Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models

P. Kügler

RICAM-Report 2012-10

Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models

Philipp Kügler[1]

1 Philipp Kügler, Mathematical Methods in Molecular and Systems Biology, Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria

∗ **E-mail: philipp.kuegler@oeaw.ac.at**

Abstract

The inference of reaction rate parameters in biochemical network models for well mixed conditions from time series concentration data is a central task in computational systems biology. The network dynamics usually are described by the chemical master equation, the Fokker Planck equation, the linear noise approximation or the macroscopic rate equation. The inverse problem of estimating the parameters of the underlying differential equation can approached in deterministic and stochastic ways and available methods often involve the difference between individual or mean concentration traces and model predictions when maximizing likelihoods, minimizing regularized least squares functionals, approximating posterior distributions or sequentially processing the data. In this article we assume that the biological reaction network can be at least partially and repeatedly observed over time such that low order statistical moments or central moments at various times for the number of molecules of the chemical species involved can be approximated from the data. Furthermore, we consider closed systems of nonlinear ordinary differential equations that approximatively describe the time evolution of the statistical moments or central moments, can be derived from the chemical master equation or their approximations and depend on the reaction rate parameters. For infering the rate parameters we then suggest to not only consider the distance between the sample mean and the mean prediction of the equations but also to take the error in higher moments explicitly into acount. Cost functions that involve higher statistical moments may form landscapes in the parameter space that have more pronounced curvatures at the minimizer and hence may weaken or even overcome parameter sloppiness and uncertainty. As a consequence both deterministic and stochastic parameter inference algorithms may be improved with respect to accuracy and efficiency. We demonstrate the concept of moment fitting for parameter inference by means of illustrative stochastic biological models from the literature.

Author Summary

Based on the chemical master equation or its approximations for stochastic biochemical reaction networks in well mixed conditions, we consider nonlinear closed systems of ordinary differential equations that depend on the reaction rate parameters and approximatively describe the time evolution of the statistical moments for the number of molecules of the chemical species involved. Under the assumption that the time course of the species molecule numbers can at least be partially and repeatedly observed, we approach the problem of parameter inference by computing moment estimates from the data and by comparing them to the predictions obtained from the differential equations. In numerical tests on three illustrative stochastic biological models from the literature we observe that problems of parameter unidentifiability or slow parameter convergence can be weakened or even overcome if not only the mean error but also the covariance error is taken into account in the parameter inference strategy. We conclude that the consideration of errors of higher moments may have a positive impact on parameter inference as the landscape in parameter space built by respective cost functions show more pronounced curvatures at the solution.

Introduction

The traditional approach to modelling of biological reaction networks is based on deterministic mass action kinetics in which the time course of the species concentrations averaged over the population is described by a set of coupled ordinary differential equations [1], often referred to as the macroscopic rate equations. For the description of intra-cellular processes characterized by a low number of reacting molecules the stochastic modelling approach [2] is an alternative that explicitly takes the discreteness and stochasticity of chemical kinetics into account. In well-mixed conditions the system dynamics are captured by the Kolmogorov differential equation, also referred to as the chemical master equation, for the transition probability kernel of a continuous time Markov process with discrete state space. Numerical solutions of the master equation, even after projection to finite state space [3], are computationally expensive, but realizations of the stochastic process can be achieved by the Gillespie algorithm and its variants [4], [2]. A stochastic differential equation approximation to the true process is given by the chemical Langevin equation [5] to which a Fokker Planck equation can be associated that describes the probability density function of the continuous state variable. As an alternative approximative description the linear noise approximation [6] features a partial differential equation for the probability distribution of the fluctuations around the deterministic part governed by the macroscopic rate equation.

Parameter estimation in differential equation models is a classic nonlinear inverse problem that arises in a variety of scientific, industrial and financial applications and is tackled both in deterministic and statistical ways [7], [8], [9], [10]. The advances of experimental biology even at the single cell level [11], [12] along with the ever growing quality and amount of species concentration data have also stimulated recent interest in the inference of reaction rate parameters in kinetic biological models. Many of the various methods suggested have in common that at some stage they compare the time series data to parameter-dependent predictions or components of the chosen differential equation model. For instance, [13], [14], [15] compare time series data to the solution of the macroscopic rate equation in the minimization of unregularized and regularized least squares functionals by deterministic and stochastic optimization routines, [16] compares finite differences of time series data to the drift term of the chemical Langevin equation in Bayesian inference, [17] compares time series data to the solution of the macroscopic rate equation or to averaged outcomes of the Gillespie algorithm in approximate Bayesian inference, [18] compares time series data to the solution of the macroscopic rate equation in maximum likelihood estimation, [19] compares time series data to the mean component of the linear noise approximation in Bayesian inference, [20] compares probability density or cumulative density functions obtained from the data to their counterparts constructed from repeated realizations of the Gillespie algorithm, and [21], [22], [23] sequentially compare time series data to the solution of the macroscopic rate equation in extended Kalman filtering or nonlinear observing. In inferring the rate parameters q from a time series data vector x, some of the available approaches explicitly take parameter dependent approximations of the mean (the first moment) $\mu^1(q)$ as well as of higher moments, e.g., of the variance $\mu^{c,2}(q)$ (the second central moment), of the state variable $x(q)$ into account, for instance, when building the (simplified) likelihood function

$$L(q) = \prod_{i=1}^{n_t} \frac{1}{\sqrt{2\pi\mu^{c,2}(t_i;q)^2}} \exp\left(-\frac{(x_i - \mu^1(t_i;q))^2}{2\mu^{c,2}(t_i;q)^2}\right)$$

