Published in IET Systems Biology
Received on 30th April 2009
Revised on 1st October 2009
doi: 10.1049/iet-syb.2009.0030

ISSN 1751-8849

Parameter identification, experimental design and model falsification for biological network models using semidefinite programming

J. Hasenauer S. Waldherr K. Wagner F. Allgöwer

Institute for Systems Theory and Automatic Control, Universität Stuttgart, Germany
E-mail: hasenauer@ist.uni-stuttgart.de

Abstract: One of the most challenging tasks in systems biology is parameter identification from experimental data. In particular, if the available data are noisy, the resulting parameter uncertainty can be huge and should be quantified. In this work, a set-based approach for parameter identification in discrete time models of biochemical reaction networks from time series data is developed. The basic idea is to determine an outer approximation to the set of parameters for which trajectories are consistent with the available data. In order to approximate the set of consistent parameters (SCP) a feasibility problem is derived. This feasibility problem is used to verify that complete parameter sets cannot contain consistent parameters. This method is very appealing because instead of checking a finite number of distinct points, complete sets are analysed. With this approach, model falsification simply corresponds to showing that the SCP is empty. Besides parameter identification, a novel set-based method for experimental design is presented. This method yields reliable predictions on the information content of future measurements also for the case of very limited a priori knowledge and uncertain inputs. The properties of the method are presented using a discrete time model of the MAP kinase cascade.

1 Introduction

Experimental design, parameter identification and model falsification are important tasks one has to deal with when constructing models of biological systems. Unfortunately, there are several open problems. In parameter identification and model falsification, sparse and noisy data sets as well as non-convexity of the underlying optimisation problem are challenging. For experimental design, classical approaches require a detailed a priori knowledge about the parameter values, which is typically not available at the beginning of the modelling process, where experimental design can have the largest impact.

In this paper, a set-based approach is presented to overcome the problems related to noisy data and non-convexity for a class of implicit non-linear discrete time systems with bounded measurement error. The method is based on the outer approximation of the set of consistent parameters (SCP), the set of parameters consistent with all available experimental data. Additionally, the application of the proposed approach to the analysis of biochemical reaction networks is illustrated with a case study.

Classical parameter identification approaches are based on the definition of the objective function and a successive modification of the parameter vector to minimise the objective and thus the difference between system and model response [1]. For the modification of the parameters, gradient-based methods are commonly applied. Hence, these standard approaches check a finite number of distinct points in parameter space. Even in cases where global optimisation methods (e.g. clustering-methods, simulated annealing or differential evolution) are employed [2–4] it can usually not be guaranteed that the optimal parameter vector is obtained with a finite number of iterations.

In particular for the task of model falsification, these classical approaches are deficient as they check only a finite number of points in parameter space. Even if exhaustive

Monte-Carlo simulations [5] are employed, which use a random sampling in the parameter space, only falsification probabilities are obtained. In cases where no parameter values are found for which the model reproduces the experimental data, it cannot be guaranteed that no such parameter values exist.

Furthermore, especially if the information content of the measured data is small, the remaining parameter uncertainties may be large and need to be quantified in order to evaluate the quality of the obtained model. This can be done via a practical identifiability analysis and the computation of the confidence intervals. Parameter confidence intervals are traditionally computed using the Fisher information matrix [6] or bootstrapping methods [7]. The Fisher information matrix is computed from the local sensitivity of the output with respect to the parameters. Hence confidence intervals computed using the Fisher information matrix are only valid locally and moreover rely on the assumption that the correct parameter is known, what is clearly not the case. Bootstrapping methods, on the other hand, are non-deterministic methods and use stochastic elements as well as repeated simulations and repeated solving of the parameter estimation problem. By this it is possible to account for non-linearity of the identification problem. However, they require detailed knowledge about measurement noise distributions and noise properties (e.g. ergodicity). Hence, their application is in some situations questionable as the prerequisites on the noise are difficult to verify [8].

Another method to perform practical identifiability analysis is based on the calculation of the SCP [8, 9]. The main advantages are that the SCP can be used to derive rigorous bounds on the parameter uncertainties. Secondly, only boundedness of the noise has to be assumed. Furthermore, in case the SCP is provided as a reduced search area for conventional optimisation-based methods, a tremendous speed-up to the parameter estimation is possible.

To compute the SCP, set-based methods have been developed during the last decade [10]. In particular, Kieffer and Walter [8] developed methodologies employing set inversion, interval analysis and constraint propagation. These methods work well if the system has a particular structure, for example if it is cooperative [11, 12], or if the dependency of the output on the parameters is known explicitly. However, if the mapping is not known, the results can get very conservative because of the strong relaxation required by interval arithmetics. One method to obtain good outer bounds for the SCP of systems of ordinary differential equation has been developed by Tucker *et al.* [13], which also uses constraint propagation. Unfortunately, one basic assumption is that the derivatives of the concentrations can be determined. For common measurements techniques used in molecular biology, the noise level typically prohibits the direct determination of measurement derivatives.

In the work of Küpfer *et al.* [9] a novel formulation of the computation of the SCP as feasibility problem is proposed and applied to compute infeasibility certificates for complete regions in parameter space. The feasibility problem is brought to a computationally efficient form by a relaxation to a semidefinite program (SDP) [14]. Related approaches have also been applied to perform global sensitivity and uncertainty analysis for (bio-)chemical reaction networks [15, 16].

The drawback of Küpfer's approach is that only information about the steady state of the system can be employed for the parameter identification and that only polynomial vector fields are considered. In this paper, we extend the application of the earlier proposed methods to perform parameter estimation for discrete time systems with rational right-hand side for which measurements of the time courses are available. The extension to time course data is crucial and in parallel to this work a first method for parameter identification and model discrimination has been proposed by Borchers *et al.* [17], considering polynomial vector fields. Additionally to the extensions on the theoretical side also a more extensive algorithm than in [9] is applied to approximate the SCP. The proposed parameter identification method is directly applicable to the model falsification problem by trying to establish emptiness of the SCP.

Besides the improvement of the parameter identification method, a first set-based experimental design method is presented in this paper. The main goal of experimental design is to select the most informative experiments for parameter identification based on a priori knowledge [6, 18]. In this work, we focus on removing the constraint that a good estimate of the real parameters has to be available for selecting the experiments. This can lead to wrong predictions of the information content. Instead, it is assumed that only an a priori consistent parameter set is known.

This problem has also been considered by Asprey and Macchietto [19] who developed a robust experimental design method based on the Fisher information matrix. In contrast to their work, the experimental design approach proposed here uses as design criterion the expected volume of the SCP, a non-local measure. Furthermore, input uncertainties are taken into account as they are common in biological applications.

The remainder of the paper is structured as follows: in Section 2, the problems of parameter identification and model falsification are formulated and the theoretical background as well as the applied algorithms are presented. Section 3 contains an explanation of the experimental design approach and the method for selecting the most informative measurements. In Section 4, the developed algorithms are illustrated by an application to the experimental design, SCP estimation and model falsification for a simple model of the MAP kinase cascade.

doi: 10.1049/iet-syb.2009.0030

Mathematical notation. The space of real symmetric $n \times n$ matrices is denoted as $\mathcal{S}^n$. $\mathcal{I}(a, b)$ denotes the integer set $\{a, a+1, \ldots, b\}$. The positive semidefiniteness of a quadratic matrix $\boldsymbol{X} \in \mathcal{S}^n$ is denoted $\boldsymbol{X} \succeq 0$ and the trace of $\boldsymbol{X}$ by $\mathrm{tr}(\boldsymbol{X})$.

