Discrimination of dynamical system models for biological and chemical processes*

Sönke Lorenz, Elmar Diederichs[1],
Regina Telgmann[1,2], Christof Schütte[1]

[1] Institut für Mathematik II, Freie Universität Berlin
Arnimallee 2-6, 14195 Berlin, Germany

[2] CiT - Computing in Technology GmbH
Oldenburger Str. 200, 26180 Rastede, Germany

May 31, 2006

Abstract

In technical chemistry, systems biology and biotechnology, the construction of predictive models has become an essential step in process design and product optimization. Accurate modelling of the reactions requires detailed knowledge about the processes involved. However, when concerned with the development of new products and production techniques for example, this knowledge often is not available due to the lack of experimental data. Thus, when one has to work with a selection of proposed models and the main tasks of early development is to discriminate these models. In this article, a new statistical approach to model discrimination is described that ranks models wrt. the probability with which they reproduce the data. The article introduces the new approach, discusses its statistical background, presents numerical techniques for its implementation and illustrates the application to examples from biokinetics.

keywords: Model discrimination, Gauss-Newton algorithm, biokinetics, dynamical systems, parameter estimation, polymerization, sensitivity analysis

Mathematical Subject Classification: 92B05, 62K05, 65P99

*E. Diederichs and R. Telgmann have been supported by the Bundesministerium für Bildung und Forschung (BMBF) within the Project Network "Mathematics for Industry". S. Lorenz acknowledges sponsering by BASF AG, Ludwigshafen, Germany.

1 Introduction

Modelling in general has become a substantial part of process design in technical chemistry and biotechnology during the last years. Detailed and predictive models, focusing on product properties and safety aspects, give the chance to successfully encounter the increasing competition on the global market by multi-objective optimization of processes and process integration models in contrast to global mass balance modelling of former years.

This basically requires an accurate modelling of chemical kinetics, ranging from standard mass products to sophisticated polymer products and biomolecules. Such kinetic models describe the reactions and reaction rates between species in a reactor on the basis of elementary reaction steps. The total reaction scheme (including side reactions and species that are both not directly measurable by experiments) with its rates and parameters is the essence of modelling work in kinetics.

The type of parametric and deterministic models exclusively considered in this article are $D-$dimensional initial value problems characterized by a set of ordinary differential equations (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = f(y(t), \theta), \quad y(t_0) = y_0. \tag{1.1}$$

with initial values y_0 and parameters $\theta^T = (\theta_1, \ldots, \theta_j, \ldots, \theta_P)$, which determine the dynamical behavior. f is called the right-hand-site (RHS) of the model. This type of model is usually being used in chemical reaction kinetics, biokinetics, systems biology or polymerization processes. Its solution at time t is denoted by $\Phi_\theta^t y_0$.

Assume that experimental data of the system under consideration is available at times $t^T = (t_1, \ldots, t_i, \ldots, t_N)$. The data $d^T = (d(t_1), \ldots, d(t_N))$ in general corresponds to measurements of model sensors $d = G(y)$. For simplicity one may assume that d means some (if not all) of the components of y. In the following all components of the data are considered. Moreover, one has to take into account that every single measurement will have a certain measurement error, i.e., every measurement $d(t_i)$ comes along with an error $\delta d(t_i)$. Typically, this error is not known explicitly. One may interpret the measurement as a *single* realization of a normally distributed random variable: its expectation value is identified with $d(t_i)$, and the deviation $\delta d(t_i)$ from the expected values is characterized by its standard deviation $\sigma_d(t_i)$. In the following, the corresponding variances σ_d^2 are abbreviated by $(\sigma_d^2)^T = \big(\sigma_d^2(t_1), \ldots, \sigma_d^2(t_N)\big)$.

However, in innovative applications, rare educt species and new production techniques are preventing to simply employ kinetic approaches or even reaction parameters from known processes. Even worse, in rapidly evolving fields like biokinetics, the experience with comparable processes is very limited and in many cases not available as quantitative knowledge. Instead, it is often necessary to identify a new, reasonable model and adapt the respective parameters on the basis of a rather limited number of experiments - in a short cycle time. At the very early stage of such a modelling process, the knowledge about the kinetics is as low as the required details of a model. Consequently, the classical textbook situation of model calibration, where merely fine tuning seems to be necessary, is only one extreme constellation in any modelling process. The main part of modelling, however, requires to work with model ideas, model alternatives of comparable quality, rough checks of the reasonability of a modelling approach, parameter estimation for intermediate models, or decisions about further experiments putting doubts on certain models.

The problem of facing many proposed models – without any kind confirmation of a single one – heavily surfaces, f.e., in the field of biokinetics. Numerous experiments consist out of many reactions with more than two reactants and catalytic components. Therefore, a huge pool of complex model candidates arises, which may differ in the number of identified reactants, parameters or catalytic components. Due to these circumstances, having an appropriate criterion for model selection is necessary to discriminate dynamical models of the form (1.1). To be more precise, if the RHS f, the initial values y_0 and repeated measurements $\{d_{(l)}(t_1), \ldots, d_{(l)}(t_N)\}_{l=1}^m$ are given, a concept of measuring the deviation between model and data by means of a functional of the following form

$$\mathcal{F}(\theta) \;=\; \text{deviation between} \left(d(t_1)_{(l)}, \ldots, d(t_N)_{(l)}\right)^T \;\text{ and }\; \left(\Phi_\theta^{t_1} y_0, \ldots, \Phi_\theta^{t_N} y_0\right)^T \tag{1.2}$$

is sought. There, a deviation can be understood in a very broad sense, ranging from, f.e., weighted residua to overlaps of probability distributions, as it will be demonstrated in this article.

This article is organized as follows: First, the general perspective of model discrimination is reviewed. Next, a new statistical approach to model discrimination is introduced where models are judged according to their probability to possibly describe the available data. Next, it is described how the new discrimination tools can be realized algorithmically and how the resulting needs can be integrated into available software platforms used in industrial modelling. Finally, the new approach is applied to some simplified examples from biokinetics. In the appendix, a brief review about the most influential approaches to model discrimination is presented, accompanied by some motivating remark to use the overlap.

2 The overlap concept

2.1 Some preliminaries on model discrimination

From a more abstract point-of-view, one can characterize the existing approaches to model discrimination using the subsequent criteria [26, 31]:

(D1) *Falsifiability*: whether there exist potential observations that are incompatible with the model, (D2) *Explanatory adequacy*: whether the theoretical account of the model helps to make sense of observed data but also established findings,

(D3) *Interpretability*: whether the components of the model, especially its parameters, are understandable and are linked to known processes,

(D4) *Faithfulness*: whether the model's ability to capture the underlying regularities comes from the theoretical principles the model purports to implement, not from the incidental choices made in its computational instantiation,

(D5) *Goodness–of–fit*: whether the model fits the observation data sufficiently well,

(D6) *Complexity and simplicity*: whether the model's description of the observed data is achieved in the simplest possible manner.

(D7) *Generalizability*: whether the model provides a good prediction of future observations.

The objective of model selection is to choose the very model out of a set of different ones (or continuum of model complexity) that performs best on future test data. However, the different discrimination criteria (D1)–(D7) cannot be applied simultaneously, since they differently validate model–data–deviations. Presently, there is no and is not going to be a general master strategy: At some point, the experimenter or modeler has to make some sort of assumption and interpretation. We review the most prominent approaches to model discrimination in some detail in the Appendix.

In the next section, it is shown how the model–data–overlap incorporates model sensitivity into the model ranking process and that it is a suitable tool for analyzing and coping with model uncertainty settings. By assessing the model's ability to take on experimental data, the model–data–overlap to be presented herein follows the discrimination strategy (D1), (D3) as well as (D5).

2.2 General statistical concept

Prior to explaining the overlap concept, some motivating arguments for its introduction as well as its major features are presented.

The core idea of the overlap approach is that the sensitivity of the set of trajectories $\{\Phi^t_{\theta,l}y_0\}^n_{l=1}$ to changes in values of the parameters θ should play a decisive role in model selection of models of the form (1.1). This has been already proposed for example in context of model validation (c.f. [35, 36, 38, 37]). However for model discrimination, this concept is based on the following four insights:

(1) Experimental data is always subject to uncertainty (e.g. random or systematic measurement errors or variations). Hence data distributions are better models than seemingly precise data values.

(2) Parameter estimation can never result in precise values. There necessarily is uncertainty which at best can be modelled by a distribution of model parameters.

(3) When considering models of type (1.1) subject to parameter distribution instead of precisely given model parameter values, we have to deal with trajectory distributions instead of single trajectories.

(4) Model discrimination for cases of type (3) should be based on the similarity between the distribution of the data and the distribution of the trajectories. A certain measure of this similarity is the overlap to be defined below.

Without information about the shape of the distribution of the parameters, model discrimination, based on pure comparison of best fits of the competing models to the available data, is unsatisfactory and constructions like confidence intervals or experimental design are even unconvincing. Therefore, the authors conclude that for model selection approaches that employ distributed parameters, the model parameters θ and at least their variances have to be estimated within the same step. This is a systematic difference to the so called textbook model discrimination, e.g. the residuum concept (see appendix) and will be illustrated in the next figure. The authors call the parameters of distribution of the model parameters θ the hyper-parameters of the model.

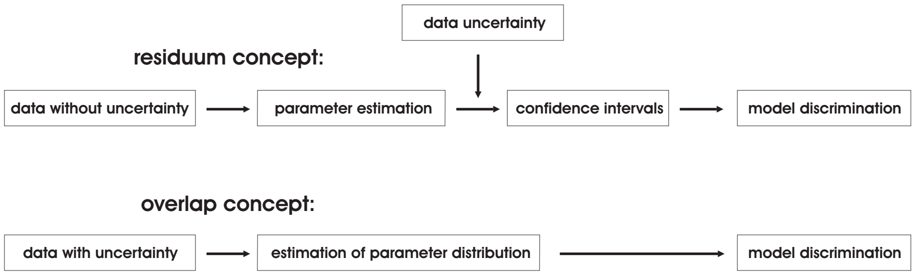


Figure 2.1: scheme of overlap concept

The assumption of distributed model parameters θ seems to be quite natural. In new experiments f.e. some parameters might be hitherto unknown and hence uncontrolled. Then there is no guarantee to have invariant parameters during the same measurement. Other parameters might show local inhomogeneities or are perturbed by non-stationary noise. If the experimenter has accepted distributed model parameters the systematic difference between measurement errors as a data property and data deviations by parameter distributions vanishes.

2.2.1 Introduction of the overlap-concept

Suppose the measurement process does not provide fixed values for all observed quantities but statistical distributions instead. This may come from physical or chemical properties of the dynamical system, from the required effort in time or the financial implications of repeated measurements. Consequently if repetition of the measurement is not advisable or will not improve the result, one should implement statistical aspects into the models under consideration also. Therefore, one has to introduce a density π_θ governing the statistics of the parameters in subsequent realizations. The family of densities π_θ should be carefully chosen on basis of prior experiences with the investigated dynamical system. In each single realization, θ is selected due to π_θ resulting in a single trajectory $\Phi^t_\theta y_0$. Thus, the parameter density π_θ induces a distribution of trajectories $\Phi^t_\theta y_0$ in the state space, developing simultaneously from the joint initial state y_0. This *model variability*, denoted $\mathcal{M}_t$ in the following, then has to be compared to the *variability* $\mathcal{D}_t$ of the measured data.