for a multivariate normal distribution or a weigthed sum of squared residuals

$$SS(q) = \sum_{i=1}^{n_t} \frac{(x_i - \mu^1(t_i;q))^2}{\mu^{c,2}(t_i;q)^2},$$

see [24], [19], [16], [25]. However, an adjustment of model parameters in order to actually also *fit higher statistical sample moments* derived from the data, e.g., by computing the difference between the sample variance $\hat{\mu}^{c,2}$ and the variance $\mu^{c,2}(q)$ predicted by the model, so far has - to the best of our knowledge - not

been considered. Furthermore, studies of parameter sensitivities and identifiability [26], [27], [28], [29], [30] typically are based on macroscopic rate equation models where the Hessian matrix of $-L(q)$ or $SS(q)$ (usually with $\mu^{c,2}(t_i;q)$ replaced by parameter independent weights) to be analyzed only involves the distance between the data x_i and the first order moment $\mu^1(t_i;q)$. Small or zero eigenvalues of the Hessian at a minimizer point to small or even vanishing curvatures of the landscape function in parameter space, a situation referred to as parameter sloppiness [27] as parameters then are only poorly constrained by the data and not uniquely identifiable. Parameter sensitivities have also been studied for stochastic chemical kinetics models [31] based on the linear noise approximation, still the distance between sample and model moments is not part of this analysis.

In this paper we present a moment fitting approach to parameter inference in stochastic biological models that to the best of our knowledge has not been studied before. First, we suppose that the state variable vector of molecule numbers can be partially - both with respect to time and state variable components - and repeatedly, say N times, observed such that the statistical moments $\hat{\mu}^o$ up to some order $\bar{k} > 1$ of interest for the observable components can be approximated from the time series data. Second, we consider closed systems of ordinary differential equations

$$\begin{aligned} \frac{\partial}{\partial t}\mu(t) &= F(\mu(t), q),\ t \in (0, t_f], \\ \mu(0) &= \mu_0 \end{aligned}$$

that describe the time evolution of the parameter dependent moment approximations $\mu(q)$ up to the order $\bar{k}$ and can be derived from the chemical master equation, the Fokker Planck equation or the linear noise approximation [32], [33], [34], [35], [36], [6]. Now, let $\mu^o(q)$ denote those components of $\mu(q)$ that only depend on the observable state variables. For solving the parameter inference problem we then suggest to utilize the distance $d(\hat{\mu}^o, \mu^o(q))$ between the sample moments $\hat{\mu}^o$ and the equation output $\mu^o(q)$ in global and local minimization techniques or approximate Bayesian methods. For $\bar{k} > 1$ the span of the eigenvalues of the Hessian matrix of $d(\hat{\mu}^o, \mu^o(q))$ may be strongly reduced in comparison to cost functions that only involve the first moment (the mean) such that problems of non-identifiability or parameter sloppiness can be relieved or even overcome. That way the efficiency and accuracy of distance based parameter inference strategies may be enhanced by higher order moment fitting. In comparison to [20], where the focus is on a comparison of probability density functions rather than on a comparison of statistical moments, model predictions based on the above mentioned ODE system are computationally much cheaper than those based on repeated realizations of the Gillespie algorithm. In a sequence of illustrative examples from the literature we demonstrate the potential benefits of the moment fitting approach.

Results

We studied the concept of moment fitting for parameter inference in stochastic biological models by means of three reference examples, see Materials and Methods for all model details, and chose $\bar{k} = 2$ as highest moment order in all of our tests.

Results for Linear Birth and Death Process

The Kolmogorov differential equation (11) in this example reads as

$$\frac{\partial \pi}{\partial t}(x,t) = q_1(x-1)\pi(x-1,t) - (q_1+q_2)x\pi(x,t) + q_2(x+1)\pi(x+1,t)$$

and implies the ODE system

$$\begin{aligned} \frac{\partial \mu^1}{\partial t}(t) &= (q_1 - q_2)\mu^1(t),\ \mu^1(0) = x_0 \\ \frac{\partial \mu^{c,2}}{\partial t}(t) &= 2(q_1 - q_2)\mu^{c,2}(t) + (q_1 + q_2)\mu^1(t),\ \mu^{c,2}(0) = 0 \end{aligned} \tag{1}$$

for the mean $\mu^1(\cdot; q)$ (the first moment) and the variance $\mu^{c,2}(\cdot; q)$ (the second central moment) approximation of the discrete state variable x, see Supporting Information for a derivation. Due to the linearity of the rate function h with respect to x, the very same system (1) is obtained if the true stochastic process is approximated by the diffusion approximation (12), (13) or by the linear noise approximation (22). Furthermore, $\mu^1(\cdot; q)$ and $\mu^{c,2}(\cdot; q)$ coincide with the true moments $m^1(\cdot; q)$ and $m^{c,2}(\cdot; q)$. Another consequence of the linearity of h is that (1) admits an analytical solution given by

$$\mu^1(t; q) = \mu^1(0)e^{(q_1 - q_2)t}, \tag{2}$$

$$\mu^{c,2}(t; q) = \mu^1(0)\frac{q_1 + q_2}{q_1 - q_2}e^{(q_1 - q_2)t}(e^{(q_1 - q_2)t} - 1). \tag{3}$$

We have simulated the true stochastic process $N = 1000$ times by means of the Gillespie algorithm [4]. Figure 1 B shows 3 (out of 1000) realizations as example, Figures 1 C and D display the sample mean $\hat{\mu}^1$ and the sample variance $\hat{\mu}^{c,2}$ at the process observation times t_j derived from the data.