2 Parameter identification and model falsification

2.1 Problem statement

Consider an implicit non-linear discrete time dynamical model for a biochemical reaction network given by the system of implicit difference equations

$$\Sigma: \begin{cases} 0 = F(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{p}), & \boldsymbol{x}^{(0)} = \boldsymbol{x}_0 \\ 0 = H(\boldsymbol{y}^{(k)}, \boldsymbol{x}^{(k)}, \boldsymbol{p}) \end{cases} \tag{1}$$

where $\boldsymbol{y}^{(k)} \in \mathbb{R}^{n_y}$ is the output vector, $\boldsymbol{x}^{(k)} \in \mathbb{R}^{n_x}$ the state vector, $\boldsymbol{u}^{(k)} \in \mathbb{R}^{n_u}$ the input vector at the kth time point and $\boldsymbol{p} \in \mathbb{R}^{n_p}$ a constant parameter vector to be estimated. In this paper, only rational functions F and H are considered, but extensions to piecewise polynomial or general smooth non-linear functions are possible [15]. Many modelling frameworks for biochemical reaction networks rely exclusively on polynomial or rational functions, which stem from the law of mass action, the Michaelis–Menten mechanism or Hill-type reaction rates with integer Hill coefficients. However, the approach proposed in this paper is not applicable to generalised mass action networks [20], which are a less commonly used formalism to describe biochemical reaction networks.

Discrete time models of reaction networks arise from discrete time modelling [21] or via time discretisation of differential equation models [22]. The advantage of discrete compared to continuous time models is the strictly algebraic mapping from $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(k+1)}$.

We will assume that the input $\boldsymbol{u}^{(k)}$ is known to be contained in a compact set $\mathcal{U}^{(k)} \subset \mathbb{R}^{n_u}$. In addition, there are constraints on the state variables, given by $\boldsymbol{x}^{(k)} \in \mathcal{X}^{(k)} \subset \mathbb{R}^{n_x}$. Such constraints are often available from conservation laws or maximal production rates for individual chemical species.

The output of the system Σ is available through possibly erroneous measurements. The measurements are given by

$$\bar{\boldsymbol{y}}^{(k)} = \boldsymbol{y}^{(k)} + \boldsymbol{e}^{(k)}, \quad k \in \mathcal{I}(0, N) \tag{2}$$

in which $\bar{\boldsymbol{y}}^{(k)} \in \mathbb{R}^{n_y}$ is the measured output, $\boldsymbol{e}^{(k)} \in \mathbb{R}^{n_y}$ the unknown measurement error and $N+1$ the number of measurement points. We assume that the measurement error at each time point is known to be contained in a known compact set $\mathcal{E}^{(k)} \subset \mathbb{R}^{n_y}$. Then one can determine the set

$$\mathcal{Y}^{(k)} = \{\boldsymbol{y} \in \mathbb{R}^{n_y} \mid \exists \boldsymbol{e} \in \mathcal{E}^{(k)} : \bar{\boldsymbol{y}}^{(k)} = \boldsymbol{y} + \boldsymbol{e}\} \tag{3}$$

which by construction contains the actual output $\boldsymbol{y}^{(k)}$ of the system at each time point k.

Let us introduce the notation $\boldsymbol{y}^{(0,k)} = (\boldsymbol{y}^{(0)}, \ldots, \boldsymbol{y}^{(k)})$, $k > 0$, for an output sequence of the model Σ, and similarly $\boldsymbol{u}^{(0,k)}$, $\boldsymbol{x}^{(0,k)}$, $\boldsymbol{e}^{(0,k)}$ and $\bar{\boldsymbol{y}}^{(0,k)}$ for input, state, measurement error and measured output sequences, respectively. Moreover, we consider sets of sequences denoted by $\mathcal{U}^{(0,k)} = \{\boldsymbol{u}^{(0,k)} | \boldsymbol{u}^{(i)} \in \mathcal{U}^{(i)}, i = 0, \ldots, k\}$, $\mathcal{X}^{(0,k)} = \{\boldsymbol{x}^{(0,k)} | \boldsymbol{x}^{(i)} \in \mathcal{X}^{(i)}, i = 0, \ldots, k\}$, $\mathcal{E}^{(0,k)} = \{\boldsymbol{e}^{(0,k)} | \boldsymbol{e}^{(i)} \in \mathcal{E}^{(i)}, i = 0, \ldots, k\}$ and $\mathcal{Y}^{(0,k)} = \{\boldsymbol{y}^{(0,k)} | \boldsymbol{y}^{(i)} \in \mathcal{Y}^{(i)}, i = 0, \ldots, k\}$.

We call a parameter vector $\boldsymbol{p} \in \mathbb{R}^{n_p}$ consistent with $(\Sigma, \mathcal{U}^{(0,N-1)}, \mathcal{X}^{(0,N)}, \mathcal{Y}^{(0,N)})$, if there exist $\boldsymbol{u}^{(0,N-1)} \in \mathcal{U}^{(0,N-1)}$ and a solution $\boldsymbol{x}^{(0,N)} \in \mathcal{X}^{(0,N)}$ of Σ such that $\boldsymbol{y}^{(0,N)} \in \mathcal{Y}^{(0,N)}$.

The first problem that is considered in this paper is to compute the SCP.

Problem 1: Given the model Σ, the set of input sequences $\mathcal{U}^{(0,N-1)}$, the set of accessible states $\mathcal{X}^{(0,N)}$ and the set of output sequences $\mathcal{Y}^{(0,N)}$, compute the set $\mathcal{P}^* \subset \mathbb{R}^{n_p}$ of all parameters which are consistent with $(\Sigma, \mathcal{U}^{(0,N-1)}, \mathcal{X}^{(0,N)}, \mathcal{Y}^{(0,N)})$.

For models of typical biochemical networks, the set $\mathcal{P}^*$ usually cannot be determined explicitly. In this work, we focus on the computation of an outer approximation $\bar{\mathcal{P}}^* \supseteq \mathcal{P}^*$, which is guaranteed to contain all consistent parameters. In this way, upper bounds on the parameter uncertainty resulting from uncertain measurement data can be obtained.

The second problem under consideration is the task of model falsification. In the proposed framework, the model falsification problem is simply the problem of proving that the SCP is empty. If this is the case, the model structure Σ cannot reproduce the experimental data for any values of the parameters $\boldsymbol{p}$.

Problem 2: Given the model Σ, $\mathcal{U}^{(0,N-1)}$, $\mathcal{X}^{(0,N)}$ and $\mathcal{Y}^{(0,N)}$, determine whether the (SCP) $\mathcal{P}^*$ is empty or not.

2.2 Theoretical background

2.2.1 Infeasibility certificates: In this section a method to compute an outer approximation to the SCP is derived. For this purpose, we introduce the feasibility

problem

(P):

$$\begin{cases} \text{find} & \boldsymbol{y}^{(0,N)} \in \mathcal{Y}^{(0,N)}, \boldsymbol{x}^{(0,N)} \in \mathcal{X}^{(0,N)} \\ & \boldsymbol{u}^{(0,N-1)} \in \mathcal{U}^{(0,N-1)}, \quad \boldsymbol{p} \in \mathcal{P} \\ \text{s.t.} & F(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(0, N-1) \\ & H(\boldsymbol{y}^{(k)}, \boldsymbol{x}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(0, N) \end{cases}$$

This feasibility problem is in the following used for the classification of a parameter test set $\mathcal{P} \subset \mathbb{R}^{n_p}$. If (P) is infeasible, $\mathcal{P}$ does not contain consistent parameters. Unfortunately, (P) is a non-linear feasibility problem and in general non-convex and therefore NP-hard to solve.

Küpfer *et al.* [9] proposed a framework for relaxing a polynomial non-convex feasibility problem to a SDP [23], and apply it to parameter estimation for steady-state measurements. Owing to inherent convexity of SDPs, these problems can be solved computationally efficiently, e.g. via primal–dual interior point methods. In the following, we present an approach which is based on the work of Küpfer *et al.* [9], and extends this work to dynamical measurements.

