# A Statistical Framework to Optimize Experimental Design for Inference Problems in Systems Biology based on Normalized Data

Von der Fakultät Konstruktions-, Produktions- und Fahrzeugtechnik
und dem Stuttgart Center for Simulation Science
der Universität Stuttgart zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von

Caterina Thomaseth

aus Padova, Italien

| | |
|---|---|
| Hauptberichterin: | Prof. Dr. rer. nat. Nicole Radde |
| Mitberichter: | Prof. Dr. rer. nat. Jens Timmer |
| | Assistant Professor Dr.-Ing. Dirk Fey |

Tag der mündlichen Prüfung: 6. April 2021

Institut für Systemtheorie und Regelungstechnik

Universität Stuttgart

2022

*To Francesco*

# Acknowledgements

The result of my PhD work is not only the outcome of an extensive work of research and cannot be simply reduced to the writing of this thesis. Indeed, it is the fruit of a wide net of human relationships and experiences, of many kinds. At this point, I want therefore to try to acknowledge all the people who contributed to let me reach this goal.

First of all, I want to thank my advisor, Prof. Nicole Radde, who always supported me in every step of my experience at the University of Stuttgart. She always acted very professionally and guided me also through difficult moments. From the academic field, I want to thank all members of the IST: professors Frank Allgöwer and Christian Ebenbauer, all the secretary and IT staff. I was very lucky to find such an inspiring place, which gave me many opportunities to grow intellectually and meet outstanding researchers from all over the world. In particular I want to mention some former colleagues, from "old" and "new" times: Debdas, Antje, Patrick, Eva, Andrei, Dirke, Carsten, Wolfgang, Daniella, Georg, Gregor, Jingbo....so many names and adventures together! Thank you for your fruitful feedbacks and support at work, but also for sharing pleasant time and friendship. Additionally, I want to thank the graduate school SimTech and all related people. Thank you for giving me nice opportunities to grow in my academic skills, to increase my network of relationships and to experience my abroad stay at the UCD in Dublin. This brings me to thank Professor Boris Kholodenko at the Institute for Systems Biology and my mentor Dirk Fey. I spent three amazing months at the ISB: thank you again for the fruitful collaboration and the professional exchange. Last but not least, I want to thank the two reviewers of my PhD thesis, professors Jens Timmer and Dirk Fey.

Leaving the academic field, there are still a few important people that I want to thank: My parents, Eleonora, Giulia, my family and closest friends. Thank you for always being closed to me even if physically far away. My last thought goes to Francesco, my husband and the person to whom I dedicate this work: Thank you to help me to become the best version of myself!

Stuttgart, January 2022
Caterina Thomaseth

# Table of Contents

**Notation** **xi**

**Abstract** **xv**

**Deutsche Kurzfassung** **xvii**

**1 Introduction** **1**
1.1 Data-driven inference problems in Systems Biology . . . . . 1
1.2 Motivation and focus . . . . . 3
1.3 Outline and contributions of this thesis . . . . . 7

**2 The impact of western blot data normalization on statistical inference** **11**
2.1 Introduction . . . . . 12
2.2 Problem formulation . . . . . 13
2.2.1 Statistical description of a knockdown experiment . . . . . 14
2.2.2 Maximum Likelihood estimates . . . . . 16
2.3 The Gaussian ratio distribution . . . . . 18
2.3.1 Statistical properties of ratio distributions of normal random variables 18
2.3.2 Structural non-identifiability of GR distributions . . . . . 21
2.3.3 Convergence properties of ML estimators for GR distributions . . . 25
2.4 Error model selection partially affects the calibrated statistical distributions 32
2.5 Summary and discussion . . . . . 36

**3 Normalization, experimental design and error model choice affect dynamical model calibration of biochemical reaction networks** **41**
3.1 Introduction . . . . . 42
3.2 Problem formulation . . . . . 44
3.2.1 Test-bed model for a reversible phosphorylation reaction . . . . . 47
3.2.2 Normalization strategies of WB time series data . . . . . 48
3.2.3 Statistical description of normalized time series data for dynamic modelling . . . . . 50

3.3 Results . . . 54
3.3.1 Increasing the amount of time points improves the quality of parameter estimates . . . 54
3.3.2 Impact of normalization strategies on the uncertainty of Maximum Likelihood estimates . . . 63
3.3.3 Statistical model comparison . . . 66
3.4 Summary and discussion . . . 69

**4 Impact of measurement noise, experimental design, and estimation methods on Modular Response Analysis based network reconstruction 73**
4.1 Introduction . . . 74
4.2 Problem formulation . . . 77
4.2.1 MAPK and p53 test-bed models with complementary dynamic behaviours . . . 80
4.3 Results . . . 81
4.3.1 Solving the MRA equations results in heavy-tailed distributions for the estimated LRCs . . . 81
4.3.2 Large perturbations tend to improve the inference of pairwise node interactions . . . 83
4.3.3 A simple control strategy is sufficient for the estimation of the LRCs 85
4.3.4 Using MRA with replicate mean values tends to outperform linear regression techniques . . . 87
4.3.5 Replicates increase precision, but not accuracy . . . 89
4.3.6 Non-linearity induces bias, but large perturbations are still required for precision . . . 90
4.3.7 Performance evaluation on the level of discrete network interactions corroborates our quantitative results . . . 92
4.4 Summary and discussion . . . 94

**5 Conclusion 97**
5.1 Summary and discussion . . . 97
5.2 Future outlook . . . 101

**6 Appendix 103**
6.1 Maximum Likelihood Estimation . . . 103
6.1.1 Practical optimization problems . . . 104
6.2 Statistical models comparison: the Akaike and Bayesian Information criteria 105
6.3 ODE model of a reversible phosphorylation reaction . . . 106

6.4 The GR error model for data normalized by the mean value: correlation coefficient . . . 106
6.5 Parametrization of the GR error model for the ODE test-bed model application 108
6.6 Impact of number of time points on the uncertainty of ML estimates . . . 110
6.6.1 Estimated parameters for increasing $K - J = 1$ . . . 112
6.6.2 Estimated parameters for increasing $K - J = 6$ . . . 115
6.6.3 Estimated parameters for increasing $K - J = 10$ . . . 118
6.7 Statistical model comparison - high noise level . . . 121
6.8 Estimated $\hat{\sigma}_{\text{MLE}}$ . . . 122
6.9 The MAPK and the p53 ODE models . . . 122
6.10 Medcouple . . . 124
6.11 MRA estimation methods . . . 125

**Bibliography** **127**

**Publications of the author** **133**

# Notation

The following acronyms and mathematical notation are used throughout the entire thesis. In addition, for clarity, chapter specific notations are listed explicitly.

## Acronyms

### Signalling pathways

| **Abbreviation** | **Full name** |
|---|---|
| EGF | Epidermal Growth Factor |
| ERK | Extracellular signal-Regulated Kinase |
| MAPK | Mitogen-Activated Protein Kinase |
| MEK | MAPK/ERK Kinase |
| NGF | Nerve Growth Factor |
| p53 | tumour suppressor protein |
| PC12 | cell line derived from a PheoChromocytoma of the rat adrenal medulla |
| Raf | Rapidly Accelerated Fibrosarcoma |

### Experimental techniques

| **Abbreviation** | **Full name** |
|---|---|
| CS | Control Strategy |
| KD | Knockdown |
| NS | Normalization Strategy |
| OE | Overexpression |
| RT-qPCR | Reverse Transcription Quantitative Polymerase Chain Reaction |
| WB | Western Blot |

## Mathematical concepts

| Abbreviation | Full name |
|---|---|
| AIC | Akaike Information Criterion |
| AUC | Area Under the Curve |
| BIC | Bayesian Information Criterion |
| GRC | Global Response Coefficient |
| ML | Maximum Likelihood |
| LRC | Local Response Coefficient |
| ODE | Ordinary Differential Equation |
| ROC | Receiver Operating Characteristic |

## Statistical quantities

| Abbreviation | Full name |
|---|---|
| CV | Coefficient of Variation |
| EM | Error Model |
| GR | Gaussian Ratio distribution |
| IQR | Interquartile Range |
| LMC | Left Medcouple |
| LN | Log-Normal distribution |
| MC | Medcouple |
| MSE | Mean Squared Error |
| N | Normal distribution |
| pdf | Probability Density Function |
| RMC | Right Medcouple |
| RV | Random Variable |
| SD | Standard Deviation |

## Algorithms

| Abbreviation | Full name |
|---|---|
| MLE | Maximum Likelihood Estimation |
| MRA | Modular Response Analysis |
| OLS | Ordinary Least Squares |
| TLS | Total Least Squares |

# Sets of numbers

| Symbol | Description |
|---|---|
| $\mathbb{N}$ | Natural numbers |
| $\mathbb{R}$ | Real numbers |
| $\mathbb{R}_{>0}$ | Positive real numbers |
| $\mathbb{R}_{+}$ | Non-negative real numbers |
| $\mathbb{R}_{+}^{N}$ | Non-negative real vectors of size $N$ |
| $\mathcal{C}^1$ | Continuous differentiable real functions |

# Mathematical variables

| Symbol | Description |
|---|---|
| $x, y, z$ | *Simple* letters: deterministic variables and random variates (realizations) |
| $\mathbf{x}, \mathbf{y}, \mathbf{z}$ | **Bold** letters: random variables |
| $p_{\mathbf{y}}(y)$ | Probability density function of $\mathbf{y}$ |
| $\mathcal{N}$ | Gaussian distribution |
| $M_r$ | Molecular mass |
| $Da$ | Dalton, unit of measurement of molecular mass |
| $\alpha_j$ | WB membrane and antibody specific constant for each replicate $j = 1, \ldots, J$ |
| $T_1$ | First step transformation: noisy raw measurements $\rightarrow$ normalized data |
| $T_2$ | Second step transformation: normalized data $\rightarrow$ inferred parameters |

## Chapter 2: Statistical inference

| Symbol | Description |
|---|---|
| $x_{\mathrm{c}}$, $\mathbf{x}_{\mathrm{c}}$ | Noisy protein amount of $x$ under control condition |
| $x_{\mathrm{k}}$, $\mathbf{x}_{\mathrm{k}}$ | Noisy protein amount of $x$ under knockdown condition |
| $\tilde{y}_{\mathrm{c}}^{j}$, $\tilde{\mathbf{y}}_{\mathrm{c}}^{j}$ | Measured optical density of $x$ under control condition for replicate $j$ |
| $\tilde{y}_{\mathrm{k}}^{j}$, $\tilde{\mathbf{y}}_{\mathrm{k}}^{j}$ | Measured optical density of $x$ under knockdown condition for replicate $j$ |
| $y_{\mathrm{k}}$, $\mathbf{y}_{\mathrm{k}}$ | Normalized data of $x$ under knockdown condition w.r.t. the control case |
| $\hat{\theta}_1$, $\hat{\boldsymbol{\theta}}_1$ | ML estimate of the N error model (scenario 1) |
| $\hat{\theta}_2$, $\hat{\boldsymbol{\theta}}_2$ | ML estimate of the LN error model (scenario 2) |
| $\hat{\theta}_3$, $\hat{\boldsymbol{\theta}}_3$ | ML estimate of the GR error model (scenario 3) |

## Chapter 3: Dynamical model calibration of biochemical reaction networks

| Symbol | Description |
|---|---|
| $\theta$ | Unknown model parameter |
| $\theta_0$ | True parameter value |
| $x_i(t_k, \theta_0)$ | True noise-free concentration of protein $x_i$ at time point $t_k$ |
| $z(t, \theta)$ | Simulated model output |
| $\tilde{x}_i(t_k)$, $\tilde{\mathbf{x}}_i(t_k)$ | Noisy concentration of protein $x_i$ at time point $t_k$ |
| $\tilde{y}_i^j(t_k)$, $\tilde{\mathbf{y}}_i^j(t_k)$ | Measured optical density of $x_i$ at time point $t_k$ for replicate $j$ |
| $y_{i,NS}(t_k), \mathbf{y}_{i,NS}(t_k)$ | Normalized measurement of $x_i$ at time point $t_k$ according to normalizations strategy NS |
| $\hat{\theta}_{\text{MLE}}$, $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ | Noisy ML estimate of model parameter $\theta$ |

## Chapter 4: MRA based network reconstruction

| Symbol | Description |
|---|---|
| $ss$ | Steady-state |
| $\bar{x}_i^0$ | True $ss$ concentration of protein $x_i$ in the control condition |
| $\bar{x}_i^j$ | True $ss$ concentration of protein $x_i$ after perturbation $j$ |
| $p_j$ | Perturbation parameter affecting the corresponding node $j$ |
| $R_{ij}^{true}$ | True analytical definition of GRC |
| $r_{ij}^{true}$ | True analytical definition of LRC |
| $\tilde{R}_{ij}$ | Noise-free approximation of GRC |
| $\tilde{r}_{ij}$ | Noise-free approximation of LRC |
| $\bar{z}_i^0$, $\bar{\mathbf{z}}_i^0$ | Noisy $ss$ concentration of protein $x_i$ in the control condition |
| $\bar{z}_i^j$, $\bar{\mathbf{z}}_i^j$ | Noisy $ss$ concentration of protein $x_i$ after perturbation $j$ |
| $R_{ij}$, $\mathbf{R}_{ij}$ | Noisy GRC obtained from noisy $ss$ data |
| $r_{ij}$, $\mathbf{r}_{ij}$ | Noisy LRC estimated from noisy GRC |

# Abstract

Inference problems in Systems Biology are primarily based on the theoretical assumption that a measured dataset comprises noisy realizations following some underlying stochastic distribution, having well-defined statistical properties. This uncertainty in the input quantities propagates through the inference process, influences the uncertainty of the estimated model parameters and subsequently affects the quality and reliability of model predictions. Understanding the mechanisms of noise propagation over an inference problem will therefore be instrumental in designing an optimal and robust experimental protocol to reduce the uncertainty of the estimated quantities of interest. This thesis investigates the underlying mechanisms of noise propagation from measured experimental data to estimated parameters by developing a statistical framework to characterize and analyse non-linear transformations of stochastic distributions. Among such non-linear transformations, data normalization, a required step for some common experimental techniques, requires specific attention, representing an additional modification of noise properties. Mathematically, the normalization step translates into ratios of two distributions. We consider standard assumptions on the distributions associated with biological raw data. In this thesis we explore three specific classes of inference problems relevant for Systems Biology applications. At first we consider the problem of statistical inference of different parametrized error models for normalized data. Subsequently, we investigate the effect of such error models when coupled with different normalization strategies on results of parameter estimation for dynamic models of biochemical reaction networks. We conclude this thesis by analysing the effects of noise propagation on Modular Response Analysis based network reconstruction. From our simulation results, we observe that non-linear noise transformations may lead to very uncertain and/or erroneous inference results. Additionally, based on the quantification of statistical measures for accuracy and precision of the inference results, we derive practical advice for an optimized and robust experimental design in order to reduce the uncertainty of the estimated quantities.

***Keywords*** — Inference problems, Systems Biology, noise propagation, ratio distributions, parameter estimation, Maximum Likelihood Estimation, statistical inference, dynamic modelling, network reconstruction, experimental design.

# Deutsche Kurzfassung

# Ein statistisches Framework zur Optimierung des experimentellen Designs für Inferenzprobleme in der Systembiologie basierend auf normalisierten Daten

Inferenzprobleme in der Systembiologie basieren in erster Linie auf der theoretischen Annahme, dass ein gemessener Datensatz verrauschte Stichproben nach einer zugrunde liegenden stochastischen Verteilung mit klar definierten statistischen Eigenschaften umfasst. Diese Unsicherheit in den Eingangsgrößen pflanzt sich im Inferenzprozess fort, beeinflusst die Unsicherheit der geschätzten Modellparameter und beeinflusst anschließend die Qualität und Zuverlässigkeit der Modellvorhersagen. Das Verständnis der Mechanismen der Rauschpropagierung bei Inferenzproblemen wird daher entscheidend dazu beitragen, ein optimales und robustes experimentelles Protokoll zu entwickeln, um die Unsicherheit der geschätzten Interessengrößen zu verringern. Diese Arbeit untersucht die zugrunde liegenden Mechanismen der Rauschpropagierung von gemessenen experimentellen Daten zu geschätzten Parametern mit Hilfe eines statistischen Rahmens zur Charakterisierung und Analyse nichtlinearer Transformationen stochastischer Verteilungen. Unter solchen nichtlinearen Transformationen erfordert die Datennormalisierung, ein notwendiger Schritt für einige gängige experimentelle Techniken, besondere Aufmerksamkeit, da eine zusätzliche Änderung der Rauscheigenschaften vorliegt. Mathematisch übersetzt sich der Normalisierungsschritt in Verhältnisse von zwei Verteilungen. Wir machen Standardannahmen über die Verteilung der biologischen Rohdaten. In dieser Arbeit analysieren wir drei spezifische Klassen von Inferenzproblemen, die für systembiologische Anwendungen relevant sind. Zuerst betrachten wir das Problem der statistischen Inferenz verschiedener parametrisierter Fehlermodelle für normalisierte Daten. Anschließend untersuchen wir die Wirkung solcher Fehlermodelle

in Verbindung mit verschiedenen Normalisierungsstrategien auf die Ergebnisse der Parameterschätzung für dynamische Modelle von biochemischen Reaktionsnetzwerken. Wir schließen diese Arbeit mit der Untersuchung der Auswirkungen der Rauschpropagierung auf die Modulare Response-Analyse basierte Netzwerkrekonstruktion ab. Anhand unserer Simulationsergebnissen stellen wir fest, dass nichtlineare Rauschtransformationen zu sehr unsicher und/oder irrigen Inferenzergebnissen führen können. Darüber hinaus leiten wir auf der Basis der Quantifizierung statistischer Messgrößen für die Genauigkeit und Präzision der Inferenzergebnisse praktische Hinweise für ein optimiertes und robustes experimentelles Design ab, um die Unsicherheit der geschätzten Größen zu reduzieren.

# 1 Introduction

## 1.1 Data-driven inference problems in Systems Biology

Biological processes are governed by highly complex regulatory schemes at the molecular level, taking place at different spatial and temporal scales. Research in the field of *Systems Biology* has brought to the stage many challenging problems, which find their motivation in the experimental observations of biological systems and are analysed from systems theoretical and mathematical perspectives. Mathematical modelling approaches can provide useful insight of such intertwined cellular systems, such as signal transduction mechanisms or gene regulation, aiming to interpret experimental results and to unravel unobserved mechanisms (Álvarez-Buylla Roces et al., 2018; Barnes and Chu, 2010). In particular, the strength of Systems Biology is the synergy between combined experimental, modelling and simulation techniques, which allows better understanding of very complex intra- and intercellular processes, despite unavoidable limitations from both experimental and computational sides (Cho and Wolkenhauer, 2003). This collaboration between theoreticians and experimental biologists opens a more advanced and interdisciplinary form of research domain. This leads to the possibility of describing complex biological phenomena by using a precise mathematical description and of simulating and predicting *in silico* the outcomes of expensive lab experiments with the help of advanced computational methods.

Mathematical modelling has therefore emerged as an effective tool which assists biologists to find answers to many open problems which have been raised based on experimental observations, and to investigate innovative therapeutic strategies for many biomedical applications. Interesting examples concern cancer mechanisms (Werner et al., 2014), secretion processes (Weber et al. (2015), Thomaseth et al. (2013)), auto-immune diseases (Kim et al., 2014), metabolic disorders (Schadt and Lum, 2006) and many more. During the last decades, a vast amount of studies was published about multi-scale and multi-dimensional models describing and analysing complex biological systems (Clarke et al., 2019; Martins et al., 2010). An extensive review is provided in Walpole et al. (2013).

One of the core tasks in this context involves inference problems, dealing with experimental data as input to estimate unknown model parameters. Such inference problems include paradigms like parameter estimation and identifiability, whose origins lie in different fields,

such as statistics (Kay, 1993; Koopmans and Reiersol, 1950; Seber and Wild, 1989), control and systems engineering (Åström and Eykhoff, 1971; Ljung, 1987). Experimental data are therefore the input information for a general inference problem. The investigated model predictions depend on the estimated unknown model parameters, which represent the output of the inference problem. Provided unavoidable error sources due to detection limitations and biological random fluctuations effecting experimental measurements, a *statistical model* assumes that the *dataset*:

$$\mathcal{D} = \{y_i, i = 1, ..., N\}, \tag{1.1.1}$$

comprises noisy variates following some underlying stochastic distribution. For the inference process it is important to take into account the statistical properties of the experimental observations, since this uncertainty in the input data propagates over the calculated quantities and effects the inferred outputs and subsequently model predictions.

In this thesis, we consider inference studies in which we describe the dataset (1.1.1) by means of a mathematical model, expressed as a deterministic *non-linear regression model*:

$$z = h(x, \theta). \tag{1.1.2}$$

The variables $z \in \mathbb{R}_+^Q$, often called *observables* or dependent variables, represent the simulated variables of the mathematical model describing the measurable quantities. These functions depend on the unknown model parameters $\theta \in \mathbb{R}^M$ to be estimated and on other variables $x \in \mathbb{R}_+^N$, usually called *regressors* or independent variables, used to explain the behaviour of $z$ (Seber and Wild, 1989). In order to maintain biological feasibility, the quantities $z$ and $x$ assume only non-negative values. Regressor variables may be fixed conditions or random quantities too, depending on the applications. For example, if we consider the application of dynamical systems modelled as Ordinary Differential Equation (ODE) systems, the observables $z$ may depend on the independent variable $t$, i.e. the time.

As mentioned in Seber and Wild (1989), especially in the biological sciences, the underlying processes are usually very complex and often not well understood. The model (1.1.2) represents in these cases an approximation of the reality and the goal is to obtain the most simple regression model which is able to fit the data in a reasonable way. An inference problem consists therefore in an optimization problem in which a function $\mathcal{C}(\mathcal{D}, z(x, \theta))$, relating measured data and simulated variables, has to be optimized:

$$\hat{\theta} = \arg \underset{\theta \in \Theta}{optimum}\, \mathcal{C}(\mathcal{D}, z(x, \theta)). \tag{1.1.3}$$

The set $\Theta \subseteq \mathbb{R}^M$ represents a subset of the parameter space, in which the optimal value is assumed to lie. Most common estimation methods are related to a minimization problem, in which a measure of the "distance" between data and model predictions has to be minimized.

In this thesis, we will focus on three concrete data-driven inference problems, which

are very relevant in the field of Systems Biology, namely statistical inference, dynamical model calibration of biochemical reaction networks and finally network reconstruction. In Section 1.2 we introduce the key topic which inspired and motivated all analyses and results presented in this thesis. Furthermore, we present the motif of the whole thesis, which will be further elaborated in the next chapters for the three application studies. Section 1.3 summarizes the outline and contents of the thesis.

## 1.2 Motivation and focus

One of the most common types of experimental data used for a semi-quantitative description of biochemical systems are *western blot data* (see e.g. Hood et al. (2019); Santos et al. (2007)). Western blot (WB), or immunoblotting, is an experimental procedure that allows the detection and quantification of specific proteins inside a complex mixture.

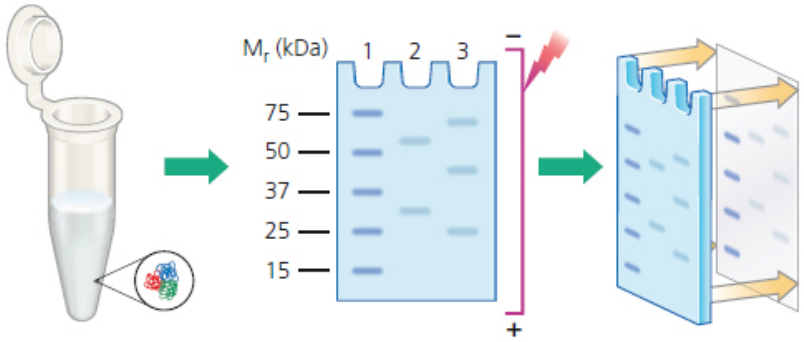


**Fig. 1.1. Western blot experimental procedure.** The lysate obtained from a sample of cells (left) undergoes the process of fractionation in order to separate proteins according to their molecular mass $M_r$, measured in $kDa$ (center). The separated proteins on the gel are then transferred (blotted) onto a membrane (right), which is then incubated sequentially with two different antibodies, leading finally to the quantification of the desired protein concentrations. Figure taken from: Bio-Rad Laboratories webpage.

In the following we highlight in a simplified way the key steps of the method, which are also schematically summarized in Figure 1.1: a sample of cells in a particular experimental condition undergoes the process of lysis in order to release the molecular content of the cells; all proteins are then separated according to their molecular mass $M_r$ by two-dimensional polyacrylamide-gel electrophoresis (fractionation); the separated proteins on the gel are transferred (blotted) onto a membrane; this membrane is then treated with a solution of labelled antibodies that attach only to the specific proteins of interest; these antibodies are then detected by a second group of antibodies that are coupled with a label emitting a fluorescent or chemiluminescent signal; finally, by quantification of these light signals, the concentration of the desired proteins can be calculated (Alberts et al., 2008; Taylor and Posch, 2014).

A peculiarity of WB data is that the measured optical densities must undergo two different steps of *normalization* for a precise quantification and to ensure a correct and reliable comparison of protein levels in different experiments (Degasperi et al., 2014; Taylor and Posch, 2014). As stated in Taylor and Posch (2014), normalization for an appropriate loading control is necessary to take into account possible irregular loading among the lanes of the blot, and therefore to assure that the observed fold changes of the protein levels reveal real changes and are not artefacts. We show an exemplary schematic representation in Figure 1.2, in which four values of one representative protein and of the corresponding loading control $\alpha$-tubulin (commonly used in cellular systems) are detected via WB (left). The raw data $\{y_1, y_2, y_3, y_4\}$, shown on the right, represent the protein response in terms of the measured optical density of the protein normalized for the loading control. Without loss of generality, we assume that these raw data, obtained after the first considered normalization step, define the dataset $\mathcal{D}$ (1.1.1), assuming that the variance in the protein used as loading control is small compared to the variance in the signal. From our discussions with biologists we arrived to the conclusion that this is a reasonable assumption, although it has not yet been investigated in full detail. Furthermore, the same assumption is also implicitly used in similar studies (Degasperi et al., 2014; Kreutz et al., 2007).

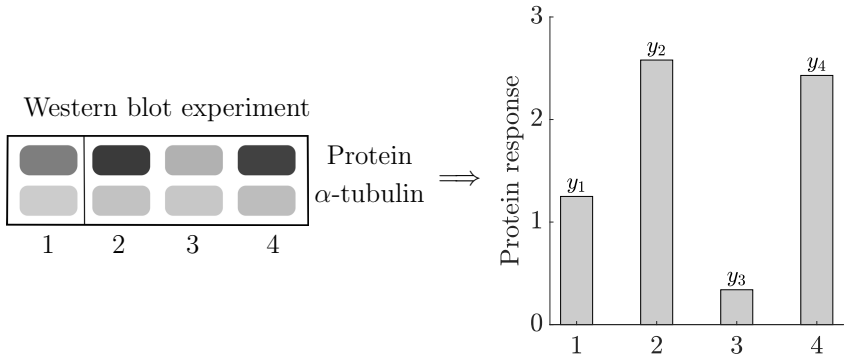


**Fig. 1.2. WB data quantification.** The amount of one exemplary protein and of the loading control $\alpha$-tubulin are measured via WB in four experimental conditions (left part). The dataset $\mathcal{D} = \{y_1, y_2, y_3, y_4\}$ (right part) represents the protein response in terms of the measured optical density of the protein normalized for the loading control.

After this, a second step of normalization is required in order to allow comparison of several replicates[1] quantified in different blots (Degasperi et al., 2014). Different strategies can be considered at this point. Commonly applied options are those with respect to some "control" experiment, e.g. the untreated culture or one representative time point in a time course study. In the case that there is no unique choice for one specific condition to be the control case, data can also be normalized to the mean value of the data on one blot or by optimal alignment for variance minimization (Degasperi et al., 2014).

[1] We refer to Blainey et al. (2014) for a nice overview on measurement replication in biological applications.

As we will explain more in detail in Chapter 2, data post-processing techniques usually entail a significant change of the statistical properties of the experimental data, an issue which has not been broadly taken into consideration in the literature yet. Relating to the inference problems considered in this thesis, data processing by normalization defines the new dataset:

$$\tilde{\mathcal{D}} = \{\tilde{y}_j, j = 1, ..., \tilde{N}\}, \quad \text{where} \quad \tilde{y}_j = T_1(y_i \in \mathcal{D}), \tag{1.2.4}$$

which is used in place of the original dataset (1.1.1) for parameter estimation via (1.1.3). The total number of available data $\tilde{N}$ may also change, as we will see in Chapter 3. The transformation function $T_1$ and which data $y_i$ of the original dataset enter as input of the transformation (1.2.4) depend on the chosen normalization strategy.

One main focus of this thesis is to characterize how the statistical properties of the transformed noisy data may change due to normalization, considering some standard assumptions on the stochastic process underlying data generation of the original dataset (1.1.1). A second main goal of this thesis is to analyse how uncertainty propagates to the estimated parameters, solution of the optimization problem (1.1.3), solved for the new dataset $\tilde{\mathcal{D}} = \{\tilde{y}_j, j = 1, ..., \tilde{N}\}$. The optimization problem provides in fact the solution $\hat{\theta}$ as a non-linear function of the normalized data:

$$\hat{\theta} = T_2(\tilde{y}_j \in \tilde{\mathcal{D}}), \tag{1.2.5}$$

implying the fact that the estimated parameter represents a sample of some underlying random distribution, too. The transformation function $T_2$ depends of course on the chosen normalization strategy via $T_1$, but also on the assumed statistical description of the normalized data, whose mathematical formulation impacts the cost function $\mathcal{C}(\tilde{\mathcal{D}}, z(x, \theta))$ for some estimation methods.

In summary, starting from noisy input raw data (1.1.1), we aim to analyse the effects of the two transformation steps $T_1$ and $T_2$ on the uncertainty of the inferred quantities, for three specific inference problems in Systems Biology. Noise propagation over the inference problem due to these two transformation steps is schematically represented in terms of continuous probability density functions (pdf) in Figure 1.3. Throughout the thesis, bold letters indicate random variables and simple letters refer to realizations.

The problem regarding the effects of data normalization and the subsequent transformations of statistical distributions, as discussed above, has remained unaddressed in the literature so far. Nevertheless, we consider it a relevant investigation since relative data are largely spread in Systems Biology studies, and not only WB data, but also multiplexed Elisa, proteomica or RT-qPCR, as mentioned in Degasperi et al. (2017).

In Degasperi et al. (2014), the authors analyse the effects of different normalization strategies for WB data on the coefficient of variation (CV) and on the results of hypothesis

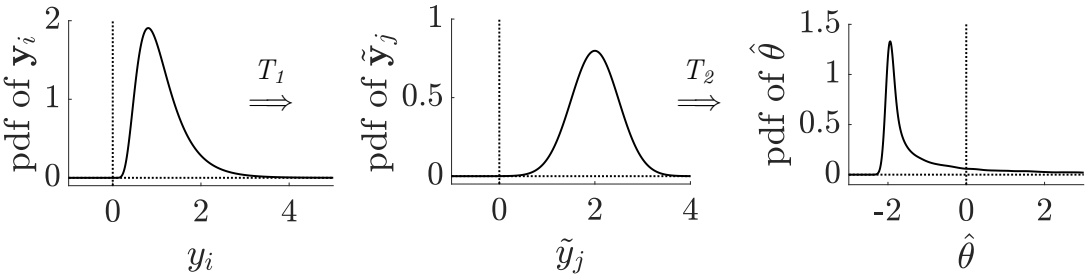


**Fig. 1.3. Noise propagation over the inference problem.** Raw noisy data are interpreted as samples of an underlying distribution, described by the probability density function (pdf of $\mathbf{y}_i$) arising from the measurement noise. We schematically show how the results of the considered inference problem are affected by the variability of the measured protein concentrations in terms of propagation of distributions over the two transformations $T_1$ and $T_2$: from the measurements (left), via normalization (center), to the estimated model parameter $\hat{\theta}$ (right).

testing, e.g. the effects on the number of false positives and false negatives, when using normalized data. This is the first work addressing the effects of normalization on the statistical properties of the data and was motivating the analysis of this doctoral thesis on the effects on the results of inference problems.

In an earlier work (Kirch et al. (2016)), we started having a closer look at dynamical models of intracellular networks, and analysed the effects of model rescaling and normalization on the results of sensitivity analysis. In particular, we considered both cases of local and global sensitivities of model outputs normalized to a reference experiment. Results show that normalization of model outputs has a large impact on the results of both sensitivity analyses and in particular complicates their interpretation.

Therefore, in this thesis, we aim to address the following questions:

- How does noise propagate from input measurement data to output estimated parameters via the two non-linear transformations $T_1$ and $T_2$, whose relationships in general cannot be solved analytically?
- How can we eventually optimize the experimental design and the computational methodology to obtain more precise and/or more accurate inference results?
- How robust is a particular experimental and computational design with respect to larger noise in the input data?

## 1.3 Outline and contributions of this thesis

The main focus of this thesis can be summarized as following: For specific inference problems of interest for Systems Biology studies, we want to develop a statistical framework to analyse noise propagation from noisy measured WB data to uncertain parameter estimates, going through the data normalization step. We evaluate the effects of non-linear transformations of statistical distributions and, based on the outcomes of the suggested approach, we derive practical recommendations to optimize the experimental design required for the considered inference problems.

Throughout the thesis, this analysis is formulated for three specific examples of inference problems: starting from the statistical inference of different error models for normalized WB data, we look at the effects on dynamical model calibration of biochemical reaction networks. We conclude applying our statistical framework to Modular Response Analysis based network reconstruction.

In the following, we outline the structure of this thesis and summarize the main results.

### Chapter 2: *The impact of western blot data normalization on statistical inference*

In this chapter, we elaborate on the statistical description of normalized WB data. Starting from two common assumptions on the distribution of noisy raw data, namely Gaussian and log-normal distributions, we regard three different classes of statistical error models for the obtained relative data. In particular, we consider ratio distributions as the straightforward formal mathematical characterization of the statistical properties of normalized data. Based on a real case study of knockdown experiments for steady-state concentrations, we analyse the impact of normalization on the results of statistical inference, i.e. parameter estimation of the statistical error model describing the fold change between the amount of a protein after knockdown with respect to the untreated control case. Making use of the Maximum Likelihood Estimation (MLE) method, we obtain parameter estimates for the different stochastic models assumed to underlie data generation. Results show that the choice of the error model partially affects the estimated distributions and their moments, even in cases of low number of replicates and large variance across them, which is very common in real studies. This indicates that the three considered error models, under plausible simplifying assumptions, can be fairly compared among each other as suitable description of normalized WB data.

**Remark.** *Parts of the content (text and pictures) presented in Chapter 2 are taken from the following publication by the author of this thesis:*

- *Thomaseth and Radde (2016).*

*The specific contributions of the author of this thesis in those publications consist in: characterizing the different scenarios of statistical description of normalized WB data, summarizing the properties of Gaussian ratio distributions from different published sources, running all numerical simulations and making the analysis of the results.*

## Chapter 3: *Normalization, experimental design and error model choice affect dynamical model calibration of biochemical reaction networks*

This chapter continues the statistical analysis presented in the previous one, extending the investigation on the effects of noise propagation on inference results in the context of dynamic modelling of biochemical reaction networks. We introduce an *in silico* study of an ODE model for a reversible phosphorylation reaction with unknown kinetic rate constants to be estimated from WB time series data. By assuming realistic noise levels for the raw measured data, we analyse the impact of the choice of different error models on the results of model calibration via MLE. Furthermore, we look at the effects of different experimental design features, like the total number of selected time points and alternative normalization strategies, and derive practical recommendations to allow more precise and more accurate parameter estimates. This study on a simple test-bed model highlights the fact that standard noise levels of detected signals and commonly used amounts of data lead to uncertain parameter estimates. Therefore, a sufficiently large amount of measured data is necessary to precisely infer model parameters and subsequently obtain reliable model predictions.

**Remark.** *The results presented in Chapter 3 are partially supported by the work presented in Wang (2018), a student thesis that was supervised by the author of this doctoral thesis. Additionally, parts of the content (text and pictures) are taken from the following publication by the author of this thesis:*

- *Thomaseth and Radde (2021).*

*All contents of this chapter are specific contributions of the author of this thesis and consist in: defining the whole theoretical problem, formulating the statistical framework for the analysis of noise propagation, running all numerical simulations and making the analysis of the results.*

## Chapter 4: *Impact of measurement noise, experimental design, and estimation methods on Modular Response Analysis based network reconstruction*

In this chapter, we present a statistical analysis of noise propagation in the context of signalling network reconstruction, which is the third and last inference problem considered in this thesis. Considering Modular Response Analysis (MRA) as theoretical method to infer and quantify protein interactions based on steady-state perturbation data, we investigate the effects of noisy data processing, experimental design and estimation methods on the uncertainty of the individual inferred protein interactions. We design an *in silico* study of the MAPK and the p53 signalling pathways with realistic noise settings. By means of statistical concepts and measures, we assess accuracy and precision of inferred pairwise interactions and resulting network structures. From our simulated results we draw clear suggestions on how to optimize the performance of MRA based network reconstruction, concerning the choice of experimental and computational features. A robustness analysis of the MRA workflow is also presented, corroborating the recommended strategy for a reliable network reconstruction.

**Remark.** *The main content (text and pictures) presented in Chapter 4 is taken from the following publication by the author of this thesis:*

- *Thomaseth et al. (2018).*

*The specific contributions of the author of this thesis in the published study consist in: developing the statistical framework for the analysis of noise propagation, selecting the proper measure for tail weight of statistical distributions, running all numerical simulations and making the analysis of the results.*

## Chapter 5: Conclusion & outlook

We conclude the thesis by summarizing all results presented in the three previous chapters, providing additional discussion points, and suggesting possible extensions of these research findings for future investigations.

## Appendix

For a better understanding of the content of this thesis, in this section we present relevant mathematical definitions, calculations, and proofs. In addition, we provide pictures illustrating some extensions of the numerical results.

# 2 The impact of western blot data normalization on statistical inference

In this chapter we investigate the impact of normalization of the data obtained from WB experiments on *statistical inference*, dealing with parameter estimation of the underlying stochastic models. Normalization is an important pre-processing step required to enable comparison across different replicates. Variations may arise in fact due to inconsistent sample preparation, unequal sample loading across lanes of a blot and various other experimental conditions. The normalization procedure takes place in two steps: first, the detected signals are normalized to a loading control, and then to a reference condition. Complications in parameter estimation for biochemical network reconstruction arise when the signals are themselves normally distributed, and hence the normalized data are described by the ratio of two normal distributions. In this chapter, we recapitulate a few important properties of such Gaussian ratio distributions and we provide plausible conditions for various approximations that facilitate further statistical analysis. Besides, we analyse the structural identifiability of its parametrization and we investigate convergence properties of the maximum-likelihood estimator by means of an *in silico* study. Results show that many samples are needed to be in the asymptotic regime. In contrast, fold changes, determined as the expected value of the inferred distributions, can relatively accurately be estimated despite large uncertainties in distribution parameters and heavy-tailedness of the ratio distribution. We illustrate the obtained results on a case study in which WB data are used to infer the fold change in a knockdown experiment.

Parts of the content of this chapter are taken from Thomaseth and Radde (2016).

## 2.1 Introduction

Western blotting, introduced by Towbin et al. (1979), is an analytical technique to detect and quantify concentrations of proteins as well as their phosphorylation states. With this technique, proteins isolated from a cell culture are transferred to a membrane and incubated with a specific primary antibody. After that the membrane is incubated with a secondary antibody, which is directed against the first antibody and serves as signal amplifier. This second antibody is chemically linked to an enzyme that catalyzes a chemiluminescence reaction. Finally, exposure of an X-ray film placed against the western blot produces bands, indicating the location of a protein-antibody complex (see Section 1.2 for more precise details about the WB experimental procedure). In the linear range, the band intensity is proportional to the amount of protein. Since proportionality factors are membrane and antibody specific, normalization has to be performed to enable comparison between different replicates. This normalization is commonly carried out in two steps. Signals are first normalized to a loading control to reduce the variance resulting from loading differences among the lanes. In a second step the data are additionally normalized to a control or reference condition. For example, if monitoring the temporal change of the phosphorylation state of a protein in an intracellular signalling cascade upon stimulation with a ligand, the state in the unstimulated case might serve as such a reference condition. Thus, normalized data are given as multiples of this reference state. Similarly, the altered concentration of a protein in a perturbation experiment, e.g. overexpression or knockdown, is normalized to the signal in the unperturbed case. Originally, western blotting was mainly used to detect differences in the amount or activity state of proteins across different conditions, but there is a continuous trend to extract also more quantitative information, such as concentrations of proteins over time or at steady-state. In recent times, experimental protocols have been improved to determine linear ranges between optical densities and protein concentrations, as well as to perform proper background corrections (Taylor et al., 2013). Coupled with these, further developments in experimental technique (see e.g. Taylor and Posch (2014)) resulted in a more frequent use of WB data for parameter estimation of quantitative models (see e.g. Weber et al. (2015)).

Researchers have also started to characterize the statistical properties of WB data (Degasperi et al., 2014; Kreutz et al., 2007). In this context, it should be mentioned that, although often not explicitly stated, the representation of different replicates using summary statistics, such as mean and variance, as well as their calculation, implicitly assumes an underlying distribution of the data. In this respect, a common assumption

consists in defining normally distributed data when it comes to represent them via summary statistics or in the application of hypothesis testing. Examples include t-test or ANOVA to determine the significance of differences in mean values (see e.g. Möller et al. (2014); Zinöcker and Vaage (2012)). On the other hand, based on a comparison of different error models, the authors in Kreutz et al. (2007) argue that WB signals rather follow a log-normal distribution.

In this study we investigate normal and log-normal distributions as statistical models underlying WB data generation. Subsequently, we show that such assumptions have different implications on the distributions of the normalized data, whose straightforward mathematical characterization is given by ratio distributions. In particular, we primarily focus on Gaussian ratio distributions. We review properties of this class of distributions and demonstrate that the parameters of the inverse problem are structurally non-identifiable. This motivated us to use a reduced parametrization and to introduce two simplifying assumptions for the estimation problem. Convergence properties of the maximum likelihood estimator using the alternative parametrization are investigated in an *in silico* study. Results show that fold changes can accurately be estimated despite large uncertainties in distribution parameters. Moreover, our simplifying assumptions highly facilitate parameter and state estimation, since otherwise those parameter estimates are not in the asymptotic limit even for large data set sizes.

We exemplify the theoretical and numerical findings with a real-world case study, in which measured data are used to determine the fold change of a protein quantified in the control as well as in a knockdown experiment (Santos et al., 2007). Using maximum likelihood (ML) estimators, we demonstrate that, under plausible simplifying premises, different assumptions on the underlying distribution of the data may lead to comparable calibrated statistical distributions of the normalized data and corresponding moments.

## 2.2 Problem formulation

A stochastic modelling framework describes data $y \in \mathbb{R}$ as random variates that are generated by an underlying stochastic model $\{\mathbf{y}_i\}_{i \in B}$, where $\mathbf{y}$ is a random variable (RV) indexed by $i \in B$, with $B$ some general set of finite or infinite conditions. This process defines the distribution $p_{\mathbf{y}_i}(y_i)$ (*probability density function* in the case of continuous variables). Statistical inference deals with the problem of estimating $p_{\mathbf{y}_i}(y_i)$ from observations. Here we consider the case of distributions that are parametrized using a set of parameters denoted by $\theta$, such that statistical inference is equivalent to the estimation of $\theta$ from observations $y$. We analyse three different stochastic models underlying data generation for the description of a knockdown (KD) experiment.

### 2.2.1 Statistical description of a knockdown experiment

We consider a WB dataset from a KD experiment, in which the amount of protein has been quantified under control and KD conditions. In order to estimate the fold change, i.e. the factor by which the protein amount is reduced compared to the amount in the control case, the data of the KD experiments are commonly presented normalized to the control case. Here we consider three different scenarios of the stochastic processes underlying experimental data generation.

- **Scenario 1:**
  **No error in the control data and KD data are normally distributed**
  This is the simplest scenario, in which data $x_\mathrm{c}$ of the control experiments are assumed to have zero error, while data $x_\mathrm{k}$ from the KD experiments are assumed to be normally distributed:

$$\begin{aligned} \mathbf{x}_\mathrm{c} &= \delta(\mu_c) \\ \mathbf{x}_\mathrm{k} &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{2.2.1}$$

- **Scenario 2:**
  **Control and KD data are log-normally distributed**
  If $\mathbf{x}_\mathrm{c}$ and $\mathbf{x}_\mathrm{k}$ follow a log-normal distribution, their logarithms are normally distributed:

$$\begin{aligned} \log \mathbf{x}_\mathrm{c} &\sim \mathcal{N}(\mu_c, \sigma_c^2) \\ \log \mathbf{x}_\mathrm{k} &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{2.2.2}$$

- **Scenario 3:**
  **Control and KD data are normally distributed**
  Here we consider the case:

$$\begin{aligned} \mathbf{x}_\mathrm{c} &\sim \mathcal{N}(\mu_c, \sigma_c^2) \\ \mathbf{x}_\mathrm{k} &\sim \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned} \tag{2.2.3}$$

We will consider two subcases of scenarios 2 and 3, namely 2A-B and 3A-B, respectively. For scenarios 2A and 3A, we assume that $\mathbf{x}_\mathrm{c}$ and $\mathbf{x}_\mathrm{k}$ are independent RVs, i.e. $\mathrm{Cov}(\mathbf{x}_\mathrm{c}, \mathbf{x}_\mathrm{k}) = \rho = 0$. On the other hand, in scenarios 2B and 3B, we consider the more general assumption in which $\mathbf{x}_\mathrm{c}$ and $\mathbf{x}_\mathrm{k}$ are correlated, i.e. $\rho \neq 0$. Indeed, correlated experimental errors may be systematically introduced due to gel and transfer inhomogeneities (Schilling et al., 2005).