In order to infer the rate parameters q_1 and q_2 we first utilized the cost (or distance) function

$$d^1(q) = \sum_{j=1}^{n_t} (\hat{\mu}_j^1 - \mu^1(t_j; q))^2 \tag{4}$$

for a comparison of the sample mean with the analytic mean expression (2). Choosing the initial parameter guess $q^0 = [1, 11]^T$ and the MATLAB [37] trust region optimization algorithm with default setting we obtained the parameter solution $\tilde{q} = [1.2392, 2.2487]^T$ after 10 iteration steps. Though the mean concentration data are perfectly fit by $\mu^1(t; \tilde{q})$, see Figure 2 A, the solution $\tilde{q}$ strongly deviates from the true parameter values $q^* = [3, 4]^T$. If $\mu^{c,2}(t, \tilde{q})$ is used to predict the sample variance, large errors can also be observed in the data space, see Figure 2 B.

The problem of non-identifiability is overcome if not only the first moment but also the second central moment is fitted to the available data, e.g., by minimizing the cost function

$$d^2(q) = \sum_{j=1}^{n_t} (\hat{\mu}_j^1 - \mu^1(t_j; q))^2 + 0.1 \sum_{j=1}^{n_t} (\hat{\mu}_j^{c,2} - \mu^{c,2}(t_j; q))^2. \tag{5}$$

Again starting from $q^0 = [1, 11]^T$ the minimization of $d^2(q)$ after 13 iteration steps led to the parameter estimate $\hat{q} = [2.9492, 3.9415]^T$ which is nearly identical to the true solution q^*. Figures 2 C and D indicate the quality of the data fit by the analytic moment expressions $\mu^1(t; \hat{q})$ and $\mu^{c,2}(t; \hat{q})$.

In a further test, we utilized the likelihood function

$$L(q) = \prod_{j=1}^{n_t} \frac{1}{\sqrt{2\pi\mu^{c,2}(t_j; q)^2}} \exp\left(-\frac{(\hat{\mu}_j^1 - \mu^1(t_j; q))^2}{2\mu^{c,2}(t_j; q)^2} \right), \tag{6}$$

which compares $\hat{\mu}_j^1$ to $\mu^1(t_j; q)$ but *does not involve* the error between $\hat{\mu}_j^{c,2}$ and $\mu^{c,2}(t_j; q)$, in a MCMC Metropolis random walk algorithm [2]. Even if we chose the favourable gamma distributions

$$q_1 \sim \Gamma(3, 1) \text{ and } q_2 \sim \Gamma(3, 1)$$

as priors for the parameters, the algorithm failed to yield acceptable marginal posterior density distributions due to the ignorance of the sample variance $\hat{\mu}^{c,2}$, see Figure 3 for details.

Results for Dimerisation Kinetics

Based on the Fokker Planck equation (13) of the diffusion modelling approach

$$\frac{\partial}{\partial t}p(\chi,t) = -\frac{\partial}{\partial \chi}\left\{(-q_1\chi(\chi-1)+q_2(\chi_0-\chi))p(\chi,t)\right\} + \frac{1}{2}\frac{\partial^2}{\partial \chi^2}\left\{(2q_1\chi(\chi-1)+2q_2(\chi_0-\chi))p(\chi,t)\right\}$$

the normal moment closure technique yields the nonlinear ODE system

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1(t) &= q_1\mu^1(t)(1-\mu^1(t)) + q_2(x_0-\mu^1(t)) - q_1\mu^{c,2}(t),\ \mu^1(0)=x_0 \\
\frac{\partial}{\partial t}\mu^{c,2}(t) &= -2q_1(2\mu^1(t)+2)\mu^{c,2}(t) - 2q_2\mu^{c,2}(t) \\
&\quad +2q_1\mu^1(t)(\mu^1(t)-1) + 2q_2(\chi_0-\mu^1(t)),\ \mu^{c,2}(0)=0
\end{aligned}$$

for the (approximative) mean $\mu^1(t;q)$ and the (approximative) variance $\mu^{c,2}(t;q)$ of the continuous state variable $\chi(t)$, see Supporting Information for the derivation. The true stochastic process was simulated $N = 1000$ times by means of the Gillespie algorithm [4] with initial molecule number $x_0 = 301$ and rate parameters $q^* = [0.005, 0.03]^T$.

We first only focused on the sample mean data $\hat{\mu}^1$ and minimized the least squares objective function

$$d^1(q) = \sum_{j=1}^{n_t}(\hat{\mu}_j^1 - \mu^1(t_j;q))^2 \tag{7}$$

by means of the MATLAB [37] trust region routine. The initial parameter guess chosen was $q^{(0)} = [0.0047, 0.0177]^T$, for any q the model predictions $\mu^1(t_j;q)$, $j = 1,\ldots,n_t$, were obtained by solving the above mentioned ODE system, and the gradient of $d^1(q)$ was provided by means of the adjoint method, see Supporting Information. Figures 4 A,C,E show that convergence of the iterates towards $\tilde{q} = [0.005, 0.0294]^T$ is obtained after 180 iteration steps. Though the mean data in this example is sufficient to obtain reliable parameter estimates (also for more distant initial guesses), a significant computational speed up is gained if not only the mean data but also the variance data $\hat{\mu}^{c,2}$ is taken into account. Figures 4 B,D,F display the performance of the same optimization algorithm with identical initial guess if applied to the minimization of the alternative objective function

$$d^2(q) = \sum_{j=1}^{n_t}(\hat{\mu}_j^1 - \mu^1(t_j;q))^2 + 0.1\sum_{j=1}^{n_t}(\hat{\mu}_j^{c,2} - \mu^{c,2}(t_j;q))^2. \tag{8}$$