For the relaxation of (P) to a SDP, the original feasibility problem is first rewritten as a quadratic feasibility problem. Therefore all equations and constraints appearing in (P) have to be polynomial. If all sets are convex polytopes and the functions F and H are polynomial in all of their arguments, this is fulfilled. In case that F and/or H are rational in their arguments, one can just multiply each equation with its least common denominator. In order to rewrite (P) as a quadratic problem, the vector $\boldsymbol{\xi} \in \mathbb{R}^{n_\xi}$ is introduced, which consists of the monomials appearing in F and H, that is

$$\boldsymbol{\xi} = (1,\ y_{i_y}^{(k)},\ x_{i_x}^{(k)},\ u_{i_u}^{(k)},\ p_{i_p},\ y_{i_y}^{(k)} x_{i_x}^{(k)},\ x_{i_x}^{(k)} p_{i_p},\ \ldots)^{\mathrm{T}} \tag{4}$$

for all $i_y \in \mathcal{I}(1, n_y)$, $i_x \in \mathcal{I}(1, n_x)$, $i_u \in \mathcal{I}(1, n_u)$, $i_p \in \mathcal{I}(1, n_p)$, $k \in \mathcal{I}(0, N)$ [14]. Using the monomial vector $\boldsymbol{\xi}$, the equality constraints

$$\begin{aligned} F(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(0, N-1) \\ H(\boldsymbol{y}^{(k)}, \boldsymbol{x}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(0, N) \end{aligned} \tag{5}$$

can be transformed to

$$\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{\xi} = 0, \quad i \in \mathcal{I}(1, n_x N + n_y(N+1)) \tag{6}$$

with $\boldsymbol{Q}_i \in \mathcal{S}^{n_\xi}$. Note that for higher order terms in $\boldsymbol{\xi}$, additional constraints have to be introduced. These lead to additional equations of the form

$$\boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{\xi} = 0, \quad i \in \mathcal{I}(n_x N + n_y(N+1) + 1, c) \tag{7}$$

where again $\boldsymbol{Q}_i \in \mathcal{S}^{n_\xi}$ and c is the total number of equality constraints.

In order to simplify notation, $\mathcal{Y}^{(k)}$, $\mathcal{X}^{(k)}$, $\mathcal{U}^{(k)}$ and $\mathcal{P}$ are restricted to polyhedral sets. Then the constraints $\boldsymbol{y}^{(0,N)} \in \mathcal{Y}^{(0,N)}$, $\boldsymbol{x}^{(0,N)} \in \mathcal{X}^{(0,N)}$, $\boldsymbol{u}^{(0,N-1)} \in \mathcal{U}^{(0,N-1)}$ and $\boldsymbol{p} \in \mathcal{P}$ can be written as

$$\boldsymbol{B}\boldsymbol{\xi} \geq 0 \tag{8}$$

with $\boldsymbol{B} \in \mathbb{R}^{n_b \times n_\xi}$, and n_b is the number of constraints that jointly describe the sets $\mathcal{Y}^{(0,N)}$, $\mathcal{X}^{(0,N)}$, $\mathcal{U}^{(0,N-1)}$ and $\mathcal{P}$.

The original feasibility problem (P) can then be restated as

$$(\mathrm{QP}): \begin{cases} \text{find} & \boldsymbol{\xi} \in \mathbb{R}^{n_\xi} \\ \text{subject to} & \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{\xi} = 0, \quad i \in \mathcal{I}(1, c) \\ & \boldsymbol{B}\boldsymbol{\xi} \geq 0 \\ & \xi_1 = 1 \end{cases}$$

A relaxation to a SDP is classically done by introducing the matrix $\boldsymbol{X} = \boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}$ and dropping the appearing non-convex constraint $\mathrm{rank}\,(\boldsymbol{X}) = 1$ [24]. This leads to the relaxed feasibility problem

$$(\mathrm{RP}): \begin{cases} \text{find} & \boldsymbol{X} \in \mathcal{S}^{n_\xi} \\ \text{subject to} & \mathrm{tr}(\boldsymbol{Q}_i \boldsymbol{X}) = 0, \quad i \in \mathcal{I}(1, c) \\ & \boldsymbol{B}\boldsymbol{X}\boldsymbol{e}_1 \geq 0 \\ & \boldsymbol{B}\boldsymbol{X}\boldsymbol{B}^{\mathrm{T}} \geq 0 \\ & \mathrm{tr}(\boldsymbol{e}_1 \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{X}) = 1 \\ & \boldsymbol{X} \succeq 0 \end{cases}$$

with $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^{n_\xi}$. Note that the relaxation may induce additional solutions. To reduce conservatism, the redundant constraint $\boldsymbol{B}\boldsymbol{X}\boldsymbol{B}^{\mathrm{T}} \geq 0$ is added, which is satisfied by every solution of (QP) and (P).

From (RP) one can derive the Lagrange dual problem

$$(\mathrm{DP}): \begin{cases} \text{maximise} & \nu_1 \\ \text{subject to} & \boldsymbol{e}_1 \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\lambda}_1 \boldsymbol{e}_1^{\mathrm{T}} + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{\lambda}_2 \boldsymbol{B} \\ & \quad + \boldsymbol{\lambda}_3 + \nu_1 \boldsymbol{e}_1 \boldsymbol{e}_1^{\mathrm{T}} + \sum_{i=1}^{c} \nu_{2,i} \boldsymbol{Q}_i = 0 \\ & \boldsymbol{\lambda}_1 \geq 0, \ \boldsymbol{\lambda}_2 \geq 0, \ \boldsymbol{\lambda}_3 \succeq 0 \end{cases}$$

with the Lagrange multipliers $\boldsymbol{\lambda}_1 \in \mathbb{R}^{n_b}$, $\boldsymbol{\lambda}_2 \in \mathcal{S}^{n_b}$, $\boldsymbol{\lambda}_3 \in \mathcal{S}^{n_\xi}$, $\nu_1 \in \mathbb{R}$ and $\boldsymbol{\nu}_2 \in \mathbb{R}^c$ [16]. Using the dual problem, one can obtain an infeasibility certificate for the original problem.

Proposition 1: Let ν_1^* be the optimal cost of (DP). If $\nu_1^* = \infty$, then the original feasibility problem (P) is infeasible.

This follows directly from weak duality. Note that any feasible solution to (DP) with $\nu_1^* > 0$ implies that (DP) is unbounded from above and Proposition 1 applies. We call such a solution an inconsistency certificate, because it gives

doi: 10.1049/iet-syb.2009.0030

a guarantee that $\mathcal{P}$ does not contain consistent parameters. For a more detailed discussion we refer to [16].

2.2.2 Reduction of computational complexity: The advantage of the formulation using the Lagrange dual is that the problem is convex and can be solved efficiently, as long as the number of optimisation variables of (DP) is trackable. However, the number of optimisation variables can be problematic already for small-scale systems, if a large number of measurement points needs to be considered. The reason is that the number of optimisation variables n_o is of order $\mathcal{O}^2((n_y + n_x + n_u + n_p)N)$ and thus grows quadratically in the number of uncertain variables and the number of time points. Furthermore, the dominating time for solving these problems is the cost for solving a linear program, which is of order $\mathcal{O}^3(n_o)$ [25]. Thus the effort for solving (DP) grows to the sixth order in the number of uncertain variables and time steps. This is of relevance in particular if the considered system or the number of measurement points are large.

To reduce the computational effort with respect to the number of time points, the original feasibility problem (P) can be split. This is done by splitting the sequences of input, state and output variables in subsequences, such that several feasibility problems, each considering only a subset of time points, are constructed.

