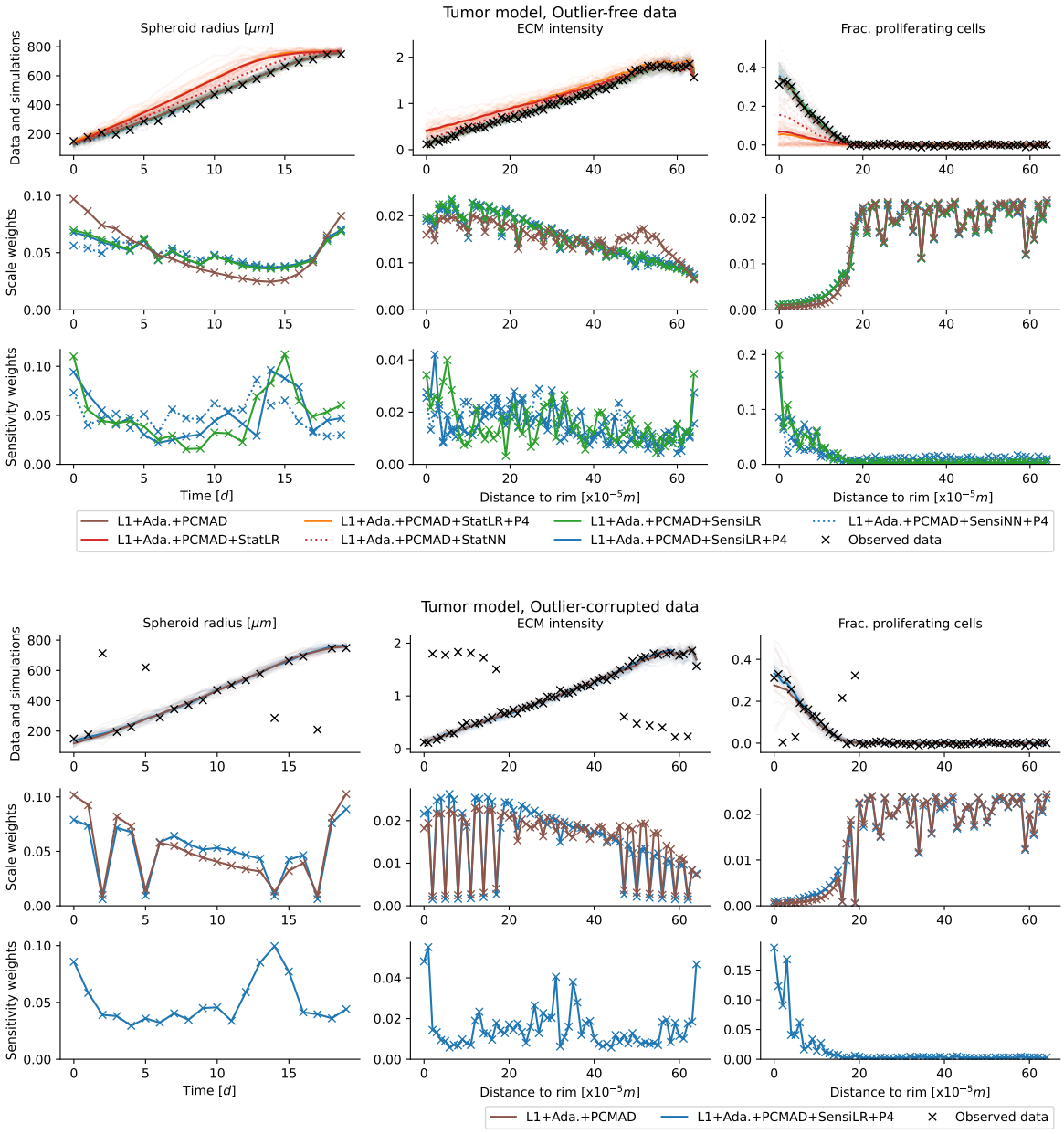

Figure 6.5: Fits as well as scale and sensitivity weights for multiple approaches on the tumor problem on outlier-free (top) and outlier-corrupted (bottom) data. The respective upper rows show the observed data (black), and, for each approach, 20 accepted simulated data sets (light lines) as well as the sample means (darker lines) from the last ABC-SMC generation. Note that these are accepted simulations, not predictions. The respective middle rows show the corresponding scale weights assigned to each data point in the last generation, normalized to sum 1, and the bottom rows sensitivity weights, respectively only for distances employing scale and sensitivity weights, and operating on the full dataset.

models did not provide a sufficiently informative low-dimensional representation, as obtained posterior marginals had wide variances (L1+Ada.+PCMAD+StatLR(+P4), Figure 6.4 top), and accepted simulations did visibly not match the observed data (Figure 6.5 top, simulations).

The overall structure of sensitivity weights assigned via linear models with and without param-

eter augmentation, as well as a neural network model, was roughly consistent across multiple runs (Figure 6.5 top, bottom row). Low weights were assigned to the fraction of proliferating cells at large distances to the rim, indicating these, as suspected, as uninformative, and counteracting the large weights resulting from scale normalization (Figure 6.5 top, middle row). In contrast, high weights were assigned to the fraction of proliferating cells close to the rim, which indeed vary considerable for other parameters, as apparent from the final simulations for e.g. L1+Ada.+PCMAD+StatLR. Further, high weights were assigned consistently to the spheroid radius after roughly two weeks, arguably when growth began to slow down. Moreover, as a tendency high weights were assigned to the ECM density close to the rim, and the farthest point, at which point, depending on the parameterization, the ECM density begins to quickly drop. In conclusion, while the obtained sensitivity weights do exhibit some noise between adjacent points, arguably due to the comparably low training sample size, overall patterns can be observed and appear to be reasonable.

Robust performance on outlier-corrupted data

The use of sensitivity weights improved performance also on outlier-corrupted data, similar to the outlier-free data (Figure 6.4 bottom). Here, we only considered L1+Ada.+PCMAD+SensiLR+P4. Accepted simulations in the final generation matched the observed data more closely than those for L1+Ada.+PCMAD (Figure 6.5 bottom, simulations). The PCMAD scheme correctly detected and assigned low weights to outliers, independent of the sensitivity weights derived from the regression model (Figure 6.5 bottom, weights), such that the combination of both methods allowed to simultaneously account for outliers and informativeness of data. The overall shapes of sensitivity weights were consistent across the outlier-free and outlier-corrupted problems.

