HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

Statistical modelling of functional data from biological systems

Ivan Kondofersky

February 2016

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

"Statistical modelling of functional data from biological systems"

Ivan Kondofersky

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:

Univ.-Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Dr. Christian Heumann, Ludwig-Maximilians-Universität München
3. Univ.-Prof. Dr. Jens Timmer, Albert-Ludwigs-Universität Freiburg

Die Dissertation wurde am 11.02.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 20.04.2016 angenommen.

Danksagung

An dieser Stelle möchte ich mich bei einigen Leuten bedanken, die mich während meiner Promotion auf verschiedensten Arten unterstützt haben.

Als allererstes gilt mein tiefster Dank Fabian Theis. Als Doktovater habe ich immer und stets Deine Unterstützung und Dein Vertrauen genossen. Für mich ist es eine große Ehre gewesen, meine Doktorarbeit unter deiner Obhut zu schreiben und ich freue mich sehr auf unsere zukünftige Zusammenarbeit.

Ja, ja, diese Doktorarbeit... Sie wäre wahrscheinlich erst gar nicht entstanden ohne die Hilfe von Christiane Fuchs. Mit Deinem unermüdlichem Einsatz hast Du mir immer wieder genau den Input gegeben, den ich gebraucht habe, um mich wissenschaftlich weiterzuentwickeln. Dass aus einer so engen beruflichen Zusammenarbeit auch eine wunderbare private Freundschaft entstanden ist, weiss ich dabei mindestens genauso hoch zu schätzen. Christiane, vielen Dank!

Selbstverständlich danke ich den beiden Gutachtern Christian Heumann und Jens Timmer für das Lesen und Bewerten meiner Dissertation.

Des Weiteren möchte ich mich auch bei den zahlreichen Kollaborationspartnern bedanken. Sie haben immer geduldig mein limitiertes biologisches Wissen tolleriert und noch geduldiger meine manchmal zu komplizierten Ausflüge in die Welt der Statistik hingenommen. Da denke ich hauptsächlich an Tina, Dennis, Simone, Andrey und Inna.

Als nächstes möchte ich betonen, dass die Atmosphäre am ICB für mich einen stets angenehmen Arbeitsalltag gebildet hat und somit einen äußerst inspierierenden Einfluss auf meine Doktorarbeit hatte. Bedanken möchte ich mich deshalb bei allen Kollegen vom ICB, insbesondere Nikola, Steffen, Katrin, Sabine und Adriana.

Zuletzt möchte ich mich auch bei denjenigen Personen bedanken, die dafür gesorgt haben, dass es meiner Seele gut geht und ich mich voll und ganz auf meine Arbeit konzentrieren kann. Bei meinen Eltern und meinem Bruder Mitko möchte ich mich vor allem für den nie schwindenden Glauben an mich bedanken. Ilina und Lea dagegen möchte ich vor allem für die große Menge an Geduld und die noch größere Menge an Zuneigung während der letzten Jahre danken.

An alle, die ich hier vergesse: Danke!

Abstract

The collection of high-resolution temporal data from complex molecular biological systems has become of great importance over the past decades. Further, the quality of available data has reached high standards, allowing a complex study of dynamical evolution in biological systems. Hypotheses are usually investigated with statistical and mathematical methods with conclusions based on the collected data. Such methods range from functional data analysis to ordinary differential equations, approximating the temporal measurements by flexible smooth functions or modelling the system and its relation to the change of this system over time mechanistically, respectively. With an increased quality and quantity of available temporal data, new research possibilities are created. Corresponding statistical or mathematical methods often perform, however, unsatisfactorily by not considering the full information. As a result, the applied methods are either not capable or unavailable of handling new research challenges. The application of an inappropriate method might lead to false conclusions and thus corrupt the established work-flow of data generation, data analysis and conclusion statements. In the present thesis, we consider the analysis of data from biological systems where knowledge about a certain system is available and yet some indeterminacies about the system remain. Because the system cannot be fully described our approach relies on modelling of the underdetermined parts with novel techniques. We identify three such situations where a novel statistical tool is able to substantially increase the modelling possibilities and at the same time reduce the system indeterminacies. First, we employ data modelling by a significance test for difference between two groups of paired temporal observations. Next, we mechanistically study and describe biological networks and consider a change in network topology through additional latent components if a network extension in required. Finally, we build up our network extension approach and generalize it to non-linear extensions by studying catalysis in the biological system. In each of the three situations, we propose a novel method which is either able to outperform existing ones or its application suggests additional aspects on the regulation and composition of the studied biological systems. Each method is thoroughly evaluated on a large number of simulated data scenarios. Moreover, we investigate real-world data examples where results suggest novel insights into the studied applications.

Zusammenfassung

Hochaufgelöste Zeitdaten aus der Molekularbiologie stehen in immer größerem Umfang zur Verfügung. Dabei ist die Qualität dieser Daten zu neuen und höheren Standards gestiegen. Datensammlungen solcher Art ermöglichen unter anderem die Untersuchung dynamischer Verläufe biologischer Systeme. Hierbei werden verschiedenste Hypothesen unter Anwendung von statistischen und mathematischen Methoden zur Datenanalyse untersucht und Schlussfolgerungen auf der Grundlage der gesammelten Daten gezogen. Solche Methoden reichen von der Analyse funktionaler Daten, bei der die zeitlichen Messungen durch flexible glatte Funktionen erklärt werden, bis hin zu gewöhnlichen Differentialgleichungen, die mechanistisch eine Verbindung zwischen dem System und der Änderung des Systems über die Zeit modellieren. Die Erhöhung der Qualität und Quantität der gemessenen Zeitdaten schafft neue Forschungsmöglichkeiten. Oft sind jedoch die statistischen oder mathematischen Methoden, die zur Analyse verwendet werden, teilweise ungeeignet, diese neuen Forschungsfragen zu untersuchen, da sie z. B. nicht den vollen Informationsgehalt der Daten ausschöpfen. Die Anwendung einer ungeeigneten Methode kann dabei zu falschen Folgerungen oder Behauptungen führen und somit den Arbeitsablauf, bestehend aus Datengenerierung, Datenanalyse und Schlussfolgerungen, negativ beeinflussen. In der vorliegenden Arbeit betrachten wir die Analyse biologischer Systeme, für die bereits ein partielles Vorwissen besteht, aber auch einige Unbestimmtheiten über das System vorhanden sind. Dementsprechend kann das System mit den ver-fügbaren Methoden nicht ausreichend beschrieben werden.Wir zeigen neuartige Ansätze der Modellierung und ermöglichen eine bessere Beschreibung der unbestimmten Teile des Systems. Wir untersuchen drei Szenarien, in denen jeweils mit einem neuen statistischen Werkzeug zusätzliche Modellierungs-möglichkeiten und gleichzeitig die Unbestimmtheiten im System reduziert werden. Zunächst beschäftigen wir uns mit Datenmodellierung, indem wir einen Signifikanztest für den Unterschied zwischen zwei Gruppen von gepaarten zeitlaufgelösten Beobachtungen entwickeln. Als nächstes charakterisieren wir die mechanistische Kopplung von biologischen Netzwerken und führen zusätzliche latente Komponenten in das Netzwerk ein, falls die Notwendigkeit zur Netzwerkerweiterung identifiziert wird. Schließlich bauen wir auf unserem Netzwerkerweiterungsansatz auf und verallgemeinern die

Methode auf nichtlineare Erweite-rungen durch Katalyse im biologischen System. In jedem der drei Fälle entwickeln wir eine neue Methode, die entweder in der Lage ist, bestehende Methoden zu übertreffen, oder mit der wir zusätzliche Aspekte zur Regulierung und Zusammensetzung des untersuchten biologischen Systems aufzeigen. Jede Methode wird durch Anwendung auf zahlreiche Simulationsstudien bewertet. Darüber hinaus wenden wir die Methoden auf reale Datenbeispiele an. Dabei zeigen die Ergebnisse interessante und neue Erkenntnisse über die untersuchten Systeme auf.

Contents

1

Introduction

Statistics is the science which deals with modelling, analysis, classification, exploration, interpretation and visualization of data. Statistical tools are of general interest and are used to combine statistics with diverse scientific fields such as biology, chemistry, psychology, sociology, economics, medicine and many more. Especially in life sciences, statistics has become an unavoidable partner for drawing conclusions and making sense of the collected experimental data. Such conclusions aim to manifest propositions for a much broader picture than only the analysed data. Here, exactly the fundamentals of statistics come into play. One basic principal of statistics states that if the data is a valid sample of a general population, the conclusions drawn from its analysis are valid for the general population.

In the present thesis, we develop novel statistical methods to study biological data and thus couple both sciences. Undoubtedly biological studies were often conducted with the help of statistics on many occasions in history (Bliss [1970]; Cleland [1967]; Mather [1943]; May *et al.* [1976]; Pearl [1977]; Ptitsyn [1969]). Interestingly, biological experiments always have some fluctuations even when performed under the exact same conditions. This is very much in concordance with statistics where models are developed which gain knowledge from these fluctuations and at the same time put them in an interpretable context. Increasing availability of biological experiments and corresponding data opens up additional possibilities for the exploration of many novel biological phenomena.

In the ages of big data (Marx [2013]; Stephens *et al.* [2015]), it is possible to study

such phenomena in a way that was not possible due to e. g. financial limitations of data generation a decade ago. For example, the cost for sequencing a whole human genome is currently estimated to be approximately US$1,000 (Stephens *et al.* [2015]), which means that it was reduced by more than ten thousand times in the range of only ten years (Wetterstrand [2015]). At the same time, technical advances allow the generated data to be of higher accuracy and the process of data generation is speeded up enormously. With data generation at such pace, many additional experimental designs can be performed realistically nowadays. This involves not only the replication of a certain experiment but also several other aspects such as studying a fine temporal development of variables and thus better understanding the underlying mechanisms of the studied phenomenon. The generation of new types of data naturally calls for the development of new techniques for analysis.

New methods are also developed to improve existing ones. This holds true not only for experimental methods where technological advances allow measurement of new types of data but it definitely also holds true for statistical and mathematical methods. Improving an established method can be advantageous is many aspects such as computational time, precision or reliability. Improvement can further be motivated by different angles or perspectives of looking at the same project and the same data.

For a deeper understanding of biological systems, a statistician formulates a *model*, which serves as a tool for assessment of one or multiple hypotheses of interest. When formulating a model for a given biological process, one aims to use the model with respect to two things. First, the model should be able to be tuned to explain the available data. Second, once this tuning is achieved, one is interested in further characteristics which can be extracted from the model. Such characteristics are model prediction, model selection or hypothesis evaluation. All of these, however, are not of much help for drawing any conclusions if the first step – the formulation and tuning of a model – is unsuccessful. On the one hand, one of the pitfalls in modelling a complex biological system is choosing a model which is heavily tailored only towards the analysed data. On the other hand, a too simple model may fail to detect important data details. A nice quote, which is often attributed to Albert Einstein states ”Everything should be made as simple as possible, but no simpler”. This nicely illustrates the modelling dilemma.

Statistical or mathematical modelling of complex biological systems in a computational context is nowadays referred to as *systems biology*. An excellent overview over systems biology is available now for little over a decade (Kitano [2002a,b]). Especially in systems biology, where usually the data is of high complexity, model building or model formulation is a key responsibility for the statistician. In this field data arises from different species, such as genes, enzymes or proteins. They are organised in complex network structures which wire the different species together and aim to produce a broader picture of the studied system. Things get even more complex when the data arising from such a network is not recorded at a steady state of the network but is dependent on time. One then goes over to models which concern the dynamics of a biological system using this temporal data. As one example for temporal data consider measurements of the same subject over the course of a certain time period. These could be monthly height and weight measurements of newborn children in the first year of age. As a next example, consider a biological experiment in which cells are cultured for a certain amount of time and small parts of the cells are hourly extracted from the culture and protein expression is measured in the cells. However, once the cells are measured they cannot be returned to the cell culture and are lost for the further experiment. These two examples demonstrate that temporal data has many varieties which have to be taken into account with respect to analysis.

In this thesis, we rely on modelling temporal data in two ways. First, we approximate time courses of single biological species based on their raw observations as smooth functions of time using methods from the field of *functional data analysis* (Ramsay & Silverman [2005]). Such smooth functions have the advantage to be extremely flexible and thus are able to explain a large variety of time courses. Second, we also use *differential equations* (Coddington & Levinson [1955]; Ross [1984]) for assessing not only the mechanic coupling of time points within one species but also the dependency of several network species to each other. They can be of great use to gain detailed insights of the mechanistic nature of a studied biological network (Aldridge *et al.* [2006]). Differential equations create a relation between a function of time and its time-derivatives. The derivative represents the rate of change over time of the given species and thus is of special interest when gaining a detailed look on a biological system. Putting together both approaches of temporal variable modelling is one of the topics which was thoroughly investigated in this thesis. Both methods are depending on parameters,

such as basis coefficients, smoothing parameter, reaction rates or initial conditions. These parameters are calibrated to produce a model fit to a given data. This is called *parameter estimation*. Furthermore, if not only parameters of a model are to be estimated but rather the model itself is not fixed, one could consider several competing models which are applied on the data and the most appropriate one is then chosen as a final model. This process is called *model selection*. *Statistical hypothesis testing* in its most general formulation presents a further variant of model selection. Here, two (or several) competing hypotheses are made and finally one comes either to the conclusion that only one of the hypotheses is valid with very high probability or the conclusion reads that with the given data one cannot favour one of the competing hypothesis over the other. The single hypotheses can be seen as competing models which are most appropriate for the analysed data.

1.1 Scientific question

As already mentioned, temporal data with biological context is readily available and is used to investigate complex and possibly novel research questions. Such questions often appear extremely interesting from multiple different points of view. First, additional biological insight brings forward the whole scientific community as for example new drugs and therapies are developed based on the provided answers of these questions. Second, if a research question or scientific idea of explaining a certain phenomenon to which no or only insufficient methods exist is investigated, this naturally calls at least for improvement of available methods or even development of new ones. Therefore, from a statistical point of view, it is immensely exciting to develop, test and apply a novel method which helps in answering complex questions and thus drive forward the overall scientific progress. Obviously, the need for a *novel* method arises from the non-existence of appropriate methods in available literature. In situations where a given question is answered with the help of a (partly) inappropriate method, the credibility of the conclusions made by this analysis is at least questionable.

In this thesis, we pursue three main research questions in such situations where lack of available methods is leading to inability of finding answers.

First, we identified the need for development of a significance test for differences in paired time-resolved observations. In biological applications, often times pairing between different groups leads to a disagreement between method assumptions and analysed data. Although this problem is well-handled in literature if the analysed data is static rather than temporal (Fahrmeir *et al.* [2007b]; Student [1908]) and even some extensions to temporal data exist (Angelini *et al.* [2007]; Berk *et al.* [2011]; Crainiceanu *et al.* [2012]; Fahrmeir *et al.* [2007a]), the case of paired time-resolved observations presents a considerably larger complexity with no available method adequately satisfying these requirements. Clearly, the correct answer of this research question should incorporate the full information available in the data, such as time dependency or pairing.

Second, we explore the general question of systematic extension of biological networks based on temporal data. Biological networks (Girvan & Newman [2002]) connect different species such as genes, microRNAs or proteins and describe the communication pattern between these species. Network topology and its identification is a frequently studied research field (Coates *et al.* [2002]; Radicchi *et al.* [2004]; Zhou & Lu [2007]). Results from studying network formations can lead to novel insights of how different parts of a biological network communicate with each other and this in turn can be of great help for the understanding of biological processes. As already mentioned, the reliability, precision and even the size of such a network depend on the quality of the measured data. Therefore, with limited data sources only small networks may be identified reliable. As more and more data from the same biological system becomes available, it seems naturally that a network extension becomes desirable. With our approach, we target networks where no additional data and no prior information concerning network extensions is available. For such networks, we are able to identify additional nodes, which significantly improve the data explanation without producing an overfit. The implication of developing a tool which is able to answer the question of systematic network extension has the potential of driving the systems biology loop by generating novel and interesting hypotheses about the structure of a given biological system.

Finally, we investigate another aspect concerning the analysis of temporal data from biological networks as we investigate the modelling of catalysis in biological systems. Catalysis (Eisenmesser *et al.* [2005]; Masel *et al.* [2001]) is a non-linear

change of interaction intensity between two nodes of a network. Often times, although catalysis is present in a studied system, these interactions are still modelled linearly and catalysis is omitted. As we already described, often times a too simplistic model may fail to recognize important aspects of the data but at the same time the available data may be non-informative for the identification of catalysis. With our approach, we ask if there is a way of efficiently inferring catalytic reactions from large biological networks. Existing methods in the field (Guyon & Elisseeff [2003]; Rickert *et al.* [2013]) either oversimplify modelling of catalysis and thus fail to robustly detect catalytic reactions or they are not suitable due to extreme computational demand when considering all modelling possibilities. Finding a good compromise between both strategies will be one of the topics of this thesis.

With these three research questions, we aim to perform a work-flow which is emblematic for (temporal-based) statistical modelling of biological systems. First, we contribute an additional method which is able to investigate and summarize statistical aspects of the studied data. Next, we strengthen the possibilities of data modelling by introduction of a general method for systematic network extension. Finally, we concentrate on more specific aspects of data modelling with a specialised method for non-linear catalysis identification in biological networks. Hence, we are able answer questions on different levels, starting from general aspects of the data (data statistics) and finishing in specialised models (data modelling).

To answer these questions, we present several tools for statistical analysis of temporal data from biological systems. These tools, answer questions regarding parameter inference, model selection and statistical hypothesis testing. We successfully cope with typical problems which arise in biological data such as high noise level and low number of observations. This is done by combination and refinement of several existing methods from functional data analysis, differential equations modelling and statistical testing.

In summary, the aim of this thesis is to advance the arsenal of available statistical tools tailored for the analysis of temporal data arising from complex biological systems.

1.2 Overview

This thesis consists of six chapters. In the current introduction, we present the general motivation of our work, organise the thesis contents and state the scientific contributions in form of publications on which parts of the thesis are based.

In Chapter 2, we discuss several aspects on biological systems. Furthermore, we present the necessary statistical background information needed for understanding the developed methods in later chapters.

Following the first two general chapters, we continue with the development of novel methods for analysis of biological systems in the next three chapters. In each chapter we identified a situation in which no or no sufficient method is available to allow a proper analysis. Each of the three methods is first presented on the base of sound statistical techniques. It is then thoroughly tested on artificial data before being applied on real-world data and conclusions about the respective studied biological system are drawn and interpreted.

In Chapter 4, we develop a latent variable approach which aims to estimate hidden variables from network-structured temporal data. Identification of such hidden variables allows a systematic extension of the studied network and leads to formulation of novel biological hypotheses. The method is applied on protein data from the JAK-STAT signalling pathway.

Chapter 5 presents a novel approach for inference of catalytic reactions in biological systems. The method is again suitable for analysis of network-structured temporal data. It strongly reduces the computational effort for estimating catalyst candidates in a given network topology. The method is applied on data from the CD95 apoptotic signalling pathway.

In Chapter 3, we develop a further method for analysis of temporal biological data. Here, we focus on statistical hypothesis testing and present a novel test which answers the question whether two groups of paired time-resolved observations are significantly different. The test is applied on metabolomics data as well as chromatin data in the application examples.

Finally, in Chapter 6 we discuss the presented methods and applications. Moreover, we show several future research potentials as an outlook of this thesis.

1.3 Scientific contributions

The major scientific contributions discussed in this thesis are listed in the following.

- Creation of a new method which makes it possible to systematically extend ordinary differential equations by additional latent components and allows for causality statements on the basis of a combination of smooth function approximation and dynamical modelling.
- Novel method for a computationally efficient inference of catalysis in ordinary differential equations on the basis of combination of smooth function approximation, dynamical modelling and similarity analysis.
- Novel statistical test for assessing differences in two groups of temporal, paired observations.
- Analysis of several biological datasets - JAK-STAT signalling pathway; CD95 apoptosis pathway; SysMBo nutritional challenges, heterochromatin formation at retrotransposons - with the above-mentioned datasets and corresponding interpretation and discussion of results.
- Development and preparation of a software package which includes an implementation of the latent component identification method.
- Development and preparation of a software package which includes an implementation of the statistical test for assessing differences in two groups of temporal, paired observations.

Parts of these contributions were already published in peer-reviewed journals. Some parts of this thesis will therefore correspond to or be identical with these publications:

- **I. Kondofersky**, C. Fuchs, and F.J. Theis (2015). Identifying latent dynamic components in biological systems. IET Systems Biology, 9, 193–203.
- **I. Kondofersky**, F.J. Theis and C. Fuchs. Inferring catalysis in biological systems, *submitted*.

- **I. Kondofersky**, T. Brennauer, T. Erdmann, H. Hauner, F. J. Theis, C. Fuchs. Significance test for difference between paired temporal observations, *in preparation*.
- D. Sadic, K. Schmidt, S. Groh, **I. Kondofersky**, J. Ellwart, C. Fuchs, F.J. Theis, and G. Schotta (2015). Atrx promotes heterochromatin formation at retrotransposons. EMBO Rep., 16, 836-850.

At the beginning of each chapter, we explicitly indicate which publications are relevant for the chapter.

Further scientific contributions

Furthermore, the author of this thesis was involved in several other research projects, which were not directly connected to the main focus of the thesis. The findings in these projects were also published in peer-reviewed journals:

- S. Wahl, C. Holzapfel, Z. Yu, M. Breier, **I. Kondofersky**, C. Fuchs, P. Singmann, C. Prehn, J. Adamski, H. Grallert, T. Illig, R. Wang-Sattler, T. Reinehr (2013). Metabolomics reveals determinants of weight loss during lifestyle intervention in obese children. Metabolomics 9(6), 1157–1167.
- A. Chursov, S.J. Kopetzky, I. Leshchiner, **I. Kondofersky**, F.J. Theis, D. Frishman, A. Shneider (2012). Specific temperature-induced perturbations of secondary mRNA structures are associated with the cold-adapted temperature-sensitive phenotype of influenza A virus. RNA Biol. 9, 1266-1274.
- S. Vlaic, A. Hoppe, N.S. Mueller, S. Braun, L.A. D'Alessandro, S, Müller, R, Meyer, S, Bohl, **I. Kondofersky**, M.U. Muckenthaler, N. Gretz, F.J. Theis, R. Guthke, H.-G. Holzhütter, U. Klingmüller, M. Boerries, H. Busch. Systematic Analysis of Time-Resolved Transcriptional Signature of the Cross-Talk Between HGF and IL6 Reveals Genetic Program of Hepatocyte Proliferation Control, *submitted*.

2

Statistical modelling of biological systems

In this thesis, we model dynamical biological systems with mathematical and statistical methods. First, we will elaborate on the mathematical and statistical modelling of biological systems which give rise to temporal data. Specifically, we will discuss spline approximations as well as differential equations. Next, we will make the reader familiar with the studied biological systems. Along this line, we will review current literature in the context of biological systems corresponding to the central dogma of molecular biology. We will also introduce signalling pathways which structure the relationships between molecules in a biological system. Finally, we will introduce the concept of catalysis of chemical reactions in such systems.

2.1 Time series and dynamic systems

In this section we will introduce the mathematical and statistical background on which this thesis builds up. The data analysed in later chapters is of time-resolved nature. This means that for one observation, we have several measurements at different time points available. Dynamical systems giving rise to such temporal data can be modelled in various ways and depending of type on the studied phenomenon, length of the available time series and studied context different mod-

elling options are available. In the following, we will briefly introduce the notation and discuss the modelling systems used throughout this thesis.

2.1.1 Notation

We present the notation used throughout this thesis in Table 2.1.

Table 2.1: Notation

Exemplary symbol	Description
$\mathbf{x}$	vector
x_i	i-th vector element
$\mathbf{A}$	matrix
A_{ij}	i-th row and j-th column element of matrix $\mathbf{A}$
$\text{diag}(\mathbf{A})$	main diagonal of quadratic matrix $\mathbf{A}$
$f(x)$	function with argument x
t	time
$f(t)$	function of time
$f(\mathbf{t})$	function $f(t)$ evaluated at vector $\mathbf{t}$
$\mathbf{x}^T$	the transpose of vector $\mathbf{x}$
$\mathcal{X} = \{\dots\}$	set
$\mathcal{X} = \{\mathbf{x}_i\}_{i=1,\dots l}$	several vectors grouped in a set
$f_{\mathbb{N}}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	density function, normal distribution
$f_{\mathbb{LN}}(x \mid \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$	density function, log-normal distribution
$I(A)$	indicator function, equal to 1 if A is true, 0 otherwise

The methods we develop in this thesis can be grouped into *spline*-based methods and *differential equations*-based methods. As we will see in later chapters both methods can also be combined. We will now introduce and discuss both methods as well as discuss further methods which are suitable for modelling temporal data.

2.1.2 Splines

Research in approximation theory (Braess [2012]; Cheney & Lorentz [1980]; Rice [1969]) is focused on approximations of functions with the least possible error and control of this error. Hereby, the approximation of a function is often done

with polynomial functions which are attractive due to simple computation and numerical advantages. Particularly Chebyshev polynomials (Fox & Parker [1968]; Mason & Handscomb [2002]) are well-suited for approximation of functions with small errors. Hereby, one e. g. constructs a Chebyshev expansion to approximate a function $f(x)$: $f(x) \approx \sum_{i=0}^{\infty} c_i T_i(x)$ with c_i being coefficients which are calculated to obtain the lowest error of approximation and $T_i(x)$ are first kind Chebyshev polynomials. This sum is calculated up until the summand $c_n T_n(c)$ which gives the n-degree polynomial approximation. Research in approximation theory is further concerned with the properties, existence, uniqueness, convergence and optimality of such approximations. Closely connected to this research field is the study of spline functions. Spline functions also aim at approximating data points with the least possible error under certain conditions and they are defined in a similar way.

Suppose we have scalar observations at time points $t_0, \ldots, t_n$ which are ordered as $t_0 \leq \ldots t_j < t_{j+1} \leq t_n$. The idea of splines is to represent these temporal measurements with a smooth function. This smooth function should be chosen flexible enough to model the studied time series. Here, splines (in contrast to e. g. a high-degree polynomial representation of the time series) present possibilities to model a time series with a high degree of smoothness while maintaining a high stability. This is due to the spline being a piecewise polynomial function of a typically low degree and thus avoiding Runge's phenomenon (Runge [1901]) of high oscillation between data points which is typical for high-degree polynomials. One particular advantage of using a smooth function over the raw measurements for further analysis is that the smooth function can be evaluated at any time point $\mathbf{t} = (t_0, \ldots, t_n)$ and not only at the time points t_i where the measurements were made. Additionally, for large amounts of temporal data (large n) a dimension reduction is achieved because the dimensionality of the smooth curve is typically chosen much smaller than n. A spline is constructed to equal

$$x(t) = \sum_{k=1}^{K} \beta_k \phi_k(t). \tag{2.1}$$

Here, $\beta_k \in \mathbb{R}$ are the *basis coefficients*, $\phi_k(t)$ are the *basis functions* and K denotes the number of basis functions. More formally, $x(t)$ is a function in the space spanned by a linear combination of $\phi_k(t)$ or $x \in \text{span}\{\phi_k(t), k \in \{1, \ldots, K\}\} = \{\sum_{k=1}^{K} \beta_k \phi_k(t), k \in \{1, \ldots, K\}, \beta_1, \ldots, \beta_K \in \mathbb{R}\}$. $x(t)$ is called a spline if certain

properties of $\phi_k(t)$, which we introduce in the next paragraph, are fulfilled. Prominent examples for non-spline constructs of type (2.1) include (Ramsay & Silverman [2005]) monomial series where $\phi_k(t) = t^{k-1}$, constant series where $\phi_k(t) = 1$ as well as Fourier series where $\phi_k(t) = \sin(k\omega t)$ if k is even and $\phi_k(t) = \cos(k\omega t)$ if k is odd.

The first step of constructing a spline is dividing the range of $\mathbf{t}$ into $K+1$ subintervals. This is done by choosing a sequence of values $\tau_0, \ldots, \tau_{K+1}$ with $t_0 = \tau_0 \leq \tau_1 \leq \ldots \leq \tau_{K+1} = t_n$. τ_k are called *knots* and they are a monotone increasing sequence. Note, that all τ_k other than τ_0 and τ_{K+1} are not bounded to equal the measurement time points $\boldsymbol{t}$. We will give some guidance as of how to place these knots later in this chapter. After a sequence of knots is chosen, $x(t)$ as defined in (2.1) is called an order-M spline ($M > 1$) if two conditions are fulfilled (Hastie *et al.* [2009]):

1. Each $\phi_k(t)$ is a piecewise polynomial of order M with local support defined by the knot sequence $\tau_0, \ldots, \tau_{K+1}$.
2. $x(t)$ has $M-2$ continuous derivatives.

The second condition is automatically fulfilled for all t with $t \notin \{\tau_0, \ldots, \tau_{K+1}\}$ due to condition 1. At the knots τ_k where two polynomials join the second condition is not necessarily fulfilled. This is achieved by introducing constraints on the basis coefficients β_k which force adjacent polynomials to have equal values at the junction points.

More specifically, a spline curve of order M can be written in a simplified form as

$$x(t) = \begin{cases} p_0(t) \text{ if } t \in [\tau_0, \tau_1] \\ \vdots \\ p_K(t) \text{ if } t \in [\tau_K, \tau_{K+1}] \end{cases} \tag{2.2}$$

and the $p_0(t), \ldots, p_k(t)$ are recursively defined and weighted polynomials of order $M-1$ with a domain defined by the knot sequence $\tau_0, \ldots, \tau_{K+1}$. Polynomials have the property of continuous derivatives which means for the whole spline is infinitely differentiable at $t \notin \{\tau_0, \ldots, \tau_{K+1}\}$. However, additional care has to be taken at the junction points and this is achieved by introducing a further constraint. One requires additionally that function $x(t)$ and the derivatives up to order $M-2$

are equal at the junction points. If this is fulfilled, $x(t)$ is called an order M spline. We will give more details on how to construct the piecewise polynomials later.

The degrees of freedom of a spline equal the order of the polynomials plus the number of interior knots. For example, an order-4 spline defined over 10 intervals will have 13 total degrees of freedom. The special case of $M = 1$ which w. l. o. g. results in a stepwise function can only fulfil condition 1 and thus it is strictly speaking not a spline. In literature, however, the term of order-1 spline (Hastie *et al.* [2009]) is used.

Several different types or systems of spline functions exist. As Ramsay & Silverman [2005] state, the most prominent and widely used one which we also use throughout this thesis is the *B-spline* basis system which was first introduced by Schoenberg [1946] and Curry & Schoenberg [1947] and made prominent approximately a decade ago by De Boor [2001]. Other spline bases include natural splines or the truncated power system or M-splines. We direct the interested reader to Schumaker [2007] or De Boor [2001] for further information.

B-splines

Before defining B-splines, we first have to extend the original knot sequence $\tau_0, \ldots, \tau_{K+1}$ by $2M$ further knots which are placed at the boundaries of this sequence. The new sequence $\xi_1, \ldots, \xi_{K+2M}$ is defined as follows:

$$\begin{aligned} &\xi_1 = \xi_2 = \ldots = \xi_M = \tau_0 \\ &\xi_{k+M} = \tau_k, k = 1, \ldots, K \\ &\xi_{K+M+1} = \xi_{K+M+2} = \ldots = \xi_{K+2M} = \tau_{K+1}. \end{aligned} \tag{2.3}$$

Using this new sequence of knots, B-splines are defined recursively. Denoting $B_{k,M}(t)$ as the k-th B-spline basis function of order M, we start with order 1:

$$B_{k,1}(t) = \begin{cases} 1 & \text{if } \xi_k \leq t < \xi_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

for $k = 1, \ldots, K + 2M - 1$. The special case of $M = 1$ produces a stepwise constant

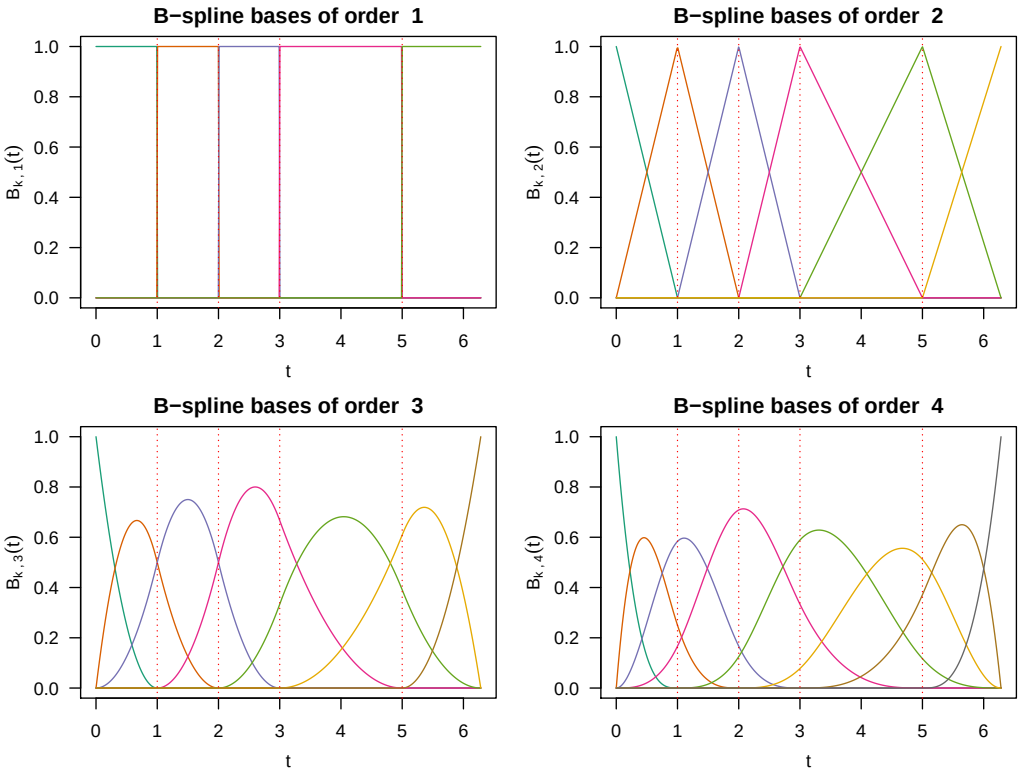

Figure 2.1: B-spline bases of different order. Dashed vertical lines mark the placement of inner knots. Each base $B_{k,M}(t)$ is non-zero over an interval spanned by $M+1$ knots.

function. Then

$$B_{k,M}(t) = \frac{t-\xi_k}{\xi_{k+M-1}-\xi_k} B_{k,M-1}(t) - \frac{\xi_{k+M}-t}{\xi_{k+M}-\xi_{k+1}} B_{k+1,M-1}(t) \qquad (2.5)$$

for $k = 1,\ldots,K+M$ fully recursively defines any order of B-splines. Note that by choosing an order-M B-spline with K interior knots, we need $M+K$ B-spline basis functions in (2.1).

Figure 2.1 illustrates B-spline bases of order 1 – 4. Here, we placed the knots at $\{0,1,2,3,5,2\pi\}$ and mark the inner knots with vertical dashed lines.

Translated to the spline definition in (2.1), we replace the basis functions $\phi_k(t)$ by B-spline bases $B_{k,M}(t)$ and omit the fixed order M in the notation. In Figure 2.2 these B-spline bases are used to approximate the sinus function based on 20 equidistant data points between 0 and 2π. Hereby, the basis coefficients β_k in

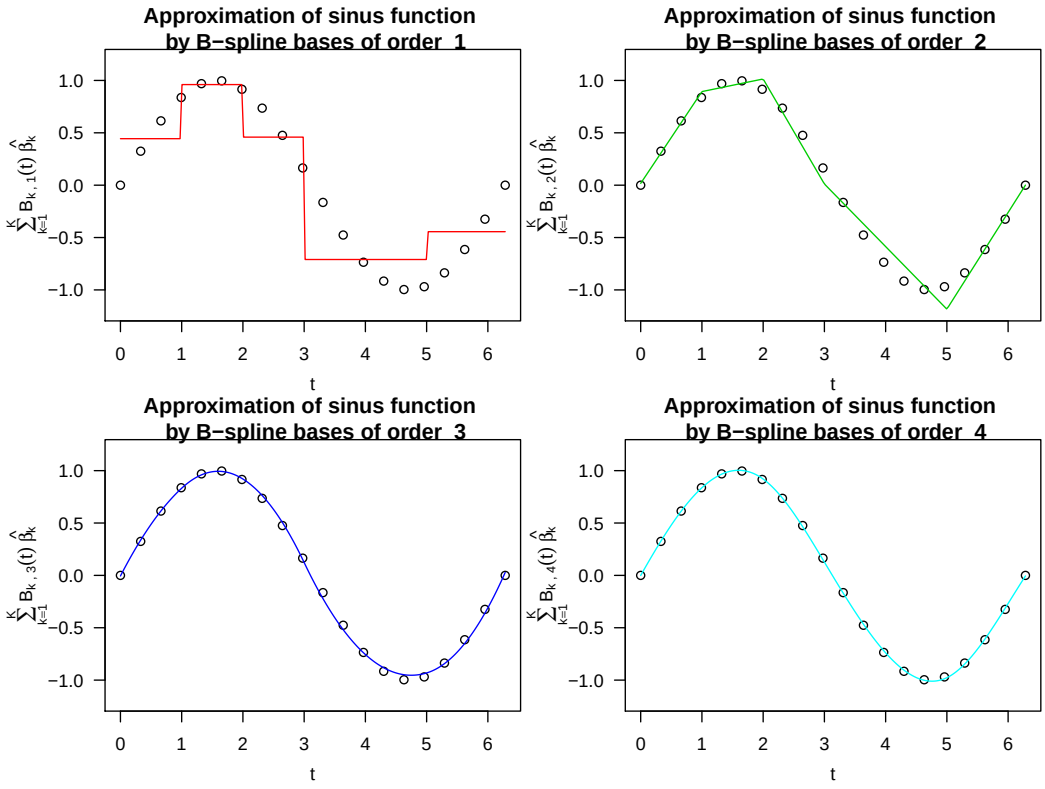

Figure 2.2: Approximation of the sinus function by B-spline bases of different order. Fitted functions are based on 20 equidistant data points between 0 and 2π of function $sin(t)$.

(2.1) are estimated with a standard least squares approach based on the 20 data points (we will discuss different ways of basis coefficients estimation in a few paragraphs). It is obvious that the lower the order M of the B-spline basis, the rougher the corresponding approximation. While B-splines of order 1 and 2 result in constant and piecewise linear approximations, respectively, order 3 and order 4 splines give smooth functions in the sense of at least one existing continuous derivative.

With (2.4) and (2.5) a system of B-spline bases is defined. Next, we will briefly discuss some properties which are important for the application of B-splines in this thesis.

B-splines are linearly independent in the vector space spanned by span$\{B_{k,M}(t), k \in \{1,\dots,K\}\} = \{\sum_{k=1}^{K} \beta_k B_{k,M}(t), k \in \{1,\dots,K\}, \beta_1,\dots,\beta_K \in \mathbb{R}\}$. This follows from the piecewise support definition and recursive formulation in (2.4) and (2.5) as

well as linear algebra where we note that $B_{k,M}(t) = B_{l,M}(t), \forall t$ if and only if $k = l$.

One very important property of B-splines with respect to computational time is called the compact support property (Ramsay & Silverman [2005]). The compact support property follows from (2.4) and (2.5) and states that each B-spline basis is equal to 0 over any interval which is outside of an interval spanned by $M+1$ adjacent knots or more formally $B_{k,M}(t) = 0 \; \forall t \notin \{\xi_k < t < \xi_{k+M}\}$. This means that the matrix of inner products of (2.1) will be sparse containing values on only M sub-diagonals to the left and right of the main diagonal. This, in turn, allows a fast computation of constructs where B-splines are used for function approximation.

The next interesting property (De Boor [2001]; Prochazkova [2005]) is shown by

$$\frac{\mathrm{d}B_{k,M}(t)}{\mathrm{d}t} = (M-1)\left(\frac{-B_{k+1,M-1}(t)}{\xi_{k+M} - \xi_{k+1}} - \frac{-B_{k,M-1}(t)}{\xi_{k+M-1} - \xi_k}\right). \tag{2.6}$$

This shows that the derivative of a B-spline of order M is a combination of B-splines of order $M-1$. Due to the recursive formulation of B-splines in (2.4) and (2.5) the lower order B-splines were already computed. Thus, the calculation of a B-spline derivative is practically not connected to any additional computational cost.

Another property of B-splines is the possibility of creating abruptly changing derivatives at certain time-points (Hastie *et al.* [2009]). This can be especially useful when studying realistic biological data where abrupt temporal changes are common after e. g. an external change of the studied system. The abrupt change in derivatives is achieved by duplicating knots. In general, if a knot is duplicated l times, this will lead to the $M-l$-th derivative to be discontinuous. This behaviour is exploited at the boundaries of the B-spline domain. In the extended sequence of knots as defined in (2.3) we introduce a duplication of M boundary knots. This means that at the boundaries the 0-th derivative or the smooth function itself is discontinuous. This is a desired property of B-splines which approximate a function only at the specified domain. Outside of this domain we do not want to model any behaviour of the approximated function and consider all modelling possibilities, even discontinuous functions.

Finally, other properties of B-splines include (Liu *et al.* [2014]):

- Positivity: $B_{k,M}(t) \geq 0, \quad t \in [t_0, t_n]$

- Unit decomposability: $\sum_k B_{k,M}(t) = 1, \quad t \in [t_0, t_n]$
- Symmetry: $B_{k,M}(t_n - t) = B_{k,K+M-k+1}(t), \quad t \in [0, t_n], k \in \{1, \ldots, K+M\}$

Knot placement

Several strategies on how to place the knots τ_k have been developed (Powell [1967]; Rice [1969]; Wold [1974]). In particular (Wold [1974]) state that one should place as few knots as possible to achieve a large dimension reduction. These few knots should be placed in such way that local extrema of a function lie approximately at the center between two adjacent knots and inflection points are close to the knots. However, these strategies work well only for a sample size of at least 30 – 40 points. In the targeted application sin this thesis, the number of measurements per observation is considerably lower (6 – 16). Additionally, Ramsay & Silverman [2005] discuss the possibilities of placing a knot at every j-th data point where j is an integer specified beforehand. A special case of this strategy is $j = 1$ resulting in *smoothing splines* which we will discuss later in this chapter. Finally, the most widely used strategy of knot placement which is implemented in many applications is to place the knots at equally spaced intervals so that $\tau_{k+1} - \tau_k = c$ with constant c. Choosing the knot sequence in a non-equidistant way may be advisable if the curvature of the approximated function is varying in different parts of the domain of t. Parts with low curvature are sufficiently approximated with few knots as opposed to parts with large curvature which are better approximated with a higher number of knots.