The resulting feasibility problems are given by

$$(\mathrm{P_j}):\begin{cases} \text{find} & \boldsymbol{y}^{(j,j+m)} \in \mathcal{Y}^{(j,j+m)},\ \boldsymbol{x}^{(j,j+m)} \in \mathcal{X}^{(j,j+m)} \\ & \boldsymbol{u}^{(j,j+m-1)} \in \mathcal{U}^{(j,j+m-1)},\quad \boldsymbol{p} \in \mathcal{P} \\ \text{s.t.} & F(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(j, j+m-1) \\ & H(\boldsymbol{y}^{(k)}, \boldsymbol{x}^{(k)}, \boldsymbol{p}) = 0, \quad k \in \mathcal{I}(j, j+m) \end{cases}$$

where $m+1$, with $1 \leq m \leq N$, is the number of sequential measurement points taken into account in the optimisation problem (P_j). Since a solution of (P) satisfies the constraints where the complete sequences from 0 to N are taken into account, we observe that if (P) is feasible, then also (P_j), $j = 0, \ldots, N-m$, are feasible. Because the reverse is in general false, considering (P_j), $j = 0, \ldots, N-m$, instead of (P) corresponds to a relaxation. Each feasibility problem (P_j) can be relaxed to its dual problem (DP_j), as shown above for (P). Using the (DP_j) we obtain:

Proposition 2: Let $\nu^*_{j,1}$ be the optimal cost of (DP_j). If

$$\sup\{\nu^*_{j,1} \,|\, \forall j \in \mathcal{I}(0, N-m)\} = \infty$$

then the original feasibility problem (P) is infeasible.

This result follows again from weak duality.

Note that splitting the original problem along the time axis solves only part of the problem, the question how m should to chosen remains. As a rule of thumb, we suggest that with increasing measurement uncertainties and a decreasing number of measured variables, m should increase in order to maintain a comparable estimation quality.

Remark 1: As the number of optimisation variables and the computational complexity for solving (DP_j), strongly increases with the number of states and parameters of the system, it is so far not possible to consider large-scale systems. To change this, an in depth analysis of the structure of (P_j) has to be performed in the future to reduce the computation time.

2.3 SCP computation and model falsification

Using the Lagrange dual problems (DP_j), a certificate for the inconsistency of (P), for a given set of parameters $\mathcal{P}_i$, can be computed. This allows us to exploit (DP_j) to determine an outer approximation $\bar{\mathcal{P}}^*$ to the SCP $\mathcal{P}^*$. In this work, this is done via a multi-dimensional bisection algorithm.

Therefore, at first an initial set $\mathcal{P}_0$, with $\mathcal{P}^* \subseteq \mathcal{P}_0$, has to be determined. In many practical applications, finding a set $\mathcal{P}_0$, with $\mathcal{P}^* \subseteq \mathcal{P}_0$, is not a restriction. A suitable set $\mathcal{P}_0$ can often be determined from physical insight into the problem. For instance, in biochemical reaction networks, all parameters are in general positive, hence already lower bounds are found.

Starting from the initial set $\mathcal{P}_0$, a recursive bisection of $\mathcal{P}_0$ is performed. For each of the resulting subsets $\mathcal{P}_i$ arising in the bisection, the corresponding dual problems (DP_j) are analysed and it is tried to compute inconsistency certificates for $\mathcal{P}_i$. Successful computation of an infeasibility certificate assures that $\mathcal{P}_i$ does not intersect the SCP. If no certificate can be obtained, $\mathcal{P}_i$ is bisected, and it is tried to obtain an infeasibility certificate for the subsets. An approximation $\bar{\mathcal{P}}^*$ of the SCP is finally given by

$$\bar{\mathcal{P}}^* = \mathcal{P}_0 \setminus \bigcup_I \mathcal{P}_I \tag{9}$$

where $\mathcal{P}_I$ are the sets for which an inconsistency certificate could be obtained.

The implementation of the algorithm is outlined as follows:

`Algorithm:` $\mathcal{P} =$ `Analyze-`$\mathcal{P}(\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{P})$

1. `If` $V(\mathcal{P}) < \epsilon$, `return` $\mathcal{P} = \mathcal{P}$

2. `Check feasibility of` $DP_j(\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{P})$, $\forall j \in \mathcal{I}(0, N-m)$

3. `If` $\sup\{\nu^*_{1,j} \,|\, j \in \mathcal{I}(0, N-m)\} = \infty$, `return` $\mathcal{P} = \emptyset$

4. `If` $\sup\{\nu^*_{1,j} \,|\, j \in \mathcal{I}(0, N-m)\} \neq \infty$:

4.1. Bisection of $\mathcal{P}$ in $\mathcal{P}_1$ and $\mathcal{P}_2$

4.2. $\mathcal{P}_1 = \text{Analyze-}\mathcal{P}(\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{P}_1)$

4.3. $\mathcal{P}_2 = \text{Analyze-}\mathcal{P}(\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{P}_2)$

4.4. Return $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$

This algorithm is called recursively until the weighted volume

$$V(\mathcal{P}) = \int_{\mathcal{P}} w(\boldsymbol{p})\, \mathrm{d}\boldsymbol{p} \tag{10}$$

of a test set $\mathcal{P}$ is smaller than a tolerance ϵ. Here, $w(\boldsymbol{p}) \geq 0$ is a weighting function used to assess the importance of different regions in parameters space. For a more detailed discussion of this bisection algorithm we refer to [10]. The algorithm is implemented in Matlab. For solving the dual problems (DP_j) the open source toolbox SeDuMi is used [26].

For the task of model falsification also the above described algorithm is applied. If $\mathcal{P}_0$ can be certified to be inconsistent, that is the algorithm returns the empty set, we have a guarantee that no parameter value $\boldsymbol{p} \in \mathcal{P}_0$ exists for which the model can fit the experimental data, and the model Σ is falsified.

Remark 2: Applying the proposed method does not require an a priori identifiability analysis. If parameters p_j are not identifiable from the data, the uncertainty in these parameters will not decrease during SCP computation. Hence, identifiability can be studied in a rigorous way using the proposed algorithm.

3 Experimental design

3.1 Problem statement

Besides the evaluation and analysis of measured data, it is of interest to predict the information content I of future experiments to perform experimental design.

Generally, experimental design aims at determining an experimental setup which allows gaining a maximal amount of additional information with respect to parameter identification [6] or model falsification [27]. In this paper, we focus on the comparison of the expected information content of different sets of input sequences $\mathcal{U}_1^{(0,N-1)}, \ldots, \mathcal{U}_M^{(0,N-1)}$, in which $\mathcal{U}_i^{(0,N-1)}$ denotes a set of experimentally feasible sequences of stimuli.

Compared to traditional approaches, it is not assumed that a good approximation of the correct parameter is known, a set-based information criterion is used and the fact that inputs can not be forced precisely is taken into account. The last point is of particular relevance in biological applications and experiments.

As a measure for information content of an experiment, we consider the volumetric ratio of the falsified and the initial parameter set. The proposed method is hence related to D-optimal experimental design, where the determinate of the covariance matrix is minimised [19].

The experimental design problem can be stated as follows:

Problem 3: Given the model Σ, an a priori consistent parameter set $\mathcal{P}_0$ and M sets of feasible input sequences $\mathcal{U}_i^{(0,N-1)}$, $i \in \mathcal{I}(1, M)$, determine the set of input sequences $\mathcal{U}_{i^*}^{(0,N-1)}$ for which the expected information content $[I](\mathcal{U}_{i^*}^{(0,N-1)}, \mathcal{P}_0)$ is maximal.