6.5 Discussion

ABC relies on an efficient comparison of informative data features via summary statistics and distance metrics. In this chapter, we discussed problems of scale-normalizing adaptive distance metrics, as well as regression-based summary statistics. More specifically, we discussed problems arising from partly uninformative data on adaptively scale-normalized distances, from parameters on different scales on regression-based summary statistics, and from parameter non-identifiability on regression model adequacy. To tackle these problems, we presented multiple solutions. Firstly, we suggested employing scale-normalizing distance metrics on top of regression-based summary statistics, to level the impact of different parameters. Secondly, and as an alternative to the first solution, we presented novel sensitivity weights derived from regression models, measuring the informativeness of data on parameters. Thirdly, we presented augmented regression targets, using e.g. higher-order moments, as a solution to parameter non-identifiability.

On a simple test problem, we demonstrated substantial improvements of the novel methods and their combination over established approaches. For the sensitivity-weighted distances, we observed robust performance on various test problems, and in particular on a systems biological application problem substantial benefits over approaches not accounting for data informativeness. Compared to regression-based summary statistics, they were more robust to potential regression model inaccuracy, and could thus be considered a practical alternative to dimension reduction

methods. The latter can however be beneficial when low-dimensional informative statistics exist and can be efficiently and accurately obtained.

Yet, there are numerous ways in which the presented methods can be improved. We would like to stress that, while we presented multiple novel approaches and overcame limitations of established methods, we do not claim our method to be out-of-the-box generally applicable.

A first and crucial aspect is the choice of regression model. While we observed simple linear models to be often sufficient, especially when trained on a high-density region, in some cases more complex models were superior, allowing to capture underlying relationships more accurately, which are however at the same time likely more prone to overfitting. An investigation of more complex models, or alternative model types such as multi-layer neural networks or Gaussian processes, would be of interest. To pick among competing model candidates, a robust model selection scheme, with out-of-sample evaluation e.g. based on cross validation, might be useful.

As more complex models usually require larger training data sets, the integrated ABC-SMC scheme used in this chapter could be extended by an increased population size in generations used for regression model training, to obtain more robust models. While increasing the training set is straightforward to do in ABC, as it only requires continued sampling, there is a trade-off of number of (expensive) forward model simulations employed for model training, and the ABC inference itself, besides a trade-off of time spent in inverse regression model training or selection, which can be not unsubstantial for e.g. high-dimensional neural networks or Gaussian processes [Jiang et al., 2017], and again the ABC inference.

While in many cases we observed advantages of delaying regression model training to later generations and thus a smaller, high-density region in which simpler models may suffice to capture underlying relationships, for simple models we observed improvements of an early use of regression methods to identify informative statistics. Thus, criteria on good time points for regression model training would be of interest. Similar to adaptive distances, there could further be benefits of repeated regression model training in multiple or all generations, especially when using efficient, e.g. linear, models.

A further evaluation may be regarding the training sample generation. While Fearnhead and Prangle [2012] employed a restricted prior sample, here we used unweighted samples from a proposal distribution, itself an approximation to the posterior. It could be beneficial to account for this discrepancy by importance weighting, which some regression models allow, or to employ alternative methods of training sample generation.

Harrison and Baker [2020] introduced an alternative distance weighting scheme that maximizes a distance between samples from the prior and the posterior approximation. More specifically, given $\tilde{N} \geq N$ samples, an objective function is formulated as, given distance weights, the Hellinger distance between samples from the prior and samples corresponding to the N closest model simulations, according to the weighted distance. This objective is then maximized over the distance weights. It uses a different ABC-SMC scheme than in this chapter, generating a fixed number of samples, out of which the closest ones according to the distance with optimized weights are then accepted. While using a different notion of informativeness of model outputs, as maximizing the difference from the prior, this approach is similar to the one presented in this chapter, such that a direct comparison, in terms of robustness, information gain, and efficiency,

would be of interest.

We employed purposely generic methods, e.g. accounting for up to fourth-order moments, although in most cases a first- or second-order model would have been sufficient, and we employed fully standardized and automated regression model training. Domain knowledge and manual tuning may be beneficial for the definition of a more tailored and efficient analysis, similar to the semi-automatic pipeline by Fearnhead and Prangle [2012]. Yet, robustly increasing the degree of automation is valuable for ease of application.

In conclusion, we presented methods accounting for data scale and informativeness in the comparison of observed and simulated data in ABC, and demonstrated improvements over established approaches. All methods have been implemented in pyABC, facilitating their straightforward application. We anticipate that methods such as the ones presented, which automatically normalize and extract features of interest, can substantially improve performance of ABC methods on a wide range of applications problems.

Chapter 7

Discussion

7.1 Summary

Mechanistic mathematical models are important tools in systems biology to describe and understand biochemical systems at a systems level and unravel underlying mechanisms [Kitano, 2002a]. Commonly, such models depend on unknown parameters that need to be estimated by comparison of model simulations with experimentally observed data [Fröhlich et al., 2019, Tarantola, 2005]. While complex and large-scale models have the potential to enable a more holistic understanding of biochemical processes, their parameterization is challenging, requiring the development of novel methods. In this thesis, we have developed multiple such methods for specific challenges as posed in Section 1.2, in the context of ordinary differential equation (ODE) models, and approximate Bayesian computation (ABC).

In Chapter 3, we combined the concept of hierarchical optimization with adjoint sensitivity analysis, to efficiently identify optimal observable transformations and noise parameters, and calculate objective function gradients for high-dimensional ODE models. Further, we derived analytical formulas for optimal scaling factors, offsets and normal noise parameters, and provided a more general perspective on hierarchical optimization. On an ODE model with thousands of states and parameters, this approach showed orders of magnitude superior performance compared to established approaches, facilitating parameter optimization and integration of heterogeneous data sets.

In Chapter 4, turning to ABC, we demonstrated the pitfalls of not properly accounting for measurement noise, and developed a broadly applicable, efficient, adaptive approach to perform exact inference with ABC methods in the presence of measurement noise. On various test problems, we observed up to orders of magnitude speed-up compared to established approaches, facilitating exact inference for complex and high-dimensional problems for which this was formerly impossible, for a broad spectrum of model types and noise models, and estimating noise parameters alongside.

In Chapter 5, we tackled the problem of data outliers in ABC, demonstrating how established approaches can give erroneous results. We developed a broadly applicable distance metric that adapts to the problem structure by scale normalization, and is robust to outliers, or even allows

for active online detection. On various test problems, we demonstrated robust and efficient performance for outlier-free and outlier-corrupted data.