The expression *model variability* shall reflect the general ability of the model's trajectory to change by means of parameter as well as initial values perturbations. $\mathcal{M}_t$ is a positive measure defined as (c.f. [29])

$$\begin{aligned} M \quad &: \quad \Gamma \times \mathbb{R} \to [0, a] \qquad a \in \mathbb{R}^+ \setminus \{0\} \\ \mathcal{M}_t(A) \quad &= \quad \frac{1}{C(t)} \int\limits_\Theta \mathbf{1}_A(\Phi^t_\theta y_0) \pi_\theta \,\mathrm{d}\theta \end{aligned} \tag{2.3}$$

for any set $A \subset \Gamma$, where Γ denotes the entire state space, π_θ the parameter density. The characteristic function $\mathbf{1}_A(x)$ is given by

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$

In order to interpret (2.3) later as a distribution, one needs to normalize the a function $C(t) \in L^\infty$[1]. By that, one also attaches a stochastic interpretation to the purely deterministic ODE-setting of (1.1). The definition of $\mathcal{D}_t$ is analogously to that of $\mathcal{M}_t$, but realized by means of data variances. For illustration consider figure 2.2.

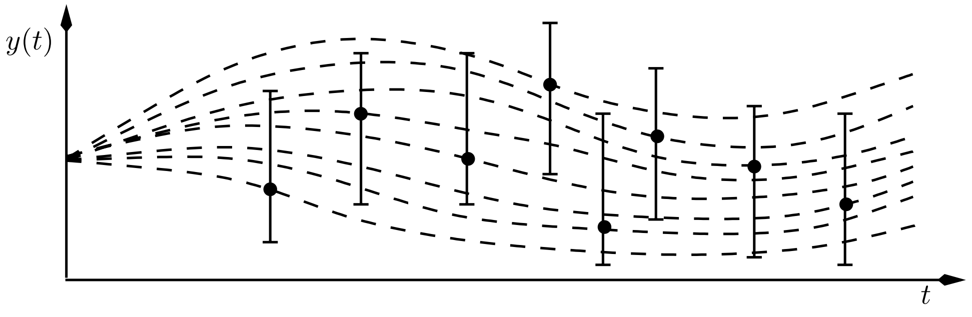


Figure 2.2: Model variability validates model–data–reproducibility: The black dots are measured data $d(t)$ with attached error bars representing some confidence interval of the data variability $\mathcal{D}_t$. Each measured data point can be explained by a single trajectory, representing a realization for θ from π_θ. Additionally, these trajectories also "pass" through confidence intervals of other data points and therefore validate the corresponding data also.

[1] It will be specified later in section 2.2.2.

Matching model variability and data spread reveals the local information of a proper model–data–fit. In other words, the *overlap* of the model variability $\mathcal{M}_t$ and the data variability $\mathcal{D}_t$ describes the goodness of data–model–reproducibility. Having a pool of proposed models, one can discriminate between them by picking the one with the highest overlap value.

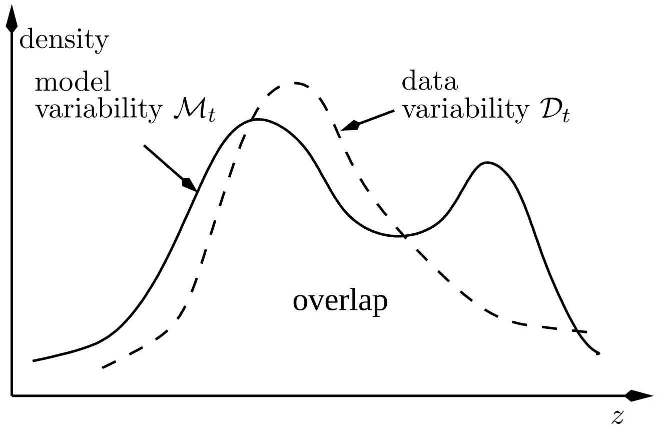


Figure 2.3: Overlap of model variability $\mathcal{M}_t$ and data variability $\mathcal{D}_t$ of measured data at a single time t. As described in section 2.2.2 $\mathcal{M}_t$ and $\mathcal{D}_t$ are normalized wrt. the Euclidean norm.

Consequently at the very beginning of the modelling process the overlap approach is useful to guide a refinement procedure, starting with some noisy data and a bunch of candidate models: Usually one wants to estimate the parameters for each model, perform additional experiments and investigate the changes of the model–data–fit, when new data is added. According to the overlap approach one can thin out the set of models step by step by estimating the parameter distributions and comparing the changes of the model–data–overlap for each candidate model.

Within the overlap approach, the local parameter estimation is an embedded part within model validation and discrimination. The hyper-parameters of the density π_θ of model parameters are chosen, such that the overlap between data and model $\mathcal{F}_\mathcal{O}$

$$\mathcal{F}_\mathcal{O} = \text{overlap of data and model variability} \tag{2.4}$$

is maximal. In contrast to the well established parameter estimation methods as least-square or maximum likelihood estimation, not the global sum of pointwise distance between the model's trajectory and data is considered, but local matching of the corresponding distributions.

In the following one has to distinguish between the residual case and the overlap case. Let $\theta_\mathcal{R}$ be the optimal model parameters in the sense of $\mathcal{F}_\mathcal{R}$, the deviation functional in the sense of (1.2) of the least-square approach[2]. Associated with

[2] This and others goodness–of–fit functionals are introduced in the Appendix

them is a single trajectory, $y_{\mathcal{R}}(t) = \Phi^t_{\theta_{\mathcal{R}}} y_0$. In contrast, one has to consider the optimal distribution $\mathcal{M}_t(\pi_{\mathcal{O}})$ resulting from the optimal parameter of the density $\pi_{\mathcal{O}}$ of model parameters. If one wants to select a single trajectory representing this distribution, one should take the average trajectory

$$y_{\mathcal{O}}(t) = \mathbb{E}\big[\mathcal{M}_t(\pi_{\mathcal{O}})\big], \tag{2.5}$$

where the expectation $\mathbb{E}[\cdot]$ is taken for each instance t separately.

To illustrate the conceptual difference between the overlap $\mathcal{F}_{\mathcal{O}}$ and the classical residual approach $\mathcal{F}_{\mathcal{R}}$, consider two different proposed models like in figure 2.4.

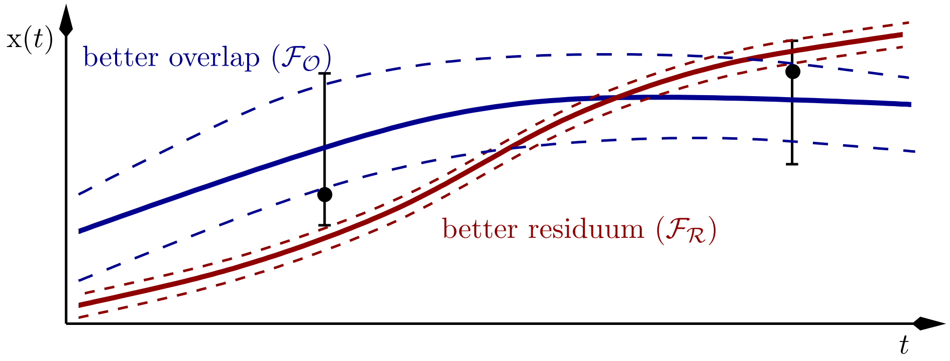

x(t)
better overlap ($\mathcal{F}_{\mathcal{O}}$)
better residuum ($\mathcal{F}_{\mathcal{R}}$)
t

Figure 2.4: Different residual and overlap interpretation: The 95 %-confidence intervals of the data variability $\mathcal{D}_t$ are symbolized by error bars, the inner 95 %-quantile of the model variability $\mathcal{M}_t(\pi_{\mathcal{O}})$ by dashed strips around each fitted trajectory. A smaller residuum (lower model) does not imply a large model–data–overlap (upper model) and vice versa.

The lower one of the proposed models shows the smaller residuum, but due to its low model variability, it does not reproduce the data distributions as well as the other one. In the residual framework, that model is preferred. In the overlap interpretation, however, one would prefer the other one with the higher model variability. Due to its higher model variability it matches the data distribution better and has therefore the higher capability of reproducing the data distributions.

Compared to figure 2.4, a more extreme setting is shown in figure 2.5. Due to non-existing model variability for the left model at the measuring point $t = 1$, there is a vanishing probability to reproduce any data given by $\mathcal{D}_{t=1}$. Assuming a negligible measurement error, the deviation of the deviation between the model and data means cannot be explained by parameter sensitivity. This interpretation challenges the quality of the model. On the other hand, the right model is capable of taking values that can be justified by $\mathcal{D}_{t=1}$.

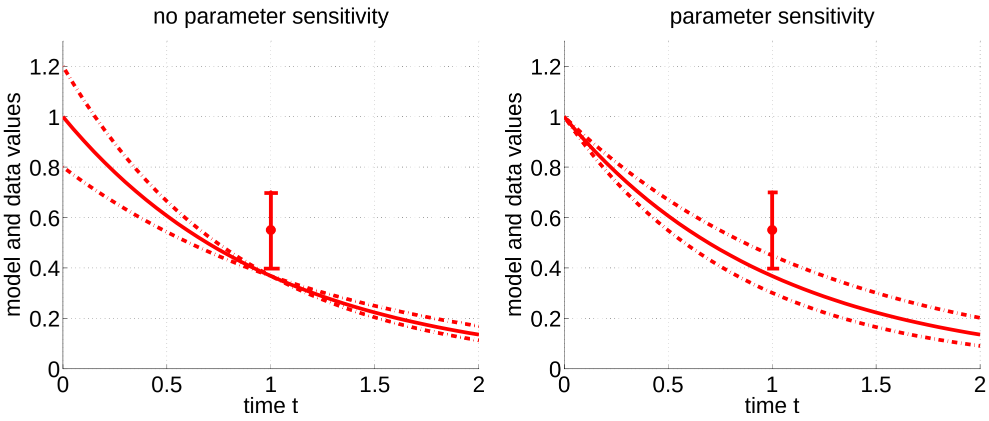


Figure 2.5: In both examples, the residuum is the same. The ability to reproduce the data is not given in the left, but in the right model.

Both illustrations show that the distance between data and model alone does not give sufficient information about the quality of the model–data–fit in the sense of reproducibility.

Next, a general advantage of the overlap approach in comparison to the distribution-free least square method is described. Consider the variance of the least-square-estimator, as introduced in (A.34):

$$\mathrm{Var}(\hat{\theta}) \quad = \quad \left(\mathbf{J}^T \mathbf{\Sigma}_{\mathcal{D}} \mathbf{J}\right)^{-1}$$

Since the variance-covariance matrix $\mathbf{\Sigma}_{\mathcal{D}}$ of the data is symmetric and the variance is invariant under orthogonal transformation, one can express $\mathrm{Var}(\hat{\theta})$ by means of the eigenvalues of $\left(\mathbf{J}^T \mathbf{\Sigma}_{\mathcal{D}} \mathbf{J}\right)^{-1}$. Hence the variance of the estimator $\hat{\theta}$ is proportional to the empirical variance of column entries of $\mathbf{J}$ (c.f. [5]). Suppose now one has to minimize $\mathcal{F}_{\mathcal{R}}(\theta)$ with the least square method. Than for least square approaches, the influence of data variances $\sigma_{d,j}(t_i)$ on the estimator $\hat{\theta}_j$ is the same for all i during the optimization, whereas the overlap method will conduct a local optimization of the objective functional for every single time step t_i. Hence the optimization solution within the residual framework, but not the overlap approach, is sensible to outliers.