The RVs $\mathbf{x}_\mathrm{c}$ and $\mathbf{x}_\mathrm{k}$ describe absolute protein amounts, which cannot be observed directly in western blots. Instead, optical densities from secondary antibodies are measured, which are assumed to be proportional to these amounts, with membrane and antibody specific

**Table 2.1.** Distributions of the raw WB signals $\tilde{\mathbf{y}}_{\mathrm{c}}^{j}$ and $\tilde{\mathbf{y}}_{\mathrm{k}}^{j}$ and of the data $\mathbf{y}_{\mathrm{c}}$ and $\mathbf{y}_{\mathrm{k}}$ normalized to the control experiment.

| | **Scenario 1** | **Scenario 2** | **Scenario 3** |
|---|---|---|---|
| $\tilde{\mathbf{y}}_{\mathrm{c}}^{j}$ | $\delta(\alpha_j\mu_c)$ | $\log\mathcal{N}(\log\alpha_j+\mu_c,\sigma_c^2)$ | $\mathcal{N}(\alpha_j\mu_c,(\alpha_j\sigma_c)^2)$ |
| $\tilde{\mathbf{y}}_{\mathrm{k}}^{j}$ | $\mathcal{N}(\alpha_j\mu_k,(\alpha_j\sigma_k)^2)$ | $\log\mathcal{N}(\log\alpha_j+\mu_k,\sigma_k^2)$ | $\mathcal{N}(\alpha_j\mu_k,(\alpha_j\sigma_k)^2)$ |
| $\mathbf{y}_{\mathrm{c}}$ | $\delta(1)$ | $\delta(1)$ | $\delta(1)$ |
| $\mathbf{y}_{\mathrm{k}}$ | $\mathcal{N}\left(\frac{\mu_k}{\mu_c},\left(\frac{\sigma_k}{\mu_c}\right)^2\right)$ | $\log\mathcal{N}(\mu_k-\mu_c,\sigma_k^2+\sigma_c^2-2\rho\sigma_k\sigma_c)$ | $\frac{\mathcal{N}(\mu_k,\sigma_k^2)}{\mathcal{N}(\mu_c,\sigma_c^2)}$ |

constants $\alpha_j$ for each replicate. Hence, the distributions of signals $\tilde{\mathbf{y}}_{\mathrm{c}}^{j}$ and $\tilde{\mathbf{y}}_{\mathrm{k}}^{j}$ that describe the observed optical densities in each replicate $j = 1,\dots J$ are given by:

$$\begin{aligned}\tilde{\mathbf{y}}_{\mathrm{c}}^{j} &= \alpha_j\mathbf{x}_{\mathrm{c}},\\ \tilde{\mathbf{y}}_{\mathrm{k}}^{j} &= \alpha_j\mathbf{x}_{\mathrm{k}}.\end{aligned} \tag{2.2.4}$$

Finally, in order to cancel out the unknown pre-factors $\alpha_j$, all replicates are normalized to the control condition, which gives normalized data $y_{\mathrm{c}}$ and $y_{\mathrm{k}}$ as:

$$y_{\mathrm{c}} = \frac{\tilde{y}_{\mathrm{c}}^{j}}{\tilde{y}_{\mathrm{c}}^{j}} = \frac{x_{\mathrm{c}}}{x_{\mathrm{c}}} = 1,\ \forall j \tag{2.2.5a}$$

$$y_{\mathrm{k}} = \frac{\tilde{y}_{\mathrm{k}}^{j}}{\tilde{y}_{\mathrm{c}}^{j}} = \frac{x_{\mathrm{k}}}{x_{\mathrm{c}}} = T_1(x_{\mathrm{k}},x_{\mathrm{c}}). \tag{2.2.5b}$$

Equation (2.2.5b) indicates how normalization represents a transformation $T_1$ of the absolute noisy data in the two experimental conditions. According to our statistical assumptions, the normalization step implies therefore a transformation of the statistical properties of the normalized data with respect to those assumed for the raw measured data.

As highlighted in Chapter 1, in general, when applying $T_1$, the analytical form of the resulting distribution is unknown. In this application study instead, given the three presented scenarios, we can easily derive the respective distributions of variables $\tilde{\mathbf{y}}_{\mathrm{c/k}}^{j}$ and $\mathbf{y}_{\mathrm{c/k}}$, which are listed in Table 2.1. For scenarios 1 and 2 the distribution of $\mathbf{y}_{\mathrm{k}}$ belongs to the same class of distributions as assumed for the original concentrations $\mathbf{x}_{\mathrm{k}}$, namely normal and log-normal distributions. Mean and variance are sufficient statistics for these distributions. In scenario 3, $\mathbf{y}_{\mathrm{k}}$ is described by the ratio distribution of two normal random variables, which is called Gaussian ratio (GR) distribution. This family of distributions does not belong to the family of exponential distributions (for details on this class see Gelman et al. (2004)), and can show some peculiarities that will be recapitulated in Section 2.3.

### 2.2.2 Maximum Likelihood estimates

We consider the inverse problem of the inference of the parameters $\theta = (\mu_k, \mu_c, \sigma_k, \sigma_c, \rho)$ of the underlying distribution $p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k})$ from a set of observations $\{y_\mathrm{k}^i\}_{i=1,\dots,N}$, for each of the three scenarios, see last row of Table 2.1. Our final goal is then to estimate the expectation value $\mathbb{E}(\mathbf{y}_\mathrm{k})$ as a measure for the KD fold change. Therefore, for each assumed error model we calculate the ML estimate $\hat{\theta}$ that maximizes the conditional probability $p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}^i|\theta)$ of the observed data $y_\mathrm{k}^i$. Assuming independence across different replicates, this reads:

$$\begin{aligned} \hat{\theta} = \underset{\theta\in\Theta}{\arg\max}\mathcal{L}(\theta) &= \underset{\theta\in\Theta}{\arg\max} \textstyle\prod_{i=1}^{N} p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}^i|\theta) \\ &= \underset{\theta\in\Theta}{\arg\min} \textstyle\sum_{i=1}^{N} -\log p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}^i|\theta). \end{aligned} \tag{2.2.6}$$

The set $\Theta \subseteq \mathbb{R}^M$ represents a subset of the parameter space, in which the estimated value is assumed to lie. For completeness, basic principles along with the properties of MLE are given in Appendix 6.1.

Under some regularity conditions, the ML estimator is often a good estimator in the sense that it is consistent, i.e. it converges in probability to the true parameters $\theta^*$ as the number of samples increases, $\hat{\theta} \xrightarrow{p} \theta^*$. This implies that $\hat{\theta}$ is asymptotically unbiased. Furthermore, it is asymptotically normal, i.e. $\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta^*)^{-1})$ with Fisher information matrix $I(\theta^*)$, which guarantees that it converges fast enough (with a rate $1/\sqrt{N}$). And last, the MLE is asymptotically efficient, meaning that $\hat{\theta}$ achieves the minimum possible variance among all unbiased estimators, or the Cramér-Rao lower bound, for large sample sizes, making it a precise estimator. These properties of ML estimators date back to Fisher (Fisher, 1922), and can be found in any statistics textbooks (see e.g. Gelman et al. (2004)).

Regularity conditions for consistency are smoothness of the likelihood function, identifiability of $\hat{\theta}$, and existence of the mean value $\mathbb{E}_{\theta^*} \log p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}|\theta)$. Furthermore, $\hat{\theta}$ must not lie on the boundary of the defined domain $\Theta$. In addition, for asymptotic normality $\mathrm{Var}_{\theta^*} \log p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}|\theta)$ has to exist. If these conditions are satisfied, then the distribution of $\hat{\theta}$ is for large sample sizes approximately normal with a small variance.

#### MLE for scenario 1

For scenario 1 we have to estimate:

$$\theta_1 = (\mu_1, \sigma_1^2) := \left(\frac{\mu_k}{\mu_c}, \left(\frac{\sigma_k}{\mu_c}\right)^2\right) \tag{2.2.7}$$

with

$$p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}^i|\theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{1}{2}\left(\frac{y_\mathrm{k}^i - \mu_1}{\sigma_1}\right)^2\right]. \tag{2.2.8}$$

This is a well-known problem (see e.g. Whittaker (1990)), and the solution is given by the sample mean and variance:

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} y_\mathrm{k}^i \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(y_\mathrm{k}^i - \hat{\mu}_1)^2. \tag{2.2.9}$$

### MLE for scenario 2

In scenario 2 the log-transformed data are normally distributed, and we have to estimate:

$$\theta_2 = (\mu_2, \sigma_2^2) := \left(\mu_k - \mu_c, \sigma_k^2 + \sigma_c^2 - 2\rho\sigma_k\sigma_c\right) \tag{2.2.10}$$

with

$$p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k}^i|\theta_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{1}{2}\left(\frac{\log y_\mathrm{k}^i - \mu_2}{\sigma_2}\right)^2\right]. \tag{2.2.11}$$

Accordingly,

$$\hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N} \log y_\mathrm{k}^i \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{1}{N}\sum_{i=1}^{N}(\log y_\mathrm{k}^i - \hat{\mu}_2)^2. \tag{2.2.12}$$

As can be seen from the equations (2.2.7), (2.2.9) and (2.2.10), (2.2.12), the individual original parameters $\theta = (\mu_k, \mu_c, \sigma_k, \sigma_c, \rho)$ are not identifiable for scenarios 1 and 2. In fact, there exist infinite combinations of values of the original parameters $\mu_k, \mu_c, \sigma_k, \sigma_c$ and $\rho$ corresponding to the calculated parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2$ and $\hat{\sigma}_2^2$, respectively. In particular, we are not able to distinguish between the two cases 2A ($\rho = 0$) and 2B ($\rho \neq 0$) and for this reason we will refer in the following to scenario 2 only. Here, it should be noted that $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ underestimate the variance in particular for small sample sizes. These estimates could be corrected by multiplication with the factor $^N/_{N-1}$, which corresponds to an unbiased variance estimator.

### MLE for scenario 3

The parameter vector that characterizes the Gaussian ratio distribution consists of four (scenario 3A) or five (scenario 3B) variables, respectively,

$$\theta_3 = (\mu_k, \mu_c, \sigma_k, \sigma_c, \rho). \tag{2.2.13}$$

The probability $p_{\mathbf{y}_k}(y_k^i|\theta_3)$ is given as:

$$p_{\mathbf{y}_k}(y_k^i|\theta_3) = \frac{b(y_k^i)d(y_k^i)}{a^3(y_k^i)\sqrt{2\pi}\sigma_k\sigma_c}\mathrm{erf}\left(\frac{b(y_k^i)}{\sqrt{2}\sqrt{(1-\rho^2)}a(y_k^i)}\right) + \frac{\sqrt{(1-\rho^2)}}{a^2(y_k^i)\cdot\pi\sigma_k\sigma_c}\exp\left(-\frac{c}{2(1-\rho^2)}\right),$$

where the terms $a(z), b(z), c, d(z)$ and $\mathrm{erf}(z)$ will be defined in equation (2.3.21), with $z$ equivalent to $y_k^i$. In the following Section 2.3 we will present the main statistical properties of this class of distributions. Unlike the previous two scenarios, there is no analytical expression for the ML estimate $\hat{\theta}_3$ in this case. The negative-log likelihood function can therefore be minimized numerically by multi-start local optimization, which is done here with a Latin hypercube sampling of the parameter space.

## 2.3 The Gaussian ratio distribution: statistical properties, structural identifiability and convergence properties of ML estimators

In this section we recapitulate some important properties of the Gaussian ratio distributions, we investigate the structural identifiability of its parametrization and analyse convergence properties of its maximum likelihood estimator by means of an *in silico* study.

### 2.3.1 Statistical properties of ratio distributions of normal random variables

Ratio distributions $\mathbf{z} = \mathbf{x}/\mathbf{y}$ of two normal RVs $\mathbf{x} \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathbf{y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and correlation $\rho$ have intensively been studies in the 60's and 70's, see e.g. Hayya et al. (1975); Hinkley (1969); Marsaglia (1965), with some refined work described in Marsaglia (2006). It was shown in Marsaglia (1965, 2006) that after rescaling and translation, the distribution of $\tilde{\mathbf{z}} = r(\mathbf{z} - s)$ equals that of $\mathbf{t} = \dfrac{a + \mathbf{v}}{b + \mathbf{w}}$, with $\mathbf{v}, \mathbf{w}$ independently standard normally distributed and $r, s, a, b$ defined as:

$$\begin{aligned} r &= \pm\frac{\sigma_Y}{\sigma_X\sqrt{1-\rho^2}} & s &= \frac{\rho\sigma_X}{\sigma_Y} \\ a &= \pm\frac{\mu_X/\sigma_X - \rho\mu_Y/\sigma_Y}{\sqrt{1-\rho^2}} & b &= \frac{\mu_Y}{\sigma_Y}. \end{aligned} \tag{2.3.14}$$

The sign '+' or '–' in the equations for $a$ and $r$ has to be chosen equally such that $a$ has the same sign of $b$, in order to ensure that $p_{\mathbf{t}}(t)$ is a proper density. In particular, we will focus on the case in which $a$ and $b$ are both non-negative. In fact we want to describe

protein concentrations with assumed positive mean values, for which reason we assume that $\mu_Y \geq 0$ and therefore $b$ has to be non-negative.

The uncorrelated case $\rho = 0$ corresponds to $s = 0$ and the accordingly simplified equations:

$$\begin{aligned} r &= \frac{\sigma_Y}{\sigma_X} \qquad & s &= 0 \\ a &= \frac{\mu_X}{\sigma_X} \qquad & b &= \frac{\mu_Y}{\sigma_Y}. \end{aligned} \tag{2.3.15}$$

In this case, the sign for both parameters $a$ and $r$ has been set to '+', since we assume that $\mu_X \geq 0$.

The density of $\mathbf{t}$ is given by:

$$p_{\mathbf{t}}(t) = \frac{e^{-\frac{1}{2}(a^2+b^2)}}{\pi(1+t^2)} \left[1 + qe^{\frac{1}{2}q^2} \int_0^q e^{-\frac{1}{2}x^2} dx\right], \tag{2.3.16}$$

with

$$q(t) = \frac{b + at}{\sqrt{1+t^2}}. \tag{2.3.17}$$

The integral contained in $p_{\mathbf{t}}(t)$ makes it slightly difficult to infer properties of the distribution directly via simple analysis. However, we can recognize some properties from equation (2.3.16). First, it can be expressed as a convex combination of the Cauchy density function (defined as $p_1(t)$) and another density ($p_2(t)$), which describes always a bimodal distribution:

$$p_{\mathbf{t}}(t) = fp_1(t) + (1-f)p_2(t), \quad f = e^{-\frac{1}{2}(a^2+b^2)} \tag{2.3.18}$$

with

$$p_1(t) = \frac{1}{\pi(1+t^2)}, \quad p_2(t) = \frac{q\int_0^q e^{-\frac{1}{2}(x^2-q^2)}dx}{\pi(1+t^2)(e^{\frac{1}{2}(a^2+b^2)}-1)}. \tag{2.3.19}$$

The Cauchy density $p_1(t)$ is independent of $a$ and $b$, while proportions $f$ and $1-f$, as well as $p_2(t)$, are functions of $a$ and $b$. The modality of the mixture distribution $p_{\mathbf{t}}(t)$ depends then on $a$ and $b$. Figure 4 of Marsaglia (2006) depicts the regions in the $(a, b)$-plane, where the density function $p_{\mathbf{t}}(t)$ is uni- or bimodal. In particular, from there it can be seen that, if the signal to noise ratio $\mu_Y/\sigma_Y$ of the variable in the denominator (represented by the parameter $b$) is large enough, say 10, then $\tilde{a} = 2.256$ serves approximately as a threshold for these two regions: $p_{\mathbf{t}}(t)$ is unimodal for $a < \tilde{a}$ and bimodal otherwise, although the second mode is often so small and quite distant from the first that it is insignificant in practice (Marsaglia, 2006).

The Cauchy distribution and also the mixture distribution $p_{\mathbf{t}}(t)$ belong to the class of heavy tailed distributions, since their tails decay slower than exponentially. As a consequence of this heavy-tailedness, the integrals $\int_{-\infty}^{\infty} t^i p_{\mathbf{t}}(t)dt$, for some $i \in \{1, 2, \ldots\}$, are infinite and hence moments do not exist. The degree of heavy-tailedness is determined

by the significance of the Cauchy component $p_1(t)$ of $p_{\mathbf{t}}(t)$ (Pham-Gia et al., 2006).

These facts hold also for $p_{\mathbf{z}}(z)$, since $\mathbf{t}$ and $\mathbf{z}$ are linearly related. Backtransformation of the distribution of $\mathbf{t}$, via scaling and translation,

$$\mathbf{z} = \frac{1}{r}\mathbf{t} + s, \tag{2.3.20}$$

gives the ratio distribution of $\mathbf{z}$ in terms of $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$, (see Hinkley (1969) for details):

$$\begin{aligned} p_{\mathbf{z}}(z) = \; & \frac{b(z)d(z)}{a^3(z)\sqrt{2\pi}\sigma_X\sigma_Y}\operatorname{erf}\left(\frac{b(z)}{\sqrt{2}\sqrt{(1-\rho^2)}a(z)}\right) \\ & + \frac{\sqrt{(1-\rho^2)}}{a^2(z)\cdot\pi\sigma_X\sigma_Y}\exp\left(-\frac{c}{2(1-\rho^2)}\right) \end{aligned} \tag{2.3.21}$$

with

$$\begin{aligned} a(z) &= \sqrt{\frac{1}{\sigma_X^2}z^2 + \frac{1}{\sigma_Y^2} - \frac{2\rho z}{\sigma_X\sigma_Y}} \\ b(z) &= \frac{\mu_X}{\sigma_X^2}z + \frac{\mu_Y}{\sigma_Y^2} - \frac{\rho(\mu_X + \mu_Y z)}{\sigma_X\sigma_Y} \\ c &= \frac{\mu_X^2}{\sigma_X^2} + \frac{\mu_Y^2}{\sigma_Y^2} - \frac{2\rho\mu_X\mu_Y}{\sigma_X\sigma_Y} \\ d(z) &= \exp\left(\frac{b^2(z) - ca^2(z)}{2(1-\rho^2)a^2(z)}\right) \\ \operatorname{erf}(z) &= \frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}\,du. \end{aligned}$$

Heavy-tailedness has several consequences for inference. The Central Limit Theorem does not apply and the sample mean is not necessarily a consistent estimator for the mean of the distribution. For example, the sample mean of Cauchy-distributed RVs is Cauchy-distributed, and hence the variance does not decrease with sample size. It has been shown in Caginalp and Caginalp (2017) that the tails of a GR distribution decrease proportional to $x^{-2}$, with a proportionality factor that depends on the parameters of the distribution. For some applications it is useful and reasonable to approximate a GR distribution with a normal distribution, which is under some conditions a good approximation around the distribution's mode, as detailed in Díaz-Francés and Rubio (2013); Shanmugalingam (1982).

Of particular interest for this study is the assumption that the CV of $\mathbf{y}$, $\sigma_Y/\mu_Y$ is small, or, equivalently, the signal to noise ratio $b = \mu_Y/\sigma_Y$ is large and $\mathbf{y} > 0$. The parameter $b$ determines the probability for $\mathbf{y}$ to be negative, $p_{\mathbf{y}}(\mathbf{y} < 0)$, and plays the role of a shape parameter of $p_{\mathbf{z}}(z)$ (Díaz-Francés and Rubio, 2013). If $b$ is large, it is reasonable to use a normal distribution for the denominator of $\mathbf{t}$ that is truncated at 0 from the left. In this case, mean and variance of $\mathbf{t}$ and $\mathbf{z}$ are well-defined (Hayya et al. (1975); Hinkley (1969); Pham-Gia et al. (2006), see also Marsaglia (2006) for examples). Approximate formulas

for mean and variance are obtained by a second-order Taylor series expansion of those quantities, leading to (Hayya et al., 1975):

$$\mathbb{E}(\mathbf{z}) \approx \tilde{\mathbb{E}}(\mathbf{z}) = \frac{\mu_X}{\mu_Y} + \frac{\sigma_Y^2 \mu_X}{\mu_Y^3} - \rho \frac{\sigma_X \sigma_Y}{\mu_Y^2} \tag{2.3.22}$$

and

$$\mathrm{Var}(\mathbf{z}) \approx \tilde{\mathrm{V}}\mathrm{ar}(\mathbf{z}) = \frac{\sigma_Y^2 \mu_X^2}{\mu_Y^4} + \frac{\sigma_X^2}{\mu_Y^2} - 2\rho \frac{\sigma_X \sigma_Y \mu_X}{\mu_Y^3}. \tag{2.3.23}$$

In the case $\rho = 0$ (scenario 3A) these formulas are simplified accordingly:

$$\mathbb{E}(\mathbf{z}) \approx \tilde{\mathbb{E}}(\mathbf{z}) = \frac{\mu_X}{\mu_Y} + \frac{\sigma_Y^2 \mu_X}{\mu_Y^3} \tag{2.3.24}$$

and

$$\mathrm{Var}(\mathbf{z}) \approx \tilde{\mathrm{V}}\mathrm{ar}(\mathbf{z}) = \frac{\sigma_Y^2 \mu_X^2}{\mu_Y^4} + \frac{\sigma_X^2}{\mu_Y^2}. \tag{2.3.25}$$

Additionally, for not too large values of the parameter $a$, the distribution $p_{\mathbf{t}}(t)$ itself becomes quite similar to a normal distribution. A practical rule, described in Marsaglia (2006), states that $\mathbf{z}$ can be approximated with a normal distribution if $a < 2.256$ and $b > 4$. Similar rules of thumb can be found in Hayya et al. (1975). All those agree in the condition that the CV of $\mathbf{y}$ has to be sufficiently small. We note here that the normal approximation might even be appropriate for parameters $a$ and $b$ for which $p_{\mathbf{t}}(t)$ falls into the bimodal region.

The RV $\mathbf{z}$, as presented in this subsection, represents then $\mathbf{y}_\mathrm{k}$ in scenario 3, as given in the last column of Table 2.1, in which case the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ are equivalent to $\mu_k, \mu_c, \sigma_k, \sigma_c$ and $\rho$.

### 2.3.2 Structural non-identifiability of GR distributions

In the following, we analyse structural identifiability of the parameter vector $\theta_3$ of the GR error model, corresponding to scenario 3. As described in the problem formulation (Section 2.2), the final goal of our statistical investigation is to estimate the expected value $\mathbb{E}(\mathbf{y}_\mathrm{k})$ from experimental observations, for all three scenarios, related to the fold change estimation of knockdown experiments. In particular for scenario 3, assuming that the conditions for the existence of mean and variance are valid, we aim to obtain an estimate for the approximating value $\tilde{\mathbb{E}}(\mathbf{y}_\mathrm{k})$ of $\mathbb{E}(\mathbf{y}_\mathrm{k})$, given in equation (2.3.22). For this purpose, we shall at first estimate the five original parameters $\mu_k, \mu_c, \sigma_k, \sigma_c, \rho$, or accordingly four in the uncorrelated case. As already mentioned in Section 2.2, in this case, unlike scenarios 1 and 2, we do not have analytical expressions for the ML estimator. We should therefore rely on numerical optimization to obtain an ML estimate for the full parameter vector.

From the theory, we demonstrate that the full inference problem with $\theta = (\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$

constitutes an ill-posed inverse problem with *structurally non-identifiable* parameters. We can, in fact, rewrite $p_{\mathbf{z}}(z)$ as a function of the equivalent parametrization $a, b, r, s$, given in the equations (2.3.14), by transformation of probability density functions. Given the monotone relationship between the RVs $\mathbf{z}$ and $\mathbf{t}$:

$$\mathbf{z} = f(\mathbf{t}) = \frac{1}{r}\mathbf{t} + s,$$

we get:

$$\begin{aligned} p_{\mathbf{z}}(z) &= p_{\mathbf{t}}(t)\big|_{t=f^{-1}(z)} \cdot \left|\frac{df^{-1}(z)}{dz}\right| = p_{\mathbf{t}}(r(z-s)) \cdot |r| \\ &= \frac{e^{-\frac{1}{2}(a^2+b^2)}}{\pi(1+t^2)} \left[1 + qe^{\frac{1}{2}q^2} \int_0^q e^{-\frac{1}{2}x^2} dx\right]\Big|_{t=r(z-s)} \cdot |r| \,. \end{aligned} \tag{2.3.26}$$

The function $q(t)$ is defined in equation (2.3.17). Therefore, given the formula of $p_{\mathbf{t}}(t)$ (equation (2.3.16)), which is parametrized only in $a$ and $b$, we obtain that $p_{\mathbf{z}}(z)$ is parametrized only in the four defined parameters $a, b, r, s$ or, in case of $\rho = 0$, in the three parameters $a, b, r$, as also mentioned in Díaz-Francés and Rubio (2013).

Thus, there exist infinite combinations of values for the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ that correspond to the same values for $a, b, r, s$. For this reason, we will consider this reduced parametrization for MLE. In particular, we can obtain the expressions of the set of equivalent solutions for the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ as functions of the uniquely determined values for $a, b, r, s$ if we set a fixed value for one of the five parameters, for example $\sigma_X$:

$$\begin{aligned} \mu_X &= \frac{a + brs}{\sqrt{r^2 s^2 + 1}} \cdot \sigma_X \\ \mu_Y &= \frac{br}{\sqrt{r^2 s^2 + 1}} \cdot \sigma_X \\ \sigma_Y &= \frac{r}{\sqrt{r^2 s^2 + 1}} \cdot \sigma_X \\ \rho &= \frac{rs}{\sqrt{r^2 s^2 + 1}}. \end{aligned} \tag{2.3.27}$$

We note that $\rho$ is always univocally determined given the estimated values of $r$ and $s$. The other four parameters (means and standard deviations) are linearly dependent among each other. In the uncorrelated case (scenario 3A) we have $s = 0$ and the formulas simplify accordingly:

$$\begin{aligned} \mu_X &= a \cdot \sigma_X \\ \mu_Y &= br \cdot \sigma_X \\ \sigma_Y &= r \cdot \sigma_X. \end{aligned} \tag{2.3.28}$$

From the set of equations (2.3.27) we can, in particular, rewrite the equations (2.3.22) and (2.3.23) as functions of the four identifiable parameters, thus giving the approximations for $\mathbb{E}(\mathbf{z})$ and $\mathrm{Var}(\mathbf{z})$, which are valid for sufficiently large values of the parameter $b$:

$$\tilde{\mathbb{E}}(\mathbf{z}) = \frac{a+brs}{br} + \frac{a+brs}{b^3 r} - \frac{s}{b^2} \tag{2.3.29}$$

$$\tilde{\mathrm{Var}}(\mathbf{z}) = \frac{(a+brs)^2}{b^4 r^2} + \frac{1+r^2 s^2}{b^2 r^2} - 2s\frac{a+brs}{b^3 r}. \tag{2.3.30}$$

Again, the case $\rho = 0$ implies $s = 0$ and the formulas simplify accordingly:

$$\tilde{\mathbb{E}}(\mathbf{z}) = \frac{a}{br} + \frac{a}{b^3 r} \tag{2.3.31}$$

$$\tilde{\mathrm{Var}}(\mathbf{z}) = \frac{a^2}{b^4 r^2} + \frac{1}{b^2 r^2}. \tag{2.3.32}$$

### Implication of non-identifiability for GR distributions

We want to understand the practical implication of *structural non-identifiability* of the GR error model under the following two simplifying assumptions:

**Assumption 1:** Control and perturbation measurements are not correlated and $s^* = 0$. In fact, this is a reasonable assumption for a proper experimental design. Although correlated experimental errors of WB data may be systematically introduced due to gel and transfer inhomogeneities, a reduction of correlated errors in WB experiments can be experimentally obtained by proper randomization of sample loading (Schilling et al., 2005).

**Assumption 2:** In addition, we assume that the two normally distributed random variables have the same standard deviation before normalization, resulting in $r^* = 1$, which is reasonable if measurement errors are the main source of noise.

Consequently, we then get indistinguishable GR distributions if we also keep $a$ and $b$ constant. We can therefore derive that, under these specific conditions, we obtain the same GR distribution when normalizing uncorrelated measured data by scaling the means and standard deviations of both numerator and denominator by the same factor. In other words, due to structural non-identifiability, we cannot distinguish between GR distributions obtained by normalizing samples from different Gaussian distributions having fixed signal to noise ratios (or equivalently coefficient of variation) for numerator and denominator. These constant values are exactly the definition of $a$ and $b$. This fact can be illustrated by choosing two different sets of values for the parameters $\mu_X, \sigma_X, \mu_Y$ and $\sigma_Y$ which produce the same values for $a$ and $b$ of the resulting GR distributions, according to equation (2.3.15) (Figure 2.1).

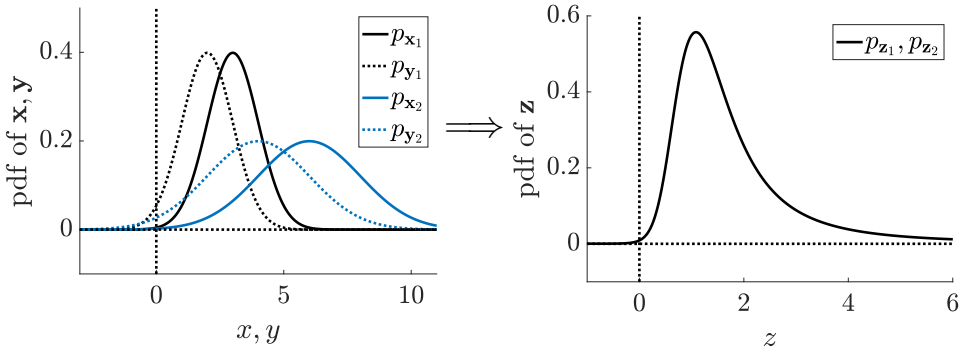


**Fig. 2.1. Structural non-identifiability of $\theta = (\mu_X, \mu_Y, \sigma_X, \sigma_Y)$.** We consider two couples of uncorrelated Gaussian RVs $\mathbf{x}_1, \mathbf{y}_1$ and $\mathbf{x}_2, \mathbf{y}_2$ defining the GR RVs $\mathbf{z}_1 = \mathbf{x}_1/\mathbf{y}_1$ and $\mathbf{z}_2 = \mathbf{x}_2/\mathbf{y}_2$. Distributions $p_{\mathbf{z}_1}(z_1)$ and $p_{\mathbf{z}_2}(z_2)$ are equal for example for $\sigma_{X_1} = \sigma_{Y_1} = 1, \sigma_{X_2} = \sigma_{Y_2} = 2, \mu_{X_1} = 3, \mu_{Y_1} = 2, \mu_{X_2} = 6$ and $\mu_{Y_2} = 4$. This corresponds to $\mathbf{x}_1 \sim \mathcal{N}(3,1), \mathbf{y}_1 \sim \mathcal{N}(2,1)$ (black curves on the left) and $\mathbf{x}_2 \sim \mathcal{N}(6,2^2), \mathbf{y}_2 \sim \mathcal{N}(4,2^2)$ (blue curves on the left). Here, the signal to noise ratios of both numerator and denominator of $\mathbf{z}_1$ and $\mathbf{z}_2$ (parameters $a$ and $b$) are kept constant and both parametrizations correspond to $(a, b, r, s) = (3, 2, 1, 0)$ (black curve on the right).

### 2.3.3 Convergence properties of ML estimators for GR distributions

In the first two parts of this section we discussed some of the statistical properties of GR distributions and investigated the structural identifiability of this error model (see Equation (2.3.26)). However, to the best of our knowledge, the inverse problem of parameter estimation of such distributions from normalized data has not been studied so far. In Morisugu et al. (2009), the authors introduce methodology to evaluate confidence intervals for the estimated approximation of the mean of a GR distribution based on samples of the normally distributed variables of numerator and denominator and corresponding sample estimates of the means, variances and covariance. These calculations are, however, not applicable in our context, since we lack information about the absolute values of the observed data before normalization.

Therefore, the open question now is whether the ML estimate of the reduced parametrization $\theta_3 = (a, b, r, s)$, obtained as solution of the optimization problem from a given set of observations $\{z^i\}_{i=1,\dots,N}$, is reliable. In other words, we want to investigate now what are the statistical properties of the ML estimator (e.g. unbiasedness) and how does it behave asymptotically.

In this subsection we present an *in silico* Monte Carlo study in which we investigate properties of the ML estimator for the parameters of the GR distribution from simulated data. It is not clear whether the conditions for asymptotic convergence of the ML estimator $\hat{\theta}$ are fulfilled in our setting and how the compact set $\Theta$ has to be chosen, if it exists at all. Moreover, $\hat{\theta}$ is not available in analytic form, and we do not know the degree of heavy-tailedness beforehand, which may also cause problems for the estimation of the mean fold change, and probably also for the distribution parameters. Thus, we decided to investigate properties of the ML estimator computationally via simulations. Therefore, we approximated the distribution of the ML estimator via calculating estimators for a large amount ($n = 10{,}000$) of simulated data sets. This was done by numerically minimizing the neg-log likelihood function. Different sizes $N = 10, 100, 1000$ of the data set $\{z^i\}_{i=1,\dots,N}$ are used to visually investigate asymptotic properties of the estimator.

We simulated a KD experiment by using the following parameter values for the two Gaussian distributions $\mathbf{x}$ and $\mathbf{y}$, defining the GR $\mathbf{z} = \mathbf{x}/\mathbf{y}$:

$$\mu_X^* = 18 \quad \sigma_X^* = \sigma_Y^* = 4 \quad \mu_Y^* = 32 \quad \rho^* = 0,$$

where we assume uncorrelated measurements and the same standard deviations for KD and control conditions, as described before. This parametrization corresponds to $(a^*, b^*, r^*, s^*) = (4.5, 8, 1, 0)$ with $\tilde{\mathbb{E}}(\mathbf{z}) = 0.57$. Approximations $\tilde{\mathbb{E}}(\mathbf{z})$ and $\tilde{\mathrm{Var}}(\mathbf{z})$ for mean and variance are valid for this parameter set and estimates $\widehat{\tilde{\mathbb{E}}}(\mathbf{z})$ and $\widehat{\tilde{\mathrm{Var}}}(\mathbf{z})$ are calculated by inserting $\hat{a}, \hat{b}, \hat{r}$

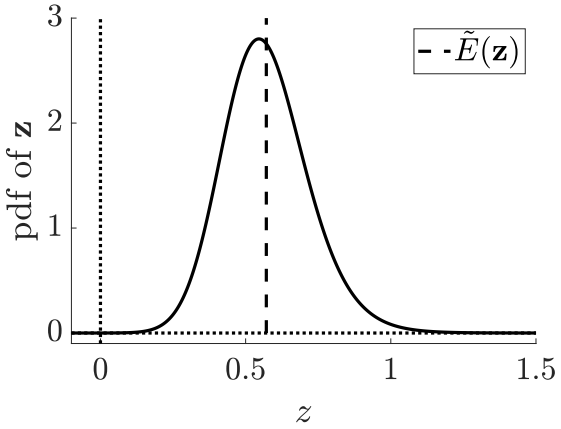


**Fig. 2.2. GR distribution describing the statistics of a KD experimental scenario used for the Monte Carlo study.** The parameters characterizing the distribution are $a^* = 4.5, b^* = 8, r^* = 1$ and $s^* = 0$. The vertical dashed line represents the approximation of the expected value of the distribution $\tilde{\mathbb{E}}(\mathbf{z})$ (equation (2.3.29)).

and $\hat{s}$ into equations (2.3.29) and (2.3.30), i.e.

$$\widehat{\tilde{\mathbb{E}}}(\mathbf{z}) = \mathbb{E}_{MLE} = \frac{\hat{a} + \hat{b}\hat{r}\hat{s}}{\hat{b}\hat{r}} + \frac{\hat{a} + \hat{b}\hat{r}\hat{s}}{\hat{b}^3\hat{r}} - \frac{\hat{s}}{\hat{b}^2} = \frac{\hat{a}}{\hat{b}\hat{r}}\left(1 + \frac{1}{\hat{b}^2}\right) + \hat{s} \tag{2.3.33}$$

and

$$\widehat{\tilde{\mathrm{Var}}}(\mathbf{z}) = \frac{(\hat{a} + \hat{b}\hat{r}\hat{s})^2}{\hat{b}^4\hat{r}^2} + \frac{1 + \hat{r}^2\hat{s}^2}{\hat{b}^2\hat{r}^2} - 2\hat{s}\frac{\hat{a} + \hat{b}\hat{r}\hat{s}}{\hat{b}^3\hat{r}}. \tag{2.3.34}$$

The GR pdf of the Monte Carlo study is represented in Figure 2.2. Of note is that this distribution is bimodal (because $a^* > 2.256$), but the second mode can be neglected.

Our assumptions $r^* = 1$ and $s^* = 0$ were included as prior information into the estimation procedure. 2D scatter plots and 1D marginals are shown in Figure 2.3. Different $N$ are indicated by different color shadings. Axes ranges correspond to boundaries that have been set for the optimization. The asymptotic theory seems to apply here, as the estimate appears to be consistent: For large $N$ the distribution mass concentrates about the true parameter value. Moreover, scatter plots could approximately be described by bivariate normal distributions. The parameters $\hat{a}$ and $\hat{b}$ are strongly positively correlated, but $\mathbb{E}_{MLE}$ can relatively accurately be estimated already for $N = 10$ data points.

Next we pose the question whether estimation is still possible if our assumptions are not used a priori. This was investigated by including first the parameter $r$ in the optimization problem, and in a second step additionally also the parameter $s$.

Respective 2D scatter plots and 1D marginals are shown in Figure 2.4. For $N = 10$ boundary effects for the chosen optimization range appear for the parameters $\hat{a}$, $\hat{b}$ and, to a small extend, also $\hat{r}$. The effect is much smaller but still present for $N = 100$. The estimator converges much slower but seems to be in the asymptotic limit for $N = 1{,}000$. Compared to Figure 2.3, correlations have changed. We can see a strong positive correlation

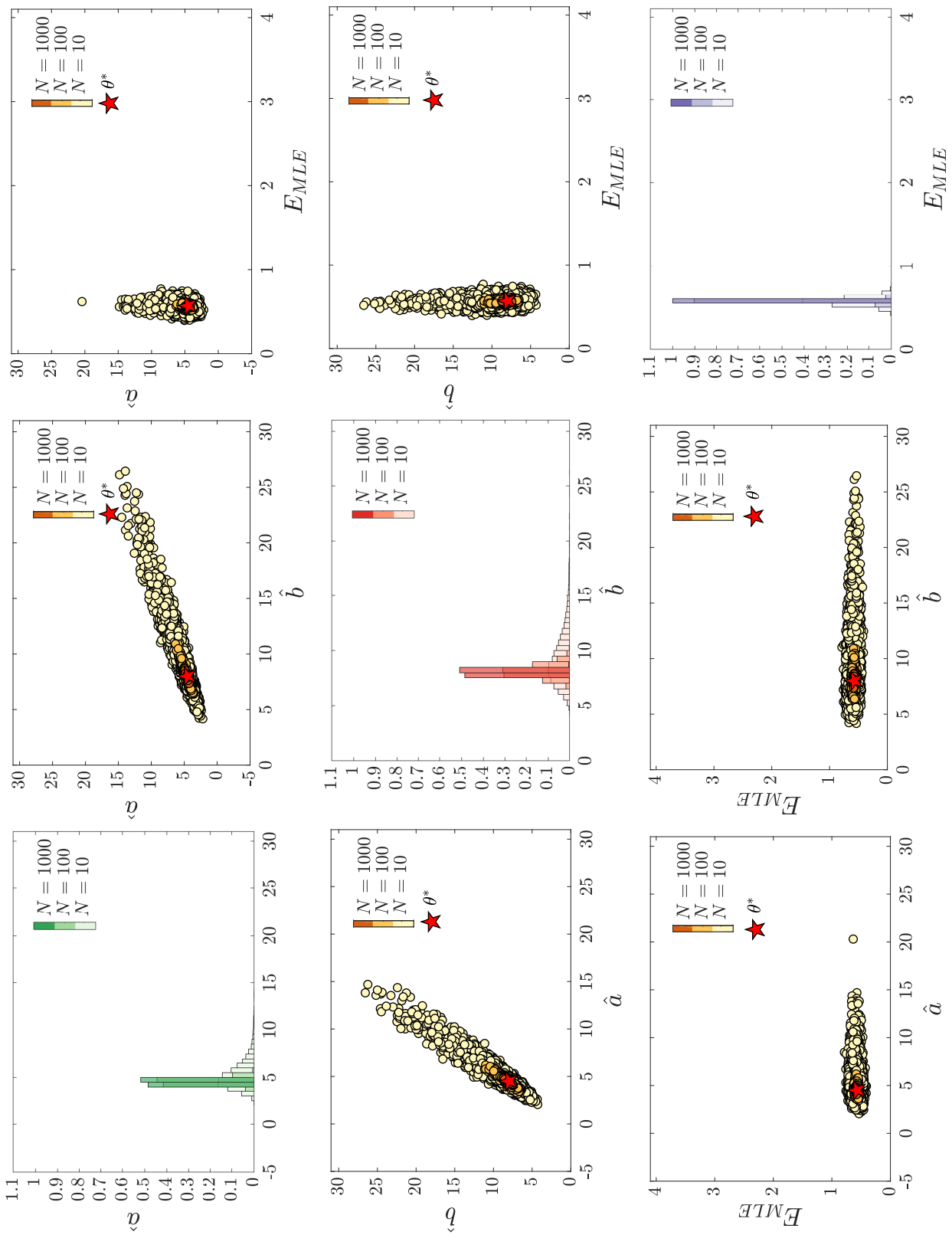


**Fig. 2.3. Effects of increasing size of the data set $N$ on the distribution of the ML estimates of the GR parametrization with $r = r^* = 1$ and $s = s^* = 0$.** 2D scatter plots and 1D marginal distributions of the parameters $\hat{a}, \hat{b}$ and $\mathbb{E}_{MLE}$ obtained from the MC study. The true parameter values $\theta^*$ is marked as red star in the 2D scatter plots.

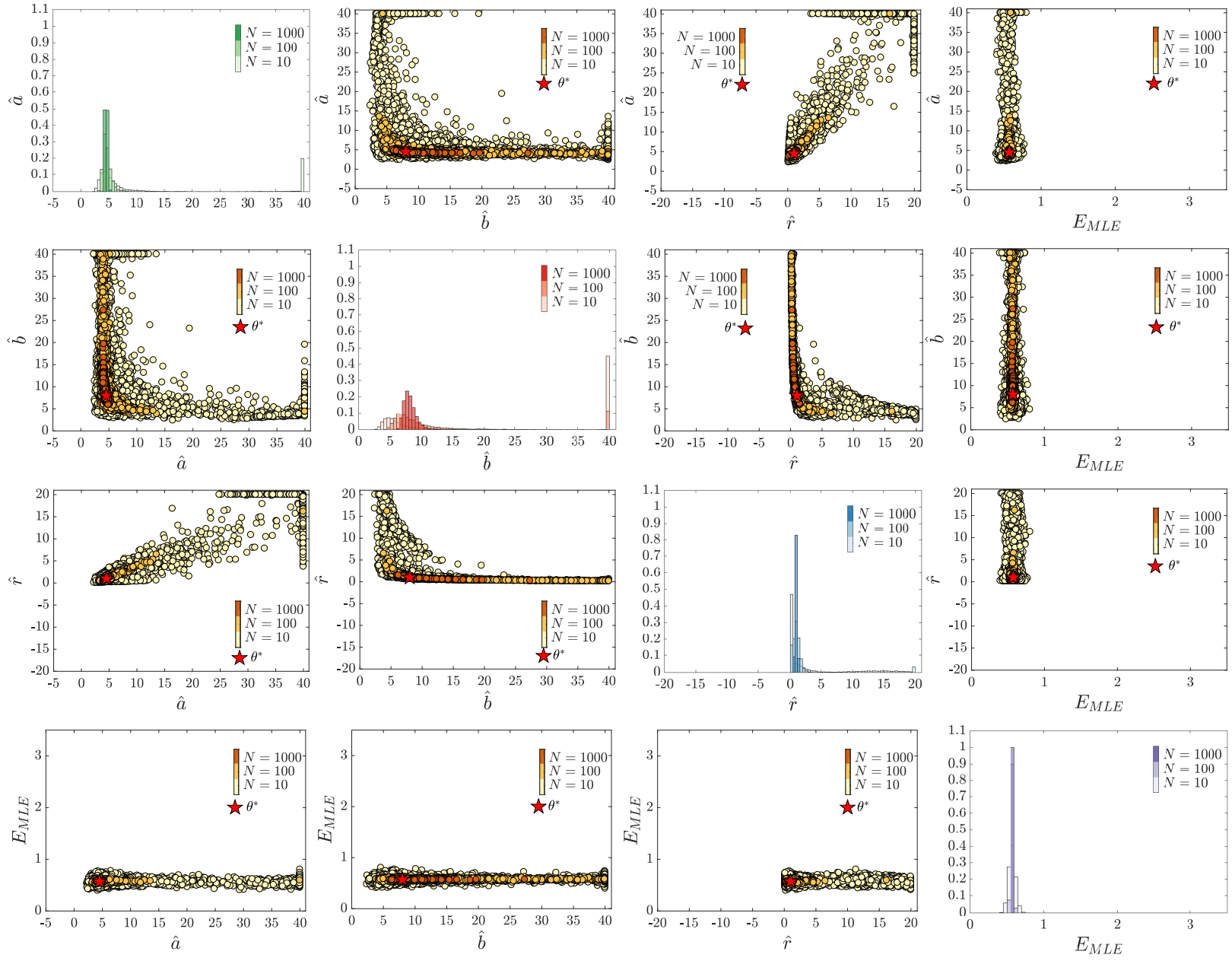


**Fig. 2.4. Effects of increasing size of the data set $N$ on the distribution of the ML estimates of the GR parametrization with $s = s^* = 0$.** 2D scatter plots and 1D marginal distributions of the parameters $\hat{a}, \hat{b}, \hat{r}$ and $\mathbb{E}_{MLE}$ obtained from the MC study. The true parameter values $\theta^*$ is marked as red star in the 2D scatter plots. Sign conditions $\hat{a}, \hat{b} > 0$ are realized by sign changes for $\hat{a}$ and $\hat{r}$ after optimization where necessary (Marsaglia, 2006).

between $\hat{a}$ and $\hat{r}$, while $\hat{a}$ and $\hat{b}$ are negatively correlated. Furthermore, the product $\hat{b}\hat{r}$ could be approximated by a constant. We observe no significant changes in the distribution of $\mathbb{E}_{MLE}$.

Finally, Figure 2.5 shows 2D scatter plots and marginal distributions when the full set of parameters is estimated. For $N = 10$ some of the estimates accumulate at the boundaries, which is the case for all four parameters $\hat{a}, \hat{b}, \hat{r}$ and $\hat{s}$ and most pronounced for the parameter $\hat{r}$, which accumulates at the lower and the upper boundaries. For $N = 100$ these boundary effects decrease, but are still present for the parameters $\hat{a}$ and $\hat{b}$ at the upper boundary. Only for $N = 1{,}000$ the boundary effects have completely vanished, and the estimates start concentrating at the correct values. However, also in this case the distributions are far from being approximately normal. Distributions of $\hat{a}$ and $\hat{s}$ are bimodal, which is in fact often caused by boundary effects and correlation between parameters. In particular, while the first mode of the distribution of $\hat{a}$ appears at $a^*$, the second mode is located at a quite high value. The parameters $\hat{b}$ and $\hat{r}$ show long tails to the right. Moreover, all four parameters seem to be correlated, $\hat{a}$ and $\hat{b}$ positively, while $\hat{a}$ and $\hat{s}$ as well as $\hat{b}$ and $\hat{s}$ negatively, and the products $\hat{b}\hat{r}$ and $\hat{r}\hat{s}$ appear to be constant. Another peculiarity of these samples is the symmetry with respect to the sign of $\hat{r}$ and $\hat{s}$, which appears for $N = 10$ and $N = 100$ and which is broken for $N = 1{,}000$. Only the estimated expectation value $\mathbb{E}_{MLE}$ is well identifiable and in the asymptotic limit.

Despite the large differences in uncertainties of the distributions for the parameters of the GR distribution in the three different settings, the interquartile range of $\mathbb{E}_{MLE}$ is very similar in all three settings (Figure 2.6a-c). The three scenarios only differ in the number and spread of the outliers. As a comparison, we have also plotted the distributions of the sample means (Figure 2.6d),

$$\hat{\mathbb{E}}_{sm}(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} z_i, \tag{2.3.35}$$

whose uncertainty decreases with increasing sample size and looks similar to Figure 2.6a, indicating that the degree of heavy-tailedness is much smaller than that of the Cauchy distribution. This is also supported by the interquartile ranges, which are $\mathrm{IQR}_{|N=10} = 0.0629$, $\mathrm{IQR}_{|N=100} = 0.0198$ and $\mathrm{IQR}_{|N=1000} = 0.0064$ and thus decrease almost with $1/\sqrt{N}$.

Figure 2.7 shows the $n = 10{,}000$ inferred pdfs in the case of estimating the full parametrization from $N = 1{,}000$ data points. The still high variances in the distribution parameters are partly reflected in the percentiles of the pdf, especially about the distribution peak.