In Chapter 6, we developed a robust approach to weight data points by informativeness using inverse regression models, and demonstrated advantages of accounting for both scale and informativeness using summary statistics or weighting schemes, yielding highly efficient ABC inference approaches.

Taking a more general perspective, our contributions, and underlying motivations, can be summarized as providing accurate inference methods efficiently and robustly applicable to a wide range of problems: We firstly developed methods scaling to higher-dimensional or more complex models, accurately accounting for measurement noise, and considered more generic situations, thus allowing for the application to a wider range of problems. Secondly, we translated these methods to efficiently and robustly applicable algorithms, in particular with automatic choice of hyperparameters and robust performance on a wide range of problems types, thus avoiding the need of manual problem-specific tuning, and enabling integrated workflows. Thirdly, we provided open-source, documented and tested implementations of our algorithms in computational software packages, thus enabling their translation to practice, and use by practitioners not familiar with the specific details.

7.2 Outlook

We discussed specific results, implications and possible extensions of the single contributions presented in this thesis already in Sections 3.4, 4.5, 5.5, and 6.5, see there for details. Taking a more general perspective, ideas for further research included e.g. generalizations, e.g. of hierarchical optimization to constraints and more general problems, allowing to cover a wider field of application problems. Algorithm improvements could enhance scalability, robustness and degree of automation, e.g. of exact ABC inference by improved hyperparameter selection avoiding manual tuning, or via improved regression model training for ABC sensitivity-weighting or summary statistics construction, to facilitate integrated workflows. Combinations of different methods could provide further synergies, e.g. of hierarchical optimization and mini-batching. Further, methods presented here could be applied in different contexts. For example, adjoint hierarchical approaches as employed in Chapter 3 could similarly be used for inference on spatial partial [Li and Petzold, 2004], stochastic [Li et al., 2020], or neural [Chen et al., 2018] differential equation models, or a hierarchical problem decomposition could similarly be employed in likelihood-based sampling via marginalization of observable and noise parameters yielding a reduced sampling problem, or conceptually even in likelihood-free methods, e.g. for noise parameters as used in Chapter 4. Machine learning methods as employed in Chapter 6 could be generally beneficial to provide insights into informativeness or as efficient surrogate models [Baker et al., 2018, Prescott and Baker, 2021].

There are countless farther-reaching and complementary challenges. While there have been substantial improvements to the calibration of complex stochastic and large-scale ODE models, there are still substantial limitations. To develop and calibrate holistic models, large datasets (e.g. Barretina et al. [2012], Eduati et al. [2017], Li et al. [2017]) are necessary, may however not be sufficient. Complementary, information to constrain parameters or formulate priors, as

e.g. provided in BRENDA [Schomburg et al., 2002] or SABIO-RK [Wittig et al., 2012] can be used. Tools to automatically retrieve, evaluate, and map information from such public databases for a given annotated model would be beneficial. Further, while the hierarchical multi-start approach in Chapter 3 provided substantial improvements in optimizer performance, it is far from an efficient and comprehensive coverage of the high-dimensional search space with several thousand parameters considered. While an exhaustive search may be, at the moment, infeasible, coverage could be improved by the development of scalable heuristic parallelized combined global-local scatter searches [Egea et al., 2014] tailored to a limited computational budget. Further, low-fidelity surrogate models could be systematically trained and used, both for ODE models and especially for expensive agent-based simulators as commonly used in ABC, e.g. via coarse discretization, using simplified mechanistic descriptions, or based on data-analytic machine learning approaches, to replace evaluations of the full high-fidelity model and guide the search [Latz et al., 2018, Prescott and Baker, 2021]. Tools allowing to systematically do so would be of interest and could improve applicability and scalability of in particular likelihood-free approaches substantially.

From a method-development perspective, benchmark problems are important to robustly and fairly evaluate and compare novel and established methods [Buchka et al., 2021]. While collections have been established e.g. for ODE models [Hass et al., 2019], similar efforts would be beneficial to enhance method development in likelihood-free inference. Further, standardization of problems and methods aids in particular reusability of data and models and interoperability of methods and analysis pipelines, as well as reproducibility of results [Wilkinson et al., 2016]. While there have been efforts of standardization on various levels, many methods are still being developed in closed environments, tailored to specific applications, or with walls raised e.g. via programming language or API, obstructing application and comparison of methods. Community standardization efforts would thus be beneficial to method and model development, and application in general.

In conclusion, in this thesis we developed efficient, accurate, robust parameter inference methods for specific challenges in the context of ODE models and ABC. These provide substantial improvements over alternative approaches, in some cases realistically facilitating efficient and accurate inference for the first time, and open up new research possibilities and perspectives. Therefore, we anticipate that the developed methods will contribute to a more holistic mechanistic understanding of processes at a systems level, in systems biology and beyond.

Bibliography

A. R. A. Anderson and V. Quaranta. Integrative mathematical oncology. *Nat. Rev. Cancer*, 8 (3):227–234, Mar. 2008.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 72(3):269–342, June 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, July 2011.

T. Bäck. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, New York and Oxford, 1996.

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

B. Ballnus. *Development and Evaluation of Sampling-based Parameter Estimation Methods for Dynamic Biological Processes*. PhD thesis, Technische Universität München, 2019.

B. Ballnus, S. Hug, K. Hatz, L. Görlitz, J. Hasenauer, and F. J. Theis. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst Biol*, 11 (63):63, 2017. doi: 10.1186/s12918-017-0433-1.

E. Balsa-Canto and J. R. Banga. AMIGO, a toolbox for advanced model identification in systems biology using global optimization. *Bioinformatics*, 27(16):2311–2313, Aug. 2011.

S. Barber, J. Voss, and M. Webster. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105, 2015.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, Jr, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber,

J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar. 2012. doi: 10.1038/nature11003.

A. Basu, I. R. Harris, N. L. Hjort, and M. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, 12 2002.

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

E. A. Bender. *An introduction to mathematical modeling*. Courier Corporation, 2012.

J. O. Berger, E. Moreno, L. R. Pericchi, M. J. Bayarri, J. M. Bernardo, J. A. Cano, J. De la Horra, J. Martín, D. Ríos-Insúa, B. Betrò, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, 1(8):9, 2017.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

M. G. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Inform. Process. Lett.*, 24(6):377–380, 1987. doi: 10.1016/0020-0190(87)90114-1.