2.2.2 Overlap notation

Analytically, the overlap will be defined as the scalar product of two measures: the model variability $\mathcal{M}_t$ and data variability $\mathcal{D}_t$ at time t. To this end, it is claimed that the first and the second moment of $\mathcal{M}_t$ and $\mathcal{D}_t$ exist and that the scalar product of the measures is limited by 1. Consequently, it is imposed that $\mathcal{M}_t, \mathcal{D}_t \in \mathbf{L}_2$. Therefore by the Cauchy-Schwarz inequality one gets

$$\mathcal{F}_{\mathcal{O}}(t) := \langle \mathcal{M}_t, \mathcal{D}_t \rangle_2 \leq \|\mathcal{D}_t\|_2 \, \|\mathcal{M}_t\|_2 \leq 1. \tag{2.6}$$

To ensure the required normalization, for the model variability $\mathcal{M}_t$ at time t introduced in section 2.2.1

$$\mathcal{M}_t(A) = \frac{1}{C(t)} \int\limits_{\Theta} \mathbf{1}_A(\Phi_\theta^t y_0)\pi_\theta \,\mathrm{d}\theta, \tag{2.7}$$

the function $C(t)$ is chosen such that $\|\mathcal{M}_t\|_2 = 1$ for each t. The choice of the Euclidean norm $\|\cdot\|_2$ is appropriate for most of the data sets. Hence in the Kolmogorov sense (2.7) it is the probability that the data distribution can be reproduced by the model wrt. its parameter distribution. In other words, within the overlap approach one chooses the very model whose probability to reproduce the data is the highest.

Unless the experimental setting dictates something else, the data distribution is assumed to be normal with $\mathcal{N}(\mu_d(t), \sigma_d^2(t))$. Then $\mathcal{D}_t$ is defined as

$$\mathcal{D}_t(x) \sim \frac{1}{\sqrt[4]{\pi_\theta}\,\sqrt{\sigma_d(t)}}\, e^{-\frac{(x-\mu_d(t))^2}{2\sigma_d^2(t)}}. \tag{2.8}$$

To calculate (2.7), one needs to model the parameter density π_θ. For illustration it is a convenient assumption to use a normal density also such that $\pi_\theta \ \sim\ \mathcal{N}(\mu_\theta,\ \mathbf{\Sigma}_\theta)$ with variance-covariance matrix

$$\mathbf{\Sigma}_\theta = \begin{pmatrix} \sigma_{\theta_1}^2 & 0 & \cdots\cdots\cdots & & 0 \\ 0 & \sigma_{\theta_2}^2 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & \cdots & 0 & \sigma_{\theta_{P-1}}^2 & 0 \\ 0 & \cdots\cdots & & 0 & \sigma_{\theta_P}^2 \end{pmatrix}. \tag{2.9}$$

and expectation value μ_θ. In order to propagate the variances, the symmetric variance-covariance matrix is transformed into its diagonal form. This causes no constraints, since variances are invariant wrt. orthogonal transformation. In contrast to classical deviation functionals like $\mathcal{F}_\mathcal{R}$ or $\mathcal{F}_\mathcal{M}$ in (A.31) or (A.35) respectively, the overlap functional $\mathcal{F}_\mathcal{O}$ does not merely depend on the data $d(t)$ and model trajectory values $\Phi_\theta^t y_0$, but also directly on the measurement standard deviations $\sigma_d(t)$ as well as on the parameter variances $\mathbf{\Sigma}_\theta$:

$$\mathcal{F}_\mathcal{O} = \mathcal{F}_\mathcal{O}(\mu_\theta, \mathbf{\Sigma}_\theta, \mu_d(t), \sigma_d(t)). \tag{2.10}$$

In the context of the overlap optimization, parameter estimation (PE) means to choose the parameters μ_θ and $\mathbf{\Sigma}_\theta$ in such a way that the overlap $\mathcal{F}_\mathcal{O}$ is maximal. Conversely, determining the model–data–overlap and validating the model wrt. the data means to conduct a PE. In comparison to the estimation of parameter uncertainties by means of confidence intervals a posteriori, the estimated parameter values μ_θ and their variances $\mathbf{\Sigma}_\theta$ are a result within the process of the overlap–model ranking and not independent from the proposed models. This is different compared to the classical estimation and discrimination approaches, where PE and model discrimination are conducted in successive steps. Moreover, in the overlap approach the parameter variance is an indication for the parameter sensitivity. Hence the overlap model discrimination concept employs the effect of a parameter perturbation $\delta\theta$

$$\theta_0 \mapsto \theta_0 + \delta\theta$$

on the model trajectory

$$y(t) \mapsto y(t) + \delta y(t).$$

2.3 Linear propagation of parameter variability

As mentioned in section 2.2.1, the parameter θ is regarded as a normally distributed random variable that is determined by $\pi_\theta \sim \mathcal{N}(\mu_\theta, \mathbf{\Sigma}_\theta)$. For complex scenarios, it is very challenging to sample the entire parameter space and propagate the resulting ensemble of trajectories within a reasonable and justifiable amount of time. Since the overlap functional has to be evaluated several times within the overlap optimization, the exact but nonlinear propagation may require tremendous computation time.

Therefore, one approximates the model variability $\mathcal{M}_t$ by considering the linearized propagation of the initial parameter perturbation $\delta\theta$ by means of the sensitivity matrix P

$$\delta y(t) = P(t;\theta_0)\,\delta\theta, \tag{2.11}$$

which is the Jacobian of the flow with respect to the parameter θ

$$P(t;\theta_0) = D_\theta\,\Phi_\theta^t y_0|_{\theta=\theta_0} = \mathbf{J}(\theta_0, t) \tag{2.12}$$

and fulfills the sensitivity equation for initial value problems (1.1)

$$P'(t;\theta_0) = \frac{\partial}{\partial y} f(y(t),\theta_0) P(t;\theta_0) + \frac{\partial}{\partial \theta} f(y(t),\theta_0)|_{\theta=\theta_0}$$

with $P(t_0,\theta_0) = 0$ (c.f. [15]).

Since the initial parameter distribution is supposed to be normal and is propagated linearly, the gained model variability is therefore normal also (c.f. [6]). Consequently, it suffices to propagate its mean and its standard deviation (c.f. [10]). In the implementation to be presented, the mean is exactly propagated by the trajectory $\Phi_\theta^t y_0$, while the variance-covariance matrix of the model variability is given by

$$\mathbf{\Sigma}_{\mathcal{M}}(\mu_\theta, \mathbf{\Sigma}_\theta, t) \quad = \quad \mathbf{J}(\theta,t)\,\mathbf{\Sigma}_\theta\,\mathbf{J}(\theta,t)^T. \tag{2.13}$$

The variance of the k^{th} dimension of the model variability at time t is the k^{th} diagonal entry of the variance-covariance matrix in (2.13) and is denoted by $\mathbf{\Sigma}_k(\mu_\theta, \mathbf{\Sigma}_\theta, t)^2$. They are calculated by using (2.13) and (2.9)

$$\mathbf{\Sigma}_k(\mu_\theta, \mathbf{\Sigma}_\theta, t)^2 = \sum_{j=1}^{P} \left(\frac{\partial}{\partial \theta_j} \left(\Phi_\theta^t y_0\right)_k \right)^2 \sigma_{\theta,j}^2 \Bigg|_{\theta_j=\theta_{0,j}}. \tag{2.14}$$

Consequently, the *linear overlap*[3] functional $\mathcal{F}_{\mathcal{L}}$ at time t is given by

$$\begin{aligned}\mathcal{F}_{\mathcal{L}}(\Phi_\theta^t y_0, \mathbf{\Sigma}_{\mathcal{M}}(\mu_\theta, \mathbf{\Sigma}_\theta, t), \mu_d(t), \sigma_d(t)) \quad &= \\ \sum_{k=1}^{D} \sqrt{\frac{2\,\sigma_{d,k}(t)\,\mathbf{\Sigma}_k(\mu_\theta, \mathbf{\Sigma}_\theta, t)}{\sigma_{d,k}(t)^2 + \mathbf{\Sigma}_k(\mu_\theta, \mathbf{\Sigma}_\theta, t)^2}}\ \exp &\left\{ \frac{-\left(\Phi_\theta^t y_0 - d_k(t)\right)^2}{2[\sigma_{d,k}(t)^2 + \mathbf{\Sigma}_k(\mu_\theta, \mathbf{\Sigma}_\theta, t)^2]} \right\} \end{aligned} \tag{2.15}$$

The linear overlap $\mathcal{F}_{\mathcal{L}}$ in (2.15) is calculated in each direction of the state space, since only information about the data is available.

2.4 Illustration for linear initial value problems

For a special class of systems (1.1), namely linear initial value problems,

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\, y(t,\theta) = \mathbf{J}(\theta)\, y(t) + b(\theta) \quad \text{with} \quad y(0) = y_0, \tag{2.16}$$

the linear overlap can be calculated using the analytical solution

$$\Phi_\theta^t y_0 = \exp\left(t\mathbf{J}(\theta)\right) y_0 + \mathbf{J}(\theta)^{-1} \left(\exp\left(t\mathbf{J}(\theta)\right) - \mathbf{1}\right) b(\theta). \tag{2.17}$$

For reasons of better readability, the dash ' shall denote the partial derivative with respect to θ and the parameter dependency with respect to θ is omitted. Then the flow derivative, needed for $\mathbf{\Sigma}_{\mathcal{M}}(\mu_\theta, \mathbf{\Sigma}_\theta, t)$ in (2.14), can be written as

$$\begin{aligned}\tfrac{\partial}{\partial\theta}\, \Phi_\theta^t y_0 \quad = \quad & \exp\left(t\mathbf{J}\right)' y_0 \,-\, \mathbf{J}^{-1}\mathbf{J}'\mathbf{J}^{-1} \exp\left(t\mathbf{J}\right) b + \mathbf{J}^{-1}\mathbf{J}'\mathbf{J}^{-1}\, b \\ & + \mathbf{J}^{-1} \exp\left(t\,\mathbf{J}\right)' b \,+\, \mathbf{J}^{-1} \exp\left(t\mathbf{J}\right) b' \,-\, \mathbf{J}^{-1}\, b'. \end{aligned} \tag{2.18}$$

Example. The two proposed models M_1

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} -2\theta & 2 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} -6 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} Y(t_0) \\ Y(t_0) \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \end{pmatrix}, \tag{2.19}$$

and model M_2

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} -2 & 2\theta \\ -1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} -6 \\ \theta \end{pmatrix}, \quad \begin{pmatrix} X(t_0) \\ Y(t_0) \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \end{pmatrix}, \tag{2.20}$$

are to be discriminated. Both models coincide for $\theta = 1$ (see figure 2.6).