3.2 Theoretical background

In order to select the most informative experiments, the expected information content $[I](\mathcal{U}_i^{(0,N-1)}, \mathcal{P}_0)$ for a given set of input sequences $\mathcal{U}_i^{(0,N-1)}$ and an a priori consistent parameter set $\mathcal{P}_0$ has to be defined. Therefore at first the information content of a particular experiment $(u^{(0,N-1)}, \bar{y}^{(0,N)})$ is defined as,

$$I(\boldsymbol{p}, \boldsymbol{u}^{(0,N-1)}, \boldsymbol{e}^{(0,N)}) = \frac{V(\mathcal{P}_0 \setminus \mathcal{P}^*(\boldsymbol{u}^{(0,N-1)}, \bar{\boldsymbol{y}}^{(0,N)}))}{V(\mathcal{P}_0)} \tag{11}$$

where $V(\mathcal{P})$ is the weighted volume of a set as defined in (10). This information content depends on the input $\boldsymbol{u}^{(0,N-1)}$ and, via the measured output $\bar{\boldsymbol{y}}^{(0,N)}$, on the system parameter $\boldsymbol{p}$ and the measurement noise $\boldsymbol{e}^{(0,N)}$. The expected information content $[I](\boldsymbol{p}, \boldsymbol{u}^{(0,N-1)})$ is obtained by marginalisation over $\boldsymbol{e}^{(0,N)}$ according to the formula

$$[I](\boldsymbol{p}, \boldsymbol{u}^{(0,N-1)}) = \frac{\int_{\mathcal{E}^{(0,N)}} I(\boldsymbol{p}, \boldsymbol{u}^{(0,N-1)}, \boldsymbol{e}^{(0,N)})\, \mathrm{d}\boldsymbol{e}}{V(\mathcal{E}^{(0,N)})} \tag{12}$$

Because the parameter $\boldsymbol{p}$ and the precise input sequence $\boldsymbol{u}^{(0,N-1)}$ are unknown, further marginalisation using the a priori information $\boldsymbol{p} \in \mathcal{P}^0$ and the set of feasible input sequences $\mathcal{U}_i^{(0,N-1)}$ is performed. This yields

$$[I](\boldsymbol{p}, \mathcal{U}_i^{(0,N-1)}) = \frac{\int_{\mathcal{U}_i^{(0,N-1)}} I(\boldsymbol{p}, \boldsymbol{u}^{(0,N-1)})\, \mathrm{d}\boldsymbol{u}^{(0,N-1)}}{V(\mathcal{U}_i^{(0,N-1)})} \tag{13}$$

the expected information content for a given set of input sequences $\mathcal{U}^{(0,N-1)}$ and the parameter $\boldsymbol{p}$, and

$$[I](\mathcal{P}_0, \mathcal{U}_i^{(0,N-1)}) = \frac{\int_{\mathcal{P}_0} [I](\boldsymbol{p}, \mathcal{U}_i^{(0,N-1)})\, \mathrm{d}\boldsymbol{p}}{V(\mathcal{P}_0)} \tag{14}$$

the expected information content for the feasible set of input sequences $\mathcal{U}_i^{(0,N-1)}$ and the set of a priori consistent parameters $\mathcal{P}_0$.

doi: 10.1049/iet-syb.2009.0030

The information measure $[I](\mathcal{P}_0, \mathcal{U}_i^{(0,N-1)})$ can now be used to determine the most informative set of input sequences:

Proposition 3: Let $[\bar{I}](\mathcal{U}_i^{(0,N-1)}, \mathcal{P}_0)$ be the expected information content for $\boldsymbol{u} \in \mathcal{U}_i^{(0,N-1)}$. If

$$[\bar{I}](\mathcal{U}_{i^*}^{(0,N-1)}, \mathcal{P}_0) \geq [\bar{I}](\mathcal{U}_i^{(0,N-1)}, \mathcal{P}_0) \quad \forall i \in \mathcal{I}(1, M) \tag{15}$$

then $\mathcal{U}_{i^*}^{(0,N-1)}$ is the maximal informative set of input sequences with respect to parameter identification.

3.3 Approximation of information measure

Unfortunately, neither the expected information content nor the SCP for a given measurement $\mathcal{P}^*$ can be computed, therefore $[I](\mathcal{P}_0, \mathcal{U}_i^{(0,N-1)})$ is approximated by $[\bar{I}](\mathcal{P}_0, \mathcal{U}_i^{(0,N-1)})$.

First of all, $\mathcal{P}^*$ is outer approximated by $\bar{\mathcal{P}}^*$ using the algorithm presented in Section 2.3. This yields a lower bound on the information content of a particular measurement

$$\bar{I}(\boldsymbol{p}, \boldsymbol{u}_i^{(0,N-1)}, \boldsymbol{e}^{(0,N)}) = \frac{V(\mathcal{P}_0 \setminus \bar{\mathcal{P}}^*(\boldsymbol{u}_i^{(0,N-1)}, \bar{\boldsymbol{y}}^{(0,N)}))}{V(\mathcal{P}_0)} \tag{16}$$

Using this approximation, the integral defining the expected information content $[I](\mathcal{U}_i^{(0,N-1)}, \mathcal{P}_0)$ is approximated via a Monte-Carlo approach

$$[\bar{I}](\mathcal{P}^0, \mathcal{U}_i^{(0,N-1)}) = \frac{1}{s_p s_u s_e} \sum_{j_1=1}^{s_p} \sum_{j_2=1}^{s_u} \sum_{j_3=1}^{s_e} \bar{I}(\boldsymbol{p}_{j_1}, \boldsymbol{u}_{j_2}^{(0,N-1)}, \boldsymbol{e}_{j_3}^{(0,N)}) \tag{17}$$

in which $\boldsymbol{p}_{j_1}$, $\boldsymbol{u}_{j_2}^{(0,N-1)}$ and $\boldsymbol{e}_{j_3}^{(0,N)}$ are obtained by drawing random samples from $\mathcal{P}_0$, $\mathcal{U}_i^{(0,N-1)}$ and $\mathcal{E}^{(0,N)}$, respectively.

Basically, the system is simulated for different parameters, input sequences and measurement errors. Based on these artificial data, the SCP is determined and used to approximate the expected information content.

It has to be emphasised that the quality of the approximation of $[\bar{I}](\mathcal{U}_i^{(0,N-1)}, \mathcal{P}_0)$ strongly depends on s_p, s_u and s_e, the number of samples. These numbers should be chosen sufficiently high, such that convergence of $[\bar{I}](\mathcal{P}^0, \mathcal{U}_i^{(0,N-1)})$ is observed. Depending on the non-linearity of the system, the required number of samples can vary strongly.

Remark 3: In this work we focus on the sets themselves and not on the probability distribution on the sets. Therefore no weighting, corresponding to an a priori probability of the measurements disturbances and the a priori consistent parameters is considered. An extension is straightforward and uses the a priori probabilities of the different variables.

4 Example

In order to illustrate the proposed experimental design, parameter identification and model falsification scheme, a simple time discrete model of the MAP kinase cascade is analysed, as illustrated in Fig. 1. The MAPK cascade plays a crucial role in cell differentiation, proliferation and other signal transduction pathways [28].

4.1 Model of the MAPK cascade

The model of the MAPK cascade considered here is build up of the three different kinases MAPKKK, MAPKK and MAPK which are unphosphorylated in the absence of signal. Phosphorylation and associated activation is performed by the upstream kinase. The difference equations describing the system dynamics are

$$\begin{aligned} x_1^{(k+1)} &= x_1^{(k)} + \Delta T\, v_1^{(k+1)}, \quad x_1^{(0)} = x_{1,0} \\ x_2^{(k+1)} &= x_2^{(k)} + \Delta T\, v_2^{(k+1)}, \quad x_2^{(0)} = x_{2,0} \\ x_3^{(k+1)} &= x_3^{(k)} + \Delta T\, v_3^{(k+1)}, \quad x_3^{(0)} = x_{3,0} \end{aligned} \tag{18}$$

in which x_1 is the concentration of MAPKKK-P, x_2 is the concentration of MAPKK-P and x_3 is the concentration of MAPK-P, all given in nM. Production and degradation of the different kinases are not considered because these happen on a slower timescale. Using mass conservation, the unphosphorylated states have been eliminated to reduce the model order.