The accuracy of the parameter estimate $\hat{q}$ obtained with (8) is the same as of $\tilde{q}$ obtained with (7), however, convergence now is already achieved after 55 iterations. The outcome $d^2(\hat{q}) > d^1(\tilde{q})$ is solely due to the additional variance term in (8) and does not allow for a comparative judgement of the inferred parameters.

Results for $p53$ Signalling System

The linear noise approximation (14) yields a nonlinear ODE system describing the temporal development of the mean approximation $\mu^1(t;q) \in \mathbb{R}^3$ and the covariance matrix approximation $\mu^{c,2}(t;q) \in \mathbb{R}^{3\times 3}$. For data generation, the true stochastic process was simulated $N = 1000$ times by means of the Gillespie algorithm [4] with the initial molecule numbers $x_0 = [10, 10, 80]^T$ and the rate parameter vector $q^* = [90, 0.002, 1.7, 1, 1.1, 0.8, 2]^T$. First, we supposed that the two components x_1 and x_2 of the state vector $x \in \mathbb{N}_0^3$ can be observed. The minimization of the objective function

$$d^1(q) = \sum_{j=1}^{n_t}(\hat{\mu}_{1,j}^1 - \mu_1^1(t_j;q))^2 + \sum_{j=1}^{n_t}(\hat{\mu}_{2,j}^1 - \mu_2^1(t_j;q))^2 \tag{9}$$

with the initial guess $q_j^{(0)} = (1 + 0.1 \cdot (-1)^j) \cdot q_j^*$, $j = 1, ..., 7$, showed that the corresponding sample mean data $\hat{\mu}_1^1$ and $\hat{\mu}_2^1$ are not sufficient to identify the true vector q^*. Though the parameter estimate $\tilde{q}$ obtained after 120 iterations is able to reproduce the data, see Figures 5 A and B, it features a maximal relative error of 156% in its second component, see Figures 6 A and B for details. For comparison, we next supposed that only the component x_1 is amenable to observations but build both the corresponding mean estimate $\hat{\mu}_1^1$ and the variance estimate $\hat{\mu}_{11}^{c,2}$ from the data, see Figures 5 A and C. The minimization of the objective function

$$d^2(q) = \sum_{j=1}^{n_t} (\hat{\mu}_{1,j}^1 - \mu_1^1(t_j; q))^2 + 0.01 \sum_{j=1}^{n_t} (\hat{\mu}_{11,j}^{c,2} - \mu_{11}^{c,2}(t_j; q))^2 \tag{10}$$

(with the same initial guess and MATLAB trust region algorithm as for (9), gradient information again provided by the adjoint method) led - after only 33 iterations - to an improved parameter estimate $\hat{q}$ whose maximal relative error (again in the second component) was reduced to 22.44%, see Figures 6 C and D for details.

Discussion

Using three test models from the systems biology literature we observed that the fitting of both mean and variance expressions described by moment differential equations to corresponding sample moments may enhance parameter identifiability and the performance of parameter identification algorithms. These observations can be understood by an examination of the cost functions that have been used for the parameter inference. In the linear birth death example the mean expression (**??**) shows that any parameter combination $(\tilde{q}_1, \tilde{q}_2)^T$ with $\tilde{q}_1 - \tilde{q}_2 = -1$ equally well explains the sample mean data as the true parameter vector $q^* = (3, 4)^T$. The failure of the MCMC approach based on the likelihood (6) can be understood from plotting the negative log-likelihood along the line $\tilde{q}_2 = \tilde{q}_1 + 1$ which assumes its minimum at the boundary imposed by parameter positivity and far away from q^*. With respect to the cost function (4) we have $d^1(\tilde{q}) = d^1(q^*)$, see Figure 7 A. This non-identifiability is also revealed by a parameter sensitivity analysis [26], [27], [28], [29], [30] based on an eigenvector decomposition of the Hessian matrix

$$H_{ij}^1(q) = \frac{\partial d^1(q)}{\partial q_i \partial q_j}, \quad i, j \in \{1, 2\}$$

of d^1. If evaluated at q^* the two (normalized) eigenvectors are $v_1(q^*) = [-\sqrt{0.5}, \sqrt{0.5}]^T$ and $v_2(q^*) = [-\sqrt{0.5}, -\sqrt{0.5}]^T$ with corresponding eigenvalues $\lambda_1(q^*) = 2.52 \cdot 10^4$ and $\lambda_2(q^*) = 0$. While $v_1(q^*)$ points towards the direction of maximal curvature (or the stiff direction), $\lambda_2(q^*) = 0$ indicates that there is no curvature at all along the direction of $v_2(q^*)$ (the soft or sloppy direction). A large eigenvalue spectrum, which in this extreme example spans infinitely many decades, is referred to as parameter sloppiness. The results of the analysis of the Hessian are also reflected in Figure 7 D which plots the level sets of (4). The cost function is minimal on a whole line, whose direction is given by $v_2(q^*)$, rather than on an isolated point. The situation significantly improves if instead of (4) the cost function (5) is chosen which also takes the mismatch between the sample covariance and the analytic expression (3) into account. The eigenvector decomposition of the Hessian matrix