[3] In order to distinguish between the overlap calculated with a linearly propagated model variability to the one with exactly propagated one, the terms *linear* as well as *nonlinear* overlap are introduced.

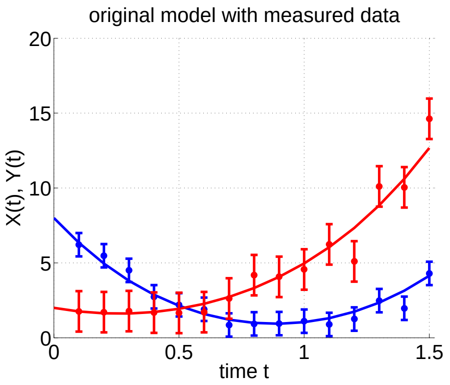


Figure 2.6: Model and data plot: For $\theta = 1$, the trajectories of M_1 and M_2 coincide. These trajectory values were taken to produce the data by perturbing them. The data is symbolized by points with attached error bar.

The data is produced artificially by taking trajectory values for $\theta = 1$ at some instances t_k and perturbing them. The variance for perturbing at each t_j was set to be proportional to the variance of the model variability $\mathbf{\Sigma}_{\mathcal{M}}(\mu_\theta = 1, \mathbf{\Sigma}_\theta = 0.15, t = t_j)$ of model 1 at each point t_j. Further, the virtual experimenter assumes the data standard deviations for the X- and Y-component to be $\sigma_X(t) = \sigma_X = \sqrt{0.25}$ and $\sigma_Y(t) = \sigma_Y = \sqrt{0.75}$, respectively.

After this generation of data, the target functionals $\mathcal{F}_{\mathcal{L}}$ from (2.15) as well as $\mathcal{F}_{\mathcal{R}}$ from (A.31) were numerically optimized with respect to $\mu_{\theta,\mathcal{L}}$, $\mathbf{\Sigma}_{\theta_{\mathcal{L}}}$ and $\theta_{\mathcal{R}}$, respectively.

	$\mu_{\theta,\mathcal{L}}$	$\mathbf{\Sigma}_{\theta,\mathcal{L}}$	$\mathcal{F}_{\mathcal{L}}$ in X	$\mathcal{F}_{\mathcal{L}}$ in Y	$\mathcal{F}_{\mathcal{L}}$ total
M_1	0.934	0.367	83.1 %	72.3 %	77.7 %
M_2	0.744	1.433	58.4 %	74.4 %	66.4 %
	$\theta_{\mathcal{R}}$	$\overline{\sigma}_{\mathcal{R}}$	ci	gof.	$\mathcal{F}_{\mathcal{R}}$
M_1	0.910	0.036	0.073	0.969	0.196
M_2	0.913	0.024	0.048	0.974	0.178

Table 2.1: Linear overlap parameter estimation by $\mathcal{F}_{\mathcal{L}}$ and residual parameter estimation $\mathcal{F}_{\mathcal{R}}$ for model M_1 and M_2. (notation: $\overline{\sigma}_{\mathcal{R}}$ = standard error, ci = half length of the 95 % confidence interval, gof. = goodness of fit in the classical χ^2_p-sense),

The linear overlap optimization by $\mathcal{F}_{\mathcal{L}}$ favors model M_1 over M_2. The results show significantly different estimated values of $\mu_{\theta,\mathcal{L}}$ and $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ for both models. Whereas $\mu_{\theta,\mathcal{L}}$ for model M_1 is reasonably close to 1, the parameter used to produce the data, $\mu_{\theta,\mathcal{L}}$ for model M_2 differs significantly and shows a very high parameter variance $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ in addition.

In contrast, the residual case leads to a different conclusion. For both models, the estimated parameters $\theta_{\mathcal{R}}$ are close together, the residual $\mathcal{F}_{\mathcal{R}}$ as well as the goodness of fit (gof.) indicate almost the same quality of fit. Only the standard error $\overline{\sigma}_{\mathcal{R}}$, corresponding to the diagonal entry of $\left(\mathbf{J}^T\mathbf{\Sigma}_{\mathcal{D}}^{-1}\mathbf{J}\right)^{-1}$ defined in (A.34), distinguishes the parameter $\theta_{\mathcal{R}}$ in M_2 to be estimated more precisely. However, in neither case, the 95%-confidence interval does include the parameter $\theta = 1$, originally used for perturbing the data. At last, the example shows the different information and interpretation of the parameter variance $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ on the one and the standard error $\overline{\sigma}_{\mathcal{R}}$ on the other hand. In the overlap concept, the model variability $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ for θ is smaller for model M_1, whereas the relation is changed in the residual framework, where the standard error $\overline{\sigma}_{\mathcal{R}}$ is smaller for $\theta_{\mathcal{R}}$ of model M_2.

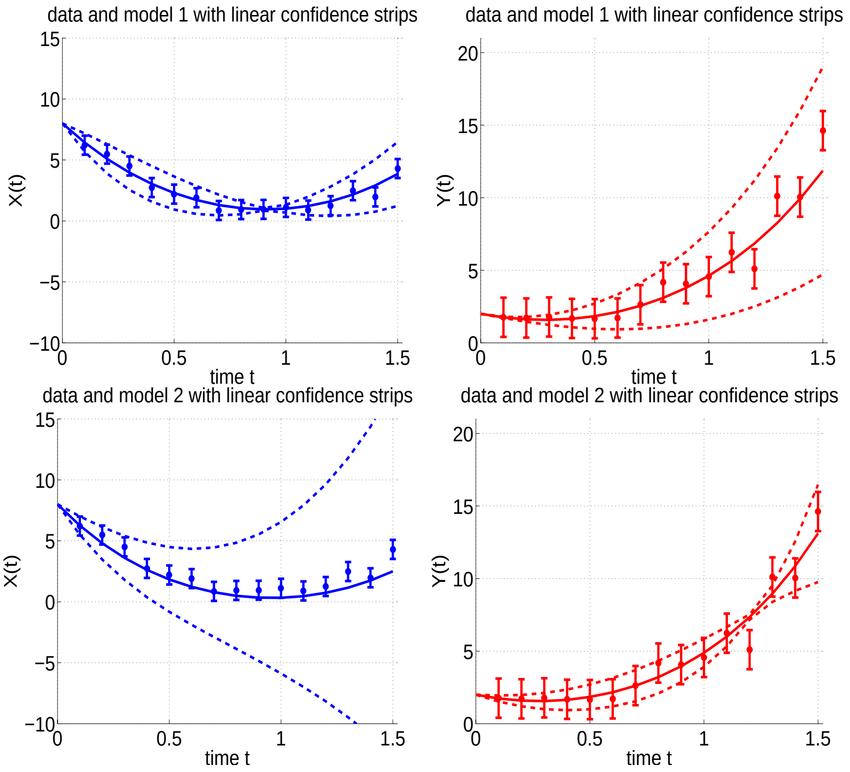


Figure 2.7: Optimal overlap of data error (symbolized by error bars) and linear model variability (symbolized by the 95%-confidence strips) for model M_1 (top two pictures) and M_2 (bottom).

The concept of model–data–fitting for the presented example is illustrated in figure 2.7. In both components, the qualitative course of the model variability strips is different. For component X of model M_1 and component Y of model M_2 the parameter sensitivity vanishes at some instances, whereas for other components, the strips are diverging. Due to the extremely diverging strips for the X-component of model M_2 the overlap compared to M_1 is worse.

2.5 Comparison to non-linear propagation

In section 2.3, it was argued that in the typical application setting, the long-lasting computation time prevents us from computing the exact parameter distribution propagation for arbitrary models. This consideration has led us to the linear overlap functional $\mathcal{F}_\mathcal{L}$ of (2.15). For very small systems, however, like model M_1 and M_2, it is justifiable to calculate $\mathcal{F}_\mathcal{O}$ instead of $\mathcal{F}_\mathcal{L}$ by sampling the parameter density π_θ, propagating the trajectories and re-assembling the model variability $\mathcal{M}$ as it is described in (2.3).

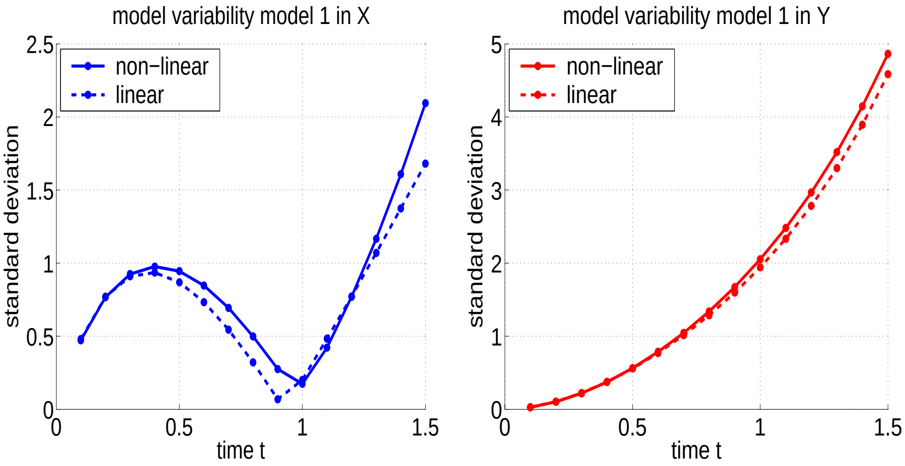
If we take the model parameter θ and its variances $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ wrt. $\mu_{\theta,\mathcal{L}}$ from table 2.1, the resulting overlap $\mathcal{F}_\mathcal{O}$ is – as expected – different to the one of $\mathcal{F}_\mathcal{L}$. As shown in table 2.2, in some cases the exact propagation yields a larger, in other ones a smaller value $\mathcal{F}_\mathcal{O}$ compared to the approximation $\mathcal{F}_\mathcal{L}$.

For the shown example, the model variability $\mathcal{M}_t$ was calculated by means of parameter sampling and consequent trajectory calculation. Due to the simple structure of the example, a simplex algorithm was used as the optimization algorithm. In each iteration step of the optimizer, a sample of 500.000 parameters was drawn. Then for each drawn parameter, the ODE in question (2.19) or (2.20) was solved, and then $\mathcal{M}_t$ of (2.7) assembled.

	in X	in Y	total
M_1 for $\mathcal{F}_\mathcal{L}$	83.1 %	72.3 %	77.7 %
M_1 for $\mathcal{F}_\mathcal{O}$	79.6 %	73.1 %	76.7 %
M_2 for $\mathcal{F}_\mathcal{L}$	58.4 %	74.4 %	66.4 %
M_2 for $\mathcal{F}_\mathcal{O}$	47.8 %	61.7 %	54.8 %

Table 2.2: Comparison of linear overlap $\mathcal{F}_\mathcal{L}$ and nonlinear overlap $\mathcal{F}_\mathcal{O}$ using the optimal parameters for $\mathcal{F}_\mathcal{L}$ in table 2.1.