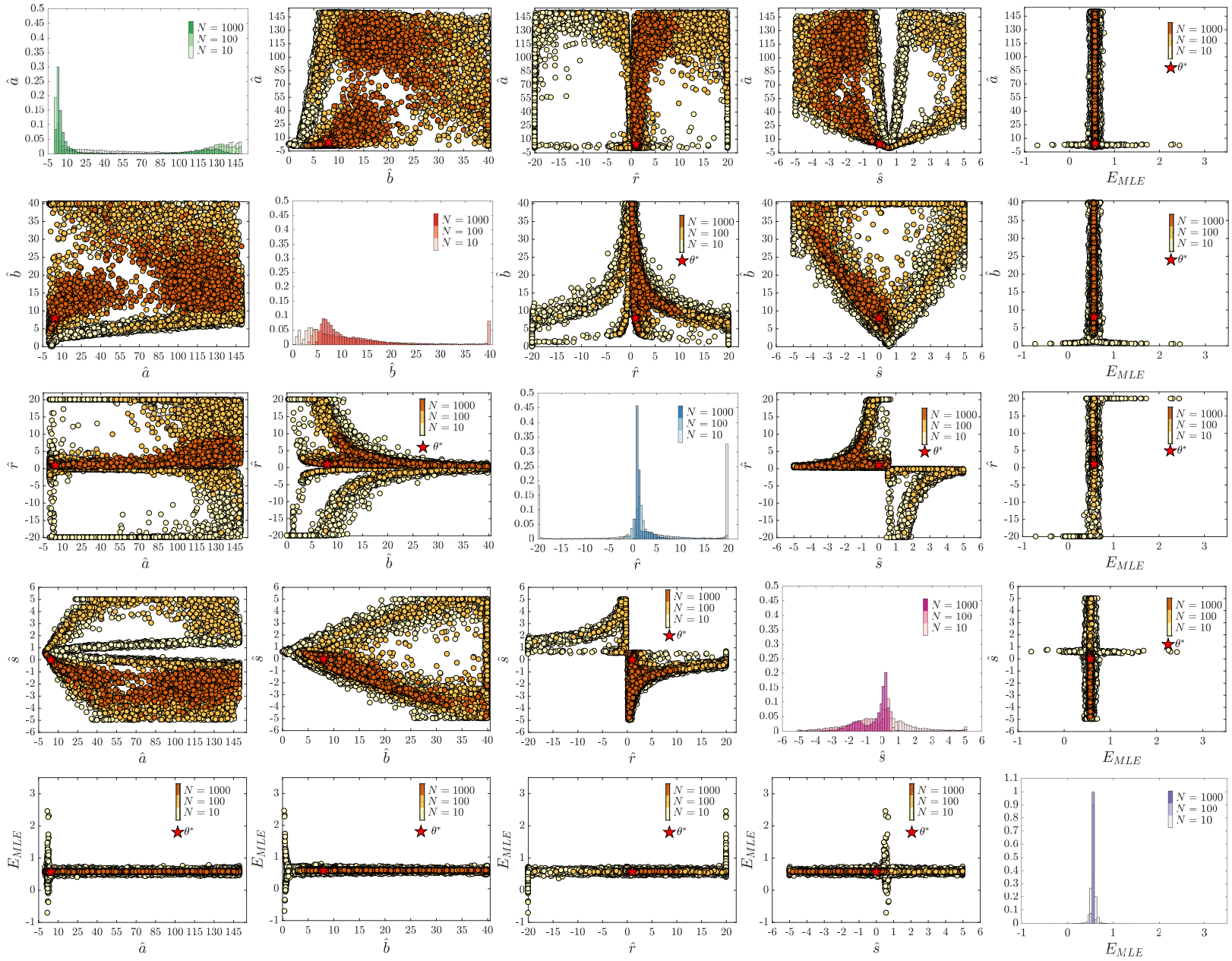


**Fig. 2.5. Effects of increasing size of the data set $N$ on the distribution of the ML estimates of the GR parametrization.** 2D scatter plots and 1D marginal distributions of the parameters $\hat{a}, \hat{b}, \hat{r}, \hat{s}$ and $\mathbb{E}_{MLE}$ obtained from the MC study. The true parameter values $\theta^*$ is marked as red star in the 2D scatter plots. Sign conditions $\hat{a}, \hat{b} > 0$ are realized by sign changes for $\hat{a}$ and $\hat{r}$ after optimization where necessary (Marsaglia, 2006).

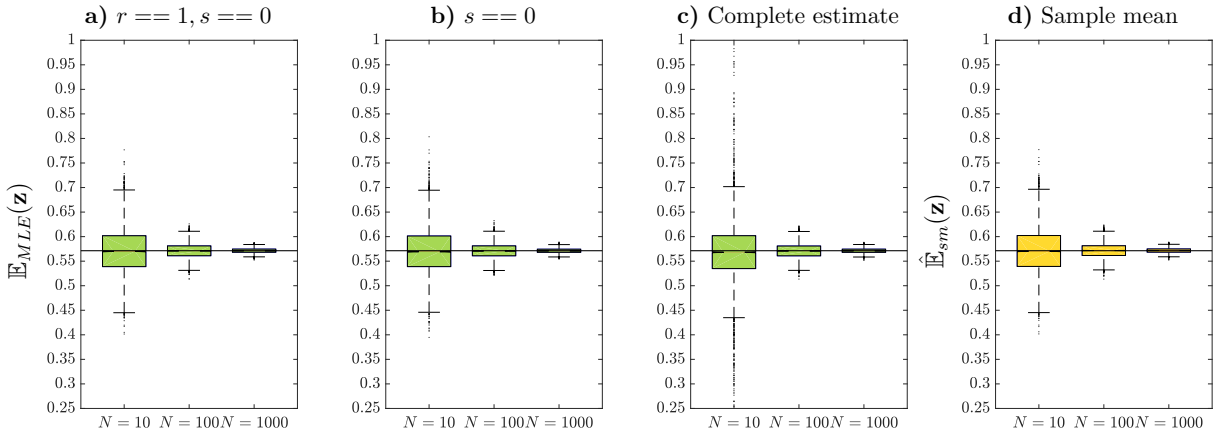


**Fig. 2.6. Effects of increasing size of the data set $N$ on estimates of the expectation value.** Boxplots of $\mathbb{E}_{MLE}$ for the three estimation settings: **(a)** Estimation of $a$ and $b$ only ($s^* = 0, r^* = 1$), **(b)** estimation of $a, b$ and $r$, **(c)** estimation of the full parametrization $a, b, r$ and $s$. **(d)** Boxplots of the sample mean $\hat{\mathbb{E}}_{sm}(\mathbf{z})$ obtained from $N$ i.i.d. samples.

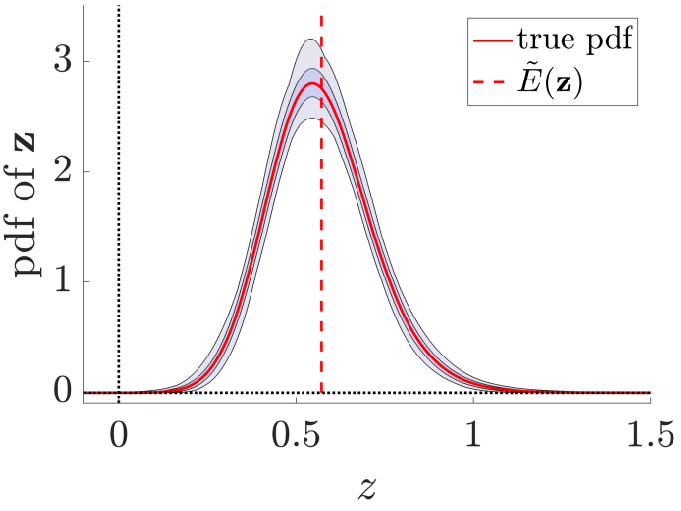


**Fig. 2.7. Percentiles of estimated pdfs for the full parametrization with** $N = 1{,}000$. Pdfs $p_{\mathbf{z}}(z)$ from the in silico study inferred with ML estimation and the full parametrization.

**Table 2.2.** Experimental silencing dataset of WB measurements normalized to the control experiment, taken from Santos et al. (2007).

| Stimulus | Raf | MEK | ERK |
|---|---|---|---|
| EGF | {0.596, 0.599, 0.411, 0.411} | {0.567, 0.398, 0.778, 0.778} | {0.440, 0.417, 0.846, 0.274} |
| NGF5 | {0.534, 0.584, 0.544, 0.590} | {0.615, 0.643, 0.757, 1.024} | {0.375, 0.621, 0.723, 0.409} |
| NGF15 | {0.489, 0.254, 0.670, 0.412} | {0.393, 1.257, 0.622, 0.953} | {0.474, 0.092, 0.234, 0.620} |

In summary, convergence properties of the parameters $a, b, r$ and $s$ of the ratio distribution are very different in all three scenarios. In particular, the asymptotic limit is still not reached for the full parametrization even with $N = 1{,}000$ data points, and a considerable uncertainty is also left in the inferred pdf. In contrast, a robust estimation of the mean of the distribution was possible in all scenarios, with surprisingly similar interquartile ranges. For practical applications we therefore recommend to work with the simplifying assumptions, if they are reasonable, and to set large boundaries for parameter estimation to avoid spurious results caused by boundary effects. Estimation of the mean value seems possible with few data points only and can either be estimated with the estimated distribution parameters or just by using the sample mean.

In the following section we want to apply our theoretical findings on a real case study with measured data of a KD experiment and compare the three statistical error models corresponding to the three statistical scenarios presented in the problem formulation.

## 2.4 Error model selection partially affects the calibrated statistical distributions

In this section we apply our methodology to a real dataset of KD experiments taken from Santos et al. (2007). In the following, all biochemical components will be introduced by means of their acronyms, due to long full names. Please refer to the Notation chapter at the beginning of the thesis for the complete list. The considered study analyses the MAPK signalling pathway in PC12 cell lines. We focus in particular on the silencing experiments, in which the three main proteins of the cascade, Raf, MEK and ERK, were subsequently silenced and concentration fold changes were quantified in each of these silencing experiments after stimulation with EGF and NGF, respectively. The data set comprises 36 measured fold changes in total, listed in Table 2.2.

In particular, we compare fold change estimation by considering the three investigated statistical scenarios (presented in section 2.2.1 and summarized in Table 2.1) and apply MLE to all three error models. As argued in the previous section, for the GR error model

**Table 2.3.** Estimator $\widehat{\mathbb{E}}(\mathbf{y}_\mathrm{k})$ of the expected value for $\mathbf{y}_\mathrm{k}$ for all error model variants.

| | Scenario 1 | Scenario 2 | Scenario 3A |
|---|---|---|---|
| $\widehat{\mathbb{E}}(\mathbf{y}_\mathrm{k})(\hat{\theta})$ | $\hat{\mu}_1$ | $e^{\hat{\mu}_2+\hat{\sigma}_2^2/2}$ | $\frac{\hat{a}}{\hat{b}\hat{r}}+\frac{\hat{a}}{\hat{b}^3\hat{r}}$ |
| Raf EGF | 0.5044 | 0.5044 | 0.5044 |
| MEK EGF | 0.6305 | 0.6318 | 0.6316 |
| ERK EGF | 0.4941 | 0.4925 | 0.4886 |
| Raf NGF5 | 0.5629 | 0.5629 | 0.5629 |
| MEK NGF5 | 0.7599 | 0.7593 | 0.7579 |
| ERK NGF5 | 0.5318 | 0.5319 | 0.5314 |
| Raf NGF15 | 0.4562 | 0.4576 | 0.456 |
| MEK NGF15 | 0.8065 | 0.8107 | 0.8043 |
| ERK NGF15 | 0.3551 | 0.3704 | 0.355 |

we will focus only on the uncorrelated case, assuming $s^* = 0$ and $r^* = 1$, focusing therefore only on scenario 3A.

As described in section 2.2.2, while the ML estimator $\hat{\theta}$ is analytically given for the normal and log-normal error models (equations (2.2.9) and (2.2.12)), it was obtained via numerical optimization of the likelihood function in the GR case, given in equation (2.3.26).

Resulting ML estimates for the respective distributions $p_{\mathbf{y}_\mathrm{k}}(y_\mathrm{k})$ are shown in Figure 2.8. It can be seen that all scenarios lead to nearly identical distributions in cases with small variance across replicates, as it is for example the case for Raf in the NGF experiment at $t^* = 5$ min (first column, second row in Figure 2.8 and consistently in Table 2.2). If the data, however, show larger variance, the distributions of the different scenarios differ slightly more evidently. In these cases it can be seen that the inferred GR distributions lie between the normal and the log-normal distributions.

Estimates for $\mathbb{E}(\mathbf{y}_\mathrm{k})$ can also be extracted from the ML estimators, as listed in Table 2.3. In particular, we recall the fact that the ML estimator of the mean for the normal error model $\hat{\mu}_1$ coincides with the sample mean (equations (2.3.35)). All estimated expected values are reasonable and we obtain almost the same value for all three scenarios. The conditions for the approximation (2.3.29) are fullfilled for scenario 3A, since estimated $\hat{b}$ values are sufficiently large (see Table 2.4, where all MLE values $\hat{\theta}$ are collected).

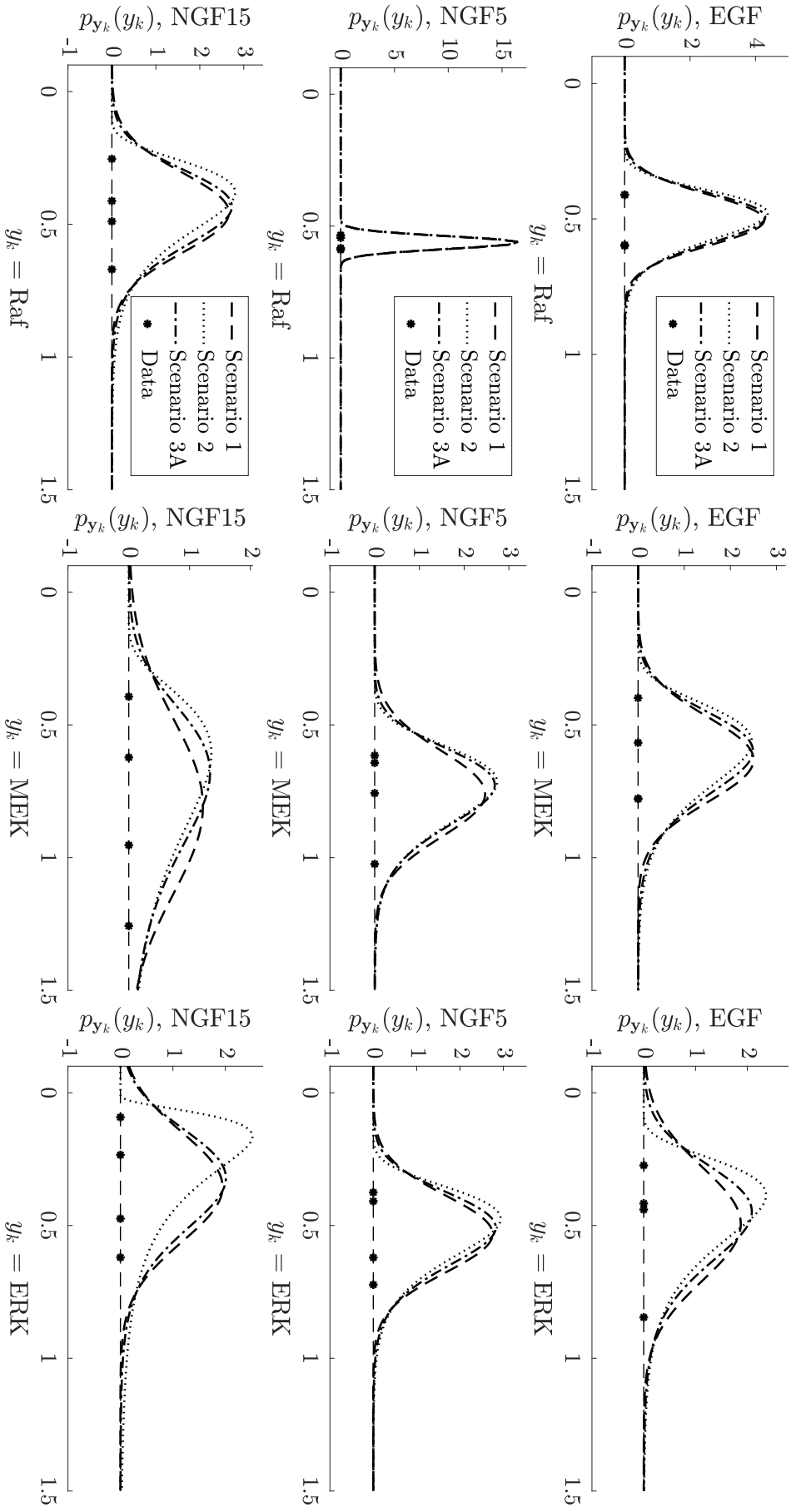


**Fig. 2.8.** ML estimates of pdfs for the entire silencing data set as presented in Santos et al. (2007), listed in Table 2.2.

**Table 2.4.** Full set of estimated parameters for the MAPK application study for all error model variants.

| | Scenario 1 | | Scenario 2 | | Scenario 3A | |
|---|---|---|---|---|---|---|
| | $\hat{\mu}_1$ | $\hat{\sigma}_1$ | $\hat{\mu}_2$ | $\hat{\sigma}_2$ | $\hat{a}$ | $\hat{b}$ |
| Raf EGF | 0.5044 | 0.0934 | -0.7018 | 0.1873 | 6.0237 | 12.0257 |
| MEK EGF | 0.6305 | 0.1593 | -0.4973 | 0.2763 | 4.5069 | 7.2709 |
| ERK EGF | 0.4941 | 0.2129 | -0.7898 | 0.4037 | 2.7183 | 5.7324 |
| Raf NGF5 | 0.5629 | 0.0240 | -0.5756 | 0.0428 | 26.8655 | 47.7490 |
| MEK NGF5 | 0.7599 | 0.1616 | -0.2954 | 0.1999 | 6.2147 | 8.3180 |
| ERK NGF5 | 0.5318 | 0.1450 | -0.6692 | 0.2754 | 4.1522 | 7.9378 |
| Raf NGF15 | 0.4562 | 0.1498 | -0.8435 | 0.3512 | 3.3408 | 7.4574 |
| MEK NGF15 | 0.8065 | 0.3275 | -0.3066 | 0.4398 | 3.0641 | 4.0426 |
| ERK NGF15 | 0.3551 | 0.2050 | -1.2649 | 0.7373 | 1.8126 | 5.2884 |

The only case which is at the limit ($\hat{b}$ slightly larger than 4) refers to the MEK data set for NGF15 stimulation (third row, second column in Figure 2.8 and eighth row in Table 2.4), which is in fact the data set showing the largest uncertainty. Estimates for $\mathrm{Var}(\mathbf{y}_k)$ can be extracted from the ML estimators as well (see Table 2.5), resulting in very similar values among the three error models.

Summarizing, inference with the GR distribution under the two considered simplifying assumptions gives very similar results as the simpler normal model in terms of estimated means and variances and inferred distributions. This indicates that we are in the regime in which the GR distribution can well be approximated with a normal distribution. Furthermore, the results obtained with this real case study indicate that the three considered error models, under plausible simplifying assumptions, can be fairly compared among each other as suitable description of normalized WB data. This is in accordance with other studies which argue that, under some conditions, GR distributions can be approximated by normal or log-normal distributions (Díaz-Francés and Rubio, 2013; Shanmugalingam, 1982).

**Table 2.5.** Estimator $\widehat{\mathrm{Var}}(\mathbf{y}_\mathrm{k})$ of the variance for $\mathbf{y}_\mathrm{k}$ for all error model variants.

| | Scenario 1 | Scenario 2 | Scenario 3A |
|---|---|---|---|
| $\widehat{\mathrm{Var}}(\mathbf{y}_\mathrm{k})(\hat{\theta})$ | $\hat{\sigma}_1^2$ | $e^{2\hat{\mu}_2+\hat{\sigma}_2^2}\left(e^{\hat{\sigma}_2^2}-1\right)$ | $\frac{\hat{a}^2}{\hat{b}^4\hat{r}^2}+\frac{1}{\hat{b}^2\hat{r}^2}$ |
| Raf EGF | 0.0087 | 0.0091 | 0.0086 |
| MEK EGF | 0.0254 | 0.0317 | 0.0262 |
| ERK EGF | 0.0453 | 0.0429 | 0.0373 |
| Raf NGF5 | 0.0006 | 0.0006 | 0.0006 |
| MEK NGF5 | 0.0261 | 0.0235 | 0.0225 |
| ERK NGF5 | 0.0210 | 0.0223 | 0.0202 |
| Raf NGF15 | 0.0224 | 0.0275 | 0.0216 |
| MEK NGF15 | 0.1072 | 0.1402 | 0.0963 |
| ERK NGF15 | 0.0420 | 0.0991 | 0.0400 |

## 2.5 Summary and discussion

In this chapter we analysed different error models for WB data and compared their effect on parameter estimation of the stochastic models underlying relative data generation. We illustrated results on an exemplary case study, in which WB data were used to infer the fold change between the amount of a protein in a knockdown experiment versus the untreated control case. We employed MLE to predict the expected value of such fold change by considering different assumptions on the distribution of the optical densities. Since WB data provide only relative information about protein amounts, data have to be preprocessed and in particular normalized in an appropriate way in order to enable comparison across replicates. As a consequence of the considered transformation $T_1$ of the data, normalized data are described by ratio distributions that depend on the assumed distribution for the unnormalized optical densities. Here we considered normal and log-normal distributions, following the assumptions of e.g. Möller et al. (2014) and Kreutz et al. (2007), respectively. The ratio of two log-normal distributions is again log-normal, which considerably simplifies the subsequent analysis. Instead, the inference problem related to the distribution arising from the ratio of two normally distributed RVs is not straightforward. We discussed some of the properties of these ratio distributions and provided plausible approximations that can be used for practical applications. We also dealt with the problem of structural identifiability for the class of GR distributions and presented an *in silico* Monte Carlo study to calibrate parameters of a GR distribution to i.i.d. samples and to analyse convergence properties

of the ML estimator with increasing sample size. Estimated parameters were also used to approximate the mean of the distribution, which was compared to the sample mean.

Based on only few replicates, for WB experiments typically two to four, it is usually not possible to decide on the type of underlying distribution. There are good arguments in favour of log-normal distributions (Kreutz et al., 2007), but on the other hand many standard tests are implicitly based on the normal distribution (Möller et al., 2014). From our results we could learn that, under plausible simplifying assumptions, which make the estimation problem for the GR distribution much more well-posed, the three considered error models can be fairly compared among each other as suitable description of normalized WB data.

We are aware that western blotting is a semi-quantitative method and that a small dataset does probably not contain the information required to select between our hypothetical scenarios. Nevertheless, WB data are more and more frequently used for the calibration of quantitative models, like ODE models, and in this context the selection of an appropriate error model is a necessary step, which is required for the inference problem. In this respect, our study shows that this decision also matters when it comes to parameter estimation of probability density functions. Our results, in fact, demonstrate that the choice of the error model affects the statistical properties of the inferred parameter distributions. According to our statistical framework, the estimated solution $\hat{\theta}$, obtained via MLE from a set of observations $\{y_{\mathrm{k}}^i\}_{i=1,\dots,N}$ (see Equation (2.2.6)), represents one sample of an underlying statistical distribution as well, whose mathematical formulation is obtained as a non-linear transformation of the distribution of the relative data:

$$\hat{\boldsymbol{\theta}} = T_2(\mathbf{y}_{\mathrm{k}}),$$

and therefore depends on the error model assumption. In the first two scenarios, the distributions of the estimated parameters $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ (see Equations (2.2.7) and (2.2.10)) are known from theory. In particular, given the estimators (2.2.9) and (2.2.12), the mean parameters $\hat{\mu}_1$ and $\hat{\mu}_2$ are samples from a normal distribution, while $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ from a $\mathcal{X}^2$ (chi-squared) distribution. This is not the case for $\hat{\boldsymbol{\theta}}_3$, for which we do not have a defined analytical form. By means of the *in silico* Monte Carlo analysis presented in subsection 2.3.3, we analysed some effects of $T_1$ and $T_2$ on $\hat{\boldsymbol{\theta}}$ in the context of statistical inference for GR error models from relative data. Estimation results of this simulation study show that different restricting assumptions on the underlying GR distribution have profound impact on the uncertainty of the inferred parameters. In the following two chapters 3 and 4, we will present the analysis on the effects of the second transformation $T_2$ on inferred parameters from relative data in the context of two different kinds of inference problems, namely dynamical model calibration of biochemical reaction networks and network reconstruction. This is an important information because, as consequence, the choice of the error model

has also an influence on variance estimation for any other model output as well, e.g. model predictions. This is the case for example of the inferred expected values for the KD fold change $\widehat{\mathbb{E}}(\mathbf{y}_{\mathrm{k}})$ considered in this chapter. Besides this, error models also impact the numerics and well-posedness of the optimization problem. Taken together, the choice of the error model influences statistical inference and model calibration, especially under the most usual circumstance of large variances across replicates.

For dynamical model calibration we see several challenges that are due to normalization. First, relative information about protein amounts can in some cases be much less informative compared to knowledge about absolute concentrations. This problem can be faced by appropriate rescaling of the model variables for a model-data comparison. However, there is no guarantee to get rid of all non-identifiabilities that are due to this relative information only. Furthermore, using GR distributions in a larger and more complex modelling framework may easily become prohibitive, since it complicates the evaluation of the likelihood function. In particular, the more general assumption of correlated samples at different experimental conditions (considered in scenario 3B) implies non-zero correlation coefficients $\rho$ among samples of data at different conditions and the corresponding reference value and requires the estimation of the complete parametrization represented by the set of four unknown parameters. This may become computationally too complex to solve, already in the case of a relatively simple model to be calibrated with a sufficient amount of experimental data. In this respect, we maintain that the simpler assumption of "ideally" uncorrelated distributions (scenario 3A) is sufficient for dynamical models calibration studies. This approximation is supported by our investigation results, showing a consistent improvement of the estimation results, when fixing the value of the parameter $s$ to 0, see Figures 2.3, 2.4 and 2.5. In this respect, experimental randomization of sample loading, suggested in Schilling et al. (2005), helps towards reduction of correlated errors, supporting our argumentation of not considering scenario 3B for dynamical model calibration purposes.

Nevertheless, from a theoretical point of view, it would be interesting and useful to deeper investigate the reasons behind the large uncertainty of the estimate of the complete parametrization of the GR error model. Possible reasons could be the larger flexibility in its shape and the heavy-tailedness of such distribution, leading to a larger probability of the presence of outliers in the dataset.

In summary, relative data and data normalization pose a challenge for model calibration, and the development of appropriate methods is important for quantitative predictive models when using these data. Till now this avenue of research related to the impact of data normalization on model calibration has been partially unexplored and therefore requires future investigations. In particular, we started analysing the effects of different normalization strategies for WB data and different assumptions of statistical error models on results of dynamical model calibration, by means of a simple test-bed ODE model. The

results of our investigations are presented in the following Chapter 3.

Overall, we believe that the statistical characterization of experimental measurements is an important part towards the quantitative description of intracellular processes, in particular concerning the estimation of uncertainty and credibility intervals for any model prediction.

# 3 Normalization, experimental design and error model choice affect dynamical model calibration of biochemical reaction networks

In this chapter we continue the statistical analysis presented in Chapter 2 and present a statistical framework to investigate noise propagation from concentration measurements to inferred parameters in the context of dynamic modelling of biochemical reaction networks. We design an *in silico* study where the unknown kinetic rate constants of an ODE model for a reversible phosphorylation reaction are estimated from WB time series data. By assuming realistic noise levels for the raw measured data, we investigate the effects of alternative normalization strategies as well as the effects of different error models on the results of dynamical model calibration via MLE. Based on Monte Carlo simulation results, we analyse the uncertainty of the ML estimators for increasing size of the experimental dataset and we derive that a sufficiently large amount of data is necessary to obtain reliable estimates of the model parameters. We also analyse how *finite-size* and *boundary* effects may lead to counterintuitive results and erroneous conclusions on the quality of the inferred solution. Based on statistical model comparison, we conclude that normalization by the mean outperforms normalization by a fixed time point. The choice of the error model does not seem to have a significant impact on model calibration results when considering a proper rescaling of model variables for a model-data comparison and including simultaneous estimation of error variances.

Some concepts and results presented in this chapter arose from the personal communication with Professor Jens Timmer and are explicitly pointed out in the text.

Parts of the content of this chapter are taken from Thomaseth and Radde (2021).

## 3.1 Introduction

In Chapter 2 we have seen how the choice of the error model has a profound impact on state estimation and on identifiability of the inferred parameters that characterize the chosen statistical distribution. In particular, we investigated the practical identifiability of the parametrization of the Gaussian ratio statistical model via a Monte Carlo study, and obtained that we should avoid the correlated case for parameter estimation.

As a meaningful continuation, Chapter 3 addresses the impact of a combined choice of the error model for normalized data and of the experimental design on the uncertainty properties of inferred parameters by considering as particular application problem the estimation of kinetic parameters characterizing dynamical models for biochemical pathways.

It is well known that ODEs are one of the most standard formalisms to model the dynamics of biochemical cellular processes in a deterministic and quantitative way. Under certain assumptions, ODE based modelling captures the average behaviour of a cell in a population. The main limitation of such mathematical modelling approach is that it can only provide an approximation to reality due to the high complexity of biological processes. Therefore, the main aim is to capture only the most relevant key players of the process in consideration, in order to virtually reproduce such a system under new experimental conditions and correctly predict its behaviour under uninvestigated scenarios.

The process of translating biochemical molecular interactions into a set of ordinary differential equations is not a trivial task, but nevertheless it follows well-established modelling tools, like the law of mass action, the Michaelis-Menten model for enzyme kinetics or the Hill equation for cooperativity effects (Alon, 2006; Murray, 2002). One of the main problems in this framework is the presence of unknown parameters in the derived kinetic equations. Since these values cannot be measured experimentally, this problem relates to the challenge of parameter estimation from noisy experimental data (Degasperi et al., 2017; Raue et al., 2013). Following the increase in size and complexity of the investigated mathematical models, many difficulties arise due to experimental and computational limitations. We refer to Fröhlich et al. (2019) for an "overview of the state-of-the-art methods for parameter and model inference" for large-scale models.

Either for large or small-scale models, most common techniques for ODE model calibration are statistical approaches that incorporate information about the stochastic nature of experimental data and formulate an *optimization problem*, whose optimal solution is the one that minimizes the error between measured data and model simulated values.
Mostly, researchers apply the MLE method, that consists in the definition of the *likelihood function*, based on a specific assumption on the distribution of the noisy data. The estimated parameter represents then the optimal value which maximizes the likelihood function, i.e. the statistical distribution of the observed data. Several studies investigate the performance

of different computation strategies and optimization algorithms for MLE (Degasperi et al., 2017; Hass et al., 2019; Raue et al., 2013).

As described in Chapter 2, there are different options to characterize the statistical properties of the process underlying noisy data generation. In particular, in this thesis we focus on the usage of WB data, also described in Chapter 2. For this class of experimental data, the most common assumptions on biological and experimental noise sources consist in either additive normally distributed variability (see e.g. Degasperi et al. (2017); Fröhlich et al. (2019); Hass et al. (2019)), or multiplicative log-normally distributed noise sources (Kreutz et al., 2007; Raue et al., 2013).

Furthermore, when working with WB data, normalization is a necessary post-processing transformation step of the measured data, in order to enable comparison across different replicates. Different normalization strategies are considered in practical problems (Degasperi et al., 2014). The different normalization options lead to different datasets used for parameter estimation and have to be accordingly related to model simulations in different ways, as it will be described in more detail in Section 3.2.3.

It is clear at this point, that the choice of the **error model** (**EM**), i.e. the assumed statistical distribution of the measured data, and the choice of the **normalization strategy** (**NS**) may impact the estimated solution of model parameters. Indeed they modify the mathematical definition of the cost function (i.e. the likelihood) for the optimization problem. Despite the extensive literature concerning ODE models calibrated from noisy WB data, a systematic study on the combined effects of different normalization strategies and statistical error models on the results of parameter estimation with MLE is still missing.

The goal of this chapter, in line with the central idea of this doctoral thesis, is to analyse how noise propagates from input data to estimated output variables and to derive a statistical framework for uncertainty analysis in the particular context of dynamical model calibration with MLE. By means of an *in silico* study, we quantify the impact of experimental and computational strategies on parameter estimates and derive practical advices for experimentalists and modellers for an optimal model calibration workflow. In particular, we investigate:

1. the effects of increasing the number of time points and/or the number of replicates on the improvement of accuracy and precision of the inference results;

2. the impact of the chosen NS on the uncertainty of the estimation results;

3. the subsequent effects of different assumptions on the statistical distribution of the measured data, leading to different cost functions for the optimization problem;

4. the robustness to different noise levels in a general experimental setup.

To compare comprehensively all experimental and computational combinations, we make use of the well-known BIC for statistical model comparison (see Appendix 6.2). Results, obtained through a simple test-bed model of a reversible phosphorylation reaction, show that we need to measure a sufficiently large amount of data to obtain reliable estimates of the model parameters. The choice of the EM does not lead to evident differences in the quality of the estimation results, while it does when considering the computational cost. Finally, normalization to the mean of all measured values outperforms normalization to a single point.

## 3.2 Problem formulation

As mentioned in the introductory section, in this chapter we consider computational studies in which cellular molecular interaction systems are described by means of ODE models. We refer therefore to dynamical models of the form:

$$\dot{x}(t,\theta) = f(x(t),\theta), \tag{3.2.1}$$

with state variables $x = (x_1, \ldots, x_N) \in \mathbb{R}_+^N$, describing the time-varying absolute protein concentrations of the considered system, and unknown kinetic parameters $\theta = (\theta_1, \ldots, \theta_M) \in \mathbb{R}_+^M$. In dynamic modelling the goal is to use the available set of experimental data to estimate the unknown kinetic parameters $\theta$ of the considered dynamic model (3.2.1). Therefore, associated to the ODE model, we introduce *model outputs* (also called *observables*) to describe the measurable quantities, whose simulated values will then be compared with the measured data in the optimization problem for parameter estimation. We define the output variables:

$$z(t,\theta) = h(x(t),\theta), \tag{3.2.2}$$

with $z = (z_1, \ldots, z_Q) \in \mathbb{R}_+^Q$, where $Q \leq N$, since in general not all states (proteins) of the investigated system are accessible experimentally (Degasperi et al., 2017; Raue et al., 2013). The vector field $f$ and the function $h$ are in general non-linear continuous functions of $x$ and $\theta$. To guarantee existence and uniqueness of the solution $x(t,\theta,x_0)$ of (3.2.1) for a specific initial condition $x_0 = x(t_0) \in \mathbb{R}_+^N$, $t_0 \in \mathbb{R}_+$, it is sufficient to assume $f \in \mathcal{C}^1$.

In our *in silico* study we assume that there exists a "true" noise-free dynamical model describing the biological process under investigation. This hypothesis is equivalent to assuming that there exists a "true" parameter value $\theta_0 \in \mathbb{R}_+^M$, for which the corresponding model $\dot{x}(t) = f(x(t),\theta_0)$ produces the noise-free protein concentrations $x(t,\theta_0) = (x_1(t,\theta_0), \ldots, x_N(t,\theta_0))$.

According to our statistical framework, time series data of protein concentrations are described by random variables $\tilde{\mathbf{x}}_i(t_k), i = 1, \ldots, N, k = 1, \ldots, K$, whose distribution is a function of the simulated noise-free state variables $x_i(t_k, \theta_0)$. To simulate noisy data, we use a realistic mixed error model consisting of a multiplicative and an independent additive part, similar to that suggested in Kreutz et al. (2007),

$$\tilde{\mathbf{x}}_i(t_k) = x_i(t_k, \theta_0) \cdot \eta + \epsilon, \quad \eta \sim \log \mathcal{N}(0, \sigma_\eta^2), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad i \in \{1, \ldots, N\},\, k \in \{1, \ldots, K\}. \tag{3.2.3}$$

For the standard deviations (SDs) of the proportional log-normally distributed error component ($\sigma_\eta$) and of the normally distributed additive part ($\sigma_\epsilon$) we assume realistic experimental values (Schilling et al., 2005). In our *in silico* study, realizations of the resulting random variable $\tilde{\mathbf{x}}_i(t_k)$ represent the simulated noisy absolute data.

According to the WB experimental setup, the quantified optical densities measuring protein concentrations represent a *scaled* version of the absolute values, with unknown scaling factors $\alpha_j$ specific for each replicate $j$ of the experiment, corresponding to one blot. In our simulation framework, this is equivalent to defining quantified optical densities as scaled versions of the noisy absolute data:

$$\tilde{\mathbf{y}}_i^j(t_k) = \alpha_j \tilde{\mathbf{x}}_i(t_k), \quad i \in \{1, \ldots, N\},\, k \in \{1, \ldots, K\},\, j \in \{1, \ldots, J\}. \tag{3.2.4}$$

To enable comparison across several technical replicates $j = 1, \ldots, J$, post-processing normalization of the measured protein concentrations is required, as described in Chapter 2. Different normalization strategies are used in practical applications (Degasperi et al., 2014). In Section 3.2.2 we will describe in detail three normalization strategies considered in this chapter. Independently from the chosen strategy, we can define the obtained normalized dataset as a transformation $T_1$ of the original dataset (3.2.4):

$$\mathbf{y}_{i,NS}(t_k) = T_1(\tilde{\mathbf{y}}_i^j(t_k)), \quad i \in \{1, \ldots, N\},\, k \in \{1, \ldots, K\},\, j \in \{1, \ldots, J\}. \tag{3.2.5}$$

The function $T_1$ is a non-linear transformation of the time series data, leading in general to a different distribution of normalized data $\mathbf{y}_{i,NS}(t_k)$, with respect to the original measured experimental data. In our *in silico* study, depending on the chosen NS, which defines the function $T_1$, the correct distribution $p_{\mathbf{y}_{i,NS}(t_k)}(y_{i,NS}(t_k))$ is obtained by transformation of the mixed error model (3.2.3), but its expression cannot be analytically derived.

The obtained normalized dataset (3.2.5) is finally used for the parameter estimation problem. This requires the assumption of a statistical EM. As we just mentioned, for the inference problem we cannot use the true underlying error model $p_{\mathbf{y}_{i,NS}(t_k)}(y_{i,NS}(t_k))$ generating the normalized dataset (3.2.5) (i.e. the "gold standard"), because we do not know its analytical expression [1]. We need therefore to assume an alternative statistical

[1] This clarification arose after the personal communication with Professor Jens Timmer.

EM, which in general is always an approximation of the real distribution of the considered dataset. In particular, the chosen EM relates the experimental data with the model outputs (3.2.2), assuming that the measurements are a noisy version of the simulated values, and therefore defines a function of the unknown model parameters:

$$p^{\mathrm{EM}}_{\mathbf{y}_{i,NS}(t_k)}(y_{i,NS}(t_k)|\theta), \quad i \in \{1,\dots,N\},\, k \in \{1,\dots,K\}.$$

Using the two common assumptions of normally or log-normally distributed WB raw data (see Chapter 2), we consider three classes of EMs for the normalized dataset, which will be described in detail in Section 3.2.3. In particular, as presented in Chapter 2, we refer to normal, log-normal or Gaussian ratio distributions. Given a set of realizations $y_{i,NS}(t_k), i \in \{1,\dots,N\}, k \in \{1,\dots,K\}$, the inference problem using the MLE method, under the assumption that the parametrization contains the true model, involves an optimization problem, whose solution is defined as:

$$\begin{aligned} \hat{\theta}_{\mathrm{MLE}} &= \arg\max_{\theta} \mathcal{L}(\theta) \\ &= \arg\max_{\theta} \prod_{\substack{k=1,\dots,K \\ i=1,\dots,N}} p^{\mathrm{EM}}_{\mathbf{y}_{i,NS}(t_k)}(y_{i,NS}(t_k)|\theta). \end{aligned} \tag{3.2.6}$$

In this stochastic scenario, if we repeat the whole measurement process, due to noise we would then obtain a different dataset $\{y_{i,NS}(t_k)\}_{\substack{i=1,\dots,N \\ k=1,\dots,K}}$, given by the realizations of the statistical processes $\mathbf{y}_{i,NS}(t_k)$, and subsequently a different ML estimate $\hat{\theta}_{MLE}$. In analogy to Chapter 2, we can therefore interpret the solution of the optimization problem (3.2.6) as a random variable too, whose distribution is obtained through a non-linear transformation $T_2$ of the normalized dataset (3.2.5):

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = T_2(\mathbf{y}_{i,NS}(t_k)). \tag{3.2.7}$$

The function $T_2$ depends on the chosen combination of the NS and EM. In our *in silico* study, the distribution of (3.2.7) is actually obtained as the concatenation of the two non-linear transformations $T_1$ and $T_2$ of the mixed error model of the noisy protein concentrations (3.2.3), whose mathematical expression cannot be derived analytically. We apply therefore a Monte Carlo approach in which we generate experimental data by means of the error model (3.2.3), and then propagate the noise via the transformations $T_1$ and $T_2$.

The utmost goal of this study consists in analysing the final distribution of the estimated parameters $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, considering the effects of different transformation functions $T_1$ and $T_2$. As it was graphically described in Figure 1.3 in the Introduction of this thesis, noise propagates in a non-linear manner from input noisy data $\tilde{y}_i^j(t_k)$ to output inferred parameters $\hat{\theta}_{\mathrm{MLE}}$. How the different probability density functions transform over the inference process depends on many factors, consisting in several features of the experimental and computational

designs. Furthermore, we also want to test the robustness of our results and therefore we analyse the effects on noise propagation for different noise levels of the experimental data, by varying the values of the SDs $\sigma_\eta$ and $\sigma_\epsilon$ in the mixed error model (3.2.3).

### 3.2.1 Test-bed model for a reversible phosphorylation reaction

For the study presented in this chapter we decided to consider a simple test-bed dynamical model, in order to understand the basic effects of experimental and computational factors on noise propagation for dynamical model calibration. This first investigation may eventually serve as starting point for later studies on more complex systems. We consider a reversible phosphorylation reaction, represented in Figure 3.1, in which a protein p can be phosphorylated into p$^*$.

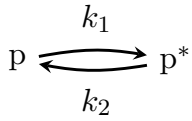


**Fig. 3.1. Reversible phosphorylation reaction.** A protein p can be phosphorylated into p$^*$, with reaction rate constants $k_1$ and $k_2$.

We assume a closed system, absence of degradation effects, and conservation of the total amount of phosphorylated and unphosphorylated proteins. By defining the state variable $x(t) \in \mathbb{R}_+$ as the phosphorylated amount with respect to the total protein concentration:

$$x(t) = \frac{p^*(t)}{p_{TOT}} = \frac{p^*(t)}{p(t) + p^*(t)}, \tag{3.2.8}$$

we obtain the following ODE describing the dynamic of $x(t)$ (for the derivation see Appendix 6.3):

$$\dot{x}(t) = k_1 - (k_1 + k_2)x(t). \tag{3.2.9}$$

In this particular case we obtain a linear ODE model, where the two parameters $k_1, k_2$ represent the two unknown reaction rate constants to be estimated, i.e. in this case we have $\theta = (k_1, k_2)$.

For a given initial condition of the state variable $x(t_0 = 0) = x_0$, we can derive the analytical solution of the differential equation (3.2.9):

$$x(t) = \frac{k_1}{k_1 + k_2} \left[1 - e^{-(k_1+k_2)t}\right] + x_0 e^{-(k_1+k_2)t}. \tag{3.2.10}$$

As described in the previous section, we make the hypothesis that a "true" parameter value $\theta_0$ of the ODE model exists, whose corresponding solution $x(t, \theta_0)$ represents the noise-free phosporylated protein concentration over time. The initial condition $x_0$ is also assumed to be known, so that we can generate a set of noisy measurements $\{\tilde{x}(t_k)\}_{k=1,\dots,K}$, $t_k > 0, \forall k$, by means of the mixed error model (3.2.3) with realistic noise settings. For one set of normalized data $\{y_{NS}(t_k)\}_{k=1,\dots,K}$, the obtained estimated parameters are $\hat{\theta}_{\text{MLE}} = (\hat{k}_{1,\text{MLE}}, \hat{k}_{2,\text{MLE}})$.

As analysed in the study thesis Wang (2018), we need to ensure that $x_0 \neq 0$ to avoid the problem of parameter non-identifiability.

### 3.2.2 Normalization strategies of WB time series data

In dynamic modelling it is common to work with time series data, that means a set of measurements of the investigated variables over time. The values collected for one run of the experiment can be compared with other replicates only after normalization.

In this thesis, we consider three different normalization strategies for time series data, which are commonly used in Systems Biology studies (Degasperi et al., 2017), similar to those presented in Degasperi et al. (2014). The first two strategies belong to the category of normalization by fixed point, while the third strategy is analogous to the category of normalization by sum. These two categories are well represented in Figure 1A-B of Degasperi et al. (2014), which we report in Figure 3.2 for convenience.

Let us consider a set of noisy data of one measured quantity $\tilde{y}^j(t_k), j = 1, \dots, J$, $k = 1, \dots, K$. Since our case study is restricted to an ODE model with one single state variable (i.e $N = 1$), from now on we do not consider the index $i \in \{1, \dots, N\}$ anymore, like we did in Equation (3.2.4). The index $j$ represents the replicate, while the index $k$ is used to indicate different time points. Referring to Figure 3.2 as example, there we have $J = 3$ total number of replicates and $K = 5$ different conditions, i.e. different time points. The first two normalization strategies consider normalization by the value at the first and last time point as reference condition, respectively, i.e. the measurements $\tilde{y}^j(t_1)$ and $\tilde{y}^j(t_K), \forall j = 1, \dots, J$. As third strategy we consider normalization by the mean of all values of the corresponding blot, i.e. $1/K \sum_{k=1}^{K} \tilde{y}^j(t_k)$ for a fixed $j \in \{1, \dots, J\}$, which is a scaled version of the normalization by the sum.

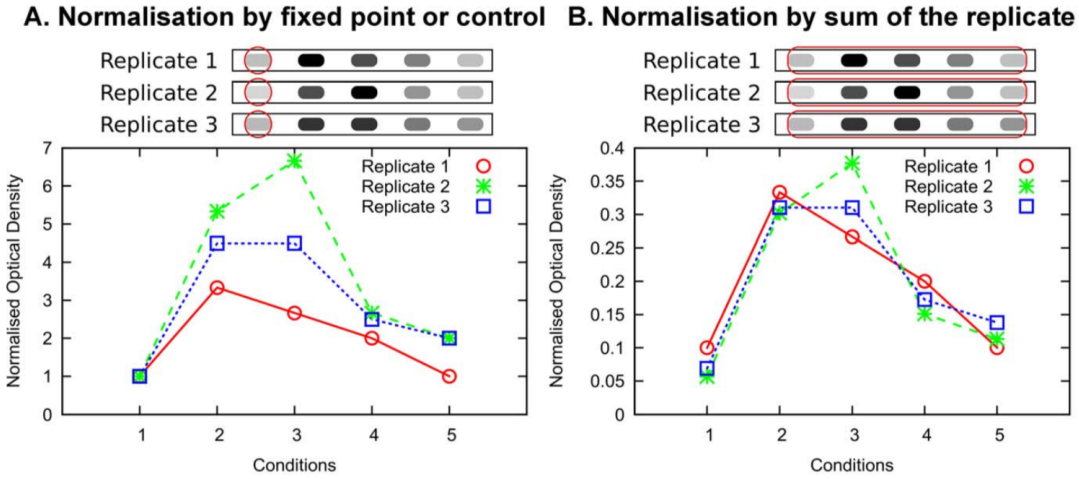


**Fig. 3.2. Normalization strategies of WB data.** The quantified optical densities from different blots at different conditions need to be normalized, in order to guarantee comparability among replicates of the same experiment. Two options are: (**A**) Normalization by fixed point and (**B**) Normalization by sum of the replicate. This Figure is taken from Degasperi et al. (2014).