M. E. Boehm, L. Adlung, M. Schilling, S. Roth, U. Klingmueller, and W. D. Lehmann. Identification of isoform-specific dynamics in phosphorylation-dependent stat5 dimerization by quantitative mass spectrometry and mathematical modeling. *Journal of Proteome Research*, 13 (12):5685–5694, 2014.

N. Bohr. I. on the constitution of atoms and molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 26(151):1–25, 1913.

A. Borowska, D. Giurghita, and D. Husmeier. Gaussian process enhanced semi-automatic approximate Bayesian computation: parameter inference in a stochastic differential equation system for chemotaxis. *Journal of Computational Physics*, 429:109999, 2021.

P. Bortot, S. G. Coles, and S. A. Sisson. Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92, 2007.

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, Mar. 2018.

G. E. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

S. Boyd and L. Vandenberghe. *Convex Optimisation*. Cambridge University Press, UK, 2004.

M. A. Branch, T. F. Coleman, and Y. Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput.*, 21(1):1–23, 1999. doi: 10.1137/s1064827595289108.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

S. Buchka, A. Hapfelmeier, P. P. Gardner, R. Wilson, and A.-L. Boulesteix. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome biology*, 22(1): 1–8, 2021.

R. H. Byrd, H. F. Khalfan, and R. B. Schnabel. Analysis of a symmetric rank-one trust region method. *SIAM J. Optim.*, 6(4):1025–1039, 1996. doi: 10.1137/s1052623493252985.

Y. Cao, S. Li, L. Petzold, and R. Serban. Adjoint sensitivity analysis for differential-algebraic equations: The adjoint dae system and its numerical solution. *SIAM journal on scientific computing*, 24(3):1076–1089, 2003.

G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

A. Chang, M. Scheer, A. Grote, I. Schomburg, and D. Schomburg. BRENDA, AMENDA and FRENDA the enzyme information system: New content and tools in 2009. *Nucleic Acids Res.*, 37(Database issue):D588–92, Jan 2009. doi: 10.1093/nar/gkn820.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

B.-E. Chérief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR, 2020.

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

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. McGraw-Hill, New York, 1955.

T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, 6:418–445, 1996.

S. Coxe, S. G. West, and L. S. Aiken. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2):121–136, 2009.

K. Csilléry, O. François, and M. G. Blum. abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, 3(3):475–479, 2012.

A. A. Cuellar, C. M. Lloyd, P. F. Nielsen, D. P. Bullivant, D. P. Nickerson, and P. J. Hunter. An overview of CellML 1.1, a biological model description language. *Simulation*, 79(12):740–747, 2003. doi: 10.1177/0037549703040939.

H. B. Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.

J. O. Dada and P. Mendes. Multi-scale modelling and simulation in systems biology. *Integr. Biol.*, 3:86–96, 2011. doi: 10.1039/c0ib00075b.

J. O. Dada, I. Spasić, N. W. Paton, and P. Mendes. SBRML: a markup language for associating systems biology data with models. *Bioinformatics*, 26:932–938, Apr. 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq069.

A. C. Daly, J. Cooper, D. J. Gavaghan, and C. Holmes. Comparing two sequential Monte Carlo samplers for exact and approximate Bayesian inference on biological models. *J. R. Soc. Interface*, 14(134):20170340, 2017.

H.-D. Dau and N. Chopin. Waste-free sequential monte carlo. *arXiv preprint arXiv:2011.02328*, 2020.

M. De La Maza and D. Yuret. Dynamic hill climbing. *AI expert*, 9:26–26, 1994.

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. doi: 10.1038/s41540-017-0023-2.

P. Del Moral. Nonlinear filtering: Interacting particle resolution. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 325(6):653–658, 1997.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. B*, 68 (3):411–436, 2006.

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE, 2005.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comp.*, 10(3):197–208, July 2000. doi: 10.1023/A:1008935410038.

C. C. Drovandi and A. N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

C. C. Drovandi, A. N. Pettitt, and M. J. Faddy. Approximate bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60 (3):317–337, 2011.

K. Durso-Cain, P. Kumberger, Y. Schälte, T. Fink, H. Dahari, J. Hasenauer, S. L. Uprichard, and F. Graw. HCV spread kinetics reveal varying contributions of transmission modes to infection dynamics. *Viruses*, 13(7), July 2021. ISSN 1999-4915. doi: 10.3390/v13071308.

R. Dutta, M. Schoengens, J.-P. Onnela, and A. Mira. Abcpy: A user-friendly, extensible, and parallel library for approximate Bayesian computation. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '17, pages 8:1–8:9, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5062-4. doi: 10.1145/3093172.3093233.

D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7(23):3910–3916, 2005.

F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, F. Kogera, M. Dorel, M. J. Garnett, N. Blüthgen, and J. Saez-Rodriguez. Drug resistance mechanisms in colorectal cancer dissected with cell type-specific dynamic logic models. *Cancer Res.*, 77(12):3364–3375, 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0078.

J. A. Egea, D. Henriques, T. Cokelaer, A. F. Villaverde, A. MacNamara, D. P. Danciu, J. R. Banga, and J. Saez-Rodriguez. MEIGO: An open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinf.*, 15(136), 2014. doi: 10.1186/1471-2105-15-136.

A. Einstein. Zur Elektrodynamik bewegter Körper. *Annalen der Physik*, 4, 1905.

O. Eriksson, A. Jauhiainen, S. Maad Sasane, A. Kramer, A. G. Nair, C. Sartorius, and J. Hellgren Kotaleski. Uncertainty quantification, propagation and characterization by bayesian analysis combined with global sensitivity analysis applied to dynamical intracellular pathway models. *Bioinformatics*, 35(2):284–292, 2019.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. B*, 74(3):419–474, 2012.

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

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.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London, Ser. A*, 222:309–368, 1922. doi: 10.1098/rsta.1922.0009.

R. Fletcher and M. J. Powell. A rapidly convergent descent method for minimization. *Comp J*, 6(2):163–168, 1963. doi: 10.1093/comjnl/6.2.163.

D. T. Frazier. Robust and efficient approximate Bayesian computation: A minimum distance approach. *arXiv preprint arXiv:2006.14126*, 2020.

D. T. Frazier, C. Drovandi, and R. Loaiza-Maya. Robust approximate Bayesian computation: An adjustment approach. *arXiv preprint arXiv:2008.04099*, 2020.

F. Fröhlich, F. J. Theis, and J. Hasenauer. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. In P. Mendes, J. O. Dada, and K. O. Smallbone, editors, *Proc. 12th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 61–72. Springer International Publishing Switzerland, Nov. 2014. doi: 10.1007/978-3-319-12982-2{_}5.