To understand the reasons of these deviations, one can, f.e., compare the standard deviations of the linearly and nonlinearly propagated model variability as shown in figure 2.8. In some cases the deviations almost coincide (c.f. Y-component of M_1), are moderately diverging (c.f. X-component of M_2 or X-component of M_1) or structurally differ (c.f. Y-component of M_2).



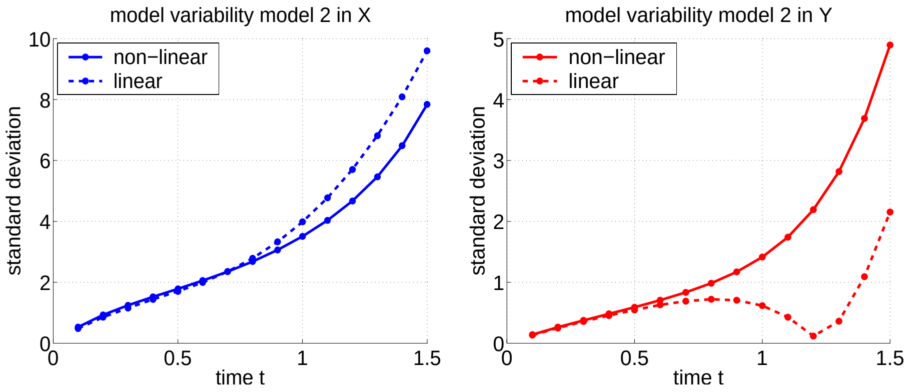


Figure 2.8: Comparison of the standard deviations of the linear and nonlinear model variability for the models M_1 and M_2.

Not only the standard deviations can differ, but also the model variability $\mathcal{M}$ can become highly non-normal. The normal property, however, was essential in the construction of $\mathcal{F}_{\mathcal{L}}$. In some cases, the normal property of the distribution is maintained as shown in figure 2.9.

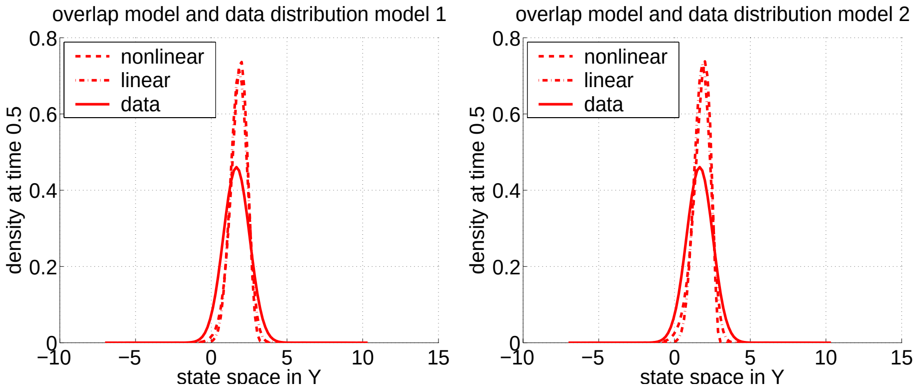


Figure 2.9: Example where non-linearly propagated model variability can be regarded as normal. Two examples for the state space in Y for model M_1 (left) and M_2 (right) are shown.

In other ones, the nonlinearly propagated model variability $\mathcal{M}$ is highly non-normal, as exemplarily seen in figure 2.10.

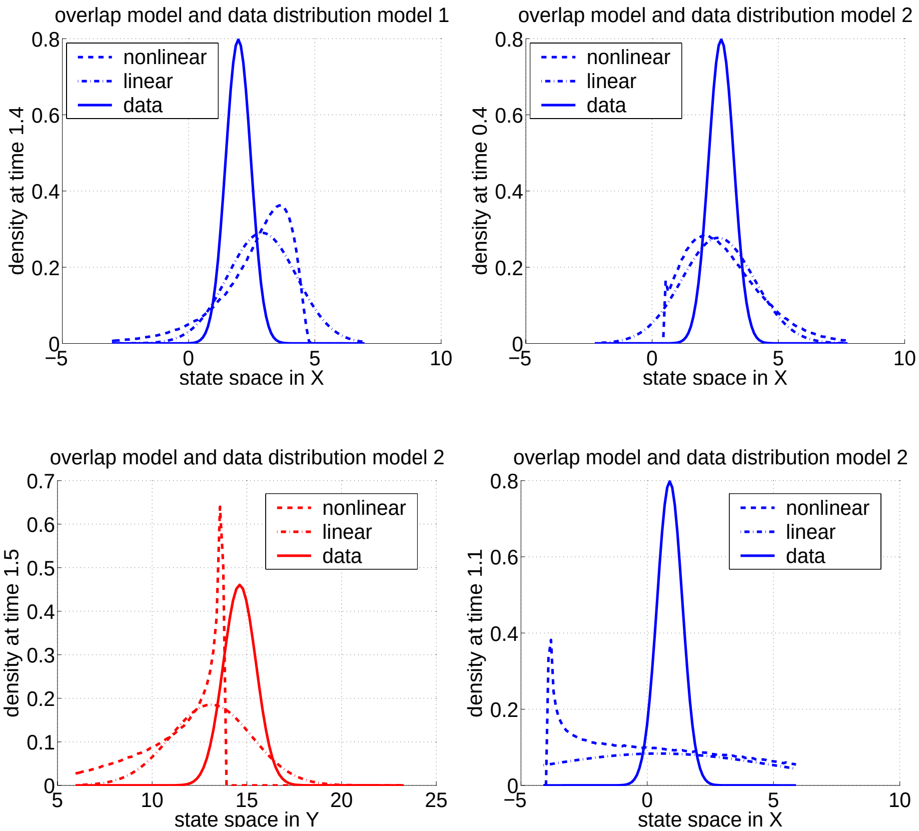


Figure 2.10: Examples where the non-linearly propagated model variability cannot be regarded as normal.

The documented effects result in different overlaps as shown in figure 2.11. The general qualitative overlap curves of the linear and the non-linear overlap are almost the same. The $X-$components in both models show a good model–data–overlap at the beginning, in the case of model M_1 roughly 1. For model M_1 it is maintained over the time course, whereas it is deteriorating for model M_2. For $Y-$component for both models show a small model–data–overlap at the beginning and at the end, but a large one in the middle. The non-linear effects of the model variability propagation seems to have a larger impact for M_2. In case of model M_1, the approximated model variability seems to be sufficient, very good coherence can be seen there, especially in the Y-component of M_1.

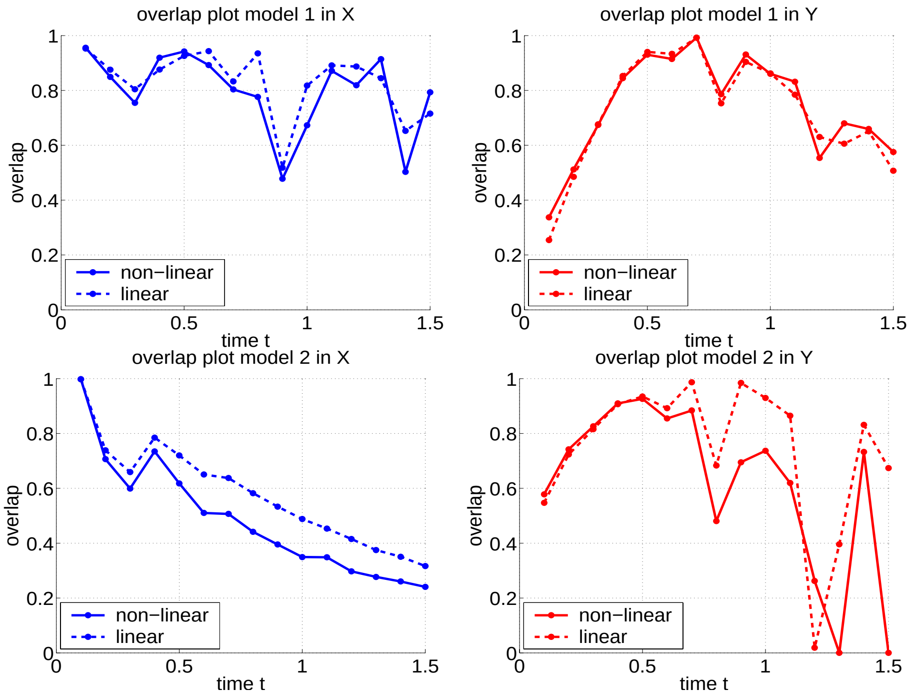


Figure 2.11: Overlap curves over time for M_1 and M_2 and each component: The linear overlap $\mathcal{F}_\mathcal{L}$ and the "exact" overlap $\mathcal{F}_\mathcal{O}$ are shown.

Non-surprisingly, the different propagation behavior results in different optimal parameters. The resulting optimal parameters for $\mathcal{F}_\mathcal{L}$ and $\mathcal{F}_\mathcal{O}$ are documented in table 2.3. The optimal parameters can differ significantly (see $\mu_{\theta,\mathcal{L}}$ versus $\mu_{\theta,\mathcal{O}}$ for M_2). Nevertheless, the qualitative result is the same: the model M_1 is discriminated to be the appropriate one.

	$\mu_{\theta,\mathcal{L}}$	$\Sigma_{\theta,\mathcal{L}}$	$\mathcal{F}_\mathcal{L}$ in X	$\mathcal{F}_\mathcal{L}$ in Y	$\mathcal{F}_\mathcal{L}$ total
M_1	0.934	0.367	83.1 %	72.3 %	77.7 %
M_1	0.916	0.312	80.0 %	73.2 %	76.8 %
	$\mu_{\theta,\mathcal{O}}$	$\Sigma_{\theta,\mathcal{O}}$	$\mathcal{F}_\mathcal{O}$ in x	$\mathcal{F}_\mathcal{O}$ in y	$\mathcal{F}_\mathcal{O}$ total
	$\mu_{\theta,\mathcal{L}}$	$\Sigma_{\theta,\mathcal{L}}$	$\mathcal{F}_\mathcal{L}$ in X	$\mathcal{F}_\mathcal{L}$ in Y	$\mathcal{F}_\mathcal{L}$ total
M_2	0.744	1.433	58.4 %	74.4 %	66.4 %
M_2	0.978	0.540	47.9 %	74.2 %	61.3 %
	$\mu_{\theta,\mathcal{O}}$	$\Sigma_{\theta,\mathcal{O}}$	$\mathcal{F}_\mathcal{O}$ in x	$\mathcal{F}_\mathcal{O}$ in y	$\mathcal{F}_\mathcal{O}$ total

Table 2.3: Comparison of PE results of linear overlap $\mathcal{F}_\mathcal{L}$ and "exact" overlap $\mathcal{F}_\mathcal{O}$ for models M_1 and M_2.

3 Algorithmic realization of the overlap concept

The calculation of the overlap $\mathcal{F}_\mathcal{O}$ according to (2.6) includes the calculation of the model variability in (2.7). As mentioned before, this would require to construct a statistically large trajectory bundle. For practical purposes, the consideration are restricted to the *linear* overlap concept introduced in section 2.3, thus replacing $\mathcal{F}_{\mathcal{O}}$ by $\mathcal{F}_{\mathcal{L}}$ of (2.15). Due to simplicity, the notation is slightly redefined from now on: $\Delta\theta$ denotes the diagonal entries of $\mathbf{\Sigma}_\theta$, with $\mathbf{\Sigma}_\theta = \mathrm{diag}\left(\sigma^2_{\theta_1}, ...\sigma^2_{\theta_j}, ...\sigma^2_{\theta_P}\right)$ and θ_0 is the expectation $\mu_{\theta,\mathcal{L}}$.