In the following we summarize the three considered normalization strategies, which provide three new different datasets:

1. Normalization by the value at the first time point:

$$y_{NS1}(t_k) = \frac{\tilde{y}^j(t_k)}{\tilde{y}^j(t_1)}, \qquad j = 1, \dots, J, \quad k = 1, \dots, K. \tag{3.2.11}$$

2. Normalization by the value at the last time point:

$$y_{NS2}(t_k) = \frac{\tilde{y}^j(t_k)}{\tilde{y}^j(t_K)}, \qquad j = 1, \dots, J, \quad k = 1, \dots, K. \tag{3.2.12}$$

3. Normalization by the mean value of all time points:

$$y_{NS3}(t_k) = \frac{\tilde{y}^j(t_k)}{\frac{1}{K}\sum_{k=1}^{K} \tilde{y}^j(t_k)}, \quad j = 1, \dots, J, \quad k = 1, \dots, K. \tag{3.2.13}$$

Referring to Equation (3.2.5), the three defined NSs relate to three different forms of the transformation $T_1$ from noisy quantified optical densities to noisy normalized data. The necessity of this transformation comes from the fact that each measured value $\tilde{y}^j(t_k)$ is a multiple of the corresponding absolute noisy protein concentration, defined as $\tilde{x}(t_k)$, with unknown scaling factor $\alpha_j$, specific for each replicate $j \in \{1, \dots, J\}$ (see Equation (3.2.4)). Therefore, with all three NSs we cancel out the scaling factors, so that the normalized intensities resemble the ratio of the absolute values. For example, in the case of the first normalization strategy, the relative intensities can be expressed as ratio of the absolute

intensities in the following way:

$$y_{NS1}(t_k) = \frac{\tilde{y}^j(t_k)}{\tilde{y}^j(t_1)} = \frac{\alpha_j \tilde{x}(t_k)}{\alpha_j \tilde{x}(t_1)} = \frac{\tilde{x}(t_k)}{\tilde{x}(t_1)}. \tag{3.2.14}$$

In our statistical framework, we obtain therefore that the first transformation $T_1$ of random variables (3.2.5) is equivalent to the ratio of the random variables $\tilde{\mathbf{x}}(t_k)$ divided by either $\tilde{\mathbf{x}}(t_1), \tilde{\mathbf{x}}(t_K)$ or $\frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{x}}(t_k)$. These three random variables used as normalization quantities have different statistical properties in a general experimental setup. This implies that the random variables $\mathbf{y}_{NSs}(t_k), s = 1, 2, 3$, describing the normalized data in the three possible scenarios, also have different statistical properties among each other. We will discuss this point further in section 3.3.2 by means of the results for our test-bed model.

We have to remark that, in the case of the first and second strategy, the obtained normalized data in the first or last time point, respectively, will be set equal to 1. These data will therefore be useless for parameter estimation, leading to a lower amount of measurements used for parameter estimation. Instead, normalization by the mean value does not change the total amount of data.

### 3.2.3 Statistical description of normalized time series data for dynamic modelling

The different normalization strategies described in the previous section lead to different datasets that we may use for calibration of dynamical models. We formulate different possible options of formal statistical descriptions for the normalized datasets, an assumption which is required for the formulation of the inference problem. These options relate to different forms of the second transformation $T_2$ of stochastic distributions from normalized data to inferred parameters, see Equation (3.2.7).

In this chapter we consider the method of MLE for parameter estimation, whose basic concepts and definitions are reported in Appendix 6.1. One key step to apply this estimation method is the definition of a *statistical error model*: we assume that all noisy measurements $\{y_{NSs}(t_k)\}, k = 1, \ldots, K$, for a given normalization strategy $NSs, s = 1, 2, 3$, are realizations of the same underlying random process. In particular, we suppose that noisy data are obtained by introducing a measurement error to the defined model outputs (3.2.2), obtaining therefore an underlying statistical distribution $p^{\mathrm{EM}}_{\mathbf{y}_{NSs}(t_k)}(y_{NSs}(t_k)|\theta)$ which is a function of the unknown model parameters $\theta$. The conditional joint probability $p_{\mathbf{y}}(y|\theta)$ of the whole dataset $\mathbf{y} = \{y_{NSs}(t_k)\}$ is called the *likelihood function*.

For the purpose of this thesis, we assume that each model output $z_q(t, \theta)$ corresponds to one state variable, normalized in the same way as the considered dataset, i.e. we assume $Q = N$. To keep the notation simple, and since we look at a one-dimensional case study,

we avoid the index referring to the specific state/output variable, considering the case $Q = N = 1$.

We introduce therefore three possible output functions corresponding to the three normalization strategies:

1. Normalization by the value at the first time point:

$$z_{NS1}(t, \theta) = h_1(x, \theta) = \frac{x(t, \theta)}{x(t_1, \theta)}. \tag{3.2.15}$$

2. Normalization by the value at the last time point:

$$z_{NS2}(t, \theta) = h_2(x, \theta) = \frac{x(t, \theta)}{x(t_K, \theta)}. \tag{3.2.16}$$

3. Normalization by the mean value of all time points:

$$z_{NS3}(t, \theta) = h_3(x, \theta) = \frac{x(t, \theta)}{\frac{1}{K}\sum_{k=1}^{K} x(t_k, \theta)}. \tag{3.2.17}$$

This choice for the model outputs is supported also by the results presented in Degasperi et al. (2017). In this work the authors compare two approaches to scale model simulations to relative measured data: 1) Introducing scaling factors to convert simulated data to the scale of the experimental data (SF approach = scaling factor) or 2) Normalizing simulated variables in the same way as the data (DNS approach = data-driven normalization of the simulation). They test both methods with different objective functions and optimization algorithms for parameter estimation of dynamical systems and conclude that the DNS approach is favourable in terms of identifiability and convergence speed of the optimization algorithms and should therefore be the preferred method in dynamic modelling studies.

With this definition of model outputs, we can directly compare the normalized dataset (see Equations (3.2.11), (3.2.12) or (3.2.13)) with the simulated data and infer model parameters $\theta$ by solving an optimization problem, whose solution gives the best possible model fit of the experimental data.

In line with the statistical analysis presented in Chapter 2, which concerns the inference of stochastic models, we consider three hypotheses on the underlying distribution of the normalized time series data $\{y_{NSs}(t_k)\}$. *Normal* and *log-normal* distributions are standard settings in Systems Biology studies. The authors in Fröhlich et al. (2019); Kreutz and Timmer (2009); Raue et al. (2013); Weber et al. (2011), for example, assume independent and identically distributed additive Gaussian noise to describe the variability of biological measurements. In many other cases (see e.g. Thomaseth et al. (2013), Kreutz et al. (2007); Limpert et al. (2001)) it is rather assumed that the main source of data variability is always positive and multiplicative. Therefore the log-normal distribution is a straightforward

description of this kind of noise, producing a noisy signal which is proportional to the noise-free quantity.

As third scenario we consider the *Gaussian ratio* distribution, whose definition and properties are presented in Chapter 2. This hypothesis follows from the implicit assumption that the absolute protein concentrations are normally distributed. We can derive it easily considering Equation (3.2.14). Assuming that each $\tilde{x}(t_k), \forall k$, is a realization of a Gaussian distribution implies that the values $y_{NS1}(t_k)$ are samples of a Gaussian ratio distribution. The same holds for the other two normalization strategies. A possible correlation between the two Gaussian random variables at numerator and denominator would represent the most general assumption, according to which different samples of measurements of a protein at different time points are correlated due to systematic experimental errors. Motivated by the *in silico* study of Chapter 2 concerning the identifiabilty of the parametrization of the GR distribution, in this analysis we decided to assume rather the "ideal" case, according to which random variables $\tilde{\mathbf{x}}(t_k)$ at different time points $t_k, k \in \{1, \dots, K\}$, are independent and therefore uncorrelated.

At this point, we want to remark the fact that none of the three investigated EMs is the true statistical model generating the data. This, in fact, would be obtained as the ratio of the mixed error model given in (3.2.3), but this kind of distribution cannot be defined analytically. The *normal*, *log-normal* and *Gaussian ratio* EMs represent therefore three approximations of the true statistical model underlying the generation of the normalized WB data, which can be fairly compared among each other.

For the definition of the statistical error model we always have to specify the considered normalization strategy ($NS1, NS2$ or $NS3$), in order to refer correctly to the corresponding output function and to the set of time indices to be included in the dataset for the inference problem. In particular, with the first normalization strategy we consider the set of time indices $\mathcal{I}_{NS1} = \{2, \cdots, K\}$, since all relative data at time point $t_1$ are equal to 1. Similarly, for the second normalization strategy we consider the set $\mathcal{I}_{NS2} = \{1, \cdots, K-1\}$, since all data at time point $t_K$ are equal to 1. Finally, in the third case the set of indices is $\mathcal{I}_{NS3} = \{1, \cdots, K\}$, considering all time points.

We present here the mathematical formulation of the three statistical error models $p^{\text{EM}}_{\mathbf{y}_{NSs}(t_k)}(y_{NSs}(t_k)|\theta)$ considered in this study:

1. Normal error model (N-EM):

$$y_{NSs}(t_k) \sim \mathcal{N}(z_{NSs}(t_k,\theta),\sigma^2), \qquad s=1,2,3, \qquad k \in \mathcal{I}_{NSs} \tag{3.2.18}$$

2. Log-normal error model (LN-EM):

$$y_{NSs}(t_k) \sim \log\mathcal{N}(\log z_{NSs}(t_k,\theta),\sigma^2), \qquad s=1,2,3, \qquad k \in \mathcal{I}_{NSs} \tag{3.2.19}$$

3. Gaussian ratio error model (GR-EM):

$$y_{NS1}(t_k) \sim \frac{\mathcal{N}(x(t_k,\theta),\sigma^2)}{\mathcal{N}(x(t_1,\theta),\sigma^2)}, \qquad \rho=0 \qquad k \in \mathcal{I}_{NS1} \tag{3.2.20}$$

$$y_{NS2}(t_k) \sim \frac{\mathcal{N}(x(t_k,\theta),\sigma^2)}{\mathcal{N}(x(t_K,\theta),\sigma^2)}, \qquad \rho=0 \qquad k \in \mathcal{I}_{NS2} \tag{3.2.21}$$

$$y_{NS3}(t_k) \sim \frac{\mathcal{N}(x(t_k,\theta),\sigma^2)}{\mathcal{N}\left(\frac{1}{K}\sum_{k=1}^{K} x(t_k,\theta),\frac{\sigma^2}{K}\right)}, \qquad \rho=\frac{1}{\sqrt{K}} \qquad k \in \mathcal{I}_{NS3} \tag{3.2.22}$$

Let's mention now some remarks concerning the three error models. First, in all three cases we assume that the parameter $\sigma \in \mathbb{R}_{>0}$, related to the SD of the considered distributions, is the same for all $k$ of the dataset. In the case of the GR-EM, we assume that $\sigma^2$ is the variance of the Gaussian RVs at numerator and denominator $\tilde{\mathbf{x}}(t_k), \forall k \in \{1,\dots,K\}$. This choice is corroborated by the results of Chapter 2, where we motivated the assumption of equal SD in the knockdown and control experimental conditions. There exist usually two options concerning the assessment of the value of $\sigma$ for measurement errors: it can be either a priori empirically determined from experimental data or it can be estimated simultaneously with the model parameters $\theta$. Several studies hint to the fact that the empirical assessment is unreliable and should be avoided, since usually a low number of technical replicates (at most four) is available, while simultaneous estimation should be preferred (Degasperi et al., 2017; Raue et al., 2013). Many parametric models are suggested for the estimation of the SD $\sigma$ (Degasperi et al., 2017; Hass et al., 2019). Therefore, in this study we implemented the estimation of $\sigma$ simultaneously to $\theta = (k_1, k_2)$ and decided to consider the most basic model, for which a unique parameter value $\hat{\sigma}_{\text{MLE}}$ is estimated from the available experimental dataset.

Concerning the GR-EM, as already mentioned, we will only assume independent Gaussian RVs at numerator and denominator in the case of different experimental conditions, in this case different time points, assumption corroborated by the analysis presented in Chapter 2. It derives that in the first two cases of $NS1$ and $NS2$ the correlation coefficient $\rho$ is equal to 0. In the case of the third normalization strategy $NS3$ we calculate the correlation coefficient $\rho = \frac{1}{\sqrt{K}}$, independently of the time point $t_k, k \in \{1,\dots,K\}$, of the

random variable at the nominator, as demonstrated in Appendix 6.4. From the assumption that all $\tilde{\mathbf{x}}(t_k)$ are independent random variables (for a given $\theta$) at different time points $t_k, k \in \{1, \ldots, K\}$, it can also be easily derived that the variance of the mean $1/K \sum_{k=1}^{K} \tilde{\mathbf{x}}(t_k)$ equals $\sigma^2/K$, as assumed in the EM (3.2.22). As we saw in Chapter 2, Section 2.3.2, the GR distribution is characterized by four identifiable parameters ($\theta_3 = (a, b, r, s)$), which in this context are related to the simulated quantities and therefore functions of the unknown ODE model parameters. In Appendix 6.5, we show the definition of such parametrization of the GR distributions for all three NSs (equations (3.2.20)–(3.2.22)), which were used for the implementation of the Likelihood function in our simulation study.

## 3.3 Results

In this section we want to apply our statistical framework to investigate how uncertainty propagates from the experimentally measured data, via normalization, to the estimated kinetic parameters of an ODE model. In particular, we want to analyse if and how we can improve the goodness of the estimation results by means of some features of the experimental design. In particular we will look at the effects of increasing the number of data points, at the differences of the three normalization strategies and of the chosen error model defining the computational cost function.

### 3.3.1 Increasing the amount of time points improves the quality of parameter estimates

**Finite-size effects and boundary effects: Computational problems may lead to erroneous conclusions under realistic experimental settings**

The first component of the experimental design that we want to investigate is the amount of time points at which to measure the considered time series dataset. In particular we want to analyse how big is the benefit of increasing the amount of time points on the obtained accuracy and precision of the estimation results with respect to the consequent increase in experimental effort and costs.

We started therefore our investigations by simulating our test-bed model of a reversible protein phosphorylation reaction using plausible biological noise levels for WB data. In Schilling et al. (2005) the authors indicate a reference value of 10% for the proportional component of the measurement noise, introduced due to errors in pipetting the cellular lysates. In Figure 3.3 we show the noisy simulated time series data for four exemplary time points $t_k, k \in \{1, 2, 3, 4\}$, obtained with the considered noise-free test-bed model (3.2.10) and the mixed error model (3.2.3) to introduce experimental noise. We recall

the fact that, in this *in silico* study, the state variable $x(t)$ resembles the percentage of phosphorylated protein with respect to the total $p_{TOT}$, therefore we always have $x(t) \leq 1$. For our simulation study we set $\theta_0 = (4, 1)$ as "true" parameter values and $x_0 = 0.3$ as initial condition. We look at two noise levels obtained by varying the values of the parameters $\sigma_\eta$ and $\sigma_\epsilon$ of the multiplicative and additive noise sources, to compare the effects for low and high measurement noise. In Figure 3.3 we can observe the variability of the random variables $\tilde{\mathbf{x}}(t_k), k \in \{1, 2, 3, 4\}$ (left part) and of their mean value $1/K \sum_{k=1}^{K} \tilde{\mathbf{x}}(t_k)$, where $K = 4$ (right part), illustrated via box plots of $n = 10,000$ simulated samples.

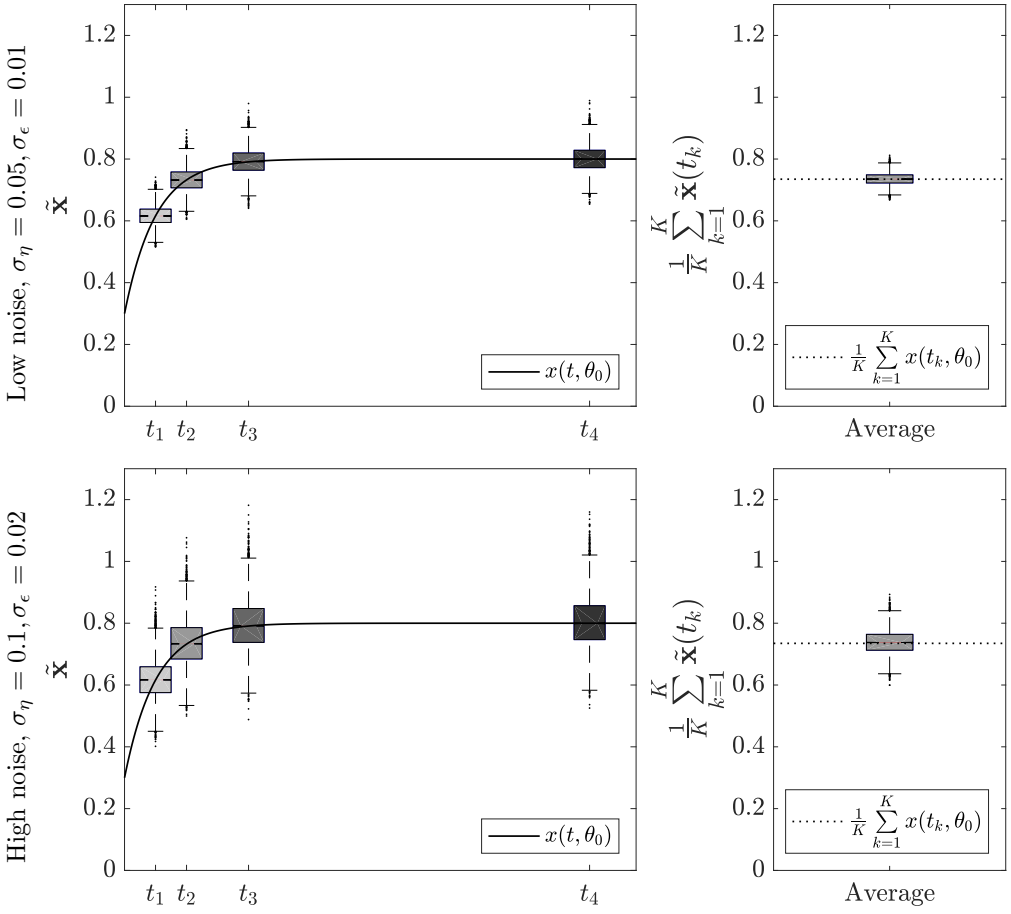


**Fig. 3.3. Noisy simulated time series data.** On the left we show the distributions (illustrated through box plots) obtained from sampled noisy realizations of the state variable $\tilde{\mathbf{x}}(t_k)$ at four exemplary time points $(t_1, t_2, t_3, t_4) = (0.2, 0.4, 0.8, 3)$. We generated $n = 10,000$ realizations via Monte Carlo simulations from the noise model (3.2.3). The continuous line represents the noise-free time course of the state variable $x(t, \theta_0)$, obtained for $x_0 = 0.3$ and $\theta_0 = (4, 1)$. On the right we show the distribution of the mean of the corresponding samples at the four time points shown on the left $1/K \sum_{k=1}^{K} \tilde{x}(t_k), K = 4$. Distributions are given for low (top) and high (bottom) noise.

From the shown simulated raw data, we obtained then $n = 10,000$ normalized datasets and solved the inference problem via MLE, for all the three considered normalization strategies. All simulations were run with the software MATLAB, and the neg-log likelihood function was minimized numerically by multi-start local optimization with a Latin hypercube sampling of the parameter space. To evaluate the quality of the obtained inference results, we considered the statistical measure of bias of the median to quantify the accuracy of the estimation and the interquartile range (IQR), representing a standard measure of the dispersion of a distribution, as an indicator for precision.

Results obtained by doubling the amount of time points from $K = 4$ to $K = 8$ are presented in Figure 3.4, where we visually summarize the two statistics considered as indicators of accuracy and precision of the inference results for both estimated model parameters $\hat{\boldsymbol{\theta}}_{\text{MLE}} = (\hat{\mathbf{k}}_{1,\text{MLE}}, \hat{\mathbf{k}}_{2,\text{MLE}})$. There we also visualize increasing noise levels with different colours. These results were obtained by considering the N-EM, as assumption for the definition of the likelihood, and the dataset obtained using the first normalization strategy $NS1$. Furthermore, the data were generated considering only $J = 1$ replicate for each measured time point.

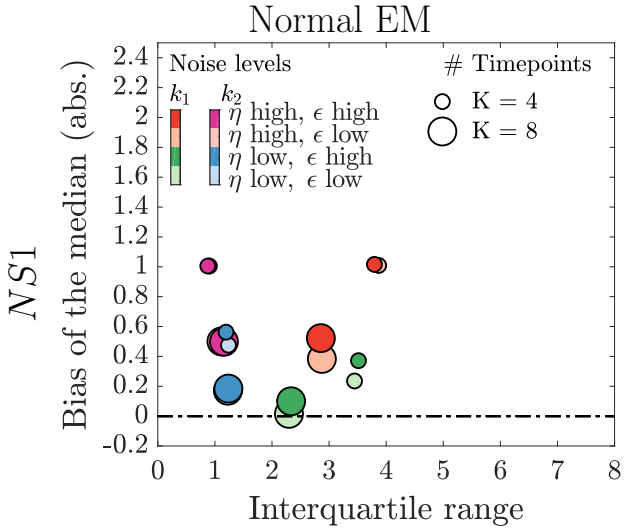


**Fig. 3.4. Effect of the amount of time points $K$ under realistic experimental settings.** Absolute values of the bias of the median versus IQR values for both estimated parameter values obtained with $K = 4$ or $K = 8$ time points and $J = 1$ replicate. These statistics are given for different noise levels $\sigma_\eta \in \{0.05$ (green/blue), 0.1 (red/magenta)$\}$ and $\sigma_\epsilon \in \{0.01, 0.02\}$ (indicated by increasing darkness). Green and red dots refer to the parameter $\hat{\mathbf{k}}_{1,\text{MLE}}$, while blue and magenta refer to $\hat{\mathbf{k}}_{2,\text{MLE}}$.

Simulation results indicate a significant improvement in terms of accuracy and precision of the estimation results in the case of $\hat{\mathbf{k}}_{1,\text{MLE}}$ when 8 measurement values are available, for both low and high noise (green and red dots). Instead, in the case of $\hat{\mathbf{k}}_{2,\text{MLE}}$ we obtain that the bias decreases but the IQR value slightly increases, which is not an expected behaviour. The overview of all nine combinations of three EMs and three NSs is given in Appendix

6.6, Figure 6.1, showing a very similar behaviour in all scenarios.

Concerning the effects of both multiplicative and additive noise components, we can observe from Figure 3.4 that the main effect is given by the multiplicative component $\eta$: In fact, dots with the same size but changing colors (green $\rightarrow$ red, blue $\rightarrow$ magenta) move substantially. Instead, dots with the same size, same colour but different shades (i.e same value for $\sigma_\eta$ but increasing value of $\sigma_\epsilon$), do not vary significantly. For this reason, in the following we will focus only on two noise levels, considering the two extreme cases in which both noise components assume the lower value (low noise: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$) or the larger (high noise: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$).

To find an explanation of the obtained counterintuitive results, we visualize the statistical distributions of both parameters by means of the box plots of the $n = 10,000$ estimated realizations (Figure 3.5a). We also investigate the corresponding marginal distributions (histograms) and scatter plots in the 2-dimensional parameter space (Figure 3.5b). In all plots we visualize both low and high noise levels with different colors.

From Figure 3.5, looking at the box plots (top) and histograms (bottom) corresponding to the second estimated parameter $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (blue and magenta for the two noise levels), we can observe that its distributions accumulate at the lower bound used for the estimation, i.e. zero. The same *boundary effect* can be observed in all the scatter plots corresponding to all combinations of three EMs, three NSs and two noise levels (see Appendix 6.6.1).

The reason behind the encountered counterintuitive results can be ascribed to a *finite-size effect*[2]. This means that the finite amount of available measured data is too small to learn about the model parameters in a sufficiently good manner. Given the high variability of many datasets, the ML estimator for $k_2$ tried to search for a solution in the negative space but the optimizer was blocked at the lower bound. This boundary effect causes a bimodal characteristic of both estimated distributions (see the histograms in Figure 3.5b), for both noise levels. As can be seen from all scatter plots, the ML estimator for $k_1$ corresponding to the peak of $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ at the lower bound is represented by the second "hill" of the bimodal distribution of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ for values larger than the true value, equal to 4.

This assumption is also supported by the interesting observation that in Equation (3.2.10), defining the solution $x(t, \theta)$ of the ODE model under investigation, the parameter $k_2$ appears only as a term in the sum with $k_1$. Subsequently, the same holds also for the output functions $z_{NSs}(t, \theta), s = 1, 2, 3$ (see Equations (3.2.15), (3.2.16) and (3.2.17)), which enter the optimization problem. Hence the boundary effect causes the bimodal characteristic of the distributions. When doubling the amount of time points $K$ from 4 to 8, while keeping the same noise level (i.e. left and right parts in Figure 3.5b), some of the samples of $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ estimated close to the 0 are "released" and the rest of the distribution

[2]The following explanations of the obtained counterintuitive results, the presented concepts about the *finite-size effect* and the subsequent structure of the results presented in this chapter arose from the personal communication with Professor Jens Timmer.

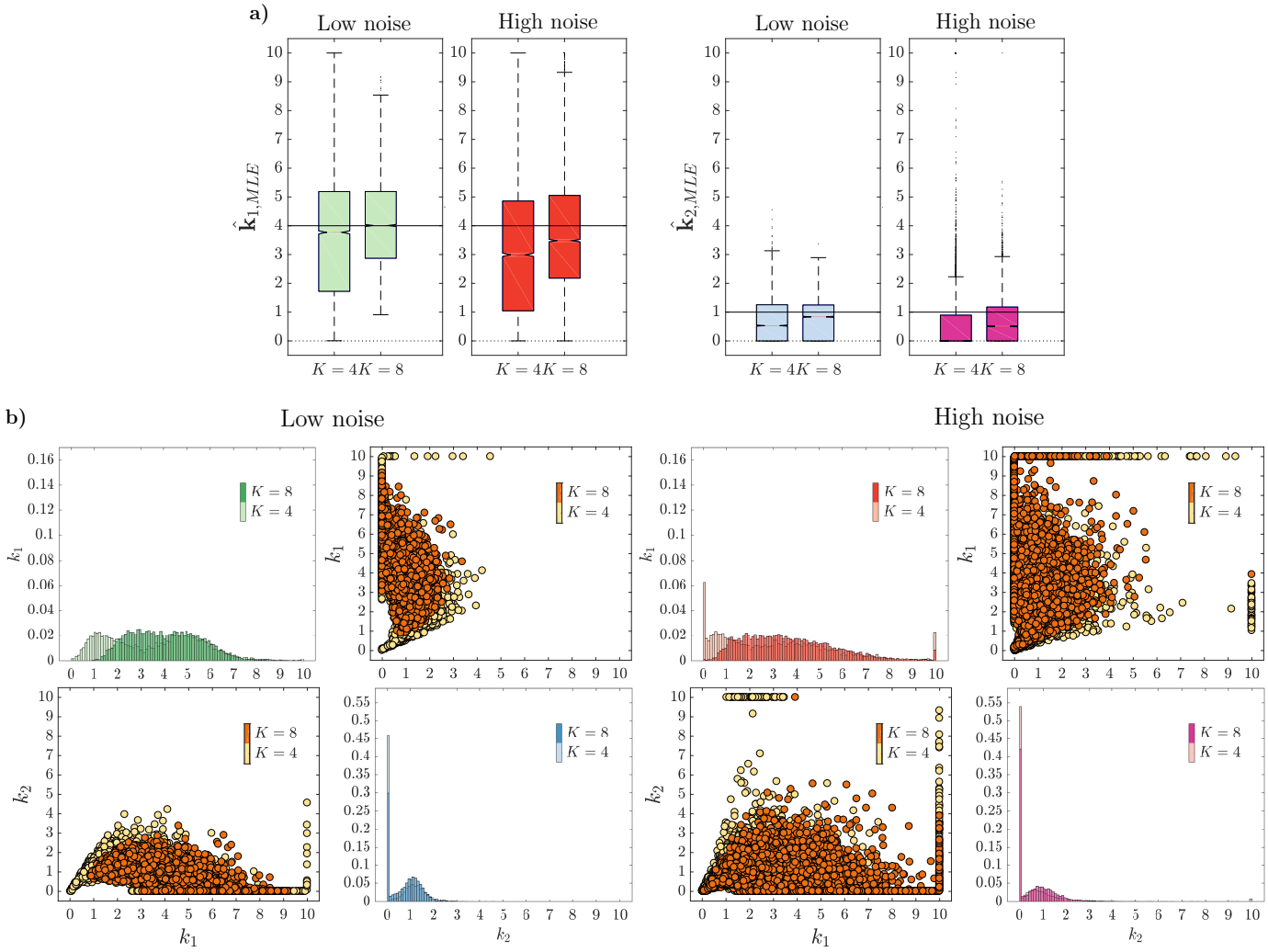


**Fig. 3.5. Effect of the amount of time points $K$ under realistic experimental settings. (a)** Box plots of the estimated parameters $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ obtained with $K = 4$ or $K = 8$ measurements of the phosphorylated protein concentration. **(b)** Marginal distributions (histograms) and symmetrical scatter plots in the 2-dimensional parameter space of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 1$, $K \in \{4, 8\}$. These results were obtained using the N-EM, the first set of normalized data ($NS1$) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$).

moves towards the "true" parameter value. This is in line with what we observe in Figure 3.5a (right part), showing that the median of the distribution of $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ gets closer to 1, while the first quartile still starts at 0, causing a larger IQR value.

At this point, we have to underline the fact that the observed increase of the IQR values for the parameter $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (Figures 3.4 and 6.1) should not lead to the erroneous conclusion of a decrease of precision of the estimation results for increasing size of the dataset. In fact, if we consider the mean squared error (MSE) as alternative measure of the estimate precision, we can observe that this quantity always decreases by doubling the amount of

measured data (see Appendix 6.6, Figure 6.2) [3].

Overall, we have learned that we cannot draw reasonable conclusions from the obtained results, since the distributions of the estimated parameters are corrupted by the boundary effects, which are related to the finite-size effect.

**Larger boundaries and increased amount of measured data help towards removing the boundary effects**

One solution to overcome the aforementioned problems is to increase the size of the measured dataset and try to obtain distributions of the parameter estimators that are not limited by the boundaries imposed in the computational problem.

In a first step, we increased the number of replicates from $J = 1$ to $J = 6$ and compared the same two values of measured time points $K = 4$ and $K = 8$. The corresponding results for all scenarios are shown in Appendix 6.6.2. There, we still observe a considerable impact of the boundary effect, motivating us to increase the size of the dataset even more.

We decided therefore to compare three numbers of time points, $K \in \{4, 8, 12\}$ and to consider $J = 10$ replicates, for a total number of 40, 80 and 120 simulated measurements, respectively. Furthermore, this time we allowed the optimizer to search also for negative values of the optimal solution and set therefore larger bounds in both directions of the parameter space. The reason for this is that we want to observe the correct distribution of the ML estimator, without boundary effects.

At this point we have to remark the fact that the estimation in the negative space was possible only for the N- and GR-EMs, while it could not be implemented for the LN-EM. We can understand this problem by looking at Equation (3.2.10), defining the solution $x(t, \theta)$ of the ODE model. Allowing negative values for the model parameter $\theta$, we may obtain some negative values $x(t_k, \theta^*)$, for some time point $t_k$ and some combination $\theta^* = (k_1^*, k_2^*)$. Hence, the output functions $z_{NSs}(t, \theta^*), s = 1, 2, 3$, may assume negative values as well. When inserting these terms into the LN Likelihood function (3.2.19), this causes an error because taking the logarithm of a negative value.

In Figure 3.6 we can see the estimation results corresponding to the three EMs, obtained with the first set of normalized data (NS1) and two noise levels (left and right columns). In particular, the histograms and symmetrical scatter plots in the 2-dimensional parameter space of $\hat{\mathbf{k}}_{1,\text{MLE}}$ and $\hat{\mathbf{k}}_{2,\text{MLE}}$ are shown, for each scenario and for the three increasing sizes of the dataset, corresponding to $K \in \{4, 8, 12\}$. Despite the large amount of data used for parameter estimation, we can observe that the boundary effects cannot be eliminated in the case of the LN-EM (Figure 3.6b), for both noise levels. Concerning the N- and GR-EMs (Figures 3.6a and 3.6c), we can see that the boundary effects disappear entirely for the low

[3] This clarification arose after the personal communication with Professor Jens Timmer.

noise level (left column). Instead, for high noise, we consider only the case for $K = 12$ to be almost unaffected.

In Appendix 6.6.3 we represent the complete set of the estimation results obtained from all three different normalized datasets (NS1, NS2, NS3), where we can observed the same behaviour for all different scenarios.

In Figure 3.7 we visually summarize the accuracy and precision of the inference results, in terms of bias of the median versus IQR values, for both estimated model parameters $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = (\hat{\mathbf{k}}_{1,\mathrm{MLE}}, \hat{\mathbf{k}}_{2,\mathrm{MLE}})$. In particular, we show the results only for the aforementioned scenarios that do not present boundary effects. We focus, in fact, on the values obtained with the N-EM (left column) and with the GR-EM (right column), for all three different normalized datasets (different rows). There, we also visualize the increasing size of the dataset with increasing size of the dots and the two noise levels with different colours. In particular, we can observe that the green and blue dots are rather close to each other for increasing sizes, respectively.

From these simulation results we can conclude that, for a very large amount of measured data, for which the distributions of the ML estimators tend to the asymptotic behaviour, the benefit of doubling or tripling the amount of measured data is rather minimal with low input noise, while this is not the case for high noise. Furthermore, we observe that the GR-EM is robuster than the N-EM considering the first and second NSs, especially for high noise level. This topic will be further analysed in the following section.

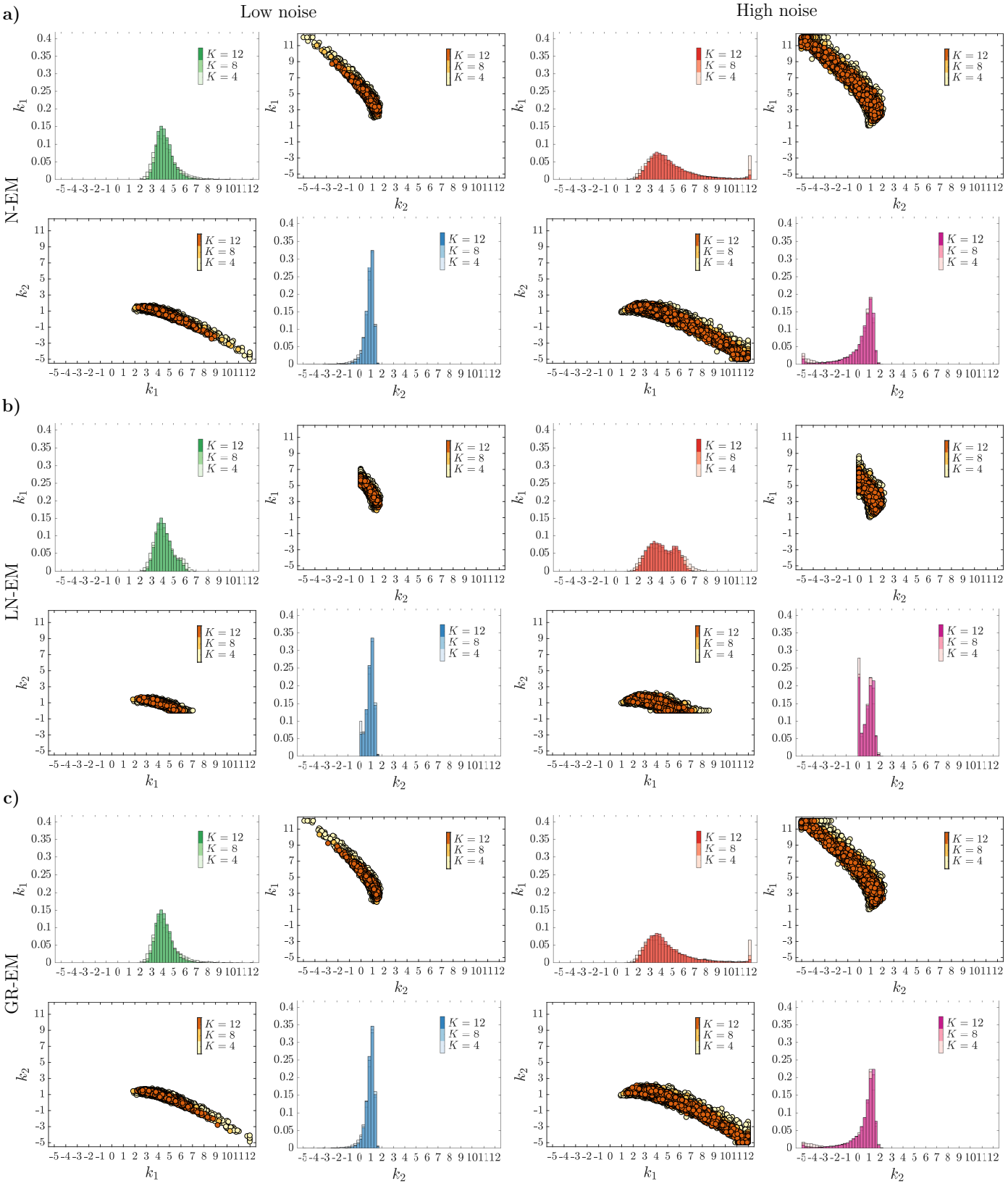


**Fig. 3.6. Effect of the amount of time points $K$.** Marginal distributions (histograms) and symmetrical scatter plots in the 2-dimensional parameter space of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 10$ and $K \in \{4, 8, 12\}$. These results were obtained using the first set of normalized data ($NS1$) and the **(a)** N-EM, **(b)** LN-EM, **(c)** GR-EM, for two noise levels, respectively (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$).

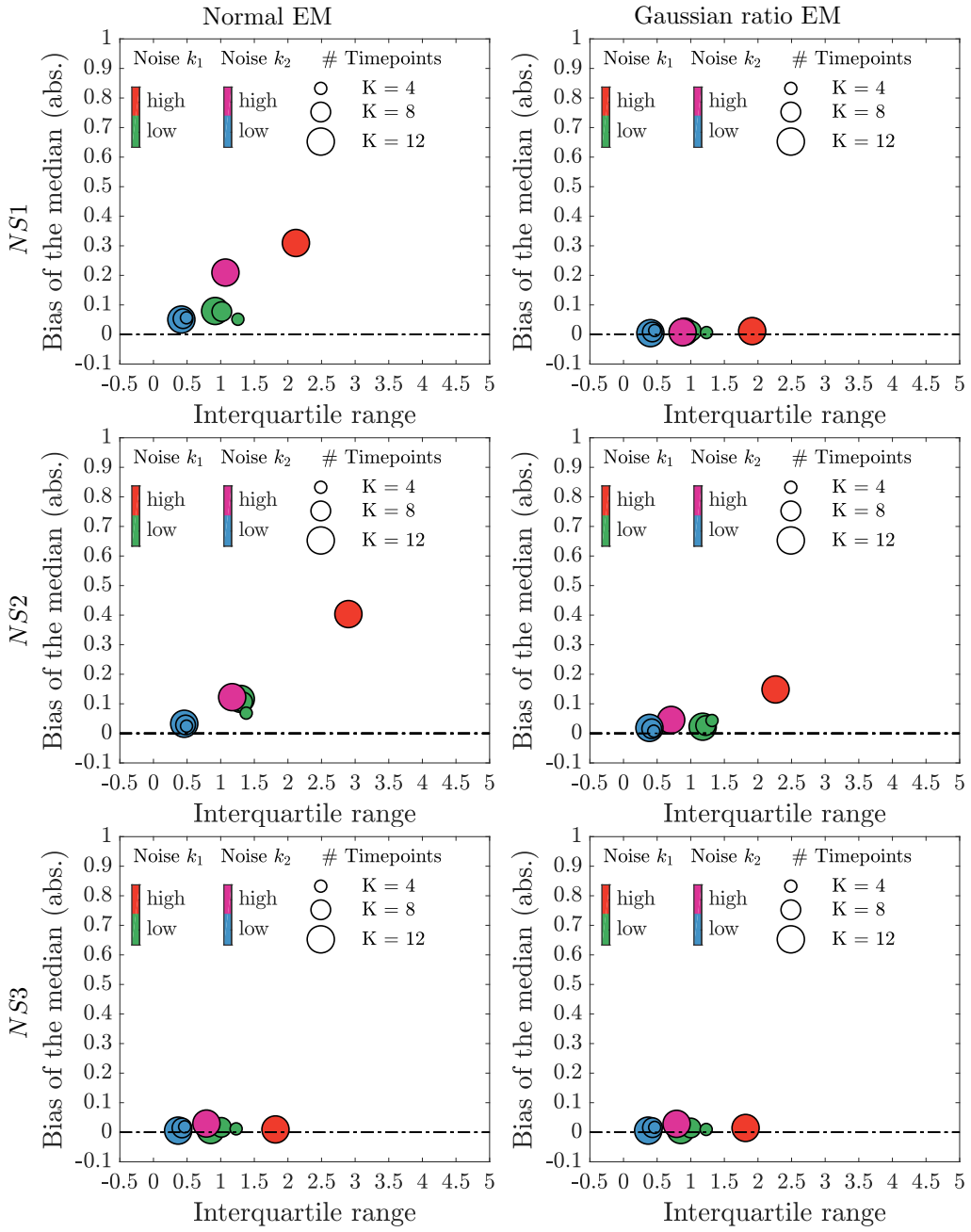


**Fig. 3.7. Effect of the amount of time points $K$.** Absolute values of the bias of the median versus IQR values for both estimated parameter values obtained with $K = 4, 8$ and 12 time points and $J = 10$ replicates, given for low noise level ($\sigma_\eta = 0.05$ and $\sigma_\epsilon = 0.01$). For high noise ($\sigma_\eta = 0.1$ and $\sigma_\epsilon = 0.02$) these statistics are given only for the case $K = 12$. Green and red dots refer to the parameter $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$, while blue and magenta refer to $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$.

### 3.3.2 Impact of normalization strategies on the uncertainty of Maximum Likelihood estimates

As discussed in Section 3.2.2, normalization modifies the statistical properties of the transformed relative data with respect to the raw measured values. This effect changes depending on the chosen normalization strategy, due to the different distributions of the random variables used as reference condition.

The question arises, how the three considered different strategies affect noise propagation from the raw concentration measurements to the estimated model parameters, while keeping fixed other features of the inference process.

Referring to our *in silico* study, in Figure 3.3 we can compare the variability of the random variables $\tilde{\mathbf{x}}(t_1), \tilde{\mathbf{x}}(t_4)$ (left part) and $1/K \sum_{k=1}^{K} \tilde{\mathbf{x}}(t_k)$, where $K = 4$ (right part), used as reference quantities for normalization in the three considered strategies, and notice how their statistical properties differ a lot among each other.

The corresponding statistical distributions (box plots) of the normalized data obtained with the three considered normalization strategies are shown in Figure 3.8, also for the two low and high noise levels, on the left and right columns, respectively. As a consequence of the specific variability of the quantity used as reference condition for normalization, the statistical properties of the obtained normalized datasets differ a lot among each other. In particular, we can observe that data normalized with the first strategy (upper row) have the largest uncertainty, followed by the second strategy and finally by the third, which has the lowest uncertainty. This is due to the fact that the normalization variable in the third case is calculated on a larger number of points and will therefore have a lower SD, as can be seen in Figure 3.3 on the right, leading to lower noise. Furthermore, the first strategy is the most sensitive to higher noise level (right column), showing many more outliers in the right tails of the distributions, which were cut off for representative reasons.

These facts are in line with what was discussed in Degasperi et al. (2014), who suggested to avoid choosing normalization points with low quantified intensities for hypothesis testing studies, since this strategy results in large CV for normalized data. Instead, normalization by the mean value redistributes the uncertainty among the random variables at the different time points.

We continue therefore our analysis investigating the influence of all three normalization strategies on the uncertainty of the estimated model parameters $\hat{\boldsymbol{\theta}}_{\text{MLE}} = (\hat{\mathbf{k}}_{1,\text{MLE}}, \hat{\mathbf{k}}_{2,\text{MLE}})$. In particular we want to analyse the combined effects on accuracy and precision of the results, and look therefore at the distributions of the inferred parameters obtained by noise propagation from the raw simulated data.

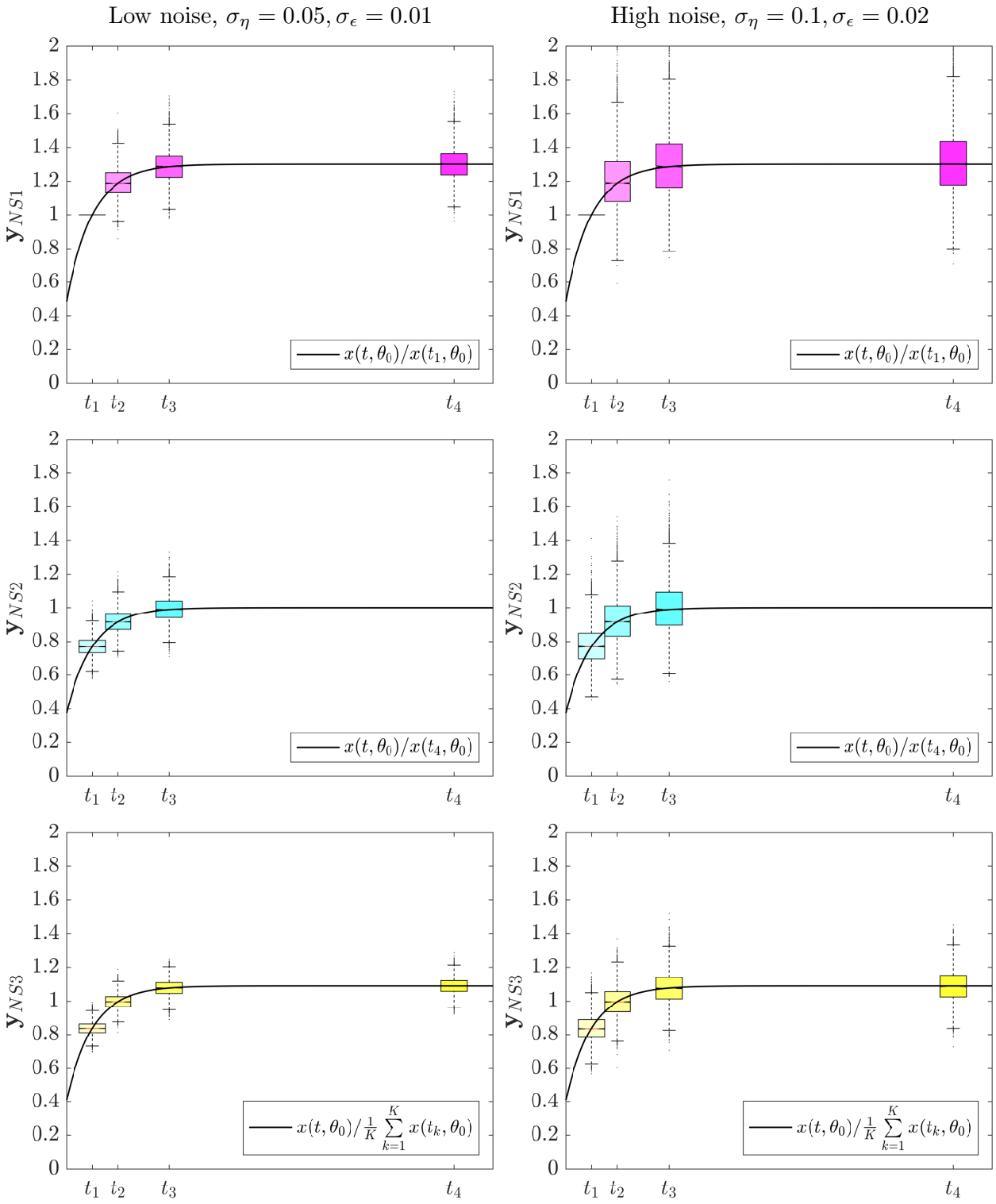


**Fig. 3.8. Noisy normalized time series data.** The shown distributions (box plots) are obtained from sampled noisy realizations of the normalized variables $\mathbf{y}_{NSs}(t_k), s = 1, 2, 3$, for the three different normalization strategies, marked with different colors, and are given for low (left) and high (right) noise.