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. doi: 10.1371/journal.pcbi.1005331.

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, Dec. 2018. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2018.10.013.

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.

F. Fröhlich, D. Weindl, Y. Schälte, D. Pathirana, Ł. Paszkowski, G. T. Lines, P. Stapor, and J. Hasenauer. AMICI: high-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics*, btab227, April 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab227. btab227.

M. Fujisawa, T. Teshima, I. Sato, and M. Sugiyama. γ-ABC: Outlier-robust approximate Bayesian computation based on a robust divergence estimator. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR, 2021.

D. Garfinkel, C. B. Marbach, and N. Z. Shapiro. Stiff differential equations. *Annual review of biophysics and bioengineering*, 6(1):525–542, 1977.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

N. A. Gershenfeld and N. Gershenfeld. *The nature of mathematical modeling*. Cambridge university press, 1999.

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*, volume 4, pages 169–193. University Press Oxford, 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.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81 (25):2340–2361, Dec. 1977. doi: 10.1021/j100540a008.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B*, 73(2):123–214, Mar. 2011. doi: 10.1111/j.1467-9868.2010.00765.x.

A. Gleim and C. Pigorsch. Approximate Bayesian computation with indirect summary statistics. *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*, 2013.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Math Comp*, 24 (109):23–26, 1970. ISSN 00255718, 10886842. doi: 10.1090/s0025-5718-1970-0258249-6.

J. H. Goldwyn, N. S. Imennov, M. Famulare, and E. Shea-Brown. Stochastic differential equation models for ion channel noise in hodgkin-huxley neurons. *Physical Review E*, 83(4):041908, 2011.

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

H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Stat. Comp.*, 16(4):339–354, 2006. doi: 10.1007/s11222-006-9438-0.

N. Hansen and A. Ostermaier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. *Evolutionary Computation*, pages 312–317, 1996.

L. A. Harris, J. S. Hogg, J.-J. Tapia, J. A. P. Sekar, S. Gupta, I. Korsunsky, A. Arora, D. Barua, R. P. Sheehan, and J. R. Faeder. BioNetGen 2.2: advances in rule-based modeling. *Bioinformatics*, 32(21):3366–3368, 07 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw469.

J. U. Harrison and R. E. Baker. An automatic adaptive method to combine summary statistics in approximate bayesian computation. *PloS one*, 15(8):e0236954, 2020.

F. Hartig, J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. Statistical inference for stochastic simulation models–theory and application. *Ecology letters*, 14(8):816–827, 2011.

J. Hasenauer. A special issue on analysis of coupled/multi-scale biological systems. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):95–96, Sept. 2015. doi: 10.1166/jcsmd.2015.1068.

J. Hasenauer, N. Jagiella, S. Hross, and F. J. Theis. Data-driven modelling of biological multi-scale processes. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):101–121, Sept. 2015. doi: 10.1166/jcsmd.2015.1069.

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. doi: 10.1038/s41540-017-0030-3.

H. Hass, C. Loos, E. Raimúndez-Álvarez, J. Timmer, J. Hasenauer, and C. Kreutz. Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, 35(17):3073–3082, 01 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz020.

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

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, July 2006. doi: 10.1038/ni0706-681.

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, September 2005. doi: 10.1145/1089014.1089020.

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

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544, Aug. 1952.

M. D. Hoffman and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI – a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006. doi: 10.1093/bioinformatics/btl485.

S. Hross, A. Fiedler, F. J. Theis, and J. Hasenauer. Quantitative comparison of competing PDE models for Pom1p dynamics in fission yeast. In R. Findeisen, E. Bullinger, E. Balsa-Canto, and K. Bernaerts, editors, *Proc. 6th IFAC Conf. Found. Syst. Biol. Eng.*, volume 49, pages 264–269. IFAC-PapersOnLine, 2016. doi: 10.1016/j.ifacol.2016.12.136.

S. Hross, F. J. Theis, M. Sixt, and J. Hasenauer. Mechanistic description of spatial processes using integrative modelling of noise-corrupted imaging data. *J. R. Soc. Interface*, 15(149): 20180600, Dec. 2018. doi: 10.1098/rsif.2018.0600.

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

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. doi: 10.1093/bioinformatics/btg015.

P. J. Hunter and T. K. Borg. Integration from proteins to organs: the Physiome Project. *Nat. Rev. Mol. Cell. Biol.*, 4(3):237–243, Mar. 2003. doi: 0.1038/nrm1054.

A. Imle, P. Kumberger, N. D. Schnellbächer, J. Fehr, P. Carrillo-Bustamante, J. Ales, P. Schmidt, C. Ritter, W. J. Godinez, B. Müller, et al. Experimental and computational analyses reveal that environmental restrictions shape HIV-1 spread in 3D cultures. *Nature Communications*, 10(1):2144, 2019.

B. Ingalls. *Mathematical modelling in systems biology: An introduction.* MIT Press, 2013.

N. Jagiella, D. Rickert, F. J. Theis, and J. Hasenauer. Parallelization and high-performance computing enables automated statistical inference of multi-scale models. *Cell Systems*, 4(2): 194–206, 02 2017. doi: 10.1016/j.cels.2016.12.002.

E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

B. Jiang. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics*, pages 1711–1721. PMLR, 2018.

B. Jiang, T.-y. Wu, C. Zheng, and W. H. Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618, 2017.

M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Eng.*, 8:447–455, May 2006.

P. Joyce and P. Marjoram. Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

R. P. Kanwal. *Generalized functions theory and technique: Theory and technique*. Springer Science & Business Media, 1998.

E.-M. Kapfer, P. Stapor, and J. Hasenauer. Challenges in the calibration of large-scale ordinary differential equation models. *IFAC-PapersOnLine*, 52(26):58–64, Dec. 2019.

J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, July 2012. doi: 10.1016/j.cell.2012.05.044.

D. Kaschek, W. Mader, M. Fehling-Kaschek, M. Rosenblatt, and J. Timmer. Dynamic modeling, parameter estimation, and uncertainty analysis in R. *J. Stat. Softw.*, 88(10), 2019.

J. Kennedy. Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer, 2011.

W. O. Kermack, A. G. McKendrick, and G. T. Walker. A Contribution to the Mathematical Theory of Epidemics. *P Roy Soc A-Math Phy*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118.

D. P. Kingma and L. J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 2015 - accepted papers*. Ithaca, 2015.

S. Kirkpatrick, C. D. Gelatt Jr, and M. P. M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983. doi: 10.1126/science.220.4598.671.