Estimation of basis coefficients

We introduced splines with the intention to approximate temporal data measurements with these flexible functions. In the following, we explain how spline theory is applied on such available measurements. Let $\boldsymbol{y} = (y_0, \ldots, y_n)^T$ denote the data observations collected at time point $t_0, \ldots, t_n$. Once a system of basis functions and other hyperparameters such as the knot sequence, number of basis functions K and the order of the basis functions are chosen, the basis coefficients $\beta_1, \ldots, \beta_K$ are calibrated using $\boldsymbol{y}$. Furthermore, for convenience, we rewrite (2.1) in matrix

notation to equal

$$x(t) = \sum_{k=1}^{K} \beta_k \phi_k(t) = \boldsymbol{\beta}^T \boldsymbol{\phi}(t). \tag{2.7}$$

Here, $\boldsymbol{\beta}$ is the vector of basis coefficients and $\boldsymbol{\phi}$ is the vector of basis functions. The evaluated basis functions at time points t_i, $\phi_k(t_i)$ can be stored in a $K \times (n+1)$ matrix which we denote by $\boldsymbol{\Phi}$. We now want to calculate an estimate for $x(t)$, which models the assumed data generating process which gives rise to the observations $\boldsymbol{y}$. Using the data, we can then estimate the coefficients $\boldsymbol{\beta}$ by minimizing

$$S(\boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_{i=0}^{n} (y_i - x(t_i))^2 = \sum_{i=0}^{n} \left(y_i - \sum_{k=1}^{K} \beta_k \phi_k(t_i) \right)^2 = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) \tag{2.8}$$

and thus using a least squares approach. As it is well known from linear model theory (Toutenburg [1992]) after differentiating (2.8) with respect to $\boldsymbol{\beta}$ and setting the derivative to $\mathbf{0}$, we arrive at the least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}. \tag{2.9}$$

The matrix $\boldsymbol{\Phi}^T \boldsymbol{\Phi}$ can be inverted as it is nonsingular because of the above-discussed property of linear independence of basis functions. Additionally, also borrowed from linear model theory, one could include a weighting matrix $\boldsymbol{W}$ into the estimation of $\boldsymbol{\beta}$. This is especially useful when dealing e. g. with nonstationary or autocorrelated errors (Ramsay & Silverman [2005]). The least squares criterion then changes to

$$S(\boldsymbol{\beta} \mid \boldsymbol{y})_{\boldsymbol{W}} = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) \tag{2.10}$$

and the estimator changes to

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = \left(\boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{y}. \tag{2.11}$$

$\boldsymbol{W}$ is a symmetric positive definite matrix which allows for unsymmetrical weighting of the contribution of the single error terms $y_i - \sum_{k=1}^{K} \beta_k \phi_k(t_i)$. This matrix can be chosen accordingly to e. g. measurement error associated with the studied data. For example, replicate observations at the same time points may be used by com-

puting the standard deviation of all replicates and then this standard deviation may be seen as an indicator for the amount of variability contained in the data for single measurements. This, in turn, may be translated to the weighting matrix $\boldsymbol{W}$. For the rest of this thesis, for the sake of notation simplicity, we will assume the weighting matrix to equal the identity matrix, $\boldsymbol{W} = \boldsymbol{I}$. This means that $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ from (2.11) equals $\hat{\boldsymbol{\beta}}$ from (2.9).

Calculating the function approximations at time points $t_0, \ldots, t_n$ is achieved by

$$\hat{x}(\boldsymbol{t}) = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.12}$$

with $\boldsymbol{H}$ called hat matrix or projection matrix. This matrix is used to calculate the *effective degrees of freedom* of the estimation process in (2.9) which equals

$$\mathrm{df} = \mathrm{trace}(\boldsymbol{H}). \tag{2.13}$$

This formulation allows the calculation of degrees of freedom in more complex scenarios such as penalization splines or smoothing splines which we will introduce shortly. In the case of a least squares fit as in (2.9) unsurprisingly the effective degrees of freedom equal K, the number of basis functions.

Choice of the number of basis functions

The choice of how many basis functions one should choose for the smooth estimation of temporal data is an important one. On the one hand, with large K the smooth approximation will fit the data very well with $\hat{x}(t)$ passing very close by or directly through the data points $\boldsymbol{y}$. However, this poses the danger of overfitting the data in the sense of fitting noise or unrealistic temporal variations. On the other hand, small K may lead to missing important data variability due to little to no flexibility of the estimated smooth function. This trade-off is well known and can be translated in the field of statistics as the *bias-variance trade off*.

The definitions of bias and variance in the context of our notation are

$$\mathrm{Bias}(\hat{x}(t)) = x(t) - \mathrm{E}\left(\hat{x}(t)\right), \tag{2.14}$$

$$\mathrm{Var}(\hat{x}(t)) = \mathrm{E}\left(\left(\hat{x}(t) - \mathrm{E}\left(\hat{x}(t)\right)\right)^2\right). \tag{2.15}$$

For large K typically (2.14) will be small and (2.15) will be large and the contrary statement holds when choosing a small K. Keeping both, bias and variance, acceptably small can be achieved by keeping a small mean squared error

$$\mathrm{MSE}(\hat{x}(t)) = \mathrm{Var}(\hat{x}(t)) + \mathrm{Bias}^2(\hat{x}(t)). \tag{2.16}$$

A good algorithm for choosing the number of basis functions K will lead to an overall low MSE. We formulated the smoothing approximation in the context of linear models in (2.8). In this context, increasing K by 1 leads to one additional coefficient which has to be estimated. Comparing this model with $K+1$ coefficients and the original one with K coefficients in terms of MSE leads to a model selection problem. Consequently, we can rely on the vast amounts of literature concerning model selection for linear models. Here, top-down methods will start with a large K and reduce it until either MSE cannot be reduced any more or, if MSE is not possible to be calculated, the smooth approximation still explains important variability features of the data. In contrast, bottom-up methods will start with a small K and increase it until the fit is not substantially improving. Combinations of both methods exist. However, all of these methods have their limitations and there is no gold standard in this case. One of the main challenges for variable selection is the discrete nature of K. Thus, in the next paragraphs we will introduce so called roughness penalties which are used for penalization and smoothing splines. This substantially decreases the problem of choosing the correct number of basis functions. For these methods the variability of $x(t)$ is controlled via a smoothing parameter and the number of basis functions merely has to be chosen sufficiently high. The estimation of this smoothing parameter is done e. g. with cross validation which acts as another way of controlling the bias-variance trade off.

2.1.3 Smoothing splines

As we already discussed the bias-variance trade-off is important for controlling the goodness of fit in terms of MSE of the function approximation. In addition to altering the number K of basis functions this can also be done by applying a penalization approach. This is done by introducing a roughness penalty and thus punishing too high data faithfulness of the approximated function. Roughness

penalties are widely known in statistics e. g. in the context of model selection. Prominent examples are ridge regression (Hoerl & Kennard [1970]), lasso (Tibshirani [1996]) and the combination of both, elastic net (Zou & Hastie [2005]). Before applying these ideas to the function approximation "roughness of a function" has to be defined. A good indicator of the variability of a function is given by its derivatives. The square of the second derivative of a function is called *curvature* (Ramsay & Silverman [2005]). It seems to be a natural indicator for roughness because the second derivative of a straight line (which is the smoothest functional approximation) is equal to 0 and thus it has a natural reference value. For this reason, a good quantifier of roughness is given by

$$\mathrm{PEN}_2(\boldsymbol{\phi}(t) \mid \boldsymbol{\beta}) = \int_t \left(\boldsymbol{\beta}^T \mathrm{D}^{(2)}(\boldsymbol{\phi}(s))\right)^2 ds. \tag{2.17}$$

The derivative operator $\mathrm{D}^{(m)}(f(t))$ denotes the componentwise derivative of function f with respect to its argument t. The notation PEN_2 with subscript is chosen to indicate that the *second* derivative is calculated. More generally, we can also formulate an equation PEN_m straightforwardly by changing the derivative operator to D^m in (2.17).

Using this result we can extend the least squares criterion (2.8) to include a penalization term:

$$S(\boldsymbol{\beta} \mid \boldsymbol{y}, \lambda) = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) + \lambda \mathrm{PEN}_2(\boldsymbol{\phi}(t)^T \boldsymbol{\beta}). \tag{2.18}$$

The parameter $\lambda \in \mathbb{R}_0^+$ is a non-negative scalar and is called *smoothing parameter*. It controls the flexibility of the approximated function and in general it holds that for $\lambda = 0$ Equation (2.8) and Equation (2.18) are equal whereas for $\lambda \to \infty$ the approximated function becomes a straight line with a zero valued second derivative. In the context of bias-variance trade-off, $\lambda \to 0$ will be associated a large variance in the sense of large curve variability and low bias as the approximated function will be close to the data points. On the other hand, $\lambda \to \infty$ will result in low variance and high bias as the approximated function will be almost a straight line.

One remarkable theorem formulated in De Boor [1972] states that the function that minimizes (2.18) is an order-4 spline as defined in (2.1) if the assumptions of existing second derivative and distinct sampling points $t_0, \ldots, t_n$ are fulfilled.

Before we can express the minimizer of (2.18) in matrix notation (2.17) has to be slightly transformed:

$$\begin{aligned}
\mathrm{PEN}_2(\boldsymbol{\phi}(t) \mid \boldsymbol{\beta}) &= \int_t \left(\boldsymbol{\beta}^T \mathrm{D}^2(\boldsymbol{\phi}(s))\right)^2 ds \\
&= \int_t \left(\boldsymbol{\beta}^T \mathrm{D}^2(\boldsymbol{\phi}(s)) \mathrm{D}^2(\boldsymbol{\phi}^T(s)) \boldsymbol{\beta}\right) ds \\
&= \boldsymbol{\beta}^T \int_t \left(\mathrm{D}^2(\boldsymbol{\phi}(s)) \mathrm{D}^2(\boldsymbol{\phi}^T(s))\right) ds \boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T \boldsymbol{R} \boldsymbol{\beta}. \qquad (2.19)
\end{aligned}$$

The matrix $\boldsymbol{R} = \int_t \left(\mathrm{D}^2(\boldsymbol{\phi}(s)) \mathrm{D}^2(\boldsymbol{\phi}^T(s))\right) ds$ depends only on the chosen system of basis functions. With this result, the analytical estimator for $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}}_\lambda = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}. \qquad (2.20)$$

The form of (2.19) has the same form as a ridge regression estimator for linear models. The hat matrix for such models is

$$\boldsymbol{H}_\lambda = \boldsymbol{\Phi} \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^T \qquad (2.21)$$

and the degrees of freedom can be calculated in the same way as in (2.13) to equal trace$(\boldsymbol{H}_\lambda)$. The subscript λ in (2.20) and (2.21) denotes that these are the basis coefficients estimator as well as the hat matrix for a previously defined smoothing parameter.

There exist different methods for choosing λ and we will discuss three which are the most commonly used in literature. The *cross validation* (CV) method and the *generalized cross validation* (GCV) method are discussed in Ramsay & Silverman [2005]. Alternatively, one can also use theory from mixture models (Wood [2000, 2004]; Wood & Augustin [2002]) for the estimation.

The general idea behind CV is to leave out some of the observations, then use the rest of the observations for model calibration and finally assess the performance of the calibrated model using the left-out observations. This is repeated until the prediction is not improved any more. Applied to smoothing spline approximation of time series data this means that we leave out a number $m < n+1$ from the observations $x_0, \ldots, x_n$, then fit a smoothing spline to the rest of the observations with a fixed starting smoothing parameter λ and finally evaluate the spline at the m

time points, corresponding to the left-out observations. A loss function such as the sum of squared residuals is then used to calculate the discrepancy between the approximation and the real data. This process is repeated for different λ parameters until no reduction of the loss function is achieved. Choosing $m = 1$ is called leave-one-out CV and this is the only variant where no randomness is included in the estimation of λ. However, his variant of CV has two downsides. First, the computational demand is large because a smoothing spline has to be estimated $n+1$ times for a single λ parameter and this might become infeasible for large n. Second, the variant tends to undersmooth the data (Ramsay & Silverman [2005]). Choosing $m > 1$ leads to randomness in the estimation of λ due to the random splitting of the two groups of observations. Usually this randomness is tolerable compared to the computational gain and to the more acceptable degree of smoothing. In practice and as recommended in literature (Kohavi [1995]) the usage of $m \approx \frac{n}{10}$, also called ten-fold CV, is recommended. In the applications discussed in this thesis, we used ten-fold CV wherever possible to estimate λ. In most real-world data scenarios time series were short consisting of 6 to 16 temporal snapshots. In these cases we generally applied leave-one-out CV.

GCV was introduced by Craven & Wahba [1979] as a simpler version of CV which is computationally attractive due to the need for the repeated re-smoothing of the smoothing splines being avoided. The GCV criterion can be expressed as

$$GCV(\lambda) = \left(\frac{n}{n - \text{trace}(\boldsymbol{H}_\lambda)}\right)\left(\frac{L(\boldsymbol{\beta} \mid \boldsymbol{y})}{n - \text{trace}(\boldsymbol{H}_\lambda)}\right) \tag{2.22}$$

with the unpenalized criterion $L(\boldsymbol{\beta} \mid \boldsymbol{y})$ as defined in (2.8). Obviously, (2.22) is only defined and makes sense if $n > \text{trace}(\boldsymbol{H}_\lambda)$. We first note that $\text{trace}(\boldsymbol{H}_\lambda) \leq \text{trace}(\boldsymbol{H}) = K$, which is obvious from Equation (2.13) and Equation (2.21). Additionally, it is reasonable to choose $K < n$ which automatically guarantees the condition $n > \text{trace}(\boldsymbol{H}_\lambda)$. We further note that the case of $K \geq n$ would result in a highly overfitting curve with an exact fit for every measurement. For that reason the case of $n > \text{trace}(\boldsymbol{H}_\lambda)$ can be assumed w. l. o. g. Although GCV is computationally more efficient than CV, it still has to be computed on a large number of λ values to find the optimal one. These values can be ordered on a simple grid or proposed by a numerical optimization. Further ideas of how to speed up this process are discussed in Ramsay & Silverman [2005].

As a further alternative for the estimation of λ, theory from linear mixed models (McCulloch & Neuhaus [2001]) can be used. To that end we first make the assumption of normally distributed basis coefficients

$$\boldsymbol{\beta} \sim \mathbb{N}(\mathbf{0}, \tau^2 \boldsymbol{R}^{-1}) \tag{2.23}$$

where $\boldsymbol{R}$ is the same as in (2.19) and again only depends on the chosen system of basis functions.

The distribution assumption in (2.23) is in contrast to the previous definitions of $\boldsymbol{\beta}$ where we assumed a fixed coefficient vector. This can be related to a Bayesian modelling perspective where parameters of a model are not assumed to be fixed but rather random variables with distributions. More detailed, in Bayesian theory one first assumes a prior distribution for the parameters and using Bayes' theorem and the data likelihood one arrives at a posterior distribution of the parameters. Only few cases exist where the posterior distribution can be assessed analytically. Therefore, for all other cases, sampling methods such as Markov chain Monte Carlo (MCMC) sampling are used to approximate the posterior distribution of the parameters. As Bayesian estimation is not in the focus of this thesis, we point the interested reader to relevant literature (Lee [2012]; Raftery *et al.* [1992]). For the rest of this thesis, we still consider the basis coefficients to be fixed and not have a distribution and only make this assumption for the following comparison with mixed models.

W. l. o. g., we assume that $\boldsymbol{R}$ is invertible. For strategies of handling a possible singularity of $\boldsymbol{R}$, see Fahrmeir *et al.* [2007a]. Then, we can consider a formulation of the log-likelihood of a linear mixed model as

$$\log L(\boldsymbol{\beta} \mid \boldsymbol{y}) = -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{\Phi\beta})^T (\boldsymbol{y} - \boldsymbol{\Phi\beta}) - \frac{1}{2\tau^2} \boldsymbol{\beta}^T \boldsymbol{R\beta} \tag{2.24}$$

with σ^2 the finite noise variance of the linear model part of the linear mixed model. The estimation of σ^2 and τ^2 is done computationally efficient using a restricted maximum likelihood approach (ReML). These estimates can then be directly translated to an estimation of λ in the form $\lambda = \frac{\sigma^2}{\tau^2}$ as discussed in Fahrmeir *et al.* [2007a].

Finally, we would also like to mention that further possibilities for estimation of the smoothing parameter exist such as bootstrap methods or model selection criteria

such as AIC or BIC measures.

2.1.4 Differential Equations

In biology differential equations represent a prominent tool for the exploration of functional behaviour of species. As the name already suggests they relate a function to its derivatives. Prominent examples in biology where differential equations are used include e. g. the Lotka-Volterra equations for the relationship between predators and prey (Lotka [1910]; Volterra [1928]). Many types of differential equations exist such as ordinary differential equations (ODE), delayed differential equations (DDE), partial differential equations (PDE) and stochastic differential equations (SDE) (Dargatz [2010]; Kuang [1993]; Michiels & Niculescu [2014]; Pons [1955]; Ross [1984]).

In the course of this thesis we focus methods and applications on ODE. For that reason, we now briefly discuss definition and properties of this type of differential equations. Let $x = x(t)$ be a function which is m times differentiable with respect to time. An m-th order ODE can be written as

$$\Psi\left(t, x, \frac{\mathrm{d}x}{\mathrm{d}t}, \ldots, \frac{\mathrm{d}^m x}{\mathrm{d}t^m}\right) = 0. \tag{2.25}$$

In this notation Ψ is a real function with $m+2$ arguments $t, x, \frac{\mathrm{d}x}{\mathrm{d}t}, \ldots, \frac{\mathrm{d}^m x}{\mathrm{d}t^m}$. x and its derivatives are functions of t called dependent variables and t is called the independent variable and usually represents the time. Similarly, a linear ordinary differential equation that can be expressed in

$$a_0(t)\frac{\mathrm{d}^m x}{\mathrm{d}t^m} + a_1(t)\frac{\mathrm{d}^{m-1} x}{\mathrm{d}t^{m-1}} + \cdots + a_{m-1}(t)\frac{\mathrm{d}x}{\mathrm{d}t} + a_m(t)x = b(t). \tag{2.26}$$

where $a_i(t)$ as well as $b(t)$ are functions which depend only on the independent variable t and where $a_0(t) \neq 0$. The applications in this thesis will mostly use linear ODE as models of temporal data. Moreover, we will focus on systems of first-order linear ODE functions which can be written as

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \psi_j(x_1, \ldots, x_N, t, \boldsymbol{\theta}) \tag{2.27}$$

for $j = 1,\ldots,N$ and ψ_i functions with $N+1$ arguments and $\boldsymbol{\theta} \in \mathbb{R}^p$ denoting a p-dimensional parameter vector. x_j is short notation for $x_j(t)$ and denotes the j-th function of t. With such functions it is possible to describe the dynamics of various biological processes involving N different species, such as genes, proteins or enzymes.

Another notation, which is well in conformity with modelling of chemical reactions is based on stoichiometry which is used to relate reactants and products to each other. The notation goes as follows:

$$\frac{d\boldsymbol{x}(t)}{dt} = \dot{\boldsymbol{x}}(t) = \boldsymbol{S}\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{\theta}) = \sum_{g=1}^{m} \boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{\theta}) \tag{2.28}$$

with $N \times m$ stoichiometry matrix $\boldsymbol{S}$, m-dimensional flux function $\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{\theta})$ with arguments $\boldsymbol{x}(t) = (x_1(t),\ldots,x_N(t))^T \in \mathbb{R}^N_{\geq 0}$ as the non-negative network component concentration functions and $\boldsymbol{\theta} \in \mathbb{R}^p$.

Let us consider two examples in the following which will nicely illustrate the concept of stoichiometry modelling. Consider the dimerization reaction

$$A + B \xrightarrow{\theta_1} C.$$

In other words, the reactants A and B react to form the product C. In this case the stoichiometry matrix S equals $(-1,-1,1)^T$ and the one-dimensional flux function equals $v = \theta_1 AB$. As a second second example, consider the following system consisting of 6 reactions of 5 species:

$$\begin{aligned} 2A &\underset{\theta_2}{\overset{\theta_1}{\rightleftharpoons}} B \\ B + C &\underset{\theta_4}{\overset{\theta_3}{\rightleftharpoons}} D \\ D &\xrightarrow{\theta_5} E \\ E &\xrightarrow{\theta_6} \emptyset. \end{aligned}$$

The corresponding stoichiometry matrix equals

$$\begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

and flux function $\boldsymbol{v} = (\theta_1 A^2, \theta_2 B, \theta_3 BC, \theta_4 D, \theta_5 D, \theta_6 E)^T$.

Finding the solution of (2.25), (2.27) and (2.28) is the main focus of ODE analysis. Most textbooks on ODE address the two most important properties of ODE solutions – existence and uniqueness. We will not go into detail and further discuss those aspects but rather point the interested reader to relevant literature (Coddington & Levinson [1955]; Gear [1971]; Ross [1980]). In general, an m-th order ODE has m linearly independent solutions. Only few ODE have an exact solution which can be calculated analytically. In most cases if solutions exist and are unique, they are found numerically. A large amount of numerical algorithms exist in literature (Hull *et al.* [1972]). Prominent examples for such algorithms are the family of Runge-Kutta solvers discussed e. g. in Butcher [1987]. In the statistical software `R` R Development Core Team [2011], one can for example use the package `deSolve` (Soetaert *et al.* [2010]) where a large variety of ODE solvers are implemented efficiently.

The unknown parameters in (2.27) are $\boldsymbol{\theta}$ and the initial conditions $\boldsymbol{x}(t_0)$. If those parameters are specified, the complete dynamics of $\boldsymbol{x}(t)$ can be described. For applications, it is possible to extract these parameters from common knowledge or literature. Generally parameter estimation is usually involved in the calibration of ODE systems.

Parameter estimation for ODE

Many possibilities exist for calibration of ODE systems. We consider observations $x_j^{\text{obs}}(t_i)$ where $i = 0, \ldots, n$ denote $n+1$ observations per species $x_j(t)$ and $j = 1, \ldots, N$ are N species. Probably the most intuitive way of calibrating an ODE system to fit these observations is to apply a least squares approach where we

estimate $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^D}{\operatorname{argmin}} \sum_{i=0}^{n}\sum_{j=1}^{N} (x_j^{\text{ode}}(t_i \mid \boldsymbol{\theta}) - x_j^{\text{obs}}(t_i))^2. \tag{2.29}$$

$x_j^{\text{ode}}(t_i \mid \boldsymbol{\theta})$ is the evaluation of the (numerically) solved ODE system with parameter vector $\boldsymbol{\theta}$ at t_i for $x_j(t)$. The least squares approach yields an estimate $\hat{\boldsymbol{\theta}}$. Finding solutions of least squares estimates is well covered in Björck [1996] or Marquardt [1963]. These estimates coincide with the estimates produced by maximum likelihood optimization if one makes an assumption of normally independent and identically distributed (iid) observation noise

$$x_j^{\text{obs}}(t_i) = x_j(t_i) + \varepsilon_{i,j}, \qquad \varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2) \tag{2.30}$$

for all $i \in \{1,\dots,N\}$ and $j \in \{0,\dots,n\}$. Here, the finite and positive parameter σ^2 is the variance of the noise terms $\varepsilon_{i,j}$ and $x_j(t_i)$ the true but unknown data generating process. Making this assumption allows us to estimate $\boldsymbol{\theta}$ with a maximum likelihood approach. The corresponding log-likelihood function is then

$$l(\boldsymbol{\theta},\sigma^2 \mid \boldsymbol{x}^{\text{ode}}(\boldsymbol{t})) = C - (n+1)N\log\sigma - \sum_{i=0}^{n}\sum_{j=1}^{N} \frac{\left(x_j^{\text{obs}}(t_i) - x_j^{\text{ode}}(t_i \mid \boldsymbol{\theta})\right)^2}{2\sigma^2}. \tag{2.31}$$

In this notation, the constant C depends only on the fixed quantities N and n. Maximization of (2.31) leads to estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$.

In (2.30) we make the assumption of normally distributed errors. Although this a common assumption widely made in literature, for real biological systems it is at least debatable and we can state two problems. First, when dealing with e. g. protein concentrations, one has measurements in $\mathbb{R}_0^+$. In this case, (2.30) is ill-defined. Second, the measurement noise is independent of the value of the measurements which is again unrealistic for biological studies where measurement noise is expected higher for larger concentrations. Both problems can be met by altering the noise terms to contribute in a multiplicative way:

$$x_j^{\text{obs}}(t_i) = x_j(t_i)\cdot\varepsilon_{i,j}, \qquad \varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathrm{LN}(\frac{-\sigma^2}{2},\sigma^2). \tag{2.32}$$

The parameters of the log-normal distribution are chosen so that $\mathrm{E}(x_j^{\mathrm{obs}}(t_i)) = x_j(t_i)$. This formulation amounts in a different log-likelihood function:

$$l(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{x}^{\mathrm{ode}}(\boldsymbol{t})) = C - (n+1)N \log \sigma - \sum_{i=0}^{n} \sum_{j=1}^{N} \frac{\left(\log x_j^{\mathrm{obs}}(t_i) - \log x_j^{\mathrm{ode}}(t_i \mid \boldsymbol{\theta}) + \frac{\sigma^2}{2}\right)^2}{2\sigma^2}. \tag{2.33}$$

Maximization of (2.33) again leads to estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$. The specification of different noise models can be also further extended to e. g. shot noise (Poisson-distributed errors, Nagaev [1995]).

Other possibilities of estimating parameters in ODE exist. For example Bayesian parameter estimation has evolved greatly with the availability of better computational resources during the past decades (Girolami [2008]; Lawrence [2010]). Especially when the modelled ODE system is underdetermined in the sense of more parameters than observables a Bayesian model can be of great help for parameter identifiability. Here, profile likelihood approaches (Kreutz *et al.* [2013]; Raue *et al.* [2009]) allow the study of marginal distributions of parameters and thus recognize identifiable parameters from partially identifiable or non-identifiable ones. Bayesian methods in general rely on Markov chain Monte Carlo (MCMC) sampling for approximation of parameter distributions. Prominent MCMC based methods for parameter estimation include Metropolis Hastings algorithm (Metropolis *et al.* [1953]) or Gibbs sampling (Geman & Geman [1984]). Another approach for parameter estimation is presented in multiple shooting (Peifer & Timmer [2007]), cross-entropy (Wang & Enright [2013]) or regression models (Brunel *et al.* [2008]).

Observability of ODE systems

Until now we assumed that all species x_i in the ODEs are *directly* observed and that *all* of them are observed. In this section, we relax this assumption. First, we now consider the number M of observed time courses to be smaller than N. Second, we allow the observed time courses $y_1, \ldots, y_M$ to be affine linear transformations of $x_1, \ldots, x_N$. W. l. o. g. we can then write $y_j(t) = \sum_{l=1}^{N} A_{l,j} x_l(t) + b_j$ for all $j = 1, \ldots, M$ with scalar constants $A_{l,j}$ and b_j collected in the $N \times M$-dimensional matrix $\boldsymbol{A}$ and the M-dimensional vector $\boldsymbol{b}$, respectively.

In this case, parameter estimation of ODE models can still be performed by optimizing one of (2.29), (2.31) or (2.33). However, some alterations to the ODE system have to be performed. First, (2.27) has to be adapted to the new observables and therefore can now be formulated as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f_j(y_1,\ldots,y_M,x_1,\ldots,x_{N-M},t,\boldsymbol{\theta}^\star) \tag{2.34}$$

for $j = 1,\ldots,N$. The new parameter vector $\boldsymbol{\theta}^\star$ extends the old parameter vector $\boldsymbol{\theta}$ by the new parameters $\boldsymbol{A}$ and $\boldsymbol{b}$. Formulation of (2.34) can be derived analytically for affine linear combinations by making use of the chain rule. Second, the loss function which is optimized for parameter estimation ((2.29), (2.31) or (2.33)) has to be adapted. For example using normally distributed additive measurement noise, the corresponding log-likelihood function version is formulated as

$$\begin{aligned} l(\boldsymbol{\theta},\boldsymbol{A},\boldsymbol{b},\sigma^2 \mid \boldsymbol{x}(\boldsymbol{t}),\boldsymbol{y}\mathrm{ode}(\boldsymbol{t})) = C - (n+1)M\log\sigma \\ - \sum_{i=0}^{n}\sum_{j=1}^{M} \frac{\left(y_j^{\mathrm{obs}}(t_i) - y_j^{\mathrm{ode}}(t_i \mid \boldsymbol{\theta},\boldsymbol{A},\boldsymbol{b})\right)^2}{2\sigma^2}. \end{aligned} \tag{2.35}$$

with $y_j^{\mathrm{ode}}(t_i \mid \boldsymbol{\theta},\boldsymbol{A},\boldsymbol{b})$ denoting the solution of the j-th equation of the ODE system evaluated at t_i for parameters $\boldsymbol{\theta}$, $\boldsymbol{A}$ and $\boldsymbol{b}$.

Overall, we arrive at parameter estimates by again optimizing a likelihood function. As M is smaller than N these estimates will be typically less accurate, especially if $M \ll N$. Furthermore, if the dimension of estimated parameters is larger than M, then non-identifiability issues will arise for some parameters. This can be handled by e. g. introducing literature-derived constraints on parameters. Details on parameter identifiability in ODE systems are discussed in Raue *et al.* [2009, 2010, 2013].

2.2 Biological systems

In the second part of this background chapter, we will briefly summarize the essentials of state of the art molecular biology, molecule structuring in signalling pathways and catalysis.

2.2.1 Molecular biology

In this thesis we will develop methods for analysis of biological data. More precisely, the data that we analyse in several applications can be classified as coming from the field of molecular biology. As the term suggests, molecular biology is the branch of the science biology which studies biological systems on molecular level. This includes the interaction of cellular systems in terms of their DNA, RNA and proteins. As postulated in the central dogma of molecular biology (Crick *et al.* [1970]; Crick [1958]), these are the key parts of a biological system which transfer genetic information in cells and thus are responsible for various processes such as cell division, cell growth or cell death. Hereby, the three processes which occur in most cells are DNA replication (DNA is copied into DNA), transcription (DNA is copied into mRNA) and translation (synthesis of proteins is directed by mRNA). It is therefore obvious that information flow within a cell is a complex process which occurs at many different scales. Understanding the various processes in a cell and thus understanding e. g. mechanisms of a certain disease evolution leads to understanding this information flow. Figure 2.3 (adapted from Ritchie *et al.* [2015]) schematically puts this into context as it shows the information transduction from DNA level ultimately to the forming of a phenotype such as different diseases or metabolic syndromes. Further details on this topic are given in Alberts *et al.* [1995]; Fasman *et al.* [1977]; Watson *et al.* [1970]. A huge amount of data has been generated in recent years (Marx [2013]) on all of the above-described scales (Ritchie *et al.* [2015]). Generating this data from single cells rather than from bulks of cells is becoming more and more available (Shapiro *et al.* [2013]). Development of appropriate methods which are able to cope with the vast amount of data is another important challenge which the scientific community is facing.

This thesis will cover application examples using proteomics and metabolomics data. Let us therefore give a rough classification of where this data stands in the process of cell information flow. Consider Figure 2.3, where the two omics are put into context with respect to other omics types such as genomics and transcriptomics. The proteome (Gooley *et al.* [1996]) is a collection of all proteins produced by an organism. Proteins are large molecules consisting of amino acid chains and they have several important cell functions. Interaction patterns between proteins are intensely studied in systems biology. Proteins are also the key players

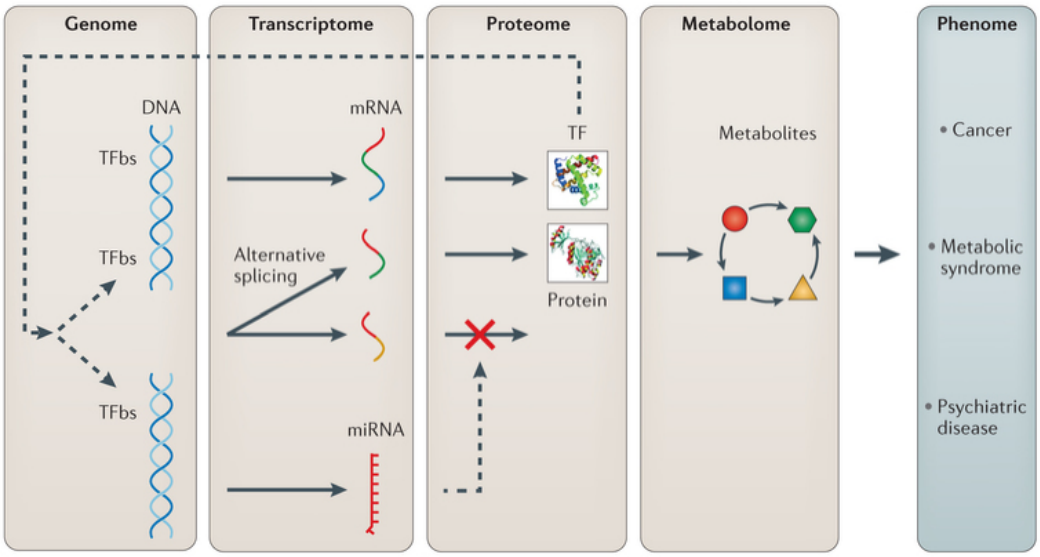

Figure 2.3: Illustration of several different omics types. DNA (stored in the genome) is transcribed to RNA (stored in the transcriptome) and RNA is translated to proteins (stored in the proteome). Proteins can act as transcription factors (TF) and activate DNA or further function as metabolites which e. g. can be associated with different phenotypes. The information flow from DNA to phenotype is indicated by arrows. Figure is adapted from Ritchie *et al.* [2015].

in cell signalling pathways (see chapter 2.2.2) as well as in catalysis (see chapter 2.2.3) of biochemical reactions in the cell. The metabolome is a collection of all small molecules in a cell. Metabolites are much smaller than e. g. proteins or DNA samples (Wishart [2007]). Examples for metabolites include alcohols, amino acids, antioxidants, vitamins or nucleotides. The study of metabolites is called metabolomics (Shulaev [2006]). It is one of the cornerstones of systems biology (alongside with genomics, transcriptomics and proteomics) and has received widespread attention in many different applications such as drug discovery (Kell [2006]), clinical toxicology (Nicholson *et al.* [2002]) and nutritional genomics (Gibney *et al.* [2005]; Trujillo *et al.* [2006]). Metabolites are measured by standard techniques used e. g. in chemistry, such as nuclear magnetic resonance or mass spectrometry. The generated and analysed data in metabolomics typically involves measurements performed on subjects under different conditions. This is then stored in a matrix with columns corresponding to the single measured metabolites and rows corresponding to the single subjects.

Relationships between proteins as well as between metabolites are often explored

using pathways. In the next section we will present some pathway examples as well as briefly review pathway properties.

2.2.2 Signalling and metabolic pathways

A signalling pathway (also called biochemical cascade) describes a series of biochemical reactions of molecules such as proteins or enzymes which cooperate to control cell functions such as division or death. For a detailed explanation of this process we can point the interested reader to the relevant literature (Lodish *et al.* [2000]). In brief, the first signal is delivered by an extracellular signalling molecule (ligand) which binds to an extracellular receptor located on the cell surface. This receptor is a transmembrane protein which has one part located outside the cell and the other inside the cell. After extracellular binding of the ligand to the receptor, the inside part of the receptor is changed. More specifically, this creates a binding site for intracellular signalling proteins and triggers a set of chemical reactions. Ultimately this leads to a certain response of the cell. The whole process is also called signal transduction and the corresponding molecules which are part of the signal transfer are organized in a signal transduction network. Depending on the activating molecule, cell and activated signal transduction network, the response can be e. g. an alteration of the shape of the cell or its ability to divide (Krauss [2006]).

Metabolic pathways are required for the cell stability and cell structure conservation. They also represent a series of chemical reactions which alter the initial metabolite. The end product of such pathways can be another metabolite which serves as an initiator of another cascade of chemical reactions. This connection between several metabolic pathways can then be organized into large metabolic networks (Jeong *et al.* [2000]). In nutritional science these can then e. g. be used to identify metabolites associated with diet or study metabolite effects on diet-disease relations (Guertin *et al.* [2014]).

Important signalling pathways include the MAPK/ERK pathway which has thoroughly been studied in cancer research (McCain [2013]; Roberts & Der [2007]) as well as the JAK-STAT signalling pathway (Arbouzova & Zeidler [2006]; Horvath [2000]; Rawlings *et al.* [2004]) which is essential for differentiation and growth of erythroid progenitor cells. A canonical representation of this pathway is shown

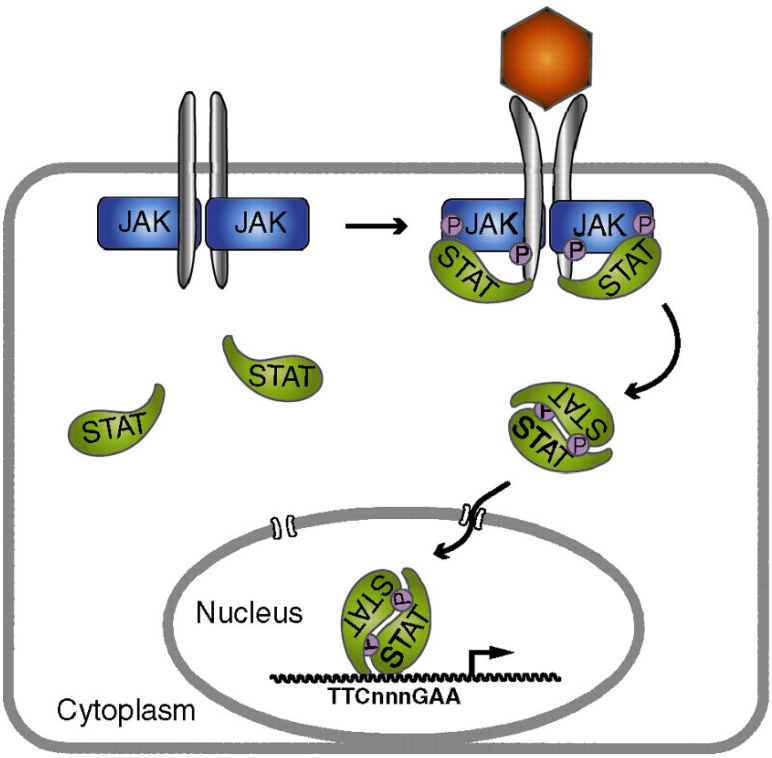

Figure 2.4: Illustration of JAK-STAT signalling pathway. Extracellular receptor activation leads to phosphorylation of JAK and provides docking sites for STAT. Next, STAT is phosphorylated and dimerized. Finally, it is translocated to the nucleus where it can bind to DNA sequences and thus drive important processes such as cell division or cell death. This signalling cascade is indicated by the arrows. Figure is adapted from Arbouzova & Zeidler [2006].

in Figure 2.4. Here, the process of extracellular receptor activation, phosphorylation of JAK and binding of STAT molecules followed by phosphorylation and dimerization of STAT and subsequent translocation to the nucleus is indicated by the black arrows. We will analyse data from the JAK/STAT pathway in Chapter 4.

Apoptosis pathways (Elmore [2007]) represent an additional class of signalling pathways which regulate a programmed cell death. In Chapter 5 we will study data from the cluster of differentiation 95 (CD95) apoptosis pathway (Huang *et al.* [1996]; Lavrik *et al.* [2007]).

Signalling and metabolic pathways usually are comprised of different recurring network motifs such as (negative or positive) feedback (discussed in Chapter 4), catalysis (discussed in Chapter 5) or cross-talks (Alon [2007]; Ashkenasy *et al.* [2004]; Cao *et al.* [2015]; Donaldson & Calder [2010]; Masoudi-Nejad *et al.*

[2012]). In the following, we will discuss one of these, catalysis, in more detail.

2.2.3 Catalysis

Until now we several times mentioned biochemical reactions between two species. Whenever such a reaction occurs, a set of some (chemical) substances, called reactants, is converted to another set of (chemical) substances, called products. The reaction rate is the speed at which such a chemical reaction is happening. The rate of a chemical reaction is characteristic at given atmospheric conditions as well as reactant concentrations. For example iron rusting can be characterized as a very slow reaction which happens over a large interval of time as compared to burn of glucose when sugar is metabolised to energy which is very fast as it happens in a fraction of seconds. For the occurrence of such processes energy plays a major role. The reactants of a chemical reaction are activated by either adding free energy (e. g. in form of heat) or spontaneously in the direction of a lower and more stable energy state. Hereby, the energy of a reaction is first increased to reach a transition state after which the energy starts decreasing and reaches the energy state of the products. The transition state of a chemical reaction is the point at which the energy of a reaction is highest.

Reaction rates can be changed when an additional (chemical) substance, called catalyst is present. In literature one divides catalysts into two categories. On the one hand, a reaction rate is increased by a catalyst. On the other hand, a reaction rate is decreased by an inhibitor. However, in the course of this thesis we will not distinguish between both terms and use catalyst as a collective word for both directions of change of reaction rates. Technically, whenever a catalyst is participating in a reaction, a different amount of energy is required to reach a transition state. A catalyst may be one of the reactants or products of a given reaction or a substance which is neither the reactant nor the product.

An example for such a catalytic concept is shown in Figure 2.5 based on the EGF pathway. Here, different proteins form complexes and some these reactions are catalysed. For example, LRIG-1 is acting as catalyst and inhibiting the formation of EGF:EGFR, whereas SRC-1 is activiting the formation of EGF:p-6Y-EGFR. Figure is adapted from Creixell *et al.* [2015].

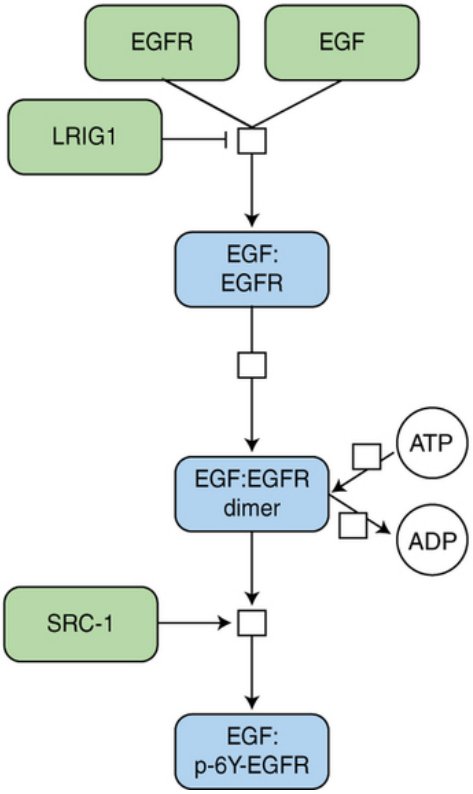

Figure 2.5: Concept of catalysis in EGF pathway. Proteins (green) form complexes (blue) which further participate in curated reactions. LRIG-1 is acting as catalyst and inhibiting the formation of EGF:EGFR, whereas SRC-1 is activiting the formation of EGF:p-6Y-EGFR. Figure is adapted from Creixell *et al.* [2015].

We can list some special forms of catalysis. The energy needed to reach the transition state is decreased by the presence of a *positive catalyst*. In contrast, the required energy for reaching the transition state is increased by a *negative catalyst* or also inhibitor. If one of the reaction products is also a reactant then it is called *autocatalyst*. Finally, an *induced catalyst* influences the rate of a reaction which without the presence of the catalyst would not be possible under ordinary conditions.

In the following chapters, we will use the described mathematical and biological background as a basis for investigating several research questions.

3

Significance test for difference between paired temporal observations

In this chapter, we introduce a novel statistical significance test for the difference of paired time-resolved observations. We construct a test statistic similar to a univariate t-test and take into account location, variability and size of the tested data. This is done by approximating the time courses with smoothing splines and then calculating and integrating over the functional mean and the functional standard deviation. The formulated test statistic has an unknown distribution and for assessing its significance we sample from the null hypothesis with preservation of the functional variability. It is the first statistical test of its kind which is suitable for time-resolved and paired data.