Results are shown in Figures 3.9 and 3.10, in which we visualize the bias of the median versus IQR values as statistical measures of accuracy and precision of the estimation results. In this analysis, we focus only on the results obtained with N- and GR-EMs. In particular, Figure 3.9 concerns the estimation results of the datasets obtained with $K = 4$ time points and low noise level. Instead, Figure 3.10 relates to the case of high noise and $K = 12$. Overall, the considered scenarios relate to all cases in which the estimation results did not present boundary effects, as analysed in Section 3.3.1.

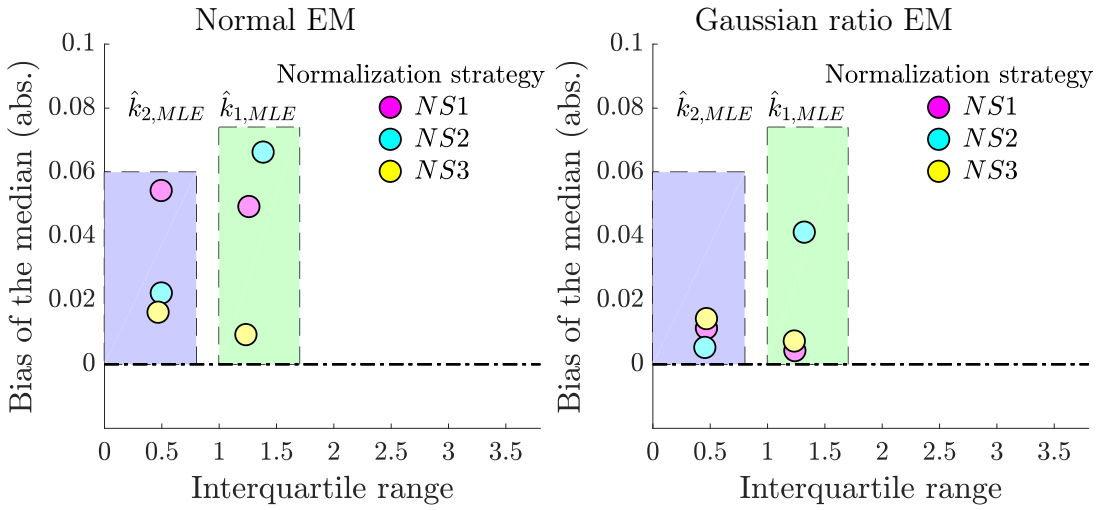


**Fig. 3.9. Effects of three alternative normalization strategies on accuracy and precision of parameter estimates.** Absolute values of the bias of the median versus IQR values for both estimated parameters $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ obtained with the three considered normalization strategies: $NS1$ by first time point (magenta), $NS2$ by last time point (blue) and $NS3$ by the mean of all time point values (yellow). These statistics were obtained assuming either the N-EM (left) or GR-EM (right) as likelihood function for the optimization problem, for $K = 4$ time points, $J = 10$ replicates and low input noise: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$.

First, we observe that the impact of the three normalization strategies is different for the two estimated parameters, therefore we cannot derive a unique statement concerning the impact of the three normalization strategies on the estimation results.

For low noise level (Figure 3.9) the different NSs affect mainly the accuracy of the estimation, while for high noise (Figure 3.10) they impact both accuracy and precision. Nevertheless, the trend is roughly maintained if considering the same EM and same parameter with increasing noise.

One general unexpected result is that the first normalization strategy (magenta dots) does not always lead to the worst results (i.e. both higher bias and higher IQR) even if the corresponding normalized data used for estimation show the largest variability. Instead, NS2 (blue dots) causes the largest IQR and also the largest bias for $\hat{k}_{1,\mathrm{MLE}}$, for both noise

levels. This is probably due to the fact that in this case we loose information of the data at steady-state, and for fixed $\hat{k}_{2,\mathrm{MLE}}$ the uncertainty of the estimated value $\hat{k}_{1,\mathrm{MLE}}$ increases.

From this analysis we can affirm that the impact of different normalization strategies is not univocal on all estimated parameters under realistic noise settings, despite the large amount of data used for the estimation. We maintain that a good compromise is to choose the third normalization strategy (by the mean value), which in general shows the lowest bias also for larger noise level.

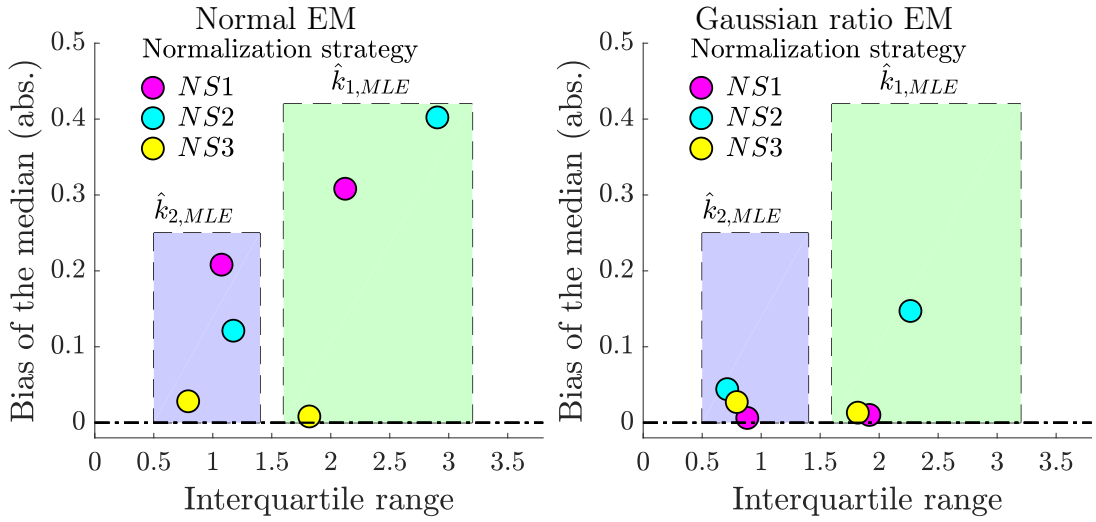


**Fig. 3.10. Effects of three alternative normalization strategies on accuracy and precision of parameter estimates.** Absolute values of the bias of the median versus IQR values for both estimated parameters $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ obtained with the three considered normalization strategies: $NS1$ by first time point (magenta), $NS2$ by last time point (blue) and $NS3$ by the mean of all time point values (yellow). These statistics were obtained assuming either the N-EM (left) or GR-EM (right) as likelihood function for the optimization problem, for $K = 12$ time points, $J = 10$ replicates and high input noise: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$.

### 3.3.3 Statistical model comparison

In the previous part of this Section, we presented a statistical analysis of our simulation *in silico* study to analyse noise propagation from experimentally measured data to the inference results of a test-bed non linear ODE model. We considered a realistic error model and plausible noise levels to generate noisy data (see Figure 3.3). In particular, we analysed the quality of the estimation results by looking at the statistical quantities considered as indicators of accuracy and precision of the inferred parameters, namely bias of the median and IQR. In a first step we analysed the effects of increasing the size of the dataset used for one estimation run and we observed that *finite-size* and *boundary* effects may lead to

counterintuitive results (see for example the results shown in Figure 3.5). In the particular case of the LN-EM, we could still not eliminate the boundary effects also for a very large size of the dataset. The reason for that was that we could not let the optimizer to search in the negative parameter space, due to numerical problems, and we should have increased the amount of considered measurements even more. Therefore, when using the LN-EM, it becomes difficult to eliminate the boundary effects when estimating model parameters whose true positive value is near the zero value, especially for high noise.

For a comprehensive comparison of the goodness of the estimation results for the other two statistical EMs (N and GR) and the three NSs, we evaluate the distribution of the BIC values calculated for each estimated parameter set corresponding to all six different estimation scenarios (i.e. three NSs and two EMs). For more details about the BIC value see Appendix 6.2. Given the definition of the BIC (see Equation (6.2.6)) we need to express the number $k$ of the estimated parameters, which equals 3 for both considered likelihood functions, namely the two kinetic parameters $k_1$ and $k_2$ and the SD $\sigma$ of each statistical model (see Equations (3.2.18) and (3.2.20)-(3.2.22)). Furthermore we have to set the value $n$ defining the number of observed data, which changes depending on the chosen NS. In particular, we have one data point less for the first two strategies $NS1$ and $NS2$ with respect to the third strategy $NS3$. We loose in fact one data point due to normalization by fixed point (see Figure 3.2).

In Figure 3.11 we show the distributions (box plots) of the BIC values obtained with the dataset corresponding to $K = 12$ time points and $J = 10$ replicates, corresponding to $n = 110$ for $NS1$ and $NS2$ and $n = 120$ for $NS3$, obtained for the low noise level of the input data ($\sigma_\eta = 0.05$ and $\sigma_\epsilon = 0.01$). From these results we observe that the goodness of the two error models is very similar among each other if using the same set of normalized data ($NS1, NS2$ or $NS3$). The lowest BIC values are obtained with the third normalization strategy, despite the larger size of the dataset (parameter $n$) that corresponds to a larger penalty factor in the BIC definition. This result supports the conclusion of Section 3.3.2, suggesting $NS3$ as the best normalization strategy to be applied for parameter estimation.

Looking at the effects of the two EMs for a fixed NS on the BIC value distributions, we can observe that the median and the spread of the distributions corresponding to the GR-EM are slightly lower than those of the N-EM for the first two NSs (compare the green and orange straight lines), while the distributions are almost identical in the case of NS3. The observed similarity between the BIC values of the N- and GR-EMs is probably due to the validity of the condition for the approximation of the GR distribution with a Gaussian distribution. As presented in Chapter 2, this approximation holds for a sufficiently large CV of the RV at the nominator of the ratio distribution, in the case of assuming uncorrelated signals.

Results obtained for the high noise level in the data ($\sigma_\eta = 0.1$ and $\sigma_\epsilon = 0.02$) are shown in Appendix 6.7. These results show the same trend as that obtained with low noise, even

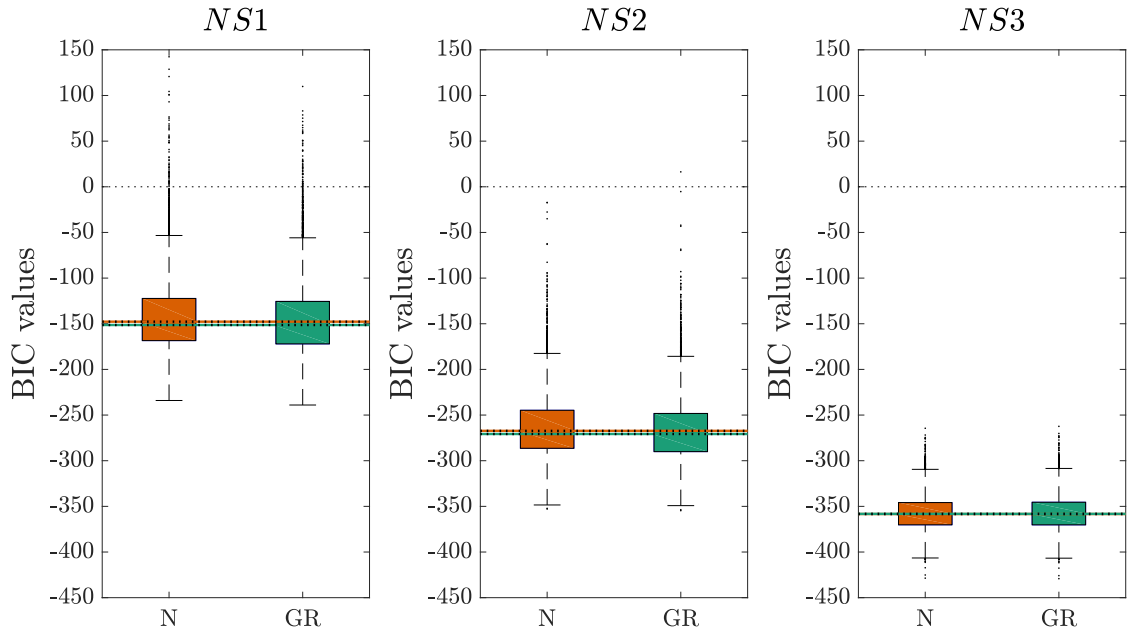


**Fig. 3.11. Statistical model comparison.** Box plots of the BIC values (6.2.6) calculated with the estimation results obtained using the different datasets from the three normalization strategies $NS1, NS2$ or $NS3$ and two EMs (N and GR). These results were obtained with $K = 12$ time points, $J = 10$ replicates and the low noise level of the input data: $\sigma_\eta = 0.05$ and $\sigma_\epsilon = 0.01$.

though in the case of higher input noise the distributions of the BIC values are shifted up to higher values, and more outliers appear in the right tails of the distributions, especially in the case of the first NS ($NS1$ on the left). Nevertheless, we can observe that the GR-EM is slightly robuster than the N-EM for larger noise if using the first or second normalization strategy.

Overall, these results confirm the conclusion of Section 3.3.2 of preferring the normalization strategy by the mean value instead of that by fixed time point, for dynamical model calibration purposes. As concerns the choice of the error model used as statistical description of the relative dataset, we suggest to select the N-EM, since it is computationally faster to optimize than the GR-EM.

## 3.4 Summary and discussion

In this chapter, we presented the results of a statistical analysis of the combined effects of different normalization strategies for WB time series data, different experimental design approaches and different assumptions on the statistical model generating the noisy data on the quality of parameter estimation via MLE for ODE models of biochemical reaction networks. We developed a statistical framework to investigate noise propagation from input measured data to output estimated parameters via Monte Carlo simulations. We considered a reversible protein phosphorylation reaction as test-bed model for our *in silico* study, which we used to generate noisy time series concentration measurements by means of a realistic noise model for WB data (Kreutz et al., 2007; Schilling et al., 2005). Nevertheless, for the inference problem we could not use the "gold standard" as definition of the error model underlying normalized WB data generation, since we do not know its analytical expression. We defined instead three alternative statistical models as approximations of the true distribution. This choice resembles a real case scenario in which the true statistical process generating the data is in general different from the EM assumed in the inference problem. We made use of statistical measures such as bias of the median and IQR to evaluate the accuracy, precision and robustness of the estimation results, as terms for comparison of the goodness of the different experimental and methodological strategies for parameter estimation.

From this study we could derive some interesting findings. First, concerning the choice of the normalization strategy, we got the clear recommendation of applying the third strategy, i.e. normalization by the mean value instead of normalization by a fixed point. Opting for this normalization strategy, in fact, leads to the best inference results in terms of accuracy and precision of the estimated parameters, independently from the chosen experimental design and noise level in the input data.

Regarding the experimental design, we analysed the effects of increasing the total amount of the measured data, and we observed, as expected, that increasing both amounts of time points and replicates leads to a general improvement of the estimation quality. A less trivial statement is, what is the minimal amount of data in order to obtain reliable ML estimates of the model parameters. Due to finite-size effects, we can in fact run into erroneous conclusions on the goodness of parameter estimation. The counterintuitive results, observed in Figure 3.4 and in Appendix 6.1, are caused in fact by boundary effects. Therefore, we cannot draw meaningful conclusions from those results and a solution is to increase further the number of measured data and to allow the optimizer to search also for negative values of the parameter space. This could not be implemented for the LN-EM. A quite unexpected result of our simulation study is that we are not able to fully eliminate the boundary effects, even for a large amount of measured data (up to 120 measurements). In

the case of the LN-EM model this was not possible, both for low and high noise levels, for which reason we decided not to consider it any more for the comparison among the three EMs. Instead, in the case of the N and GR assumptions, this was possible for low input noise, since a small part of the distribution of the ML estimates was allowed to spread also in the negative space. By the way, the problem could not be totally eliminated for the larger noise level, in which case we considered only the largest amount of data (120) for the subsequent comparison of the results.

A surprising result was to observe no significant differences among the two investigated statistical EMs, namely N and GR, as concerns the quality of the inference results. The only observed difference is that, in the case of high input noise (see the figure in Appendix 6.7), the GR-EM seems to be slightly more robust than the N one. An interesting study of Maier et al. (2017) reveals that using heavier-tailed distributions, such as Laplace, Huber, Cauchy or Student's t, instead of the N distribution for parameter estimation of dynamical systems, improves the quality of the inference results in terms of robustness against outlier-corrupted datasets. Motivated by these findings, we speculate that using the GR distribution as error model to describe normalized WB data, which belongs to the class of heavy-tailed distributions, may be beneficial for ODE model calibration in the case of the presence of outliers in measured data. This would be an interesting extension of our statistical analysis and, related to this, we should be able to provide the necessary gradients and Hessian matrices of the likelihood function in order to ensure an efficient optimization, as stated in Maier et al. (2017). In fact, the main bottleneck of applying the GR distribution in the likelihood function is the computational cost.

We are aware that our conclusions are based on the results obtained with a simple test-bed model of one single equation and only two kinetic parameters. Common Systems Biology studies consider larger models containing many unknown parameters to be estimated from noisy relative data. Nevertheless, our simulation results highlight the importance of taking noise transformation into account when dealing with data post-processing techniques like normalization. The crucial fact is that measured input data generate uncertain estimated parameters that subsequently lead to model predictions affected by uncertainty. In this respect, our analysis lays the foundations for a bigger awareness that noise transformation via normalization may lead to a significant uncertainty of the inference results, even for a significant amount of the measured data, which may then lead to wrong or unreliable model predictions.

A further aspect impacting the uncertainty of model predictions is the intrinsic uncertainty of the calibrated EM, which is related to the SD of the assumed distribution. In general, it is not trivial how to set this value, for which an empirical estimate is commonly used. We opted, instead, for the simultaneous estimation of the unknown additional parameter $\sigma$, characterizing all investigated EMs (see Equations (3.2.18)–(3.2.22)), as described in

Section 3.2.3. The corresponding obtained estimates $\hat{\sigma}_{\mathrm{MLE}}$ were not shown in the Results section and are given for completeness in Appendix 6.8. In this case, we cannot compare directly the obtained estimates with some "true" parameter value, because the SD of the error model used to simulate the noisy data has a different meaning from the parameter $\sigma$ characterizing the EMs used in the inference problem.

# 4 Impact of measurement noise, experimental design, and estimation methods on Modular Response Analysis based network reconstruction

*Modular Response Analysis* (MRA) is a method to reconstruct signalling networks from steady-state perturbation data that has frequently been used in different settings. Since these data are usually noisy due to multi-step measurement procedures and biological variability, it is important to investigate the effect of this noise onto network reconstruction.

In this chapter we present a systematic study to investigate propagation of noise from concentration measurements to network structures, in an analogous way to what we presented in the previous chapter concerning dynamic modelling of biochemical reaction networks. Therefore, we design an *in silico* study of the MAPK and the p53 signalling pathways with realistic noise settings. We make use of statistical concepts and measures to evaluate accuracy and precision of individual inferred interactions and resulting network structures. Our results allow to derive clear recommendations to optimize the performance of MRA based network reconstruction: First, large perturbations are favourable in terms of accuracy even for models with non-linear steady-state response curves. Second, a single control measurement for different perturbation experiments seems to be sufficient for network reconstruction, and third, we recommend to execute the MRA workflow with the mean of different replicates for concentration measurements rather than using computationally more involved regression strategies.

The main content of this chapter is taken from Thomaseth et al. (2018).

## 4.1 Introduction

Advanced experimental techniques have facilitated our mechanistic understanding of intracellular processes in the last decades. However, the problem of network reconstruction from experimental data remains a challenging task, for which many different approaches have been suggested. Among those, Modular Response Analysis (MRA) has been proven successful in many applications (Gong et al., 2015; Santos et al., 2007; Speth et al., 2017; Stelniec-Klotz et al., 2012). MRA uses steady-state data of experiments in which each node of a network is perturbed successively (see Figure 4.1a). These steady-state data are transformed into quantitative pairwise interaction strengths, denoted Local Response Coefficients (LRCs), which define the network structure. This is done in a two-step process, in which first concentration measurements are transformed into Global Response Coefficients (GRCs), which are then used to calculate the LRCs. MRA is an elegant method that gives reliable results in case that concentrations can be accurately measured and measurement noise can be neglected (Kholodenko et al., 2002).

MRA is however often applied in settings in which one has to deal with real noisy experimental data and few replicates, for example when using western blotting to investigate signalling pathways, as exemplified in Santos et al. (2007); Stelniec-Klotz et al. (2012). In these studies the authors addressed the issue of measurement noise by using statistical approaches, like Monte Carlo simulations or ML, to estimate interaction strengths and respective uncertainties. Despite the extensive usage of MRA, the effect of noise in the input data on network reconstruction is not completely understood. Recent developments include statistical reformulations of the MRA that have been suggested to address the issue of noisy and sparse/insufficient data (Santra et al., 2018). A further extension combines the classical deterministic MRA framework with advanced nonparametric single-cell data resampling to discriminate between direct and indirect connectivities (Kang et al., 2015).

Despite the broad literature tackling the issue of experimental noise, a comprehensive and systematic study on how network inference via MRA is affected by noise is still required. In particular, some of the studies do not consider noise propagation from the input data to the estimated LRCs, but start directly with the GRCs (Andrec et al., 2005). Furthermore, a realistic statistical characterization of the MRA variables (measured data, GRCs and LRCs) and a robustness analysis in a general experimental setup are also still missing.

In this chapter, following the central motif of this doctoral thesis, we develop a statistical framework to analyse noise propagation in the context of MRA based network reconstruction (Figure 4.1b), proceeding with a similar scheme as we did for the study on dynamic modelling for biochemical reaction networks, presented in Chapter 3.

By comparing different experimental and computational strategies in an *in silico* study, we derive recipes for experimentalists and modellers regarding an optimal MRA workflow design. In particular, we investigate:

1. how non-linear transformations and mathematical approximations of the MRA framework affect noise propagation;

2. the influence of perturbation strength, control strategy and number of replicates on the uncertainty of the estimated interactions;

3. the effects of different estimation methods on the performance of the network inference problem.

To evaluate the resulting network structure, we apply a performance evaluation method that was proposed in Bansal et al. (2006) and is schematically depicted in Figure 4.1c. It works similar to a Receiver Operating Characteristic (ROC) and its Area Under the Curve (AUC) value for evaluating the performance of a classifier but, additionally, taking correctness of the sign of an inferred interaction into account. A correctly identified network has an AUC value of 1 (whole square), the random case corresponds on average to an AUC value of 0.25 (darker grey triangle).

Results are given for two test-bed examples of well-known signalling pathways, a model for the MAPK pathway and a model for the tumour suppressor protein p53. These models show very different non-linear properties regarding their steady-state behaviour in dependence of perturbation strengths. Our results show that large perturbations and few technical replicates, combined with a simple control strategy and a basic estimation method, lead to an optimal 'bias-variability' trade-off of the estimated pairwise interactions and also give robust results regarding network reconstruction.

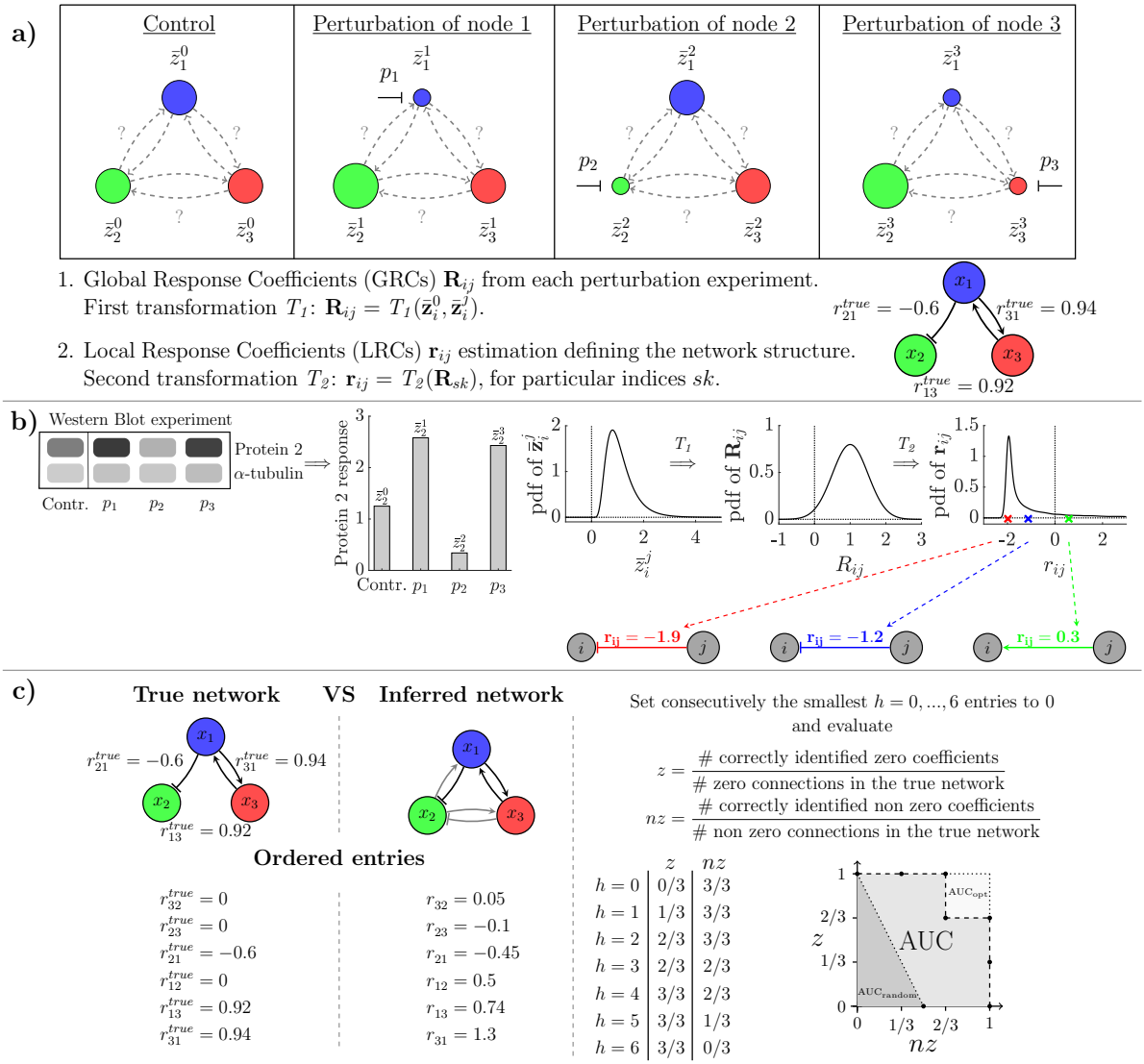


**Fig. 4.1. Variability and noise in measurements affect network reconstruction via Modular Response Analysis. (a)** MRA workflow for a three-node network. After subsequent perturbation of all nodes and quantification of concentration fold changes, LRCs $\mathbf{r}_{ij}$ are calculated via a two-step non-linear transformation. **(b)** One exemplary realization (one replicate) of the noisy measurement for protein 2 is evaluated in all experimental conditions (left part). These values are interpreted as samples of the corresponding distributions, described by the probability density function (pdf of $\bar{\mathbf{z}}_i^j$) arising from the measurement noise. On the right we describe how network reconstruction is affected by the variability of the measured protein concentrations in terms of propagation of distributions from the measurements $\bar{\mathbf{z}}_i^j$ and $\bar{\mathbf{z}}_i^0$ via the GRCs $\mathbf{R}_{ij}$ to the LRCs $\mathbf{r}_{ij}$.
**(c)** Performance evaluation of inferred network structures is done by using the assessment method presented in Bansal et al. (2006), which compares the inferred structure with a reference structure. Similar to ROC analysis, the Area Under the Curve (AUC) serves as a normalized measure for the fit quality, and varies from the average value 0.25 in the random case (darker grey triangle) up to the optimal value 1 for a correctly identified network (whole square).

## 4.2 Problem formulation

MRA is a mathematical approach to reveal interaction strengths from steady-state perturbation data of a dynamic network of interacting modules. In our examples each module consists of a single protein, and we refer to them as nodes (Kholodenko et al., 2002). Considering the network at equilibrium, pairwise interaction strengths between nodes are characterized by quantifying the immediate change of the activity of one node of the network caused by a small change of another node, whereas the rest of the network is unaffected. Local Response Coefficients (LRCs) express these local effects among $N$ nodes and are defined as the fractional change of the steady-state concentration of node $i$ ($\bar{x}_i$) with respect to that of node $j$ ($\bar{x}_j$), while keeping all other nodes $\bar{x}_k$, $k \neq i, j$, at a constant level,

$$\text{LRCs}: \quad r_{ij}^{true} = \frac{\partial \ln \bar{x}_i(\bar{x}_j, \bar{x}_k)}{\partial \ln \bar{x}_j}, \quad \bar{x}_k = \text{const}, \, k \neq i, j, \quad i \neq j, \quad i, j = 1, \ldots, N. \tag{4.2.1}$$

These LRCs describe pairwise interactions between nodes when they act in isolation and are not directly accessible. A perturbation of one parameter $p_j$, which specifically affects the activity of node $j$, spreads over the whole network, thus generating a global change of the equilibria of all nodes. This global change can be quantified from fold change measurements of concentrations relative to the unperturbed system. Formally, Global Response Coefficients (GRCs) are defined as the total derivative of the logarithm of the steady-state variables ($\ln \bar{x}_i$) with respect to the perturbed parameter ($p_j$) (see exemplary network in Figure 4.1a),

$$\text{GRCs}: \quad R_{ij}^{true} = \frac{\mathrm{d} \ln \bar{x}_i(p_j)}{\mathrm{d}\, p_j} = \frac{1}{\bar{x}_i(p_j)} \frac{\mathrm{d}\, \bar{x}_i(p_j)}{\mathrm{d}\, p_j}, \quad i, j = 1, \ldots, N. \tag{4.2.2}$$

The corresponding MRA equations (Kholodenko et al., 2002)

$$\sum_{j=1, j\neq i}^{n} r_{ij}^{true} R_{jk}^{true} = R_{ik}^{true} \tag{4.2.3}$$

establish a mathematically exact relationship between the GRCs and LRCs and can be used to extract LRCs from GRCs.

In our *in silico* study we assume that the investigated dynamical system can be described by a true underlying noise-free ODE model $\dot{x} = f(x, p)$, with state variables $x = (x_1, \ldots, x_N) \in \mathbb{R}_+^N$ and parameters $p = (p_1, \ldots, p_N) \in \mathbb{R}_+^N$. The state variable $x_i$ represents the activity of node $i$. The parameters $p_j > 0$, $j = 1, \ldots, N$, are all equal to one in the nominal setting (control experiment), and can be varied to simulate the perturbation experiment affecting the corresponding node $j$. These parameters often affect preserved quantities such as total protein concentrations or production rates. The true LRCs $r_{ij}^{true}, i \neq j$, are obtained by calculating the 'normalized' entries of the Jacobian

matrix at steady-state ($ss$), as described in Kholodenko et al. (2002),

$$r_{ij}^{true} = -\left(\frac{\partial f_i(x,p)}{\partial x_j} \Big/ \frac{\partial f_i(x,p)}{\partial x_i}\right) \cdot \left(\frac{x_j}{x_i}\right)\Bigg|_{ss}, \quad i \neq j, \quad i,j = 1,\ldots,N. \tag{4.2.4}$$

When the underlying ODE system is not known, LRCs can be inferred from concentration measurements via two non-linear transformations. In a first transformation $T_1$, differential GRCs are estimated from the steady states obtained in the control experiment ($\bar{x}_i(p_j) =: \bar{x}_i^0$) and respective steady states in the perturbation experiments ($\bar{x}_i(p_j + \Delta p_j) =: \bar{x}_i^j$),

$$T_1: \quad R_{ij}^{true}\Delta p_j \approx \tilde{R}_{ij} = \frac{\bar{x}_i^j - \bar{x}_i^0}{\frac{1}{2}\left(\bar{x}_i^j + \bar{x}_i^0\right)} = 2 \cdot \frac{\bar{x}_i^j - \bar{x}_i^0}{\bar{x}_i^0 + \bar{x}_i^j}, \quad i,j = 1,\ldots,N, \tag{4.2.5}$$

where we have approximated the derivative in (4.2.2) with finite differences and $\bar{x}_i^0$ with the average of $\bar{x}_i^0$ and $\bar{x}_i^j$. The $N \cdot (N-1)$ LRCs are then obtained via substituting these $\tilde{R}_{ij}$ into equation (4.2.3), which corresponds to solving $N$ linear systems with $N-1$ equations in $N-1$ independent variables each (Kholodenko et al., 2002; Kholodenko and Sontag, 2002),

$$\sum_{j=1, j\neq i}^{n} \tilde{r}_{ij}\tilde{R}_{jk} = \tilde{R}_{ik}, \quad k \neq i; \quad i,k = 1,\ldots,N. \tag{4.2.6}$$

We note here that $\Delta p_k$ cancels out since it appears as a factor on both sides in this system. Due to the approximation (4.2.5), the values $\tilde{r}_{ij}$ obtained in this way are also an approximation of the true LRCs ($r_{ij}^{true}$), and depend in particular on the perturbation strengths. In the following we will always consider $\tilde{R}_{ij}$ directly and thus refer to this measure simply as GRC. A second non-linear transformation, defined as $T_2$, provides a solution for all coefficients $\tilde{r}_{ij}$. As shown in the Supplementary material S2 of Thomaseth et al. (2018) for $N = 3$, we can rewrite equation (4.2.6) as a linear system,

$$\vec{y}(\tilde{R}_{ij}) = A(\tilde{R}_{ij}) \cdot \vec{x}, \tag{4.2.7}$$

in which the vector $\vec{x}$ contains all unknowns $\tilde{r}_{ij}$, while the vector $\vec{y}(\tilde{R}_{ij})$ and the matrix $A(\tilde{R}_{ij})$ are functions of the GRCs. The solution of this linear equation system, assuming that $A$ has linearly independent columns, is given by

$$T_2: \quad \vec{x} = (A^T A)^{-1} A^T \vec{y}. \tag{4.2.8}$$

The variables $\bar{x}_i^j, \tilde{R}_{ij}$ and $\tilde{r}_{ij}$ are assumed to be continuous functions of the perturbation parameters $p_j$. For the following analysis, we refer to the difference

$$\Delta r_{ij}(p_j) = |\tilde{r}_{ij}(p_j) - r_{ij}^{true}|, \quad i \neq j, \quad i,j = 1,\ldots,N \tag{4.2.9}$$

as intrinsic bias which results from the approximations (4.2.5) and (4.2.8).

Our two considered test-bed models for the MAPK and the p53 signalling pathways (see

Appendix 6.9) significantly differ in the courses of $\Delta r_{ij}$ over a large range of perturbation strengths $p_j$ and thus can be considered as complementary examples concerning the approximation quality (4.2.8) and the validity of results in dependence of the perturbation strengths $p_j$.

According to our statistical methodology, following the discussion in Chapter 3, Section 3.2, concentration measurements are described by random variables $\bar{\mathbf{z}}_i^{0,j}, i, j = 1, \ldots, N$, whose distribution is a function of the noise-free steady-state values $\bar{x}_i^{0,j}$ (see Figure 4.1a-b). In the same way as presented in Chapter 3, we consider a realistic error model consisting of a multiplicative and an independent additive part:

$$\bar{\mathbf{z}}_i^{0,j} = \bar{x}_i^{0,j} \cdot \eta + \epsilon, \quad \eta \sim \log \mathcal{N}(0, \sigma_\eta^2), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad i, j = 1, \ldots, N. \tag{4.2.10}$$

The parameters $\sigma_\eta$ and $\sigma_\epsilon$ denote the SDs of the proportional ($\log \eta$) and additive ($\epsilon$) measurement errors, respectively, $\bar{x}_i^{0,j}$ are the simulated noise-free steady-state concentrations in the control (0) and perturbed ($j$) conditions, and $\bar{\mathbf{z}}_i^{0,j}$ the resulting random variables representing the noisy simulated data.

We note here that due to the experimental procedure we often do not directly obtain concentrations but fold changes only, as explained in more detail in Chapter 2. Thus, measurements $\bar{z}_i^0$ and $\bar{z}_i^j$ are realizations of random variables which are proportional to the real absolute concentrations. As shown in Figure 4.1b (left), we assume that measurements refer to signals detected via western blotting which have been normalized to a loading control. Without loss of generality, we neglect the proportionality factor $\alpha$ specific for each blot. In fact, the GRCs calculated via equation (4.2.5), and hence the LRCs, are independent of these factors, as long as the two samples $\bar{z}_i^0$ and $\bar{z}_i^j$ have been quantified in the same blot.

Due to the two non-linear transformations (4.2.5) and (4.2.8), the GRCs and LRCs are also random variables, which we express as $\mathbf{R}_{ij}$ and $\mathbf{r}_{ij}$ (see Figure 4.1a-b). Given measurements of $\mathbf{R}_{ij}$, a solution of equation (4.2.7) is obtained by applying estimation methods. The simplest choice is to use Ordinary (Linear) Least Squares, whose solution has exactly the same form as equation (4.2.8). Changing the method corresponds to changing the operator $T_2$, which remains a non-linear function of the GRCs in all cases.

Since it is impossible to derive the distributions of the LRCs directly from the error model (4.2.10) of the measurements, we applied a Monte Carlo approach in which we used our error model to simulate experimental data and propagated these to respective LRCs via the transformations $T_1$ and $T_2$.

### 4.2.1 MAPK and p53 test-bed models with complementary dynamic behaviours

We used two test-bed examples to investigate noise propagation and impacts of the experimental design on the MRA estimates. The models are similar in that they both consist of three states (nodes) with two positive and one negative interactions (Figure 4.2). However, both models feature very different equations, dynamics, and non-linear properties (see Appendix 6.9 for details), allowing us to judge the generality of our results and to investigate the impacts of moderate and strong non-linearities.

A model of signal transduction of the MAPK pathway upon EGF stimulation is illustrated in Figure 4.2a. It consists of a three-tiered cascade of phosphorylation-dephosphorylation cycles in which pRaf phosphorylates and thereby activates MEK, which then activates ERK, which negatively feeds back to Raf (Kholodenko et al., 2010). Both MEK and ERK require phosphorylation at two sites to become fully active, which is for simplicity assumed to happen in a single reaction step for both proteins (Kholodenko, 2006). Variables $x_1$, $x_2$ and $x_3$ represent protein activities. All reactions are modelled using Michaelis Menten type equations (Appendix Figure 6.15a). This system exhibits moderate non-linearity for the chosen parameters (Appendix Figure 6.15a).

The p53 model (Figure 4.2b) is based on the core signalling system of the DNA damage response (Purvis et al., 2012). Here, ATM activates p53 by phosphorylation and protein stabilisation, p53 activates MDM2 by inducing gene expression, and MDM2 mediates a negative feedback loop to p53 by promoting p53 degradation (Fey et al., 2016). In contrast to the MAPK system, the p53 system exhibits a strong degree of non-linearity (Appendix Figure 6.15b), including three ultra-sensitive Hill type equations for the reaction kinetics (Appendix Figure 6.15b).

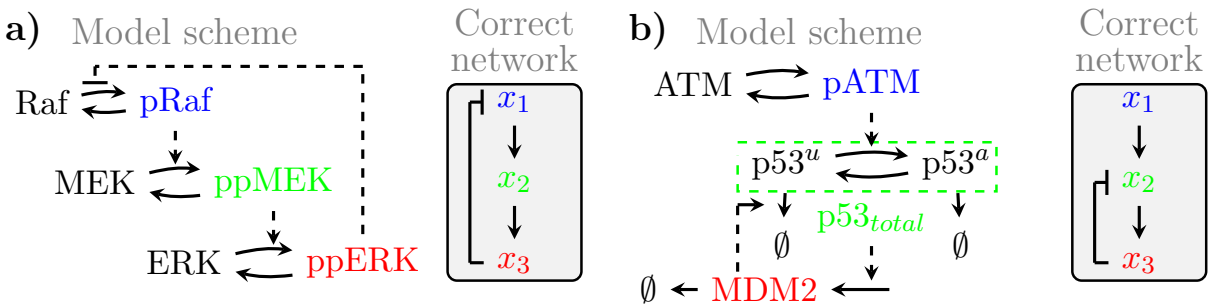


**Fig. 4.2. Two test-bed models.** Shown are the reaction kinetic schemes (left) and the arising network structure (right) for **(a)** the MAPK system; **(b)** the p53 system. The coloured nodes (blue, green, red) indicate the measured species, which define the states of the models.

## 4.3 Results

### 4.3.1 Solving the MRA equations results in heavy-tailed distributions for the estimated LRCs

We started our study by investigating the propagation of noise from the concentration measurements to the estimated LRCs as a basis to deduce strategies for an optimal design of experiments and estimation methods.

Therefore, we simulated the MAPK model with noise parameters that are in a biologically plausible range for western blot data (Schilling et al., 2005; Taylor and Posch, 2014)[1]. Exemplary results are shown in Figure 4.3a, where resulting distributions are illustrated by box blots. While the variability of the resulting distributions of the GRCs is comparable to those of the inputs (Figure 4.3a centre), we observe a much higher variability in the distributions of the LRCs (Figure 4.3a right), which is mainly manifested in the number of outliers and the range covered by them. The complete set of distributions is given in the Supplementary material of Thomaseth et al. (2018) (Figures S2, S3, S4) and shows that these results are representative.

Driven by our analysis, we decided at that point to consider, besides standard measures for statistical dispersion such as interquartile ranges, also the amount of spread of the outliers, which is a measure for the degree of heavy-tailedness of an underlying distribution. Normalization of a signal obtained from a western blot to a signal of a respective control experiment indeed corresponds to a transformation that may result in heavy-tailed distributions (Thomaseth and Radde (2016)). Since the tails of such distributions usually follow a power-law decay, the probability mass in the tails exceeds that of a Gaussian distribution, whose tails decay exponentially. As a consequence, samples from heavy-tailed distributions will contain more outliers which are spread over a larger range. A characteristic feature of heavy-tailed distributions is the fact that some or all moments do not exist. This severely impedes network reconstruction in our framework, since empirical estimators of moments are unstable due to the high occurrence of outliers. Empirical moments like the sample mean, the sample variance, or skewness and kurtosis, which are standard measures of asymmetry and tail-heaviness, do not provide meaningful estimates under these circumstances.

Thus, we decided to evaluate left and right medcouples (LMC and RMC) (Brys et al., 2006) as suitable measures of left and right tail weights. The medcouple (MC) function (see Appendix 6.10) was proposed as an efficient measure of the asymmetry of a univariate continuous distribution alternative to the classical skewness estimator (Brys et al., 2004). The medcouple applied to one single side of the distribution leads to LMC and RMC,

[1] All simulation results presented in this chapter were run with the software MATLAB.

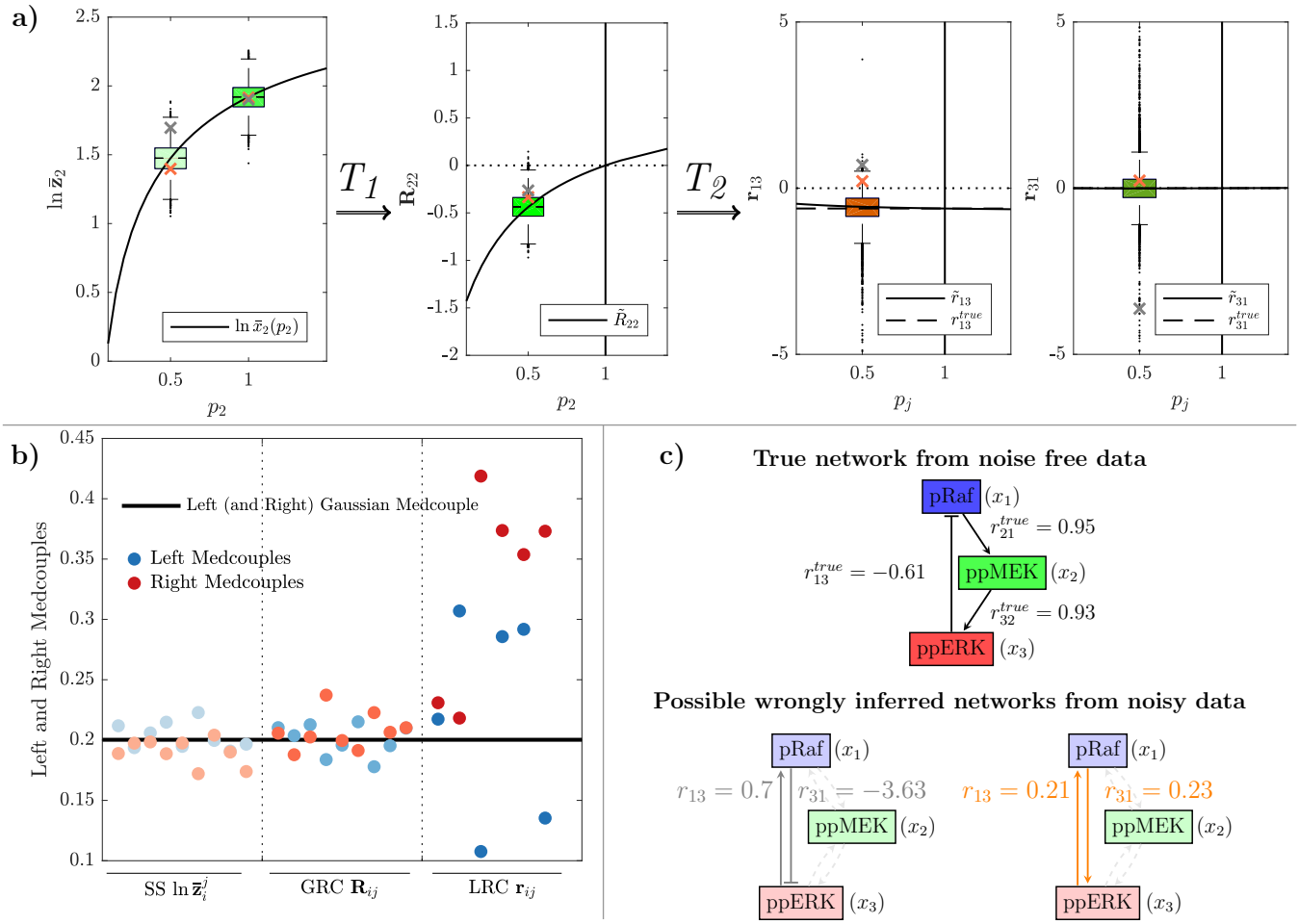


**Fig. 4.3. Non-linear propagation of measurement error in the MAPK test-bed model. (a)** The curves represent the dependencies of steady states (left), GRCs (centre) and LRCs (right) over the changing parameters $p_j$. Exemplary, on the left we show the distributions (boxplots) obtained from sampled noisy realizations of the steady-state $\bar{\mathbf{z}}_2$ in the control experiment ($p_2 = 1$) and in the 50% knockdown experiment of node 2 ($p_2 = 0.5$). We generated $n = 10,000$ realizations via Monte Carlo simulations from the noise model (4.2.10) with parameters $\sigma_\eta = 0.1$ and $\sigma_\epsilon = 0.2$. The ODE model with numerical values used for simulations is given in the Appendix 6.9. The variability in the measured steady states (left) translates into variability of the calculated GRCs (centre), which then translates into variability of the LRCs (right). Two sample points have been indicated by an orange and a grey cross and tracked during the transformations to illustrate consequences for network inference from individual samples. **(b)** Propagation of LMC and RMC values during the two-step transformation process reveals that heavy-tailedness is mainly introduced by the transformation $T_2$. Numerical values are given in the Supplementary Table S1 in Thomaseth et al. (2018). **(c)** True network structure of the MAPK test-bed model as obtained via equation (4.2.4). As a comparison, also the two realizations of the LRCs $\mathbf{r}_{13}$ and $\mathbf{r}_{31}$ that result from the two tracked orange and grey sample points are shown.

which are monotonically increasing functions of tail-heaviness. They are robust to outliers, since they only depend on quantiles and hence are suitable for heavy-tailed distributions.