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

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov. 2002b.

E. Klinger and J. Hasenauer. A scheme for adaptive selection of population sizes in Approximate Bayesian Computation - Sequential Monte Carlo. In J. Feret and H. Koeppl, editors, *Computational Methods in Systems Biology. CMSB 2017*, volume 10545 of *Lecture Notes in Computer Science*. Springer, Cham, 2017.

E. Klinger, D. Rickert, and J. Hasenauer. pyABC: distributed, likelihood-free inference. *Bioinformatics*, 34(20):3591–3593, 10 2018. doi: 10.1093/bioinformatics/bty361.

E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice*. Wiley-VCH, Weinheim, 2005. ISBN 978-3-527-31078-4.

A. Korkut, W. Wang, E. Demir, B. A. Aksoy, X. Jing, E. J. Molinelli, Ö. Babur, D. L. Bemis, S. O. Sumer, D. B. Solit, et al. Perturbation biology nominates upstream–downstream drug combinations in raf inhibitor resistant melanoma cells. *Elife*, 4:e04640, 2015.

A. Kramer, J. Hasenauer, F. Allgöwer, and N. Radde. Computation of the posterior entropy in a Bayesian framework for parameter estimation in biological networks. In *Proc. IEEE Multi-Conf. Syst. Contr.*, pages 493–498, Yokohama, Japan, Sept. 2010. doi: 10.1109/CCA.2010.5611198.

P. F. Lang, S. Shin, and V. M. Zavala. Sbml2julia: interfacing SBML with efficient nonlinear Julia modelling and solution tools for parameter optimization, 2020.

J. Latz, I. Papaioannou, and E. Ullmann. Multilevel sequential2 Monte Carlo for Bayesian inverse problems. *Journal of Computational Physics*, 368:154–178, 2018.

J. R. Leis and M. A. Kramer. The simultaneous solution and sensitivity analysis of systems described by ordinary differential equations. *ACM Transactions on Mathematical Software (TOMS)*, 14(1):45–60, 1988.

O. Lenive, P. D. Kirk, and M. P. Stumpf. Inferring extrinsic noise from single-cell gene expression data using approximate bayesian computation. *BMC Systems biology*, 10(1):81, 2016.

C. Leonhardt, G. Schwake, T. R. Stögbauer, S. Rappl, J. T. Kuhr, T. S. Ligon, and J. O. Rädler. Single-cell mRNA transfection studies: Delivery, kinetics and statistics by numbers. *Nanomedicine: Nanotechnology, Biology, and Medicine*, 10(4):679–688, May 2014. doi: 10.1016/j.nano.2013.11.008.

C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. I. Stefan, J. L. Snoep, M. Hucka, N. Le Novère, and C. Laibe. BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol*, 4:92, 2010.

J. Li, W. Zhao, R. Akbani, W. Liu, Z. Ju, S. Ling, C. P. Vellano, P. Roebuck, Q. Yu, A. K. Eterovic, L. A. Byers, M. A. Davies, W. Deng, Y. N. V. Gopal, G. Chen, E. M. von Euw, D. Slamon, D. Conklin, J. V. Heymach, A. F. Gazdar, J. D. Minna, J. N. Myers, Y. Lu, G. B. Mills, and H. Liang. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell*, 31(2):225–239, 2017/05/08 2017. doi: 10.1016/j.ccell.2017.01.005.

S. Li and L. Petzold. Adjoint sensitivity analysis for time-dependent partial differential equations with adaptive mesh refinement. *Journal of Computational Physics*, 198(1):310–325, 2004.

X. Li, T.-K. L. Wong, R. T. Chen, and D. Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pages 3870–3882. PMLR, 2020.

D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1):503–528, 1989. doi: 10.1007/bf01589116.

J. S. Liu, R. Chen, and W. H. Wong. Rejection control and sequential importance sampling. *J. Am. Stat. Assoc.*, 93(443):1022–1031, 1998.

A. C. Lloyd. The regulation of cell size. *Cell*, 154(6):1194–1205, 2013.

C. Loos, S. Krause, and J. Hasenauer. Hierarchical optimization for the efficient parametrization of ODE models. *Bioinf.*, 34(24):4266–4273, July 2018. doi: 10.1093/bioinformatics/bty514.

C. Maier, C. Loos, and J. Hasenauer. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725, Mar. 2017. doi: 10.1093/bioinformatics/btw703.

C. Maier, N. Hartung, J. de Wiljes, C. Kloft, and W. Huisinga. Bayesian data assimilation to support informed decision making in individualized chemotherapy. *CPT: pharmacometrics & systems pharmacology*, 9(3):153–164, 2020.

T. Maiwald and J. Timmer. Dynamical modeling and multi-experiment fitting with PottersWheel. *Bioinformatics*, 24(18):2037–2043, July 2008.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavare. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328, Dec. 2003.

G. M. Martin, D. T. Frazier, and C. P. Robert. Computing Bayes: Bayesian computation from 1763 to the 21st century. *arXiv preprint arXiv:2004.06425*, 2020.

L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

M. L. Martins, S. C. Ferreira Jr., and M. J. Vilela. Multiscale models for biological systems. *Curr. Opin. Colloid Interface Sci.*, 15(1–2):18–23, Apr. 2010. doi: 10.1016/j.cocis.2009.04.004.

T. McKinley, A. R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *Int. J. of Biostat.*, 5(1), 2009.

P. Mendes, S. Hoops, S. Sahle, R. Gauges, J. Dada, and U. Kummer. *Computational Modeling of Biochemical Networks Using COPASI*, chapter 2. Part of the Methods in Molecular Biology. Humana Press, 2009.

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

B. Miasojedow, E. Moulines, and M. Vihola. An adaptive parallel tempering algorithm. *J. Comput. Graph. Stat.*, 22(3):649–664, 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.

U. Münzner, E. Klipp, and M. Krantz. A comprehensive, mechanistically detailed, and executable model of the cell division cycle in saccharomyces cerevisiae. *Nat. Commun.*, 10, 2019.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

R. M. Neal. *Probabilistic inference using Markov chain Monte Carlo methods.* Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

R. M. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics. Chapman & Hall / CRC Press, London, United Kingdom, 2011.

J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7(4): 308–313, 1965. doi: 10.1093/comjnl/7.4.308.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994. doi: https://doi.org/10.1016/S1573-4412(05)80005-4.