The developed test is applied on a large number of different artificially created datasets and its dependence on several influencing factors such as noise, number of time points per sample, number of samples per group and fraction of missing time points per sample are investigated. Furthermore, the test is compared in terms of power and receiver operator characteristic (ROC) curves to two other methods which do not account for the sample pairing in the two investigated groups. Finally, the test is used to quantify data in two real-world data scenarios. First, a setting from nutritional sciences is studied. Second, differences in genetic loci of wild-type and Atrx knock-out embryonic stem cells are assessed.

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

- I. Kondofersky, T. Erdmann, T. Brennauer, H. Hauner, F. J. Theis, C. Fuchs. Significance test for difference between paired temporal observations, *in preparation*.
- D. Sadic, K. Schmidt, S. Groh, I. Kondofersky, J. Ellwart, C. Fuchs, F.J. Theis, and G. Schotta (2015). Atrx promotes heterochromatin formation at retrotransposons. EMBO Rep., 16, 836850.

3.1 State of the art

Studying biological processes is often done by collecting temporal observations (Kholodenko [2006]; Smith *et al.* [2015]; Zhang *et al.* [2005]). As an example consider longitudinal studies which study obesity and insulin resistance over time (Jess *et al.* [2008]; McCormack *et al.* [2013]; Mihalik *et al.* [2012]). In such studies, often a hypothesized effect is investigated by collecting data either under different conditions or from two different groups of origin, called wild-type and knockout group or control and treatment group. Often this effect is assessed by performing univariate significance tests for each time point where data was collected and applying a multiple comparison correction to adjust for the number of time points (Fathers *et al.* [2005]; Lohr *et al.* [2014]; Nishino *et al.* [2011]; Prajapati *et al.* [2009]; Schikowski *et al.* [2013]; Weber *et al.* [2015]). However, this approach has multiple problems such as interpretation of different significance findings at different time points. Instead, a solid conclusion about an overall difference in both groups should be based on the full information available: all temporal measurements, group association and possibly pairing within groups. To our knowledge, the current literature does not provide a tool which is able to use the full information and at least some of the available information (e. g. time dependency or observation pairing) is not used. In this chapter we develop a novel test which uses the full information and allows to draw conclusions about the overall difference of two groups of temporal observations. We call it **t**ime-resolved **p**aired **d**ifferences **t**est, or short *TPDT*.

Several methods which account for the time dependency between single measurements, were developed throughout the past decade.

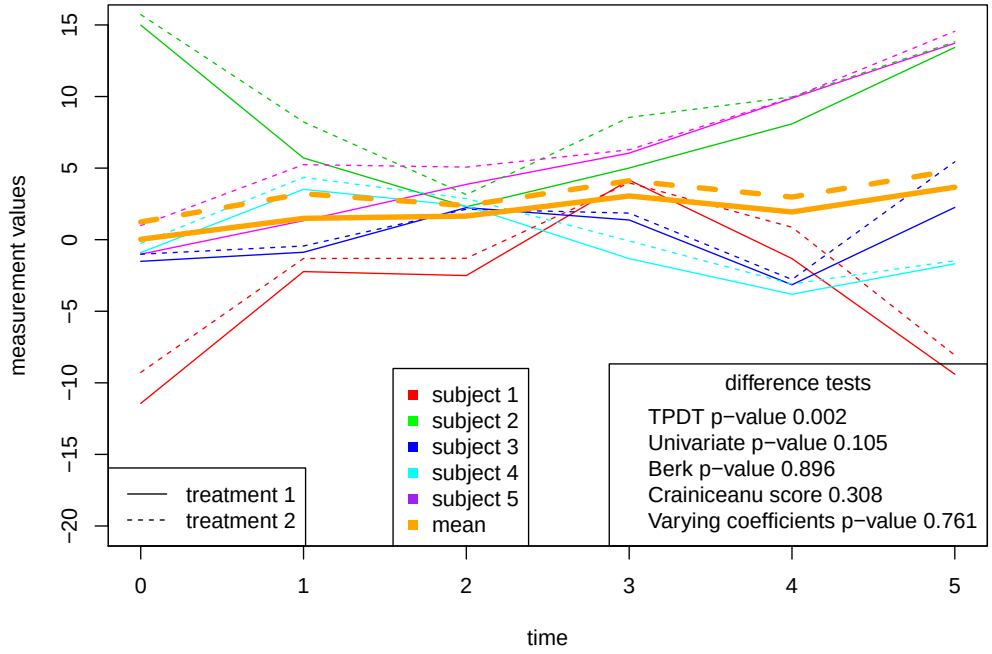

Figure 3.1: Artificial example for difference assessment in temporal measurements. Treatment 1 (solid lines) and treatment 2 (dashed lines) are performed on five different subjects denoted by the different colours. The overall mean of both treatments is shown in orange colour and thicker lines. Four tests were applied on this data with only TPDT correctly rejecting the null hypothesis of no differences between both groups.

One of the first to start developing such a test by using smooth functions were Storey *et al.* [2005], however the developed test did not have the ability to deal with missing, repeated or non-synchronized time points. A Bayesian approach by Angelini *et al.* [2007] was proposed where the testing procedure could be used for the analysis of two competing smooth curves. However, the test does not allow to compare two bundles of time series data representing two different groups. Recently, a test which was able to handle multiple time series per group was proposed by Berk *et al.* [2011]. It relies on mixture models theory to approximate smooth functions which appropriately represent the time dependency of the measured data. The test is carried out by comparing the parameters of these smooth functions. Finally, Crainiceanu *et al.* [2012] formulate a test which is

Table 3.1: Comparison of different significance tests which may be applied on temporal data.

Properties	Univariate t-test	Varying coefficients	Storey 2005	Angelini 2007	Berk 2011	Crainiceanu 2012	TPDT
time-resolved	no	yes	yes	yes	yes	yes	yes
missing	no	yes	no	yes	yes	yes	yes
repeated	no	yes	no	yes	yes	yes	yes
non-synchronized	no	yes	no	yes	yes	yes	yes
paired	yes	no	no	no	no	yes	yes
global decision	no	yes	no	no	yes	no	yes

easy to implement and allows for detailed investigation of the difference between two groups. The method uses bootstrap based confidence intervals, which allow for the quantification of single time periods where a difference between the two studied groups may occur. The above-described tests as well as two other commonly available alternatives, univariate t-test and varying coefficients model, can be conveniently grouped alongside with the newly developed TPDT by different properties in Table 3.1.

The different properties we judged the tests on were the following:

- time-resolved: can the test be applied on time resolved observations?
- missing: can the test handle missing observations?
- repeated: can the test handle repeated observations?
- non-synchronized: can the test still be applied if the observations are not made at the exact same time points?
- paired: does the test consider pairing between the two groups?
- global decision: Does the test come to a decision whether there is a global difference between the time series?

Table 3.1 demonstrates that there is no test yet which is able to extract the full information out of the data. Especially when observations are paired and one is interested in a global difference over the whole time series, no tests are available to incorporate this information into the test procedure. A possible pairing may have an effect on the end result, which, depending on the dataset, is of different magnitude. We can further elaborate this statement with a theoretical example which demonstrates the importance of pairing consideration in statistical testing. First, notice that a significance test usually takes into account both location, e. g.

the mean difference in two groups, and variability, e. g. the noise contained in the measurements. An overall difference between two groups is more sensibly identifiable when the difference location is large and at the same time the difference in variability is low. If the pairing in the observations is ignored, the location measure does not change. However, the variability measure is affected due to the simple relationship $\text{var}(X-Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y)$ for two random variables X and Y. On the one hand, if X and Y are positively correlated, then the variability of the differences will be larger than the sum of the variability. On the other hand, if X and Y are negatively correlated the variability of the differences will be larger.

To further demonstrate the state of the art methodology for assessing temporal differences, consider the artificially created example of two different treatments shown in Figure 3.1. Here, we created a data situation where we consider differently shaped time-resolved observations of two treatments. For each subject, we first simulate one treatment outcome and then use this simulated time course and shift it upwards by a nearly constant value of 0.8 (with small noise added) at each time point. This results in paired time-resolved observations. The two treatments are shown in solid versus dashed lines and the temporal measurements of five different subjects are shown in different colours. This means that data of each subject is available one time with application of treatment 1 and one time with application of treatment 2. The subject-specific effect is present as the temporal behaviour is different across subjects. For example, the green lines tend to have high measurement values at both limits of the considered time scale whereas the red lines have high measurement values at the middle of the time scale.

The mean at each time point is shown in orange and it is obvious that there exists a difference between the two treatments. However, available tests (which we will address in more detail later in Section 3.4) fail to find a significant difference (all p-values larger than 0.1). In contrast, the newly developed TPDT is able to correctly reject the null hypothesis with a low p-value of 0.002.

Similar to the above mentioned tests, our approach also begins by approximating the time courses of the samples by using the raw data measurements. Next, we use these time-courses and develop a statistical framework and compute a test statistic which measures the difference between two groups of multiple paired smooth functions. The distribution of this test statistic is approximated with a resampling technique which allows us to compute a p-value and thus in turn makes

the decision whether to reject the null hypothesis of no difference in the two groups feasible. For simplicity, we discuss applications and theoretical aspects of the test by considering temporal observations. Note, however, that the proposed test does not need time-resolved observations but any kind of data which can be described by a smooth function.

The rest of this chapter is organized as follows: In Section 3.2 we develop the method and present details on the mathematical computation and the statistical hypothesis. Next, we apply the developed statistical test in on artificial as well as real-world data. Here, we first assess the general applicability of the developed test in Section 3.3 and then compare it to other commonly used approaches in literature in Section 3.4. Analysis of data from nutritional sciences in Section 3.5.1 and embryonic stem cell data in Section 3.5.2 are also demonstrated. Finally, we conclude the chapter in Section 3.6.

3.2 Methods

3.2.1 Notation and spline representation

We consider two paired groups of variables $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1,\dots,N}$ and $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i=1,\dots,N}$. The index denotes the pairing of the two groups and contains the information that $\boldsymbol{x}_l$ and $\boldsymbol{y}_k$ are connected or paired only if $l = k$. We assume that each $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ represents time-resolved measurements of possibly different lengths denoted by $x_i(t_{j_1}^{(x,i)})$ and $y_i(t_{j_2}^{(y,i)})$. The measurements were made at discrete time points $t_{j_1}^{(x,i)}$ and $t_{j_2}^{(y,i)}$ with $j_1 \in \{0,\dots,J_{xi}\}$ and $j_2 \in \{0,\dots,J_{yi}\}$. Written in vector form, the time points are $\boldsymbol{t}_{xi} = (t_0^{(x,i)},\dots,t_{J_{xi}}^{(x,i)})$ and $\boldsymbol{t}_{yi} = (t_0^{(y,i)},\dots,t_{J_{yi}}^{(y,i)})$. An example of such measurements is shown in Figure 3.2A. We do not require that the measurements of both variable groups or within the groups are synchronized in the sense that the time points at which measurements were made are equal over all variables or the number of measurements per variable is equal. However, mainly for simpler notation, we require that $t_0^{(x,i)} = t_0^{(y,i)}$ as well as $t_{J_{xi}}^{(x,i)} = t_{J_{yi}}^{(y,i)}$.

We are interested in the detection of time-resolved differences between both groups $\mathcal{X}$ and $\mathcal{Y}$ and want to explore it in a functional context. We therefore define the fol-

lowing hypotheses:

$$\begin{aligned} H_0 &: \text{groups are equal,} \\ H_1 &: \text{groups are different.} \end{aligned} \tag{3.1}$$

More specifically, we are interested in the *time-resolved* differences in both groups. We assume that the measurements $x_i(t_{j_1}^{(x,i)})$ and $y_i(t_{j_2}^{(y,i)})$ represent local snapshots of a smooth time course of the variables. In a first step, we approximate this time course by smoothing splines as already discussed in Chapter 2.1.3:

$$\begin{aligned} \hat{x}_i(t) &:= \sum_{k=1}^{K_x} \hat{\beta}_{xki}\phi_{kx}(t) = \hat{\boldsymbol{\beta}}_{xi}\boldsymbol{\phi}_x(t) \\ \hat{y}_i(t) &:= \sum_{k=1}^{K_y} \hat{\beta}_{yki}\phi_{ky}(t) = \hat{\boldsymbol{\beta}}_{yi}\boldsymbol{\phi}_y(t) \end{aligned} \tag{3.2}$$

where $\boldsymbol{\phi}_x = (\phi_{1x}, \ldots, \phi_{K_x x})^T$ and $\boldsymbol{\phi}_y = (\phi_{1y}, \ldots, \phi_{K_y y})^T$ are known basis functions, e. g. B-spline basis functions and K_x, and K_y are the respective number of basis functions. $\hat{\boldsymbol{\beta}}_{xi} = (\hat{\beta}_{x1i}, \ldots, \hat{\beta}_{xK_x i})^T$ and $\hat{\boldsymbol{\beta}}_{yi} = (\hat{\beta}_{y1i}, \ldots, \hat{\beta}_{yK_y i})^T$ represent optimized coefficient vectors. There are several ways how these coefficients can be optimized as discussed in Chapter 2.1.2. The corresponding equation for such an optimization in the current notation is (analogously also for $\hat{\boldsymbol{\beta}}_{yi}$)

$$\hat{\boldsymbol{\beta}}_{xi} = \underset{\boldsymbol{\beta}_{xi}}{\operatorname{argmin}} \left(\left(x_i(\boldsymbol{t}_{xi}) - \boldsymbol{\beta}_{xi}\boldsymbol{\phi}_x(\boldsymbol{t}_{xi})\right)^T \left(x_i(\boldsymbol{t}_{xi}) - \boldsymbol{\beta}_{xi}\boldsymbol{\phi}_x(\boldsymbol{t}_{xi})\right) + \lambda_{xi} \int_t \left(\boldsymbol{\beta}_{xi}\ddot{\boldsymbol{\phi}}_x(s)\right)^2 \mathrm{d}s \right) \tag{3.3}$$

where $\ddot{\boldsymbol{\phi}} = (\ddot{\phi}_{1x}, \ldots, \ddot{\phi}_{1x})^T$ denotes the vector of twice differentiated basis functions with respect to time. We estimate λ_{xi} using cross validation (see Chapter 2.1.3). Figure 3.2B shows such estimated smooth curves.

With the above-described equations (3.2) and (3.3) we are able to represent the raw measurements as smooth functions. For that reason, typical problems such as non-synchronized time points of the measurements or missing values are effectively addressed. In a next step, after smooth representation of $\mathcal{X}$ and $\mathcal{Y}$, we proceed by constructing a test statistic for difference assessment using these smooth functions.

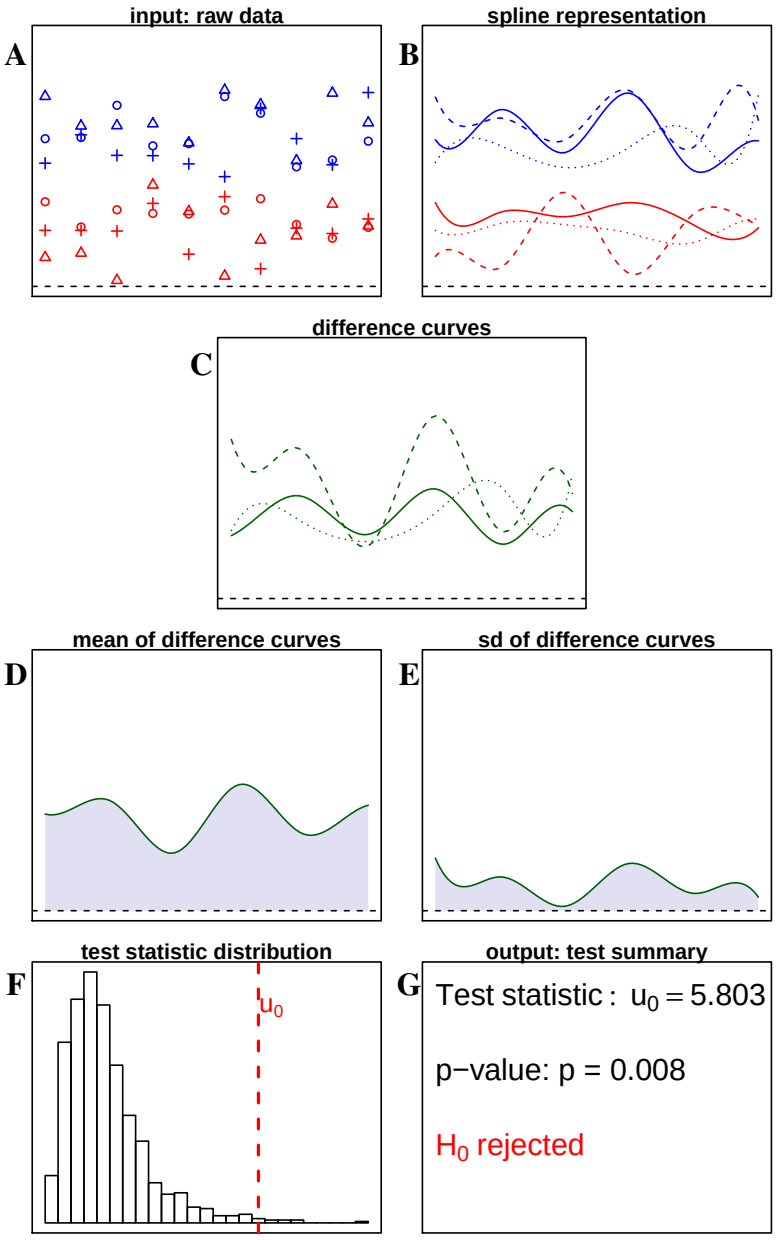

Figure 3.2: TPDT concept. Decision about the null hypothesis (H_0) of no differences between two groups of temporal observations is made. **A**: raw measurements of each subject are denoted by different point shapes, treatments are shown in different colours; **B**: splines are fitted to the raw measurements; **C**: difference curves of each subject are calculated; **D**: the functional mean of the difference curves is calculated and the corresponding integral (shaded area) approximated; **E**: the functional standard deviation of the difference curves is calculated and the corresponding integral (shaded area) approximated; **F**: distribution of test statistic u is approximated with resampling; **G**: test outcomes such as p-value and hypothesis decision are extracted.

3.2.2 Test statistic u

After estimation of $\hat{x}_i(t)$ and $\hat{y}_i(t)$, we calculate the difference curves (Figure 3.2C)

$$\hat{d}_i(t) = \hat{y}_i(t) - \hat{x}_i(t). \tag{3.4}$$

With these difference curves, we are able to transform the question whether the two groups of paired time-resolved observations differ significantly from each other by a pre-defined function $\mu_0(t)$ into the question whether the difference curves significantly differ from an arbitrary function $\mu_0(t)$. This also allows a reformulation of the hypotheses (3.1):

$$\begin{aligned} H_0 &: \bar{d}(t) = \mu_0(t) \\ H_1 &: \bar{d}(t) \neq \mu_0(t). \end{aligned} \tag{3.5}$$

The answer has to take into account noise, variability, location and size of the considered dataset. Similar to a univariate t-test for paired observations, we define the test statistic to equal

$$u := \sqrt{N}\frac{D}{S} = \sqrt{N}\frac{\int_{t_0}^{t_n} \mid \bar{\hat{d}}(s) - \mu_0(s) \mid ds}{\int_{t_0}^{t_n} \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{d}_i(s) - \bar{\hat{d}}(s)\right)^2}\, ds}. \tag{3.6}$$

with a functional mean difference curve $\bar{\hat{d}}(t) = \frac{1}{N}\sum_{i=1}^{N}\hat{d}_i(t)$. The test statistic u captures the location differences of both groups in D and it takes the variability of both groups into account in S. We approximate the integrals through finite differences (Ramsay & Silverman [2005]). Both quantities are shown as the grey areas in the Figure 3.2D and Figure 3.2E. Additionally, we use a correction term for the number of curves in the two groups in the form of $\sqrt{N}$ in order to extract similar magnitudes of u for different sizes of datasets. The arbitrary curve $\mu_0(t)$ can be used to answer questions whether the two groups differ significantly from each other and an additional offset $\mu_0(t)$, which may also be constant over time. A result of $u = 0$ means that $D = 0$ for the observed data and this in turn means that the observed functional means of both groups are exactly the same. A result of $u > 0$ means that there are differences between $\mu_0(t)$ and $\bar{\hat{d}}(t)$ for the observed data. The significance of these differences will be assessed with a resampling

approach in the following. We reformulate the hypotheses a further time to:

$$\begin{aligned} H_0 &: u = 0 \\ H_1 &: u > 0. \end{aligned} \tag{3.7}$$

3.2.3 Resampling functional curves

Calculation of the test statistic u as described in (3.6) reveals whether there are differences between the two considered groups of time-resolved observations. This is representative for the current dataset but it is not generalisable due to the current dataset being a random subsample of all possible datasets. The distribution of u would allow to make general test decisions, however it is unknown. Therefore, we apply a resampling approach to approximate the distribution of u. To that end, we simulate time-resolved measurements under the null hypothesis of no difference between both groups. The simulated curves are chosen in such way that both, the smooth curves corresponding to the original data $\mathcal{X}$ and $\mathcal{Y}$, as well as the simulated curves contain the same amount of variability S and are of the same size N in a sense that is explained further below. W. l. o. g. assume that $\mu_0(t) = 0$ and the considered smooth curves $\hat{d}_i(t)$ belong to the same class of functions, $f(x) = \sum_k \beta_k \phi_k(x)$, where $\phi_k(x)$ are cubic B-splines which span over the same interval. We simulate curves from this class by adding a normally distributed noise with mean 0 to the basis coefficients:

$$\begin{aligned} d_i^{\text{sim}}(t) &= d_i(t) + \boldsymbol{\varepsilon}_i \\ &= \sum_{k=1}^{K_d} \hat{\beta}_{dki} \phi_{kd}(t) + \sum_{k=1}^{K_d} \varepsilon_{ki} \phi_{kd}(t) \\ &= \sum_{k=1}^{K_d} (\hat{\beta}_{dki} + \varepsilon_{ki}) \phi_{kd}(t) \end{aligned} \tag{3.8}$$

with K_d the number of basis functions used to represent the difference curves, $\hat{\beta}_{dki}$ the difference basis coefficients, ϕ_{kd} the corresponding basis functions (estimated in (3.3)), $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \ldots, \varepsilon_{K_d i})^T$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{K_d}(0, \Sigma)$ with covariance matrix Σ. As we show later, Σ is chosen to equal the empirical covariance matrix of the basis coefficients which are extracted from the splines which approximate the observed measurements. The expectation of $d_i^{\text{sim}}(t)$ equals the estimated functional mean

$\hat{d}_i(t)$:

$$\begin{aligned} \mathrm{E}\left[d_i^{\mathrm{sim}}(t)\right] &= \mathrm{E}\left[\sum_{k=1}^{K_d}(\hat{\beta}_{dki}+\varepsilon_{ki})\phi_{kd}(t)\right] \\ &= \sum_{k=1}^{K_d}(\hat{\beta}_{dki}+\mathrm{E}\left[\varepsilon_{dki}\right])\phi_{kd}(t) \\ &= \sum_{k=1}^{K_d}(\hat{\beta}_{dki}+0)\phi_{kd}(t) \\ &= \hat{d}_i(t). \end{aligned} \tag{3.9}$$

An example for simulated random curves is shown in Figure 3.3. We chose a sinusoidal shaped curve demonstrating the randomizing of a functional curve. We added normal noise with zero mean and variance equal to 1 to the basis coefficients and repeated this for 1000 times. The mean curve of the resulting simulated curves is very close to the function from which data was generated for this large sample size. If we look at only the mean curve of the first 20 random curves it is already very closeto the function from which data was generated for this moderately large sample size.

As we stated above, the resampling of the test statistic is done under the null hypothesis of no difference between both groups. While resampling, we want to preserve the variability contained in the data for which the test is applied. Therefore, the covariance matrix Σ is estimated from the fitted basis coefficients of the difference curves:

$$\hat{\Sigma} = \mathrm{cov}(\hat{\boldsymbol{\beta}}_d) \tag{3.10}$$

with $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{d1i},\ldots,\hat{\beta}_{dK_dN})^T$. Using this empirical covariance matrix, we simulate N curves and calculate a test statistic u_b as in (3.6) which is based on these simulated curves. Repeating this procedure B times (e. g. $B = 10^6$) results in B different test statistics which are used to approximate the distribution of u (Figure 3.2F). In a last step, we apply the percentile method (Efron & Tibshirani [1994]) and consider the fraction of $(u_1,\ldots,u_B)^T$ which have a more extreme value than the test statistic computed on the original smooth curves u. This fraction also gives the final estimate for a p-value $\hat{p}$ of the test:

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} I(u_b > u) \tag{3.11}$$

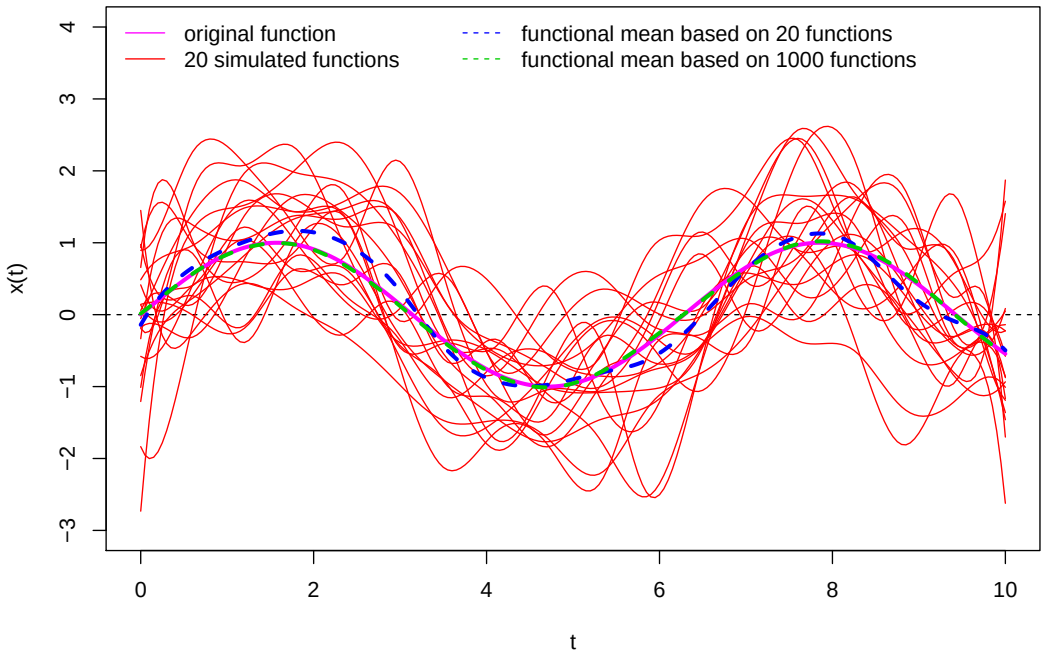

Figure 3.3: Simulated random curves. The curve which is used for sampling has a sinusoidal form and is shown in purple. 20 single simulated curves are shown in red and the simulation based mean curves based on 20 (blue) or 1000 (green) curves show good agreement with the data generating curve.

with indicator function $I(A)$. This p-value can be interpreted in the usual way as the probability of having observed a difference between $\mathcal{X}$ and $\mathcal{Y}$ given H_0 is true only by chance. Therefore, it serves as a tool to decide whether to reject the null hypothesis or not. For a given significance level α, we can look at the approximated $1-\alpha$ quantile $\hat{q}_{1-\alpha} = \lceil B \cdot (1-\alpha) \rceil$ -value of the sorted test statistics. The final test decision is then made:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } u \geq \hat{q}_{1-\alpha} \\ \text{Do not reject } H_0 & \text{otherwise.} \end{array} \tag{3.12}$$

In summary, with the newly developed TPDT we are able to identify whether two paired groups of time-resolved measurements significantly differ in location from each other and summarize this result in a single scalar p-value (Figure 3.2G).

In the next section we thoroughly test the developed TPDT on several artificially created data scenarios in order to investigate the general applicability of the test and compare it to other available tests. We first study the effect of different data

settings, such as noise level, number of subjects, number of time points, smoothing parameters for the spline estimation as well as data missingness. This is studied by considering vertical, horizontal and multiplicative shifts of two groups. Next, we compare TPDT to two other statistical tests, which are used to compare differences in a functional context between two groups. We do this by comparing the area under the receiver operator characteristic (ROC) and power of the three tests, which are again applied on different artificially created datasets.

3.3 Parameter influence on TPDT

In this subsection we will demonstrate TPDT on synthetic data. Furthermore, we will assess the test dependency on several parameters: noise, number of subjects, number of time points, difference in the created groups and missing observations.

We consider two groups of temporal observations – group 1 and group 2. Data for group 1 is generated based on the function

$$f_1(t) = 2t\sin(t) + 10. \tag{3.13}$$

We sample snapshots of this function at n equidistant time points within the interval $[0, 10]$. Subsequently, we add normally distributed noise with 0 mean and variance σ_1^2 to each snapshot. Next, we randomly delete a fraction of m observations to create a missing data scenario and thus form one time-resolved sample of group 1. This process is repeated N times to obtain N similar time-resolved samples of this group. The parameters n, σ_1^2, m and N are varied (see Table 3.2) with the purpose of investigating a large number of different simulation scenarios and thus get different datasets onto which to apply our test.

In a next step, we consider a second group of time-resolved measurements by introducing shifts of different types and applying them on the data corresponding to group 1. Here, we consider four different scenarios:

1. vertical shift v_1 (group 2.1)
2. horizontal shift v_2 (group 2.2)
3. slope shift v_3 (group 2.3)

4. combined shift ν_4 (group 2.4)

The corresponding functions from which we generate data in group 2 are

$$f_{2.1}(t) = 2t\sin(t) + 10 + \nu_1, \tag{3.14}$$
$$f_{2.2}(t) = 2t\sin(t + \nu_2) + 10, \tag{3.15}$$
$$f_{2.3}(t) = 2(t + \nu_3)\sin(t) + 10, \tag{3.16}$$
$$f_{2.4}(t) = 2(t + \nu_4)\sin(t + \nu_4) + 10 + \nu_4. \tag{3.17}$$

We introduce pairing between observations from group 1 and group 2 in the following way. We take the (noisy) time-resolved samples already obtained in group 1 and first shift them by a chosen value of $\nu_i, i \in \{1,\ldots,4\}$, and subsequently add normally distributed noise with variance σ_2^2 to each of these shifted snapshots. Using a (noisy) sample from group 1 for the generation of a sample of group 2 naturally introduces a pairing between the two samples. Therefore, following the above-described protocol of data generation, we obtain matched samples from group 1 and group 2.

A comparison of the functions for $f_{2.1}(t)$ to $f_{2.4}(t)$ for different values of $\nu_i, i \in \{1,\ldots,4\}$ is shown in Figure 3.4. We observe that with the same values for $\nu_i, i \in \{1,\ldots,4\}$ in each scenario, we introduce differences between both groups of varying magnitude. Along this line, we expect that ν_3 and ν_4 will have the largest effect when we investigate differences in both groups and ν_1 will have the smallest effect.

We conducted a large number of TPDT comparisons by varying the parameters n, N, σ_1, σ_2, m and $\nu_i, i \in \{1,\ldots,4\}$. The varied parameters are summarized in Table 3.2. The ranges of variation of these parameters were chosen to resemble realistic data situations, such as e. g. 10 snapshots per time-resolved sample or a low amount of samples per group and a moderate percentage of missing values.

Choosing a value of 0 for the shift parameters $\nu_i, i \in \{1,\ldots,4\}$ results in the correct test decision being to not reject the null hypothesis because the data generating process for both groups is the same. On the other hand, if we choose a positive value for $\nu_i, i \in \{1,\ldots,4\}$, the correct test decision is to reject the null hypothesis. For each combination of parameters, we simulate a dataset and perform TPDT on both groups and thus obtain a different p-value for each parameter combination.

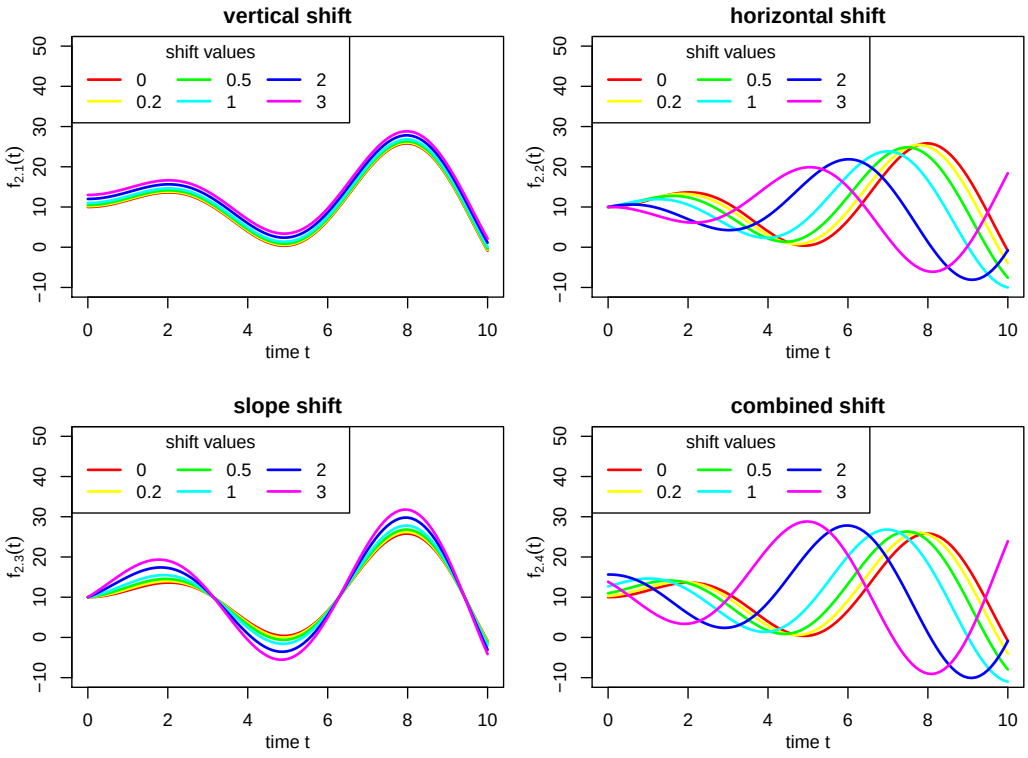

Figure 3.4: Functions from which data is simulated. The different types of shifts $\nu_i, i \in \{1,\ldots,4\}$ are shown in the four panels.

We continue with discussion of these results in the following. Only a subsample of the results is discussed and shown in Figure 3.5 – Figure 3.8. The complete set of results is available upon request. For the discussion, we choose a significance level $\alpha = 0.05$ which means that whenever a p-value is calculated to be lower than 0.05, the test decision is to reject the null hypothesis of no differences.

It holds for all simulated parameters that the null hypothesis was almost never rejected when there was no difference between both groups (blue boxplots). This is an indication for low type II errors as we will demonstrate in Section 3.4.

Figure 3.5 and Figure 3.6 indicate that the two noise parameters σ_1^2 and σ_2^2 play a major role if the difference between both groups is small ($\nu_1 \leq 0.5$ or $\nu_2 \leq 0.2$) and especially if the number of samples per group is also small ($N < 6$). If we introduce a large amount of noise for those cases, TPDT does not reject the null hypothesis reliably.

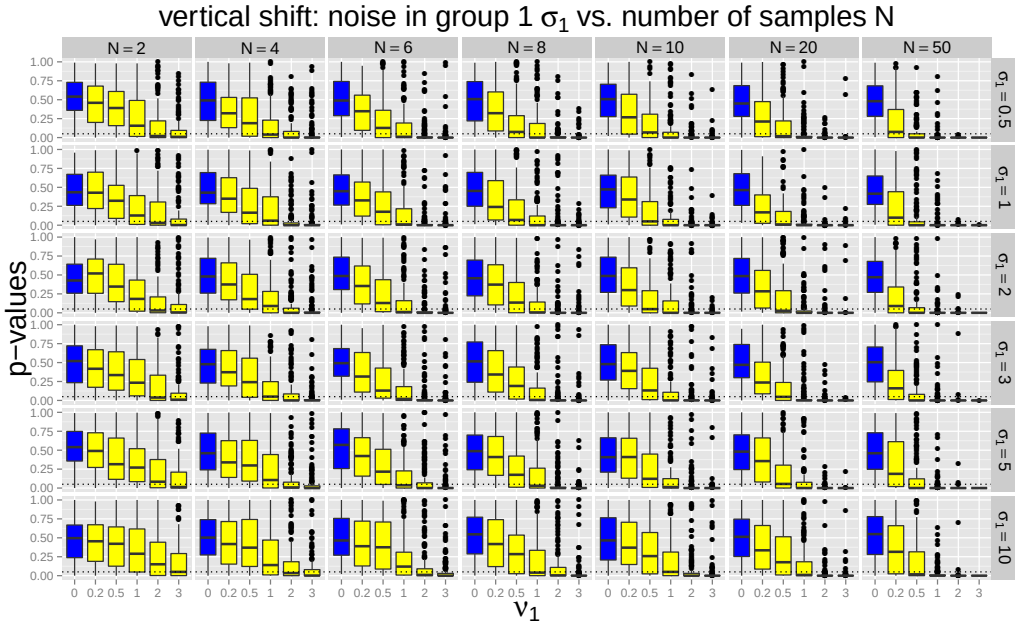

Figure 3.5: Influence of σ_1^2 on TPDT performance (scenario: vertical shift). Blue: null hypothesis correct, yellow: alternative hypothesis correct.

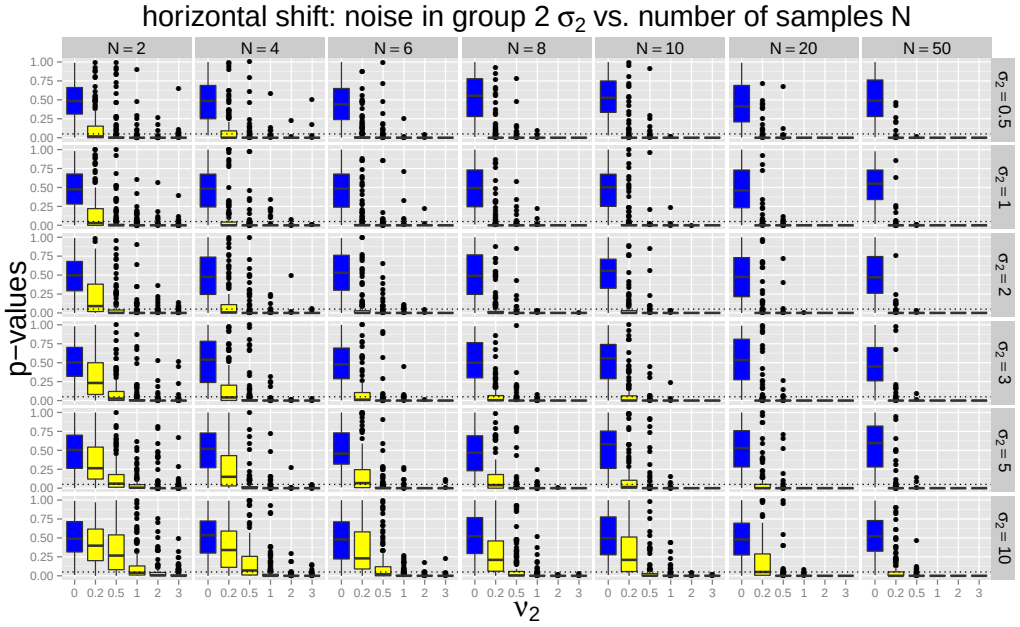

Figure 3.6: Influence of σ_2^2 on TPDT performance (scenario: horizontal shift). Blue: null hypothesis correct, yellow: alternative hypothesis correct.

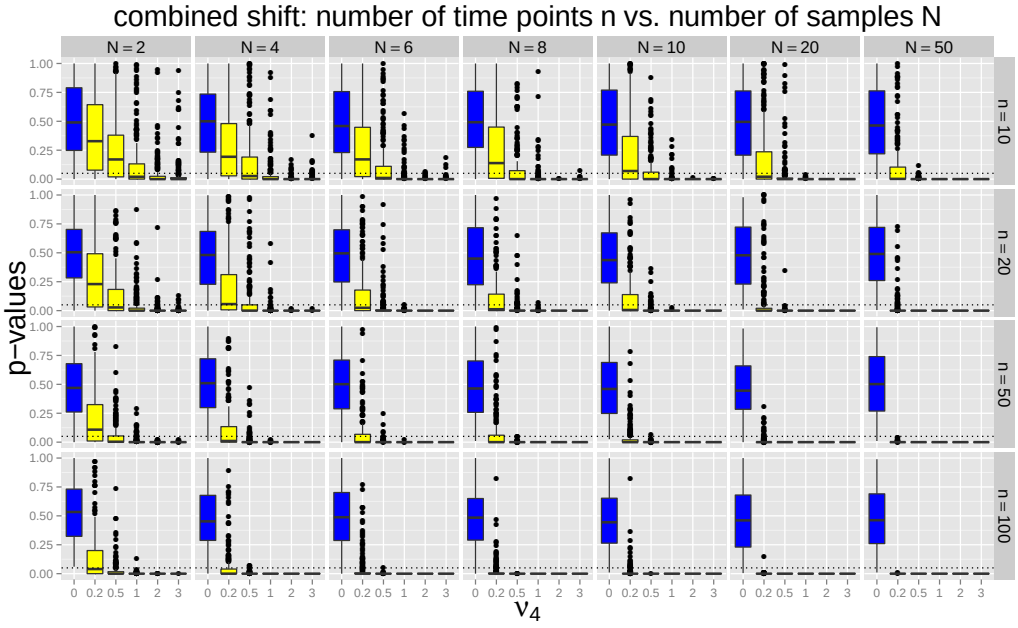

Figure 3.7: Influence of n on TPDT performance (scenario: combined shift). Blue: null hypothesis correct, yellow: alternative hypothesis correct.

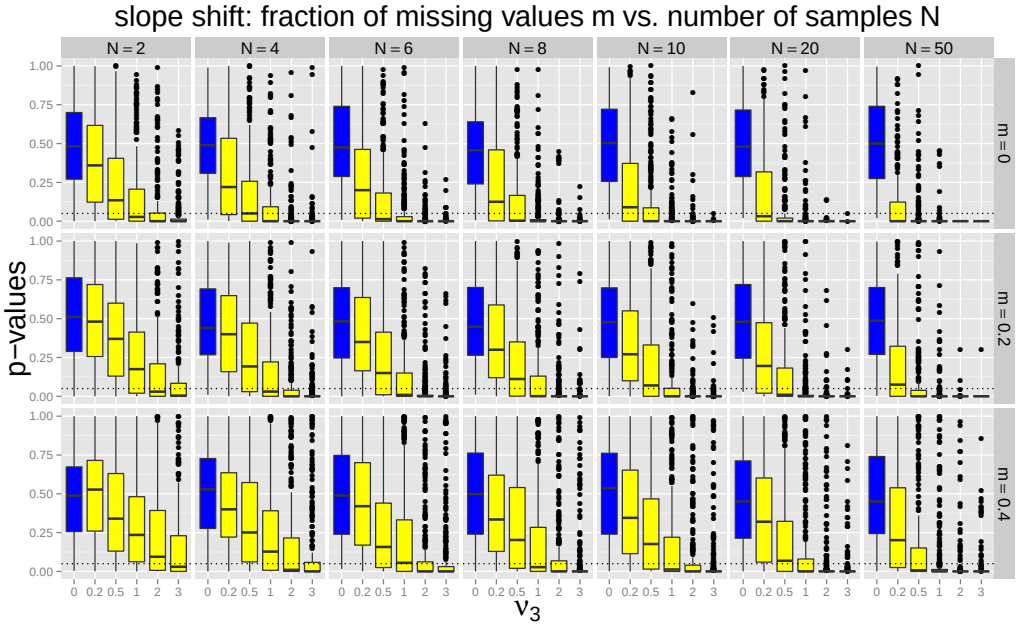

Figure 3.8: Influence of m on TPDT performance (scenario: slope shift). Blue: null hypothesis correct, yellow: alternative hypothesis correct.