In the following, it is distinguished between two different models for the reaction fluxes:

Model 1: The first model of the MAPK cascade is a simple chain of phosphorylations. M_0 controls the activation of MAPKKK, MAPKKK-P controls the activation of MAPKK and so on. The reaction fluxes v_i are modelled

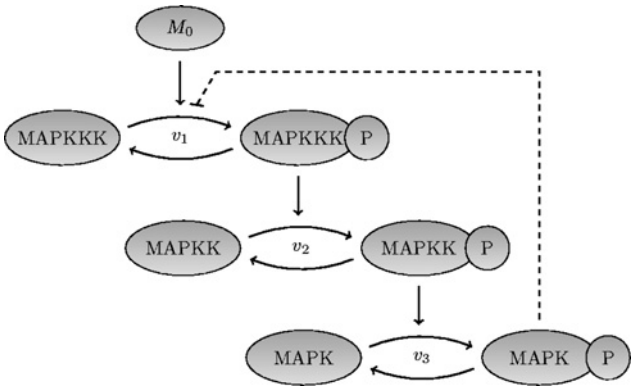

Figure 1 *Illustration of the MAP-kinase-cascade*

using mass-action kinetics

$$\begin{aligned} v_1^{(k)} &= k_{-1}x_1^{(k)} - k_1(M_1 - x_1^{(k)})u^{(k)} \\ v_2^{(k)} &= k_{-2}x_2^{(k)} - k_2(M_2 - x_2^{(k)})x_1^{(k)} \\ v_3^{(k)} &= k_{-3}x_3^{(k)} - k_3(M_3 - x_3^{(k)})x_2^{(k)} \end{aligned} \tag{19}$$

The concentration of the enzyme M_0 can be interpreted as the input to the system and is denoted by u.

Model 2: In the second model, besides the phosphorylation cascade included in model 1, also a feedback inhibition from MAPK-P onto the phosphorylation of MAPKKK is considered (dashed line in Fig. 1). This yields the modified reaction flux v_1

$$v_1^{(k)} = k_{-1}x_1^{(k)} - k_1(M_1 - x_1^{(k)})\frac{u^{(k)}}{1 + k_4 x_3^{(k)}} \tag{20}$$

v_2 and v_3 are equivalent to those in model 1. The modification of v_1 yields a rational system which can be transformed to a polynomial one by multiplication with the denominator $1 + k_4 x_3^{(k)}$, as explained in Section 2.2.1.

The two discrete time models are the time discretisation, using an implicit Euler scheme, of the continuous time models, describing the signalling pathways. The nominal parameter values are given in Table 1.

In the following, model 1 is assumed to describe the process and measurement data are generated using model 1. Besides parameter identification for model 1 a goal is to falsify model 2.

It is assumed that the concentrations of all three phosphorylated kinases are measurable

$$\boldsymbol{y}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)})^{\mathrm{T}} + \boldsymbol{e}^{(k)} \tag{21}$$

in which $\boldsymbol{e}$ is the uniformly distributed measurement noise with $\boldsymbol{e} \in \mathcal{E}^{(0,N)}$ and

$$\mathcal{E}^{(0,N)} = \{\boldsymbol{e}^{(0,N)} | -\bar{\boldsymbol{e}} \le \boldsymbol{e}^{(k)} \le \bar{\boldsymbol{e}},\ \forall k \in \mathcal{I}(0, N)\} \tag{22}$$

in which $\bar{\boldsymbol{e}} = (0.02, 100, 100)^{\mathrm{T}}$ nM. This absolute error corresponds to a relative error of 15% and is realistic for Western Blots which are typically used to measure protein concentrations [29].

To reduce the problem size and to simplify the visualisation of the result it is assumed that the ratios of forward to backward reaction rates $r_i = k_i/k_{-i}$ are known, for instance from previously performed steady-state measurements. Furthermore, the total amount of the different kinases M_i is considered to be known. Therefore, for both models just the absolute values, here k_{-i} for $i = 1$, 2, 3, are in the following considered as uncertain

$$\boldsymbol{p} = (k_{-1}, k_{-2}, k_{-3})^{\mathrm{T}} \tag{23}$$

The initial set $\mathcal{P}^0$ for the estimation is set to

$$\mathcal{P}_0 = \{\boldsymbol{p} \in \mathbb{R}^3 \,|\, 10^{-3} \le p_i \le 10,\ \forall i \in \{1, 2, 3\}\} \tag{24}$$

Thus initial parameter uncertainties of four orders of magnitudes are considered. This is realistic for biological systems.

4.2 Experimental design for the MAPK cascade

Before any experiment is performed, the experiment with the highest expected information content is selected using model 1 and the set of a priori consistent parameters $\mathcal{P}_0$. Experimental constraints are that measurements can only be performed at eight different points in time, $N = 7$, and only pulses in M_0 with a nominal concentration of 10^{-3} nM are feasible. The design variable is the pulse length yielding

$$\mathcal{U}_i^{(1,7)} = \left\{ u^{(1,7)} \middle| \begin{array}{l} u^{(k)} \in \mathcal{U}^{\mathrm{h}},\ \forall k \in \mathcal{I}(1, i), \\ u^{(k)} \in \mathcal{U}^{\mathrm{s}},\ \forall k \in \mathcal{I}(i+1), \\ u^{(k)} \in \mathcal{U}^{\mathrm{l}},\ \forall k \in \mathcal{I}(i+2, 7) \end{array} \right\}, \forall i = \mathcal{I}(1, 7) \tag{25}$$

in which $\mathcal{U}^{\mathrm{l}}$ and $\mathcal{U}^{\mathrm{h}}$ denote the set of low and high enzyme

Table 1 Actual parameter values

Parameter	Value	Units	Parameter	Value	Units
k_1	5	1/(min nM)	k_{-1}	0.05	1/min
k_2	2	1/(min nM)	k_{-2}	0.1	1/min
k_3	0.001	1/(min nM)	k_{-3}	0.1	1/min
k_4	0.1	1/nM	ΔT	4	min
M_0	0.001	nM	M_1	3	nM
M_2	1200	nM	M_3	1200	nM

doi: 10.1049/iet-syb.2009.0030

concentration

$$\mathcal{U}^{\mathrm{l}} = \{u | 0 \leq u \leq 0.1 \times 10^{-3}\}$$
$$\mathcal{U}^{\mathrm{h}} = \{u | 0.9 \times 10^{-3} \leq u \leq 1.1 \times 10^{-3}\} \quad (26)$$

Here an input uncertainty of 10^{-4} nM is considered, which may arise from errors in pipettation, filtration and measurement. Furthermore, uncertainty and inaccuracies in the stimulus removal time are modelled by assuming the input directly after switching off as unknown with the bound $u^{(i+1)} \in \mathcal{U}^{\mathrm{s}} = \{u | 0 \leq u \leq 1.1 \times 10^{-3}\}$.

For all sets of input sequences $\mathcal{U}_i^{(1,7)}$, the expected information content is computed. This is done according to (17) using the explained Monte-Carlo method. For the weighting function (10), used to determine the information content of a single artificial measurement, $w = \prod_{i=1}^{3} p_i^{-1}$ is chosen. This enforces a uniform weighting on the logarithmically scaled axes. The resulting expected information content for the different input sequences is depicted in Fig. 2. The highest expected information content is obtained for a step length of six and the lowest one for a step length of 1. For a step length of 6, the expected amount of the parameter set $\mathcal{P}_0$ which can be qualified as inconsistent is 97.7%. For a step of length one, it is only 93.1%. The corresponding difference in the expected size of the consistent parameters $\bar{\mathcal{P}}^*$ is thus approximately a factor of three.