LMC and RMC values are put into context by comparison with the respective values for a standard Gaussian and Cauchy distribution, which are 0.2 and 0.5, respectively.

LMC and RMC values for concentration measurements, GRCs and LRCs are depicted in Figure 4.3b. LMCs and RMCs for the distributions of the measurement data and of the GRCs are comparable to those of a Gaussian distribution. Respective values for the distributions of the estimated LRCs are considerably larger, indicating that heavy-tailedness is mainly introduced by the transformation $T_2$. This increase might have severe consequences for network reconstruction, since it distorts estimation of moments of the LRCs such as the mean and the variance from samples. Evaluation of LMC and RMC values for the p53 test-bed model reveals similar results (Figure S5 in the Supplementary material of Thomaseth et al. (2018)). Interestingly, MRA does not markedly affect the interquartile range (IQR) over the two transformations, which is a frequently used bulk-measure of variability (Figure S6 in the Supplementary material of Thomaseth et al. (2018)).

We conclude that MRA amplifies the variability of the measurement noise in terms of degree of heavy-tailedness, while the IQR is not as much affected. Since heavy-tailedness is directly related to the occurrence of samples in the tails, which appear as outliers in the box plots, this impedes network reconstruction, as illustrated with two sample points indicated with orange and grey crosses and respective wrongly inferred network structures (Figure 4.3c).

The question arises how we can optimize network reconstruction by influencing the distribution of the LRCs via experimental design and/or estimation procedures. In a first step we analyse how to best design the experiments regarding the choice of the perturbation strengths and the control strategy and subsequently investigate how to best handle multiple replicates.

### 4.3.2 Large perturbations tend to improve the inference of pairwise node interactions

Since the GRCs and LRCs are defined as derivatives, a precise approximation via finite differences theoretically requires infinitesimal small perturbations, which is not feasible in practice. Moreover, noise deteriorates estimation of derivatives particularly from small differences. The question arises whether we are able to define perturbation strengths that constitute a good trade-off. For the MAPK test-bed model we observe that the noise-free approximated solution for the LRCs $\tilde{r}_{ij}$ is robust over a large range of perturbation parameters $p_j$ and does not deviate much from the corresponding true value $r_{ij}^{true}$ (see right of Figure 4.3a and Figure S4 in the Supplementary material of Thomaseth et al. (2018)). This is different for the p53 test-bed model (Figure S9 in the Supplementary material of Thomaseth et al. (2018)) and might also not be the case for other systems, which we usually

don't know a priori.

In order to answer our question, we compare the variability of the estimated LRCs resulting from different perturbation strengths. Therefore, we consider three knockdown experiments with downregulation of the 80%, 50% and 25% of the total protein concentrations with respect to the control experiment, and one overexpression experiment with 150% of total protein concentrations, resulting in a set of values for the perturbation strengths $p_j \in \{0.2, 0.5, 0.75, 1.5\}$.

As can be seen in Figure 4.4a and Figure S10 in the Supplementary material of Thomaseth et al. (2018), the distributions of the estimated coefficients differ significantly in the four scenarios. The spread of the estimated coefficients is smallest for $p_j = 0.2$ and rapidly increases with decreasing perturbation strength, i.e. when $p_j$ approaches one. The spread of the overexpression experiment is comparable to the 25% knockdown experiment, which is probably a result of the fact that we are in the saturated regime. We also observe a small and perturbation-dependent bias in the empirical estimate of the medians of all distributions.

To investigate the influence of the perturbation strength on accuracy and precision of the estimation more comprehensively, we collected values of the bias of the median and of the LMCs and RMCs for the 80% and the 25% knockdown experiments. Results are shown in Figure 4.4b, where we have also visualized different noise levels $\sigma_\eta$ and $\sigma_\epsilon$ with different colours and corresponding different shades. We observe a 'bias-spread' trade-off between large and small perturbations. A low bias and a low LMC can only be obtained with large perturbations (large dots), while small perturbations lead to higher LMC values.

Increasing noise levels affect the bias markedly only for the small perturbation (small dots), which is true for all coefficients (Figure S11 in the Supplementary material of Thomaseth et al. (2018)). The influence of increasing noise levels on LMC and RMC values is visible but moderate in the 80% knockdown experiment, while in case of 25% knockdown a marked effect can only be seen for very small multiplicative noise $\sigma_\eta$ (Figure 12 in the Supplementary material of Thomaseth et al. (2018)). Intriguingly, in most of the cases these quantities behave non-monotonically with respect to noise. Increasing noise does not necessarily imply larger bias or medcouples, which is probably due to the non-linear transformations $T_1$ and $T_2$. For the p53 test-bed model we observe similar trends (Figures S13, S14, S15, S16 in the Supplementary material of Thomaseth et al. (2018)), even though we observe a large bias of the median also for the large perturbation experiments here. From this analysis we conclude that larger perturbations are generally preferable, since they reduce the risk to infer erroneous network interactions.

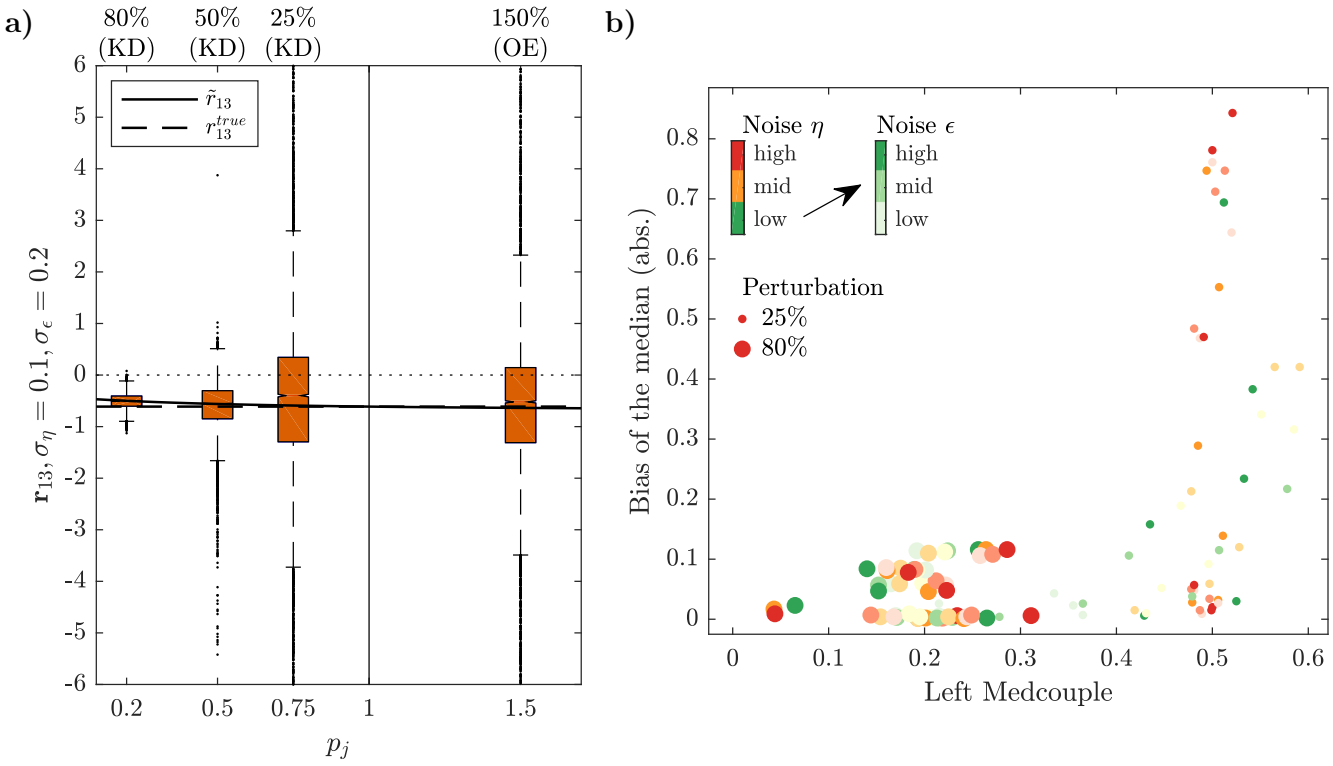


**Fig. 4.4. Effects of different perturbation strengths on network reconstruction.** **(a)** Boxplots of the estimated LRC $\mathbf{r}_{13}$ of the MAPK test-bed model, describing the negative feedback from ERK to Raf, for different perturbation strengths: 80%, 50%, 25% knockdowns (KD) and 150% overexpression (OE) of the total protein concentrations. **(b)** Absolute values of the bias of the median versus LMC values for the entire set of LRC values obtained with large (80%) or small (25%) knockdown strengths of the total protein concentrations. These statistics are given for different noise levels $\sigma_\eta \in \{0.05 \text{ (green)}, 0.1 \text{ (yellow)}, 0.2 \text{ (red)}\}$ and $\sigma_\epsilon \in \{0.1, 0.2, 0.5\}$ (indicated by increasing darkness).

### 4.3.3 A simple control strategy is sufficient for the estimation of the LRCs

The second component of the experimental design under investigation is the control strategy (CS). Here we compare a single control for a node for all three perturbations (Figure 4.5a left) versus individual controls for each perturbation (Figure 4.5a right). The steady-state variable $\bar{x}_i^0$ of the control experiment appears in the GRCs $\tilde{R}_{ij}$ of all perturbation experiments $j = 1, \ldots, N$ (equation (4.2.5)). Simulating the first control strategy thus translates into using the same realization of the random variable $\bar{\mathbf{z}}_i^0$ to calculate the realizations $R_{ij}, j = 1, \ldots, N$, for fixed $i$, and results in block-wise positive correlations between the GRCs, as can be seen in the $\mathbf{R}_{ij}$ scatter plot matrix in Figure 4.5a (left). These correlations disappear when performing multiple independent controls $\bar{z}_i^0$ for each perturbation experiment $j$ (Figure 4.5a right).

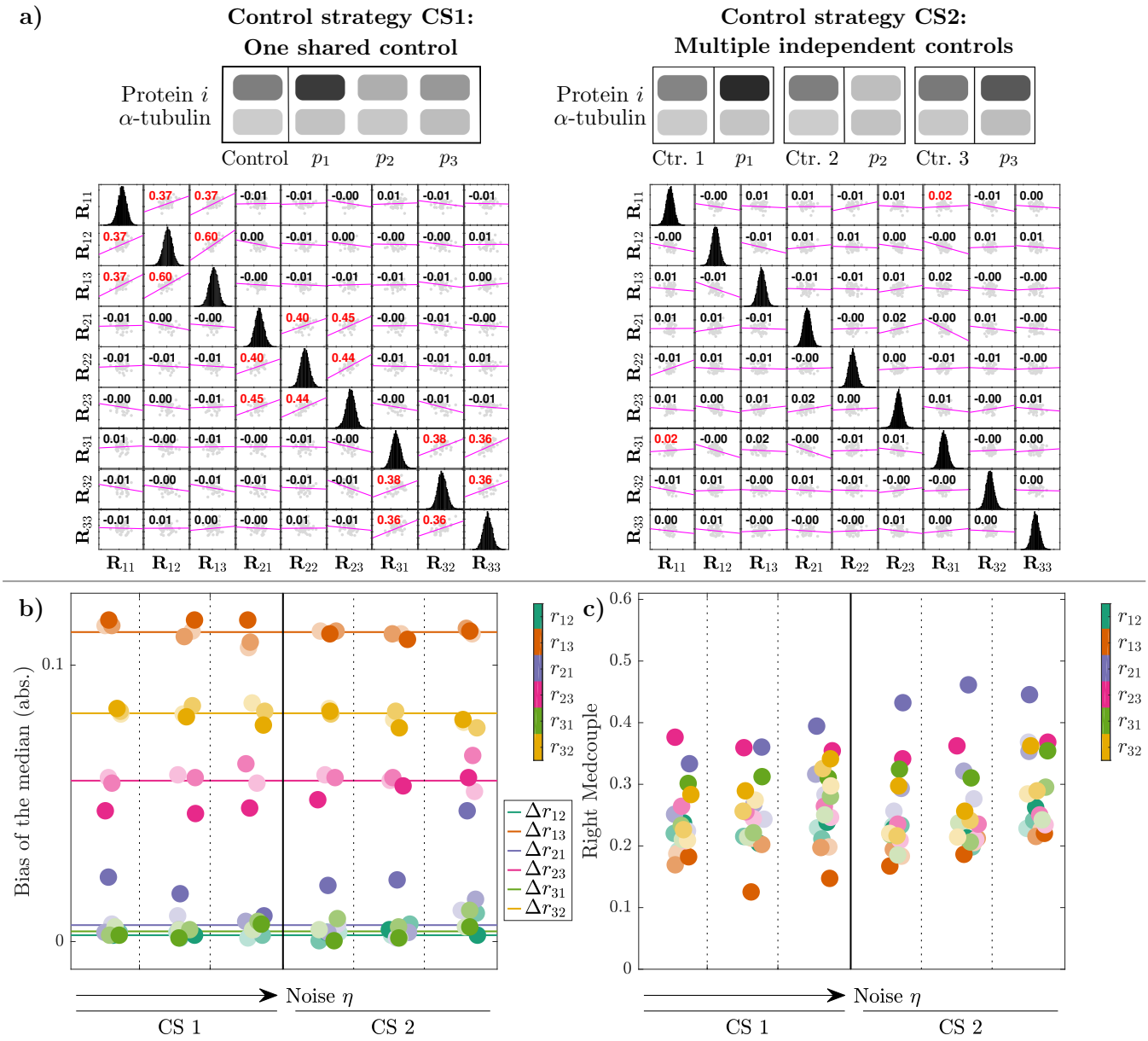


**Fig. 4.5. Effects of two alternative control strategies. (a)** The first strategy (CS1) considers one single control realization for the calculation of all samples $\mathbf{R}_{ij}$ with $j = 1, 2, 3$, while the second strategy (CS2) performs independent control experiments for each perturbation experiment. Corresponding correlations can be seen in the $\mathbf{R}_{ij}$ scatter plot matrices. **(b)** Absolute values of the bias of the medians of the resulting LRCs $\mathbf{r}_{ij}$ for both control strategies CS1 (left) and CS2 (right) in dependence of different noise levels $\sigma_\eta \in \{0.05, 0.1, 0.2\}$ and $\sigma_\epsilon \in \{0.1, 0.2, 0.5\}$. For every value of $\sigma_\eta$, i.e. for each specific column, the three shades of the LRCs correspond to the three (increasing) values of the SD of the additive noise $\sigma_\epsilon$. Lines indicate intrinsic bias values for each LRC. **(c)** Same illustration for the RMC values.

The choice of the control strategy determines the correlation among the coefficients $\mathbf{R}_{ij}$ with the same index $i$, but it is unclear whether it also has a marked effect on the LRCs $\mathbf{r}_{ij}$ and thus on network inference. In order to resolve this issue, we used the bias of the

medians, the IQR, LMCs and RMCs as statistical measures of the distributions of the LRCs to compare the two control strategies. Since large perturbations have already turned out to be advantageous for MRA analysis, we simulated an 80% knockdown experiment and additionally also analysed the effect of increasing noise levels. As can be seen in Figure 4.5b-c, we do not detect significant differences between the two control strategies with respect to the 'bias-spread' trade-off. In fact, bias and RMCs behave similarly in the two cases. As before, we do not see a marked effect of increasing noise levels on the LRC distribution measures. These observations also generalize to the IQRs and the LMCs (Figure S18 in the Supplementary material of Thomaseth et al. (2018)).

The horizontal lines shown in Figure 4.5b represent the absolute values of the differences between the true LRCs and the LRCs resulting from the noise-free approximation, defined as $\Delta r_{ij}$ (equation (4.2.9)). For realistic noise levels, as used here, the bias of the medians is centred around this corresponding intrinsic bias, showing that the main contribution to the bias is caused by the approximation (4.2.5) rather than by the measurement noise. As before, there is no clear monotonic relation visible between the considered statistics and the levels of additive and multiplicative noise, respectively. The p53 test-bed model behaves very similar in this analysis (Figures S19 and S20 in the Supplementary material of Thomaseth et al. (2018)).

Taken together, since we could not observe marked differences of the LRC statistics between the two control strategies in both models, we advice experimenters to use the first control strategy of taking a single control measurement for a node for all corresponding perturbations, since this requires less samples.

### 4.3.4 Using MRA with replicate mean values tends to outperform linear regression techniques

Generally, perturbation data contain several replicates of the same experiment. This raises the question of how to best handle these replicates during the MRA workflow. One solution is to calculate the mean over the replicates. Another, is the use of linear regression, for which several techniques have been suggested. The most common choice is to solve equation (4.2.7) by applying a least squares method, like Ordinary Least Squares (OLS) and Total Least Squares (TLS) (Andrec et al., 2005) (see Appendix 6.11). But whether regression, and if so which, is better than using the mean over all replicates remains unclear.

Therefore, we aim to solve the question about which estimation method, combined with the proper experimental design and data normalization, allows the best results in terms of accuracy, precision and robustness of the LRCs estimates.

We compare results obtained with three replicates, which is the typically required number in many biological studies. In our simulations we mimic replicates by drawing independent realizations $\bar{z}_i^{0,j}$, providing different realizations of the GRCs $R_{ij}$. We considered the methods of taking the mean over the three obtained GRCs replicates and solving the linear regression problem (4.2.7), or determining GRC values for individual replicates and then applying either OLS or TLS from noisy values $R_{ij}$, delivering one estimate of the LRCs $r_{ij}, i, j = 1, \ldots, N$. Moreover, we consider yet another experimental approach, in which we take multiple sample data not by repeating the same perturbation experiment but by varying the perturbation strengths $p_j, j = 1, 2, 3$. Our choice is to mix three realizations obtained using three different knockdown strengths: 80%, 50% and 25% KD of the total protein concentrations. The results of our analysis are summarized in Figure 4.6.

These results confirm that the experimental design with the large perturbation is superior compared to small perturbations or a combination of different perturbation strengths (see also Figures S21 and S22 in the Supplementary material of Thomaseth et al. (2018)). The considered measures for dispersion (LMC, RMC and IQR) are low and robust to noise for all three estimation approaches.

Interestingly, the mixture approach also delivers good results in terms of 'bias-spread' trade-off (right part of Figure 4.6, Figures S21 and S22 in the Supplementary material of Thomaseth et al. (2018)). As before, the OLS method results in a larger bias, but the dispersion measures are more sensitive to increasing noise if using TLS.

We can confirm that the experimental approach with the small perturbation strength delivers unsatisfactory results, leading to a high risk to reconstruct an erroneous network structure. Compared to the other two experimental designs, the bias is much larger and sensitive to noise with all three estimation methods: This holds especially true for the three non-zero coefficients $r_{21}$, $r_{32}$ and $r_{13}$ (central part of Figure 4.6b and Figure S22 in the Supplementary material of Thomaseth et al. (2018)). The measures for dispersion are low for all coefficients only for very low noise and if using OLS.

Summarizing our results, we obtained the best estimation results in terms of accuracy, precision and robustness to noise by performing large perturbations and a simple control strategy. In terms of efficiency we recommend to use the simplest estimation method, which means to solve the regression problem (4.2.7) with the GRC means. In comparison, the mixture approach seems to be suboptimal in terms of 'bias-spread' trade-off, but it might be beneficial for systems with higher non-linearities, as discussed later in Subsection 4.3.6.

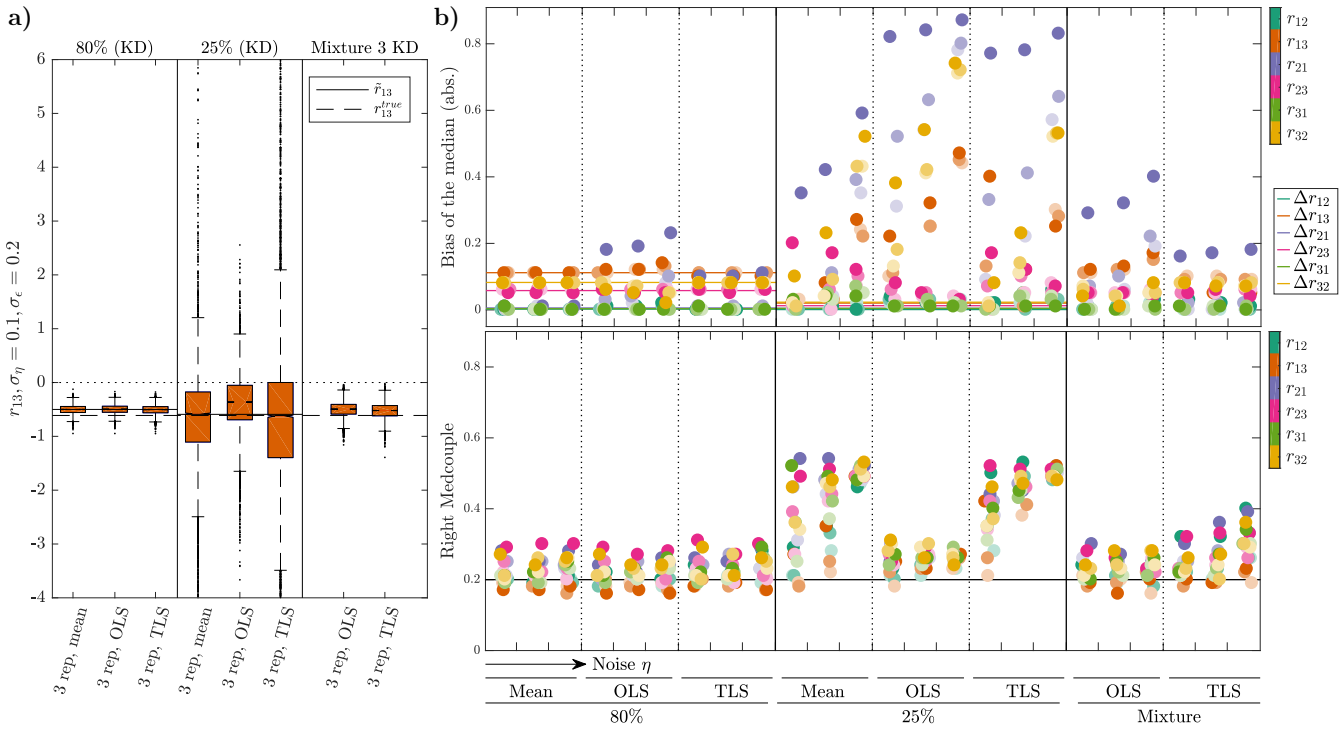


**Fig. 4.6. Effects of different estimation methods for the linear regression problem with multiple replicates. (a)** Boxplots of the estimated LRC $\mathbf{r}_{13}$ of the MAPK model for two different perturbation strengths and different strategies to handle three replicates. First, the LRCs are calculated by taking the mean values of the GRCs. Second, GRC replicates are taken individually into account and LRCs are obtained by solving OLS or TLS, respectively. The third column illustrates results from a mixture of measurements from three knockdown experiments with different perturbation strengths. **(b)** Absolute bias values of the estimated medians and RMC values for all LRCs and increasing levels of multiplicative and additive noise, $\sigma_\eta \in \{0.05, 0.1, 0.2\}$ and $\sigma_\epsilon \in \{0.1, 0.2, 0.5\}$. For every value of $\sigma_\eta$, i.e. for a specific column, the three shades of the LRCs correspond to three (increasing) values of the SD of the additive noise $\sigma_\epsilon$.

### 4.3.5 Replicates increase precision, but not accuracy

The choice of the number of replicates is another important question for experimental design because of the trade-off between the experimental effort and cost, and the quality of the inferred results (Blainey et al., 2014). We address this issue by investigating how much the estimation of the LRCs is improved by increasing the number of replicates. We compare results obtained with one, three and six replicates. For each Monte Carlo run, we proceed by taking the mean value of these GRCs to further calculate one realization $r_{ij}$ of the LRCs, which we have seen to be the most efficient estimation method, combined with large perturbations and the simple control strategy.

Results are depicted in Figure 4.7. As expected, the precision of the estimation increases with the number of replicates, for both test-bed models. This manifests in a decrease of the

considered measures for statistical dispersion, which are RMC and LMC values and the IQR, for all coefficients and noise levels and both test-bed models (Figure 4.7a top, Figures S26 and S28 in the Supplementary material of Thomaseth et al. (2018)). In particular, RMC and LMC values converge to the value 0.2 of the standard Gaussian distribution. This effect is robust against increased multiplicative noise levels $\eta$.

In contrast, the biases in the medians are neither much affected by the number of replicates nor by the level of multiplicative noise, as can be seen in Figure 4.7a (bottom) and Figure S28 in the Supplementary material of Thomaseth et al. (2018). In some cases increased additive noise $\epsilon$ (indicated by a darker shade of the coloured dots) leads to a larger bias, but not in a monotonic manner. As before, the medians rather coincide with the noise-free approximated values $\tilde{r}_{ij}$ (see also Figures S25, S27 c), whose deviations from the true values result from the choice of a large perturbation, showing that the bias in the medians is again dominated by the error of the approximation (4.2.5).

Summarizing, increasing the number of replicates reduces the dispersion of the distribution and therefore increases precision, but the bias cannot be eliminated, which restricts the accuracy of the estimates. We consider three replicates to be a good bias-spread trade-off, since all RMC values decrease below 0.3 when going from a single measurement to three replicates, while the decrease is much less pronounced when going from three to six replicates. Thus we recommend to use at least three replicates, and to include more depending on how much experimental effort is acceptable.

### 4.3.6 Non-linearity induces bias, but large perturbations are still required for precision

In the MAPK model, the steady states show an approximately linear behaviour in dependence of the perturbation strengths in all cases (see Figure S1a in the Supplementary material of Thomaseth et al. (2018)), suggesting that the linear approximations (4.2.5) and (4.2.8) do not induce unduly large errors even for large perturbations. This was confirmed by our simulation results. When applying the MRA in practice, however, the course of the steady states of the system for varying perturbation strengths is not known, and it could also be highly non-linear. Do our recommendations and guidelines for an optimal performance of network reconstruction via MRA still hold true for such cases? In order to address this question, we used the p53 test-bed model as an example of a system whose steady states are non-linear functions of perturbation strengths (see Figure S1b in the Supplementary material of Thomaseth et al. (2018)). In this case the approximated LRCs $\tilde{r}_{ij}$ are sensitive to the choice of the perturbation strength and it is not clear a priori whether they are a good approximation of the true values $r_{ij}^{true}$ (compare Figures S4 and S9 of the Supplementary material of Thomaseth et al. (2018)).

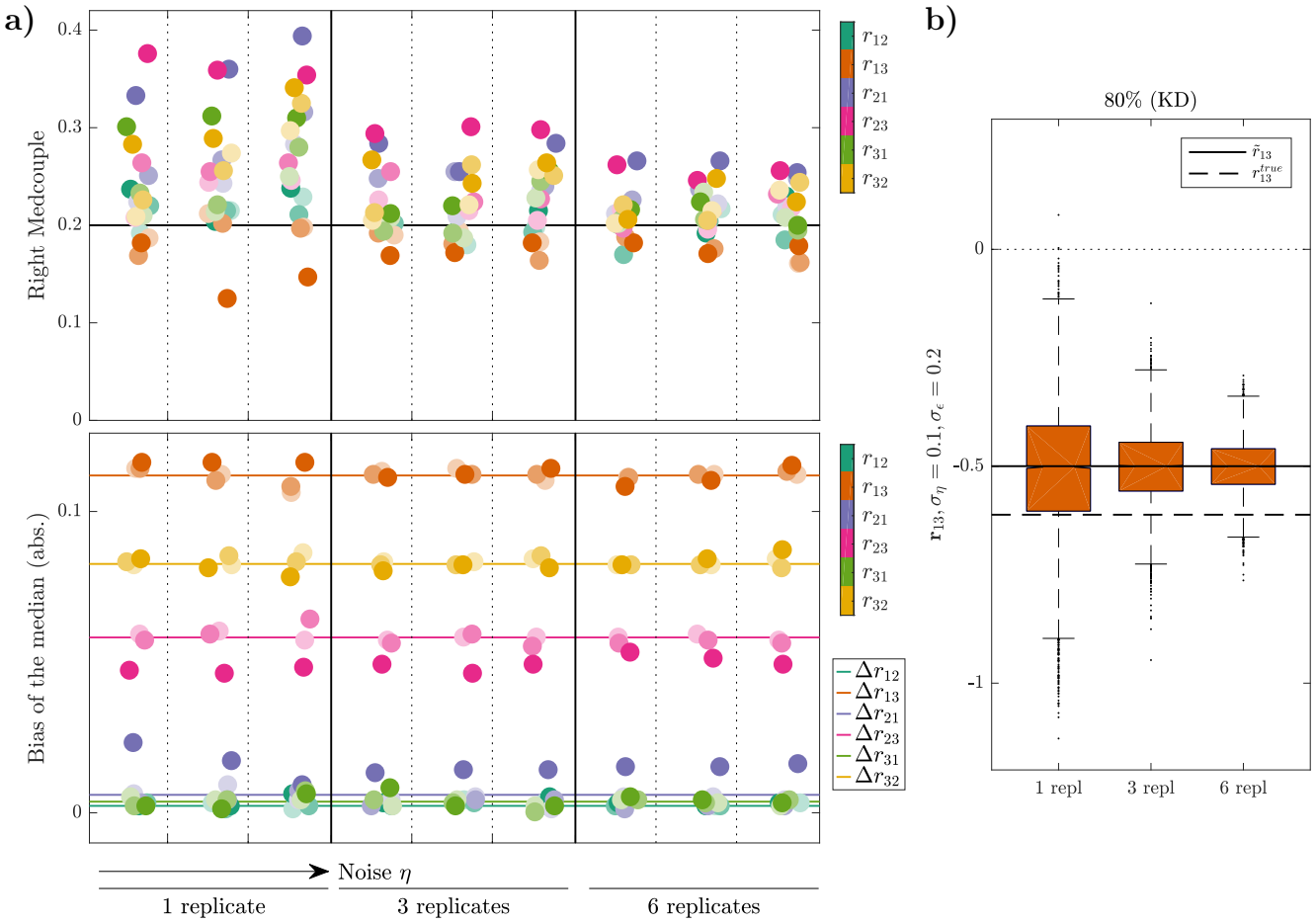


**Fig. 4.7. Effects of different numbers of replicates on network reconstruction.** **(a)** RMCs and absolute bias values of the medians of the estimated distributions for the LRCs of the MAPK model for one, three and six replicates and different noise levels in the 80% KD perturbation experiments. Noise levels have been set to $\sigma_\eta = \{0.05, 0.1, 0.2\}$ (different columns) and $\sigma_\epsilon = \{0.1, 0.2, 0.5\}$ (coded in different shades). The black line indicates the RMC value for a normal distribution and the coloured lines show the intrinsic bias values for each LRC. **(b)** Exemplary boxplots of the LRC $\mathbf{r}_{13}$ for different numbers of replicates.

We applied our MRA workflow to this test-bed model and performed the same analysis as before with the MAPK model. Summarizing, the results show that the most critical part is indeed the appearance of a large bias in the median of the distributions of the estimated $\mathbf{r}_{ij}$ if applying large perturbation experiments (see Figures S23 (left parts) and S24a in the Supplementary material of Thomaseth et al. (2018)). This effect is related to the intrinsic bias $\Delta r_{ij}(p_j)$ and cannot be reduced by an increase in the number of replicates (see Figures S28a in the Supplementary material of Thomaseth et al. (2018)).

Nevertheless, the goal is to estimate the correct network structure, and therefore it is important to minimize the dispersion of the distributions of the estimated $\mathbf{r}_{ij}$. This holds especially if the intrinsic bias is significant for some of the LRCs, which is the case in the

p53 example (Figure S23 in the Supplementary material of Thomaseth et al. (2018)). In such cases it is necessary that the approximated LRCs $\tilde{r}_{ij}$ have the same sign as the true values, leading to qualitatively correctly estimated interactions. The trend of the spread of the estimated distributions shows that in general the lowest dispersion is still obtained with the largest perturbation experiment, in a similar way for all three computational approaches (see Figures S24b-d in the Supplementary material of Thomaseth et al. (2018)). In all these cases this behaviour is robust to increasing noise levels.

From these results we conclude that, due to the noise sensitivity, larger perturbations are generally still preferable, even for highly non-linear systems, since they reduce the risk to infer erroneous network interactions.

### 4.3.7 Performance evaluation on the level of discrete network interactions corroborates our quantitative results

So far, we have investigated the influence of different experimental designs, estimation methods and noise levels on the statistical properties of the estimated LRCs. We have in particular focused on the bias of the median and on LMC and RMC values as measures for accuracy and precision of the individual estimates. In a final analysis step we transfer these results onto network inference, where the set of inferred LRCs is used to decide upon the network structure. The simplest way to do this is to arrange all LRCs according to their absolute value and to define a threshold for an interaction to be present or not. Sensitivity and specificity can then be calculated for an inferred network by a comparison with the true or a reference network. Doing this with varying threshold values, the Area Under the Curve (AUC) value is then an aggregated measure for the overall performance of the inference method independent of the threshold parameter. For such an analysis, it is not sufficient to look at each LRC separately. Here we applied an assessment method proposed in Bansal et al. (2006), which is similar to a receiver-operator analysis, but also takes the signs of the inferred interactions into account. Depending on the percentage of correctly identified interactions, a normalized measure for the fit quality is assigned to an inferred network structure (see Figure 4.1c), which is 0 in the worst and 1 in the best case. This overall measure for fit quality was determined for the different scenarios considered before and the distribution of this measure was investigated by sampling $n = 10,000$ network structures for each setting.

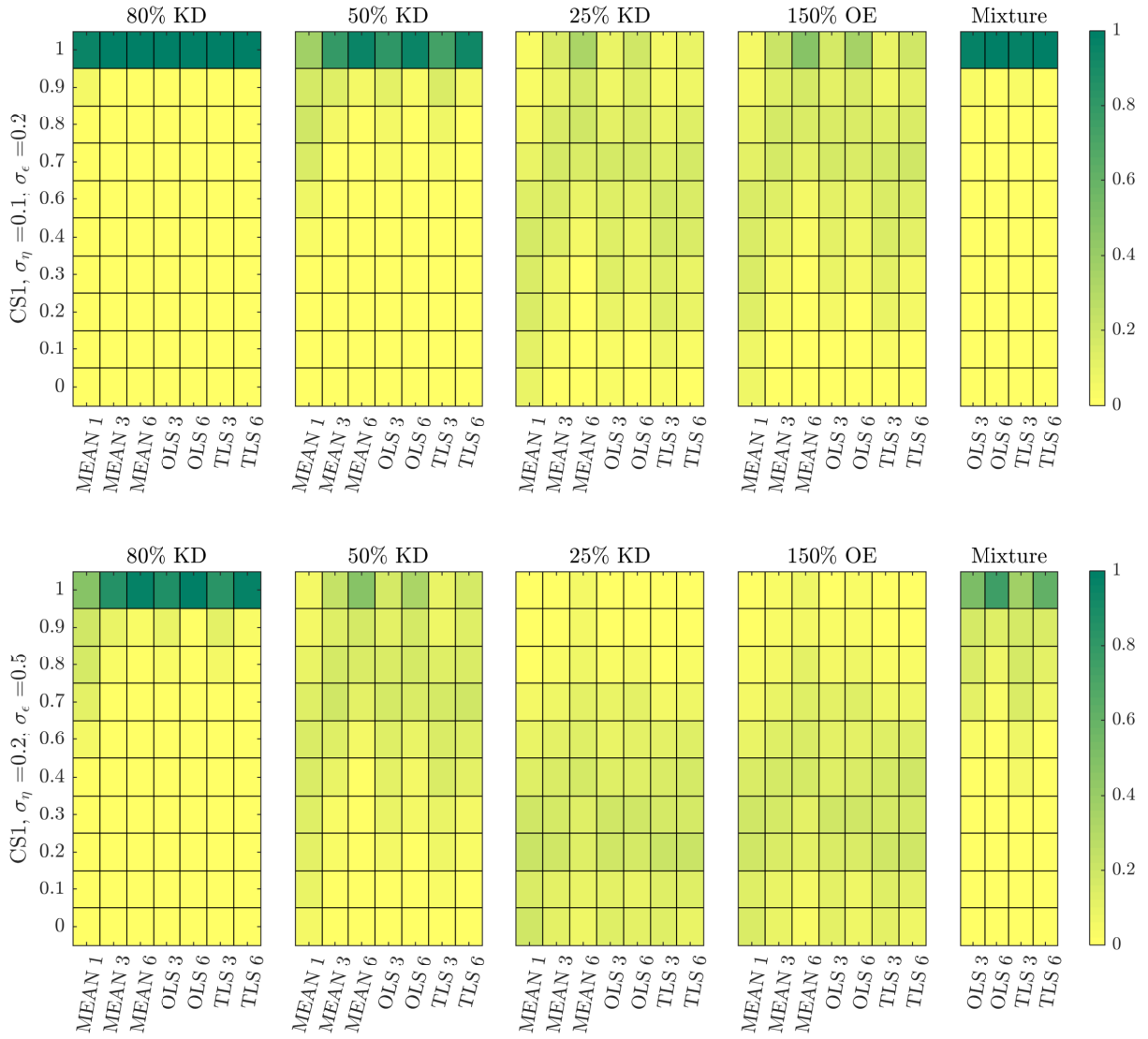


**Fig. 4.8. Performance evaluation of all MRA settings for network inference obtained with the MAPK test-bed model.** Empirical distributions of fit quality of the inferred networks for different experimental designs and computational strategies for intermediate (top row) and high (bottom row) noise levels.

Results are shown in Figure 4.8. Here, colour-coded empirical probability distributions of the discrete set of fit-quality values are shown for different settings. The first and second row depict results for intermediate and high noise levels, respectively. Different computational strategies and different perturbation strengths are compared. It can be seen that network inference works quite well in the 80% KD experiments for both noise levels and almost independent from the number of replicates and from the strategy to handle replicates. For intermediate noise levels, also the mixture method and the 50% KD perform very well, but are more sensitive to increasing noise levels. Both 25% KD and 150% OE perform worse in all scenarios. It can also be seen that results when considering replicates are not

markedly different across almost all scenarios if averaging over the GRCs or when using OLS or TLS. We also compared these statistics across the two control strategies CS1 and CS2 (e.g. compare Figures S29a-b and S29c-d of the Supplementary material of Thomaseth et al. (2018)), which shows that the simple control strategy (CS1) is sufficient and there is no need to evaluate multiple independent control samples for each perturbation experiment. Similar results were obtained for the p53 model (Figure S30 in the Supplementary material of Thomaseth et al. (2018)). Taken together, these results further confirm the conclusions drawn from the quantitative analyses in the previous sections: Firstly, due to noise, large perturbation are preferable, even for systems with a high degree of non-linearity. Secondly, it suffices to use a simple experimental strategy with one unperturbed control as reference for all perturbed conditions.

## 4.4 Summary and discussion

In this chapter, we went through a comprehensive analysis of the effects of different experimental and estimation approaches for MRA on the goodness of network inference from noisy data, in terms of accuracy, precision and robustness. Our results led to some interesting findings. First, Monte Carlo simulations of concentration measurements with a realistic noise model for western blot data clearly show a strong increase of heavy-tailedness, quantified in terms of LMC and RMC values, in the transformation from the GRCs to the LRCs, while respective values for concentration measurements and GRCs are quite similar (Figure 4.3). This is a very relevant result, since heavy-tailedness deteriorates estimation of moments from samples, inducing as consequence a high risk of wrong outcomes for the network inference problem. In extreme cases, i.e. when certain moments are not defined, a stable estimation is not possible, even for large sample sizes. At least, this implies that concentrations and GRCs can be estimated more accurately from concentration measurements than LRCs. Second, for both test-bed models large perturbations are more favourable than smaller ones. Estimation of LRCs and hence network inference is much more accurate when using large perturbations. This is a non-trivial result, since estimation of the LRCs is done via a finite difference approximation of the GRCs in the MRA workflow, for which small differences would be beneficial in the absence of noise, since large differences imply an intrinsic bias. In particular, Figure 4.4 shows a clear clustering of inferred LRCs according to the perturbation strengths: Results from 80% knockdown simulations show small biases and small LMC values, while 25% knockdowns show a much higher spread of bias values and consistently high LMC values. This leads to the clear advice to use large perturbations in the MRA workflow, even when the underlying model system features a considerable amount of non-linearity. Furthermore, regarding experimental design in

the number of controls, our results indicate that a single control for different perturbation experiments, as often applied in practice, is sufficient (Figure 4.5). While a single control causes correlations between GRCs, GRC marginals are not much affected. In particular, there is not much difference in the bias and RMC values of inferred LRC values among the two control strategies. The bias values are dominated by the intrinsic bias, and this is true for all tested noise levels. Regarding the required number of measurements and the estimation method, we advice to use the mean of at least three replicates. The spread of inferred LRCs decreases monotonically with the number of replicates, while the bias of the median is dominated by the intrinsic bias (Figures 4.6 and 4.7). Finally, our conclusions also hold true for the overall network inference problem, as evaluated in this study by a normalized quality measure for a classification problem (Figure 4.8). Our results in particular show that learning the network topology is possible with very high accuracy also for high noise levels in our setting with the 80% knockdown experiments and few replicates.

As pointed out in the introduction, the effect of noise and variability in the data used for MRA network inference had already been an issue in earlier studies (Andrec et al., 2005; Santos et al., 2007). In a later work (Santra et al., 2013), the authors developed an advanced version of MRA, combining it with a Bayesian model selection algorithm, relaxing also the restriction of required number of perturbation experiments to equal the number of nodes of the network. However, none of these are comparable in considering propagation of noise from concentration measurements via GRCs and LRCs to network topology inference in a consistent stochastic framework with realistic noise assumptions. These studies also use Monte Carlo techniques, but start with i.i.d. normal distributions directly on the GRC values, and also completely neglect the effects of heavy-tailedness. They are also lacking concrete recommendations for experimental design and computational methodology regarding MRA based network inference.

As with all inference methods, our methodology has some limitations. The MRA framework itself assumes a continuous functional dependence between perturbation parameters and steady states of the system. This excludes for instance systems which exhibit bifurcations, as they appear for example in positive feedback systems which exhibit multi-stability. For those systems, the theory only holds as long as the perturbation does not induce a switch of the system to a different fixed point branch. It might be difficult to decide whether this is indeed the case in real settings, where the underlying dynamical system is not known. Furthermore, there might be potential for improvement regarding methodology to solve the regression problem to calculate the LRCs from the GRCs. Methods like feasible generalized least squares or MLE might be beneficial in this respect. Finally, evaluation of our findings and recommendations in a setting with real experimental data is an open issue for the future.

# 5 Conclusion

In this chapter, we recapitulate the central theme of the entire thesis and summarize chapter-wise findings. Additionally, we debate some critical aspects, make general remarks, and discuss few potential prospective extensions of the research presented in this thesis.

## 5.1 Summary and discussion

A general *inference problem* requires experimentally measured datasets as input information for the estimation of the unknown parameters of a mathematical model, which is used to describe the physical system under investigation.

Particularly in the field of *Systems Biology*, starting from experimental observations, biologists and theoreticians cooperate to unravel complex phenomena by means of mathematical modelling and simulation frameworks. Like any experimentally measured dataset, biological data suffer from noise sources due to intrinsic variability and measurement techniques. In this thesis we developed a *statistical framework* to investigate how noise propagates from the experimental data over the calculations of an inference problem, eventually affecting the uncertainty of the estimated parameters. This understanding is essential for a fundamental analysis of data-driven inference problems, since it contributes towards the optimization of the experimental design, the development of robust theoretical and simulation frameworks and finally the achievement of reliable model predictions.

Dealing with the stochastic nature of biological data is a well-established procedure for inference problems in Systems Biology (Cho and Wolkenhauer (2003); Raue et al. (2013), Thomaseth et al. (2017)). Nevertheless, a comprehensive analysis of the mechanisms of noise propagation in terms of transformations of statistical distributions from the input data to the estimated outputs was still missing in the literature.

This thesis focuses in particular on investigating the role of *data normalization* with respect to noise propagation, a post-processing step required for some common experimental techniques like western blotting (Degasperi et al., 2014; Taylor and Posch, 2014). The work presented in this thesis was inspired by published results on the effects of data normalization on hypothesis testing (Degasperi et al., 2014) and sensitivity analysis (Kirch et al. (2016)). Anyway, there are no studies in the literature taking into account the change

of the statistical properties of the normalized data with respect to the raw data and the possible resulting effects on the uncertainty of inference results.

Our analysis begins with Chapter 2, where we investigate the impact of noise transformation due to normalization on **statistical inference**, meaning the estimation of the unknown parameters characterizing the statistical model assumed to describe the noisy absolute protein concentrations. There, we provide an overview of WB measurement technique and explain the reason behind the requirement of the post-processing step of normalization. As core assumptions of our statistical framework, we consider the two most common hypotheses on the nature of noise in the measured raw data, namely log-normal and normal. In Kreutz et al. (2007) the authors maintain that WB data are generated by a log-normal distribution, based on the analysis of a large real dataset. Anyway, the second hypothesis of Gaussian error model is widely considered in the literature as well. Based on these assumptions, we consider *ratio distributions* as the straightforward formal mathematical characterization of the statistical properties of normalized data. We derive three different classes of such ratio distributions, namely normal, log-normal and Gaussian ratio. Among them, we primarily focus on the Gaussian ratio distribution and largely discuss its statistical properties, such as bimodality and heavy-tailedness. These properties complicate the characterization of summary statistics to quantify the mean and variance of the distribution, but, under particular conditions, approximation formulas are provided. Additionally, we investigate the structural identifiability for this class of ratio distributions and identify a reduced parametrization which better characterizes it. We make use of a statistical framework to analyse noise propagation from the measured sampled data to the statistical distributions of the inferred parameters. Estimation results obtained by means of a Monte Carlo simulation study show that different assumptions on the underlying GR distribution have profound impact on parameter inference. Thus, from this study we advice to use GR distributions for inference problems only with constraints, if only few datapoints are available, as done here, or in cases where heavy-tailedness is explicitly known to be present. Finally, a comparison of the three classes of error models for the description of normalized WB data is illustrated through a real dataset of WB knockdown data normalized with respect to the unperturbed control case, taken from Santos et al. (2007). Results of statistical model calibration via MLE (see Figure 2.8) show that the choice of the error model for normalized data has a partial impact on the estimation results. Obtained estimates of the expected values of the knockdown fold change are also similar among each other. From these results we could learn that the three considered error models can be fairly compared among each other as suitable description of normalized WB data.