The algorithmic problem is to determine the model parameters θ and their variances $\mathbf{\Sigma}_{\theta,\mathcal{L}}$ wrt. $\mu_{\theta,\mathcal{L}}$ so that the linear overlap is maximal, i.e.

$$(\theta_0, \Delta\theta) := \underset{(\mu_{\theta,\mathcal{L}}, \mathbf{\Sigma}_{\theta,\mathcal{L}})}{\arg\max} \left(\mathcal{F}_{\mathcal{L}}(\mu_{\theta,\mathcal{L}}, \mathbf{\Sigma}_{\mathcal{M}}(\mu_{\theta,\mathcal{L}}, \mathbf{\Sigma}_{\theta,\mathcal{L}}, t), \mu_d(t), \sigma_d(t))\right). \tag{3.21}$$

In principle one could adopt very different strategies to solve this minimization problem, f.e., stochastic techniques like simulated annealing or other approaches to global optimization. However, the costly evaluations of $\mathcal{F}_{\mathcal{L}}$ suggest the application of a Gauss-Newton-type minimization, more specifically a (damped) Gauss-Newton algorithm with statistical dimension reduction (c.f. [18]).

Let $\mathbf{F}$ be the matrix with

$$\begin{aligned}(F)_{ik} &= \text{overlap at time } t_i \text{ in the } k^{th} \text{ component} \\ &= \left\langle (\mathcal{D}_{t_i})_k, (\mathcal{M}_{t_i})_k \right\rangle \end{aligned} \tag{3.22}$$

which is linked to the linear overlap functional $\mathcal{F}_{\mathcal{L}}$ by the matrix norm $\|.\|$

$$\mathcal{F}_{\mathcal{L}} = \|\mathbf{F}\|.$$

Further, let z abbreviate

$$z = (\theta, \Delta\theta).$$

Then the general Gauss-Newton-algorithm solves a series of linearized problems

$$\|\mathbf{J}_F(z)\Delta z - \mathbf{F}(z)\|_2 = \min_{\Delta z} \tag{3.23}$$

with Δz being an update for $z_{\text{new}} = z_{\text{old}} + \Delta z$ and $\mathbf{J}_F$ being the (componentwise) Jacobian of $\mathbf{F}$, defined in (3.22), wrt. z (c.f. [15]).

In comparison to the standard application of (damped) Gauss-Newton strategies to residuum minimization ([14]), one faces a new challenge: Due to the dependency of the overlap on the variances $\mathbf{\Sigma}_{\mathcal{M}}(\theta_0, \Delta\theta, t)$ of the model, the parameter variances $\Delta\theta$ are to be optimized simultaneously. Consequently, (a) the dimension of the optimization problem is doubled in contrast to residuum optimization for the same model, (b) statistical correlations between parameters and the associated numerical problems will be more pronounced (since one has to expect correlations between a parameter θ_j and its variance $\Delta\theta_j$), and (c) one will see that the numerical effort for the evaluation of the Jacobian for each Gauss-Newton step increases quadratically.

Next, we have to discuss the numerical computations that are necessary to evaluate the Jacobian $\mathbf{J}_F$ of $\mathbf{F}$ as given by (3.23) with respect to θ and $\Delta\theta$, resulting in the composition of $\mathbf{J}_1$ representing the componentwise partial derivatives wrt. to θ and $\mathbf{J}_2$ the one wrt. $\Delta\theta$.

$$\begin{aligned}
\mathbf{J}_F &= (\mathbf{J}_1, \mathbf{J}_2) \\
\mathbf{J}_1 &= \left[\frac{\partial F_{ik}}{\partial \Phi_\theta^{t_i} y_0} \cdot \frac{\partial \Phi_\theta^{t_i} y_0}{\partial \theta_j} + \frac{\partial F_{ik}}{\partial \left(\Delta\theta(t_i)\right)_k} \cdot \frac{\partial \left(\Delta\theta(t_i)\right)_k}{\partial \theta_j}\right]_{(\theta_0, \Delta\theta)} \\
\mathbf{J}_2 &= \left[\frac{\partial F_{ik}}{\partial \left(\Delta\theta(t_i)\right)_k} \cdot \frac{\partial \left(\Delta\theta(t_i)\right)_k}{\partial \sigma_{\theta_j}}\right]_{(\theta_0, \Delta\theta)}
\end{aligned}$$

The computational effort for evaluation of the Jacobian will increase like P^2 (with $\theta \in \mathbb{R}^P$) for $\mathcal{F}_{\mathcal{L}}$ instead like P for the residuum $\mathcal{F}_{\mathcal{R}}$. The numerical evaluation of the derivatives involved is realized by numerical differentiation as it has been implemented within Presto Kinetics$^{\text{TM}}$.

Presto Kinetics[4] is a professional software tool used within research and development. This software package focusses on the modelling and dynamic simulation of arbitrary kinetic reactions ([42]). It provides general reaction step patterns for reaction kinetics and biokinetics as well as possibilities for the input of arbitrary ODE-systems. Its philosophy has also been proved to be very applicable within research context (c.f. [9, 12, 24, 25]). It contains a quite general Gauss-Newton framework for parameter estimation for dynamical systems with damping strategy, convergence monitor, and update strategy as given in [15, 22]. This framework has been extended to implement and test the stochastically damped Gauss-Newton approach to overlap optimization as presented in the following. Details on this implementation will be published in a forthcoming publications (c.f. [41]).

In order to determine the initial values μ_θ and $\mathbf{\Sigma}_\theta$ of the parameters and their variances for the Gauss-Newton iteration, one may, f.e., use box search. In the following it is assumed that there is a unique (local) maximum of $\mathcal{F}_{\mathcal{L}}$ in the vicinity of these initial values. With this preparations, the stochastically damped Gauss-Newton-algorithm consists of the following steps:

i) Initially set $l = 0$.

ii) Compute $\mathcal{F}_{\mathcal{L}}(\theta_0^{(l)}, \mathbf{\Sigma}(\theta^{(l)}, \Delta\theta^{(l)}, t), \mu_d(t), \sigma_d(t))$ by means of $\mathbf{F}$. Compute a set of realizations of the Jacobian $\mathbf{J}_F$ at $(\theta_0^{(l)}, \Delta\theta^{(l)})$ by numerical differentiation.

iii) Conduct dimension reduction by means of a truncated singular value decomposition (c.f. [18]).

iv) Compute the increment Δz to z by solving

$$\mathbf{J}_F(\theta_0^{(l)}, \mathbf{\Sigma}(\theta_0^{(l)}, \Delta\theta^{(l)}, t))\Delta z^T = \mathcal{F}_{\mathcal{L}}(\theta_0^{(l)}, \mathbf{\Sigma}(\theta_0^{(l)}, \Delta\theta^{(l)}, t), \mu_d(t), \sigma_d(t))$$

in the sense of (3.23) thus incorporating the dimension reduction.

[4] Presto Kinetics$^{\text{TM}}$ is a registered trademark by Dr. Michael Wulkow CiT GmbH Rastede

v) Set

$$(\theta_0^{(l+1)}, \Delta\theta^{(l+1)}) = (\theta_0^{(l)}, \Delta\theta^{(l)}) + \kappa\,\Delta z$$

with damping parameter κ. Verify monotony as reported in [22].

vi) Test convergence by means of the stopping criteria given in [14]. If not converged, set $l = l + 1$ and iterate from ii) onwards.

4 Numerical experiments

The overlap concept will now be applied to discriminating biokinetic models. Biokinetics describe chemical reactions performed by and between microorganisms like bacteria (c.f. [7]). To this end one has to conduct a model ranking by means of numerically prepared data. The data were generated by one model representing a dynamical system with distributed model parameters.

A simple example for a dynamical system modelling such biological processes is the following example

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = \mu(X(t), S(t)) \cdot X(t) - k_{\mathrm{d}} \cdot X(t) \tag{4.24}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = -\frac{\mu(X(t), S(t))}{y_{\mathrm{xs}}} \cdot X(t) - m_{\mathrm{s}} \cdot X(t). \tag{4.25}$$

S denotes the substrate that is transformed into biomass X, where $S(t)$ and $X(t)$ denote the associated concentrations. The parameter y_{xs} represents the ratio of mass of cells formed to mass of substrate consumed, k_{d} the deterioration rate of the biomass and m_{s} the one of the substrate that is consumed by the biomass. The rate of the transformation process is described by the kinetic $\mu(X, S)$. The following three types of kinetics

$$\mu(X(t), S(t)) = \mu_{\max} \cdot S(t)^r \tag{4.26}$$

$$\mu(X(t), S(t)) = \mu_{\max} \cdot \frac{S(t)}{k_{\mathrm{s}} + S(t)} \tag{4.27}$$

$$\mu(X(t), S(t)) = \mu_{\max} \cdot \frac{S(t)}{k_{\mathrm{s}} \cdot X(t) + S(t)} \tag{4.28}$$

are to be discriminated. The parameters $\mu_{\max}$ denote the maximal growth rate, k_s the half-saturation concentration. The kinetics given in (4.27) and (4.28) are known as the Monod and Contois kinetics, respectively (c.f. [7]). The third candidate in (4.26), the action mass kinetics, does show a significantly different behavior, f.e., asymptotically one observes $S \to \infty$. It therefore can be considered as a certain case of a chemically inappropriate model.

Now the overlap model ranking for the three models (4.27), (4.28) and (4.26) is illustrated. As in section 2.4, artificial data, generated on the basis of the Monod kinetics, was used again. The overlap is calculated at 14 time points: $t_1 = 0.1, \ldots, t_{14} = 1.4$. In order to generate the data, 14 sets of values of model parameters θ were randomly drawn according to a given normal parameter density π_θ, each for every time point. This is also shown in the last two lines of table 4.4 under ”original”. Next, the model for the Monod kinetics is evaluated wrt. the time points and sets of parameter values. Then the artificial data points were set to be $d(t_1) = \Phi^{t_1}_{\theta(1)} y_0$, ..., $d(t_{14}) = \Phi^{t_{14}}_{\theta(14)} y_0$. Further, the corresponding data variances σ_d^2, which were not provided by the model, were set to be $\sigma_d^2(t_i) = 0.15$ for all $i = 1$, ..., 14.

The following computations have been performed by Presto Kinetics, as mentioned in section 3. One will see, that even though the Jacobians $\mathbf{J}_F$ in the Gauss-Newton-steps (3.23), are relatively small for our investigated models, namely of dimension 10×28 corresponding to the number of parameters (including their variances) times the numbers of measured data for each dimension, the main numerical problem are the mentioned correlations in the matrix. This is going to be investigated at the end of the this section.