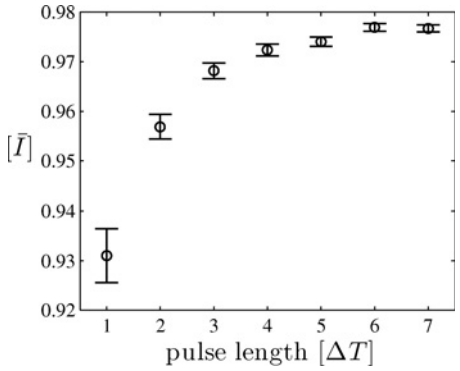

Figure 2 *Expected information content $[\bar{I}]$ and corresponding variance for the different set of input sequences, $\mathcal{U}_1^{(1,7)}, \ldots, \mathcal{U}_7^{(1,7)}$*

Discretisation $\Delta T = 4$ min (Table 1)

4.3 Parameter identification

For the examination of the properties of the developed scheme for the SCP computation an artificial experiment is now performed. To generate artificial measurement data, model 1 is simulated using the nominal parameter values and the nominal input sequence with a pulse length of six. The resulting output is corrupted by random measurement noise according to (22). The obtained artificial experimental data are depicted in Fig. 3. The approximated information content of this measurement is 0.994, thus 99.4% of $\mathcal{P}_0$ could be shown to be inconsistent with the measurement data.

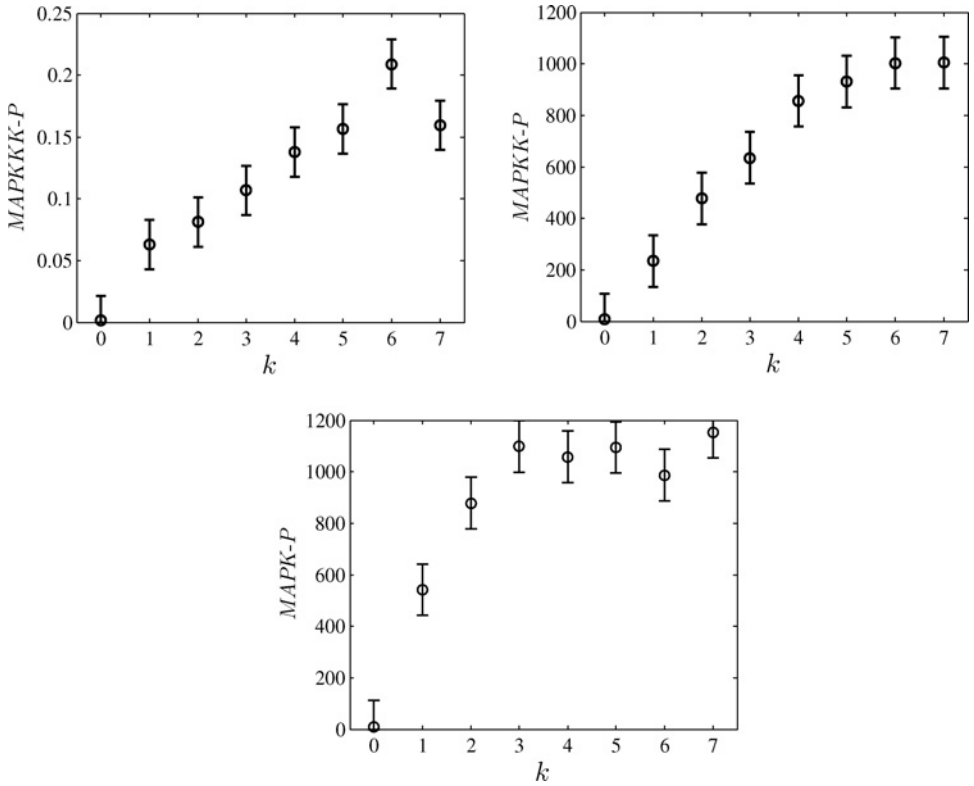

Figure 3 *Artificial experimental data for MAPKKK-P, MAPKK-P and MAPK-P and corresponding error bounds*

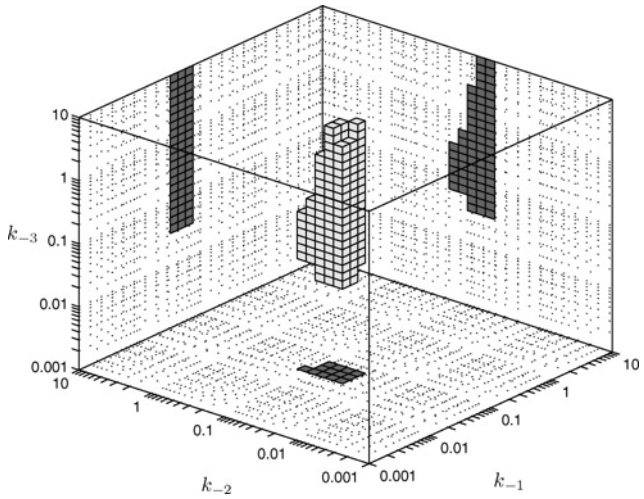

Figure 4 *Approximation $\bar{\mathcal{P}}^*$ (light grey) of SCP and the projections of $\bar{\mathcal{P}}^*$ (dark grey) on the planes for the set of input sequences with the highest expected information content, $\mathcal{U}_6^{(1,7)}$*

For the artificial experimental data depicted in Fig. 3, the SCP of model 1 is computed using the algorithm outlined in Section 2. The obtained result is shown in Fig. 4.

In order to evaluate the effect of the experimental design on the estimated SCP, also for the input with the lowest expected information content an artificial experiment is performed and the corresponding SCP approximated. The results can be found in Fig. 5. Here 98.6% of $\mathcal{P}_0$ can be ruled out and hence the volume of the remaining parameter set consistent with the experimental data differs by a factor of 2.4. This can be seen, also in case of extreme limited a priori knowledge, the proposed experimental design approach can help to select the most informative experiments.

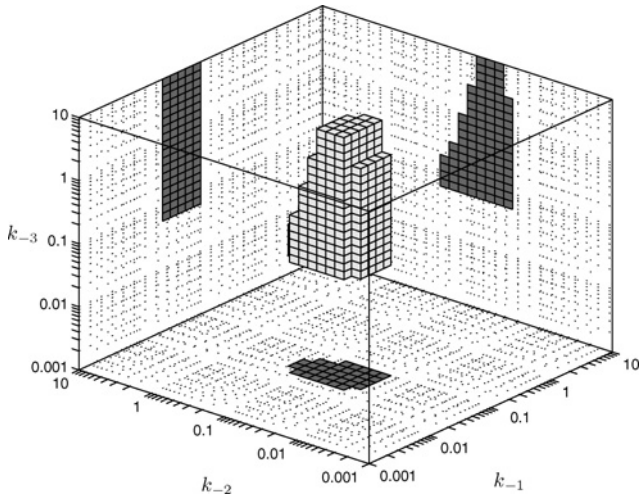

Figure 5 *Approximation $\bar{\mathcal{P}}^*$ (light grey) of SCP and the projections of $\bar{\mathcal{P}}^*$ (dark grey) on the planes using for the set of input sequences with the lowest expected information content, $\mathcal{U}_1^{(1,7)}$*

To rate the quality of the obtained outer approximation $\bar{\mathcal{P}}^*$ of the SCP using the proposed algorithm, 1000 plausible points in parameter space are computed using a sampling-based approach. A detailed analysis uncovers that 70.7% of the hypercubes that $\bar{\mathcal{P}}^*$ is built of contain at least one parameter sample and thus $\bar{\mathcal{P}}^*$ is a fairly good approximation of the SCP.