$$H_{ij}^2(q) = \frac{\partial d^2(q)}{\partial q_i \partial q_j}, \quad i, j \in \{1, 2\}$$

of (5) evaluated at q^* yields the eigenvectors $v_1(q^*) = [-0.749, 0.663]^T$ and $v_2(q^*) = [-0.663, -0.749]^T$ with corresponding eigenvalues $\lambda_1(q^*) = 1.44 \cdot 10^5$ and $\lambda_2(q^*) = 306.14$. Though the directions of largest and smallest curvature are similar as before, the Hessian now is regular with a narrow span of the

eigenvalues. A plot of the level sets, see Figure 7 D, indicates that - as opposed to d^1 - the function d^2 admits a landscape with a sharp trough in the neighborhood of the isolated minimizer. The unique parameter identifiability is also evident from the quadratic behaviour of the cost function d^2 along the line $\tilde{q}_2 = \tilde{q}_1 + 1$ (with direction $[-\sqrt{0.5}, -\sqrt{0.5}]^T$), see Figure 7 C, with its global minimizer corresponding to q^*. These results hold true if the weighting parameter w_2 in the definition of d^2 is changed from 0.1 to, e.g., 1 or 0.01.

For the dimerisation example the level sets of the cost function (7) are shown in Figure 8 A. In this example the mean sample data is sufficient to uniquely determine the model parameter vector q^*, at least if a nearby initial guess is chosen. But even then the iterates of a gradient based optimizer may be forced to slowly wander along the elongated and flat valley of d^1 before they reach the unique minimizer. In comparison, the level sets of the alternative cost function (8) show a considerably smaller ratio between the major and minor axes of the ellipses, see Figure 8 B, such that the iterates may faster approach the minimizer. The condition number of the Hessian matrix $H^1(q^*)$ for d^1 is given by $\kappa_1(q^*) = \lambda_{\max}(q^*)/\lambda_{\min}(q^*) = 1911.3$, while for the Hessian $H^2(q^*)$ for d^2 a reduction by 25% is achieved. The observed algorithmic improvement is in agreement with gradient based optimization theory [38] according to which the rate of convergence improves if the condition number of the Hessian matrix, also reflected in the contour plots of the level sets, decreases.

Similar conclusions can be also be drawn from the $p53$ example in which we put focus on the practicability of our approach in case of partial state observations. The parameters in this example have different units and varying scales. As a consequence we build the Hessians of the cost functions d^1 and d^2

$$H^k_{ij}(q) = \frac{\partial d^k(q)}{\partial \log(q_i) \partial \log(q_j)}, \quad i,j \in \{1,2\}, \quad k = 1,2$$

by differentiation with respect to $\log(q)$ in order to take relative changes in parameter values into account [27]. A relative comparison of the two condition numbers $\kappa_1(q^*)$ and $\kappa_2(q^*)$ yields

$$\frac{\kappa_2(q^*) - \kappa_1(q^*)}{\kappa_1(q^*)} = -0.85$$

which corresponds to a reduction of the eigenvalue ratio $\lambda_{\max}(q^*)/\lambda_{\min}(q^*)$ by 85% when switching from d^1 to d^2. This improvement is even more pronounced if the derivatives in the Hessian are taken with respect to q and is also clearly reflected in 6.

Based on accepted mathematical descriptions of stochastical reaction networks in well-mixed conditions we have introduced the concept of moment fitting for the nonlinear inverse problem of parameter inference for the scenario of repeated measurements. In numerical tests for simple reference examples we observed that if the common comparison of sample mean data with the parameter-dependent mean expression derived from the model is augmented by consideration of higher moments such as covariance, improvements both with respect to parameter estimate accuracy and algorithmic efficiency are achieved. For staying focused we did not add measurement errors to our data obtained by the Gillespie algorithm. One option for handling of that situation would be to expand the moment cost function by a stabilizing penalty term and to consider a regularized functional [7]

$$d_{reg}(q) = d(q) + \lambda \cdot \text{penalty}(q).$$

As the eigenvalues of the Hessian of a cost function d also reflect the degree of ill-conditioning of the parameter inference problem, our examples show that already the consideration of higher moments may have a benefitial effect with respect to the amplification of measurement errors. Other possible expansions of the moment fitting approach presented in this paper include the consideration of more elaborate state observation operators, unknown initial conditions or reaction rate parameters that themselves are treated as stochastic quantities.