Despite the fact that western blotting is a semi-quantitative measurement technique, signals detected via WB are broadly used as training datasets for dynamic modelling studies. In this regard, Chapter 3 deals with the problem of investigating noise propagation in the context of **dynamical model calibration** for biochemical reaction networks from relative time series data. This translates into the inference problem of estimating the unknown kinetic parameters of an ODE model used to describe the dynamics of the investigated biological system. In particular, we present a statistical framework to analyse noise propagation in terms of non-linear transformations of statistical distributions, in a similar way to what we introduced in Chapter 2. By means of a realistic mixed error model, we generated *in silico* time series data of the absolute protein concentrations. To imitate the real experimental scenario, according to which only scaled amounts of the absolute values can be quantified, we consider three types of normalization strategies commonly used for WB data. We apply MLE to infer model parameters from the different normalized datasets. In particular, to define the likelihood function required for the optimization problem, we consider the three classes of statistical distributions presented in Chapter 2 as possible error models underlying relative data generation, namely normal, log-normal and Gaussian ratio distributions. In order to analyse how distributions may be transformed from the raw experimental data to estimated parameters via normalization, we run Monte Carlo simulations and apply all combinations of normalization strategies and error models to solve the inference problem. Finally, we characterize the quality of the estimates by means of statistical measures to evaluate precision and accuracy of the obtained distributions. Our analysis is illustrated by means of a simulation study of a test-bed model for a reversible phosphorylation reaction. Our findings, based on a statistical model comparison, highlight the fact that normalization by fixed time point should be avoided and we should rather opt for the strategy of normalization by the mean value. This result holds true independently from the input noise level. Instead, the choice of the error model used to describe the normalized data surprisingly does not play a significant role in the considered inference problem. Furthermore, we could derive some practical advices concerning the total amount of measured data, with the aim of obtaining more precise and accurate parameter estimates. Overall, this study highlights the fact that standard noise levels of detected signals and commonly used amounts of data lead to uncertain parameter estimates, which profoundly impact the reliability of model predictions.

In Chapter 4 we investigate the effects of noise propagation for the last class of inference problems considered in this thesis, namely **network reconstruction** for biological applications. For this purpose, we consider the Modular Response Analysis (MRA) approach, a theoretical method based on steady-state perturbation data, which has been frequently used in the literature to reconstruct biological signalling networks. Following a similar scheme and methodology of the simulation study described in Chapter 3 for dynamical model

calibration, we went through a comprehensive analysis of noise propagation from noisy measured data to estimated Local Response Coefficients (LRCs), quantities that characterize the interactions between the nodes of a network. By means of Monte Carlo simulations and realistic noise assumptions, we could detect the appearance of heavy-tailedness in the estimated LRCs, fact which entails a high risk of wrong inference results. Through the quantification of statistical measures, we investigated the effects of different experimental and estimation approaches on the goodness of MRA-based inference of network topology. By means of our results, we could then derive useful practical advice to increase the reliability and robustness of this method for network reconstruction.

As mentioned at the beginning of this section, the content of this thesis represents a pioneer statistical analysis of noise propagation, which we expressed as a multi-step non-linear transformation of random variables over the sequence of calculations of different inference problems. In particular, our interest to understand the possible statistical implications of WB data normalization on the results of parameter estimation problems inspired our whole study. Our investigation started with the definition of the different classes of ratio distributions presented in Chapter 2. Therefore, we analysed at first the effects on statistical model calibration from normalized data and then extended these concepts to the investigation of the effects on parameter estimation for dynamical models. Finally, the last presented study on MRA-based network reconstruction arose from the fruitful collaboration with Prof. Boris Kholodenko and his team at Systems Biology Ireland, University College Dublin.

The considered methodological frameworks for parameter inference are broadly used in the literature, but we are conscious that each considered experimental and estimation approach has its own limitations, which could be improved at a later stage. For example, concerning dynamical model calibration, it would be an interesting extension to consider Bayesian estimation from the posterior distribution instead of the MLE method.

We are aware that our statistical framework cannot be formally generalized to any kind of inference problem, but rather specifically applied to concrete case studies, taking into account the experimental and computational specifications and the considered estimation method. Anyway, we tried to formulate it in an abstract way, by defining the two transformation levels $T_1$ and $T_2$ which generalize well on the three presented scenarios and possibly to other application studies. In particular, a remarkable contribution of this thesis was to characterize the three classes of ratio distributions, which represent the analytical solution of the first transformation $T_1$ for most of the Systems Biology applications. Therefore, there is still open space for many utilisations of our statistical analysis and some possible extensions of this study are discussed in the next section.

Overall, thanks to this investigation, we were able to show the relevance of taking into account the effects of noise transformation on inference results. In particular, this study highlights the fact that non-linear transformations of statistical distributions may lead to very uncertain and/or erroneous estimation results. Finally, from our statistical analysis we were able to give practical recommendations on how to optimize the experimental and methodological design of the considered inference problems.

## 5.2 Future outlook

As mentioned in the previous section, the analysis presented in this thesis may be generalized to different classes of inference problems, in which the noisy measured raw data are first transformed by some post-processing technique, for example, normalization, and after that, the transformed data are used to solve an optimization problem to estimate the unknown model parameters. Applying the statistical framework presented in this thesis would then allow to investigate how the statistical properties of the random variables are transformed at the different levels, via the two transformations $T_1$ and $T_2$, and how this may impact the uncertainty of the estimated parameters. Some interesting applications may be, for example, the estimation of cellular decision making related variables given noisy data of the underlying molecular process (Balázsi et al., 2011) or the calibration of finite mixture models from censored or uncensored data (Geissen et al., 2019; Steele and Raftery, 2010).

As future investigations it would be relevant to address some open issues concerning the results presented in this thesis. First of all, as concerns dynamical model calibration and MRA-based network reconstruction, it would be necessary to test our theoretical findings and practical suggestions in settings with real available experimental datasets.

As non-trivial problem, another interesting future direction would be to single out the factors that induce the appearance of heavy-tailedness in the inferred parameters. This was the case for the estimated LRCs in Chapter 4, under some experimental conditions. Instead, we did not observe this property in the estimated parameters $\hat{\theta}_{MLE}$ of the ODE model in Chapter 3. This may depend on the cost function applied for the optimization problem and probably MLE combined with the simultaneous estimation of error variances represents a robuster approach in this respect. Another point is that the estimated LRCs may have both positive or negative values, while we considered parameters, characterizing dynamic modelling problems, assuming only positive values.

Our statistical approach is primarily based on Monte Carlo analysis. To explore the whole distribution of all analysed random variables, we need therefore to sample a large set of realizations, and propagate the noise by solving the inference problem. This approach may be quite time consuming for rather simple case studies and, therefore, be suboptimal

for more complex models with a larger amount of unknown parameters. This would be, for example, the case of an ODE system with a large amount of state variables, whose solution has to be obtained with numerical integration. For this reason, as future study, it would be interesting to develop a formal tool to obtain functional relationships between statistical quantities, e.g. moments, characterizing the transformed distributions. A first application could be an extension of the results presented in Chapter 2, with the goal to explore the statistical properties of MLE results of statistical inference for the Gaussian ratio error model. In this regard, the work in von Luxburg and Franz (2007), presenting a geometric method to determine confidence sets for the ratio of the mean values of two random variables $\mathbf{x}$ and $\mathbf{y}$, could be a first reference to look at. In this respect, another option may be the calculation of the unscented transform via a set of selected sigma points. This represents in fact an established method to estimate the propagation of means and covariances of probability distributions, when applying non-linear transformations (Julier and Uhlmann, 2002).

# 6 Appendix

## 6.1 Maximum Likelihood Estimation

We present here a brief overview of the method of *Maximum Likelihood* for parameter estimation of dynamic models. Parts of the text of this section are taken from the Master thesis Thomaseth (2012).

We introduce both the formal definition and the practical optimization problem that has to be implemented, describing some related questions and problems that arise in the search of the optimal solution. For further and more specific details about the theory we refer to standard statistical texts, such as Ljung (1987) and Seber and Wild (1989).

This is a statistical framework for parameter estimation, relying on the hypothesis that observations are realizations of stochastic variables.

We consider a set of observations $y_i \in \mathbb{R}^n$, $i = 1, ..., N$, where $n$ is the number of measured outputs, and $i$ represents the index of the $N$ observed experiments. We collect then all values in the vector $\mathbf{y} \in \mathbb{R}^q$, which is simply the sequence of all observations $y_i$, with $q = n \cdot N$.

Suppose that $\mathbf{y}$ is a random vector, distributed with unknown probability density function that belongs to the parametrized family $\{p_{\mathbf{y}}(y|\theta), \theta \in \Theta \subseteq \mathbb{R}^M\}$.

The likelihood function of the set of observations $y_0 = \{y_i,\, i = 1, ..., N\} \in \mathbb{R}^q$ is the function $\mathcal{L}_{y_0} : \Theta \to \mathbb{R}_+$ defined by:

$$\mathcal{L}_{y_0}(\theta) = p_{\mathbf{y}}(y_0|\theta). \tag{6.1.1}$$

The "Maximum Likelihood principle", introduced by Gauss in 1809 and subsequently popularized by R.A. Fisher, suggests to take as estimate of $\theta$, referring to the observed data $y_0$, the vector $\hat{\theta}_{MLE} \in \Theta$ that maximizes $\mathcal{L}_{y_0}(\theta)$:

$$\mathcal{L}_{y_0}(\hat{\theta}_{MLE}) = \max_{\theta \in \Theta} \mathcal{L}_{y_0}(\theta), \tag{6.1.2}$$

that means:

$$\hat{\theta}_{MLE}(y_0) = \arg\max_{\theta \in \Theta} \mathcal{L}_{y_0}(\theta), \tag{6.1.3}$$

assuming implicitly that the maximum exists. In this way the value of the vector $\hat{\theta}_{MLE}$ is the one that maximizes the probability to see "a posteriori" the observation $y_0$.

For the goal of this thesis, we consider the special case when the observation $y_0 = (y_1, \ldots, y_N)$ consist of independent realizations of the random vector. Therefore, the joint probability density function can be expressed as the product of the probability densities of the individual observations Åström (1980):

$$\mathcal{L}_{y_0}(\theta) = p(y_1|\theta)p(y_2|\theta)\ldots p(y_N|\theta). \tag{6.1.4}$$

### 6.1.1 Practical optimization problems

As concerns the practical solution of this optimization problem the main question that arises is where the solution $\hat{\theta}_{MLE}$ has to be searched in the parameter space. In a general framework we expect that the desired result should be a global one, but most of the times finding the global maximum is a very difficult and complex problem. This occurs especially if the dimension $M$ of the given parameter vector $\theta$ is large and in some cases if the likelihood function is a very irregular function, with many local maxima and minima or with stiffness properties.

Moreover in a biological framework the estimated parameter values should be compatible with their biological meaning, e.g. half-lives, synthesis rates, diffusion rates, and a partial knowledge of the biochemical context under study can be useful to set some constraints for parameters, e.g. at least positivity.

For these reasons to implement the optimization problem of interest the solution can not be easily searched in the entire space $\mathbb{R}^M$ and we need constraints for our problem. In this sense we need to impose bounds for each parameter that has to be estimated, and it would be reasonable to set these bounds in a region where we expect that the solution should lie.

This can be interpreted as an *a priori* information about the distribution of parameters, and in a statistical framework this information can be expressed by a probability density function $p(\theta)$, that represents the a priori knowledge about $\theta$ before having seen the data $y$, and for this reason it is defined **prior distribution** over parameters.

From a practical point of view imposing bounds on parameters is a useful strategy for ensuring convergence of MLE optimization algorithms by avoiding, during intermediate optimization steps, inadmissible parameter values, e.g. negative values under positivity constraints, that may either hinder recovery to the admissible parameter region or even cause failure of numerical algorithms such as integration procedures. In a probabilistic context, imposing *hard* bounds on parameters by means of lower and upper limits, e.g. $\theta_{min} \leq \theta \leq \theta_{max}$, can be interpreted as assuming a uniform prior distribution on parameters, i.e. $\theta \sim \mathcal{U}(\theta_{min}, \theta_{max})$. If the parameter bounds are wide enough that the maximum of the likelihood function is attained inside the admissible parameter region then the bounds are not influential and the MLE estimate coincides with the Maximum a Posteriori (MAP) estimate, i.e. the parameter value that maximizes the posterior distribution of the

parameters given the data, under the assumption of a uniform prior. This links the MLE and the Bayesian inference Ljung (1987).

As last consideration, we underline the fact that building quantitative dynamic models for intracellular processes is only possible for specific parts of a cell, for which we have to assume that they function autonomously and can be described in isolation. Anyway external manipulations made on the specific subsystem do not act only locally but have certainly multiple effects on other parts spread all over the cell. These effects could involve unmodelled components that are not considered in the simplified model, and there could be unexpected results that cannot be explained by the model under study.

It is clear how choosing model constraints and bounds for parameters has a very important meaning and at the same time it consists in a very difficult task in the construction of predictive models.

## 6.2 Statistical models comparison: the Akaike and Bayesian Information criteria

Both the *Akaike Information Criterion* and the *Bayesian Information Criterion* are criteria for model selection based on the likelihood function, which favours models with the lowest AIC/BIC values.

The concept of AIC is related to the *information theory*, indicating the relative amount of information lost by a considered model when claiming to describe the statistical process that generated the measured data: the less information a model loses, the higher the quality of that model. The AIC value is defined as:

$$AIC = 2k - 2\ln(\mathcal{L}(\hat{\theta}_{MLE})), \tag{6.2.5}$$

with:

- $k = \#$ estimated parameters.

As can be read from Equation (6.2.5), the AIC rewards goodness of fit, but it also includes a penalty that is an increasing function of the number of estimated parameters, in order to penalize overfitting.

The BIC has a very similar definition to the AIC, but considers the total amount of observations used for the estimation as prefactor in the penalty term for the number of parameters:

$$BIC = \ln(n)k - 2\ln(\mathcal{L}(\hat{\theta}_{MLE})), \tag{6.2.6}$$

with:

- $n = \#$ observed data

- $k = \#$ estimated parameters

By comparing equations (6.2.5) and (6.2.6), it is evident that the penalty term of the BIC is larger than that of the AIC when $\ln(n) > 2$, i.e. $n > e^2 \approx 7.4 \Leftrightarrow n \geq 8$.

## 6.3 ODE model of a reversible phosphorylation reaction

From Figure 3.1, we derive the following ODE for the phosphorylated protein concentration:

$$\begin{aligned} \dot{p}^*(t) &= k_1 p(t) - k_2 p^*(t) \\ &= k_1 p_{TOT} - k_1 p^*(t) - k_2 p^*(t) \\ &= k_1 p_{TOT} - (k_1 + k_2) p^*(t). \end{aligned} \tag{6.3.7}$$

In particular, we assume total mass concentration. By defining the state variable $x(t)$ as the phosphorylated amount with respect to the total protein concentration (see Equation (3.2.8)), we divide Equation (6.3.7) by $p_{TOT}$ and easily obtain Equation (3.2.9), given by:

$$\dot{x}(t) = k_1 - (k_1 + k_2) x(t).$$

## 6.4 The GR error model for data normalized by the mean value: correlation coefficient

We consider the set of normalized measurements obtained with the third normalization strategy $y^j_{N3}(t_k)$, which can be written as the ratio of the absolute protein concentration $\tilde{x}^j(t_k)$ and the corresponding mean value of all time points $\frac{1}{K}\sum_{k=1}^{K} \tilde{x}^j(t_k)$. We consider therefore the random variable:

$$\mathbf{y}_{N3}(t_k) = \frac{\tilde{\mathbf{x}}(t_k)}{\frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{x}}(t_k)}.$$

We want to prove that the correlation between numerator and denominator is equal to $\rho = \frac{1}{\sqrt{K}}$.

For simplicity we write the proof for $k = 1$ and $K = 4$, i.e. we consider the random variable:

$$\mathbf{y}_{N3}(t_1) = \frac{\tilde{\mathbf{x}}(t_1)}{\frac{1}{4}\left(\tilde{\mathbf{x}}(t_1) + \tilde{\mathbf{x}}(t_2) + \tilde{\mathbf{x}}(t_3) + \tilde{\mathbf{x}}(t_4)\right)}.$$

The results can then be generalized $\forall k \in \{1, \ldots, K\}$ and for any value of $K \in \mathbb{N}$.

We simplify the notation and define the variables:

- $X_k = \tilde{\mathbf{x}}(t_k), k = 1, 2, 3, 4$

- $X = X_1 + X_2 + X_3 + X_4 = \tilde{\mathbf{x}}(t_1) + \tilde{\mathbf{x}}(t_2) + \tilde{\mathbf{x}}(t_3) + \tilde{\mathbf{x}}(t_4)$.

We want to calculate the correlation between $X_1$ and $X$:

$$Corr(X_1, X) = \frac{Cov(X_1, X)}{\sqrt{Var(X_1)}\sqrt{Var(X)}}. \tag{6.4.8}$$

- $$\begin{aligned} Cov(X_1, X) &= \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X - \mathbb{E}(X))] \\ &= \mathbb{E}\left[(X_1 - \mu_1)\left(X - \frac{1}{4}\sum \mu_k\right)\right] \\ &= \mathbb{E}[X_1 X] - \mathbb{E}[\mu_1 X] - \mathbb{E}[X_1 \frac{1}{4}\sum \mu_k] + \mathbb{E}[\mu_1 \frac{1}{4}\sum \mu_k] \\ &= \mathbb{E}[X_1 \frac{1}{4}(X_1 + X_2 + X_3 + X_4)] - \mu_1 \mathbb{E}[X] - \mathbb{E}[X_1]\frac{1}{4}\sum \mu_k + \mu_1 \frac{1}{4}\sum \mu_k \\ &= \mathbb{E}[\frac{1}{4}X_1X_1 + \frac{1}{4}X_1X_2 + \frac{1}{4}X_1X_3 + \frac{1}{4}X_1X_4] + \\ &\quad - \mu_1 \frac{1}{4}\sum \mu_k - \mu_1 \frac{1}{4}\sum \mu_k + \mu_1 \frac{1}{4}\sum \mu_k \\ &= \mathbb{E}[\frac{1}{4}X_1X_1 + \frac{1}{4}X_1X_2 + \frac{1}{4}X_1X_3 + \frac{1}{4}X_1X_4] - \mu_1 \frac{1}{4}\sum \mu_k \\ &= \frac{1}{4}\mathbb{E}[X_1^2] + \frac{1}{4}\mathbb{E}[X_1X_2] + \frac{1}{4}\mathbb{E}[X_1X_3] + \frac{1}{4}\mathbb{E}[X_1X_4] - \mu_1 \frac{1}{4}\sum \mu_k \\ &= \frac{1}{4}\mathbb{E}[X_1^2] + \frac{1}{4}\mathbb{E}[X_1]\mathbb{E}[X_2] + \frac{1}{4}\mathbb{E}[X_1]\mathbb{E}[X_3] + \frac{1}{4}\mathbb{E}[X_1]\mathbb{E}[X_4] - \mu_1 \frac{1}{4}\sum \mu_k \\ &= \frac{1}{4}\mathbb{E}[X_1^2] + \frac{1}{4}\mu_1(\mu_2 + \mu_3 + \mu_4) - \frac{1}{4}\mu_1^2 - \frac{1}{4}\mu_1(\mu_2 + \mu_3 + \mu_4) \\ &= \frac{1}{4}\mathbb{E}[X_1^2] - \frac{1}{4}\mu_1^2 \\ &= \frac{1}{4}Var(X_1) = \frac{1}{4}\sigma^2. \end{aligned}$$
- $Var(X) = \dfrac{\sigma^2}{4}$

$$\implies Corr(X_1, X) = \frac{\frac{1}{4}\sigma^2}{\sigma \cdot \frac{1}{2}\sigma} = \frac{1}{2}$$

In general for $K$ timepoints we obtain:

$$Corr(X_k, X) = \frac{\frac{1}{K}\sigma^2}{\sigma \cdot \frac{1}{\sqrt{K}}\sigma} = \frac{1}{\sqrt{K}}$$

## 6.5 Parametrization of the GR error model for the ODE test-bed model application

For the parameter estimation study applied to the ODE test-bed model presented in Chapter 3, we have to implement the Likelihood function for the three considered EMs (see equations (3.2.18), (3.2.19) and (3.2.20)–(3.2.22)), which depend on the specific simulated model outputs and on the chosen NS. In the particular case of the GR distribution we have to define the parametrization $\theta_3 = (a, b, r, s)$, specifically for all three NSs. In particular, we compare equations (3.2.20)–(3.2.22) with the general case given in Section 2.3, which defines the GR distribution of the RV $\mathbf{z} = \mathbf{x}/\mathbf{y}$, ratio of the two normal RVs $\mathbf{x} \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathbf{y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ with correlation $\rho$. We also need equations (2.3.14) for the definition of the four identifiable parameters $a, b, r, s$ given $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$.

First, we look at the parameters $s$ and $r$, which can be univocally determined for all three NSs.

- **NS1** and **NS2**: from the two assumptions of uncorrelated RVs $\tilde{\mathbf{x}}(t_k) \sim \mathcal{N}(x(t_k, \theta), \sigma^2)$ ($\rho = 0$) with the same SDs ($\sigma_X = \sigma_Y$), valid for different time points, we obtain:
  - $s = 0$
  - $r = 1$
- **NS3**: In this case we can prove that $\rho = \frac{1}{\sqrt{K}}$ and that $\sigma_Y = \frac{1}{\sqrt{K}}\sigma_X$ (see equation (3.2.22)). We obtain then:
  - $s = 1$
  - $r = \frac{1}{\sqrt{K-1}}$

As concerns the remaining two parameters $a$ and $b$, we obtain that both depends on the unknown parameters to be estimated, while $a$ also depends on the time point of the corresponding measurement $t_k$, for all three different normalization strategies:

- **NS1**:
  - $a_k(\theta, \sigma) = \dfrac{x(t_k, \theta)}{\sigma}, k = 2, \ldots, K$
  - $b(\theta, \sigma) = \dfrac{x(t_1, \theta)}{\sigma}$
- **NS2**:
  - $a_k(\theta, \sigma) = \dfrac{x(t_k, \theta)}{\sigma}, k = 1, \ldots, K-1$
  - $b(\theta, \sigma) = \dfrac{x(t_K, \theta)}{\sigma}$
- **NS3**:
  - $a_k(\theta, \sigma) = \dfrac{(x(t_k, \theta) - x_{mean}(\theta))\sqrt{K}}{\sigma\sqrt{K-1}}, k = 1, \ldots, K$
  - $b(\theta, \sigma) = \dfrac{x_{mean}(\theta)\sqrt{K}}{\sigma}$

## 6.6 Impact of number of time points on the uncertainty of ML estimates

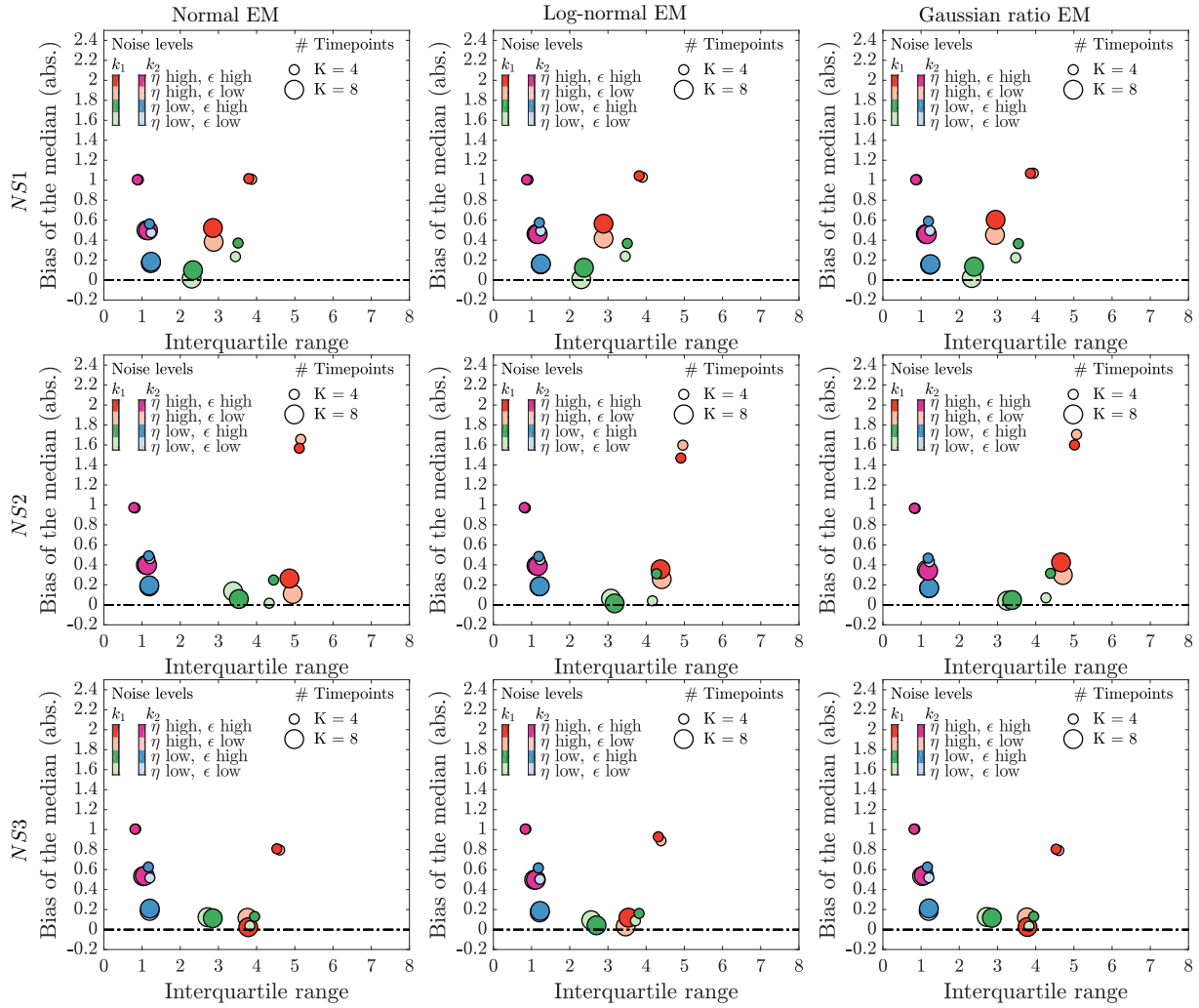


**Fig. 6.1. Effect of the amount of time points $K$ under realistic experimental settings.** Absolute values of the bias of the median versus IQR values for both estimated parameter values obtained with $K = 4$ or $K = 8$ time points and $J = 1$ replicate. These statistics are given for four different noise levels obtained combining the value $\sigma_\eta \in \{0.05$ (green/blue), 0.1 (red/magenta)$\}$ and $\sigma_\epsilon \in \{0.01, 0.02\}$ (indicated by increasing darkness). Green and red dots refer to the parameter $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$, while blue and magenta refer to $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$.

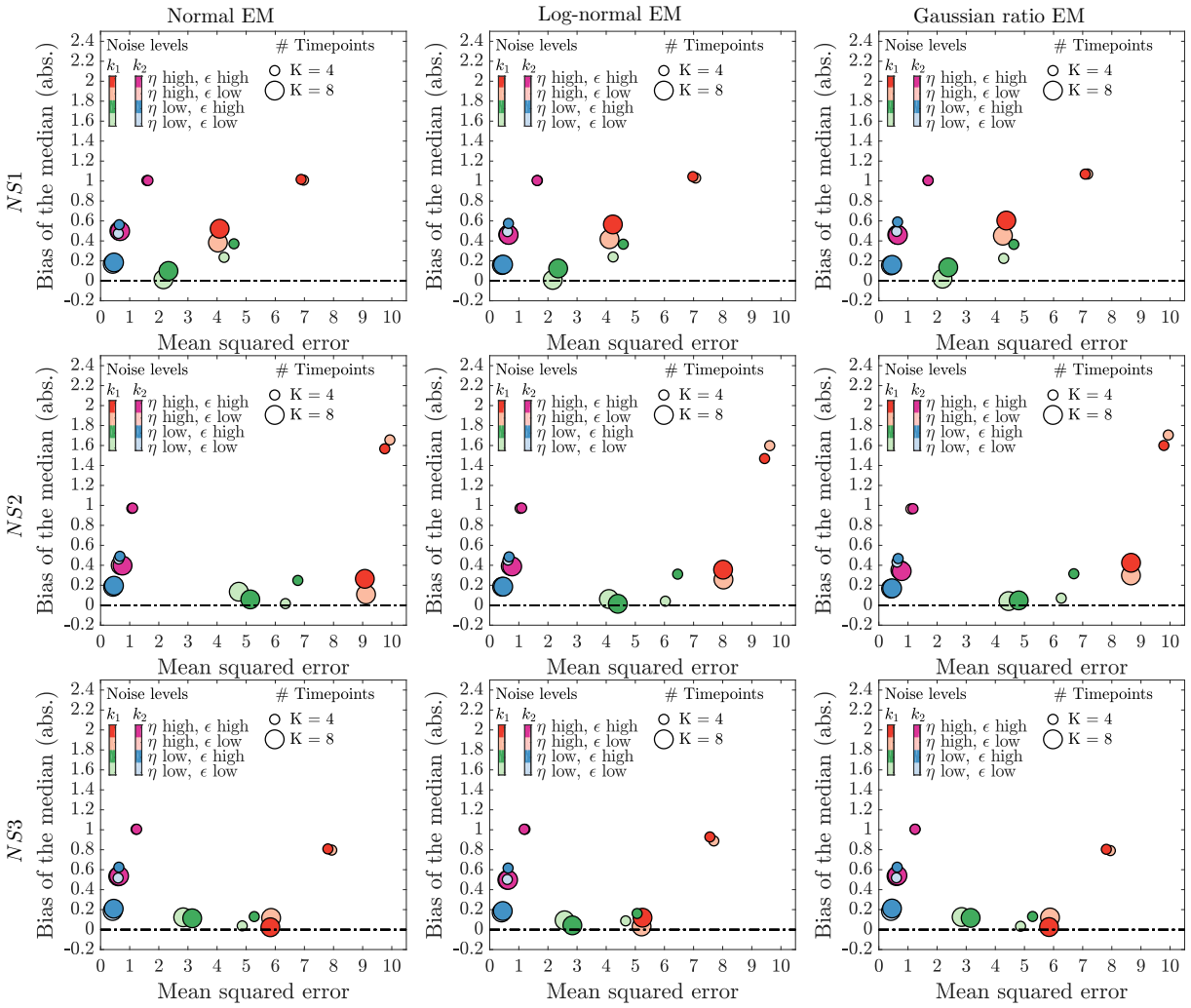


**Fig. 6.2. Effect of the amount of time points $K$ under realistic experimental settings.** Absolute values of the bias of the median versus MSE values for both estimated parameter values obtained with $K = 4$ or $K = 8$ time points and $J = 1$ replicate. These statistics are given for four different noise levels obtained combining the value $\sigma_\eta \in \{0.05$ (green/blue), 0.1 (red/magenta)$\}$ and $\sigma_\epsilon \in \{0.01, 0.02\}$ (indicated by increasing darkness). Green and red dots refer to the parameter $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$, while blue and magenta refer to $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$.

### 6.6.1 Estimated parameters for increasing $K - J = 1$

- $J = 1, K \in \{4, 8\}$, **Normal error model**

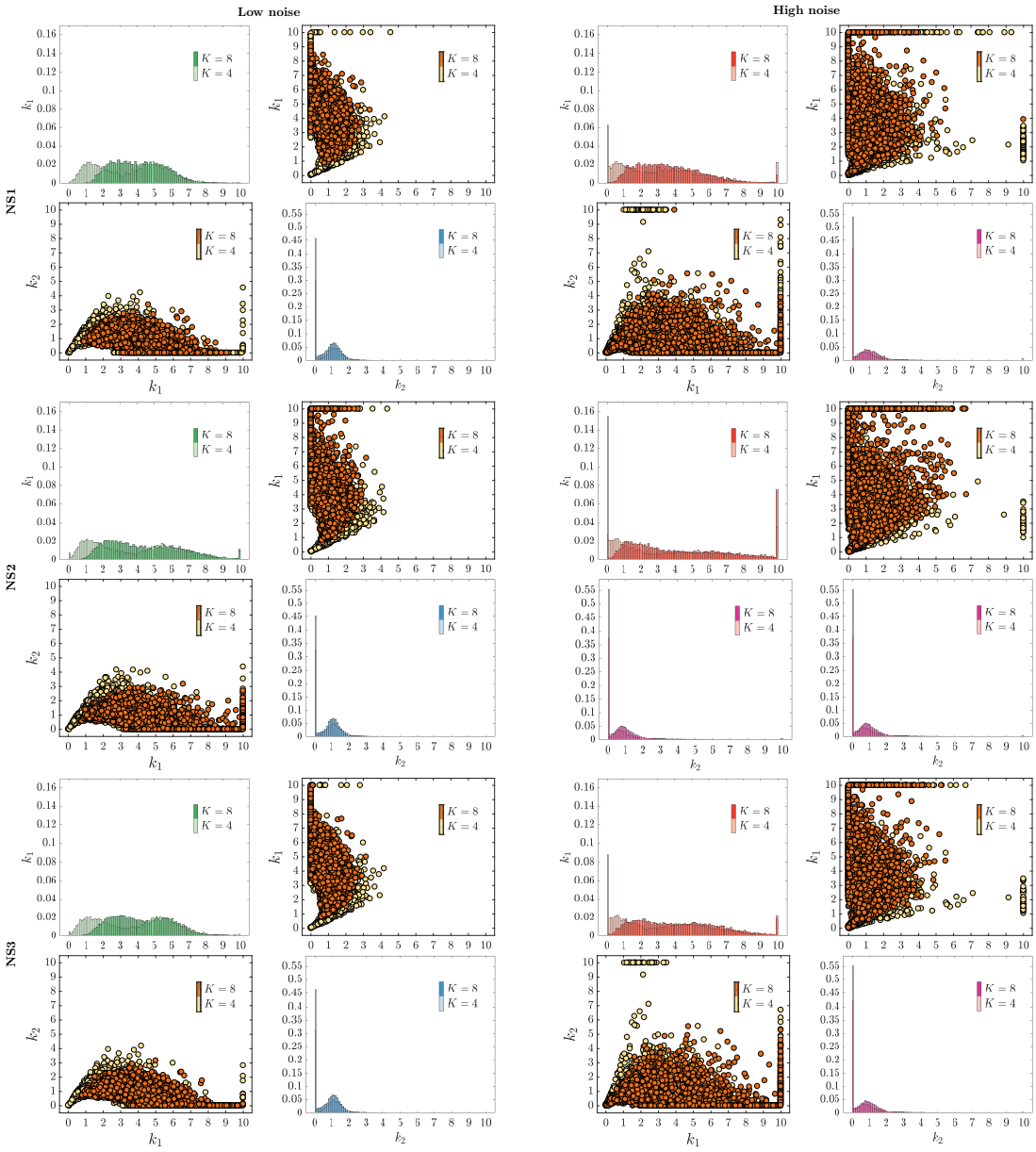


**Fig. 6.3. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 1$, $K \in \{4, 8\}$ and assuming the normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 1, K \in \{4, 8\}$, **Log-normal error model**

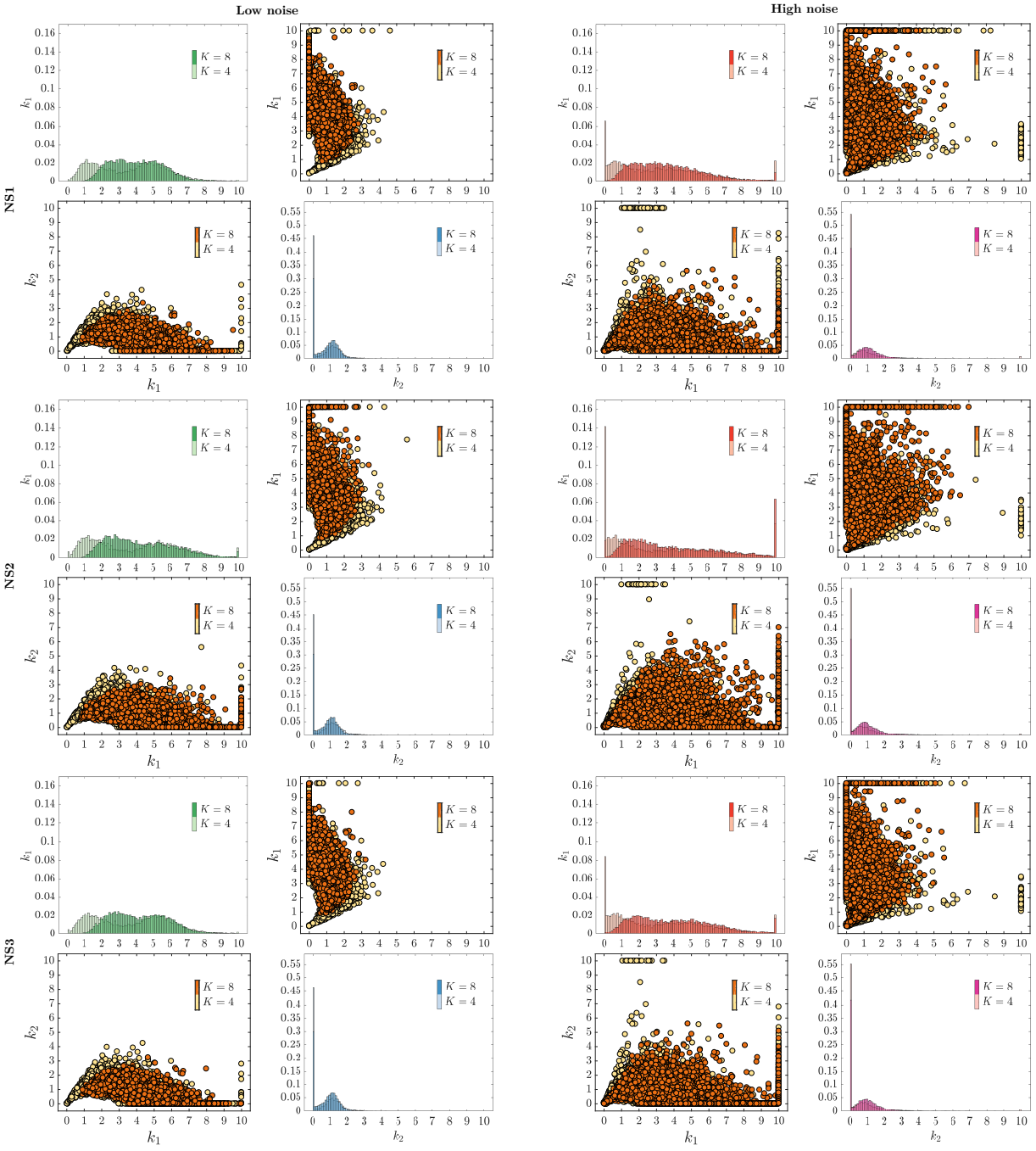


**Fig. 6.4. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 1$, $K \in \{4, 8\}$ and assuming the log-normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 1, K \in \{4, 8\}$, **Gaussian ratio error model**

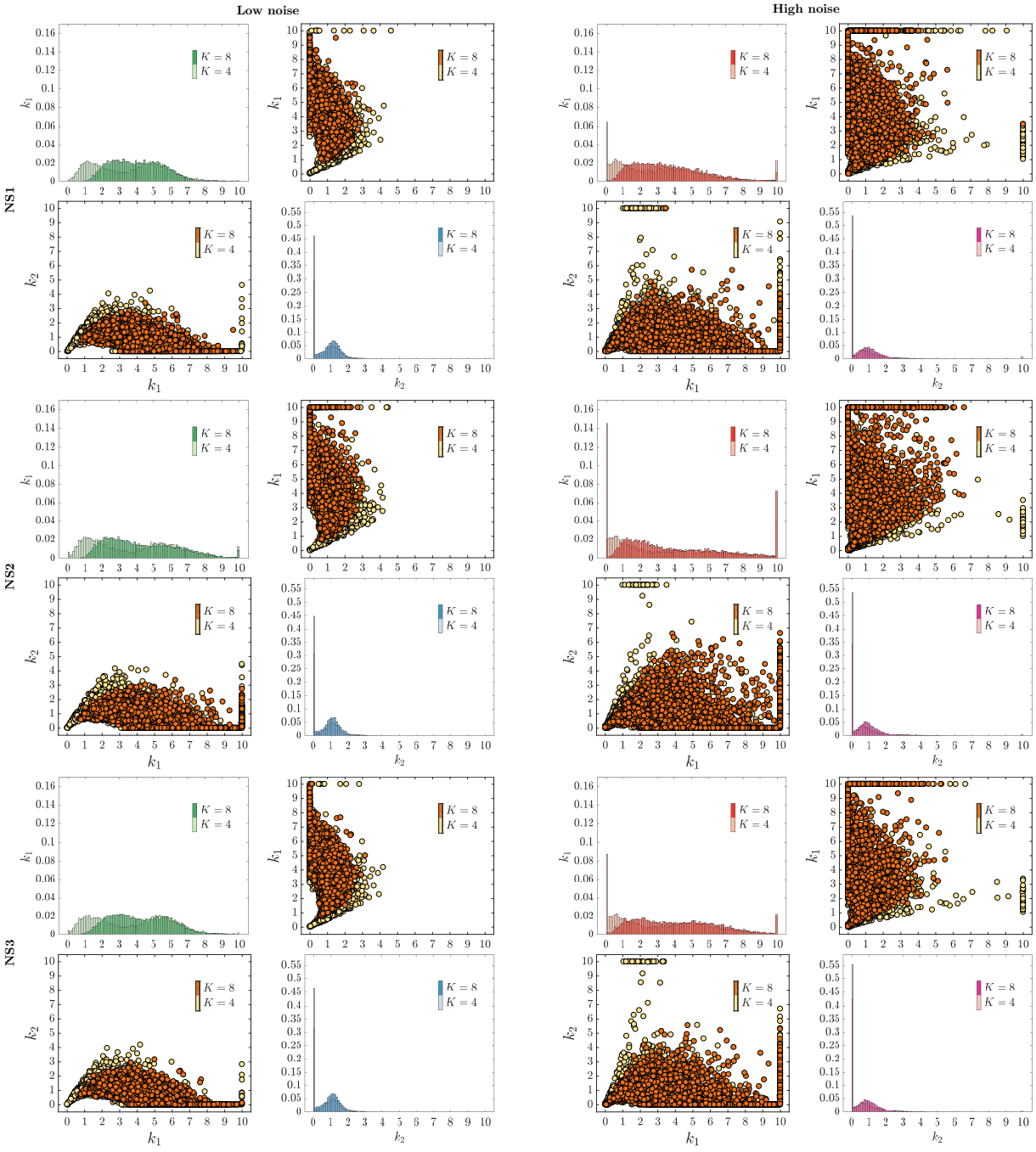


**Fig. 6.5. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 1$, $K \in \{4, 8\}$ and assuming the Gaussian ratio error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

### 6.6.2 Estimated parameters for increasing $K - J = 6$

- $J = 6, K \in \{4, 8\}$, **Normal error model**

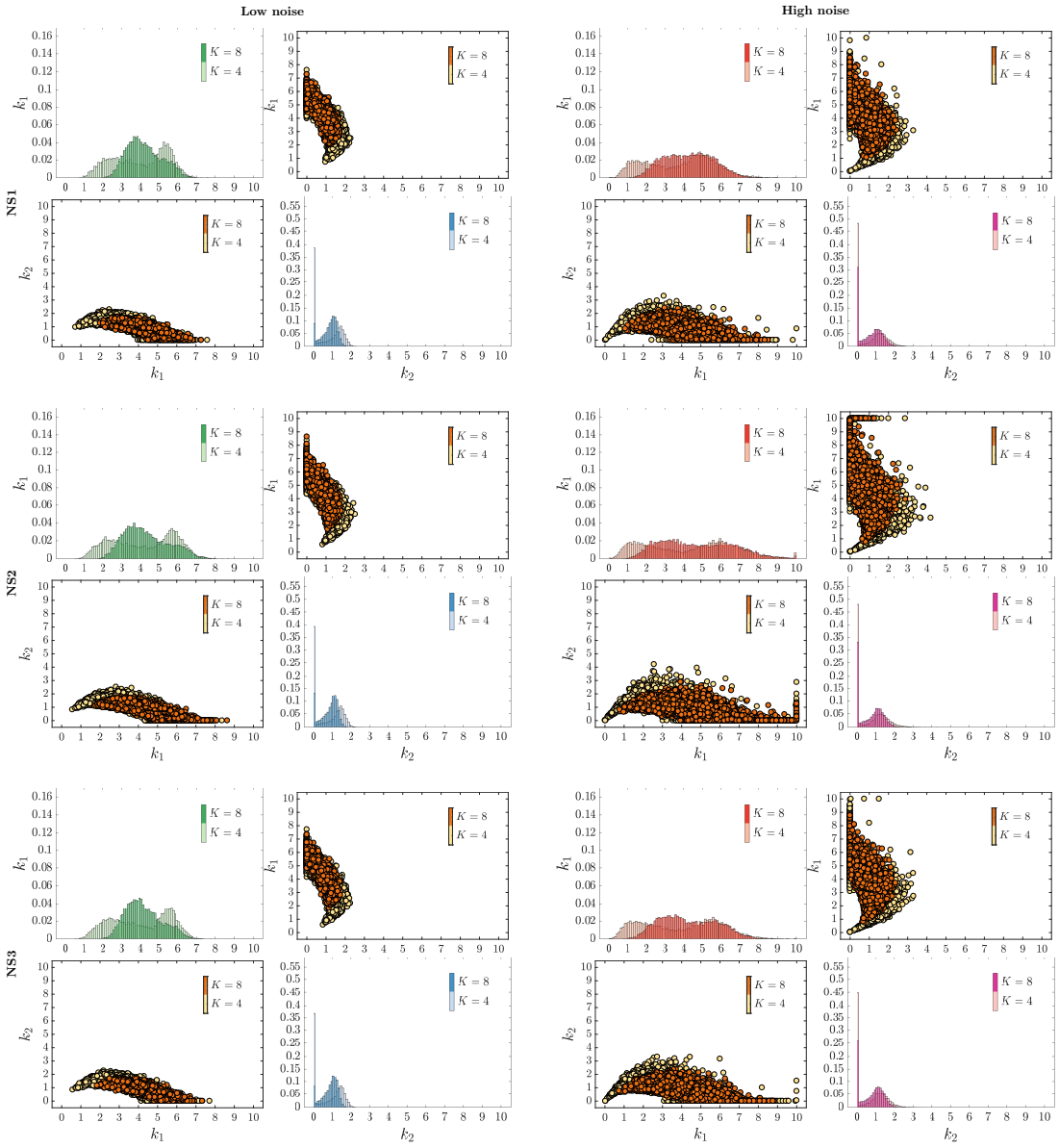


**Fig. 6.6. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 6$, $K \in \{4, 8\}$ and assuming the normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 6, K \in \{4, 8\}$, **Log-normal error model**