Table 3.2: Parameters are varied and for each combination, we simulate and analyse the corresponding data with TPDT.

parameter	varied values
number of time points n	$\{10, 20, 50, 100\}$
number of paired samples in each group N	$\{2, 4, 6, 8, 10, 20, 50\}$
noise in group 1 σ_1	$\{0.5, 1, 2, 3, 5, 10\}$
noise in group 2 σ_2	$\{0.5, 1, 2, 3, 5, 10\}$
percentage of missing snapshots m	$\{0, 20, 40\}$
amount of difference $\nu_i, i \in \{1, \ldots, 4\}$	$\{0, 0.2, 0.5, 1, 2, 3\}$

The number of sampled time points n only has a substantial effect on the test decision if the number of observations per group N is small. For few time points ($n \leq 20$) and low sample size ($N \leq 10$), a large difference between both groups is needed for a correct test decision (Figure 3.7). For a large number of time points $n = 100$, TPDT correctly rejects the null hypothesis in almost all cases as long as the number of subjects is not too low ($N < 5$).

Finally, the ratio of missing values m seems to not have a large effect on the simulation results again only if the shift values (ν_3) are low (Figure 3.8). Overall, the ratio of missing snapshots does not have a large impact on the test decision.

We further investigate this data in the next chapter, where we apply TPDT and other tests on the same data discussed in the present section. In the following, we introduce and describe other available tests and comment on possible adaptations we perform in order to make all tests comparable. Subsequently, we compare TPDT to the other methods by considering power and ROC analyses.

3.4 Comparison of TPDT to other methods

In this section, we compare TPDT to other methods which are specifically designed for detection of time-resolved differences. Additionally, we also implemented an adapted version of a univariate t-test, which is commonly used in literature. The comparison is done using the data simulated in Section 3.3 and calculating the power and ROC curves of each of the tests for different values of $\nu_i, i \in \{1, \ldots, 4\}$. Furthermore, we will also compare the tests for a shuffled version of the same data where we intentionally remove the pairing between samples. Prior to these comparisons, we first introduce and describe the considered tests.

3.4.1 Other available methods

Moderated functional t-type statistic after Berk et al.

Berk *et al.* [2011] introduce an approximative time-resolved test for differences between two groups of time-resolved samples. A test statistic is calculated by applying a mixture model, and the distribution of this test statistic is then approximated by bootstrap methods.

The authors first define a functional mixed-effects model (Guo [2002]) for each of both groups as

$$y_k(t_{ij}) = \mu_k(t_{ij}) + \nu_{ik}(t_{ij}) + \varepsilon_{kij} \tag{3.18}$$

where $y_k(t_{ij})$ are the measurements for group $k, k \in \{1,2\}$, $\mu_k(t_i j)$ is the fixed mean function for group k evaluated at t_{ij}, $\nu_{ik}(t_{ij})$ is the i-th random realization of an underlying Gaussian process with zero mean for group k evaluated at t_{ij} and ε_{kij} are additive error terms with group specific variance σ_k^2. The number of samples per group are denoted by n_k and the number of temporal observations per sample are denoted by m_i, thus $i \in \{1,\ldots,n_k\}$ and $j \in \{1,\ldots,m_i\}$. $\mu_k(t)$ and $\nu_{ik}(t)$ are represented by smoothing splines (see Chapter 2.1.3) and $\hat{\mu}_k(t)$ and $\hat{\nu}_{ik}(t)$ are estimated by a penalised generalised log-likelihood ansatz. After obtaining these estimates, a test statistic Ft is calculated

$$\mathrm{Ft} = \frac{l_2}{se + se_m}. \tag{3.19}$$

The l_2 term is the square root of $\int_{t_{\min}}^{t_{\max}} (\hat{\mu}_1(t) - \hat{\mu}_2(t))^2 \, dt$ where the integration is done by applying the trapezoidal rule on a fine grid between both time limits $t_{\min}$ and $t_{\max}$. The term se is the functional standard error and is defined as

$$se = \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \tag{3.20}$$

and $\hat{s}_k^2, k \in \{1,2\}$ are the sample variance estimates for group k defined as

$$\hat{s}_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} \int_{t_{\min}}^{t_{\max}} (\hat{\nu}_{ki}(t))^2 \, dt. \tag{3.21}$$

Finally, the term se_m is a small value which is added to the denominator in (3.19) and is used to slightly moderate the Ft statistic. This is done in order to handle cases where the functional standard error *se* is very small and division in (3.19) would lead to inflated Ft statistics if even for small estimation errors. However, se_m is only of relevance if multiple variables are investigated for temporal differences. In the considered simulation studies this is not the case and thus we set this value to 0.

After computation of Ft, its distribution under the null hypothesis of no differences is approximated by resampling (as done also for TPDT). This is done by the following procedure which produces simulated observations:

1. Select one of $\mu_1(t)$ or $\mu_2(t)$ randomly as a mean curve.
2. Choose individual curves with replacement from v_{ki} randomly and add these individual curves to the selected mean curve.
3. Choose error terms with replacement from ε_{kij} and add them to the individual and mean curves.
4. Calculate an Ft statistic for this generated data.

This is repeated B times and results in B different Ft statistics. The Ft statistic which is based on the original data is then compared to those B statistics and a raw p-value is approximated.

The R package *sme* (Berk [2013]) implements this procedure and we used mainly the functions from this R package to calculate test statistics and p-values with this test. We had to implement slight adaptations of the provided functions due to e. g. computational errors when considering missing data.

Bootstrap-based joint pointwise confidence intervals after Crainiceanu et al.

Crainiceanu *et al.* [2012] propose a construction of pointwise and joint confidence intervals by extracting the covariance matrix of the data and then using it to obtain confidence intervals for the mean difference curve.

The authors first start by calculating two functional means, $\mu_1(t)$ and $\mu_2(t)$, and suggest to use penalised splines for the functional representation. Next, a difference function $d(t) = \mu_1(t) - \mu_2(t)$ is formulated and a functional bootstrap ap-

proach is used to calculate estimates of $d(t)$. This is done by randomly selecting one sample from group 1 and one sample from group 2, fitting penalized splines to these samples and finally computing the difference function. If the samples are paired, then only the sample from group 1 is selected randomly and the matched sample from group 2 is automatically chosen as second sample. This process is repeated $B = 1000$ times resulting in $B = 1000$ estimates $\hat{d}(t)$. These estimates are then evaluated at a grid of time points of length n and stored in the $n \times B$ matrix S. The column mean of this matrix is denoted by $\bar{d}$ and the covariance of the samples is an $n \times n$ matrix Σ. Once both $\bar{d}$ and Σ are calculated, the following algorithm is used to create joint confidence intervals:

1. Simulate d_i from a multivariate Gaussian distribution $\mathcal{N}_p(\bar{d}, \Sigma)$.
2. Calculate $x_i = \max_j \{\frac{|d_i - \bar{d}|}{\sqrt{\Sigma_{j,j}}}\}$.
3. Repeat steps 1 and 2 a total of N times and obtain $q_{1-\alpha}$, the $1-\alpha$ empirical quantile of the sample $\{x_i, i = 1 \ldots, N\}$.
4. Obtain n joint pointwise confidence intervals $\bar{d} \pm q_{1-\alpha}\sqrt{\Sigma_{j,j}}$for each $j \in \{1, \ldots, n\}$.

This method is especially useful if one wants to assess specific time intervals where a difference between both groups can be observed. However, the authors do not comment on how to construct a p-value with their method. For reasons of comparison we construct a score, which we can use for ROC comparisons later. The score is standardized in the interval $(\frac{1}{B}, 1 - \frac{1}{B})$ and low values stand for strong evidence against the null hypothesis given the current data. The score is constructed with the following protocol:

1. The above-described calculation of pointwise confidence intervals is performed with $\alpha = 0$.
   - If $\exists j : 0 \notin [\bar{d} - q_{1-\alpha}\sqrt{\Sigma_{j,j}}, \bar{d} + q_{1-\alpha}\sqrt{\Sigma_{j,j}}]$, set score $s = \frac{1}{B}$
   - Else: continue with protocol number 2.
2. The above-described calculation of pointwise confidence intervals is performed with $\alpha = 1$.
   - If $\forall j : 0 \in [\bar{d} - q_{1-\alpha}\sqrt{\Sigma_{j,j}}, \bar{d} + q_{1-\alpha}\sqrt{\Sigma_{j,j}}]$, set score $s = 1 - \frac{1}{B}$
   - Else: continue with protocol number 3.

3. The above-described calculation of pointwise confidence intervals is performed on a fine grid of α values.

    - Numerical optimization: find $\alpha^\star$, such that the number of j with $0 \in [\bar{d} - q_{1-\alpha^\star}\sqrt{\Sigma_{j,j}}, \bar{d} + q_{1-\alpha^\star}\sqrt{\Sigma_{j,j}}]$ is non-zero but minimal. Set score $s = \alpha^\star$.

In other words, in protocol number 3, we vary $\alpha^\star$ and the pointwise confidence intervals in such a way that the null hypothesis is rejected at a minimal number of time points and set the score s to this time point. The first two protocol numbers handle cases at both limits of α.

Adapted version of univariate paired t-test

We implemented a further procedure which we can use to investigate the null hypothesis of no differences in both groups. We call it an adapted version of the univariate paired t-test. It consists of subsequent application of the t-test (Student [1908]) at each time point which results in n different p-values. These p-values are then adjusted for multiple testing by using an false discovery rate correction (Benjamini & Hochberg [1995]) which in turn results in n adjusted p-values. Finally, we choose the smallest of all adjusted p-values as the overall p-value for the whole time-series data. As discussed in the beginning of this chapter, this approach is often used due to its simplicity and computational effectiveness. However, it clearly has the following shortcomings (among others):

1. The time dependency is not taken into account.
2. It is not applicable if the time series are not synchronized.
3. It is not applicable if data snapshots are missing.

Nevertheless, we implemented a version of this test for reasons of comparison.

Varying coefficients model

As a last option to investigate differences in two groups of time-resolved observations, we consider a varying coefficients model (Fahrmeir *et al.* [2007a]; Fan &

Zhang [2008]; Hastie & Tibshirani [1993]). Such model can be formulated as

$$\boldsymbol{y} = \alpha_0 + f_1(\boldsymbol{t}) + f_2(\boldsymbol{t})\boldsymbol{x} + \beta_1\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{3.22}$$

with data $\boldsymbol{y}$, coefficients $(\alpha_0, \beta_1)^T$, smooth spline functions $f_i(\boldsymbol{t}), i \in \{1,2\}$, dummy variable $\boldsymbol{x}$ and error vector $\boldsymbol{\varepsilon}$. The dummy variable $\boldsymbol{x}$ is used to distinguish observations of both groups. The interpretation of this model is as follows:

- α_0 is the overall intercept which is used to centre the data and make the model identifiable,
- $f_1(\boldsymbol{t})$ is the non-linear time effect for group 1 (corresponding to cases $\{\boldsymbol{x} \mid x_j = 0\}_{j=0,\ldots,n}$),
- $f_1(\boldsymbol{t}) + f_2(\boldsymbol{t}) + \beta_1$ is the non-linear time effect for group 2 (corresponding to cases $\{\boldsymbol{x} \mid x_j = 1\}_{j=0,\ldots,n}$).

Translated to the research question of this manuscript, difference between both groups is suggested if the estimation of f_2 is significantly different from 0. Estimation of the coefficients in $f_i(\boldsymbol{t}), i \in \{1,2\}$ as well as $(\alpha_0, \beta_1)^T$ is performed with a restricted maximum likelihood (REML) ansatz. Significance of the smooth terms is assessed with an approximated p-value (details in Wood [2013]). For an overall quantification of differences between both groups, we use the p-value corresponding to $f_2(\boldsymbol{t})$.

3.4.2 Comparison measures

In the following, we introduce and describe use statistical power and ROC curves as measures for the performance of the different tests.

Statistical power

The statistical power is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true.

$$\text{pow} = \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true}). \tag{3.23}$$

In the field of biostatistics, power is often called sensitivity. In the context of TPDT, power or sensitivity is the probability of rejecting the null hypothesis of no differences given that the two groups are indeed different.

We will further illustrate the concept of statistical power by a small worked example. We assume that we used a statistical test to predict a binary feature for 100 samples and got the test results shown in Table 3.3.

Table 3.3: Illustrative example for better understanding the concept of statistical power.

	H_1 is true	H_0 is true
H_0 rejected	60	20
H_0 not rejected	15	5

For this worked example, the power is equal to $\frac{60}{60+15} = 0.8$.

Receiver operating characteristic

The second measure which is commonly applied for test evaluation is specificity.

$$\text{spec} = \mathbb{P}(\text{do not reject } H_0 \mid H_0 \text{ is true}). \tag{3.24}$$

The interpretation of specificity in context of TPDT is the probability of not rejecting the null hypothesis of no differences given that there truly are no differences between both groups. In the worked example summarized in 3.3, the specificity is equal to $\frac{5}{5+20} = 0.2$.

The two values pow and spec are used to create an ROC curve which is a fundamental tool for test evaluation. It is created by visualizing pow and $(1 - \text{spec})$ for different cut-off values α. It holds that for very low values of α the specificity of a test will be close to 1 due to the null hypothesis almost never being rejected regardless whether it is true or not. At the same time the power of the test will be very low for the same reason. On the other hand, if α is chosen close to 1, specificity will be low and sensitivity or power large. Overall, it is desirable for a test to have high power for all specificity values. This can be measured by computation of the area under the ROC curve (AUC). AUC is a value $\in [0,1]$ and it holds that a larger AUC represents a test with higher predictive power. A value of AUC at 0.5 is equivalent to random guessing.

In the following, we employ the concept of ROC and AUC and reanalyse the artificial data simulated in 3.3 (see (3.14) - (3.16) and Table 3.2).

3.4.3 ROC comparisons

We apply the five discussed statistical tests (TPDT, Berk et al., Crainiceanu et al., adapted t-test and varying coefficients) on the simulated data and compare all datasets which were produced under H_0 with $\nu_i = 0, i \in \{1,\dots,4\}$ to datasets which were simulated under H_1 with $\nu_i > 0, i \in \{1,\dots,4\}$. We first perform an ROC analysis for two shift values $\nu_i = 0.5, i \in \{1,\dots,4\}$ and $\nu_i = 1, i \in \{1,\dots,4\}$ shown in Figure 3.9 and Figure 3.10, respectively.

The ROC curves for the four scenarios - vertical, horizontal, slope and combined shift - reveal interesting aspects about the test performance of the single tests. TPDT is outperforming the other tests in terms of AUC in most scenarios. While the varying coefficients model is comparable when the complete samples are shifted either horizontally or with a combination of all shifts, its performance drops substantially for the slope shift and the vertical shift scenario. For those two scenarios the adapted t-test is strongest competitor for TPDT and a clear second best test. While the joint intervals approach as well as the Moderated Ft statistic perform satisfactorily for the horizontal and combined shift scenarios, their performance clearly breaks down for the vertical and slope shift scenarios. Overall, these results strongly favour TPDT over the other tests.

For the next ROC curve analysis, we investigated datasets where a clearer difference between both groups was introduced with $\nu_i = 1, i \in \{1,\dots,4\}$. The results, shown in Figure 3.10, show higher AUC values as the difference in both groups is larger. TPDT is again outperforming the other tests for the vertical and slope shift scenarios. For the horizontal and combined shift, the AUC values associated to TPDT and the varying coefficients model are roughly equal and are slightly higher than the AUC values of the other three tests. In conclusion, the results for the larger shift of $\nu_i = 1, i \in \{1,\dots,4\}$ again point at TPDT as most appropriate test as it outperforms all other tests in two of the considered scenarios and is only matched by the varying coefficients model in the other two scenarios.

Overall, the results discussed in the previous and present sections make us confident of the strong performance of TPDT when studying time-resolved paired

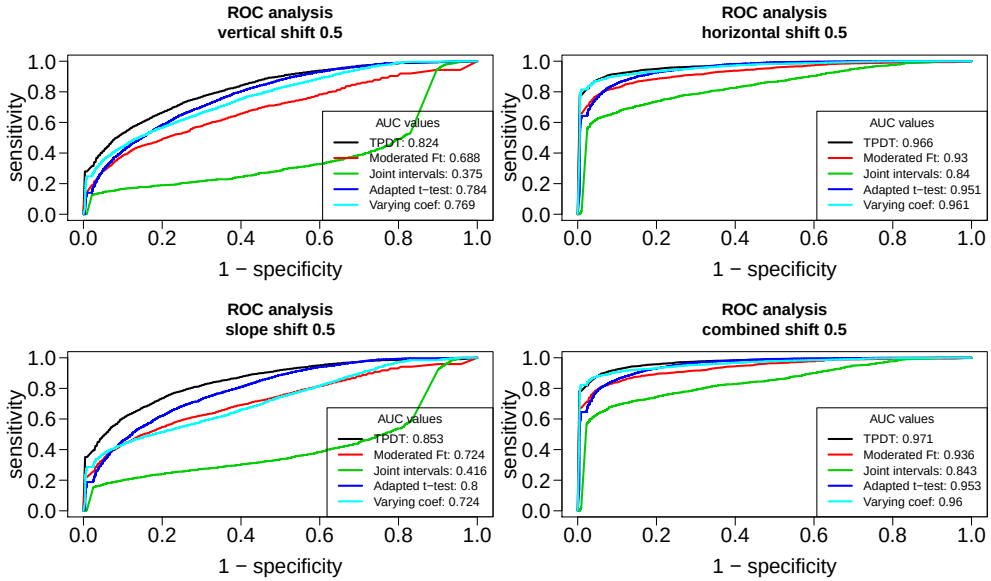

Figure 3.9: ROC curve comparison for $\nu_i = 0.5, i \in \{1,\ldots,4\}$ and different tests (black: TPDT, red: Moderated Ft statistic test (Berk et al.), green: Joint bootstrap confidence intervals test (Crainiceanu et al.), blue: Adapted t-test, cyan: Varying coefficients model) with artificially generated *paired* time-resolved samples. Results of four investigated scenarios vertical shift, horizontal shift, slope shift and combined shift are shown from left to right.

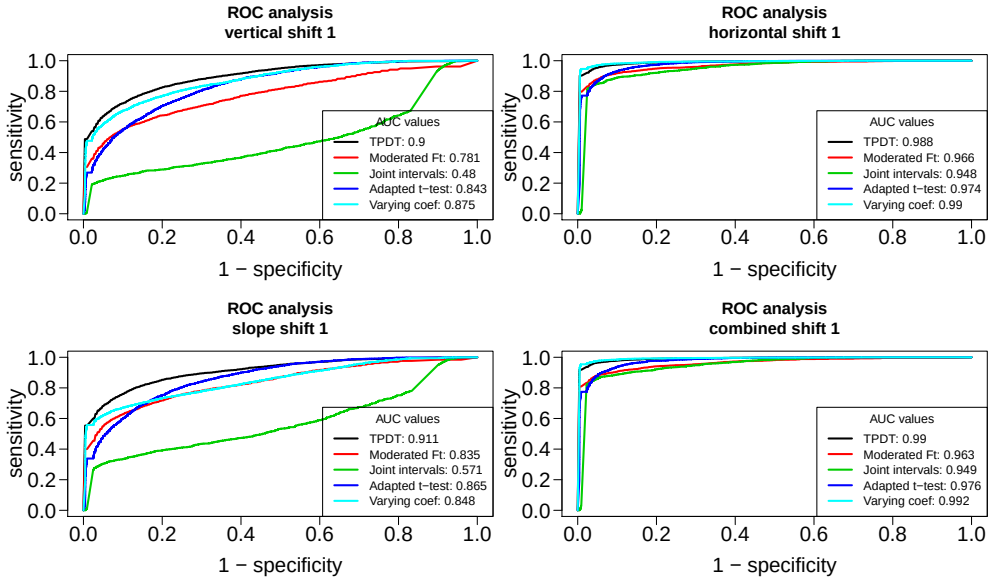

Figure 3.10: ROC curve comparison for $\nu_i = 1, i \in \{1,\ldots,4\}$ and the four different tests (black: TPDT, red: Moderated Ft statistic test, green: Joint bootstrap confidence intervals test, blue: Adapted t-test) with artificially generated *paired* time-resolved samples. Results of three investigated scenarios vertical shift, horizontal shift and slope shift are shown from left to right.

observations. In Section 3.5, we apply our test on data which is exactly of this format.

3.4.4 Power comparisons

As a last comparison of tests, we investigated the statistical power of each test for a grid of $\nu_i, i \in \{1, \ldots, 4\}$ values. Looking only at the power of a test is not of high value if it is the only measure of comparison. A hypothetical test which always rejects the null hypothesis regardless of whether the null hypothesis is true or not would result in a power value of 1. However, this hypothetical test would have no predictive value and would not be of real interest. In this simulation study, we already performed a ROC analysis and discussed results which show that the considered tests are associated with high AUC values and thus have predictive value. Therefore, an analysis of power in the present study is a valuable asset for a broader comparison of the considered tests for temporal differences.

The results of the power comparisons are shown in Figure 3.11. They convincingly demonstrate the dominating performance of TPDT as opposed to all other tests with exception of the varying coefficients model in terms of power. This is most obvious for the vertical and slope shift scenario where TPDT reaches a power above 0.5 already at very low values of ν_1 and ν_3 and clearly dominates the other tests. Compared to the varying coefficients model, we note slightly higher power for TPDT in the region $\nu_1 \leq 1$ and $\nu_3 \leq 1$. For larger value of the shift parameters, both test perform similarly. For the other two scenarios, the horizontal shift and the combined shift, the power of all tests is generally high as these are the scenarios where temporal differences are most distinguishable. Here, it holds that TPDT power rises above 0.9 even for small ν_2 or ν_4 values. For these scenarios the varying coefficients model has a similar performance in terms of power. Both tests reach a power very close to 1 for shift values $\nu_2 \geq 1$ and $\nu_4 \geq 1$ which indicates practically a correct test decision in almost all simulated examples.

In the following, we apply TPDT on two real-world data examples.

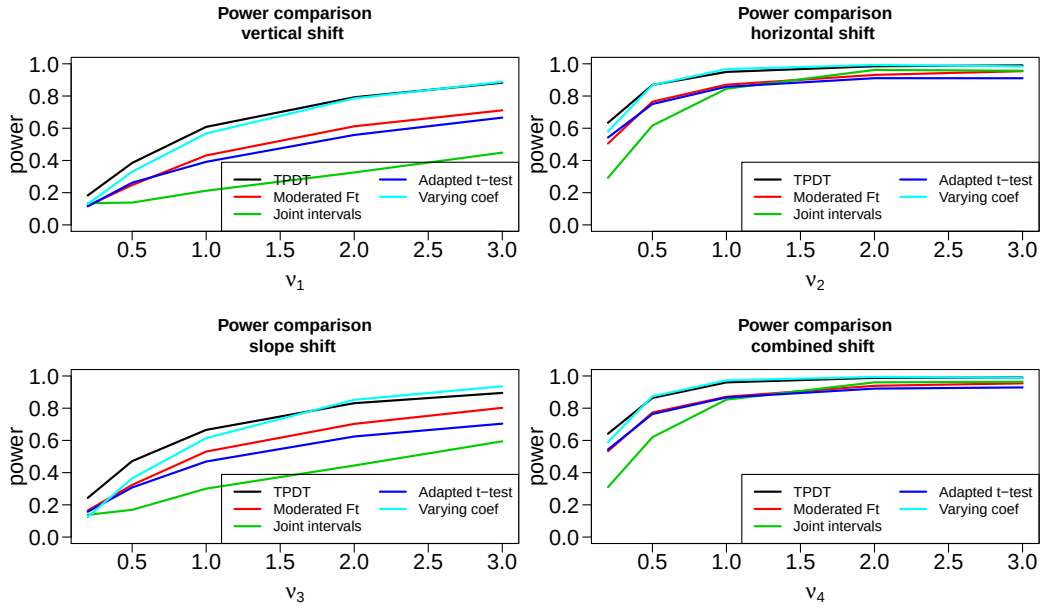

Figure 3.11: Comparison of power for the artificially created datasets in section 3.3. Tests indicated by different colours are applied and power (3.23) is computed at significance cut-off $\alpha = 0.05$. Results of four investigated scenarios vertical shift, horizontal shift, slope shift and combined shift are shown in the different panels.

3.5 Applications

The results presented in the previous two sections make us confident that TPDT will be applicable in general settings when analysing differences of two groups of time-resolved samples. In the following, we apply TPDT on real-world data.

3.5.1 Dietary effects on postprandial metabolism

In this subsection, we will apply TPDT on human metabolite data collected in a pilot study where probands were subject to four different dietary challenges:

1. Non-standardized Western Diet (NWD)
2. Standardized Western Diet (SWD)
3. Healthy Breakfast (HB)

4. Oral Lipid Test (OLT)

The data consists of six probands which took part in all four challenges on different days and blood metabolite measurements were collected at 0, 1, 2, 4, 6 and 8 hours after consumption of one of the four menus. In challenge HB, there were no measurements at the last time point of 8 hours. Probands were homogeneously chosen to be of the same gender (male), in the same age range (between 40 and 60 years old), non-smokers, perform less then 5 hours sports per week and have a BMI between 20 and 27. Metabolite data was collected using two platforms: targeted metabolomics measurements were carried out by using the AbsoluteIDQTM p180 kit (Römisch-Margl *et al.* [2012]; Zukunft *et al.* [2013]); non-targeted metabolomics profiles were measured using a previously described method of Metabolon Inc. (Evans *et al.* [2009]). Overall, after cleaning the data by omitting time-resolved samples with more than 50 % missing values, there were approximately 400 time-resolved metabolite variables for each challenge available.

Important research questions within this project were the quantification of time-resolved differences between the two menus NWD and SWD as well as further comparisons between NWD or SWD and the other two menus HB and OLT. The time-resolved measurements are paired due to data belonging to the same six probands in each dietary challenge. We thus have six group comparisons (NWD vs. SWD, NWD vs. HB, etc.) where we applied TPDT on each of the approximately 400 metabolite variables. In the following, we discuss results for the first comparison, namely NWD vs. SWD. In Appendix B, we also present the tabulated results of the other five challenges.

The comparison between NWD and SWD is of special interest. If we find many metabolites that have significant differences within the two challenges, this will mean that nutrition standardization is an important part of future study design in nutritional sciences. On the other hand, if we do not find evidence for differences in those two groups, we could postulate that a standardization of nutrition is not necessary in future study designs (given a low sample size and similar variation as observed in the present data). This in turn would mean that one can concentrate the often times limited financial resources on the data collection of one larger group rather than one standardized and one non-standardized. It would not only facilitate data collection but also would make data modeling and analysis less challenging.

The standardization of studies is a disputed topic in the field of nutritional science (Walsh *et al.* [2006]; Winnike *et al.* [2009]).

With TPDT, we analysed all metabolites of both platforms *biocrates* and *metabolon*. Due to performing the statistical test on different metabolites, we corrected the results for multiple testing. We did this by controlling the false discovery rate (FDR) (Benjamini & Hochberg [1995]). As a main result we only found differences in postprandial time-courses of isobutyrylcarnitine (adjusted p-value < 0.0305, see Figure 3.12). Time-courses show lower postprandial plasma-isobutyrylcarnitine levels after dietary standardization and similar trends in some subjects for leucine and isoleucine. Additionally, with the targeted metabolomics approach we measured acylcarnitines with a chain length of 4 carbons (C4), potentially also including isobutyrylcarnitine. However, C4 did not reach significance after FDR correction for multiple testing, although differences were indicated prior to the correction for multiple testing (non-adjusted p-value = 0.01). All other metabolite time courses were not associated with significant differences between the two challenges. An explanation of these significant differences for isobutyrylcarnitine may be the trend for higher carbohydrate and fiber intake in the standardization phase compared with the habitual diet of the study participants. Dietary fiber is discussed to delay nutrient absorption and, therefore, might lead to a higher local protein synthesis and oxidation in the small intestine (Pirman *et al.* [2007]; Ten Have *et al.* [2007]). Consequently, amino acid levels in the portal vein and plasma might be reduced (Ten Have *et al.* [2007]). Isobutyryl-CoA is known to be an intermediate of valine metabolism (Luís *et al.* [2011]), therefore, valine levels might be associated with isobutyrylcarnitine levels in plasma.

We see this result as an indication against standardization of nutritional challenges. However, we state this with caution since the study consisted of only 6 probands and we had a large number of tested metabolites (> 400), which led to strong multiple testing corrections. Furthermore, we want to stress that we do not interpret the results in the sense that both groups of metabolites are equal which would mean that not rejecting the null hypothesis means that it is true (this would be a question which can be answered with an equivalence test). Nevertheless, our suggestion remains to omit standardization of nutritional challenges if the data is collected in the course of a pilot study as it was the case in the present analysed data.

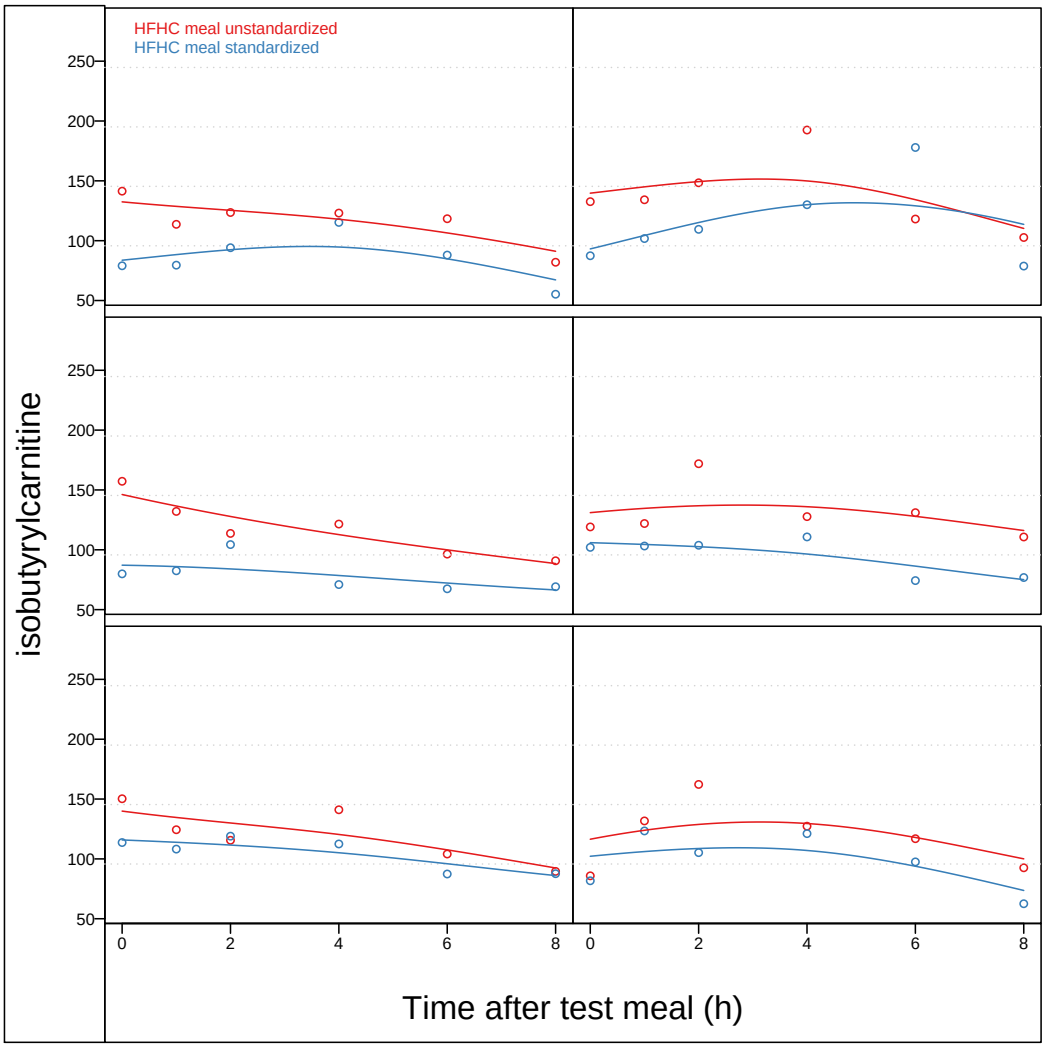

Figure 3.12: Postprandial time-courses of isobutyrylcarnitine. Time-courses are shown for metabolites measured with non-targeted metabolomics separately for each subject. Lines show the fitted smoothing splines of the high-fat, high-carbohydrate (HFHC) meal without previous dietary standardization (red) and after three-day dietary standardization (blue). Dots represent single measurements.

As stated above, we present results of the other five comparisons in Appendix B. There, we see multiple metabolites with significant differences, which is an indication that even with a low sample size and strong multiple testing correction, we are able to identify metabolites which may explain the differences in the selected nutritional challenges.

3.5.2 Promotion of heterochromatin formation at retrotransposons

The study aimed to better understand how retrotransposon sequences are targeted for silencing and which are the main contributors involved in both establishing as well as maintaining the heterochromatic state. This is of importance since more than 50 % of mammalian genomes consist of retrotransposon sequences and their silencing is crucial to verify genomic stability (Bourc'his & Bestor [2004]) and transcriptional integrity (Rowe *et al.* [2013]; Wilkins [2010]). Recent studies have also involved retrotransposons in development of human diseases, such as cancer (Helman *et al.* [2014]).

To identify novel players in heterochromatin establishment and maintenance on retrotransposons we chose the class of intracisternal A-particle (IAP) retrotransposons as a model system. We systematically tested sequence elements of IAP retrotransposons for their ability to induce heterochromatin formation and identified a small region of 160bp (SHIN) which is sufficient to trigger silencing. Based on this sequence we developed a small hairpin RNA (shRNA) screen and identified the chromatin remodeler Atrx as strong modifier of IAP silencing. Atrx was initially identified as the gene responsible for the X-linked alpha thalassemia / mental retardation (ATR-X) syndrome (Gibbons *et al.* [1995]).

In the course of the data analysis, we used TPDT to investigate the difference in the two functional groups (wild-type vs. Atrx knock-out embryonic stem cells) for several selected loci and corrected for multiple testing controlling the FDR (Benjamini & Hochberg [1995]). Single loci p-values are shown in Figure 3.13. The data we investigated was the consumption of the DNA in a specific locus which changed with increase of micrococcal nuclease (MNase) concentration. Thus, the time axis in the previous examples is now changed to a Mnase concentration axis

locus type	primer	u.statistic	p.value	fdr.corr.
IAP retrotransposons	IAP SHIN region	9.180	0.003	0.016
	IAP region 2	4.111	0.019	0.048
	IAP region 3	3.780	0.013	0.046
intergenic sites	intergenic Chr.1	10.150	0.001	0.005
H3K9me3 + Atrx	Polrmt	4.640	0.015	0.046
	Nnat	5.154	0.011	0.046
	Ezr	10.131	0.000	0.005
silent genes	F8	1.141	0.340	0.509
	Six3	0.275	0.869	0.869
	Tspan32	0.543	0.727	0.818
active gene	Oct4	2.428	0.071	0.127
major satelittes	major satellite	0.335	0.805	0.852
telomeric regions	Tel11	0.561	0.636	0.764
	Tel5	3.294	0.033	0.075
	TelX	1.095	0.367	0.509

Figure 3.13: Results of TPDT applied to study the differences between wild-type and Atrx knock-out embryonic stem cells and several selected genomic loci. Column 1: selected loci; column 2: primer; column 3: u statistic of TPDT; column 4: TPDT p-value; column 5: fdr-adjusted TPDT p-value. The yellow highlighted adjusted significant p-values all correspond to either IAP retrotransposons (blue) or intergenic sites (red). Control regions (gray) do not show significant differences between the two groups.

for analysis. For each locus, we had three replicates available and for each replicate we had wild-type as well as Atrx knock-out snapshots available, which means that the data is paired. An example for data corresponding to one such locus, is shown in Figure 3.14. Here, the MNase concentration measurements for the IAP SHIN region and the fitted smoothing splines for this data are visualized. With TPDT, we calculated an adjusted p-value of 0.016 for this example. Interestingly, we found significant differences only for IAP retrotransposons and intergenic sites but no significant differences for the control genes such as the primer for silent genes F8, Six3 and Tspan32 as well as all other tested regions.

The TPDT analysis contributed to the overall analysis of this data and to the conclusion of Atrx being crucial for fast and efficient establishment of heterochro-

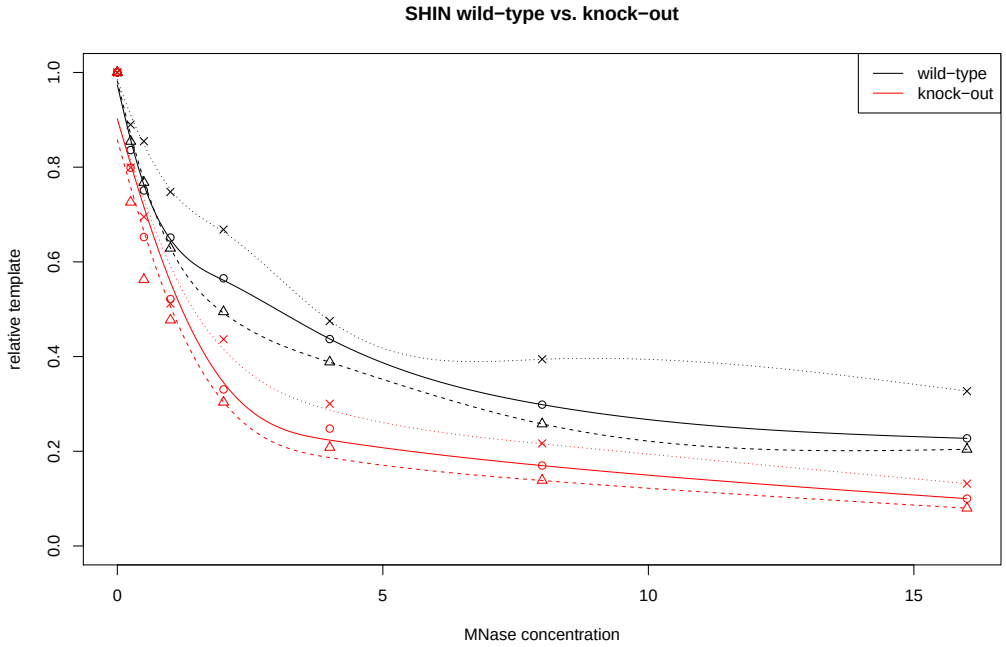

Figure 3.14: Raw data points and smoothing splines of IAP SHIN region. Different symbols and different types of lines correspond to different replicates. A significant difference (adjusted p-value 0.016 is found with TPDT for this locus.)

matin. Furthermore, we found that chromatin accessibility in Atrx knock-out cells was significantly increased on IAP elements, demonstrating that Atrx is important for proper heterochromatin organization. Thus the analysis of our data provides strong evidence for a general role of Atrx for establishment and robust maintenance of heterochromatin domains.

3.6 Discussion

We propose TPDT, a novel statistical test for assessing the difference in paired, time-resolved data. To our knowledge, this is the first test which is able to analyse time-resolved *and* paired samples based on the whole time scale (global decision) and also can be applied conveniently to assess the significance of multiple variables. We identified a demand for such a test and believe that it will be of great help when analysing paired, time-resolved samples. For interpretability, we set up the test in a similar fashion as a univariate t-test for paired samples and extended

the corresponding equations to a functional context. This allows us to extract a test statistic u, which quantifies the difference between two groups of samples. However, the distribution of this test statistic is unknown and thus we had to resort to resampling techniques to approximate the distribution of the test statistic. This was done by sampling from the null hypothesis while preserving the functional variability of the measurements. Overall, we are able to summarize the test result in a single approximated p-value, which in turn allows for easy interpretation and decision making with regard to the rejection of the null hypothesis of no difference between the two studied groups.

In simulations, we show that the test can be successfully applied on very general scenarios. We specifically investigate situations where artificially created data is subject to a high amount of noise, the number of observed time points is low and the number of samples per group is low and still are able to extract convincing results. Furthermore, we investigate ROC curves and power of the proposed test and also compare it to other commonly used significance tests. Here, we are able to outperform the other tests both in power and AUC in most simulations.

TPDT is applied on data arising from two major projects. In both projects, the data is exactly of the above-described nature: time-resolved and paired samples. This means that we can analyse the data using all the provided information only with the proposed test. First, we study the difference in nutritional challenges where a large number of metabolite measurements from two groups of human probands is collected. With TPDT, we are able to find only one out of several hundred significant metabolites, isobutyrylcarnitine, when studying the differences in the two nutritional challenges standardized and non-standardized Western Diet. For other challenge comparisons, we found a substantially higher number of metabolites which had significant differences. In the second project, we looked into the promotion of heterochromatin formation at retrotrasposons. Specifically, the establishment and maintenance of heterochromatin state was studied by comparing wild-type and Atrx knock-out embryonic stem cells and the digestion of several DNA loci by interaction with micrococcal nuclease. TPDT was applied together with multiple other methods to analyse the data. The results of the test suggest a significant difference between wild-type and knock-out cells for IAP retrotransposons and intergenic sites but no significant differences for all other investigated

loci. Together with other analysis tools we were able to conclude that Atrx is a major player for establishment and maintenance of heterochromatin domains.

Overall, the developed test allows for assessing the difference between two groups of time-resolved paired samples. It is easy to apply and thus accessible to a large community. Whenever this type of data is collected, we believe that TPDT is the most appropriate way to correctly analyse it.

4

Identifying latent dynamic components in biological systems

In systems biology, a general aim is to derive regulatory models from multivariate readouts, thereby generating predictions for novel experiments. We consider the case where a given model fails to predict a set of observations from a biological system with acceptable accuracy and ask the question whether this is due to the model lacking important external regulations. Examples for such external entities range from microRNAs to metabolic fluxes. This chapter describes the development of a novel method which aims to systematically extend biological networks by additional latent components. We demonstrate that the time course of this additional component can be inferred from data in a fully automatic way without using any prior knowledge or requiring manual method guidance by the researcher. As this identification of a latent component provides novel insights for extension of the structure of a given biological system, the result can be used as guidance for future experiments.

Our approach is a two-step procedure which is a combination of functional data analysis and differential equations. In the first step, we approximate raw observations arising from the biological network with splines and calculate the spline derivative. In combination with the network structure described by ordinary differential equations, a rough and unweighted estimate for the time course of the hidden component is calculated. In step two, we obtain final estimates for the time course of the hidden component by iteratively performing maximum likelihood

parameter estimation for ordinary differential equations. As an additional byproduct of the developed method, estimates for the noise level in the system as well as interaction strengths and directions between the hidden component and the rest of the network are obtained.