For the generated data, the linear overlap $\mathcal{F}_\mathcal{L}$ was optimized and the parameters $\mu_{\max}$, k_d, y_xs, m_s, k_s and r as well as their variances were estimated. The results of the PE for the linear overlap optimization are shown in table 4.4. Additionally, a classical residual based PE with respect to $\mathcal{F}_\mathcal{R}$ was conducted.

model	entity	$\mu_{\max}$	k_s or r	y_xs	m_s	k_d
(4.26)	expectation $\mu_{\theta,\mathcal{L}}$	0.331	0.999	0.573	0.254	$1.213 \cdot 10^{-2}$
	variance $\mathbf{\Sigma}_{\theta,\mathcal{L}}$	0.076	0.150	0.094	0.648	$1.132 \cdot 10^{-5}$
	parameter $\theta_\mathcal{R}$	0.633	0.397	0.564	1.345	$2.179 \cdot 10^{-4}$
	standard error $\overline{\sigma}_R$	0.181	0.239	0.222	0.393	$1.014 \cdot 10^{-3}$
(4.27)	expectation $\mu_{\theta,\mathcal{L}}$	1.386	0.737	0.573	0.091	$6.102 \cdot 10^{-2}$
	variance $\mathbf{\Sigma}_{\theta,\mathcal{L}}$	0.203	0.084	0.101	0.048	$3.030 \cdot 10^{-3}$
	parameter $\theta_\mathcal{R}$	1.262	0.793	0.613	0.377	$9.208 \cdot 10^{-6}$
	standard error $\overline{\sigma}_R$	0.262	0.639	0.125	0.278	$1.236 \cdot 10^{-4}$
(4.28)	expectation $\mu_{\theta,\mathcal{L}}$	2.028	0.981	0.525	0.059	$9.872 \cdot 10^{-2}$
	variance $\mathbf{\Sigma}_{\theta,\mathcal{L}}$	0.421	0.001	0.222	0.003	$1.145 \cdot 10^{-2}$
	parameter $\theta_\mathcal{R}$	1.560	0.902	0.819	0.791	$1.106 \cdot 10^{-4}$
	standard error $\overline{\sigma}_R$	0.394	0.779	0.155	0.192	$1.334 \cdot 10^{-3}$
original	expectation $\mu_{\theta,\mathcal{L}}$	1.400	0.730	0.600	0.090	$6.000 \cdot 10^{-2}$
	variance $\mathbf{\Sigma}_\theta$	0.100	0.080	0.100	0.010	$1.000 \cdot 10^{-2}$

Table 4.4: Linear overlap parameter estimation wrt. $\mathcal{F}_\mathcal{L}$ and classical parameter estimation wrt. $\mathcal{F}_\mathcal{R}$ for biokinetics (4.26), (4.27) and (4.28).

As expected, the parameters $\theta_\mathcal{L}$ including their variances $\mathbf{\Sigma}_{\theta,\mathcal{L}}$, were best estimated for the Monod kinetics, the ones for Contois stayed reasonable close, whereas the one for mass action kinetics differ in magnitude compared to the parameters used to generate the data. Figure (4.12) shows the trajectories calculated for (4.27) corresponding to the expectation determined by π_θ.

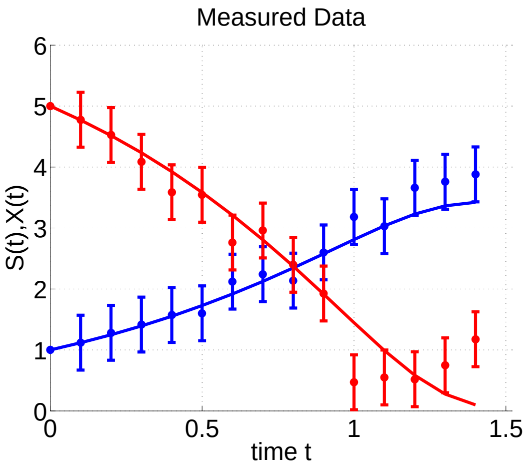


Figure 4.12: Monod kinetics with the prepared data. The substrate concentration $S(t)$ shown in the decreasing, the biomass concentration $X(t)$ in the increasing trajectory. The error bars denote the 95% confidence interval of the data.

The estimation methods by $\mathcal{F}_{\mathcal{L}}$ on the one and by $\mathcal{F}_{\mathcal{R}}$ on the other also allow for different interpretations in the parameters. For the Monod kinetics, the parameter k_{d} shows an influence in the overlap setting, but was estimated lower by the residuum by the factor 1000.

Returning to the numerical problems, one should recall that the Gauss-Newton-algorithm described in section 3 converges locally. Hence, one does need a qualified guess for the starting points for the algorithm. The problem surfaces in the context of overlap estimation even more, since the target functional for our applied kinetics is nonlinear and the $\theta^{(l)}$ and $\Delta\theta^{(l)}$ interact with each other. The numerical simulations show that these correlations affect the numerical condition of the Gauss-Newton-algorithm as well as the problem of identification dominant model parameters. The authors call a model parameter dominant, if and only if whose values dominate the dynamics of the model. In order to find uncorrelated and dominant model parameters, it shall be mentioned that no apriori knowledge about parameter correlations wrt. the nonlinear model under investigation is available. Moreover, one expects the parameter correlations to be dependent on the choice of π_θ. Hence, it is reasonable to identify a set of dominant and pairwise statistically model parameters by analyzing the dynamics of the model as a preliminary step in model ranking. This preliminary step will be the subject of further investigations.

Since the parameters gained by the overlap estimation are assumed to be of the same order of magnitude, a qualified guess proved to be the estimated parameters $\theta_{\mathcal{R}}$ by a least-square estimation as described in A.31. The starting values for the parameter sensibility $\mathbf{\Sigma}_\theta$ could also be chosen to be the estimated parameter standard deviation $\overline{\sigma}_R$. This choice provides an acceptable convergence of the overlap optimization. Choosing parameters that are far away from this choice, frequently result in trapping at a local minima.

In table 4.5, the targets functionals for the parameters estimated above are documented.

model		$\mathcal{F}_\mathcal{L}$	$\mathcal{F}_\mathcal{O}$		residuum
(4.26)	in X	68.97 %	62.64 %	$\mathcal{F}_\mathcal{R}$	0.144
	in S	58.15 %	58.68 %	avg.traj.	4.578
	total	63.56 %	60.66 %		
(4.27)	in X	77.99 %	68.15 %	$\mathcal{F}_\mathcal{R}$	0.159
	in S	66.96 %	55.87 %	avg.traj.	0.275
	total	72.47 %	62.02 %		
(4.28)	in X	31.71 %	19.76 %	$\mathcal{F}_\mathcal{R}$	0.144
	in S	38.45 %	51.52 %	avg.traj.	1.329
	total	35.09 %	35.64 %		

Table 4.5: Linear overlap parameter estimation wrt. $\mathcal{F}_\mathcal{L}$ and classical parameter estimation wrt. $\mathcal{F}_\mathcal{R}$ for the three kinetics. The values for $\mathcal{F}_\mathcal{O}$ were calculated for the optimized parameters by $\mathcal{F}_\mathcal{L}$ of table 4.4 to document the deviation between linear and nonlinear propagation. The residuum in the avg.traj. (left row) is the squared distance between the mean value of the data distribution and the average trajectory of (2.5) in the overlap setting.

Just knowing the values of table 4.5, the virtual experimenter has to discriminate between the three model candidates. By merely looking at the classical residuum $\mathcal{F}_\mathcal{R}$, no model can be favored. However, according to the overlap information $\mathcal{F}_\mathcal{L}$, the Contois kinetic of (4.28) ought to be rejected. For the remaining two candidates, the squared distance between the mean values of the data distribution and the average trajectory of (2.5), abbreviated table 4.5 by avg.traj., favors the Monod kinetics (4.27). The average trajectory is closer to the data for Monod than for the other one. Combining criteria from the residual on the one and the overlap framework on the other, one is able to discriminate the given model candidates.

Table 4.5 also documents again the differences between a linear $\mathcal{F}_\mathcal{L}$ and "exact" $\mathcal{F}_\mathcal{O}$ overlap calculation. How to improve the nonlinear propagation is subject of further papers to come (c.f. [23]).

A Some approaches to model discrimination

In this section, a brief review on some of the most influential ideas in model discrimination is given and compared to the overlap-approach.

Bayesian approach to model discrimination. There also is a Bayesian approach to model discrimination that has been developed in order to incorporate knowledge gained from other sources, namely apriori information about the parameters or the models (c.f. [30]). This is typically done by Bayesian parameter estimation [19] or by Bayesian factors [11, 27].

The very idea of Bayesian parameter estimation is to calculate the aposteriori distribution for the parameters θ using the Bayes formula:

$$P[\theta \,|\, d] = \frac{P[d \,|\, \theta]}{\int\limits_{\Theta} P[d \,|\, \theta] P[\theta] \, \mathrm{d}\theta} \cdot P[\theta] = \frac{P[d \,|\, \theta]}{P[d]} \cdot P[\theta], \tag{A.29}$$

Here the first factor of (A.29) is the *standardized likelihood*, $P[\theta]$ is the prior distribution for the parameters and $P[d \,|\, \theta]$ the likelihood. The conditional aposteriori distribution of θ can be used in the model discrimation process, if $P[d \,|\, \theta]$ was estimated from the data [39].

A *Bayesian factor* $B(M_i, M_j)$ for two models M_i and M_j is defined as the ration of the posterior odds and the model priors resulting in the ratio of the model likelihoods for two models

$$B(M_i, M_j) = \frac{P[M_i \,|\, d] \;\, P[M_j]}{P[M_j \,|\, d] \;\, P[M_i]} = \frac{P[d \,|\, M_i]}{P[d \,|\, M_j]}, \tag{A.30}$$

where $P[M_i]$ is the model prior for M_i, $P[d \,|\, M_i]$ the model likelihood for model M_i. The marginal distribution of the data d under model M, the Bayesian factors, therefore, choose the very model for which the marginal likelihood of the data is maximum. Again, $P[d \,|\, M_i]$ has to be estimated from the data and $P[M_i \,|\, d]$ can be calculated using the Bayes formula.

Residuum and least squares estimation. In parameter estimation and model discrimination methods, the residual, measuring the distance between data and model, is used to describe the model–data–fit. For non-constant data variance $\sigma_d^2(t_i)$, most commonly the weighted residual is considered (c.f. [8])

$$\mathcal{F}_{\mathcal{R}}(\theta) \;=\; \sum_{k=1}^{D} \sum_{i=1}^{N} \left[\frac{d_k(t_i) - (\Phi_\theta^{t_i} y_0)_k}{\sigma_d(t_i)} \right]^2 . \tag{A.31}$$

The parameter estimation mostly evolves around *least squares estimation*: the parameters θ are chosen to minimize the residuum $\mathcal{F}_{\mathcal{R}}$. The least-squares technique can be best illustrated in the case of linear regression models (c.f. [33, 40]). The relationship between random variables is given in terms of a model, concatenating the dependent (endogenous) variable y, which is explained by the model, and the independent (exogenous, explanatory) variable θ, which explains or predicts the dependent variables through the model. A linear regression model is given by

$$y = \mathbf{X}\theta + \epsilon \qquad\qquad \epsilon \sim \mathcal{N}(0, \mathbf{\Sigma}_d) \tag{A.32}$$

where $\mathbf{X} \in \mathbb{R}^{D \times P}$ is the design matrix and ϵ the random perturbation, that models the measurement uncertainty.