I. Newton. *Philosophiae naturalis principia mathematica*, volume 2. typis A. et JM Duncan, 1833.

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. doi: 10.1007/b98874.

M. A. Nunes and D. J. Balding. On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol.*, 9(1), 2010.

B. G. Olivier and J. L. Snoep. Web-based kinetic modelling using JWS Online. *Bioinf.*, 20(13): 2143–4, Sep 2004. doi: 10.1093/bioinformatics/bth200.

J. Owen, D. J. Wilkinson, and C. S. Gillespie. Likelihood free inference for Markov processes: a comparison. *Stat. App. Gen. Mol Biol*, 14(2):189–209, 2015.

D. B. Özyurt and P. I. Barton. Cheap second order directional derivatives of stiff ODE embedded functionals. *SIAM J. Sci. Comput.*, 26(5):1725–1743, 2005. doi: 10.1137/030601582.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. R. Penas, P. González, J. A. Egea, J. R. Banga, and R. Doallo. Parallel metaheuristics in computational biology: An asynchronous cooperative enhanced scatter search method. *Procedia Comput. Sci.*, 51:630–639, 2015. doi: 10.1016/j.procs.2015.05.331.

D. R. Penas, P. González, J. A. Egea, R. Doallo, and J. R. Banga. Parameter estimation in large-scale systems biology models: a parallel and self-adaptive cooperative strategy. *BMC bioinformatics*, 18(1):52, 2017.

U. Picchini. Inference for SDE models via approximate Bayesian computation. *Journal of Computational and Graphical Statistics*, 23(4):1080–1100, 2014.

M. J. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994.

M. J. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 2009.

D. Prangle. Adapting the ABC distance function. *Bayesian Analysis*, 12(1):289–309, 2017. doi: 10.1214/16-BA1002.

C. Predescu, M. Predescu, and C. V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *The Journal of Chemical Physics*, 120(9): 4119–4128, 2004.

T. P. Prescott and R. E. Baker. Multifidelity approximate Bayesian computation with sequential Monte Carlo parameter sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2): 788–817, 2021.

L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. *J. Comp. Graph. Stat.*, 27(1):1–11, 2018.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

E. Raimúndez, S. Keller, G. Zwingenberger, K. Ebert, S. Hug, F. J. Theis, D. Maier, B. Luber, and J. Hasenauer. Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines. *PLoS Comput. Biol.*, 16(3):e1007147, 2020.

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, May 2009. doi: 10.1093/bioinformatics/btp358.

A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philos T Roy Soc A*, 371(1984), 2013a. doi: 10.1098/rsta.2011.0544.

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, Sept. 2013b. doi: 10.1371/journal.pone.0074335.

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. doi: 10.1093/bioinformatics/btv405.

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, July 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, June 2012. doi: 10.1016/j.csda.2011.11.020.

L. M. Rios and N. V. Sahinidis. Derivative-free optimization: A review of algorithms and comparison of software implementations. *J. Global Optim.*, 56(3):1247–1293, Jul 2013. ISSN 1573-2916. doi: 10.1007/s10898-012-9951-y.

C. Robert and G. Casella. A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, pages 102–115, 2011.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

E. Ruli, N. Sartori, and L. Ventura. Robust approximate Bayesian inference. *Journal of Statistical Planning and Inference*, 205:10–22, 2020.

J. Schaff, C. C. Fink, B. Slepchenko, J. H. Carson, and L. M. Loew. A general computational framework for modeling cellular structure and function. *Biophys. J.*, 73(3):1135–1146, Sept. 1997.

Y. Schälte and J. Hasenauer. Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. *Bioinformatics*, 36(Supplement 1):i551–i559, 7 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa397.

Y. Schälte, P. Stapor, and J. Hasenauer. Evaluation of derivative-free optimizers for parameter estimation in systems biology. *IFAC-PapersOnLine*, 51(19):98–101, 2018. 7th Conference on Foundations of Systems Biology in Engineering FOSBE 2018.

Y. Schälte, E. Alamoudi, and J. Hasenauer. Robust adaptive distance functions for approximate Bayesian inference on outlier-corrupted data. *bioRxiv*, 2021a.

Y. Schälte, F. Fröhlich, P. Stapor, J. Vanhoefer, D. Wang, D. Weindl, P. J. Jost, P. Lakrisenko, E. Raimúndez, D. Pathirana, L. Schmiester, P. Sädter, L. Contento, E. Dudkin, K. Meyer, S. Merkt, and J. Hasenauer. ICB-DCM/pyPESTO: pyPESTO 0.2.5, May 2021b. https://doi.org/10.5281/zenodo.2553546.

L. Schmiester, Y. Schälte, F. Fröhlich, J. Hasenauer, and D. Weindl. Efficient parameterization of large-scale dynamic models based on relative measurements. *Bioinformatics*, 36(2):594–602, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz581. URL `https://doi.org/10.1093/bioinformatics/btz581`.

L. Schmiester, D. Weindl, and J. Hasenauer. Parameterization of mechanistic models from qualitative data using an efficient optimal scaling approach. *J. Math. Biol.*, 81(2):603–623, July 2020. doi: 10.1007/s00285-020-01522-w.

L. Schmiester, Y. Schälte, F. T. Bergmann, T. Camba, E. Dudkin, J. Egert, F. Fröhlich, L. Fuhrmann, A. L. Hauber, S. Kemmer, P. Lakrisenko, C. Loos, S. Merkt, W. Müller, D. Pathirana, E. Raimúndez, L. Refisch, M. Rosenblatt, P. L. Stapor, P. Städter, D. Wang, F.-G. Wieland, J. R. Banga, J. Timmer, A. F. Villaverde, S. Sahle, C. Kreutz, J. Hasenauer, and D. Weindl. PEtab—interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology*, 17(1):1–10, January 2021a. doi: 10.1371/journal.pcbi.1008646.

L. Schmiester, D. Weindl, and J. Hasenauer. Efficient gradient-based parameter estimation for dynamic models using qualitative data. *Bioinformatics*, btab512, 2021b. doi: 10.1093/bioinformatics/btab512.

I. Schomburg, A. Chang, and D. Schomburg. Brenda, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 2002.

G. Schwarz. Estimating the dimension of a model. *Ann Stat*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

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*, volume 6, pages 257–269. ASME, 2005. ISBN 0-7918-4743-8. doi: 10.1115/DETC2005-85597.

A. B. Shiflet and G. W. Shiflet. *Introduction to computational science: modeling and simulation for the sciences*. Princeton University Press, 2014.