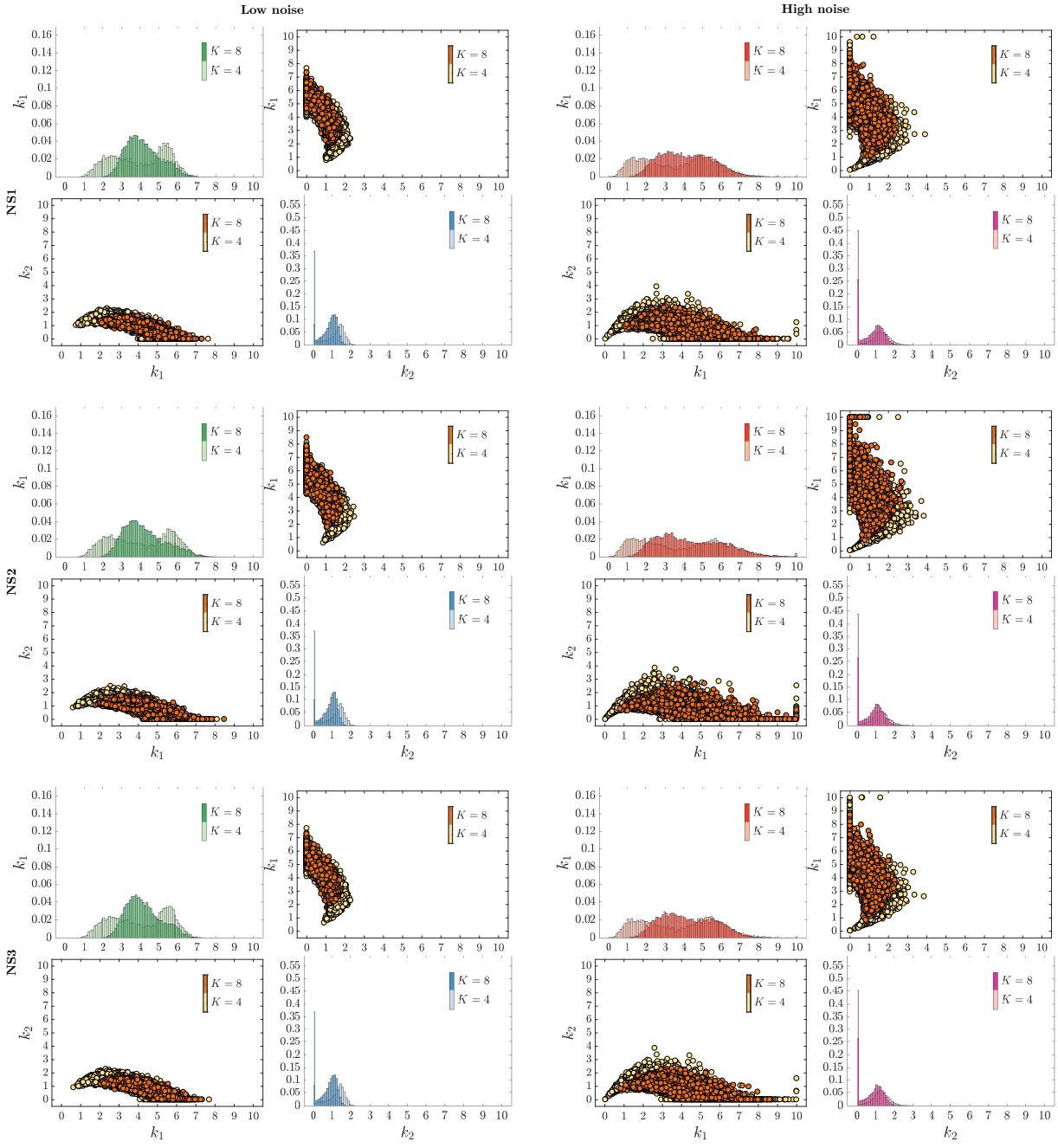


**Fig. 6.7. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 6$, $K \in \{4, 8\}$ and assuming the log-normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 6, K \in \{4, 8\}$, **Gaussian ratio error model**

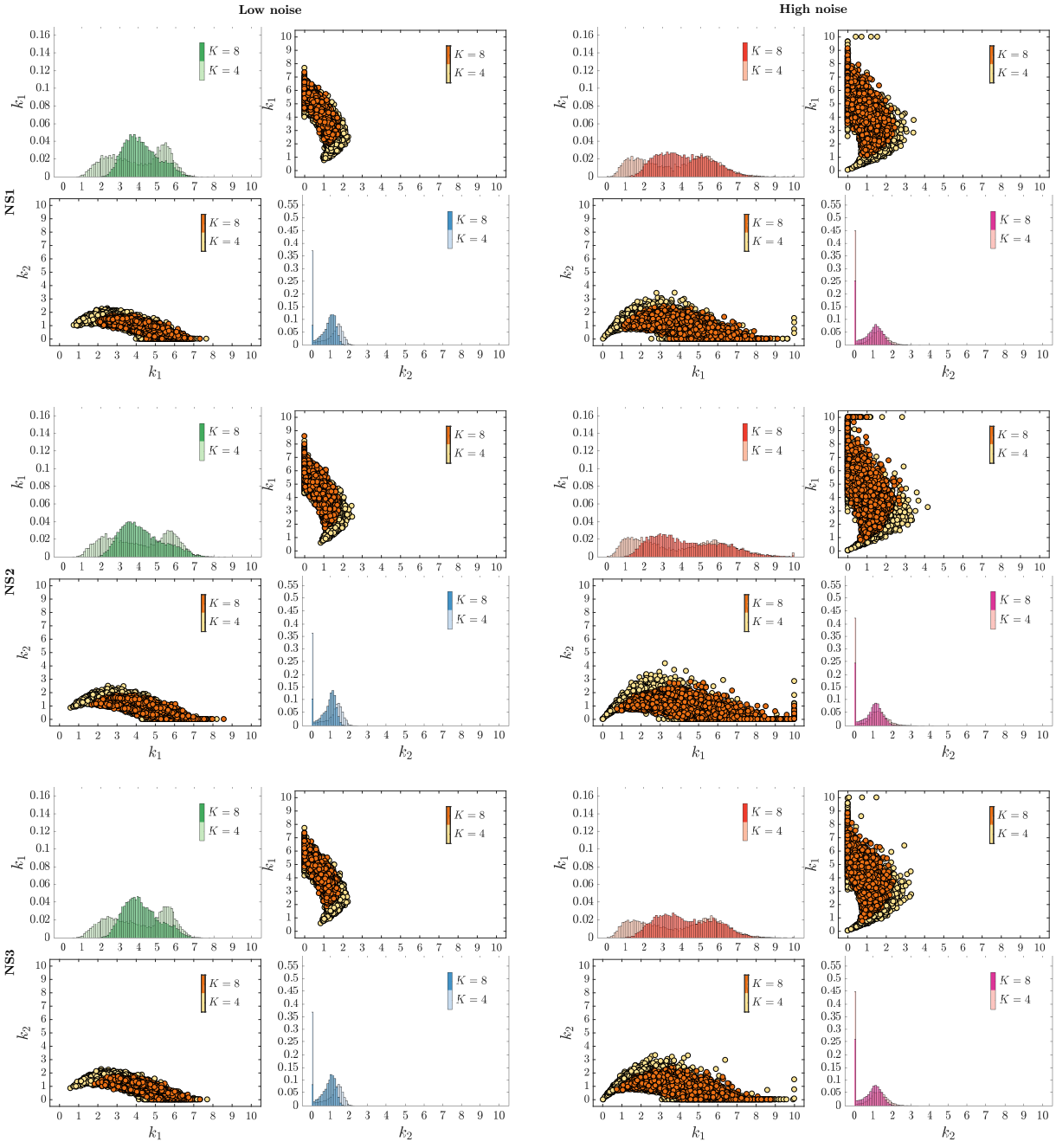


**Fig. 6.8. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 6$, $K \in \{4, 8\}$ and assuming the Gaussian ratio error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

### 6.6.3 Estimated parameters for increasing $K - J = 10$

- $J = 10, K \in \{4, 8, 12\}$, **N-EM**

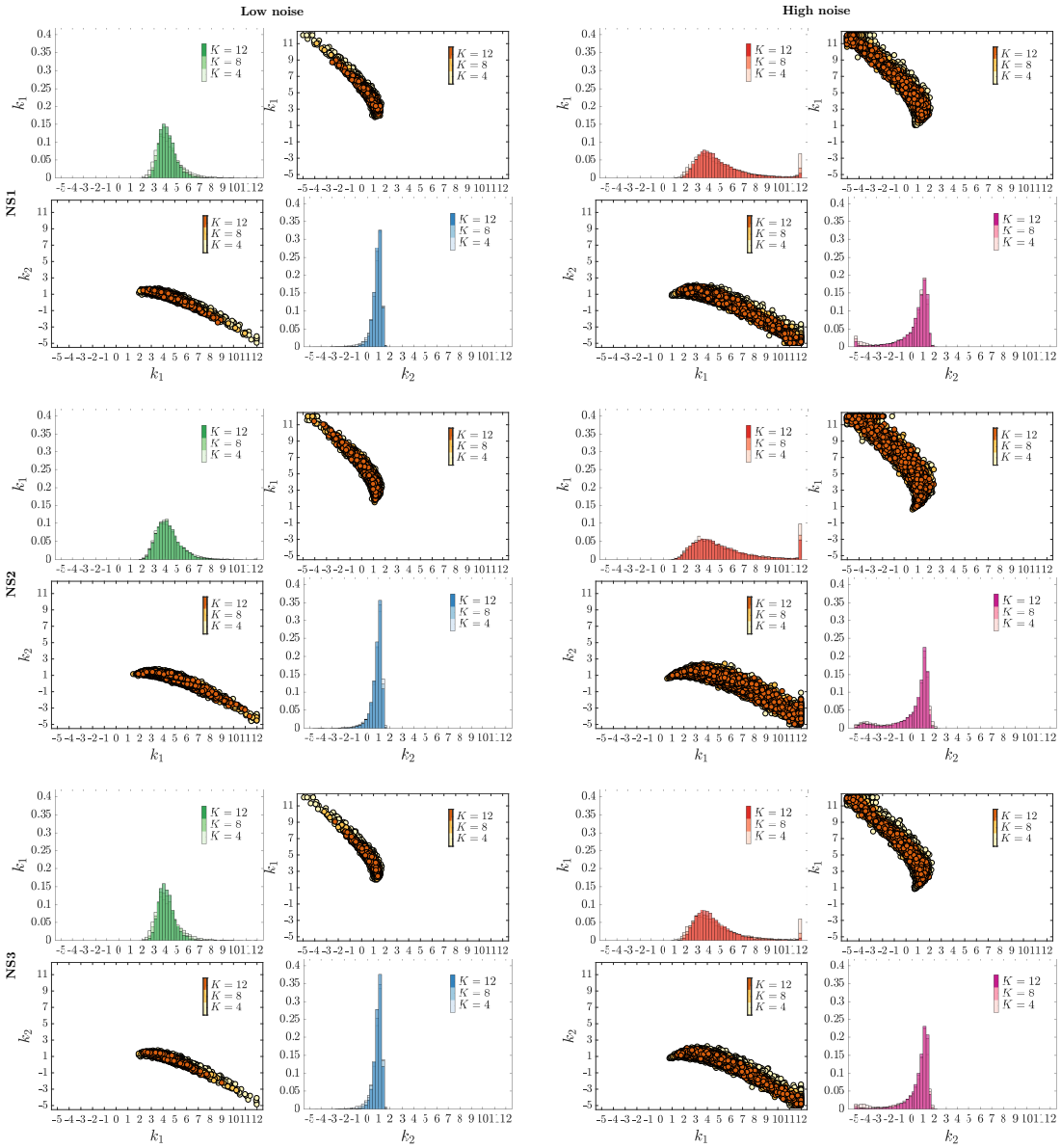


**Fig. 6.9. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathbf{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathbf{MLE}}$, obtained for $J = 10$, $K \in \{4, 8, 12\}$ and assuming the normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 10, K \in \{4, 8, 12\}$, **LN-EM**

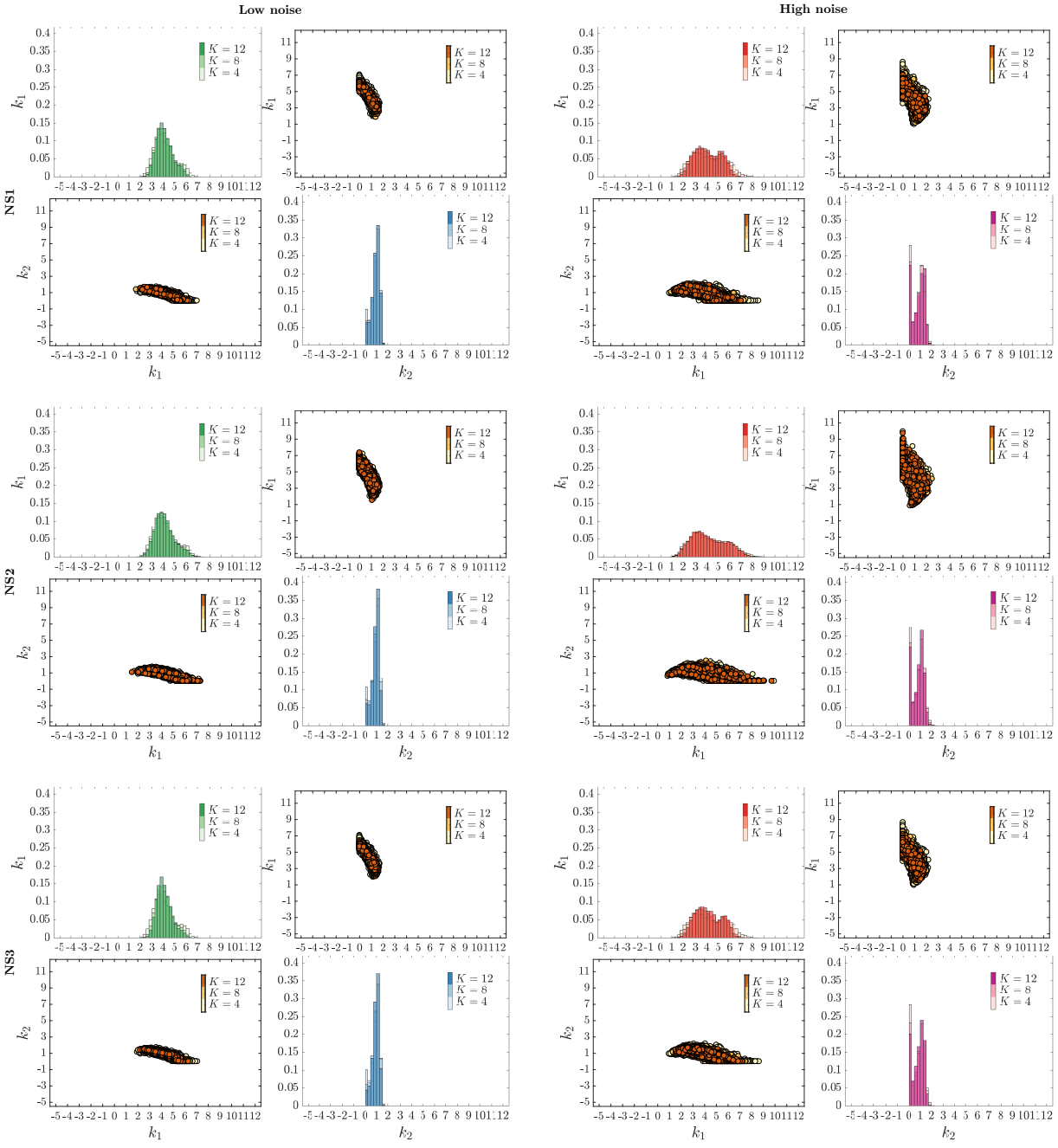


**Fig. 6.10. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 10$, $K \in \{4, 8, 12\}$ and assuming the log-normal error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

- $J = 10, K \in \{4, 8, 12\}$, **GR-EM**

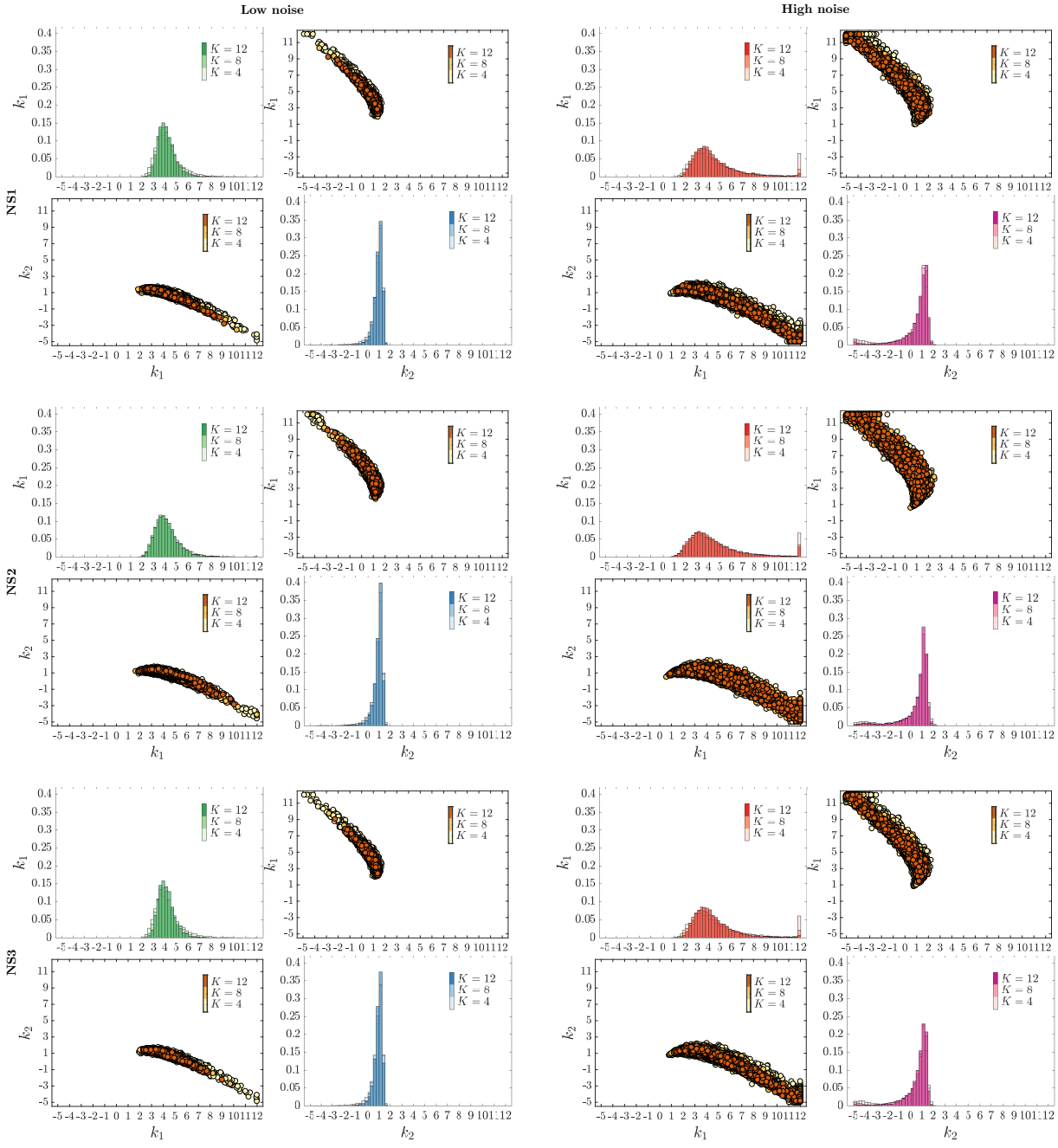


**Fig. 6.11. Marginals and scatter plots of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$, obtained for $J = 10$, $K \in \{4, 8, 12\}$ and assuming the Gaussian ratio error model.** The estimated parameters were obtained with the three different sets of normalized data (rows) and two noise levels (low on the left: $\sigma_\eta = 0.05, \sigma_\epsilon = 0.01$, high on the right: $\sigma_\eta = 0.1, \sigma_\epsilon = 0.02$ For each scenario the plot shows the two marginal distributions (histograms) of $\hat{\mathbf{k}}_{1,\mathrm{MLE}}$ and $\hat{\mathbf{k}}_{2,\mathrm{MLE}}$ (on the diagonal) and the symmetrical scatter plots in the 2-dimensional parameter space.

## 6.7 Statistical model comparison - high noise level

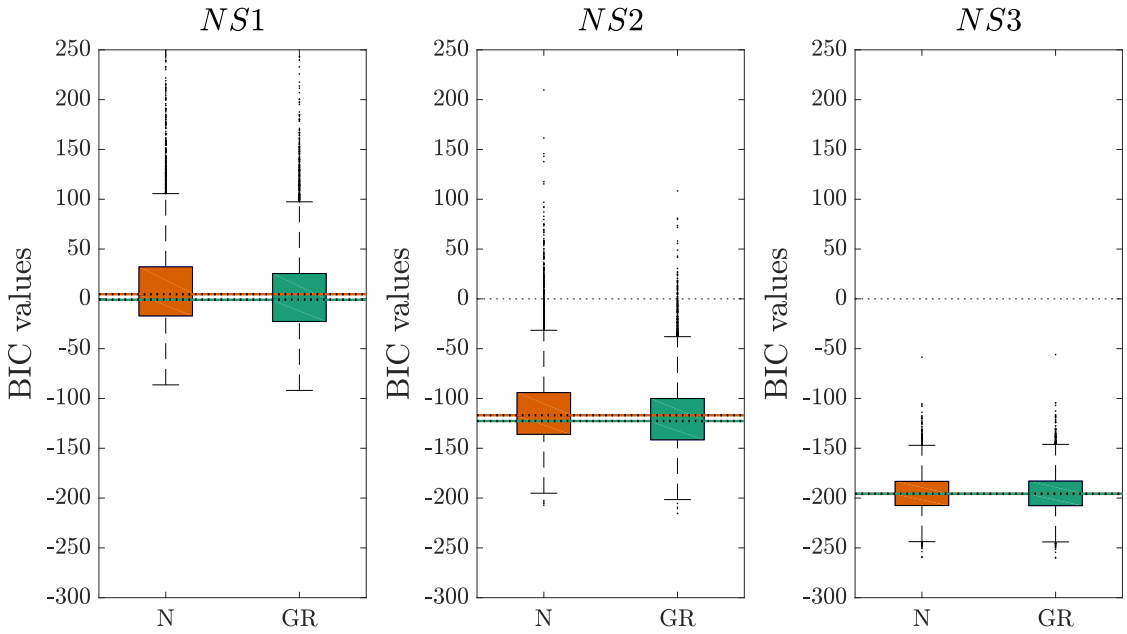


**Fig. 6.12. Statistical model comparison.** Box plots of the BIC values (6.2.6) calculated with the estimation results obtained using the different datasets from the three normalization strategies $NS1, NS2$ or $NS3$ and two EMs (N and GR). These results were obtained with $K = 12$ time points, $J = 10$ replicates and the high noise level of the input data: $\sigma_\eta = 0.1$ and $\sigma_\epsilon = 0.02$.

## 6.8 Estimated $\hat{\sigma}_{\mathbf{MLE}}$

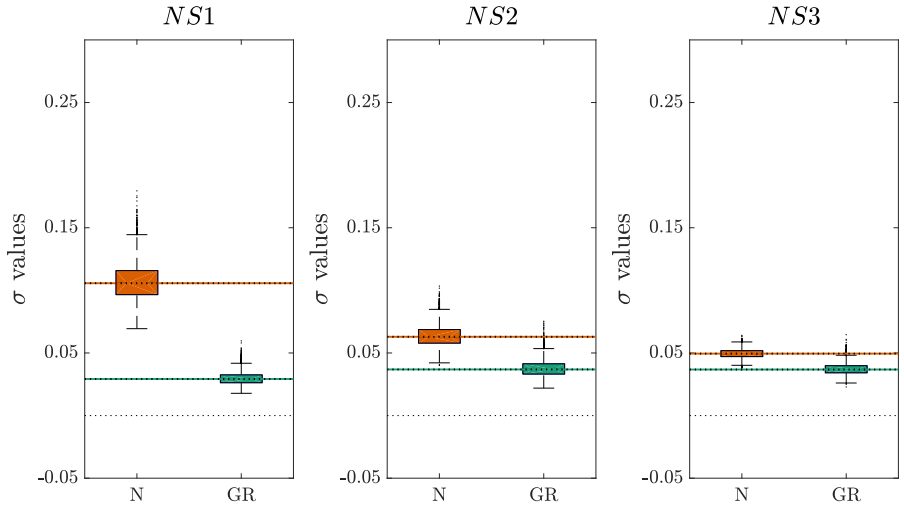


**Fig. 6.13. Box plots of $\hat{\sigma}_{\mathbf{MLE}}$** corresponding to the SD of the assumed N-EM or GR-EM, obtained using the different datasets from the three normalization strategies $NS1, NS2$ or $NS3$. These results were obtained with $K = 12$ time points, $J = 10$ replicates and the low noise level of the input data: $\sigma_\eta = 0.05$ and $\sigma_\epsilon = 0.01$.

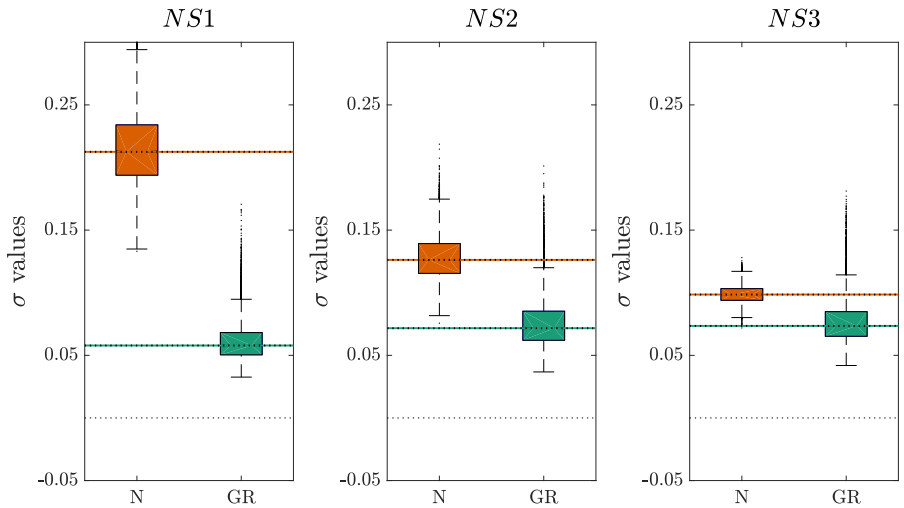


**Fig. 6.14. Box plots of $\hat{\sigma}_{\mathbf{MLE}}$** corresponding to the SD of the assumed N-EM or GR-EM, obtained using the different datasets from the three normalization strategies $NS1, NS2$ or $NS3$. These results were obtained with $K = 12$ time points, $J = 10$ replicates and the high noise level of the input data: $\sigma_\eta = 0.1$ and $\sigma_\epsilon = 0.02$.

## 6.9 The MAPK and the p53 ODE models

Figure 6.15 shows the two test-bed models which are used in the study presented in Chapter4. Both models are based on previously published models for the MAPK (Kholodenko, 2006)

and p53 (Fey et al., 2016) systems. Parameters have been chosen in order to obtain an EGF-induced MAPK model that behaves relatively linear and a p53 DNA-damage response model that behaves strongly nonlinear. This reflects the observed behaviour of these systems (Kholodenko et al., 2010; Purvis et al., 2012).

A model of signal transduction of the MAPK pathway upon EGF stimulation is illustrated in Figure 6.15a. It consists of a three-tiered cascade of phosphorylation-dephosphorylation cycles in which pRaf phosphorylates and thereby activates MEK, which then activates ERK, which negatively feeds back to Raf. Both MEK and ERK require phosphorylation at two sites to become fully activated, which is for simplicity assumed to happen in a single reaction step for both proteins. This system is described by a dynamical model in the form of ODEs $\dot{x} = f(x, \theta)$, $x \in \mathbb{R}^3_+$. The state variables $x(t) = [x_1(t), x_2(t), x_3(t)]^\top$ refer to the active states of the three proteins pRaf, ppMEK and ppERK. The input $u(t) =$ EGF is set to a constant value $u(t) = 1$. The initial conditions of the state variables are set to zero, i.e. there are no active states at time $t = 0$. The total concentrations of the three states are $\mathrm{Raf_{TOT}}, \mathrm{MEK_{TOT}}, \mathrm{ERK_{TOT}} = 20$. The kinetic rates and other parameter values are: $k_1 = 5, K^+_{m1} = 20, V_{m1} = 10, K^-_{m1} = 20, k_2 = 3, K^+_{m2} = 20, V_{m2} = 10, K^-_{m2} = 20, k_3 = 1, K^+_{m3} = 20, V_{m3} = 10, K^-_{m3} = 20, K_{mf} = 5$. Perturbation parameter $p_1$, $p_2$ and $p_3$ correspond to fold changes in total protein amounts. One after the other they are set to some defined value smaller or greater than 1, the same for all perturbation experiments. To obtain the simulation results presented in the manuscript, we considered the values $p_j \in \{0.2, 0.5, 0.75, 1.5\}, \forall j = 1, 2, 3$, that represent 80%, 50% and 25% knockdown or 50% overexpression of the total protein concentrations.

Figure 6.15b illustrates activation of p53 by pATM. Active p53 triggers expression of MDM2, which is in turn involved in the degradation of p53, resulting in a negative feedback loop. Both p53 and MDM2 are subject to synthesis and degradation, and model variables $x_1$, $x_2$ and $x_3$ correspond to pATM, p53 total amount and MDM2, respectively. As before, perturbation parameters $p_1$, $p_2$ and $p_3$ describe fold changes in ATM total amount, and in p53 and MDM2 synthesis rates.

The initial conditions of the state variables are set to zero, i.e. there are no active states at time $t = 0$. The input is set to a constant value $u(t) = 1$. The kinetic rates and other parameter values are: $k_1 = 3, K_1 = 0.5, k_2 = 5, K2 = 0.5, k_3 = 1, n_5 = 5, K_5 = 0.1, k_4 = 1, KD = 0.01, k_6 = 1, n_6 = 5, K_6 = 0.5, k_7 = 1$. The total concentration of ATM is $\mathrm{ATM_{TOT}} = 1$.

In both subfigures, graphs indicate the steady states of the system variables as functions of the perturbation parameters. While these curves can well be approximated by linear functions in the first case, they show a more pronounced non-linear behaviour in the second case.

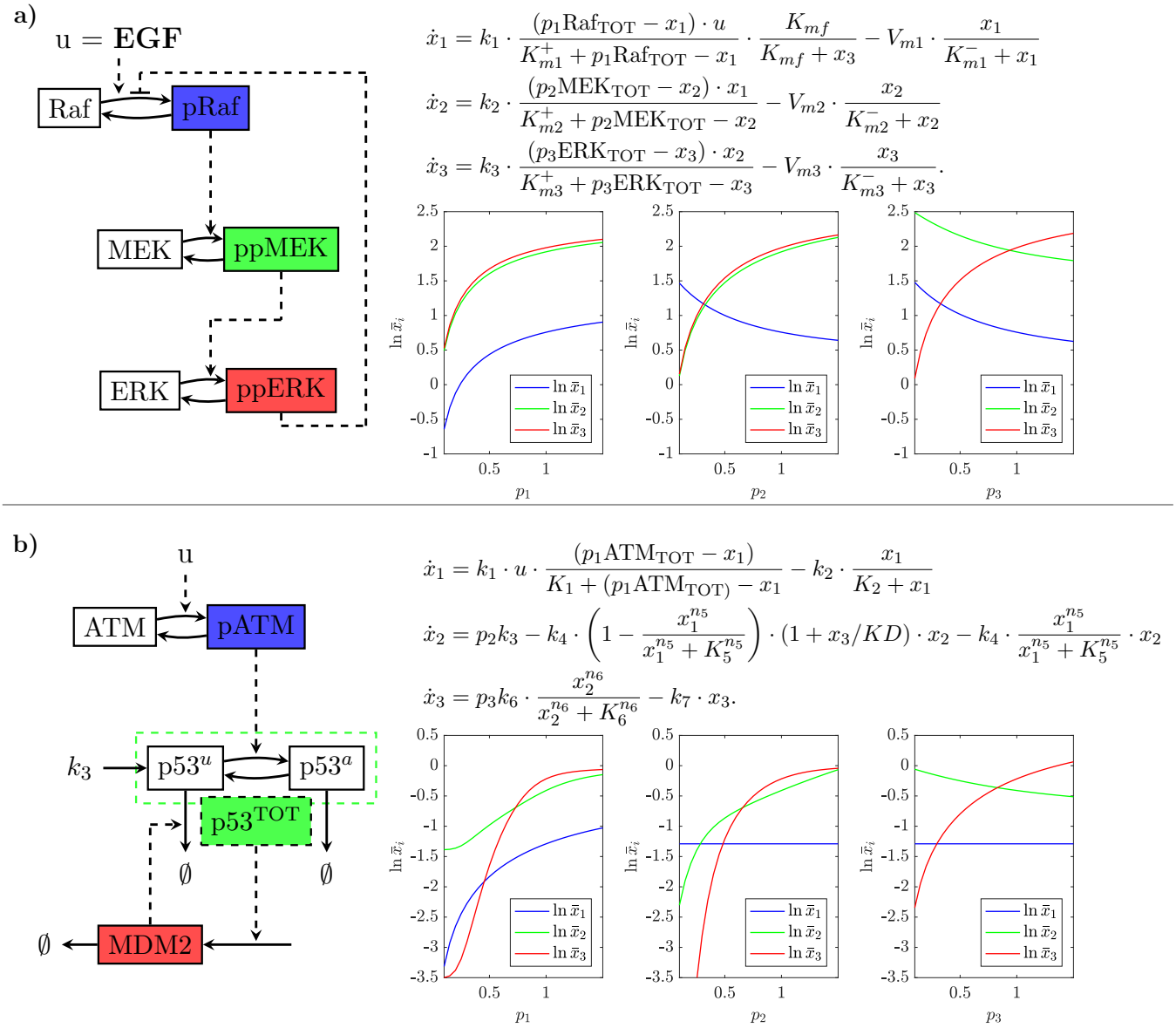


**Fig. 6.15. Two test-bed models.** Shown are the reaction kinetic schemes (left), the ODE system (right top) and the dependencies of the logarithm of the state variables on the perturbation parameters (right bottom) for **(a)** the MAPK system; **(b)** the p53 system.

## 6.10 Medcouple

Given a set of $n$ independent samples $\{x_1, ..., x_n\}$ from a continuous univariate distribution, with median $m_n$, the medcouple is defined as

$$\mathrm{MC} = \underset{x_i \le m_n \le x_j}{\mathrm{med}} h(x_i, x_j), \quad \text{with} \quad h(x_i, x_j) = \frac{(x_j - m_n) - (m_n - x_i)}{x_j - x_i}, \forall\, x_j \neq x_i. \quad (6.10.9)$$

The kernel function $h(x_i, x_j)$ measures the (normalized) difference between the distances of $x_i$ and $x_j$ to the median. The medcouple represents a robust measure of the asymmetry

of a distribution, which can be computed also for distributions without finite moments, which is not the case for the classical skewness coefficient (Brys et al., 2004). As robust measure of tail weight, the authors propose to apply the medcouple only to one single side of the distribution, leading to Left Medcouple (LMC) and Right Medcouple (RMC) (Brys et al., 2006):

$$\text{LMC} = -\text{MC}(x < m_n) \quad \text{and} \quad \text{RMC} = \text{MC}(x > m_n). \tag{6.10.10}$$

The calculation of such quantities for all datasets in our study was performed with the MATLAB toolbox LIBRA (Verboven and Hubert, 2010), developed by the same authors, which can be downloaded from:
`https://wis.kuleuven.be/stat/robust/LIBRA/LIBRA-home`.

## 6.11 MRA estimation methods

For the estimation problem, we have to solve equation (4.2.7), $\mathbf{y} = A \cdot \mathbf{x}$, which is a linear regression model, in the unknown variable $\mathbf{x}$. Assuming no error in the regression variables, i.e. in the entries of the matrix $A$, and i.i.d. normal errors in the variable $\mathbf{y}$, we obtain the well known Ordinary Least Squares (OLS) solution, given in equation (4.2.8). However, this assumption is wrong, since the entries in the matrix $A$ are also affected by noise, being samples of GRCs. One option is to consider error-in-variables models, such as Total Least Squares (TLS), whose computation requires singular value decomposition and is presented in Andrec et al. (2005).

# Bibliography


Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular biology of the cell*. Garland Science, 5th edition.

Alon, U. (2006). *An Introduction to Systems Biology*. Mathematical and computational biology series. Chapman & Hall/CRC.

Álvarez-Buylla Roces, M. E., Martínez-García, J. C., Dávila-Velderrain, J., Domínguez-Hüttinger, E., and Martínez-Sánchez, M. E. (2018). *Modeling Methods for Medical Systems Biology*, volume 1069 of *Advances in Experimental Medicine and Biology*. Springer International Publishing AG.

Andrec, M., Kholodenko, B. N., Levy, R. M., and Sontag, E. (2005). Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. *J. Theor. Biol.*, 232:427–441.

Åström, K. J. (1980). Maximum likelihood and prediction error methods. *Automatica*, 16:551–574.

Åström, K. J. and Eykhoff, P. (1971). System identification – a survey. *Automatica*, 7:123–162.

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

Bansal, M., Della Gatta, G., and Di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22:815–822.

Barnes, D. J. and Chu, D. (2010). *Introduction to Modeling for Biosciences*. Springer.

Blainey, P., Krzywinsky, M., and Altman, N. (2014). Replication. *Nature Methods*, 11(9):879–880.

Brys, G., Hubert, M., and Struyf, A. (2004). A robust measure of skewness. *Comp. & Graph. Stat.*, 13(4):996–1017.

Brys, G., Hubert, M., and Struyf, A. (2006). Robust measures of tail weight. *Comp. Stat. & Data Anal.*, 50:733–759.

Caginalp, C. and Caginalp, G. (2017). The quotient of normal random variables and application to asset price fat tails. *Physica A*, 499:457–471.

Cho, K. H. and Wolkenhauer, O. (2003). Analysis and modelling of signal transduction pathways in Systems Biology. *Biochem. Soc. Trans.*, 31(6).

Clarke, R., Tyson, J. J., Tan, M., Baumann, W. T., Jin, L., Xuan, J., and Wang, Y. (2019). Systems biology: perspectives on multiscale modeling in research on endocrine-related cancers. *Endocrine-related cancer*, 26(6):345–368.

Degasperi, A., Birtwistle, M. R., Volinsky, N., Rauch, J., Kolch, W., and Kholodenko, B. N. (2014). Evaluating strategies to normalize biological replicates of Western blot data. *PLOS One*, 9(1):1–11.

Degasperi, A., Fey, D., and Kholodenko, B. N. (2017). Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *npj Systems Biology and Applications*, 3(20):1–9.

Díaz-Francés, E. and Rubio, F. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Stat Papers*, 54:309–323.

Fey, D., Kuehn, A., and Kholodenko, B. N. (2016). On the personalised modelling of cancer signalling. *IFAC-PapersOnLine*, 49(26):312–317.

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

Fröhlich, F., Loos, C., and Hasenauer, J. (2019). *Scalable Inference of Ordinary Differential Equation Models of Biochemical Processes. In Sanguinetti, G. and Huynh-Thu, V.A. Gene Regulatory Networks. Methods and Protocols.*, volume 1883 of *Methods in Molecular Biology*. Humana Press, New York, NY, 1st edition.

Geissen, E. M., Hasenauer, J., and Radde, N. (2019). Inference of finite mixture models and the effect of binning. *Statistical applications in genetics and molecular biology*, 18(4).

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science. Chapman & Hall, CRC, 2 edition.

Gong, C., Zhang, Y., Shankaran, H., and Resat, H. (2015). Integrated analysis reveals that STAT3 is central to the crosstalk between HER/ErbB receptor signaling pathways in human mammary epithelial cells. *Mol. Biosyst.*, 11:146–158.

Hass, H., Loos, C., Raimundez Alvarez, E., Timmer, J., Hasenauer, J., and Kreutz, C. (2019). Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, 35(17):3073–3082.

Hayya, J., Armstrong, D., and Gressis, N. (1975). A note on the ratio of two normally distributed variables. *Management Sci*, 21(11):1338–1341.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56(3):635–639.

Hood, F. E., Klinger, B., Newlaczyl, A. U., Sieber, A., Dorel, M., Oliver, S. P., Coulson, J. M., Blüthgen, N., and Prior, I. A. (2019). Isoform-specific Ras signaling is growth factor dependent. *Molecular Biology of the Cell*, 30:1108–1117.

Julier, S. J. and Uhlmann, J. K. (2002). Reduced sigma point filters for the propagation of means and covariances through nonlinear transformations. *Proceedings of the 2002 American Control Conference (IEEE Cat. No.CH37301)*, 2:887–892.

Kang, T., Moore, R., Li, Y., Sontag, E., and Bleris, L. (2015). Discriminating direct and indirect connectivities in biological networks. *Proc. Natl. Acad. Sci.*, 112:12893–12898.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc.

Kholodenko, B. N. (2006). Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology*, 7(3):165.

Kholodenko, B. N., Hancock, J. F., and Kolch, W. (2010). Signalling ballet in space and time. *Nature Reviews Molecular Cell Biology*, 11(6):414.

Kholodenko, B. N., Kiyatkin, A., Bruggeman, F. J., Sontag, E., Westerhoff, H. V., and Hoeak, J. B. (2002). Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci.*, 99(20):12841–12846.

Kholodenko, B. N. and Sontag, E. (2002). Determination of functional network structure from local parameter dependence data. *Web Archive arXiv:physics/0205003*.

Kim, H. Y., Kim, H. R., and Lee, S. H. (2014). Advances in Systems Biology approaches for autoimmune diseases. *Immune Network*, 14(2):73–80.

Koopmans, T. C. and Reiersol, O. (1950). Identification of structural characteristics. *Ann. Math. Statist.*, 21:165–181.

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

Kreutz, C. and Timmer, J. (2009). Systems biology: experimental design. *FEBS Journal*, 276:923–942.

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352.

Ljung, L. (1987). *System identification. Theory for the user*. Prentice-Hall, Inc.

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

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *J. Amer. Stat. Assoc.*, 60(309):193–204.

Marsaglia, G. (2006). Ratios of normal variables. *J. Stat. Software*, 16(4):1–10.

Martins, M. L., Ferreira Jr, S. C., and Vilela, M. J. (2010). Multiscale models for biological systems. *Current opinion in colloid & Interface Science*, 15(1-2):18–23.

Möller, Y., Siegemund, M., Beyes, S., Herr, R., Lecis, D., Delia, D., Kontermann, R., Brummer, T., Pfizenmaier, K., and Olayioye, M. (2014). EGFR-targeted TRAIL and a Smac mimetic synergize to overcome apoptosis resistance in KRAS mutant colorectal cancer cells. *PLOS One*, 9(9):1–12.

Morisugu, H., Romero, J., and Moriguchi, T. (2009). Confidence interval for the ratio of two normal variables (an application to value of time). *Interdisciplinary Information Sciences*, 15(1):37–43.

Murray, J. D. (2002). *Mathematical Biology. I. An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer.

Pham-Gia, T., Turkkan, N., and Marchand, E. (2006). Density of the ratio of two normal random variables and applications. *Communications in Statistics - Theory and Methods*, 35(9):1569–1591.

Purvis, J. E., Karhohs, K. W., Mock, C., Batchelor, E., Loewer, A., and Lahav, G. (2012). p53 dynamics control cell fate. *Science*, 336(6087):1440–1444.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., and Timmer, J. (2013). Lessons learned from quantitative dynamical modeling in Systems Biology. *PLOS One*, 8(9):1–17.

Santos, S. D. M., Verveer, P. J., and Bastiaens, P. I. H. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biol.*, 9:324–330.

Santra, T., Kolch, W., and Kholodenko, B. N. (2013). Integrating Bayesian variable selection with Modular Response Analysis to infer biochemical network topology. *BMC Syst. Biol.*, 7:57.

Santra, T., Rukhlenko, O., Zhernovkov, V., and Kholodenko, B. N. (2018). Reconstructing static and dynamic models of signaling pathways using Modular Response Analysis. *Current Opinion in Syst. Biol.*, 9:11–21.

Schadt, E. E. and Lum, P. Y. (2006). Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *Journal of Lipid Research*, 47:2601–2613.

Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J., and Klingmüller, U. (2005). Computational processing and error reduction strategies for standardized quantitative data in biological networks. *FEBS Journal*, 272:6400–6411.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear regression*. John Wiley & Sons, Inc.

Shanmugalingam, S. (1982). On the analysis of the ratio of two correlated normal variables. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 31(3):251–258.

Speth, Z., Islam, T., Banerjee, K., and Resat, H. (2017). EGFR signaling pathways are wired differently in normal 184a1l5 human mammary epithelial and MDA-MB-231 breast cancel cells. *J. Cell Commun. Signal.*, 11:341–356.

Steele, R. J. and Raftery, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. *Frontiers of statistical decision making and bayesian analysis*, 2:113–130.

Stelniec-Klotz, I., Legewie, S., Tchernitsa, O., Witzel, F., Klinger, B., Sers, C., Herzel, H., N., B., and Schäfer, R. (2012). Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS. *Mol. Syst. Biol.*, 8:601.

Taylor, S. C., Berkelman, T., Yadav, G., and Hammond, M. (2013). A defined methodology for reliable quantification of western blot data. *Mol. Biotechnol.*, 55:217–226.

Taylor, S. C. and Posch, A. (2014). The design of a quantitative Western blot experiment. *BioMed Res. Int.*, 2014:361590.

Towbin, H., Staehelin, T., and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *PNAS*, 76(9):4350–4354.

Verboven, S. and Hubert, M. (2010). Matlab library LIBRA. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:509–515.

von Luxburg, U. and Franz, V. H. (2007). A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations, and bootstrap. *Statistica sinica*, 19(3):1095–1117.

Walpole, J., Papin, J. A., and Peirce, S. M. (2013). Multiscale computational models of complex biological systems. *Annual review of biomedical engineering*, 15:137–154.

Wang, Q. (2018). Analysis of the effects of experimental data normalization on model calibration. *Study thesis.*

Weber, P., Hasenauer, J., Allgöwer, F., and Radde, N. (2011). Parameter estimation and identifiability of biological networks using relative data. *Proceedings of the 18th World Congress The International Federation of Automatic Control*, 44(1):11648–11653.

Weber, P., Hornjik, M., Olayioye, M. A., Hausser, A., and Radde, N. (2015). A computational model of PKD and CERT interactions at the trans-Golgi network of mammalian cells. *BMC Syst. Biol.*, 9(9).

Werner, H. M. J., Mills, G. B., and Ram, P. T. (2014). Cancer Systems Biology: a peak into the future of patient care? *Nat. Rev. Clin. Oncol.*, 11(3):167–176.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. John Wiley & Sons.

Zinöcker, S. and Vaage, J. T. (2012). Rat mesenchymal stromal cells inhibit T cell proliferation bot not cytokine production through inducible nitric oxide synthase. *Frontiers Immunol.*, 3(62):1–13.

# Publications of the author

Kirch, J., Thomaseth, C., Jensch, A., and Radde, N. (2016). The effect of model rescaling and normalization on sensitivity analysis on an example of a MAPK pathway model. *EPJ Nonlinear Biomedical Physics*, 4(3):1–23.

Thomaseth, C. (2012). Modeling SMS driven conversion of ceramide to sphingomyelin reveals the existence of a positive feedback mechanism. *Master thesis*.

Thomaseth, C., Fey, D., Santra, T., Rukhlenko, O., Radde, N., and Kholodenko, B. (2018). Impact of measurement noise, experimental design, and estimation methods on Modular Response Analysis based network reconstruction. *Scientific Reports*, 8(16217):1–14.

Thomaseth, C., Jensch, A., and Radde, N. (2017). Sampling-based bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines. *BMC Syst. Biol.*, 11(11):1–15.

Thomaseth, C. and Radde, N. (2016). Normalization of Western blot data affects the statistics of estimators. *IFAC-PapersOnLine*, 49(26):56–62.

Thomaseth, C. and Radde, N. (2021). The effect of normalization and error model choice on maximum likelihood estimation for a biochemical reaction network. Under review.

Thomaseth, C., Weber, P., Hamm, T., Kashima, K., and Radde, N. (2013). Modeling sphingomyelin synthase 1 driven reaction at the Golgi apparatus can explain data by inclusion of a positive feedback mechanism. *J. Theor. Biol.*, 337:174–180.