However, the most desirable extension of the present study would be an error analysis for the estimated parameters that in practical applications guides the choice of the number N of repeated observations, the choice of the underlying modelling concept, the choice of the moment order $\bar{k}$ to be used and/or the choice of the weights w_k in the cost functions. Currently, the parameters are estimated based on a comparison of the sample moments $\hat{\mu}$ with the mathematical description $\mu(\cdot;q)$ defined by the moment ODE system (21). But both $\hat{\mu}$ and $\mu(\cdot;q)$ only represent approximations of the actual moments $m(\cdot;q)$ either due to the finite sampling number N or due to surrogates for the chemical master equation along with the truncation of the moment expansions at some finite $\bar{k}$. Considering monostable chemical reaction networks, error estimates of $\mu(\cdot;q)$ based on the chemical Langevin equation (12) are derived in [39] for $\bar{k}=2$ in terms of the reaction volume raised to negative exponents. It is also argued in [39] that the mean and variance predictions of the chemical Langevin equation (12) are more accurate than those obtained from the linear noise approximation (22). Furthermore, [36] considers the derivative matching technique for moment closure and provides error estimates for time derivatives of $\mu(\cdot;q)$ at $t=0$ in terms of the norm of the initial state vector $x(0)$ raised to the power of $-\bar{k}$. A Taylor series argument then guarantees that the trajectories of $\mu(\cdot;q)$ and $m(\cdot;q)$ will remain close at least locally in time. Finally, [32] considers the closure method of setting all moments above $\bar{k}$ equal to zero and shows that the error of the mean $\mu^1(\cdot;q)$ then is of order $t_f^{\bar{k}}$ while the error of the second central moment $\mu^{c,2}(\cdot;q)$ is of order $t_f^{\bar{k}-1}$ as the length t^f of the time interval approaches zero. On the other hand, the central limit theorem [40] may be utilized to estimate the standard error of the sample mean $\hat{\mu}^1$ in terms of the sample size N, while error estimates for the sample covariance matrix $\hat{\mu}^{c,2}$ may be obtained from [41], [42]. It is the subject of future research to study how these results can be integrated into error bounds for the parameter estimates $\hat{q}$ of the moment fitting approach that may be exploited in algorithmic realizations.

Materials and Methods

Modelling Strategies

In well-mixed conditions a network of l coupled chemical reactions $R_1,...,R_l$ involving n chemical species $X_1,...,X_n$ is characterized by the formalism [2]

$$R_j:\ \mathrm{s_{-j1}X_1}+\cdots+\mathrm{s_{-jn}X_n}\xrightarrow{q_j}\mathrm{s_{+j1}X_1}+\cdots+\mathrm{s_{+jn}X_n},$$

where the integers s_{-ji} and s_{+ji}, $i=1,\ldots,n$ denote the numbers of molecules consumed and produced in a single step of reaction R_j. If $x\in\mathbb{N}_0^n$ represents the vector of species molecule numbers and $s_{ij}=s_{+ji}-s_{-ji}$ denote the components of the stoichiometric matrix $S\in\mathbb{Z}^{n\times l}$, then the state vector is updated according to $x\to x+S_{:,j}$ whenever reaction R_j fires. Each reaction R_j is associated with a rate law (or hazard function) $h_j(x,q_j)$ and a stochastic rate constant q_j.

Let $\pi(x,t)$ denote the probability of being in state x at time t given the initial condition $x(0)$. Then, the time evolution of $\pi(x,t)$ is described by the Kolmogorov differential equation (or chemical Master equation)

$$\frac{\partial\pi}{\partial t}(x,t)=\sum_{j=1}^{l}\left\{\pi(x-S_{:,j},t)h_j(x-S_{:,j},q_j)-\pi(x,t)h_j(x,q_j)\right\}. \tag{11}$$

With $h(x,q)=(h_1(x,q_1),\ldots,h_l(x,q_l))^T\in\mathbb{R}^l$ the diffusion approximation to the true process is based on the chemical Langevin equation [5]

$$d\chi=Sh(\chi,q)dt+\sqrt{S\mathrm{diag}(h(\chi,q))S'}dW, \tag{12}$$

where $d\chi$ is the change in state $\chi(t)\in\mathbb{R}^n$ in an infinitely small time interval dt and dW is the increment of a n-dimensional Wiener process. In the stochastic differential equation (12) the stochastic perturbations

are modelled by a state and rate parameter dependent Gaussian noise and the associated probability density function $p(\chi,t)$ is described by the Fokker-Planck equation

$$\frac{\partial p}{\partial t}(\chi,t) = -\sum_{i=1}^{n} \frac{\partial}{\partial \chi_i} \left\{ [Sh(\chi,q)]_i \, p(\chi,t) \right\} + \frac{1}{2} \sum_{i,k=1}^{n} \frac{\partial^2}{\partial \chi_i \partial \chi_k} \left\{ \left[S \mathrm{diag}(h(\chi,q)) S^T \right]_{ik} p(\chi,t) \right\}. \tag{13}$$

An alternative approximative description of the stochastic process is given by the linear noise approximation [6] which is derived from a Taylor expansion of (11) in powers of $1/\sqrt{\Omega}$ where Ω denotes the volume of the reactive system. This leads to a decomposition of the molecule concentration vector $c(t) = x(t)/\Omega \in \mathbb{R}^n$ according to

$$c(t) = \varphi(t) + \frac{1}{\sqrt{\Omega}} \xi(t) \tag{14}$$

into a deterministic part φ that solves the macroscopic rate equation

$$\frac{\partial \varphi}{\partial t}(t) = Sh(\varphi, q) \tag{15}$$

and a stochastic process ξ described by a linear diffusion equation

$$d\xi = S \frac{\partial h}{\partial \varphi}(\varphi, q) \xi dt + S \sqrt{\mathrm{diag}(h(\varphi,q))} dW$$

with the increment dW of a l-dimensional Wiener process.

Finally, the deterministic modelling approach [1], [43] ignores (if the justifying assumptions are satisfied) random fluctuations due to the stochasticity of the reactions and describes the time course of the species concentration vector $c(t)$ by the set of ordinary differential equations

$$\frac{\partial c}{\partial t}(t) = Sh(c, q) \tag{16}$$

which corresponds to (15) with the setting $\varphi(t) = c(t)$.