The method performance when dealing with typical difficulties associated to experimental data such as low number of observations, partially observed networks, high noise level and missing data are investigated via simulations. Results from these simulations show that a hidden component can successfully and consistently be inferred from data with a low error rate. The method is also applied to a signalling pathway model where we analyse real-world data and obtain promising results.

This chapter is based on and in part identical with the following publication:

- I. Kondofersky, C. Fuchs, and F.J. Theis (2015). Identifying latent dynamic components in biological systems. IET Systems Biology, 9, 193–203.

4.1 State of the art and research questions

We present a novel approach for extension of biological systems which is applied on time-resolved measurements arising from biological networks. As discussed in Chapter 2.1, for prediction of time-resolved, dynamical network behaviour, mathematical models are employed that typically involve several unknown parameters in addition to the network components. A popular modelling approach for time-resolved measurements is given by ODEs that represent the dynamics of and dependencies between the components of the network. The parameters describing the dynamics in an ODE must be inferred statistically, and in the case of several competing network models, the most appropriate model can be chosen by model selection methods. Hence, one deals with a mathematical modelling problem and a statistical estimation problem, simultaneously (Emmert-Streib *et al.* [2014]).

In such an analysis, ODEs directly arise from the network topology, i. e. the modeller specifies the components of the network and possible interactions. In many applications, the key elements of the dynamics of interest have been previously determined in various studies and are well-known from the literature. It is possible, however, that some interaction partners or connections remain unspecified. For

example, in addition to transcription factors modulating gene regulation, strong evidence indicates that microRNAs play an important role in transcription and translation processes (Hornstein & Shomron [2006]). Translation can also be influenced by external stimuli like drugs (Borowiak *et al.* [2009]; Chickarmane & Peterson [2008]; Lin *et al.* [2009]). Consequently, a mathematical model may be insufficient to explain the dynamics of interest, i. e. discrepancies with the measured data which are not simply due to measurement error may be evident even with the best model fit.

A promising way of addressing such discrepancies is given by employing additional network components to extend the proposed model. Our main focus in this chapter is systematic model extension. A substantial amount of work has been conducted in the past years in this field.

Ambroise *et al.* [2009] identify additional links in undirected graphs with Gaussian graphical models. These links represent model extensions and are systematically identified using an l_1-penalized likelihood. However, the proposed algorithm is not applicable to dynamical data.

Gao *et al.* [2008] and Honkela *et al.* [2010] also consider a model extension, this time for dynamical data. Similar to the approach to be presented here, they describe their models in terms of ODEs with a latent variable. Using Gaussian processes, they infer the time course of this variable and predict its behaviour. However, they do not model entire networks, which may possibly involve numerous links between components, but rather focus only on transcription and translation of single genes and on analytical solutions of the specific ODE models.

Furthermore, model extension by latent variables is utilized in the context of latent confounder modelling (Hoyer *et al.* [2006]; Ramb *et al.* [2013]). Here, the most frequently used method is structural equation modelling (SEM) (Bollen [1998]; Monecke & Leisch [2012]). SEM allows the identification of multiple latent variables and their relationship with observed variables by exploiting the data covariance structure. SEM is mainly formulated for single time points, and an extension to dynamical data is quite limited and often not possible.

In contrast to the just described model extensions, however, we do not want to change and possibly misspecify the ODE system but flexibly include a hidden influence and do this in a data-driven and systematic way.

In the present chapter, we address the problem of poor model quality in dynamical models by considering the effect of hidden influences on the network. We do not assume a functional form for the putative time courses of such hidden processes, but flexibly estimate their dynamics and interaction strengths. Wherever a hidden influence is observed that substantially improves the model's ability to represent the data, we attempt to provide a biological meaning with the help of experimental collaborators. Thus, we can guide the design of additional experiments in a detailed manner by providing a quantification of the hidden time courses as well as relative interaction rates between the hidden components and the existing network. The proposed method is applicable to Lipschitz continuous ODE models, e. g. gene regulation models or signal transduction models.

This chapter is organized as follows. We first present the ODE models considered and the means by which a hidden influence is included therein as well as a schematic representation of the developed method. The hidden component and the model parameters are statistically estimated in a two-step procedure. Furthermore, we discuss parameter uncertainty and model selection for the current setting. We then apply the developed technique to different scenarios and to a real-world data example – the JAK2-STAT5 signaling pathway. Finally, we conclude the chapter and discuss strengths and limitations of the proposed method.

4.2 Approach

In this section, we highlight the main ideas of the present study. Systematic network extension is illustrated by considering a small motif example. Next, we generalize this extension to networks of size N.

Consider a simple motif like the one presented in Figure 4.1. We use the schematic representation of small network motifs shown in Figure 4.1 to illustrate our method as follows. Figure 4.1A shows a simple network motif comprising two components x_1 and x_2, which influence each other, as indicated by the corresponding arrows. With the proposed method, we estimate a hidden component h, shown in Figure 4.1B, which may substantially contribute to the network dynamics, but was not previously considered. Thus, we call h a *hidden influence*. Figure 4.1C stresses that not all components must be observed.

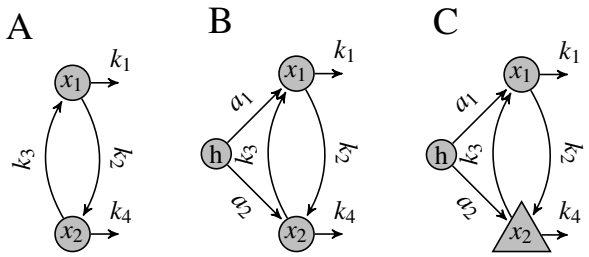

Figure 4.1: Example network motifs (circles: observed and hidden components; triangle: unobserved or indirectly observed component). **A**: a motif without hidden components, i. e. all components are observed. **B**: a motif with a single hidden component (h), where all other components are observed. **C**: a motif with a single hidden component (h) and partially observed components. Partially observed networks are discussed in Section 4.3.5.

We describe the network dynamics with ODEs. We assume them to be Lipschitz continuous; thus, the existence and uniqueness of an ODE solution are guaranteed. For motif A in Figure 4.1, the corresponding equations are

$$\dot{x}_i(t) \quad = \quad \psi_i(\boldsymbol{k}, \boldsymbol{x}(t)) \tag{4.1}$$

with parameter vectors $\boldsymbol{k} = (k_1, \ldots, k_L)^T$, $k_l \in \mathrm{R}_{\geq 0}$, non-negative state vector $\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T$, $x_i(t) \in \mathrm{R}_{\geq 0}$, derivatives with respect to time $\dot{x}_i(t)$, possibly non-linear functions $\psi_i : \mathrm{R}^p_{\geq 0} \times \mathrm{R}^N_{\geq 0} \to \mathrm{R}$ and suitable initial values $x_i(t_0)$, where $t \geq t_0$ represents the time. Equation (4.1) may also be represented with a stoichiometry matrix and flux function as already discussed in Equation (2.28). The functions ψ_i generate the network structure and may include several combinations of the state variables $\boldsymbol{x}(t)$ such as linear combinations, Michaelis-Menten kinetics, complex formation and others. The connection between the state variables is described by the parameters $\boldsymbol{k}$.

The components $x_i(t)$ may be observed or unobserved. In addition to the motif in Figure 4.1A, we now assume a time-varying hidden component $h(t)$ that acts linearly on $\dot{x}_i(t)$, as shown in Figure 4.1B. The system of differential equations then changes to

$$\dot{x}_i(t) \quad = \quad \psi_i(\boldsymbol{k}, \boldsymbol{x}(t)) + a_i h(t) \tag{4.2}$$

with weights $\boldsymbol{a} = (a_1, \ldots, a_N)^T$, $a_i \in \mathrm{R}$. Positive weights a_i in this context represent activation of the i-th component, whereas a negative value of a_i implies inhibition.

A similar model was considered, e. g. in Blöchl & Theis [2009]. Other than for $x_i(t)$, we do not assume any parametric structure for the hidden component. The time course of $h(t)$ cannot be observed directly.

Six elements determine the model: the components x_i and their time derivatives $\dot{x}_i$, the parameter vectors $\boldsymbol{k}$ and $\boldsymbol{a}$, the dependency describing functions ψ_i and the hidden influence h. We will extend established models from the literature by adding hidden components and applying our estimation method described in Section 4.3. For reasons of simplicity, we assume that the reaction rates $\boldsymbol{k}$ and dependency functions ψ_i are known. Both assumptions can also be relaxed, as is demonstrated later in Section 4.4.4 where we additionally estimate $\boldsymbol{k}$ and recover a missing feedback, thus altering the network structure.

The objective of our study is to estimate a_i and $h(t)$, and examine if they improve the ability of the model to represent the data; this also requires the estimation of $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$. In our analysis we exploit the following connection between h and all other components:

$$h(t) = \frac{\dot{x}_i(t) - \psi_i(\boldsymbol{k}, \boldsymbol{x}(t))}{a_i} \tag{4.3}$$

for all t and i with $a_i \neq 0$. The hidden influence can then be estimated according to two major steps as follows. First, we fit penalization splines to the measurements of $\boldsymbol{x}$. This allows a direct computation of the time derivatives $\dot{\boldsymbol{x}}$ such that the right-hand side of Equation (4.3) is known up to a scaling factor a_i. These factors are then estimated via likelihood maximization, utilizing the differential equation structure. A flowchart that illustrates the details of the developed method is shown in Figure 4.2 on page 81.

4.3 Methods

This section describes the above-mentioned two-step procedure for the estimation of the hidden influence h. As a basis, we assume observations $x_i^{\text{obs}}(t_j)$ of all components $x_1, \ldots, x_N$ at discrete time points $t_0, \ldots, t_n$. The first step is presented in Section 4.3.1, where we use spline functions to approximate the time courses of $x_1, \ldots, x_N$ and their time-derivatives. In a second step, we define a noise model for the data and estimate the weights a_i using likelihood maximization in Section 4.3.2. In Section 4.3.3, we discuss uncertainty and the fit quality of the

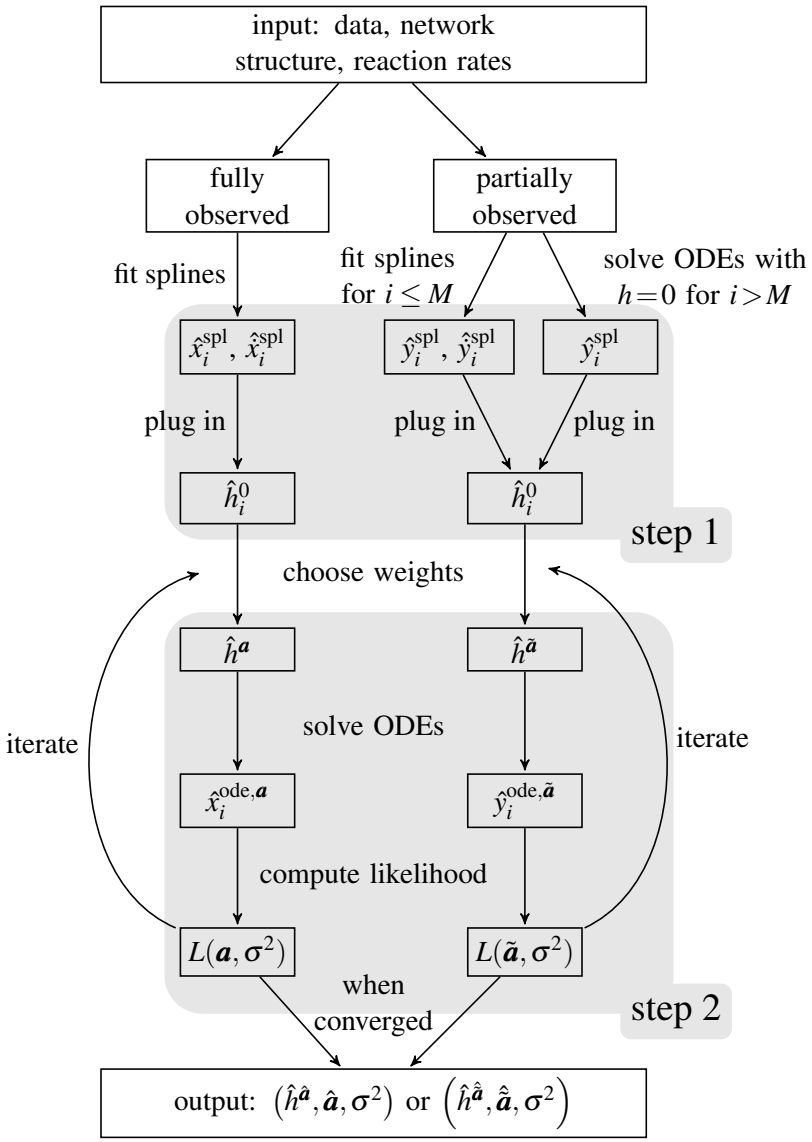

Figure 4.2: Flowchart illustrating the details of the developed method. The method involves estimating hidden components according to two major steps (indicated by gray boxes). We distinguish between fully and partially observed networks. If at least one network component is unobserved, the procedure requires a preliminary step where the system of ordinary differential equations (ODEs) is solved without considering a hidden component. As a next step in both scenarios, penalization splines are fitted to the measurements using cross-validation for the estimation of the smoothing parameter. Finally, a maximum likelihood loop is performed until convergence to estimate the time course of the hidden component and its interaction weights as well as the noise parameter σ^2.

parameters of interest. Next, in Section 4.3.4, we perform model selection on the considered networks. Finally, in Section 4.3.5, we extend the estimation methods to the case of not fully but partially observed networks. The different steps are shown in Figure 4.2 where we highlight the two-step procedure with grey background boxes and summarize the whole method with respect to network observability by dividing the flowchart into a left (fully observed network) and right (partially observed network) branch.

4.3.1 Spline estimation for observed time courses and their time-derivatives

As discussed in Chapter 2.1.2 splines are a convenient way to approximate the time course of a series of measurements in a functional form and are successfully used to model time-resolved, biological data, e. g. by Bar-Joseph *et al.* [2003]. Recall, that they arise as a linear combination of known basis functions and basis coefficients as described e. g. in De Boor [2001]. Applied to the model given in (4.2), for a given smoothing parameter λ_i, the basis coefficients $\beta_{1i},\ldots,\beta_{Ki}$ are chosen such that the following term is minimized (Ramsay & Silverman [2005]):

$$\left(\boldsymbol{x}_i^{\text{obs}} - \sum_{k=1}^{K} \beta_{ki}\boldsymbol{\phi}_k\right)^T \left(\boldsymbol{x}_i^{\text{obs}} - \sum_{k=1}^{K} \beta_{ki}\boldsymbol{\phi}_k\right) + \lambda_i \int_{t_0}^{t_n} \left(\sum_{k=1}^{K} \beta_{ki}\ddot{\phi}_k(s)\right)^2 ds. \tag{4.4}$$

In this notation, $\boldsymbol{x}_i^{\text{obs}} = (x_i^{\text{obs}}(t_0),\ldots,x_i^{\text{obs}}(t_n))^T$ is the vector of the measured data, and the basis functions evaluated at the observation times are denoted by $\boldsymbol{\phi}_k = (\phi_k(t_0),\ldots,\phi_k(t_n))^T$. As the index i in λ_i suggests, a penalization parameter is chosen for each component separately.

In this study, we choose a sufficiently large K and estimate λ_i using leave-one-out cross-validation. See Chapter 2.1.3 for more details on cross validation.

Minimization of (4.4) yields optimal coefficients $\hat{\beta}_{ki}$, and consequently an approximation for the time course of the observed components

$$\hat{x}_i^{\text{spl}}(t) := \sum_{k=1}^{K} \hat{\beta}_{ki}\phi_k(t) \tag{4.5}$$

and their time derivatives

$$\dot{\hat{x}}_i^{\text{spl}}(t) := \frac{\partial \hat{x}_i^{\text{spl}}(t)}{\partial t} = \sum_{k=1}^{K} \hat{\beta}_{ki} \dot{\phi}_k(t). \tag{4.6}$$

The estimation of the time derivatives given by (4.6) plays an equally important role as the estimation of the splines given by (4.5) for the final estimation of the time course of the hidden component given by (4.3). Thus, particularly for the analysis of very noisy data, additional smoothing techniques, such as a higher penalization order in (4.4), may be considered. The results presented in Section 4.4 are based on cubic B-splines defined on an equally spaced time grid. Recall (Chapter 2.1.2) that these functions are twice continuously differentiable such that the penalization term in (4.4) is well-defined. In practice, the integral in (4.4) is approximated through finite differences.

With the approximations given by (4.5) and (4.6), we can now estimate the numerator in (4.3):

$$\hat{h}_i^0(t_j) = \dot{\hat{x}}_i^{\text{spl}}(t_j) - \psi_i(\boldsymbol{k}, \hat{\boldsymbol{x}}^{\text{spl}}(t_j)). \tag{4.7}$$

For the estimation of the denominator in (4.3), we apply a likelihood approach, as described in the following.

4.3.2 Maximum likelihood estimation

Given a weight vector $\boldsymbol{a}$, the hidden influence can be approximated as

$$\hat{h}^{\boldsymbol{a}}(t_j) = \frac{1}{N^{\boldsymbol{a}}} \sum_{\{i: a_i \neq 0\}} \frac{\hat{h}_i^0(t_j)}{a_i}, \tag{4.8}$$

where $N^{\boldsymbol{a}}$ is the number of non-zero weights a_i. The case $\boldsymbol{a} = \boldsymbol{0}$ can be excluded without loss of generality because it indicates the absence of a hidden influence extending the network. This approximation of h will later be plugged into (4.2) where it is multiplied with a_i. If $\hat{h}_i^0$ is the true numerator of (4.3), the time courses $\hat{h}_i^0$ on the right side will all be identical up to a scaling factor. However, because it is an approximation there will be differences between them in practice. For this reason we consider the pointwise weighted average in (4.8), which presents a natural choice of a summary statistic. If it holds that the single estimates $\hat{h}_i^0(t_j)$

strongly differ from each other, then the weighted average $\hat{h}^{\boldsymbol{a}}(t_j)$ will be inaccurate, and this in turn will be reflected in the likelihood function and the corresponding information criterion that we later formulate in (4.13) and (4.16), respectively, thus leading to the rejection of the proposed model.

Plugging in h into (4.2) and multiplying it by a_i introduces a non-identifiability. Because of

$$a_i \hat{h}^{\boldsymbol{a}} = (\xi a_i) \left(\frac{\hat{h}^{\boldsymbol{a}}}{\xi} \right) \tag{4.9}$$

for any $\xi \neq 0$, the weights a_i are non-identifiable. For this reason, we restrict $\boldsymbol{a}$ to $\sum_i \mid a_i \mid = 1$. In the special case of a network consisting of only one component x_1, we only estimate the interaction direction of the hidden influence, i. e. $a \in \{-1, 1\}$.

In most biological applications, the data contains noise of different origins, such as measurement noise or technical noise (Paulsson [2004]; Raser & O'Shea [2005]). The most common assumption is that measurement errors are independent and normally distributed with mean zero and constant variance $\sigma^2 > 0$. This was already introduced and discussed in (2.30) in Chapter 2.1.4. Translated to the notation in this chapter, the assumption reads as:

$$x_i^{\text{obs}}(t_j) = x_i(t_j) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathbb{N}\left(0, \sigma^2\right). \tag{4.10}$$

In applications, $x_i(t_j)$ often has a positive domain, and in this case, (4.10) might be ill-defined. Note that we do not restrict our methods to only this type of noise. In Appendix A.1, we also specifically derive all the equations given in this section for log-normally distributed multiplicative noise. The distribution of ε_{ij} immediately propagates to the measurements:

$$x_i^{\text{obs}}(t_j) \big| x_i(t_j) \overset{\text{iid}}{\sim} \mathbb{N}\left(x_i(t_j), \sigma^2\right). \tag{4.11}$$

While the true time course $x_i(t)$ is unknown, it has already been approximated in (4.5). This approximation, however, does not contain any information about $\boldsymbol{a}$, which we seek to estimate in the following. Hence, we introduce another approximation for $x_i(t)$, this time exploiting the ODE structure given in (4.2): For a given $\boldsymbol{a}$, we plug in $\hat{h}^{\boldsymbol{a}}$ from (4.8) into the ODE given in (4.2) and solve the differential equations either analytically or numerically, as described e. g. in Ross [1984].

This yields $\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)$ and leads to the approximate distribution

$$x_i^{\text{obs}}(t_j)\big|\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j) \overset{\text{iid}}{\sim} \mathbb{N}\Big(\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j), \sigma^2\Big). \tag{4.12}$$

Overall, we arrive at the conditional likelihood function

$$L(\boldsymbol{a}, \sigma^2 \mid x_i^{\text{obs}}(t_j)) = \prod_{i=1}^{N}\prod_{j=0}^{n} f_{\boldsymbol{a},\sigma^2}(x_i^{\text{obs}}(t_j)), \tag{4.13}$$

where $f_{\boldsymbol{a},\sigma^2}$ is the probability density function corresponding to the chosen error specification.

Additionally, a conditional estimate for σ^2 can be derived analytically:

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{N(n+1)} \sum_{i=1}^{N}\sum_{j=0}^{n} \Big(x_i^{\text{obs}}(t_j) - \hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)\Big)^2. \tag{4.14}$$

The parameters $\boldsymbol{a}$ and σ^2 are jointly estimated using (4.14) and a numerical optimization of (4.13). Furthermore, unknown initial conditions $\boldsymbol{x}(t_0)$ are treated as unknown parameters and are equivalently estimated.

4.3.3 Parameter uncertainty

We further explore our likelihood approach with respect to parameter uncertainty. The overall estimation performance of the unknown parameters σ^2 and $\boldsymbol{a}$ can be analysed by calculating the Cramer-Rao lower bound (CRLB) (Cramér [1945]; Rao [1945]) which is defined as the inverse expected Fisher information matrix. This theoretical value describes a lower bound for the mean squared error (MSE) of a given parameter. To that end, we look at the diagonal elements of the expected Fisher information matrix, which, in the case of a normally distributed error, have the following form:

$$I_k(\boldsymbol{a}, \sigma^2) = \begin{cases} \frac{1}{\sigma^2}\sum_i\sum_j \Big(\frac{\partial}{\partial a_k}\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)\Big)^2 & k \leq N \\ N(n+1)/\big(2\sigma^4\big) & k = N+1. \end{cases} \tag{4.15}$$

In practice, we solve the ODEs numerically. Here, a sophisticated ODE solver,

such as the Runge-Kutta 4th order method (Butcher [1987]) can be employed to produce accurate estimates. However, an analytical derivation of the CRLB for such a method becomes very complex because of the complicated recursive formulation of the ODE solution. For exemplary purposes, we outline a derivation for a specific small example using the Euler method for solving the ODEs in Appendix A.2.

Large values on the diagonal of the expected Fisher information matrix represent parameters with a small CRLB. These parameters can be estimated accurately with an (asymptotically) efficient estimator. For the parameters a_l, the respective l-th diagonal element increases if

- σ^2 is small, i. e. the data are subject to a small amount of noise,
- $\left(\frac{\partial}{\partial a_k}\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)\right)^2$ is large, i. e. the ODE solution is sensitive to changes in the parameter a_k and
- n and/or N are large, i. e. the data arise from a large number of time points and different (observed) species.

For the parameter σ^2, we look at the $(N+1)$th diagonal element of (4.15), which increases if

- σ^2 is small, i. e. the data are subject to a small amount of noise and
- n and/or N are large, i. e. the data arise from a large number of time points and different (observed) species.

We can conclude that, as expected, the estimation accuracy will suffer if we apply our method to small networks, few observations, conditions indicative of a weak influence of the hidden component and large noise. As indicated in Section 4.3.2, we estimate the parameters with a maximum likelihood approach. The estimation is asymptotically efficient (Zacks [1971]); thus, the CRLB is asymptotically achieved. However, the approximation of the time courses using splines as described in Section 4.3.1, introduces additional uncertainty. In Appendix A.2, we examine this loss of accuracy for a given showcase network and various parameter combinations, thereby concluding that our method produces estimates that are close to the CRLB.

4.3.4 Model selection

The vector $\boldsymbol{a}$ controls the interaction strength between the hidden influence h and the network components x_i. If a weight a_i is estimated to be close to zero, it will have a negligible effect on the network and will probably improve the model fit only slightly. In such a case, one may ask whether the inclusion of this parameter a_i is worth the involved estimation effort or whether one should simply set this component equal to zero, thus reducing the complexity of the model.

We already discussed in Chapter 2 that for quantification of the trade-off between improved model fitting and increased model complexity the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) can be considered:

$$\begin{aligned} \mathrm{AIC}(\hat{\boldsymbol{\theta}}) &= -2\log(L(\hat{\boldsymbol{\theta}})) + 2\dim(\hat{\boldsymbol{\theta}}) \\ \mathrm{BIC}(\hat{\boldsymbol{\theta}}) &= -2\log(L(\hat{\boldsymbol{\theta}})) + \log((n+1)N)\dim(\hat{\boldsymbol{\theta}}). \end{aligned} \tag{4.16}$$

In these equations, $\hat{\boldsymbol{\theta}}$ denotes a vector containing all parameter estimates, $L(\hat{\boldsymbol{\theta}})$ is the likelihood function (4.13) evaluated at $\hat{\boldsymbol{\theta}}$ and $\dim(\hat{\boldsymbol{\theta}})$ is the number of estimated parameters.

To consider the complexity of the overall estimation procedure, the vector $\boldsymbol{\theta}$ can be chosen to include all unknowns determined in our two-step approach, i. e. all λ_i, β_{ik} and σ^2. In our considerations, however, the number of variables is constant apart from the number of non-zero a_i. Hence, we can replace $\dim(\hat{\boldsymbol{\theta}})$ by $N^{\boldsymbol{a}}$ as defined in (4.8), to compare different models.

The models that we are considering with our method are all of a nested type. The special case of $\boldsymbol{a} = \mathbf{0}$ is the null model and is nested within all other models with arbitrary $\boldsymbol{a}$. Regarding the decision of which values a_i to set equal to zero, we follow three conventional variable selection methods: best subset selection, forward stepwise selection and backward stepwise selection. See Chapter 2 for more details.

In the forward stepwise selection, we begin with the model given in (4.2), which contains no interactions between the hidden influence and the network components, i. e. all a_i equal zero. In the second step, N models are estimated, where, for each of the models a different element of $\boldsymbol{a}$ is non-zero while the others are held

equal to zero. If the best model outperforms the selected model from the previous step, this model is accepted, and in the subsequent step, another component of $\boldsymbol{a}$ is set to a non-zero value. This step is repeated until no increase in model performance is achieved with a more complicated model. Once a component a_i is chosen to be non-zero, it will remain non-zero in all subsequent steps.

Backward stepwise selection is an analogy of forward stepwise selection wherein the initial model selected is the most complicated model for which all interactions between the hidden and the other components are estimated. In each subsequent step, a single entry of $\boldsymbol{a}$ is fixed to zero until no lower value of AIC/BIC is achieved.

Finally, in the best subset selection, the AIC or BIC is computed for all possible models, and the model with the best score is chosen.

In Section 4.4, we employ the BIC for model choice on synthetic and real data because this criterion penalizes the model complexity more than the AIC.

4.3.5 Partially observed network components

In the estimation procedure discussed in Section 4.3.1 and Section 4.3.2, we assumed that the components x_i were *directly* observed and that *all* of them were observed. In the following, we consider the case where the observed time courses are affine linear transformations $y_1, \ldots, y_M$ of $x_1, \ldots, x_N$ and the number M of observed time courses is smaller than the total number of network components N. The flowchart shown in Figure 4.2 illustrates the single steps of the estimation procedure. Chapter 2.1.4 outlined a general strategy how to deal with parameter estimation for partially observed networks.

As an example for non-direct observations in the context of the current chapter, consider the motif depicted in Figure 4.1C. It is assumed to follow exactly the same dynamics as that in Figure 4.1B and can therefore be described in terms of the ODEs given in Equation (4.2). Suppose that one can now only measure time courses of the observation functions $y_1(t) = x_1(t)$ and $y_2(t) = bx_2(t) + c$ for scalars $b \neq 0$ and c. The ODEs given in (4.2) can then be translated to

$$\dot{y}_m(t) = \eta_m(\boldsymbol{\kappa}, \boldsymbol{y}(t)) + \tilde{a}_m h(t) \tag{4.17}$$

with appropriate η_m, $\tilde{a}_m$ depending on $\boldsymbol{a}$, b and c and $\boldsymbol{\kappa}$ being the collection of the interaction rates $\boldsymbol{k}$ and transformation parameters b and c. Because the observation functions $\boldsymbol{y}$ are affine linear transformations of the network components $\boldsymbol{x}$, we can extract the hidden influence following Equations (4.7) and (4.8):

$$\hat{h}^{\tilde{\boldsymbol{a}}}(t_j) = \frac{1}{N^{\tilde{\boldsymbol{a}}}} \sum_{\{i:\tilde{a}_i \neq 0\}} \frac{\hat{y}_i^{\mathrm{spl}}(t_j) - \eta_i(\boldsymbol{\kappa}, \hat{\boldsymbol{y}}^{\mathrm{spl}}(t_j))}{\tilde{a}_i}. \tag{4.18}$$

Note that, for non-linear observation functions, we cannot directly apply our method possibly due to, for example, quadratic or higher order terms of $h(t)$ in Equation (4.17); however, in the above case, one can proceed in a manner analogous to that given in Sections 4.3.1 and 4.3.2 for the estimation of h and $\tilde{\boldsymbol{a}}$ if both y_1 and y_2 are observed.

As an example of partial observation, we can assume that only y_1 is observed. Because the two-dimensional ODE system given in (4.1) contains no redundant equation, the dynamics of interest are fully described by a network of only two components. Hence, in addition to the observed variable y_1, we include one latent component y_2 in our analysis, e. g. $y_2 = x_2$ or $y_2 = bx_2 + c$, as discussed above. More generally, we consider a network with observed components $y_1, \ldots, y_M$ and unobserved components $y_{M+1}, \ldots, y_N$. The estimation of a hidden influence and its weights changes slightly as opposed to the fully-observed case because there is no spline approximation possible for the time courses of $y_{M+1}, \ldots, y_N$.

In this case, we approximate $y_1, \ldots, y_M$ and their derivatives as before (see (4.5) and (4.6)). Furthermore, we approximate $y_{M+1}, \ldots, y_N$ by their solutions of the N-dimensional ODE system given by (4.17) with $h \equiv 0$. For simplicity, we denote these approximations by $\hat{y}_i^{\mathrm{spl}}$ for all i, although there are no splines involved for $i > M$. The starting values $y_i(t_0)$ are treated as additional unknown parameters. Because of the ODE-based derivation of $\hat{y}_i^{\mathrm{spl}}$ for the latent variables, estimation of the corresponding $\tilde{a}_i$ is not feasible in the first step of the estimation procedure. Hence, we restrict the components of $\tilde{\boldsymbol{a}}$ to be zero for $i > M$. For a given weight vector, the hidden influence is estimated through (4.18). In the second step, the likelihood function results as in (4.13) as a product over all observed components ($i \in \{1, \ldots, M\}$) and observation times ($j \in \{0, \ldots, n\}$). Maximization of the likelihood function yields estimates for h and $\tilde{\boldsymbol{a}}$ for all $i \in \{1, \ldots, N\}$.

4.4 Simulation studies

In this section, we demonstrate several different applications of our method. In Section 4.4.1, the prediction of the time course of a hidden component is evaluated. In Section 4.4.2 and Section 4.4.3 we present further simulations which concern the method performance when dealing with unusual shaped time courses of hidden components or missing data, respectively. Finally, in Section 4.4.4, we present our method as a tool that guides the reconstruction of a previously misspecified network.

4.4.1 Synthetic examples with unimodal latent components

To evaluate the performance of our method, we conduct several simulation studies. All test runs are performed with the statistical software R (R Development Core Team [2011]). We examine the robustness of our method by varying the noise intensity of the simulated data. Additionally, we evaluate networks of different sizes and study the dependence of the results on the number of unobserved components.

The parameters $\boldsymbol{k}$ and $\boldsymbol{a}$ are chosen at random for each simulation run, and conditioned on these, we generate artificial data at 30 equally spaced time points. We use log-normal noise (see Appendix A.1), and the three noise levels that we consider are low ($\sigma = 0.01$), medium ($\sigma = 0.1$) and high ($\sigma = 0.3$). In the simulation, we allow only linear interactions between the network components which, indicates that the structure of the ODEs can be summarized as

$$\dot{x}_i(t) = \sum_{u=1}^{N} \left(k_{iu} x_u(t) - k_{ui} x_i(t)\right) + a_i h(t) \tag{4.19}$$

with uniformly distributed k_{iu} in $[0,1]$ for describing the reaction strength between the i-th and u-th component and uniformly distributed a_i in $[-1,1]$.

We use the same hidden influence for each simulation run, thus producing comparable results. After application of our estimation procedure, the resulting fit quality is measured by:

$$s = \frac{1}{n+1} \sum_{j=0}^{n} \mid \hat{h}(t_j) - h(t_j) \mid . \tag{4.20}$$

We estimate rates $\boldsymbol{a}$ with the forward selection technique. As illustrated in Figure 4.3, results of 100 simulations indicate that a smaller network size and a smaller fraction of observed components lead to increasingly poor model fitting performance.

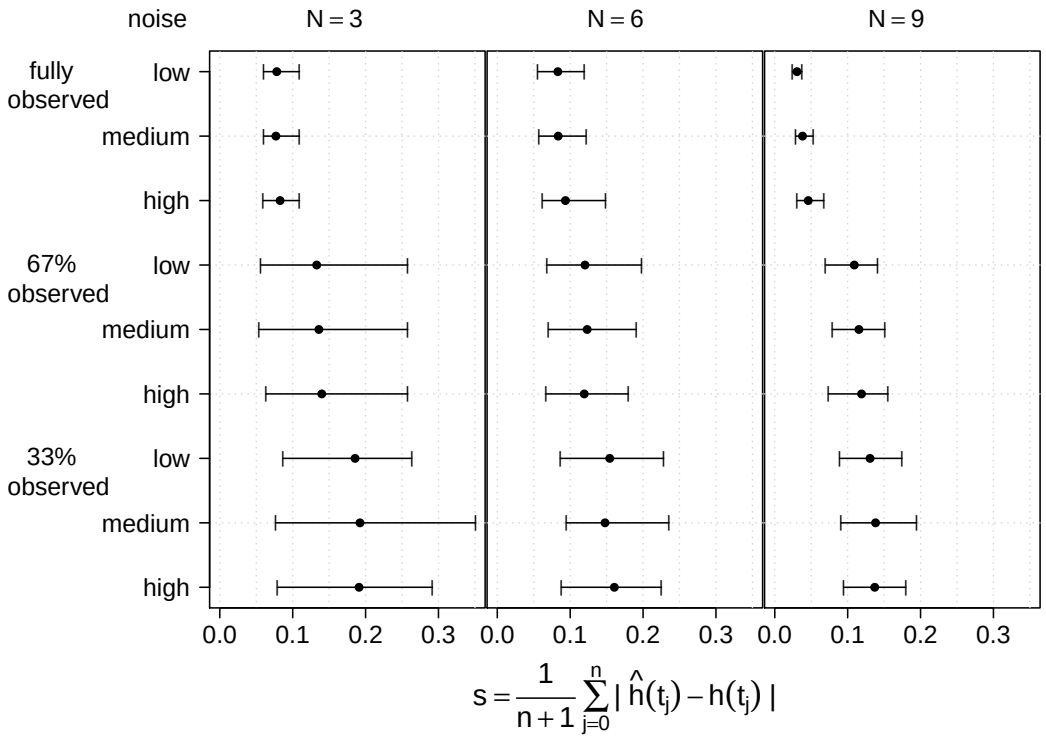

Figure 4.3: Description of the simulation studies used to evaluate the performance of our method. For 27 different combinations of network size (3, 6 and 9 components), noise intensity ('low' $\equiv \sigma = 0.01$, 'medium' $\equiv \sigma = 0.1$ and 'high' $\equiv \sigma = 0.3$) and ratio of observed to unobserved components (100% observed, 67% observed and 33% observed), 100 different simulated networks are created and the mean as well as the 5% and 95% quantiles of the error measurement in (4.20) are displayed for each combination. All interaction rates between components are chosen randomly. It holds that, the smaller the value of s, the better the estimated time course.

Only small differences are observed between low and high noise intensities, indicating that our method can accommodate a high degree of noise while extracting the relevant information from the data. Additionally, it appears that the network size plays only a minor role with regard to the estimation quality of our method because the scores for larger networks decrease only slightly.

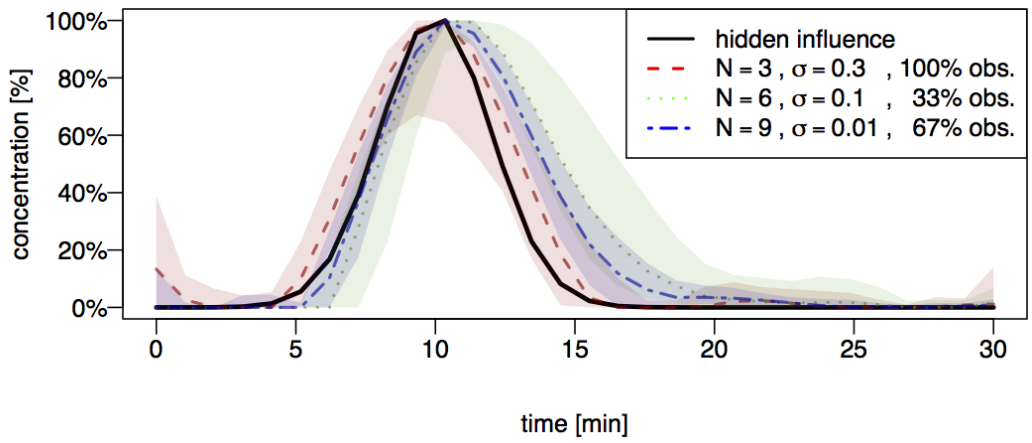

Figure 4.4: Time course of the hidden influence used in the simulations (black solid line) and the mean and 5% and 95% pointwise quantile courses of three exemplary simulation scenarios with different parameters (dashed lines) defined as follows: a fully observed network of size 3 with high noise intensity (red); a partially observed network (33%) of size 6 with medium noise intensity (green); a partially observed network (67%) of size 9 with low noise intensity (blue). Mean and confidence intervals are based on 100 estimates $\hat{h}(t)$.

Our approach also yields estimates of the time course of the hidden component, which we can compare with the true hidden component used to generate the data. Figure 4.4 shows the mean and 5% and 95% pointwise quantile time courses of three exemplary simulation scenarios. The shape of the hidden influence is reproduced satisfactorily, albeit differently. For a network comprising 3 components and a high degree of noise, the estimates produce additional fluctuations that are not present in the true time course and the confidence intervals are very broad. For a larger network size (6 or 9 components), the estimates become more stable and recover the peak of the true time course; however, the second part of the peak is slightly overestimated due to the network being partially observed.

In the following, we continue with extensive simulation from the same network as in Equation (4.19) and focus on a comparison between unimodal and bimodal shapes of the hidden influence.

4.4.2 Synthetic examples with bimodal latent components

As a further test for our method on a larger number of scenarios we design another simulation similar to Section 4.4.1. Networks are simulated according to Equation (4.19), which is formulated as

$$\dot{x}_i(t) = \sum_{u=1}^{N} (k_{iu}x_u(t) - k_{ui}x_i(t)) + a_i h(t). \tag{4.21}$$

The parameters $\boldsymbol{k}$ and $\boldsymbol{a}$ are chosen at random for each simulation run (total 100 simulation runs); conditioned on these, we generate artificial data at 100 equally spaced time points resulting in longer time series than the simulations in Section 4.4.1. We use log-normal noise and the three noise levels we consider are once again low ($\sigma = 0.01$), medium ($\sigma = 0.1$) and high ($\sigma = 0.3$). Here, we additionally investigate a bimodal hidden influence shape and compare the estimation quality to the unimodal case from Section 4.4.1.

In Figure 4.5 we show the results of this simulation alongside with confidence intervals of the hidden time course estimates. We confirm the finding that lower noise and higher network size lead to the best results. Additionally we can conclude that the signal in the sense of time course shape (regardless whether bimodal or unimodal) is successfully recovered in all simulations. The estimation of the bimodal time course has a tendency to be slightly worse. Additionally, we note a considerably worse estimation of the bimodal time course for small sample sizes $N \leq 3$. A closer inspection of the fitted time courses (not shown) reveals that for these small networks the estimation of both peaks is imbalanced. One of the peaks is recovered with high accuracy at the cost of larger error for estimation of the second peak. This also leads to an inaccurate shape of the time course of the hidden component in the middle of the time scale. This effect disappears as more data is available in larger networks and both peaks are estimated at balanced strength.

Finally, we note that the estimation at the two ends of the time interval are produced with a higher error. This is due to the fact that the splines are fitted with a higher uncertainty when there is not enough neighbouring observations to support their approximation.

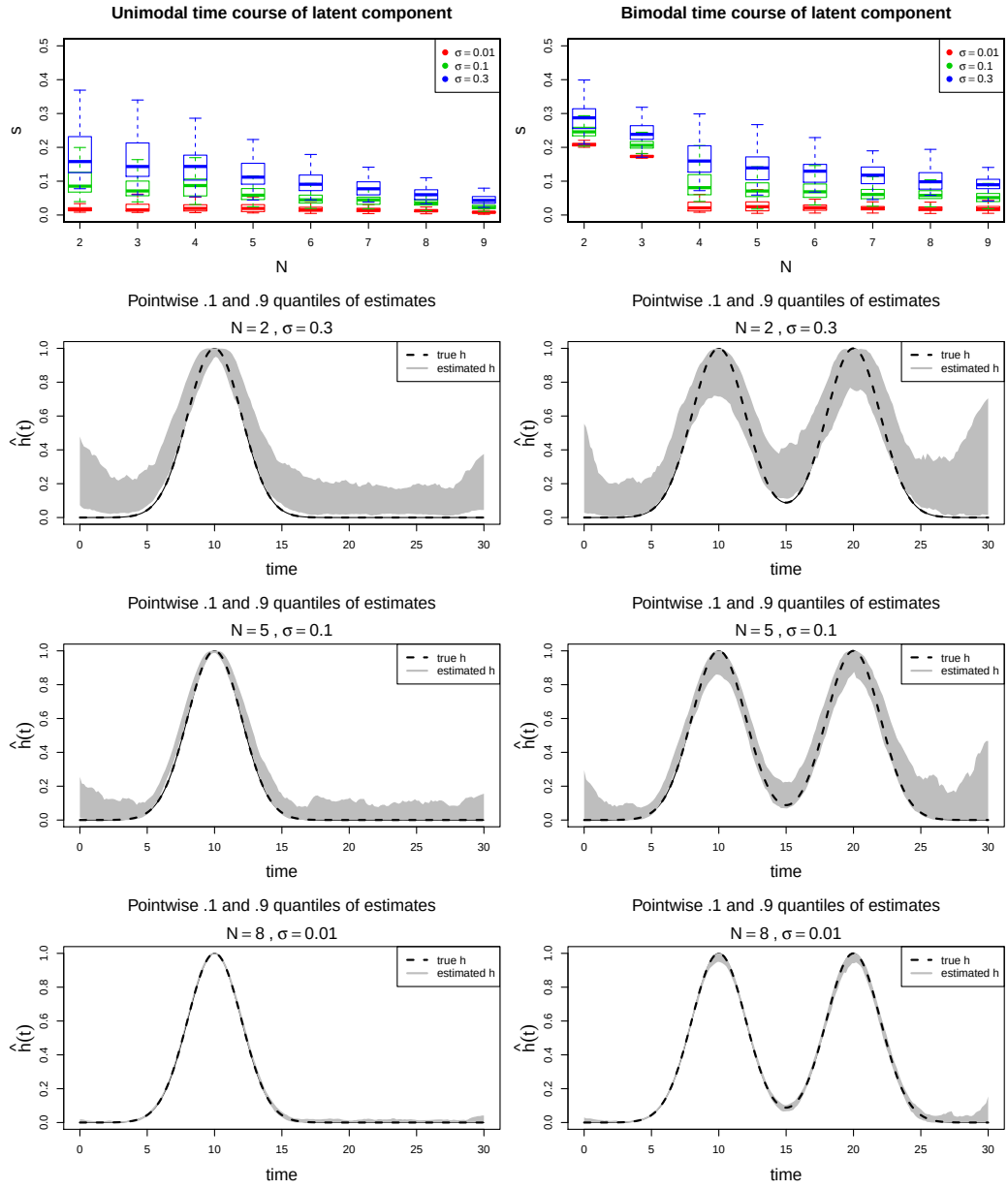

Figure 4.5: Estimating hidden time course from simulated data; left panels: unimodal time course of hidden component, right panels: bimodal time course of hidden component. Top panels show boxplots of error measures s (see Equation (4.20)) for networks of different combinations of size and noise. Each boxplot is based on 100 different s values. Lower panels show the 0.1 and 0.9 pointwise quantiles of estimated time courses as shaded area alongside with true time course used to generate the data as dashed line.