Compared to the sampling-based approach the main advantage of our method is that an outer approximation of the SCP is obtained. Thus all consistent parameters have to be contained in $\bar{\mathcal{P}}^*$. The traditional approach on the other hand yields an inner approximation of the SCP and important solutions can be missed. The computation time of both approaches is comparable for this example.

4.4 Model falsification

In the previous subsections, it has been shown how the proposed method can be applied for SCP computation. In this section the focus lies on the falsification of models. Therefore for model 2 the SCP is computed using the artificial experimental data shown in Fig. 3, which are obtained by simulating model 1.

This computation takes here only 2 min until the SCP-computation algorithm returns that $\bar{\mathcal{P}}^*$ is empty. Hence, no parameter in the considered a priori consistent parameter set $\mathcal{P}_0$ can reproduce the measured dynamics within the measurement error bounds. Thus model 2 is falsified.

Compared to standard algorithms which sample the set $\mathcal{P}_0$ the computational effort is small. Additionally and even more important, it can be guaranteed that no consistent parametrisation of model 2 exists. Standard algorithms just provide a falsification probability.

5 Conclusions

In this paper an approach for parameter identification, model falsification and experimental design using set-based methods based on work of Kuepfer *et al.* [9] is developed. For the classification of complete parameter sets a feasibility problem is defined, which is relaxed to a computationally efficient SDP. Based on this SDP, an algorithm to outer approximate the set of all consistent parameters of discrete time dynamical processes with rational right-hand side is developed.

The resulting approximation to the SCP directly describes the uncertainty in the parameters resulting from uncertain measurements. The result can also be applied for model falsification. Owing to these properties, the proposed approach provides valuable information to the modeler. Furthermore, the set-based approach is applied to define a measure for the information content of a specific experimental measurement. This measure can be used to select the most informative experimental setup with respect

doi: 10.1049/iet-syb.2009.0030

to parameter identification, even in case of extremely limited a priori information and uncertain input sequences.

To illustrate the method, it is applied to a simple model of the MAPK cascade. This example highlights the advantages of the set-based approach for parameter identification, model falsification and experimental design over classical methods.

6 Acknowledgments

The authors thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology at the University of Stuttgart as well as the German Federal Ministry for Education and Research (BMBF).

7 References

[1] LENNART L.: 'System identification: theory for the user' (Prentice Hall PTR, December 1998)

[2] BALSA-CANTO E., PEIFER M., BANGA J.R., TIMMER J., FLECK C.: 'Hybrid optimization method with general switching strategy for parameter estimation', *BMC Syst. Biol.*, 2008, **2**, p. 26

[3] MOLES C.G., MENDES P., BANGA J.R.: 'Parameter estimation in biochemical pathways: a comparison of global optimization methods', *Genome Res.*, 2003, **13**, (11), pp. 2467–2474

[4] RODRIGUEZ-FERNANDEZ M., EGEA J.A., BANGA J.R.: 'Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems', *BMC Bioinf.*, 2006, **7**, p. 18

[5] ROBERT C.P., CASELLA G.: 'Monte Carlo statistical methods' (Springer, 2004)

[6] BALSA-CANTO E., ALONSO A.A., BANGA J.R.: 'Computational procedures for optimal experimental design in biological systems', *IET Syst. Biol.*, 2008, **2**, (4), pp. 163–172

[7] JOSHI M., SEIDEL-MORGENSTERN A., KREMLING A.: 'Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems', *Metabolic Eng.*, 2006, **8**, (5), pp. 447–455

[8] KIEFFER M., WALTER E.: 'Interval analysis for guaranteed nonlinear parameter and state estimation', *Math. Comput. Model. Dyn. Syst.*, 2005, **11**, (2), pp. 171–181

[9] KUEPFER L., SAUER U., PARRILO P.A.: 'Efficient classification of complete parameter regions based on semidefinite programming', *BMC Bioinf.*, 2007, **8**, p. 12

[10] JAULIN L., KIEFFER M., DIDRIT O., WALTER E.: 'Applied interval analysis' (Springer, Heidelberg, 2001)

[11] JOHNSON T., TUCKER W.: 'Rigorous parameter reconstruction for differential equations with noisy data', *Automatica*, 2008, **44**, pp. 2422–2426

[12] WALTER E., KIEFFER M.: 'Guaranteed nonlinear parameter estimation in knowledge-based model', *J. Comput. Appl. Math.*, 2007, **199**, (2), pp. 277–285

[13] TUCKER W., KUTALIK Z., MOULTON V.: 'Estimating parameters for generalized mass action models using constraint propagation', *Math. Biosci.*, 2007, **208**, pp. 607–620

[14] PARRILO P.A.: 'Semidefinite programming relaxations for semialgebraic problems', *Math. Program. B*, 2003, **96**, pp. 293–320

[15] HASENAUER J., RUMSCHINSKI P., WALDHERR S., BORCHERS S., ALLGÖWER F., FINDEISEN R.: 'Guaranteed steady-state bounds for uncertain chemical processes'. Proc. Int. Symp. on Advanced Control of Chemical Processes, Adchem, 2009, pp. 674–679

[16] WALDHERR S., FINDEISEN R., ALLGÖWER F.: 'Global sensitivity analysis of biochemical reaction networks via semidefinite programming'. Proc. 17th IFAC World Congress, 2008, pp. 9701–9706

[17] BORCHERS S., RUMSCHINSKI P., BOSIO S., WEISMANTEL R., FINDEISEN R.: 'Model discrimination and parameter estimation via infeasibility certificates for dynamical biochemical reaction networks'. 15th IFAC Symp. on System Identification, 2009, pp. 245–250

[18] KREMLING A., FISCHER S., GADKAR K., *ET AL.*: 'A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions', *Genome Res.*, 2004, **14**, pp. 1773–1785

[19] ASPREY S.P., MACCHIETTO S.: 'Designing robust optimal dynamic experiments', *J. Process Control*, 2002, **12**, pp. 545–556

[20] VOIT E.O.: 'Computational analysis of biochemical systems' (Cambridge University Press, Cambridge, UK, 2000)

[21] FAISAL S., LICHTENBERG G., WERNER H.: 'A polynomial approach to structural gene dynamics modelling'. Proc. 16th IFAC World Congress, Prague, 2005

[22] DEUFLHARD P., BORNEMANN F.: 'Scientific computing with ordinary differential equations' (Springer, New York, USA, 2002)

[23] VANDENBERGHE L., BOYD S.: 'Semidefinite programming', *SIAM Rev.*, 1996, **38**, pp. 49–95

[24] FUJIE T., KOJIMA M.: 'Semidefinite programming relaxation for non-convex quadratic programs', *J. Global Optim.*, 1997, **10**, pp. 367–380

[25] BOYD S., VANDENBERGHE L.: 'Convex optimization' (Cambridge University Press, Cambridge, UK, 2004)

[26] STURM J.F.: 'Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones', *Optim. Methods Softw.*, 1999, **11**, pp. 625–653

[27] CHEN B.H., ASPREY S.P.: 'On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations', *Ind. Eng. Chem. Res.*, 2003, **42**, pp. 1379–1390

[28] ORTON R.J., STURM O.E., VYSHEMIRSKY V., CALDER M., GILBERT D.R., KOLCH W.: 'Computational modelling of the receptor-tyrosine-kinase-activated mapk pathway', *Biochem. J.*, 2005, **392**, pp. 249–261

[29] SZALLASI Z.: 'Biological data acquisition or system level modeling – an exercise in the art of compromise', in SZALLASI Z., STELLING J., PERIWAL V. (EDS.): 'System modeling in cellular biology' (MIT Press, 2006)

doi: 10.1049/iet-syb.2009.0030