Moments of the Random State Variable

Depending on the chosen modelling approach the state variable of a stochastic biochemical reaction network is described as a discrete or a continuous random quantity. In the discrete case associated to the Kolmogorov differential equation (11) the first order moments [40] of the n-dimensional state variable $x(t)$ are given by

$$m^1_{r_1,\ldots,r_n}[x(t)] = E\left[\prod_{i=1}^{n} x_i^{r_i}(t) \right] = \sum_{\tilde{x} \in \mathcal{X}} \left\{ \prod_{i=1}^{n} \tilde{x}_i^{r_i}(t) \right\} \pi(\tilde{x}, t), \text{ with } r_1 + \cdots + r_n = 1,$$

where E is the expectation operator and $\mathcal{X}$ denotes the countable state space. Using this formalism the n-dimensional mean vector $m^1[x(t)]$ of $x(t)$ is described by

$$m^1[x(t)] = \begin{pmatrix} m^1_{1,0,\ldots,0}(x,t) \\ \vdots \\ m^1_{0,\ldots,0,1}(x,t) \end{pmatrix} = \begin{pmatrix} E[x_1(t)] \\ \vdots \\ E[x_n(t)] \end{pmatrix} = \begin{pmatrix} \sum_{\tilde{x} \in \mathcal{X}} \tilde{x}_1(t) \pi(\tilde{x}, t) \\ \vdots \\ \sum_{\tilde{x} \in \mathcal{X}} \tilde{x}_n(t) \pi(\tilde{x}, t) \end{pmatrix}.$$

In general, the k-th order moments [40] are given by

$$m^k_{r_1,\ldots,r_n}(x,t) = E[\prod_{i=1}^{n} x_i^{r_i}(t)] = \sum_{\tilde{x} \in \mathcal{X}} \left\{ \prod_{i=1}^{n} \tilde{x}_i^{r_i}(t) \right\} \pi(\tilde{x}, t), \text{ with } r_1 + \cdots + r_n = k \tag{17}$$

and the k-th central moments [40] are given by

$$m^{c,k}_{r_1,\ldots,r_n}(x,t) = E[\prod_{i=1}^{n}(x_i(t) - E[x_i(t)])^{r_i}] = \sum_{\tilde{x}\in\mathcal{X}}\left\{\prod_{i=1}^{n}(\tilde{x}_i(t) - E[x_i(t)])^{r_i}\right\}\pi(\tilde{x},t)$$

with $r_1 + \cdots + r_n = k$. For instance, the covariance matrix $m^{c,2}[x(t)] \in \mathbb{R}^{n\times n}$ of $x(t)$ is described by the 2-nd order central moments according to

$$m^{c,2}[x(t)] = \begin{pmatrix} m^{c,2}_{2,0,\ldots,0}(x,t) & m^{c,2}_{1,1,\ldots,0}(x,t) & \ldots & m^{c,2}_{1,0,\ldots,1}(x,t) \\ m^{c,2}_{1,1,\ldots,0}(x,t) & m^{c,2}_{0,2,\ldots,0}(x,t) & \ldots & m^{c,2}_{0,1,\ldots,1}(x,t) \\ \vdots & & \ddots & \\ m^{c,2}_{1,0,\ldots,1}(x,t) & m^{c,2}_{0,1,\ldots,1}(x,t) & \ldots & m^{c,2}_{0,0,\ldots,2}(x,t) \end{pmatrix}.$$

In case the state space of the biological system is modelled as a continuous random variable χ its moments are described the same way after switching from the sumation over a probability mass function $\pi(x,t)$ to an integration over a probability density function $p(\chi,t)$ as, e.g., defined by the Fokker Planck equation (13). For instance, the k-th order moments then are given by

$$m^{k}_{r_1,\ldots,r_n}(\chi,t) = E[\prod_{i=1}^{n}\chi_i^{r_i}(t)] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left\{\prod_{i=1}^{n}\tilde{\chi}_i^{r_i}(t)\right\}p(\tilde{\chi},t)d\tilde{\chi}_1\ldots d\tilde{\chi}_n, \text{ with } r_1+\cdots+r_n = k. \quad (18)$$

Differential Equations for the Moments

Based on the differential equations (11) and (13) for the probability mass and density functions, differential equations that describe the time evolution of the k-th order moments (17) and (18) or their centered counterparts can be derived. For instance, the time evolution of the mean vector $m^1[x(t)] \in \mathbb{R}^n$ of the discrete state $x(t) \in \mathbb{N}_0^n$ is given by

$$\frac{\partial}{\partial t}m^1[x(t)] = \sum_{\tilde{x}\in\mathcal{X}}\tilde{x}\frac{\partial \pi}{\partial t}(\tilde{x},t) = SE[h(x(t),q)], \quad (19)$$

see [2], [32]. Note that (19) corresponds to the deterministic rate equation (16) only in case of a propensity function $h(x,q)$ that is linear in x as only then

$$E[h(x(t),q)] = h(m^1[x(t)],q)$$

holds. Another example is the time evolution of the k-th order moments of a continuous state variable $\chi(t) \in \mathbb{R}$ which based on the Fokker Planck equation (13) with $n = 1$ is given by

$$\begin{aligned}\frac{\partial}{\partial t}m^k[\chi(t)] &= \frac{\partial}{\partial t}\int_{-\infty}^{\infty}\tilde{\chi}^k p(\tilde{\chi},t)\,d\tilde{\chi} \\ &= k\int_{-\infty}^{\infty}\tilde{\chi}^{k-1}Sh(\tilde{\chi},q)p(\tilde{\chi},t)\,d\tilde{\chi} + \frac{k(k-1)}{2}\int_{-\infty}^{\infty}\tilde{\chi}^{k-2}S\mathrm{diag}(h(\tilde{\chi},q))S^T p(\tilde{\chi},t)\,d\tilde{\chi} \\ &= kE[\chi(t)^{k-1}Sh(\chi(t),q)] + \frac{k(k-1)}{2}E[\chi(t)^{k-2}S\mathrm{diag}(h(\chi(t),q))S^T]. \quad (20)\end{aligned}$$