Overall, the second set of simulations showed the ability of our method to estimate irregularly shaped time courses with high accuracy. In the following, we investigate the behaviour of our method when dealing with missing data.

4.4.3 Missing data

In the following, we investigate the behaviour of our method when parts of the data are missing. Specifically, we consider some of the time snaphots of a given time series to be missing. Missing data situations are common when dealing with time-resolved real-world biological data. Reasons for missing data may be unavailability of a test subject at a given time point or a broken experimental vessel. We consider the case when the data are missing completely at random also referred to as the MCAR case in literature (Rubin [1976]; Wothke [2000]). In short, this is a situation where the probability of a data point to be missing is independent from all variables (in the studied context $\boldsymbol{x}(t)$ and $h(t)$) and parameters (in the studied context initial conditions, $\boldsymbol{a}$, $\boldsymbol{k}$ and all spline parameters) in the model.

We simulate data from a small network motif of size $N = 3$. Equation (4.19) in this case becomes

$$\dot{x}_i(t) = \sum_{u=1}^{3} (k_{iu}x_u(t) - k_{ui}x_i(t)) + a_i h(t). \tag{4.22}$$

For simulation, we again randomly choose the parameters $\boldsymbol{a}$ and $\boldsymbol{k}$. However, for estimation of $h(t)$, we consider $\boldsymbol{k}$ to be known. This allows us to focus the method performance evaluation solely on the estimation accuracy with respect to missing data. Once these parameters are fixed, we generate artificial data at one of $n \in \{10, 30, 50\}$ time points. Normally distributed noise is then added to these time points and we consider different options for the standard deviation, $\sigma \in \{0.05, 0.1, 0.3, 0.5\}$. After simulation of these observations we randomly delete a fraction m of observations from each component $x_i(t), i \in \{1, 2, 3\}$ with $m \in \{0, 10, 30, 50, 70\}$. For all combinations of n, σ and m, we simulate 100 datasets. One example for the data generation at $n = 30$, $\sigma = 0.1$ and $m = 50$ is shown in Figure 4.6.

The estimation accuracy of $\hat{h}(t)$ for each of these datasets is measured by the score s as defined in (4.20). Results of this simulation study are summarized in Table 4.1.

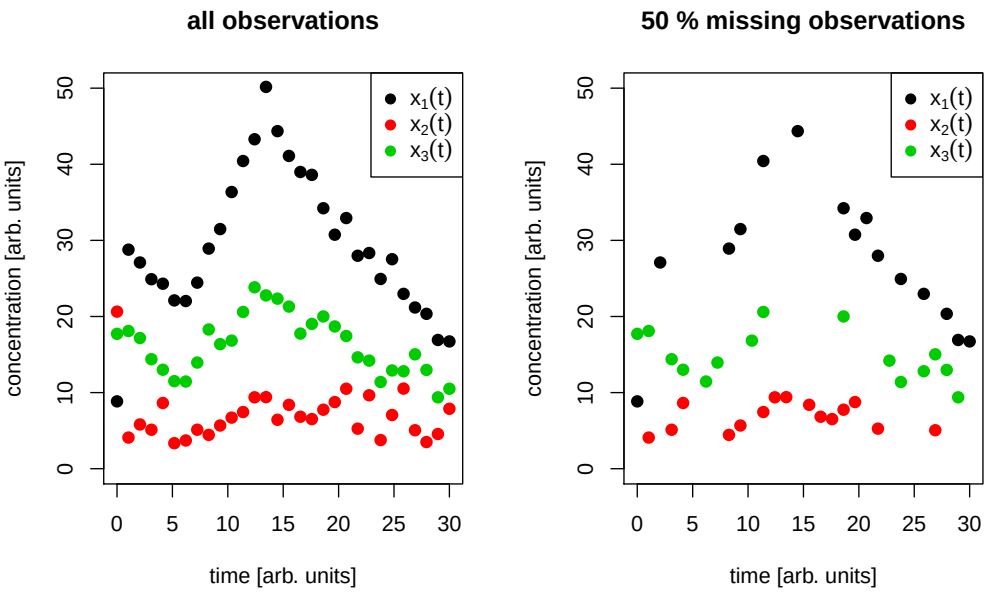

Figure 4.6: Creating a MCAR data scenario. Left: all simulated observations; right: 50% of observation per variable $x_i(t), i \in \{1,2,3\}$ are discarded.

The values in this table are the average score based on the 100 above-described datasets. We observe that a higher number of sampled time points tend to lower the score s. For example, if 10% of the observations are missing, the average score is approximately three times lower for $n = 50$ and $\sigma = 0.1$ than for $n = 10$ and $\sigma = 0.1$. On the other hand, higher noise level is resembled in larger score values. Both effects are in concordance with the previously discussed simulations in Section 4.4.1 and Section 4.4.2. Interestingly, for a low fraction of missing values ($m < 50$) there is only a low effect on the score. For example, scores almost do not change if 10% of data is missing and they change by less than 10% if 30% of the data is missing. However, for half or more of the data missing, the average scores are considerably increased. The reason for this is that the chance for data from the peak shown in Figure 4.6 completely missing is increased if the overall percentage of missing data is increased. If this peak is not present in the data, it is not possible to estimate the shape of the hidden component and this results in a more or less random guess. Nevertheless, we are confident that in situations where up to 30% of the data is missing, our method will not suffer a dramatic estimation accuracy.

Table 4.1: Average score s for different combinations of number of time points n, noise level σ and fraction of missing values m. Values are based on 100 randomly simulated datasets for each combination of parameters.

all data observed				
	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
$n = 10$	0.24	0.31	0.46	0.52
$n = 30$	0.14	0.23	0.36	0.42
$n = 50$	0.04	0.12	0.25	0.31
10% missing data				
	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
$n = 10$	0.25	0.34	0.47	0.53
$n = 30$	0.15	0.23	0.39	0.43
$n = 50$	0.05	0.11	0.26	0.33
30% missing data				
	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
$n = 10$	0.28	0.37	0.52	0.57
$n = 30$	0.17	0.27	0.40	0.45
$n = 50$	0.08	0.15	0.29	0.35
50% missing data				
	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
$n = 10$	0.35	0.45	0.58	0.64
$n = 30$	0.25	0.36	0.48	0.53
$n = 50$	0.15	0.23	0.37	0.43
70% missing data				
	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
$n = 10$	0.48	0.56	0.70	0.76
$n = 30$	0.37	0.45	0.60	0.68
$n = 50$	0.25	0.34	0.51	0.54

4.4.4 Recovering misspecified networks with a latent variable

Our method can be used for a guided repair of a wrongly specified network. We demonstrate this using artificial data in a further example. In this example, the network from which we simulate time-dependent observations consists of four players that are connected with each other in a forward cascade ending with a feedback loop, as shown in Figure 4.7A. However, we assume that the initial hypothesis suggests a network structure with a missing feedback loop. Furthermore, we do not assume known reaction rates $\boldsymbol{k}$; thus, we incorporate the fitting of $\boldsymbol{k}$ into the application of our method.

Figure 4.7 shows that we can simultaneously reconstruct the misspecified network structure and estimate $\boldsymbol{k}$ very well. The model without a feedback loop is best estimated with parameters $\hat{\boldsymbol{k}} = (0.05, 0.06, 0.01)^T$ and has a BIC value of 1376.78 (Figure 4.7B-1). The identified latent component has a positive interaction with the first species $x_1(t)$ and a negative interaction with the last species $x_4(t)$. This suggests that a feedback loop might be missing in the network specification. The corresponding BIC value is 857.48. The estimate $\hat{\boldsymbol{k}} = (0.15, 0.29, 0.20)^T$ is close to the true $\boldsymbol{k}$ (Figure 4.7B-2,3). The constellation of interactions between the hidden component and $\boldsymbol{x}$ suggests a feedback loop. Inclusion of this loop further improves the BIC value to 854.15 and slightly alters $\hat{\boldsymbol{k}} = (0.14, 0.30, 0.20)^T$ (Figure 4.7B-4). Subsequent application of our method does not identify a latent component which significantly improves the model fit (Figure 4.7B-5,6).

This example demonstrates the ability of our method to recover misspecified network structures. We repeated the presented example with random data 100 times and concluded the same missing feedback in 97% of the repetitions (results not shown). However, in general networks, misspecifications may occur in very a complex manner; thus, overall it will be difficult to always apply our method under all conditions. Nevertheless, even if the network structure cannot be recovered completely, a hidden component may indicate which network components are candidates for refining the network structure and whether inhibition or activation of certain network components are more likely to improve a given model.

4.5 Application: JAK2–STAT5 signalling pathway

The simulation studies in Section 4.4 have shown that our estimation procedure can reliably detect and quantify a hidden influence on a given network. We now focus on models and real-world data from the literature. A prominent and well-studied example is the erythropoietin (Epo) signalling pathway which tranduces Epo stimulation via JAK2-STAT5 (Darnell [1997]). Epo signalling plays an important role in proliferation, differentiation and survival of erythroid progenitor cells (Klingmüller *et al.* [1996]). After binding of the Epo hormone to its receptor, STAT5 can also bind. Subsequently, dimerization of STAT5 results in a translocation to the nucleus where the STAT5 dimer acts as a transcription factor.

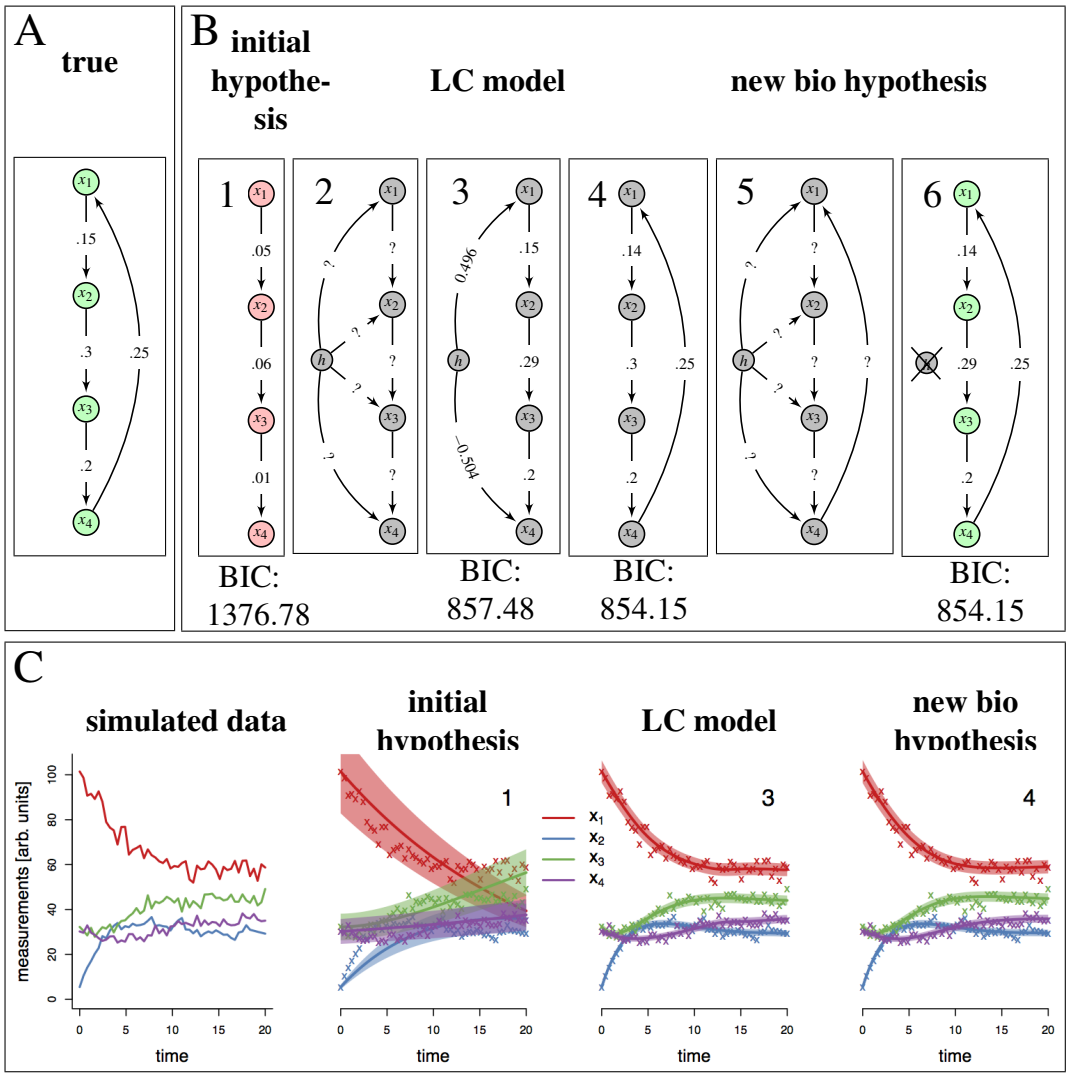

Figure 4.7: Recovering a misspecified network with a hidden component and simultaneously fitting the interaction rates $\boldsymbol{k}$. **A**: the true network from which the data is simulated. **B**: schematic representation of the workflow. 1) misspecified network. 2)/3) latent component (LC) model suggests a feedback loop. 4)/5)/6) the model with feedback loop cannot be further improved. **C**: corresponding data and model fits.

Several models exist which explain the molecular dynamics in various ways (Müller *et al.* [2004]; Swameye *et al.* [2003]; Timmer *et al.* [2004]; Toni & Stumpf [2010]). We analyze immunoblotting data which have already been analysed with a basic model by Swameye *et al.* [2003] using the following system of ODEs:

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 \mathrm{EpoR_A} \\ \dot{x}_2 &= -k_2 x_2^2 + k_1 x_1 \mathrm{EpoR_A} \\ \dot{x}_3 &= -k_3 x_3 + 0.5 k_2 x_2^2 \\ \dot{x}_4 &= +k_3 x_3 . \end{aligned} \tag{4.23}$$

Here, the different states of STAT5 are cytoplasmic unphosphorylated STAT5 (denoted by x_1), cytoplasmic phosphorylated monomeric STAT5 (x_2), cytoplasmic phosphorylated dimeric STAT5 (x_3) and STAT5 in the nucleus (x_4). $\mathrm{EpoR_A}$ describes the Epo-induced tyrosine phosphorylation which can be measured up to a scaling factor. The initial values are $x_1(0) > 0$ (to be estimated) and $x_2(0) = x_3(0) = x_4(0) = 0$.

In the above mentioned literature, the model given in (4.23) is further refined by, e. g. introducing an additional transition from nuclear STAT5 to the cytoplasmic unphosphorylated state, thus completing the loop from x_1 to x_4, or introducing time delays. These model refinements typically lead to an improved representation of the measured data, confirmed by, e. g. likelihood ratio tests, information criteria (AIC/BIC) or Bayes factors. To start from the best-known model, we extend the refined model by incorporating a hidden influence. As a first step, we consider

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 \mathrm{EpoR_A} + a_1 h, \\ \dot{x}_2 &= -k_2 x_2^2 + k_1 x_1 \mathrm{EpoR_A} + a_2 h, \\ \dot{x}_3 &= -k_3 x_3 + 0.5 k_2 x_2^2 + a_3 h. \end{aligned} \tag{4.24}$$

Here, we do not consider the fourth row of (22) because we have no information about x_4 as we use the measurements of experiment number 1 provided as supporting material in Swameye *et al.* [2003]. These measurements describe the total amount of cytoplasmic tyrosine phosphorylated STAT5, that is, $y_1 = k_5(x_2 + 2x_3)$, the total amount of cytoplasmic STAT5, $y_2 = k_6(x_1 + x_2 + 2x_3)$, and the Epo-induced tyrosine phosphorylation, $y_3 = k_7 \mathrm{EpoR_A}$. All three mea-

sured time-varying variables were experimentally quantified up to scaling factors denoted by k_5, k_6 and k_7. Evidently, only transformations of the ODE components $x_1,\ldots,x_3$ are observed. Furthermore, a system comprising only y_1, y_2 and y_3 cannot be described in closed form. For that reason, we also include the auxiliary variable x_3. The differential equations for the observed and latent components are as follows:

$$\begin{aligned}
\dot{y}_1 &= \frac{k_1 k_5 y_2 y_3}{k_6 k_7} - \frac{k_1 y_1 y_3}{k_7} - 2k_3 k_5 x_3 + k_5(a_2 + 2a_3)h, \\
\dot{y}_2 &= -2k_3 k_6 x_3 + k_6(a_1 + a_2 + 2a_3)h, \\
\dot{x}_3 &= -k_3 x_3 + \frac{k_2 y_1^2}{2k_5^2} - \frac{2k_2 y_1 x_3}{k_5} + 2k_2 x_3^2 + a_3 h.
\end{aligned} \tag{4.25}$$

We further refine the model by completing the loop from x_4 to x_1 and including a time delay, as has been done previously (Nikolov *et al.* [2007]). The authors suggested the use of a linear chain trick (Fall [2002]) and introduced a delayed loop. Thus, two (or possibly more) additional variables in the system of differential equations are introduced:

$$\begin{aligned}
\dot{x}_1 &= -k_1 x_1 \mathrm{EpoR_A} + 2k_4 z_2 + a_1 h, \\
\dot{x}_2 &= -k_2 x_2^2 + k_1 x_1 \mathrm{EpoR_A} + a_2 h, \\
\dot{x}_3 &= -k_3 x_3 + 0.5 k_2 x_2^2 + a_3 h, \\
\dot{x}_4 &= +k_3 x_3 - k_4 z_2, \\
\dot{z}_1 &= \frac{1}{\tau}(x_3 - z_1), \\
\dot{z}_2 &= \frac{2}{\tau}(z_1 - z_2).
\end{aligned} \tag{4.26}$$

Analogously, we can transform these equations to counterparts depending on y_1,

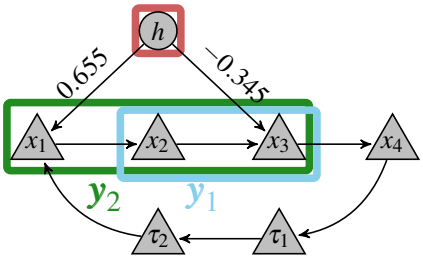

Figure 4.8: Schematic representation of the JAK2-STAT5 signaling pathway. The four different states of STAT5 are regulated by a latent component h with different weights, as estimated by our method. The observed variables y_1 and y_2 are linear combinations of the single states x_1 to x_3. τ_1 and τ_2 represent artificial delay variables.

y_2 and y_3:

$$\begin{aligned}
\dot{y}_1 &= \frac{k_1 k_5 y_2 y_3}{k_6 k_7} - \frac{k_1 y_1 y_3}{k_7} - 2k_3 k_5 x_3 + k_5(a_2 + 2a_3)h, \\
\dot{y}_2 &= -2k_3 k_6 x_3 + 2k_4 k_6 z_2 + k_6(a_1 + a_2 + 2a_3)h, \\
\dot{x}_3 &= -k_3 x_3 + \frac{k_2 y_1^2}{2k_5^2} - \frac{2k_2 y_1 x_3}{k_5} + 2k_2 x_3^2 + a_3 h, \\
\dot{z}_1 &= \frac{1}{\tau}(x_3 - z_1), \\
\dot{z}_2 &= \frac{2}{\tau}(z_1 - z_2).
\end{aligned} \tag{4.27}$$

This representation captures the dynamics of the observed variables. The right-hand side of (4.27) depend on the observed components y_1 to y_3, the hidden component h, the unobserved component x_3 and the two artificially introduced delay variables z_1 and z_2. For this reason, we must estimate x_3, z_1 and z_2 prior to h. This is achieved by numerically computing the solution of the model given in (4.26) without considering the hidden component (i. e. $\boldsymbol{a} = \mathbf{0}$) and using the approximations for x_3, z_1 and z_2 arising from this model. Once these quantities are determined, we estimate the three transformed weighting coefficients $\tilde{\boldsymbol{a}} = (k_5(a_2 + 2a_3), k_6(a_1 + a_2 + a_3), a_3)$ and use the estimates for x_3, z_1 and z_2 as input in the new iteration. This procedure is repeated until convergence. Once $\tilde{\boldsymbol{a}}$ is successfully obtained, we simply calculate $\boldsymbol{a}$ from $\tilde{\boldsymbol{a}}$ up to the scaling factors k_5 and k_6.

Figure 4.8 shows a schematic description of the estimated model given in (4.27). According to our estimation performed by best subset selection, the hidden com-

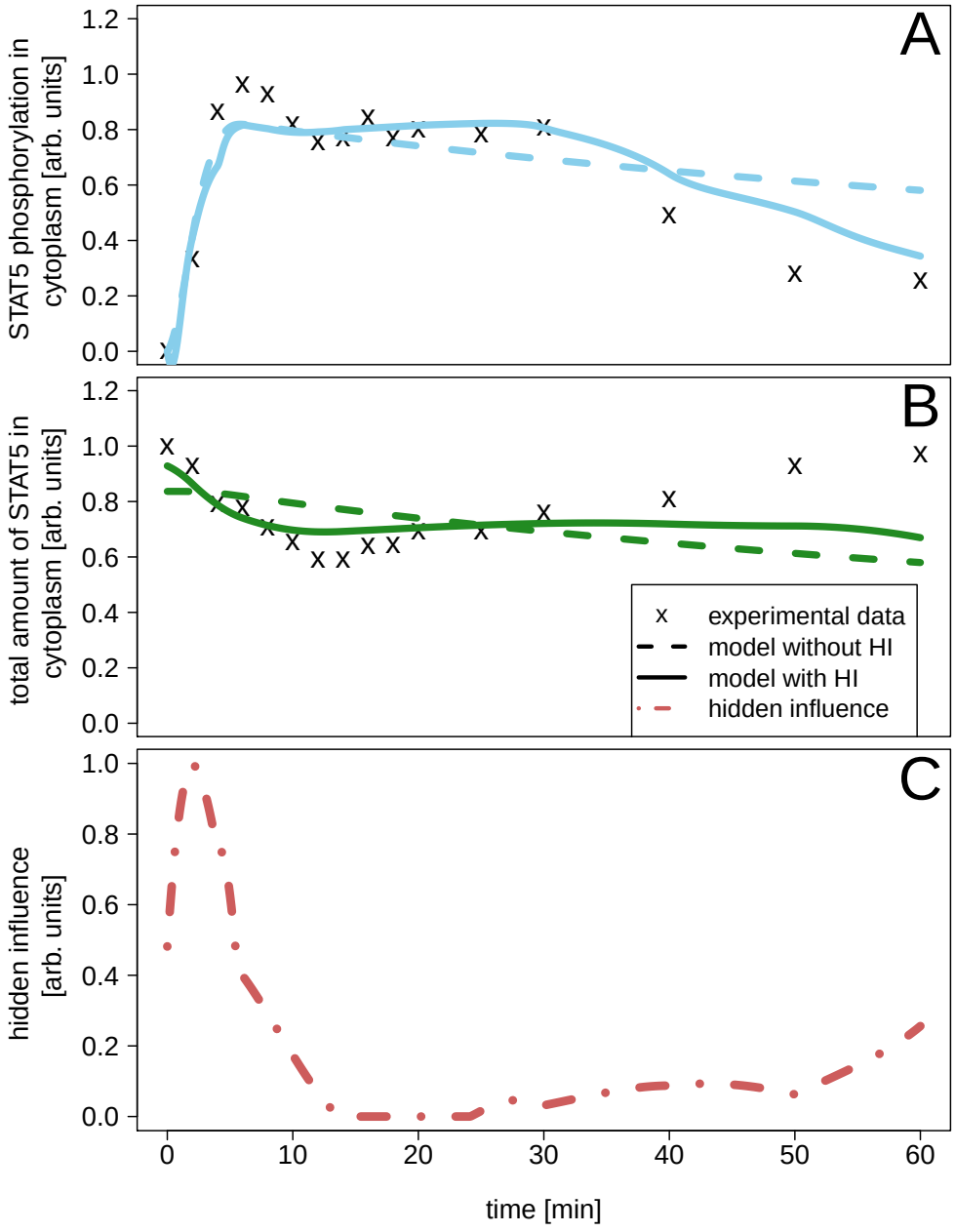

Figure 4.9: **A**: experimental data and model fitting for STAT5 phosphorylation in cytoplasm (y_1). **B**: experimental data and model fitting for the total amount of STAT5 in cytoplasm ($y2$). The model that includes no hidden component is indicated by the dashed lines, whereas that which includes a hidden component is indicated by solid lines. The model with a hidden influence produces a time course that better fits the experimental data. **C**: estimated time course of the hidden component, which exhibits a strong peak at the beginning of the experiment, quickly drops to 0 and begins to increase again after 30 min.

ponent interacts only with the first and third state of STAT5. Interestingly, the interaction direction (activating x_1 and inhibiting x_3) hints at a translocation of STAT5 from its nuclear state to the cytoplasm as also hypothesized by e. g. Swameye *et al.* [2003].

Figures 4.9A and 4.9B show the experimental data and the estimated time courses of y_1 and y_2. The model with a hidden component h outperforms the model without h because it best represents the experimental data. Most importantly, the time course produced with a hidden component is considerably more flexible but does not overfit the data. The time course of the estimated hidden component (third panel of Figure 4.9C) exhibits large values at the beginning of the experiment, decreases and then begins increasing after 30 minutes. Our interpretation of this behaviour is that an external quantity should be present at the beginning of the experiment (or shortly after); thus, the entire signalling pathway is kick-started. This external stimulus depletes completely and its influence slowly begins increasing after 30 min, bringing the entire system into equilibrium with the inhibition of the dimerized STAT5, and simultaneously the activation of the monomeric STAT5.

4.6 Discussion

The main objective of this chapter is to provide a new method for model extension by introducing a hidden component to known networks. With the proposed method, we can not only derive the relative time course of the hidden component but also predict the influence of the hidden component on all other network components.

We first fit splines to the observed components or to observation functions that are affine linear transformations of the former. On the basis of the observation error distribution, we apply maximum likelihood estimation and model selection. By doing so, we can estimate a combination of the time course of a hidden component and the weights that lead to the best model in terms of data faithfulness without overfitting.

The method is applied to artificial data to test robustness and applicability. The results suggest a robust and good performance for the identification of the time

course of the hidden component even in situations with high level of noise, irregularly shaped time courses of the hidden component or missing data.

One application of our method is the detection of misspecified networks. As a demonstration, we choose a network which includes a feedback loop that is missing in the model specification. The loop is successfully recovered, thus providing a promising application variant of our technique. Our method, however, is not a tool for general network inference in its current form. An automation of the process by combining theory from network topology estimation with the proposed latent variable model presents a possible extension in future work.

We applied the method to the well-studied JAK2-STAT5 signalling pathway. Extension of the model with a latent component was performed on a system of ODEs with introduced delay. Our method improves the model quality in terms of BIC and produces results which are in conformity with other methods suggested in the literature.

For the method presented here, we intentionally chose to separate the two major estimations into two steps, and both steps can be associated with two major modelling perspectives (Emmert-Streib *et al.* [2014]). While fitting the spline parameters can be associated with a statistical perspective exploiting the network structure for inference of the latent time-course and its interaction weights is closely connected to the mathematical modelling perspective. Model selection and thus network prediction, brings the method back to the statistical perspective. Formulating the problem as a joint optimization of all parameters involved (reaction rates, spline parameters and noise parameters) is possible. This, however, leads to a considerably more complex and computationally intensive method.

As we demonstrate in Appendix A.2, the performance of our method depends on the quality of the spline approximation. This quality will typically suffer if the modelled data are sparse, contain extreme outliers, are corrupted by a high amount of noise or the chosen spline representation cannot resemble fluctuations of the observed time-series appropriately.

The results of the proposed method can be employed as a promising aid for guiding future experiments, thus helping to complete the systems biology loop (de Ridder *et al.* [2013]; Endler *et al.* [2009]) between experimental data and model analysis.

5

Inferring catalysis in biological systems

In this chapter we further investigate the communication patterns between several species, such as genes, enzymes or proteins. As already demonstrated in the previous chapter, time-resolved communication between such species can be structured by reaction networks. Mathematical modelling of data arising from such networks often reveals important details, thus helping better to understand the studied system. In many cases, however, corresponding models still deviate from the observed data. This may be due to unknown but present catalytic reactions. From a modelling perspective, the question of whether a certain reaction is catalysed and which active catalyst is observed, leads to a large increase of model candidates. For large networks the calibration of all possible models becomes computationally infeasible very fast.

We present a novel method for inference of catalysis from biological systems. It can be summarized in three major steps. First, we extend a given network by a number of additional components which is equal to the number of total interactions within a network. Next, we infer the time courses of the hidden components with the help of spline approximation and a least squares approach. Finally, the inferred time courses are compared to the time courses of the original network components. This comparison results in a similarity score which describe the likelihood of a certain network component catalysing a certain network reaction. The scores are standardized in the unit interval and can be used to identify the most

probable catalysis candidates for each reaction and thus consider only a small number of models for a given system. This is especially useful when the studied biological network is large and considering all model possibilities results in high computational demand. Furthermore, the method also provides parameter estimates for the reaction rates of network interactions.

The method is applied on artificial data with the aim to assess its general applicability and it is also compared to other possible model selection techniques. Results confirm that with our method, we are able to substantially reduce the number of candidate models for a given system without discarding the correct model from which the artificial data was generated. This holds true independent from the network size and also for non-informative data with few observations. Finally, we apply the method to real-world data arising from the CD95 apoptotic pathway and provides new insights into apoptosis regulation.

This chapter is based on and in part identical with the following publication:

- I. Kondofersky, F.J. Theis and C. Fuchs. Inferring catalysis in biological systems, *submitted*.

5.1 State of the art and research questions

A central objective in systems biology is to derive a mathematical model, which is used to explain multivariate readouts and thus serve as a tool for detailed investigation of a given biochemical process ([Aloy & Russell, 2006; Kitano, 2002a]). Although there are many ways in constructing such models, they generally share the well-known dilemma of models being always only an approximation of reality. This means that regardless of the quality of the model performance, there always remains uncertainty when explaining a biological phenomenon (Slezak *et al.* [2010]). This uncertainty may arise from different sources, some of which are: the collected data may be subject to various kinds of noise; parameters of complex models may be unidentifiable and thus lead to equal quality of several competing models; the model topology may be specified in a wrong way, e. g. providing a too extreme simplification of reality.

Describing the connection between several variables can be conveniently done using networks or pathways (Barabasi & Oltvai [2004]). This has successfully been

applied in the field of biology in past decades (Jeong *et al.* [2000]). For example, signalling pathways are known to be the core mechanism of numerous biochemical processes, such as cell differentiation, cell death or cell division. Additionally, the intracellular behaviour of small molecules can be described in a detailed manner ([Artavanis-Tsakonas *et al.*, 1999; Vogel & Sheetz, 2009]). Small differences in this behaviour may determine the cell fate and thus are of major importance for the overall understanding of the modelled system. Interactions between single components of such networks can occur in various complexities e. g. linear, higher-order or catalytic reactions. The identification of catalytic reactions can be especially challenging if the catalyst of an interaction is not known.

Considering the possibility that reactions are catalysed expands the model candidate space in an exponential way. To address this challenge, some established model selection techniques, such as greedy stepwise model selection or full best-subset model selection, can be applied. However, these model selection techniques often fail to find the most appropriate model for given data due to either not taking correlation of network components into account or overfitting to data. This means that reducing the model candidate space often comes at a high price of reduced method performance.

Recently, a novel scheme of catalysis identification has been proposed by Rickert *et al.* [2013]. Here, the authors suggest a model reduction technique which is a graphical approach, taking into account the network topology of the system. Although their approach is able to vastly reduce the model candidate space, this reduction is mostly achieved by eliminating catalysis from certain reactions due to biological prior knowledge rather than performing a statistical comparative study. Furthermore, their approach needs user input suggesting which reactions should be investigated for catalysis.

This manuscript proposes a novel approach for identification of catalysis in biological systems. We first extend the known network by including hidden components and estimate their time courses with a combination of smoothing splines and least squares approach. In the next step, we compare those time courses of the hidden components to the time courses of network components. Here, we measure similarity between two time courses based on correlation and L^2-distance and associate each comparison with a score. Subsequently, we choose a threshold and only consider components with high scores to be relevant catalyst candidates. The

reduced number of model candidates is finally calibrated and the best model is chosen via maximum likelihood.

This chapter is organised as follows. In Section 5.2 we define our modelling approach and explain how we estimate model parameters. This ultimately leads to building a score for every network component, which describes its affinity to catalyse a certain reaction. Section 5.3 applies the developed technique to different simulated scenarios and Section 5.4 to a real data example - the CD95 apoptosis pathway. Section 5.5 discusses strengths and limitations of the proposed method.

5.2 Methods

In this section, we present the developed method for inferring catalytic reactions biological systems. We first describe the types of systems we aim to study with this method in the context of catalysis. Then, we introduce the individual steps of the estimation procedure. In brief, we model an extended system with external or hidden catalysts and afterwards compare these external catalysts to observed network components by construction of a similarity score. This allows us to preselect only a small number of model candidates, which we then compare in more detail with a likelihood approach. Overall, this results in obtaining the most appropriate model for the data without wasting computational resources.

5.2.1 Mathematical formulation of catalysis

We consider N-dimensional ODEs with m reaction fluxes, as formulated in Equation (2.28) which we repeat here:

$$\frac{d\boldsymbol{x}(t)}{dt} = \dot{\boldsymbol{x}}(t) = \boldsymbol{S}\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k}) = \sum_{g=1}^{m} \boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k})$$

with $N \times m$ stoichiometry matrix $\boldsymbol{S}$, m-dimensional flux function $\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k})$ with arguments $\boldsymbol{x}(t) = (x_1(t),\ldots,x_N(t))^T \in \mathbb{R}^N_{\geq 0}$ as the non-negative network component concentration functions and $\boldsymbol{k} \in \mathbb{R}^p$ as the reaction rate constants as already defined in Section 2.1.4. We assume the individual flux functions $v_g(\boldsymbol{x}(t);\boldsymbol{k})$ to

be linear combinations of $\boldsymbol{x}(t)$. The components $x_i(t)$ may be observed or unobserved.

Modelling network dynamics as in (2.28) presents a general way of describing biological systems. However, this description is often not sufficient to explain the observed network dynamics. To improve the discrepancies between model fit and observed data, one can choose different strategies. Approaches range from construction of more complex interactions such as Michaelis Menten kinetics, time-varying reaction rates or complex formations introducing external latent variables (Chapter 4). Catalysis is an additional way of improving the model fit and at the same time maintaining a low level of model complexity. Furthermore, it represents the modelling of a realistic scenario since catalysis is an often-occurring pattern in many biological systems (Masel *et al.* [2001]). A catalytic reaction can be included into (2.28) by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{S}(\boldsymbol{v}(\boldsymbol{x}(t);\boldsymbol{k}) \circ \boldsymbol{h}(t)) = \sum_{g=1}^{m} \boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t);\boldsymbol{k}) h_g(t) = \boldsymbol{\psi}(\boldsymbol{S}, \boldsymbol{v}, \boldsymbol{x}(t), \boldsymbol{k}, \boldsymbol{h}(t)) \quad (5.1)$$

with $\circ$ denoting the Hadamard product (componentwise multiplication), $\boldsymbol{h}(t) = (h_1(t), \ldots, h_m(t))^T \in \mathbb{R}^m_{\geq 0}$ representing the concentration of the non-negative catalysts and $\boldsymbol{\psi}$ as a summarizing function for the right-hand side of the ODE. We will later estimate the unknown catalysts $\boldsymbol{h}(t)$. We further restrict our models to $\boldsymbol{h}(t)$ having a meaningful effect on $\boldsymbol{\psi}$ and thus require

$$\forall \varepsilon > 0, \forall t \geq t_0, \forall h_2(\boldsymbol{t}) \in U_\varepsilon(h_1(\boldsymbol{t})) : \boldsymbol{\psi}(h_1(\boldsymbol{t})) \neq \boldsymbol{\psi}(h_2(\boldsymbol{t})) \quad (5.2)$$

with $U_\varepsilon(h_1(\boldsymbol{t})) = \{h_2(\boldsymbol{t}) \in \mathbb{R}_{\geq 0} :|| \, h_1(\boldsymbol{t}) - h_2(\boldsymbol{t}) \, ||^2 < \varepsilon\}$ and $|| \cdot ||^2$ denoting the L^2 norm. Furthermore, without loss of generality, for the rest of the manuscript we assume $\boldsymbol{S}$ and $\boldsymbol{v}$ to be known in parametric form, e. g. from literature. This assumption can be relaxed and $\boldsymbol{S}$ and $\boldsymbol{v}$ can also be estimated with our method, which increases the number of unknown parameters. The assumption seems reasonable since we want to apply our method to well-studied systems where information about $\boldsymbol{S}$ and $\boldsymbol{v}$ is available.

5.2.2 Estimation of hidden catalysts

In the first part of the proposed method, we estimate $\boldsymbol{h}(t)$ and interaction parameters $\boldsymbol{k}$. Here, we approximate the observed time courses of $\boldsymbol{x}(t)$ by smoothing splines (compare Section 2.1.3) resulting in an estimtate $\hat{\boldsymbol{x}}(t)$. This also presents an immediate approximation of $\dot{\boldsymbol{x}}(t)$ as $\hat{\dot{\boldsymbol{x}}}(t) = \frac{\partial}{\partial t}\hat{\boldsymbol{x}}(t)$. Subsequently, we plug in these approximations into (5.1) and estimate the parameters $\boldsymbol{h}(t)$ and $\boldsymbol{k}$ by

$$(\hat{\boldsymbol{k}}, \hat{\boldsymbol{h}}(t)) = \underset{\boldsymbol{k}, \boldsymbol{h}(t)}{\operatorname{argmin}} \left[|| \, \hat{\dot{\boldsymbol{x}}}(t) - \psi(\boldsymbol{S}, \boldsymbol{v}, \hat{\boldsymbol{x}}(t), \boldsymbol{k}, \boldsymbol{h}(t)) \, ||^2 \right]. \tag{5.3}$$

In general, (5.3) has more unknown parameters than the ODE dimension and thus some estimated parameters in (5.3) may not be identifiable. One possibility to reduce the estimated parameter space is to set non-identifiable parameter entries in $\boldsymbol{k}$ to a constant, e. g. to 1. Such non-identifiable parameters can occur wherever the ODE has entries such as $\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t); \boldsymbol{k}) \boldsymbol{h}_g(t)$ where both $\boldsymbol{k}$ and $\boldsymbol{h}_g(t)$ cannot be estimated simultaneously due to $\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t); \boldsymbol{k}) \boldsymbol{h}_g(t) = (a\boldsymbol{h}_g(t)) \cdot \frac{\boldsymbol{s}_{\cdot,g} v_g(\boldsymbol{x}(t); \boldsymbol{k})}{a}$, $\forall a \in R_{\neq 0}$ and without additional prior information or constraints. Approximations for the non-identifiable entries in $\boldsymbol{k}$ will be found in the second step of the proposed method. After elimination of such non-identifiable parameters, (5.3) is numerically optimized e. g. by a gradient descent method. The result of this first step are the approximations of the components of $\boldsymbol{h}(t)$, which can be grouped in a set:

$$\mathcal{H} = \{\hat{h}_g(t_j)\}_{g=1,\ldots,m; j=0,\ldots,n}. \tag{5.4}$$

Once these approximations are found, we perform similarity analysis to relate them to the network components which we describe in the following.

5.2.3 Relating hidden catalysts to network components

In the second step, we compare the entries of $\mathcal{H}$ to the set

$$\mathcal{X} = \{X_{ij} \mid X_{ij} = \hat{x}_i(t_j) \text{ if } i \leq N, X_{N+1}(t_j) = 1\}_{i=1,\ldots,N+1; j=0,\ldots,n}, \tag{5.5}$$

which contains the spline-approximated time courses of the network components $\hat{\boldsymbol{x}}$ with an additional component $x_{N+1}(t)$, which is equated to 1 for all t and is thus comparable to the intercept term in a regression context. This comparison between

entries in $\mathcal{H}$ and $\mathcal{X}$ is done in terms of two different measures of similarity. On the one hand, we measure similar time-course shapes of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{h}}$ by calculating the Pearson correlation coefficient between entries in $\mathcal{H}$ and $\mathcal{X}$, resulting in the set of correlations $\mathcal{C}$:

$$\mathcal{C} = \{C_{ig} \mid C_{ig} = \mathrm{cor}(\mathcal{X}_{i\cdot}, \mathcal{H}_{g\cdot})\}_{i=1,\ldots,N;g=1,\ldots,m}. \tag{5.6}$$

On the other hand, the proximity between $\mathcal{X}$ and $\mathcal{H}$ is measured by the L^2 distance and these values are collected in a set $\mathcal{L}$:

$$\mathcal{L} = \{\min_{\kappa_{ig}\in\mathrm{R}}(|| \mathcal{X}_{i\cdot} - \kappa_{ig}\mathcal{H}_{g\cdot}) ||^2)\}_{i=1,\ldots,N+1;g=1,\ldots,m} \tag{5.7}$$

with scaling parameters κ_{ig}, which are used to find the best scaling of $\mathcal{H}_{g\cdot}$ so that the L^2-distance to $\mathcal{X}_{i\cdot}$ is minimized. Recall that while optimizing (5.3), we set the non-identifiable parameters in $\hat{\boldsymbol{k}}$ equal to 1. With the optimization in (5.7), these parameters can now be estimated as the minimizers in (5.7).