Assume that the variance of the data is known and given by the variance-covariance matrix $\mathbf{\Sigma}_d$. Then it can be shown (c.f. [8]) that the weighted least squares estimation for parameters θ can be calculated analytically by

$$\hat{\theta} = \left(\mathbf{X}^T \mathbf{\Sigma}_d^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{\Sigma}_d^{-1} y. \tag{A.33}$$

The point-estimator $\hat{\theta}$ in (A.33) happens to be the one with the smallest variance of all unbiased linear estimators (BLUE-estimator), namely:

$$\mathrm{Var}(\hat{\theta}) = \left(\mathbf{X}^T \mathbf{\Sigma}_d^{-1} \mathbf{X}\right)^{-1} . \tag{A.34}$$

Equation (A.34) can be used to link the variance $\mathbf{\Sigma}_d$ of the data to the quality of the estimated parameters θ, which is expressed by the variance of the estimation of $\hat{\theta}$. Thus (A.34) characterizes the uncertainty of the exact value for the parameter to estimate. Often the minimization of (A.31) is used to select one of a pool of competing linear and nonlinear models.

Maximum likelihood estimation (MLE). Within the concept of maximum likelihood estimation, the *likelihood function* $\mathcal{F}_M$ is considered.

$$\mathcal{F}_M(\theta) = \mathbf{P}\left(\|d - \Phi_\theta^{t=\tau} y_0\|\right) , \tag{A.35}$$

Here $\tau \in \mathbb{R}^+ \setminus \{0\}$ is an arbitrary constant value and $\mathbf{P}$ a probability measure. For linear regression models f.e., it measures the probability of the distance between model and data. Parameter estimation in the context of MLE means: the parameters θ are chosen to maximize the likelihood function $\mathcal{F}_M(\theta)$.

If the model is linear with respect to the parameters θ and the data errors are assumed to be normally distributed with known covariance, then the maximum likelihood estimation coincides (c.f. [8]) with the weighted least squares estimation of (A.33).

Discussion. In comparison to the overlap-approach the Bayesian approaches to model selection abstains from a clear distinction between deterministic input variables and distributed model parameters as made in (1.1) and favors a complete stochastic concept of a model. Consequently a concept of a model which differs slightly from (1.1) holds. Furthermore, due to the unavoidable estimation of distributions in the Bayesian approaches a model of the form (1.1) obviously can be translated into a sequence of stochastically dependent Bayesian models in time. Due to this fact, the procedure of model selection in the non-stationary case expressed in Bayesian terms will cause additional problems for model selection. Instead the authors have demonstrated that the overlap-approach to discriminate competing models of the form (1.1) is much less complex. Hence, no Bayesian approaches to model selection was discussed in this paper.

According to [13, 16, 21], there are three main sources of model uncertainty:

(U1) Uncertainty about the structure of the model;

(U2) Uncertainty about estimates of model parameters, assuming that one knows the structure of the model;

(U3) Unexplained random variation in the observed variables even when one knows the structure of the model and the values of the model parameters.

According to [13], the notion of model uncertainty based on nescience in model structure of (U1) can be broken down further, namely

(S1) Model misspecification (e.g. omitting a variable by mistake),

(S2) Specifying a general class of models of which the best model is a special, but unknown case or

(S3) Being confronted with two or more models of quite different structures.

Linear regression methods (c.f. [34]) and even more local polynomial regression methods (c.f. [17]) are popular statistical methods for dealing with the aspects (U2), (U3) as well as (S2). However, for cases like (U1), (S1) or (S3), the existing concepts do not allow strong inferences as in the previously mentioned cases. This circumstance is illustrated in the following for regression models considering a well known model selection criterion which deals which different models in the line of (D5) and (D6).

A basic concept for model selection is to employ some very general goodness–of–fit criteria combined with a penalty for model complexity:

$$\mathcal{F}_{\text{trade off}} = -2l(\theta) + \gamma P, \tag{A.36}$$

where P the number of parameters and

$$l(\theta) = \log P[d|\theta] \tag{A.37}$$

is the log–likelihood, which is maximized for the maximum-likelihood estimation of the parameters. For $\gamma = 2$ this is the Akaike-Information-Criterion (AIC) (c.f. [1, 2, 3, 4])

$$\mathcal{F}_{\text{AIC}} = -2l(\theta) + 2P. \tag{A.38}$$

The AIC is a large sample approximation of the discrepancy between the assumed true model and the fitted model in terms of the Kullback–Leibler distance (c.f. [28]). Model selection in terms of AIC means to choose the very model among the candidates that is closest in the Kullback–Leibler sense. However, the AIC tends to accept the most complex model (c.f. [32]) and is not always asymptotically constistent. [5]

The above type of model discrimination, along with the majority of textbook model discrimination approaches, consequently starts with a parameter estimation by means of least-square- or maximum likelihood approaches for each model. Additionally, it is assumed that the noise is of smaller magnitude than the effect. In a second step, the quality of the goodness-of-fit is evaluated and optimized if necessary. In this step one has to be aware of over- and underfitting: When a model has too many degrees of freedom the model fit will include parts of the noise as well as the structure contained in the distributed data.

[5] Mallow's C_p-criterion for selection among models with different numbers of parameters is a special case of AIC.

Overfitting typically leads to high goodness-of-fit. On the other hand when the model is not complex enough, it cannot capture the structure in our data, no matter how much data are available. Underfitting typically leads to bad predictions due to biased models [20].

Generally speaking, no master approach to the textbook model discrimination, parameter selection or model validation exists. Instead one hitherto needs a mixture of tailored solutions, which depends on the stage of modelling, on the type of model uncertainty one expects and on the strategy one chooses to cope with it. Hence a unified indicator for localized model ranking, which is able to deal with small amounts of data is of notably interest.

References

[1] H. Akaike. Statistical predictor identication. *Ann. Inst. Statist. Math.*, 21:203–217, 1970.

[2] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaksi, editors, *2nd International Symposium on Information Theory*, pages 267–281, 1973.

[3] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[4] H. Akaike. Factor analysis and aic. *Psychometrika*, 52:317–332, 1987.

[5] I. C. Araújo and M. P. de Oliveira. Volume and variance in the linear statistical model. *Linear algebra and its applications*, 357:303–306, 2002.

[6] S. F. Arnold. *The theory of linear models and multivariante analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1981.

[7] J. E. Bailey and D. F. Ollis. *Biochemical engineering Fundamentals.* McGrawHill, Inc., second edition, 1986.

[8] Y. Bard. *Nonlinear parameter estimation.* Academic Press, Inc., 1974.

[9] C. Barner-Kowollik, J. F. Quinn, D. R. Morsley, and T. P. Davis. Modeling the reversible addition-fragmentation chain transfer process in cumyl dithiobenzoate-mediated styrene homoppolymerizations: Assessing rate coefficients for addition-fragmentation equilibrium. *Journal of Polymer Science*, 39:1353–1365, 2001.

[10] P. Billingsley. *Probability and Measure.* Wiley Series in probability and mathematical statistics. John Wiley & Sons, Inc., 1979.

[11] G. E. P. Box and W. J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, February 1967.

[12] M. Busch. Modeling kinetics and structural properties in high-pressure fluid-phase polymerization. *Macromolecular theory simulation*, 10(5):408–429, 2001.

[13] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, 158(3):419–466, 1995.

[14] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Springer Series Computational Mathematics*. Springer, 2004.

[15] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equiations - Texts in Applied Mathematics*. Springer, 2002.

[16] D. Draper, J. S. Hodges, E. E. Morris, C. N. Morris, and D. B. Rubin. A research agenda for assessment and propagation of model uncertainty. Technical Report N-2683-RC, Rand Cooperation, Santa Monica, 1987.

[17] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, 1996.

[18] R. Fierro, G. Golub, P. Hansen, and D. O'Leary. Regularization by truncated total least squares. *SIAM Journal on Scientific Computing*, 18(4):1223–1241, July 1997.

[19] A. B. Gelman, J. S. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2000.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistcs. Springer, 2001.

[21] J. S. Hodges. Uncertainty policy analysis and statistics. *Statistical Science*, 2:259–291, 1987.

[22] A. Hohmann. *Inexact Newton Methods for Parameter Dependent Nonlinear Problems*. Reports on Mathematics, Shaker Verlag, Aachen, 1994.

[23] I. Horenko, S. Lorenz, C. Schütte, and W. Huisinga. Adaptive approach for nonlinear sensitivity analysis of reaction kinetics. *Journal of Computational Chemistry*, 26(9):941–948, July 2005.

[24] R. A. Hutchinson. Modeling of chain length and long–chain branching distributions in free–radical polymerization. *Macromolecular Theory Simulation*, 10(3):144–157, 2001.

[25] P. D. Iedema, C. Willems, G. van Vliet, W. Bunge, S. M. P. Mutsers, and H. C. J. Hoefsloot. Using molecular weight distributions to determine the kinetics of peroxide-induced degradation of polypropylene. *Chemical Engineering Science*, 56:3659–3669, 2001.

[26] A. M. Jacobs and J. Grainiger. Models of visual word recognition – sampling the state of the art. *Journal of Experimental Psychology: Human perception and performance*, 29:1311–1334, 1994.

[27] R. E. Kaas and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995.

[28] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[29] A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise – Stochastic Aspects of Dynamics*, volume 97 of *Applied Mathematical Sciences*. Springer Verlag New York, second edition, 1994.

[30] W. Ledermann and E. Lloyd, editors. *Handbook of Applicable Mathematics*, volume VI – part B, chapter Bayesian Statistics. John Wiley & Sons, Inc., 1984.

[31] J. Myung and M. A. Pitt. Model comparison methods. In L. Brand and M. L. Johnson, editors, *Numericals computer methods. Part D*, volume 383 of *Methods in Enzymology*, pages 351–366. Elsevier Inc., 2004.

[32] A. E. Raftery. Choosing models for cross-classification. *American Sociological Review*, 51:145–146, 1986.

[33] C. R. Rao and H. Toutenburg. *Linear Models – Least Squares and Alternatives*. Springer Series in Statistics. Springer, 1995.

[34] A. Rencher. *Linear Models in Statistics*. Wiley and Sons, 2001.

[35] A. Saltelli, K. Chan, and M. Scott, editors. *Special Issue on sensitivity analysis*, volume 117. Computer Physics Communications, 1999.

[36] A. Saltelli, K. Chan, and M. Scott, editors. *Sensitivity analysis*. Probability and Statistics series. John Wiley & Sons, 2000.

[37] A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 15(4):377–395, 2000.

[38] A. Saltelli, S. Tarantola, and K. Chan. A quantitative, model independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, 1999.

[39] D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, London, Sydney, 1992.

[40] J. E. Stapleton. *Linear statistical models*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 1995.

[41] R. Telgmann. *Numerical aspects of the model–overlap–concept*. PhD thesis, Freie Universität Berlin, 2006. in progress.

[42] M. Wulkow and R. Telgmann. *Presto Kinetics – Simulation of kinetic models (Manual)*. Dr. Michael Wulkow CiT GmbH, 2003.