D. Silk, S. Filippi, and M. P. H. Stumpf. Optimizing threshold-schedules for sequential approximate Bayesian computation: Applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.*, 12(5):603–618, Oct. 2013.

S. A. Sisson and Y. Fan. ABC samplers. In *Handbook of Approximate Bayesian Computation*, pages 87–123. Chapman and Hall/CRC, 2018.

S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 104(6):1760–1765, Jan. 2007. doi: 10.1073/pnas.0607208104.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018.

K. Smallbone and P. Mendes. Large-scale metabolic models: From reconstruction to differential equations. *Ind. Biotechnol.*, 9(4):179–184, 2013. doi: 10.1089/ind.2013.0003.

S. L. Spencer and P. K. Sorger. Measuring and modeling apoptosis in single cells. *Cell*, 144(6):926–939, Mar. 2011. doi: 10.1016/j.cell.2011.03.002.

P. Städter, Y. Schälte, L. Schmiester, J. Hasenauer, and P. L. Stapor. Benchmarking of numerical integration methods for ODE models of biological systems. *Scientific Reports*, 11(1):2696, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82196-2. URL `https://doi.org/10.1038/s41598-021-82196-2`.

N. J. Stanford, M. Scharm, P. D. Dobson, M. Golebiewski, M. Hucka, V. B. Kothamachu, D. Nickerson, S. Owen, J. Pahle, U. Wittig, et al. Data management in computational systems biology: exploring standards, tools, databases, and packaging best practices. In *Yeast Systems Biology*, pages 285–314. Springer, 2019.

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, 2018.

P. Stapor, L. Schmiester, C. Wierling, B. M. Lange, D. Weindl, and J. Hasenauer. Mini-batch optimization enables training of ode models on large-scale datasets. *bioRxiv*, 2019. doi: 10.1101/859884. URL `https://www.biorxiv.org/content/early/2019/11/30/859884`.

J. Starruß, W. de Back, L. Brusch, and A. Deutsch. Morpheus: A user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics*, 30(9):1331–1332, Jan. 2014. doi: 10.1093/bioinformatics/btt772.

E. M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, 2005.

Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *The Journal of Chemical Physics*, 113(15):6042–6051, 2000.

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, Feb 2003.

S. Syga, D. David-Rus, Y. Schälte, M. Meyer-Hermann, H. Hatzikirou, and A. Deutsch. Inferring the effect of interventions on COVID-19 transmission networks. *arXiv preprint arXiv:2012.03846*, 2020.

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

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

C. Thomaseth and N. Radde. Normalization of Western blot data affects the statistics of estimators. *IFAC-PapersOnLine*, 49(26):56–62, 2016.

T. Toni and M. P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 10 2010.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6:187–202, 7 2009.

T. E. Turner, S. Schnell, and K. Burrage. Stochastic approaches for modelling in vivo reactions. *Comput. Biol. Chem.*, 28(3):165–178, July 2004.

E. van der Vaart, D. Prangle, and R. M. Sibly. Taking error into account when fitting models using approximate bayesian computation. *Ecological Applications*, 28(2):267–274, 2018. doi: 10.1002/eap.1656.

J. Vanhoefer, M. R. A. Matos, D. Pathirana, Y. Schälte, and J. Hasenauer. yaml2sbml: Human-readable and -writable specification of ode models and their conversion to sbml. *Journal of Open Source Software*, 6(61):3215, May 2021. doi: 10.21105/joss.03215.

A. Vaz and L. Vicente. A particle swarm pattern search method for bound constrained global optimization. *J. Global Optim.*, 39(2):197–219, 2007. doi: 10.1007/s10898-007-9133-5.

D. Venzon and S. Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):87–94, 1988.

A. F. Villaverde, F. Froehlich, D. Weindl, J. Hasenauer, and J. R. Banga. Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics*, page bty736, 2018.

A. F. Villaverde, E. Raimúndez-Álvarez, J. Hasenauer, and J. R. Banga. A comparison of methods for quantifying prediction uncertainty in systems biology. *IFAC-PapersOnLine*, 2019.

A. F. Villaverde, D. Pathirana, F. Fröhlich, J. Hasenauer, and J. R. Banga. A protocol for dynamic model calibration. *arXiv preprint arXiv:2105.12008*, 2021.

W. Vousden, W. M. Farr, and I. Mandel. Dynamic temperature selection for parallel tempering in markov chain monte carlo simulations. *Mon. Not. R. Astron. Soc.*, 455(2):1919–1937, 2016.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1):25–57, 2006. doi: 10.1007/s10107-004-0559-y.

D. Waltemath, R. Adams, F. T. Bergmann, M. Hucka, F. K. A. K. Miller, I. I. Moraru, D. Nickerson, J. L. Snoep, and N. Le Novère. Reproducible computational biology experiments with SED-ML – The Simulation Experiment Description Markup Language. *BMC Syst. Biol.*, 5 (198), Dec. 2011. doi: 10.1186/1752-0509-5-198.

D. J. Warne, R. E. Baker, and M. J. Simpson. Multilevel rejection sampling for approximate bayesian computation. *Computational Statistics & Data Analysis*, 124:71–86, 2018.

D. J. Warne, R. E. Baker, and M. J. Simpson. A practical guide to pseudo-marginal methods for computational inference in systems biology. *Journal of theoretical biology*, 496:110255, 2020.

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, Aug. 2011. doi: 10.3182/20110828-6-IT-1002.01007.

D. J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.*, 10(2):122–133, Feb. 2009. doi: 10.1038/nrg2509.

D. J. Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2011.

M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Sci Data*, 3:160018, Mar 2016. doi: 10.1038/sdata.2016.18.

R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Stat. Appl. Gen. Mol. Bio.*, 12(2):129–141, May 2013. doi: 10.1515/sagmb-2013-0010.

S. S. Wilks. Mathematical statistics. *John Wiley and Sons*, 1962.

U. Wittig, R. Kania, M. Golebiewski, M. Rey, L. Shi, L. Jong, E. Algaa, A. Weidemann, H. Sauer-Danzwith, S. Mir, O. Krebs, M. Bittkowski, E. Wetsch, I. Rojas, and W. Müller. SABIO-RK–database for biochemical reaction kinetics. *Nucl. Acids Res.*, 40(D1):D790–D796, 2012. doi: 10.1093/nar/gkr1046.

S. Zacks. *Parametric statistical inference: basic theory and modern approaches*, volume 4. Elsevier, 2014.

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, Aug. 2012. doi: 10.1073/pnas.1205707109.