The two sets, $\mathcal{C}$ and $\mathcal{L}$, measure two different aspects of similarity (shape and proximity), which are suitable for comparing two time series. Furthermore, *smaller* values in $\mathcal{L}$ and *larger* values in $\mathcal{C}$ correspond to higher similarities. Therefore, they are combined and weighted to form a set of scores $\mathcal{S}$, which can be used to easily identify catalysis candidates. To construct such a set, single entries of $\mathcal{C}$ and $\mathcal{L}$ are combined and scaled in the unit interval. Formally, we build

$$\mathcal{S} = \frac{1}{2}\left\{\frac{\max(\mathcal{L}_{i\cdot}) - \mathcal{L}_{ig}}{\max(\mathcal{L}_{i\cdot}) - \min(\mathcal{L}_{i\cdot})} + \frac{\mathcal{C}_{ig} - \min(\mathcal{C}_{i\cdot})}{\max(\mathcal{C}_{i\cdot}) - \min(\mathcal{C}_{i\cdot})}\right\}_{i=1,\ldots,N+1;g=1,\ldots,m} \tag{5.8}$$

with $\max(\mathcal{L}_{i\cdot}) := \max_{g'}\{\mathcal{L}_{ig'} \mid g' = 1,\ldots,m\}$ and $\min(\mathcal{L}_{i\cdot})$, $\max(\mathcal{C}_{i\cdot})$ and $\min(\mathcal{C}_{i\cdot})$ defined in the same way. The special cases of $\max(\mathcal{L}_{i\cdot}) = \min(\mathcal{L}_{i\cdot})$ and $\max(\mathcal{C}_{i\cdot}) = \min(\mathcal{C}_{i\cdot})$ can be excluded without loss of generality. If one of those cases occurs, it means that we cannot distinguish between all candidates either on basis of distance or correlation. In this case all candidates describe the data equally and no candidate reduction can be achieved. The set $\mathcal{S}$ is constructed from $\mathcal{C}$ and $\mathcal{L}$ with equal contribution, respectively. One could of course also consider the inclusion of a weighting parameter which favours e. g. the correlation measure more strongly. Overall, $\mathcal{S}$ has values in the unit interval with a value of 1 in $\mathcal{S}_{i,g}$ meaning that $\mathcal{X}_{i\cdot}$ is best correlated and has the lowest L^2 distance (after scaling) to $\mathcal{H}_{j\cdot}$, making $\mathcal{X}_{i\cdot}$

the most obvious catalyst candidate for the g-th reaction. It is possible that multiple entries of $\mathcal{X}_{i\cdot}$ have a high score close to 1 which may then all be considered as catalyst candidates. We define a threshold τ, which is used to filter components with high associations for catalysts of a given reaction by the rule:

$$\mathcal{S}_{i,g} > 1 - \tau \Rightarrow \mathcal{X}_{i\cdot} \text{ candidate for } g\text{-th reaction.} \tag{5.9}$$

The index τ can be used in various ways. If τ equals 1, all components are classified as possible catalyst candidates (no model reduction), whereas if τ equals 0, at most one component per reaction is chosen as a possible catalyst (and only if it outperforms all other components in distance *and* correlation measure). Generally, τ can be used to control the trade-off between a large number of acceptable models and a high probability of finding the most appropriate model with the proposed algorithm. In practice and for the examples presented in this manuscript, we found that setting τ to 0.1 presents a reasonable choice.

5.2.4 Choice of most appropriate model from reduced model candidates with maximum likelihood

After performing the described steps above, a reduced set of models M_τ is obtained as a subset of all possible models, M. Additionally, we obtain the set $\bar{M}_\tau$, which describes the models which are not considered to be appropriate for the characterization of the studied system. It holds that $M_\tau \subseteq M$ and for large systems and for small τ we usually obtain $\mid M_\tau \mid \ll \mid M \mid$. Without loss of generality, we can assume that $\mid M_\tau \mid > 1$ and we still need to find the most appropriate model from the set M_τ. In this context, we apply a maximum likelihood optimization scheme to determine the model of choice. Therefore, we first specify an error distribution of the observed data as (compare (4.10) in Section 4.3.1):

$$x_i^{\text{obs}}(t_j) = x_i(t_j) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

In applications, $x_i^{\text{obs}}(t_j)$ often has a positive domain in which case this equation might be ill-defined. One possible solution for this might be log-normally distributed multiplicative noise as discussed in Chapter 2. Such error model is

straightforward here, however, for reasons of notation simplicity, we only consider normally distributed errors in the manuscript and in the example section. The distribution of ε_{ij} immediately propagates to the measurements:

$$x_i^{\text{obs}}(t_j)\big|x_i(t_j) \overset{\text{iid}}{\sim} \mathbb{N}\left(x_i(t_j), \sigma^2\right).$$

While the true time course $x_i(t)$ is unknown, it has already been approximated by smoothing splines and we can plug in this approximation :

$$x_i^{\text{obs}}(t_j)\big|\hat{x}_i(t_j) \overset{\text{iid}}{\sim} \mathbb{N}\left(\hat{x}_i(t_j), \sigma^2\right).$$

With this last approximation, we are now able to formulate a likelihood function, which measures the overall agreement between model and data depending on the model parameters

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{g=0}^{n} f_{\mathbb{N}}(x_i^{\text{obs}}(t_j) \mid \boldsymbol{\theta})$$

with $\boldsymbol{\theta}$ representing the conglomerate of parameters $(\boldsymbol{k}, \sigma^2)^T$. This likelihood function can be maximized with a gradient descent method and the parameters corresponding to this optimum are then called $\hat{\boldsymbol{\theta}}$. The dimension of $\hat{\boldsymbol{\theta}}$ does not change regardless of the number of reactions catalysed and the different combinations of catalytic reactions. Therefore, comparing models only by comparing likelihoods instead of using e. g. information criterion is possible in this setting.

In the next section, we will test the developed method on several artificial datasets and also apply it on real-world data from a biochemical pathway.

5.3 Simulation studies

In this section, we apply our method on artificially generated data. We perform two excessive simulations in which we test the applicability and effectiveness of our method. First, we simulate random networks of different size and estimate the reaction catalysts in those networks. Second, we fix the network size at $N = 5$ and test our method by comparing it to two other common approaches in model selection – the computationally demanding best subset selection and the greedy forward selection. All computations were performed using the open source software

R (R Development Core Team [2011]), version 3.2.1 and associated packages fda (Ramsay *et al.* [2009]) for the smoothing spline estimation and deSolve (Soetaert *et al.* [2010]) for estimating ODEs.

5.3.1 Random networks and random catalysts

We use several simulation runs to test the general applicability of the proposed method. To that end, we consider networks consisting of 2 to 10 nodes and sampled from

$$\dot{x}_i(t) = \sum_{g=1}^{N} \left(k_{ig} x_g(t) h_{ig}(t) - k_{gi} x_i(t) h_{gi}(t)\right), \tag{5.10}$$

where the reaction rates k_{ig} are chosen randomly from $\mathbb{U}(\frac{-1}{N}, \frac{1}{N})$ and the catalysts $h_{ji}(t)$ are chosen randomly to equal one of $(\mathbf{1}, \boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_N(t))$ with equal probability. Furthermore, the initial values $\boldsymbol{x}(0)$ are chosen randomly from $\mathbb{U}(1, 100)$. To achieve more realistic sparse networks, we randomly delete approximately a fraction of $\frac{2}{N}$ of the possible reactions by setting the corresponding reaction rates k_{ig} to 0. After forward simulation of the randomly chosen network, we add normally distributed measurement noise $\varepsilon \sim \mathbb{N}(0, \sigma^2)$ to the simulated time snapshots and arrive at the observed measurement points used for further analysis. The number of observed time points per component $\boldsymbol{x}_i(t)$ and the noise parameter σ^2 are also chosen at random for each simulation run from $\mathbb{U}(10, 30)$ and $\mathbb{U}(1, 15)$, respectively. In the described setting, we run 100 simulation runs per fixed network size and estimate the catalyst of each reaction. Results are shown in Figure 5.1.

Figure 5.1A shows violin plots of the fraction of models to be estimated after applying the latent catalyst method depending on the network size. Additionally, the average time needed to compute either all possible combinations for a given network size or the reduced set of models is shown with solid lines. Here, we observe that computing all possible combinations for networks of size 2 or 3 is faster than computing only a reduced number of models. This can be explained by the computational time needed to fit the splines and the relatively low number of possible candidate models for such small network sizes. With increasing network size, this relationship switches very fast and already at network size 5 the computational time needed to identify the correct catalysts with our method is less than 0.1% of

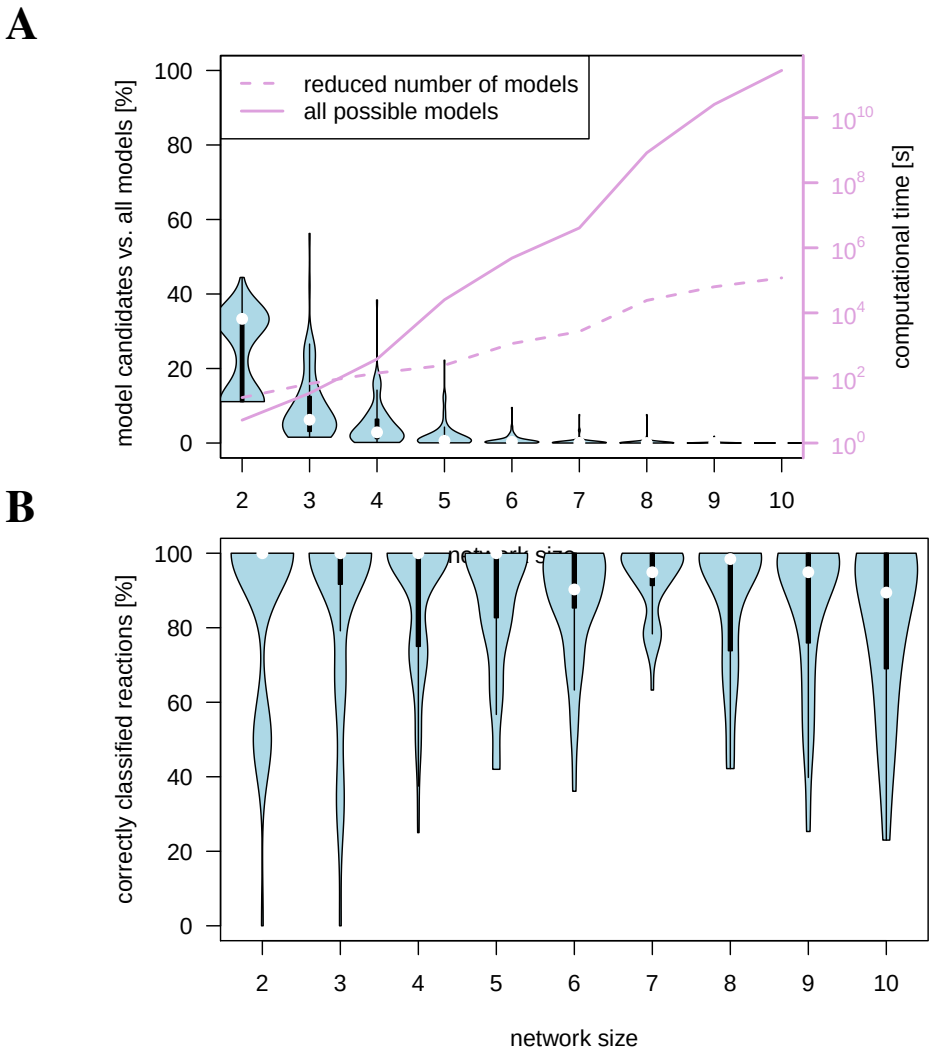

Figure 5.1: Results of simulation 1. **A**: Violin plots show fraction of model candidates chosen by application of the latent catalysts method compared to all possible models. The reduction of model candidates becomes more pronounced for larger networks. Additionally, lines show the average computational time in (log-scale) needed to estimate either all possible models (solid line) or the reduced set of model candidates (dashed line). **B**: Violin plots of the fraction of correctly classified catalysts. The reduced model candidates include the correct model that was used to generate the data in almost all simulation runs. This is consistent for all studied network sizes.

the time needed to compute all possible models. In the violin plots, a value of 100% means that no reduction of model candidates was achieved with our method and all possible models have to be computed. We observe a dramatic decrease of model candidates for networks consisting of more than 4 nodes. This shows the efficiency of our method, which potentially allows a reduction of computational time from days to minutes depending on the studied system.

This efficiency would not be meaningful if the reduced number of models did not include the correct model, which was used to generate the data. However, as Figure 5.1B suggests, in most simulation runs the correct model is part of the reduced model candidates, this is consistent for all studied network sizes. Although the method may also miss the correct model in certain simulation scenarios with e. g. large noise or many similarly shaped component dynamics, we observe a median of above 90% correctly identified catalysts by applying our method. Additionally, we note that we used a threshold parameter $\tau = 0.1$ for all simulations. If we set this parameter to a higher value, we will capture more correct models in the model candidates, however at the cost of lower efficiency and higher computational demand.

5.3.2 Catalysis in common network motifs in systems biology

Figure 5.2 shows artificial networks with and without catalytic interactions. This network consists of 5 nodes $x_1 - x_5$ and a total of 7 regulatory interactions between those nodes. In Figure 5.2A, we first show a version of the network with no cat-

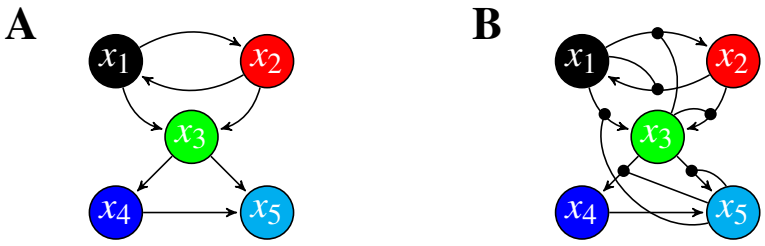

Figure 5.2: Network of interest for simulation 2. **A**: "core network" with no catalytic interactions. **B**: network with catalytic interactions from which data is sampled.

alytic reactions in order to demonstrate the general connection between the nodes. In Figure 5.2B, we include catalysis in the network structure and use this network

to simulate artificial data. We chose this network to further investigate our method performance because it captures several patterns which are commonly observed in systems biology. First, x_1 and x_2 are engaged in a mutual activation pattern and whichever of the two dominates this pattern also dominates the interaction with x_3. Second, a typical motif is presented by the interaction between x_3 and x_5 for which there is a direct interaction and at the same time an indirect or lagged interaction through x_4. Finally, we observe both layers to be connected by the key node x_3 and several catalytic connections which contribute to the overall interaction pattern of the studied network.

We sampled data from the network shown in Figure 5.2B by randomly choosing initial values $x_1(0) \sim \mathbb{U}(0,100), \ldots, x_5(0) \sim \mathbb{U}(0,100)$ at equidistant time points between $t_0 = 0$ and $t_n = 1$ with $t_{i+1} - t_i = 0.1$ and interaction weights from $\mathbb{U}(-1,1)$. We also added normally distributed measurement noise $\varepsilon \sim \mathbb{N}(0,10)$ to each simulated data point as shown in Figure 5.3.

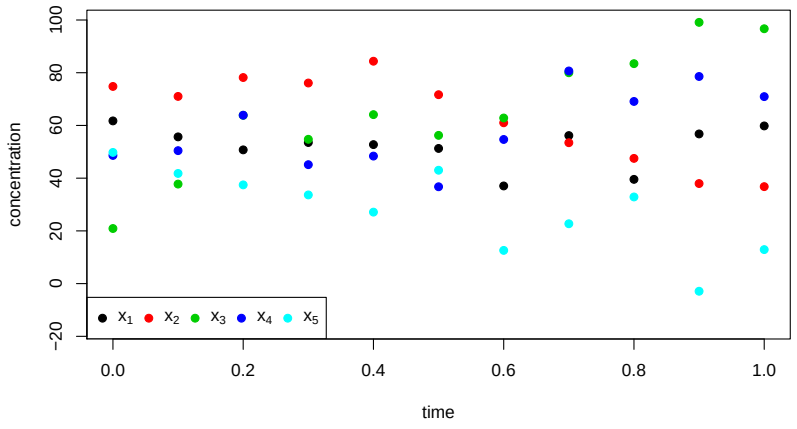

Figure 5.3: Simulated data from network shown in Figure 5.2B and used for simulation study in section 5.3.2.

The next step in this simulation was to apply three techniques in order to estimate the correct catalyst for each interaction. First, we applied a very extensive search for the best model in which we fitted all possible models. In this case there are $6^7 \approx 300000$ different models. For each model, we optimized a (log)-likelihood function. The models were then ordered by the optima of the log-likelihood values with the most appropriate model having the highest log-likelihood value. Results

Table 5.1: Results of fitting all possible models for the network shown in Figure 5.2. Here, we see the best six models and the worst model with respect to negative log-likelihood value. The model used to generate the data is highlighted in red.

rank	$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$	-(log-likelihood)
1	x_3	x_3	x_4	x_1	x_4	x_4	x_2	324.34
2	x_3	x_5	x_4	x_3	x_4	x_5	x_2	325.37
3	x_3	x_3	x_1	x_1	x_5	x_2	1	326.50
4	x_3	x_4	x_1	x_1	x_5	x_2	x_2	326.57
5	x_3	x_1	x_5	x_3	x_5	x_5	1	326.75
6	x_3	x_1	x_5	x_1	x_5	x_4	x_2	327.65
279936	1	1	1	x_2	x_1	1	x_5	1331.40

of selected models are shown in Table 5.1. Here, we present the seven different reactions in one column each and show the catalysts of these reactions in the rows. In this notation, a 1 denotes an uncatalysed, linear reaction. The model used to generate the data is ranked on the fifth place with other top-ranked models being very similar in topology.

Second, we applied a greedy forward selection method. Here, the idea is to start from the null model with no catalysis in the network (Figure 5.2A) and subsequently allow for one catalytic reaction after another. For the studied network, it means that we calculate the log-likelihood of only one model in the first step, then 35 models in the second step (we have 5 possible catalysts for 7 different interactions) with 35 corresponding log-likelihoods. Subsequently, we choose *one* catalyst for *one* interaction corresponding to the model with the highest log-likelihood

Table 5.2: Results of applying a forward model selection to the network shown in Figure 5.2. The best model, corresponding to the highest log-likelihood value, is achieved in step 2. The data-generating (”true”) model does not equal the chosen one.

steps	$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$	-(log-likelihood)
step 0	1	1	1	1	1	1	1	436.57
step 1	x_3	1	1	1	1	1	1	394.10
step 2	x_3	1	1	1	x_5	1	1	386.92
step 3	x_3	1	1	1	x_5	1	x_2	411.28

value and move on to the third step where another catalyst is selected from 30 different models in the same manner. We stop when the log-likelihood is not longer increased by a subsequent step. The results of this procedure are shown in Table 5.2. The stepwise model selection stops after inclusion of three catalysed reactions and fails to identify the correct model by far.

Finally, we applied our method and selected model candidates with threshold $\tau = 0.1$. With our approach we select 288 model candidates and compute the cor-

Table 5.3: Results of application of our latent catalyst approach to the network shown in Figure 5.2. Each reaction has a different number of possible components which may act as a catalyst. The data generating model is highlighted in red.

$x_1 \to x_2$	$x_2 \to x_1$	$x_1 \to x_3$	$x_2 \to x_3$	$x_3 \to x_4$	$x_3 \to x_5$	$x_4 \to x_5$
$\{x_3\}$	$\{x_1, x_3, x_4, x_5\}$	$\{x_1, x_4, x_5\}$	$\{x_1, x_3\}$	$\{x_4, x_5\}$	$\{x_2, x_4, x_5\}$	$\{1, x_2\}$

responding log-likelihood. The component candidates for each model are shown in Table 5.3. The model used to generate the data is included in these model candidates.

Application of the three different model selection techniques revealed different aspects. On the one hand, the forward selection is very fast due to the low number of models being fitted, however it fails in detecting a model that can fit the data reasonably well. On the other hand, the best subset selection does not only find the correct model which was used to generate the data shown in Figure 5.3 but it also finds four models which fit the data more appropriately. This can be explained by the fact that we added a high amount of measurement noise to the true ODE solutions and thus created data situations where the data generating model is not anymore the model that best fits the data. Nevertheless, we believe that this represents a scenario which is much more realistic for real-world applications than looking at the true ODE solutions as measurements where the data generating model will fit the data best by a large margin. The computational cost of this procedure is very high even for this medium-sized example as it runs a total of roughly 76 days on a single core machine (faster with parallelisation). Finally, our approach with modelling latent catalysts also reveals the best model which fit the data best. This is achieved in a very efficient way by reducing the possible model candidates to 288, which is a reduction by 99.897%.

5.4 Application: CD95 apoptosis signalling model

In this section, we apply our method to real-world data collected from the cluster of differentiation 95 (CD95) signalling pathway (Lavrik *et al.* [2007]). This pathway is relevant for regulation of cell death decisions and is mediated via proteins FADD and procaspase-8 as well as its cleavage products p43/p41 and p18 (see Figure 5.4). The pathway can be summarized in the following steps: after extracellular binding of the CD95 ligand to its receptor, FADD is recruited to CD95.

This creates the death inducing signalling complex (DISC), and procaspase-8, c-FLIP long (c-$FLIP_L$) and c-FLIP short (c-$FLIP_S$) can bind to it. This results in the formation of different types of dimers: procaspase-8 homodimers (p8hod), procaspase-8 heterodimer (p8hed) and c-$FLIP_L$ heterodimers. Next, the procaspase-8 part of the dimers is split and is in its active form of p43 homodimer (p43hod) and p43 heterodimer (p43hed). Finally, p43hod is processed to form the cleavage product p18. Next, procaspase-8 homo- and hetero- dimers undergo autocatalytic processing resulting in the formation of the p43 homodimer (p43hom) and p43 heterodimer (p43hod), respectively. The latter along with the cleavage product of procaspase-8, p43 comprises the cleavage product of c-FLIP, p43-FLIP. All of the steps described above have been reported in literature ([Fricker *et al.*, 2010; Kischkel *et al.*, 1995; Lavrik *et al.*, 2007; Neumann *et al.*, 2010]). The last three reactions highlighted in green in Figure 5.4 are known to be possibly catalysed (Rickert *et al.* [2013]). The focus of our work lies in the analysis of a small core motif containing 5 species and 3 reactions. The experimental data used in this manuscript provides measurements of the total p43, total p18 and total procaspase-8 concentration for two time-resolved experiments over a total of 6 time points each. The two experiments differ from each other in the amount of ligand used. We modelled both experiments separately thus obtaining two sets of results for the present data.

Our approach resulted in a very strong reduction of model candidates. Without our approach, a total of 216 models need to be computed and compared for each of the two experiments. With our approach, we are able to narrow down the number of model candidates to 3 and 12 for experiment 1 and 2, respectively. These models are presented in Table 5.4. Here, we ranked the models by their corresponding negative log-likelihood. For both experiments, we calculated some models that clearly outperform all others in this measure (candidate 1 for experiment 1 and candidates 1–3 for experiment 2). Furthermore, we make the observation that there is a large difference of the number of model candidates which were identified with our method in both experiments. Intriguingly, when more ligand is present in the system (experiment 2), non-catalysed splitting of procaspase-8 heterodimer and non-catalysed processing of p43 homodimer are emerging as reactions contributing to the increase of model candidates. This is intuitively understandable because the more ligand is present at the beginning of the experiment, the more procaspase-8 and p43 will be produced and thus a catalysis appears less necessary

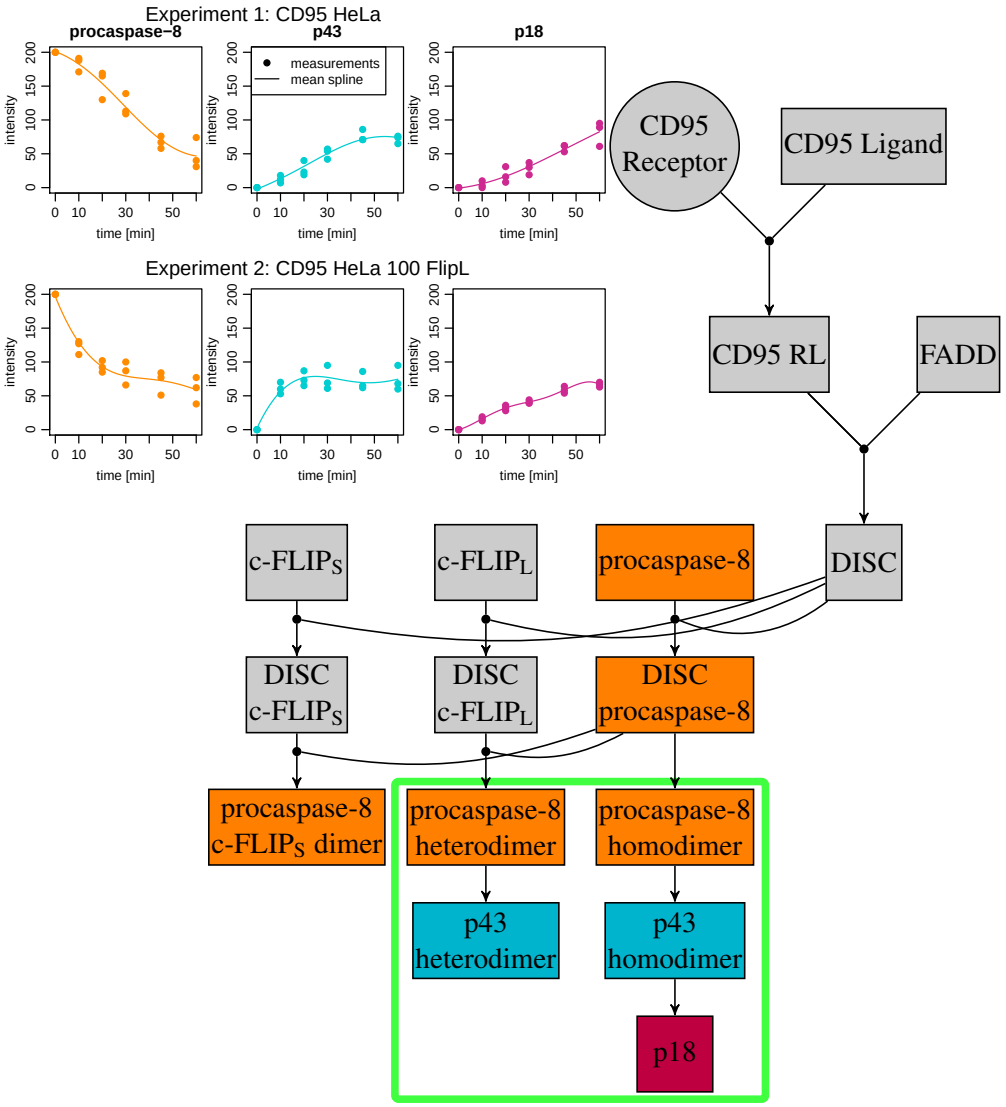

Figure 5.4: Data and schematic representation of CD 95 pathway. Dots show three replicates of two experiments at each time point of procaspase-8 (orange), p43 (blue) and p18 (red). Corresponding lines show the mean spline approximations. In the pathway, rectangles denote proteins, the extracellular receptor is denoted by a circle. Coloured rectangles indicate the different experimental measurements: total amount procaspase-8 (orange), total amount p43 (blue) and total amount p18 (red). The green box indicates the core motif, which is analysed by our method.

Table 5.4: Results of application of the latent catalyst approach to CD95 pathway. The three possibly catalysed reactions are shown on top of the table and the possible catalysts associated with the respective reactions are shown in the rows. All suggested model candidates are shown for experiment 1 (low amount of ligand) and experiment 2 (high amount of ligand). Models are ranked by their negative log-likelihood.

experiment 1				
	p8hed → p43hed	p8hod → p43hod	p43hod → p18	-(log-likelihood)
candidate 1	p8hed	1	p8hed	66.95
candidate 2	p8hed	p43hed	p8hed	71.05
candidate 3	p8hed	p43hod	p8hed	71.09
experiment 2				
	p8hed → p43hed	p8hod → p43hod	p43hod → p18	-(log-likelihood)
candidate 1	p8hed	p43hod	p8hed	66.53
candidate 2	p8hed	1	1	68.23
candidate 3	p8hed	p43hod	1	69.45
candidate 4	p8hed	1	p8hed	76.00
candidate 5	1	p43hed	1	78.06
candidate 6	p8hed	p43hed	1	78.42
candidate 7	1	p43hod	1	78.58
candidate 8	1	p43hed	p8hed	78.92
candidate 9	p8hed	p43hed	p8hed	83.42
candidate 10	1	1	1	84.67
candidate 11	1	1	p8hed	97.86
candidate 12	1	p43hod	p8hed	98.13

in the system. The results further suggest four possible catalysts: procaspase-8 heterodimer as a possible catalyst of the splitting of procaspase-8 heterodimer (autocatalysis), procaspase-8 homodimer and p43 heterodimer as catalysts for the splitting of procaspase-8 homodimer and finally procaspase-8 heterodimer as catalyst for the processing of p43 homodimer. These results are in good agreement with previous analysis of the data (Rickert *et al.* [2013]), where three of the four proposed catalysts from our approach are suggested by the authors. The additional catalysed reaction (procaspase-8 heterodimer catalysing processing of p43 homodimer) was excluded prior to application of the proposed model reduction scheme. Additionally, we also see model candidates where the reactions are not catalysed especially for the experiment with high amount of ligand. We can therefore conclude that by adding more ligand at the beginning of the experiment catalytic reactions play only a minor role in the CD95 pathway. The presence of a low amount of ligand, however, enforces catalysis in the studied system. Upon overcoming the apoptotic threshold and upon high stimulation the additional catalytic reactions are not required any more.

5.5 Discussion

Modelling becomes more complex if catalysis is considered since the number of model candidates increases exponentially. Depending on the size of the studied system, available approaches such as best subset model selection or stepwise selection schemes become infeasible due to the high amount of computational time involved. Additionally, with an increase of model candidates, the hazard of overfitting becomes larger. We proposed a novel and efficient method to incorporate catalysis into the modelling of biological systems. With our approach, we efficiently reduce the number of model candidates to a manageable number with a low probability of missing the most appropriate model for given data. We do this by extending the studied system by latent catalyst components. Subsequently, we estimate those components with a combination of different methods: smoothing splines, ODE modelling and likelihood estimation. Finally, we compare those estimates to all other components of the studied system and each comparison is associated with a score between 0 and 1. This score can then be used to identify relevant components which may act as catalysts for a given reaction. Another byproduct of our approach is the automatic identification of model parameters during the estimation steps.

We studied the proposed method on several simulation settings and noted a substantial decrease of model candidates and at the same time we were able to recover the true models in almost all performed simulations. The application of our method to the CD95 apoptosis pathway confirmed previous results in literature and additionally identified different catalysts for some reactions. We could also conclude that the presence of ligand at the beginning in this system is also a factor which seems to drive the importance of catalysis at later stages.

Overall, we are confident that our method is a useful tool which can be used to gain additional knowledge out of network-based and time-resolved measured data and allows for different conclusions regarding catalysis. Based on such findings, we expect additional hypotheses for future research to be generated and thus lead to a better understanding of the studied system.

6

Discussion and Outlook

This thesis provides novel statistical methods for the analysis of biological systems based on functional data. These methods extend the available statistical tools in two fields. First, we contribute to the field of statistical hypothesis testing by introducing a novel statistical test for differences in two functional groups with paired observations. Second, we propose two novel schemes for an systemised improvement of the topology of a studied system in terms of model fit. In the following, we briefly summarize the key topics discussed in this thesis and provide clues as to how to further pursue the presented ideas.

6.1 Summary

We develop a significance test for the difference between two groups of paired temporal observations in Chapter 3. Other available methods fail to make use of the full information contained in the data such as the pairing between observations or the time dependency contained in the data. With the proposed test, we are able to use the full data information and provide an approximative p-value which can be used to answer the question of global differences between the two groups. The test is based on representation of the temporal observations by smooth functions. The functional mean and functional standard deviation of these smooth functions are used to compute a test statistic. The distribution of this test statistic is approximated by sampling from the null hypothesis of no differences between both groups with preservation of the functional variability. Subsequently, the percentile

method is applied to approximate the final p-value of the test based on the resampled test statistic distribution.

The method is flexible enough to be applied even on temporal data with a low resolution and is also powerful enough to produce meaningful results e. g. in situations with low sample sizes per group or high variability. Compared to other available methods, which do not use the full information of the data, our test is able to clearly outperform the competitors in almost all considered synthetic examples.

We apply the developed test on two real-world data examples.

First, we analyse a pilot study from the field of nutritional science where two different meal challenges are posed to the same study participants on two different days thus incorporating a pairing in the studied data. With our approach we are able to use the full information in the available data and identify solely one out of several hundred metabolites which showed significant difference between both groups. We see this result as a clear indication that meal standardization cannot reveal substantial differences at least in pilot studies with a low number of samples.

Second, we analyse heterochromatin data, where we are able to show that compared to wild-type cells chromatin accessibility in Atrx knock-out cells is significantly increased on IAP elements. This result also demonstrates the role of Atrx as crucial for fast and efficient establishment of heterochromatin. Overall, our test provides strong evidence for a general role of Atrx for establishment as well as robust maintenance of heterochromatin domains.

Next, in Chapter 4, we develop a novel method for the identification of latent components in biological systems. We target biological systems with a temporal resolution which can be modelled with ODEs and systematically extend these systems by the addition of a latent component. We impose hardly any restrictions concerning the shape of the time course of this component and allow and estimate interactions between the latent component and all other parts of the biological system. Additionally, we also demonstrate the ability of our method to reconstruct a misspecified structure of a biological system and thus contribute to the identification of the correct underlying topology. The method is proposed as a combination of dynamical modelling in the sense of ODEs and functional data analysis in the sense of spline approximation of temporal data. Additionally, we estimate system

parameters with a likelihood approach and are able to identify existing interactions with a model selection approach.

Applied on synthetic data, the method is shown to cope with typical difficulties connected to the analysis of biological systems such as missing data, low temporal resolution, large fraction of unobserved parts of the studied system, and high magnitude of noise.

We also apply the method on real-world data arising from the JAK-STAT signalling pathway. Here, we are able to show that proposed modelling approaches of the studied pathway are connected to insufficient model performance and propose to extend the pathway by an additional latent component. We also present the probable time course of this component and based on its shape we give clues on the role it might be playing in the studied system. With this information at hand, guidance for additional future experiments is provided and thus the systems biology loop can be further advanced for the JAK-STAT signalling pathway.

Finally, in Chapter 5, we introduce a novel way for verification of catalysis in biological systems. With our approach we aim to investigate possible reactions being catalysed in already established biological structures. With a growing system size, the model space expands exponentially if all combinations of catalytic and non-catalytic reactions are considered and therefore the computation of all possible models becomes infeasible for larger systems simply due to the computational demand. As we show in Chapter 5, greedy algorithms which investigate only one reaction at a time often fail to find the most appropriate model for given data. With our approach we are able to reduce the model space significantly and at the same time do not discard model candidates which present eligible model candidates. Hereby, we build up on our latent variable approach from Chapter 4 and first identify non-linear interactions between established parts of the biological system and additionally introduced latent components. Once the latent components time-courses are estimated, we compare them with the time-courses of the other components of the system and use a similarity score to decide which latent component can be replaced by already established components. Overall, we arrive at a reduced model candidate space by employing a threshold which controls the magnitude of model reduction.

Synthetic data examples demonstrate an excellent performance of our approach as opposed to other available methods.

We apply the identification of catalytic reactions to data arising from the CD95 apoptosis pathway. Here, one of the results is the effective reduction of possible models by more than 90%. Furthermore, we could also conclude that the presence of ligand at the beginning in this biological system is also a factor which seems to drive the importance of catalysis at later stages as our approach shows that even low amount of stimuli can drive the cell into apoptosis.

6.2 Outlook

This work presents the basis for statistical modelling of functional data in biological systems. Furthermore, based on this work additional research questions can be investigated in future research. We now discuss some of these additional topics.

Concerning the significance test developed in Chapter 3, we identify three further research questions connected to this topic.

First, note that a major component of the test is the time-course estimation through splines. We generally recommend the usage of smoothing splines. Here a smoothing parameter is estimated with leave-one-out cross validation. However, other computationally inexpensive methods such as mixed models (Wood [2004]) or generalized cross validation (Hastie & Tibshirani [1990]) are also possible estimation techniques and may lead to slightly different results. We performed several robustness analyses with regard to the smoothness of the approximated time-courses and found that it did not play a major role for the studied simulations. However, we can imagine scenarios where the smoothness is of high importance for the successful application of TPDT and think that this is one possibility to further study our method.

Second, one limitation of our method is the unknown distribution of the test statistic. Assessing this distribution through resampling is the computationally most expensive part of the test. One way of approaching this problem is the extension of our test to a Bayesian setting where one could impose prior distributions on the spline parameters (basis coefficients and smoothing parameter). Markov chain Monte Carlo sampling would then lead to a more efficient estimation of the distribution of the test statistic. Whether this gains computational time can be assessed in future work, however we think that the test statistic approximation can be made

more robust in this way. Therefore, this extension presents a promising possibility in the further development of the proposed test.

Third, a further extension of TPDT is given in the comparison of multiple groups. In a univariate context and under certain assumptions this question can be answered with ANOVA. Extensions of ANOVA to time-resolved measurements have been proposed (Cuevas *et al.* [2004]; González-Rodríguez *et al.* [2012]). However, the pairing of samples from each group is not considered in any of the proposed approaches and could be seized by a TPDT extension. More specifically, as the development of TPDT was motivated towards an extension of a paired samples t-test, in a setting with multiple groups the appropriate method for an extension seems to be the repeated measures ANOVA (Gueorguieva & Krystal [2004]).

With regard to our approach of identification of latent components in biological systems, we again propose three topics, which present targets for future research.

First, it generally holds that a key element in model building is the estimation of parameters and possibly topology from data. In Chapter 4, we propose to interpret model estimation as a latent variable problem in a dynamical system. We target applications in which latent variables are influencing observations but not vice versa. A coupling in the latent component in the sense of feedback of observed network components to the latent component is possible; however, we mainly see two limitations of this approach. First, additional assumptions about the latent component must be made, and second, including the latent component into the system of ODEs limits its shape and does not allow for additional flexibility. Nevertheless, inclusion of a feedback to the latent component presents a target for future research.

Second, we currently limit our method and only allow linear model extensions. We already partly covered non-linear extensions of this method in Chapter 5. Nevertheless, further extension of our method to more general settings contains additional research potential. More specifically, one could consider multiple independent latent variables that influence a system of differential equations. In this scenario, in addition to the estimation of the latent time courses, we study possible ways of separation of the single variables. Here, blind source separation techniques (Blöchl & Theis [2009]) present one possible ansatz for this task.

Third, one might investigate additional model selection methods that are applicable to our method and and compare them to the already implemented ones. Examples are established methods such as likelihood ratio tests, lasso ([Tibshirani, 1996]), and elastic net ([Zou & Hastie, 2005]). Extending the method to Bayesian theory would further allow the application of Bayes factors ([Kass & Raftery, 1995]) and thermodynamic integration ([Kirk *et al.*, 2013; Schmidl *et al.*, 2012]).

Finally, in our approach for estimation of catalytic reactions, we see additional three topics which contain fruitful research potential.

First, we restricted all analyses and applications to a linear catalysis in the form of $x(t) \cdot h(t)$. This can, however, be extended to more general, non-linear settings in the form of $h(x(t))$. Furthermore, the generalization of the form of $h(t)$ can be performed in a shape limitation approach. For example, it is realistic from a biological point of view to introduce a threshold value for the concentration of the hidden catalyst which forces the intensity of this catalyst to be 0 if this threshold value is not reached and to be described by smooth functions only after the threshold value is reached. This would resemble biological systems where an external stimulus such as drug intake is added to the system only after a certain amount of time and not at the beginning of the study or experiment.

Second, we assume that all possible catalysts are already *part of the network*. This assumption can be relaxed and we could allow for *external catalysts*, which were previously not part of the modelled species. We studied this network extension in Chapter 4 and a combination of both proposed methods could be a desirable feature on the method side.

Third, we also see another possible extension within the formulation of the whole method in Bayesian fashion. Here, we think of formulating multinomial prior distributions describing the probability of a specific reaction to be catalysed by a certain component. With proper sampling methods one then could potentially approximate the distributions of the similarity scores introduced in this manuscript.

In conclusion, understanding biology is a process which is studied in close collaboration between experimentalists and method or data scientists for many decades now. Both, experimental techniques as well as available methods have undergone an enormous advancement throughout this time. The gain that is provided by better methods is two-fold. On the one hand, improved methods allow a more detailed

insight into various biological processes, on other hand new types of data arising from better experimental techniques pose new demands for its analysis. With this work we contribute to the overall advancement on the method side and thus add significant value to the interpretation of results on the data side.

A

Further theory and simulations for latent causes approach

A.1 Log-normally distributed multiplicative noise

Normally distributed error terms, as in (4.10), imply that the noise level is independent of the magnitude of the measurements. Although this assumption is commonly made, it is not appropriate for all biological applications, e. g. when concentrations are measured over time. This assumption is particularly problematic for concentrations close to zero because the model description of $x_i^{\text{obs}}(t_j)$ can then easily become negative. This difficulty is avoided by assuming multiplicative log-normally distributed noise:

$$x_i^{\text{obs}}(t_j) = x_i(t_j) \cdot \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathbb{LN}\left(-\frac{\sigma^2}{2}, \sigma^2\right).$$

The parameter choice of the log-normal distribution is motivated because of the implication $\mathbb{E}\left(x_i^{\text{obs}}(t_j)\middle|x_i(t_j)\right) = x_i(t_j)$.

The distribution of ε_{ij} immediately propagates to the measurements:

$$x_i^{\text{obs}}(t_j)\middle|x_i(t_j) \overset{\text{iid}}{\sim} \mathbb{LN}\left(\log(x_i(t_j)) - \frac{\sigma^2}{2}, \sigma^2\right).$$

As discussed in Section 4.3.2 we exploit the ODE structure of our system which

leads to the approximate distribution

$$x_i^{\text{obs}}(t_j)\big|\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j) \overset{\text{iid}}{\sim} \mathbb{LN}\left(\log(\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)) - \frac{\sigma^2}{2}, \sigma^2\right).$$

The parameter σ^2 is estimated as

$$\hat{\sigma}^2_{\text{ML}} = -2 + 2\sqrt{1 + \frac{\sum_{i=1}^{N}\sum_{j=0}^{n}\left[\log\frac{x_i^{\text{obs}}(t_j)}{\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)}\right]^2}{N(n+1)}}.$$

The diagonal elements of the expected Fisher information matrix are derived as

$$I_k(\boldsymbol{a},\sigma^2) = \begin{cases} \frac{1}{\sigma^2}\sum_i\sum_j\left(\frac{\frac{\partial}{\partial a_k}\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)}{\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)}\right)^2 & k \le N \\ \frac{N(n+1)}{2}\left(\frac{1}{\sigma^4} + \frac{1}{2\sigma^2}\right) & k = N+1. \end{cases}$$

A.2 Example for parameter uncertainty calculation

In this section we evaluate the goodness of fit for parameters associated with a specific example. For simplicity, we choose a small network comprising two species and normally distributed measurement noise. The two parameters of interest are a_1 and σ^2. We analytically compute the CRLB for both parameters and examine whether it is achieved using a simulation with known parameters.

The specific network is described by (4.2) and Figure 4.1B and we consider a network size $N = 2$ and linear combinations of the components. Without loss of generality, we assume $a_2 = 1 - a_1$ and $a_1 \in (0,1)$.

The CRLB for parameter a_1 involves $\frac{\partial}{\partial a_1}\hat{x}_i^{\text{ode},\boldsymbol{a}}(t_j)$. The value of this term depends on the numerical method chosen for solving the ODE. If the Euler method is used, the recursive solution has the form

$$\begin{aligned} \hat{x}_1^{\boldsymbol{a}}(t_{j+1}) &= \hat{x}_1^{\boldsymbol{a}}(t_j) + \Delta(c_1\hat{x}_1^{\boldsymbol{a}}(t_j) + c_2\hat{x}_2^{\boldsymbol{a}}(t_j) + a_1\hat{h}^{\boldsymbol{a}}(t_j)) \\ \hat{x}_2^{\boldsymbol{a}}(t_{j+1}) &= \hat{x}_2^{\boldsymbol{a}}(t_j) + \Delta(c_3\hat{x}_2^{\boldsymbol{a}}(t_j) + c_4\hat{x}_1^{\boldsymbol{a}}(t_j) + (1-a_1)\hat{h}^{\boldsymbol{a}}(t_j)), \end{aligned}$$

with $c_1 = -k_2 - k_1$, $c_2 = k_3$, $c_3 = -k_4 - k_3$, $c_4 = k_2$, Δ denoting the time step, and $\hat{x}_i^{\boldsymbol{a}}$ being a shortened version of $\hat{x}_i^{\boldsymbol{a},\text{ode}}$. We assume that the initial values $x_i(t_0)$ do

not depend on a_1. One can then show by full induction that, for $j = 1, \ldots, n$,

$$\begin{aligned} \frac{\partial x_1(t_j)}{\partial a_1} &= \frac{F_1(t_j)}{(1-a_1)^2} - \frac{F_2(t_j)}{a_1^2} \\ \frac{\partial x_2(t_j)}{\partial a_1} &= \frac{F_3(t_j)}{a_1^2} + \frac{F_4(t_j)}{(1-a_1)^2}, \end{aligned} \tag{A.1}$$

with

$$\boldsymbol{F}(t_{j+1}) = \boldsymbol{F}(t_j) + \Delta \begin{pmatrix} c_1 F_1(t_j) + c_2 F_4(t_j) + \frac{1}{2}\hat{h}_2^0(t_j) \\ c_1 F_2(t_j) - c_2 F_3(t_j) \\ c_3 F_3(t_j) - c_4 F_2(t_j) - \frac{1}{2}\hat{h}_1^0(t_j) \\ c_3 F_4(t_j) + c_4 F_1(t_j) \end{pmatrix}$$

and $\boldsymbol{F}(t_0) = (0,0,0,0)^T$. $\hat{h}_i^0(t)$ are the unweighted estimates of the time course of the hidden component, as described in (4.7). The entries of $\boldsymbol{F}(t_j)$ are independent of a_1. For given σ^2, $\boldsymbol{k}$, a_1, spline approximations of $\boldsymbol{x}(t)$ and data dimensions, Equation (A.1) allows the analytical computation of the expected Fisher information $I(a_1)$, and hence that of the CRLB $I^{-1}(a_1)$. For this specific example, the expected Fisher information matrix has a diagonal form and $I^{-1}(\sigma^2)$ equals $\frac{2\sigma^4}{N(n+1)}$.

The just derived CRLBs are lower bounds for the MSE of the maximum likelihood estimates. Our estimation procedure, however, consists of two steps: spline approximation and maximum likelihood estimation, each entailing uncertainty in the parameter estimates. We hence computed Monte Carlo estimates for the MSE of a_1 and σ^2 using two different approaches. First, we used the true hidden time course during the estimation procedure. The resulting empirical MSE is the one resulting from the maximum likelihood step and is bounded below by I^{-1}. Second, we estimated the hidden time course as well. The resulting MSE is slightly larger and accounts for the uncertainty of the overall estimation procedure.

Results of the simulation are shown in Table A.1. We examine different combinations of σ^2 and $\boldsymbol{a}$. For each combination, we simulate 500 time courses at 100 time points and estimate the parameters. We numerically compute the MSE of these 500 estimates. In the table, we show this MSE and the corresponding CRLB for a given parameter combination.

The results of Table A.1 show that the CRLB is achieved for σ^2 if we consider

Table A.1: Results of the parameter uncertainty simulation. MSE and corresponding CRLB (in parentheses) for the two parameters of interest.

		MSE (CRLB)	
		$\boldsymbol{a} = \begin{pmatrix} .5 \\ .5 \end{pmatrix}$	$\boldsymbol{a} = \begin{pmatrix} .9 \\ .1 \end{pmatrix}$
maximum likelihood approximation			
$\sigma = 1$	a_1	$9.90 \times 10^{-6} (1.63 \times 10^{-6})$	$3.73 \times 10^{-9} (9.27 \times 10^{-10})$
	σ^2	$1.07 \times 10^{-2} (1.00 \times 10^{-2})$	$1.02 \times 10^{-2} (1.00 \times 10^{-2})$
$\sigma = .5$	a_1	$2.33 \times 10^{-6} (4.08 \times 10^{-7})$	$9.04 \times 10^{-10} (2.32 \times 10^{-10})$
	σ^2	$6.19 \times 10^{-4} (6.25 \times 10^{-4})$	$6.38 \times 10^{-4} (6.25 \times 10^{-4})$
$\sigma = .1$	a_1	$9.06 \times 10^{-8} (1.62 \times 10^{-8})$	$3.54 \times 10^{-11} (9.27 \times 10^{-12})$
	σ^2	$1.04 \times 10^{-6} (1.00 \times 10^{-6})$	$9.83 \times 10^{-7} (1.00 \times 10^{-6})$
maximum likelihood and spline approximation			
$\sigma = 1$	a_1	$1.55 \times 10^{-4} (1.24 \times 10^{-5})$	$5.23 \times 10^{-6} (1.13 \times 10^{-7})$
	σ^2	$7.59 \times 10^{-2} (1.00 \times 10^{-2})$	$1.33 \times 10^{-1} (1.00 \times 10^{-2})$
$\sigma = .5$	a_1	$1.62 \times 10^{-5} (3.13 \times 10^{-6})$	$9.57 \times 10^{-7} (3.44 \times 10^{-8})$
	σ^2	$4.29 \times 10^{-3} (6.25 \times 10^{-4})$	$5.44 \times 10^{-3} (6.25 \times 10^{-4})$
$\sigma = .1$	a_1	$6.22 \times 10^{-7} (1.25 \times 10^{-7})$	$2.69 \times 10^{-8} (1.08 \times 10^{-9})$
	σ^2	$6.24 \times 10^{-6} (1.00 \times 10^{-6})$	$4.31 \times 10^{-6} (1.00 \times 10^{-6})$

only maximum likelihood approximation. If we additionally consider the uncertainty introduced by the spline approximation, the ratio between MSE and CRLB increases slightly. For a_1, we observe a similar result in that the ratio increases if we consider spline approximation, while, nevertheless, providing MSE values that are very close to the corresponding CRLB. Another interesting result is the increase in fit quality for smaller σ^2 values, as we already discussed in Section 4.3.3.

B

Additional TPDT examples

In this chapter, we will show additional information about the studied nutritional challenges in Section 3.5.1. Recall that overall we have four different challenges: Non-standardized Western Diet (NWD), Standardized Western Diet (SWD), Healthy Breakfast (HB) and Oral Lipid Test (OLT).

In the main part of the manuscript we investigated one of the six challenge comparisons in detail. Here, we show the results of TPDT applied on the other five challenge comparisons. For these additional comparisons we found a large number of metabolites which showed significant differences in both groups. Table B.1 summarizes the number of significant metabolites for each challenge. Table B.2 – Table B.6 show the specific metabolites for each challenge as well as the corresponding p-values.

Table B.1: Number of significant (after multiple testing correction) metabolites per challenge.

Challenge	Number of significant metabolites	Detailed information
NWD vs. SWD	1	Chapter 3
NWD vs. HB	7	Table B.2
NWD vs. OLT	8	Table B.3
SWD vs. HB	10	Table B.4
SWD vs. OLT	13	Table B.5
HB vs. OLT	18	Table B.6

Non-standardized Western Diet vs. Healthy Breakfast

For the comparison of NWD and HB challenges we identified 7 metabolite time-courses which showed significant differences after FDR correction. The results are shown in Table B.2.

Table B.2: Metabolites with significant differences found with TPDT for challenge comparison NWD vs. HB.

metabolite name	*u*-statistic	p-value	adjusted p-value
CMPF	6.13	0.00	0.01
isoleucine	4.14	0.00	0.01
N-methyl proline	7.26	0.00	0.00
stachydrine	8.35	0.00	0.00
X - 09789	6.38	0.00	0.01
X - 11360	4.51	0.00	0.01
X - 18913	4.45	0.00	0.05

Non-standardized Western Diet vs. Oral Lipid Test

For the comparison of NWD and OLT challenges we identified 8 metabolite time-courses which showed significant differences after FDR correction. The results are shown in Table B.3.

Table B.3: Metabolites with significant differences found with TPDT for challenge comparison NWD vs. OLT.

metabolite name	*u*-statistic	p-value	adjusted p-value
C18.2	5.98	0.00	0.00
3-methylxanthine	5.31	0.00	0.02
pro-hydroxy-pro	4.43	0.00	0.04
theobromine	7.48	0.00	0.00
X - 11261	5.06	0.00	0.02
X - 11360	4.52	0.00	0.01
X - 12850	6.00	0.00	0.02
X - 13429	4.97	0.00	0.01

Standardized Western Diet vs. Healthy Breakfast

For the comparison of SWD and HB challenges we identified 10 metabolite time-courses which showed significant differences after FDR correction. The results

are shown in Table B.4.

Table B.4: Metabolites with significant differences found with TPDT for challenge comparison SWD vs. HB.

metabolite name	*u*-statistic	p-value	adjusted p-value
C3	4.71	0.00	0.04
2-hydroxydecanoic acid	5.42	0.00	0.00
CMPF	5.74	0.00	0.01
caprate (10:0)	4.64	0.00	0.01
catechol sulfate	7.55	0.00	0.01
glycocholate	4.76	0.00	0.01
hippurate	4.46	0.00	0.02
N-methyl proline	6.06	0.00	0.00
pro-hydroxy-pro	5.36	0.00	0.01
stachydrine	12.97	0.00	0.00

Standardized Western Diet vs. Oral Lipid Test

For the comparison of SWD and OLT challenges we identified 13 metabolite time-courses which showed significant differences after FDR correction. The results are shown in Table B.5.

Table B.5: Metabolites with significant differences found with TPDT for challenge comparison SWD vs. OLT.

metabolite name	*u*-statistic	p-value	adjusted p-value
C18.2	4.89	0.00	0.01
PC.ae.C40.2	3.98	0.00	0.04
lysoPC.a.C17.0	4.40	0.00	0.03
3-indoxyl sulfate	4.12	0.00	0.02
oleoylcarnitine	3.94	0.00	0.02
phenol sulfate	4.35	0.00	0.02
pro-hydroxy-pro	5.99	0.00	0.01
theobromine	9.77	0.00	0.00
X - 11261	6.10	0.00	0.01
X - 11529	3.66	0.00	0.04
X - 11538	3.73	0.00	0.03
X - 13429	5.02	0.00	0.03
X - 13871	5.21	0.00	0.00

Healthy Breakfast vs. Oral Lipid Test

For the comparison of HB and OLT challenges we identified 18 metabolite time-courses which showed significant differences after FDR correction. The results are shown in Table B.6.

Table B.6: Metabolites with significant differences found with TPDT for challenge comparison HB vs. OLT.

metabolite name	u-statistic	p-value	adjusted p-value
C18.1	4.26	0.00	0.01
C18.2	5.11	0.00	0.03
Asn	4.22	0.00	0.04
Cit	5.16	0.00	0.00
Thr	5.75	0.00	0.01
1-oleoylglycerophosphocholine	3.88	0.00	0.01
2-hydroxydecanoic acid	7.10	0.00	0.00
arginine	3.05	0.00	0.02
citrulline	4.32	0.00	0.04
linoleate (18:2n6)	3.36	0.00	0.05
N-methyl proline	7.47	0.00	0.00
oleoylcarnitine	3.72	0.00	0.02
phenol sulfate	4.21	0.00	0.02
stachydrine	7.18	0.00	0.02
theobromine	7.90	0.00	0.00
X - 11261	6.03	0.00	0.02
X - 11470	3.42	0.00	0.05
X - 16480	6.86	0.00	0.00

References

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K., WATSON, J.D. & GRIMSTONE, A. (1995). Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, **20**, 210–210. 33

ALDRIDGE, B.B., BURKE, J.M., LAUFFENBURGER, D.A. & SORGER, P.K. (2006). Physicochemical modelling of cell signalling pathways. *Nature cell biology*, **8**, 1195–1203. 3

ALON, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**, 450–461. 36

ALOY, P. & RUSSELL, R.B. (2006). Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, **7**, 188–197. 108

AMBROISE, C., CHIQUET, J. & MATIAS, C. (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, **3**, 205–238. 77

ANGELINI, C., DE CANDITIIS, D., MUTARELLI, M. & PENSKY, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **6**. 5, 41

ARBOUZOVA, N.I. & ZEIDLER, M.P. (2006). JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions. *Development*, **133**, 2605–2616. 35, 36

ARTAVANIS-TSAKONAS, S., RAND, M.D. & LAKE, R.J. (1999). Notch signaling: cell fate control and signal integration in development. *Science*, **284**, 770–776. 109

ASHKENASY, G., JAGASIA, R., YADAV, M. & GHADIRI, M.R. (2004). Design of a directed molecular network. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10872–10877. 36

BAR-JOSEPH, Z., GERBER, G.K., GIFFORD, D.K., JAAKKOLA, T.S. & SIMON, I. (2003). Continuous representations of time-series gene expression data. *Journal of Computational Biology*, **10**, 341–356. 82

BARABASI, A.L. & OLTVAI, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, **5**, 101–113. 108

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300. 60, 68, 70

BERK, M. (2013). *sme: Smoothing-splines Mixed-effects Models*. 58

BERK, M., EBBELS, T. & MONTANA, G. (2011). A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics*, **27**, 1979–1985. 5, 41, 57

BJÖRCK, A. (1996). *Numerical methods for least squares problems*. Siam. 30

BLISS, C. (1970). *Statistics in biology. Vol. 2.*. McGraw-Hill Book Company, New York & London. 1

BLÖCHL, F. & THEIS, F.J. (2009). Estimating hidden influences in metabolic and gene regulatory networks. In *Independent Component Analysis and Signal Separation*, 387–394, Springer. 80, 131

BOLLEN, K.A. (1998). *Structural equation models*. Wiley Online Library. 77

BOROWIAK, M., MAEHR, R., CHEN, S., CHEN, A., TANG, W., FOX, J., SCHREIBER, S. & MELTON, D. (2009). Small molecules efficiently direct endodermal differentiation of mouse and human embryonic stem cells. *Cell Stem Cell*, **4**, 348–358. 77

BOURC'HIS, D. & BESTOR, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, **431**, 96–99. 70

BRAESS, D. (2012). *Nonlinear approximation theory*, vol. 7. Springer Science & Business Media. 12

BRUNEL, N.J. *et al.* (2008). Parameter estimation of odes via nonparametric estimators. *Electronic Journal of Statistics*, **2**, 1242–1267. 31

BUTCHER, J.C. (1987). *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*. Wiley-Interscience. 29, 86

CAO, Y., WANG, H., OUYANG, Q. & TU, Y. (2015). The free-energy cost of accurate biochemical oscillations. *Nature Physics*. 36

CHENEY, E.W. & LORENTZ, G.G. (1980). *Approximation theory III*. Academic Press. 12

CHICKARMANE, V. & PETERSON, C. (2008). A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS One*, **3**, e3478. 77

CLELAND, W. (1967). The statistical analysis of enzyme kinetic data. *Advances in Enzymology and Related Areas of Molecular Biology, Volume 29*, 1–32. 1

COATES, M., CASTRO, R., NOWAK, R., GADHIOK, M., KING, R. & TSANG, Y. (2002). Maximum likelihood network topology identification from edge-based unicast measurements. In *ACM SIGMETRICS*, vol. 30, 11–20, ACM. 5

CODDINGTON, E.A. & LEVINSON, N. (1955). *Theory of ordinary differential equations*. Tata McGraw-Hill Education. 3, 29

CRAINICEANU, C.M., STAICU, A.M., RAY, S. & PUNJABI, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in medicine*, **31**, 3223–3240. 5, 41, 58

CRAMÉR, H. (1945). *Mathematical methods of statistics*, vol. 9. Princeton university press. 85

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403. 25

CREIXELL, P., REIMAND, J., HAIDER, S., WU, G., SHIBATA, T., VAZQUEZ, M., MUSTONEN, V., GONZALEZ-PEREZ, A., PEARSON, J., SANDER, C. *et al.* (2015). Pathway and network analysis of cancer genomes. *Nature methods*, **12**, 615. 37, 38

CRICK, F. *et al.* (1970). Central dogma of molecular biology. *Nature*, **227**, 561–563. 33

CRICK, F.H. (1958). The biological replication of macromolecules. In *Symp. Soc. Exp. Biol*, vol. 12, 138–163. 33

CUEVAS, A., FEBRERO, M. & FRAIMAN, R. (2004). An ANOVA test for functional data. *Computational statistics & data analysis*, **47**, 111–122. 131

CURRY, H.B. & SCHOENBERG, I.J. (1947). On spline distributions and their limits-the polya distribution functions. In *Bulletin of the American Mathematical Society*, vol. 53, 1114–1114, AMER MATHEMATICAL SOC 201 CHARLES ST, PROVIDENCE, RI 02940-2213. 15

DARGATZ, C. (2010). *Bayesian inference for diffusion processes with applications in life sciences*. Ph.D. thesis, lmu. 27

DARNELL, J. (1997). Stats and gene regulation. *Science*, **277**, 1630–1635. 98

DE BOOR, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, **6**, 50–62. 23

DE BOOR, C. (2001). *A practical guide to splines*, vol. 27. Springer Verlag. 15, 18, 82

DE RIDDER, D., DE RIDDER, J. & REINDERS, M.J. (2013). Pattern recognition in bioinformatics. *Briefings in Bioinformatics*. 105

DONALDSON, R. & CALDER, M. (2010). Modelling and analysis of biochemical signalling pathway cross-talk. *arXiv preprint arXiv:1002.4062*. 36

EFRON, B. & TIBSHIRANI, R.J. (1994). *An introduction to the bootstrap*. CRC press. 49

EISENMESSER, E.Z., MILLET, O., LABEIKOVSKY, W., KORZHNEV, D.M., WOLF-WATZ, M., BOSCO, D.A., SKALICKY, J.J., KAY, L.E. & KERN, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438**, 117–121. 5

ELMORE, S. (2007). Apoptosis: a review of programmed cell death. *Toxicologic pathology*, **35**, 495–516. 36

EMMERT-STREIB, F., DEHMER, M. & HAIBE-KAINS, B. (2014). Untangling statistical and biological models to understand network inference: the need for a genomics network ontology. *Frontiers in genetics*, **5**. 76, 105

ENDLER, L., RODRIGUEZ, N., JUTY, N., CHELLIAH, V., LAIBE, C., LI, C. & LE NOVÈRE, N. (2009). Designing and encoding models for synthetic biology. *Journal of The Royal Society Interface*, **6**, S405–S417. 105

EVANS, A.M., DEHAVEN, C.D., BARRETT, T., MITCHELL, M. & MILGRAM, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*, **81**, 6656–6667. 67

FAHRMEIR, L., KNEIB, T. & LANG, S. (2007a). *Regression*. Springer. 5, 26, 60

FAHRMEIR, L., KÜNSTLER, R., PIGEOT, I. & TUTZ, G. (2007b). *Statistik*. Springer. 5

FALL, C. (2002). *Computational Cell Biology*, vol. 20. Springer Verlag. 101

FAN, J. & ZHANG, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**, 179. 60

FASMAN, G.D., SOBER, H.A. *et al.* (1977). *Handbook of biochemistry and molecular biology*, vol. 4. CRC press Cleveland. 33

FATHERS, K.E., STONE, C.M., MINHAS, K., MARRIOTT, J.J., GREENWOOD, J.D., DUMONT, D.J. & COOMBER, B.L. (2005). Heterogeneity of Tie2 expression in tumor microcirculation: influence of cancer type, implantation site, and response to therapy. *The American journal of pathology*, **167**, 1753–1762. 40

FOX, L. & PARKER, I.B. (1968). *Chebyshev polynomials in numerical analysis*, vol. 29. Oxford university press London. 13

FRICKER, N., BEAUDOUIN, J., RICHTER, P., EILS, R., KRAMMER, P.H. & LAVRIK, I.N. (2010). Model-based dissection of CD95 signaling dynamics reveals both a pro- and antiapoptotic role of c-FLIPL. *The Journal of cell biology*, **190**, 377–389. 122

GAO, P., HONKELA, A., RATTRAY, M. & LAWRENCE, N.D. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70–i75. 77

GEAR, C.W. (1971). *Numerical initial value problems in ordinary differential equations*. Prentice Hall PTR. 29

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 721–741. 31

GIBBONS, R.J., PICKETTS, D.J., VILLARD, L. & HIGGS, D.R. (1995). Mutations in a putative global transcriptional regulator cause X-linked mental retardation with α-thalassemia (ATR-X syndrome). *Cell*, **80**, 837–845. 70

GIBNEY, M.J., WALSH, M., BRENNAN, L., ROCHE, H.M., GERMAN, B. & VAN OMMEN, B. (2005). Metabolomics in human nutrition: opportunities and challenges. *The American journal of clinical nutrition*, **82**, 497–503. 34

GIROLAMI, M. (2008). Bayesian inference for differential equations. *Theoretical Computer Science*, **408**, 4–16. 31

GIRVAN, M. & NEWMAN, M.E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**, 7821–7826. 5

GONZÁLEZ-RODRÍGUEZ, G., COLUBI, A. & GIL, M.Á. (2012). Fuzzy data treated as functional data: A one-way anova test approach. *Computational Statistics & Data Analysis*, **56**, 943–955. 131

GOOLEY, A.A., HUGHES, G., HUMPHERY-SMITH, I., WILLIAMS, K.L. & HOCHSTRASSER, D.F. (1996). From proteins to proteomes: Large scale protein identification by twodimensional electrophoresis and amino acid analysis. *Biotechnology*, **14**, 1. 33

GUEORGUIEVA, R. & KRYSTAL, J.H. (2004). Move over anova: Progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry*, **61**, 310–317. 131

GUERTIN, K.A., MOORE, S.C., SAMPSON, J.N., HUANG, W.Y., XIAO, Q., STOLZENBERG-SOLOMON, R.Z., SINHA, R. & CROSS, A.J. (2014). Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *The American journal of clinical nutrition*, ajcn–078758. 35

GUO, W. (2002). Functional mixed effects models. *Biometrics*, **58**, 121–128. 57

GUYON, I. & ELISSEEFF, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**, 1157–1182. 6

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796. 61

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., HASTIE, T., FRIEDMAN, J. & TIBSHIRANI, R. (2009). *The elements of statistical learning*, vol. 2. Springer. 14, 15, 18

HASTIE, T.J. & TIBSHIRANI, R.J. (1990). *Generalized additive models*, vol. 43. CRC Press. 130

HELMAN, E., LAWRENCE, M.S., STEWART, C., SOUGNEZ, C., GETZ, G. & MEYERSON, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome research*, **24**, 1053–1063. 70

HOERL, A.E. & KENNARD, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67. 23

HONKELA, A., GIRARDOT, C., GUSTAFSON, E.H., LIU, Y.H., FURLONG, E.E., LAWRENCE, N.D. & RATTRAY, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, **107**, 7793–7798. 77

HORNSTEIN, E. & SHOMRON, N. (2006). Canalization of development by microRNAs. *Nature Genetics*, **38**, S20–S24. 77

HORVATH, C.M. (2000). Stat proteins and transcriptional responses to extracellular signals. *Trends in biochemical sciences*, **25**, 496–502. 35

HOYER, P.O., SHIMIZU, S. & KERMINEN, A.J. (2006). Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. *arXiv preprint cs/0603038*. 77

HUANG, B., EBERSTADT, M., OLEJNICZAK, E.T., MEADOWS, R.P. & FESIK, S.W. (1996). NMR structure and mutagenesis of the Fas (APO-1/CD95) death domain. *Nature*, **384**, 638–641. 36

HULL, T., ENRIGHT, W., FELLEN, B. & SEDGWICK, A. (1972). Comparing numerical methods for ordinary differential equations. *SIAM Journal on Numerical Analysis*, **9**, 603–637. 29

JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z.N. & BARABÁSI, A.L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654. 35, 109

JESS, T., ZIMMERMANN, E., KRING, S.I.I., BERENTZEN, T., HOLST, C., TOUBRO, S., ASTRUP, A., HANSEN, T., PEDERSEN, O. & SØRENSEN, T.I. (2008). Impact on weight dynamics and general growth of the common FTO rs9939609: a longitudinal Danish cohort study. *International journal of obesity*, **32**, 1388–1394. 40

KASS, R.E. & RAFTERY, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795. 132

KELL, D.B. (2006). Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug discovery today*, **11**, 1085–1092. 34

KHOLODENKO, B.N. (2006). Cell-signalling dynamics in time and space. *Nature reviews Molecular cell biology*, **7**, 165–176. 40

KIRK, P., THORNE, T. & STUMPF, M.P. (2013). Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*, **24**, 767 – 774. 132

KISCHKEL, F., HELLBARDT, S., BEHRMANN, I., GERMER, M., PAWLITA, M., KRAMMER, P. & PETER, M. (1995). Cytotoxicity-dependent APO-1 (Fas/CD95)-associated proteins form a death-inducing signaling complex (DISC) with the receptor. *The EMBO journal*, **14**, 5579. 122

KITANO, H. (2002a). Computational systems biology. *Nature*, **420**, 206–210. 3, 108

KITANO, H. (2002b). Systems biology: a brief overview. *Science*, **295**, 1662–1664. 3

KLINGMÜLLER, U., BERGELSON, S., HSIAO, J. & LODISH, H. (1996). Multiple tyrosine residues in the cytosolic domain of the erythropoietin receptor promote activation of STAT5. *Proceedings of the National Academy of Science*, **93**, 8324–8328. 98

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, vol. 14, 1137–1143. 25

KRAUSS, G. (2006). *Biochemistry of signal transduction and regulation*. John Wiley & Sons. 35

KREUTZ, C., RAUE, A., KASCHEK, D. & TIMMER, J. (2013). Profile likelihood in systems biology. *FEBS Journal*, **280**, 2564–2571. 31

KUANG, Y. (1993). *Delay differential equations: with applications in population dynamics*. Academic Press. 27

LAVRIK, I.N., GOLKS, A., RIESS, D., BENTELE, M., EILS, R. & KRAMMER, P.H. (2007). Analysis of CD95 threshold signaling triggering of CD95 (FAS/APO-1) at low concentrations primarily results in survival signaling. *Journal of Biological Chemistry*, **282**, 13664–13671. 36, 121, 122

LAWRENCE, N.D. (2010). *Learning and inference in computational systems biology*. MIT press. 31

LEE, P.M. (2012). *Bayesian statistics: an introduction*. John Wiley & Sons. 26

LIN, T., AMBASUDHAN, R., YUAN, X., LI, W., HILCOVE, S., ABUJAROUR, R., LIN, X., HAHM, H., HAO, E., HAYEK, A. *et al.* (2009). A chemical platform for improved induction of human iPSCs. *Nature Methods*, **6**, 805–808. 77

LIU, C., MEN, X., YANG, A. & CHANG, J. (2014). Generalized B-spline curve surface and its properties. *Journal of Chemical & Pharmaceutical Research*, **6**, 64–70. 18

LODISH, H.F., BERK, A., ZIPURSKY, S.L., MATSUDAIRA, P., BALTIMORE, D., DARNELL, J. *et al.* (2000). *Molecular cell biology*, vol. 4. Citeseer. 35

LOHR, M., PROHL, A., OSTERMANN, C., LIEBLER-TENORIO, E., SCHROEDL, W., AEBY, S., GREUB, G. & REINHOLD, P. (2014). A bovine model of a respiratory Parachlamydia acanthamoebae infection. *Pathogens and disease*. 40

LOTKA, A.J. (1910). Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, **14**, 271–274. 27

LUÍS, P.B., RUITER, J.P., IJLST, L., DE ALMEIDA, I.T., DURAN, M., MOHSEN, A.W., VOCKLEY, J., WANDERS, R.J. & SILVA, M.F. (2011). Role of isovaleryl-CoA dehydrogenase and short branched-chain acyl-CoA dehydrogenase in the metabolism of valproic acid: implications for the branched-chain amino acid oxidation pathway. *Drug metabolism and disposition*, dmd–110. 68

MARQUARDT, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, **11**, 431–441. 30

MARX, V. (2013). Biology: The big challenges of big data. *Nature*, **498**, 255–260. 1, 33

MASEL, R.I. *et al.* (2001). *Chemical kinetics and catalysis*. Wiley-Interscience New York. 5, 111

MASON, J.C. & HANDSCOMB, D.C. (2002). *Chebyshev polynomials*. CRC Press. 13

MASOUDI-NEJAD, A., SCHREIBER, F. & KASHANI, Z. (2012). Building blocks of biological networks: a review on major network motif discovery algorithms. *IET systems biology*, **6**, 164–174. 36

MATHER, K. (1943). Statistical analysis in biology. 1

MAY, R.M. *et al.* (1976). Simple mathematical models with very complicated dynamics. *Nature*, **261**, 459–467. 1

MCCAIN, J. (2013). The MAPK (ERK) pathway: Investigational combinations for the treatment of BRAF-mutated metastatic melanoma. *Pharmacy and Therapeutics*, **38**, 96. 35

MCCORMACK, S.E., SHAHAM, O., MCCARTHY, M.A., DEIK, A.A., WANG, T.J., GERSZTEN, R.E., CLISH, C.B., MOOTHA, V.K., GRINSPOON, S.K. & FLEISCHMAN, A. (2013). Circulating branched-chain amino acid concentrations are associated with obesity and future insulin resistance in children and adolescents. *Pediatric obesity*, **8**, 52–61. 40

MCCULLOCH, C.E. & NEUHAUS, J.M. (2001). *Generalized linear mixed models*. Wiley Online Library. 26

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H. & TELLER, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087–1092. 31

MICHIELS, W. & NICULESCU, S.I. (2014). *Stability, Control, and Computation for Time-Delay Systems: An Eigenvalue-Based Approach*, vol. 27. Society for Industrial and Applied Mathematics. 27

MIHALIK, S.J., MICHALISZYN, S.F., DE LAS HERAS, J., BACHA, F., LEE, S., CHACE, D.H., DEJESUS, V.R., VOCKLEY, J. & ARSLANIAN, S.A. (2012). Metabolomic profiling of fatty acid and amino acid metabolism in youth with obesity and type 2 diabetes evidence for enhanced mitochondrial oxidation. *Diabetes Care*, **35**, 605–611. 40

MONECKE, A. & LEISCH, F. (2012). semPLS: structural equation modeling using partial least squares. *Journal of Statistical Software*, **48**, 1–32. 77

MÜLLER, T., FALLER, D., TIMMER, J., SWAMEYE, I., SANDRA, O. & KLINGMÜLLER, U. (2004). Tests for cycling in a signalling pathway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 557–568. 100

NAGAEV, K. (1995). Influence of electron-electron scattering on shot noise in diffusive contacts. *Physical Review B*, **52**, 4740. 31

NEUMANN, L., PFORR, C., BEAUDOUIN, J., PAPPA, A., FRICKER, N., KRAMMER, P.H., LAVRIK, I.N. & EILS, R. (2010). Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. *Molecular systems biology*, **6**, 352. 122

NICHOLSON, J.K., CONNELLY, J., LINDON, J.C. & HOLMES, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature reviews Drug discovery*, **1**, 153–161. 34

NIKOLOV, S., GEORGIEV, G., KOTEV, V. & WOLKENHAUER, O. (2007). Stability analysis of a time delay model for the JAK-STAT signalling pathway. *Series on Biomechanics*, **23**, 52–65. 101

NISHINO, Y., YAMADA, Y., EBISAWA, K., NAKAMURA, S., OKABE, K., UMEMURA, E., HARA, K. & UEDA, M. (2011). Stem cells from human exfoliated deciduous teeth (shed) enhance wound healing and the possibility of novel cell therapy. *Cytotherapy*, **13**, 598–605. 40

PAULSSON, J. (2004). Summing up the noise in gene networks. *Nature*, **427**, 415–418. 84

PEARL, R. (1977). *The biology of population growth*. Ayer Publishing. 1

PEIFER, M. & TIMMER, J. (2007). Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *Systems Biology, IET*, **1**, 78–88. 31

PIRMAN, T., RIBEYRE, M.C., MOSONI, L., RÉMOND, D., VRECL, M., SALOBIR, J. & MIRAND, P.P. (2007). Dietary pectin stimulates protein metabolism in the digestive tract. *Nutrition*, **23**, 69–75. 68

PONS, O. (1955). Partial differential equations. 27

POWELL, M. (1967). Curve fitting by cubic splines, Rep. TP307. *Atomic Energy Research Establishment*. 19

PRAJAPATI, S.I., MARTINEZ, C.O., BAHADUR, A.N., WU, I.Q., ZHENG, W., LECHLEITER, J.D., MCMANUS, L.M., CHISHOLM, G.B., MICHALEK, J.E., SHIREMAN, P.K. *et al.* (2009). Near-infrared imaging of injured tissue in living subjects using IR-820. *Molecular imaging*, **8**, 45. 40

PROCHAZKOVA, J. (2005). Derivative of B-spline function. In *Proceedings of the 25th Conference on Geometry and Computer Graphics. Prague, Czech Republic*. 18

PTITSYN, O. (1969). Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *Journal of molecular biology*, **42**, 501–510. 1

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 29, 90, 116

RADICCHI, F., CASTELLANO, C., CECCONI, F., LORETO, V. & PARISI, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2658–2663. 5

RAFTERY, A.E., LEWIS, S., BERNARDO, J., BERGER, J., DAWID, A. & SMITH, A. (1992). Bayesian statistics. *Bayesian statistics*. 26

RAMB, R., EICHLER, M., ING, A., THIEL, M., WEILLER, C., GREBOGI, C., SCHWARZBAUER, C., TIMMER, J. & SCHELTER, B. (2013). The impact of latent confounders in directed network analysis in neuroscience. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**, 20110612. 77

RAMSAY, J.O. & SILVERMAN, B.W. (2005). *Functional data analysis*. Springer, New York. 3, 14, 15, 18, 19, 20, 23, 24, 25, 47, 82

RAMSAY, J.O., WICKHAM, H., GRAVES, S. & HOOKER, G. (2009). *fda: Functional Data Analysis*. R package version 2.1.2. 116

RAO, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81–91. 85

RASER, J.M. & O'SHEA, E.K. (2005). Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010–2013. 84

RAUE, A., KREUTZ, C., MAIWALD, T., BACHMANN, J., SCHILLING, M., KLINGMÜLLER, U. & TIMMER, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929. 31, 32

RAUE, A., BECKER, V., KLINGMÜLLER, U. & TIMMER, J. (2010). Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **20**, 045105. 32

RAUE, A., KREUTZ, C., THEIS, F.J. & TIMMER, J. (2013). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **371**, 20110544. 32

RAWLINGS, J.S., ROSLER, K.M. & HARRISON, D.A. (2004). The JAK/STAT signaling pathway. *Journal of cell science*, **117**, 1281–1283. 35

RICE, J.R. (1969). *The approximation of functions*, vol. 2. Addison-Wesley Reading, Mass. 12, 19

RICKERT, D., FRICKER, N., LAVRIK, I.N. & THEIS, F.J. (2013). Systematic complexity reduction of signaling models and application to a CD95 signaling model for apoptosis. In *Systems Biology of Apoptosis*, 57–84, Springer. 6, 109, 122, 124

RITCHIE, M.D., HOLZINGER, E.R., LI, R., PENDERGRASS, S.A. & KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, **16**, 85–97. 33, 34

ROBERTS, P. & DER, C. (2007). Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, **26**, 3291–3310. 35

RÖMISCH-MARGL, W., PREHN, C., BOGUMIL, R., RÖHRING, C., SUHRE, K. & ADAMSKI, J. (2012). Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. *Metabolomics*, **8**, 133–142. 67

ROSS, S. (1984). *Differential Equations*, vol. 3. New York: John Wiley & Sons. 3, 27, 84

ROSS, S.L. (1980). *Introduction to ordinary differential equations*. John Wiley & Sons. 29

ROWE, H.M., KAPOPOULOU, A., CORSINOTTI, A., FASCHING, L., MACFARLAN, T.S., TARABAY, Y., VIVILLE, S., JAKOBSSON, J., PFAFF, S.L. & TRONO, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome research*, **23**, 452–461. 70

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592. 95

RUNGE, C. (1901). Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, **46**, 20. 13

SCHIKOWSKI, T., SCHAFFNER, E., MEIER, F., PHULERIA, H.C., VIERKÖTTER, A., SCHINDLER, C., KRIEMLER, S., ZEMP, E., KRÄMER, U., BRIDEVAUX, P.O. *et al.* (2013). Improved air quality and attenuated lung function decline: modification by obesity in the SAPALDIA cohort. *Environ Health Perspect*, **121**, 1034—1039. 40

SCHMIDL, D., HUG, S., LI, W.B., GREITER, M.B. & THEIS, F.J. (2012). Bayesian model selection validates a biokinetic model for zirconium processing in humans. *BMC Systems Biology*, **6**, 95. 132

SCHOENBERG, I.J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions, part b: On the problem of osculatory interpolation, a second class of analytic approximation formulae. *Quart. Appl. Math*, **4**, 112–141. 15

SCHUMAKER, L. (2007). *Spline functions: basic theory*. Cambridge University Press. 15

SHAPIRO, E., BIEZUNER, T. & LINNARSSON, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, **14**, 618–630. 33

SHULAEV, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, **7**, 128–139. 34

SLEZAK, D.F., SUÁREZ, C., CECCHI, G.A., MARSHALL, G. & STOLOVITZKY, G. (2010). When the optimal is not the best: parameter estimation in complex biological models. *PloS one*, **5**, e13283. 108

SMITH, M.J., MARSHALL, C.B., THEILLET, F.X., BINOLFI, A., SELENKO, P. & IKURA, M. (2015). Real-time NMR monitoring of biological activities in complex physiological environments. *Current opinion in structural biology*, **32**, 39–47. 40

SOETAERT, K., PETZOLDT, T. & SETZER, R.W. (2010). Solving differential equations in R: package deSolve. *Journal of Statistical Software*, **33**. 29, 116

STEPHENS, Z.D., LEE, S.Y., FAGHRI, F., CAMPBELL, R.H., ZHAI, C., EFRON, M.J., IYER, R., SCHATZ, M.C., SINHA, S. & ROBINSON, G.E. (2015). Big data: Astronomical or genomical? *PLoS Biol*, **13**, e1002195. 1, 2

STOREY, J.D., XIAO, W., LEEK, J.T., TOMPKINS, R.G. & DAVIS, R.W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 12837–12842. 41

STUDENT (1908). The probable error of a mean. *Biometrika*, 1–25. 5, 60

SWAMEYE, I., MÜLLER, T., TIMMER, J., SANDRA, O. & KLINGMÜLLER, U. (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1028. 100, 104

TEN HAVE, G., ENGELEN, M., LUIKING, Y.C. & DEUTZ, N. (2007). Absorption kinetics of amino acids, peptides, and intact proteins. *Int J Sport Nutr Exerc Metab*, **17**, S23–36. 68

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. 23, 132

TIMMER, J., MÜLLER, T., SWAMEYE, I., SANDRA, O. & KLINGMÜLLER, U. (2004). Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, **14**, 2069–2080. 100

TONI, T. & STUMPF, M. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, **26**, 104–110. 100

TOUTENBURG, H. (1992). *Lineare modelle*. Springer. 20

TRUJILLO, E., DAVIS, C. & MILNER, J. (2006). Nutrigenomics, proteomics, metabolomics, and the practice of dietetics. *Journal of the American dietetic association*, **106**, 403–413. 34

VOGEL, V. & SHEETZ, M.P. (2009). Cell fate regulation by coupling mechanical cycles to biochemical signaling pathways. *Current opinion in cell biology*, **21**, 38–46. 109

VOLTERRA, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *J. Cons. Int. Explor. Mer*, **3**, 3–51. 27

WALSH, M.C., BRENNAN, L., MALTHOUSE, J.P.G., ROCHE, H.M. & GIBNEY, M.J. (2006). Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *The American journal of clinical nutrition*, **84**, 531–539. 68

WANG, B. & ENRIGHT, W. (2013). Parameter estimation for ODEs using a cross-entropy approach. *SIAM Journal on Scientific Computing*, **35**, A2718–A2737. 31

WATSON, J.D. *et al.* (1970). Molecular biology of the gene. *Molecular biology of the gene.*. 33

WEBER, M., WU, T., HANSON, J.E., ALAM, N.M., SOLANOY, H., NGU, H., LAUFER, B.E., LIN, H.H., DOMINGUEZ, S.L., REEDER, J. *et al.* (2015). Cognitive deficits, changes in synaptic function, and brain pathology in a mouse model of normal aging. *eNeuro*, **2**, ENEURO.0047. 40

WETTERSTRAND, K. (2015). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). available at: www.genome.gov/sequencingcosts. accessed [18.09.2015]. 2

WILKINS, A.S. (2010). The enemy within: an epigenetic role of retrotransposons in cancer initiation. *Bioessays*, **32**, 856–865. 70

WINNIKE, J.H., BUSBY, M.G., WATKINS, P.B. & O'CONNELL, T.M. (2009). Effects of a prolonged standardized diet on normalizing the human metabolome. *The American journal of clinical nutrition*, **90**, 1496–1501. 68

WISHART, D.S. (2007). Current progress in computational metabolomics. *Briefings in Bioinformatics*, **8**, 279–293. 34

WOLD, S. (1974). Spline functions in data analysis. *Technometrics*, **16**, 1–11. 19

WOOD, S.N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 413–428. 24

WOOD, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**. 24, 130

WOOD, S.N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, **100**, 221–228. 61

WOOD, S.N. & AUGUSTIN, N.H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling*, **157**, 157–177. 24

WOTHKE, W. (2000). Longitudinal and multigroup modeling with missing data. 95

ZACKS, S. (1971). *The theory of statistical inference*, vol. 34. Wiley New York. 86

ZHANG, Y., WOLF-YADLIN, A., ROSS, P.L., PAPPIN, D.J., RUSH, J., LAUFFENBURGER, D.A. & WHITE, F.M. (2005). Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Molecular & Cellular Proteomics*, **4**, 1240–1250. 40

ZHOU, J. & LU, J.A. (2007). Topology identification of weighted complex dynamical networks. *Physica A: Statistical Mechanics and Its Applications*, **386**, 481–491. 5

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320. 23, 132

ZUKUNFT, S., SORGENFREI, M., PREHN, C., MÖLLER, G. & ADAMSKI, J. (2013). Targeted metabolomics of dried blood spot extracts. *Chromatographia*, **76**, 1295–1305. 67