The simple example $l = n = 1$ with $h(\chi,q) = q\chi^2$ in (20) yields

$$kE[\chi(t)^{k-1}Sh(\chi(t),q)] = kqE[\chi(t)^{k+1}] = kqm^{k+1}[\chi(t)]$$

and implies the dependency of the ordinary differential equation for $m^k[\chi(t)]$ on the higher moment $m^{k+1}[\chi(t)]$. This type of dependency is typical for the moments $m^k_{r_1,\ldots,r_n}[\chi(t)]$ or $m^k_{r_1,\ldots,r_n}[x(t)]$ whenever $h(x,q)$ is a nonlinear function of x and may render the *exact* differential equations for their time evolution impossible to solve analytically or numerically. The problem may be overcome by the technique of moment closure [32], [33], [34], [35], [36] which sets moments or central moments above a certain order $\bar{k}$ of interest equal to zero or alternatively replaces them by expressions depending only on moments up to order $\bar{k}$. As a result one obtains (manually or supported by symbolic computation tools [34], [44]) a self-contained (or closed) set of coupled ordinary differential equations

$$\mu_t(t) = F(\mu(t), q), \tag{21}$$

in which $\mu(t) = \left(\mu^1(t), \mu^2(t), \ldots, \mu^{\bar{k}}(t)\right)$ or $\mu(t) = \left(\mu^1(t), \mu^{c,2}(t) \ldots, \mu^{c,\bar{k}}(t)\right)$ *approximatively* describes the time evolution of the true moments $m(t) = \left(m^1_{r_1,\ldots,r_n}[x(t)], m^2_{r_1,\ldots,r_n}[x(t)], \ldots, m^{\bar{k}}_{r_1,\ldots,r_n}[x(t)]\right)$ or $m(t) = \left(m^1_{r_1,\ldots,r_n}[x(t)], m^{c,2}_{r_1,\ldots,r_n}[x(t)], \ldots, m^{c,\bar{k}}_{r_1,\ldots,r_n}[x(t)]\right)$. In our notation we emphasize the dependency of μ on the rate parameters by writing the solution of (21) (to be supplemented by appropriate initial conditions) as $\mu(t;q)$. The final form of (21) depends on the underlying modelling approach, the choice of $\bar{k}$ and the choice of the closure technique.

As an alternative to the moment closure approach, a closed form (21) for the approximative description of the moment time course for $\bar{k} = 2$ can be obtained from the linear noise approximation [6], [31], [19] which with $\mu(t;q) = \left(\mu^1(t;q), \mu^{c,2}(t;q)\right)$ as approximation of $m\left(m^1[x(t)], m^{c,2}[x(t)]\right)$ yields the nonlinear ODE system

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1(t) &= Sh(\mu^1(t), q), \\
\frac{\partial}{\partial t}\mu^{c,2}(t) &= Sh_x(\mu^1(t), q)\mu^{c,2}(t) + \mu^{c,2}(t)h_x(\mu^1(t), q)^T S^T \\
&\quad + S\mathrm{diag}(h(\mu^1(t), q))S^T,
\end{aligned} \tag{22}$$

where $h_x(x,q) \in \mathbb{R}^{l\times n}$ denotes the Jacobian matrix of $h(x,q)$ with respect to x.

Test Models

Linear Birth and Death Process

A classic and illustrative reaction system widely studied in the literature is the linear birth and death process [45], [40], [2] for the species X with molecule number $x \in \mathbb{N}_0$. The birth and death reactions are given by

$$\begin{aligned}
\mathrm{R}_1 &: \ \mathrm{X} \xrightarrow{q_1} 2\,\mathrm{X} \\
\mathrm{R}_2 &: \ \mathrm{X} \xrightarrow{q_2} \emptyset
\end{aligned}$$

with the associated stoichiometric matrix S and the rate functions h

$$S = \begin{pmatrix} 1 & -1 \end{pmatrix}, \quad h(x,q) = \begin{pmatrix} h_1(x,q_1) \\ h_2(x,q_2) \end{pmatrix} = \begin{pmatrix} q_1 x \\ q_2 x \end{pmatrix}.$$

In all simulations of the discrete stochastic dynamics of the model we chose the rate parameters $q^* = (q_1^*, q_2^*)^T = (3,4)^T$ and the initial condition $x(0) = 50$.

With respect to the parameter inference problem we suppose that the state variable

$$x = \{x(t) \mid t \in [0, t_f]\} \in \mathcal{F}([0,t_f], \mathbb{N}_0)$$

can only be partially observed at the n_t discrete times

$$t_{j+1} = t_j + (j+1)\frac{t_f}{n_t},\ j = 0, \ldots, n_t - 1 \tag{23}$$

with $t_0 = 0$ of an (without loss of generality) equidistant grid over the time interval $[0, t_f]$. This then gives rise to a state observation operator

$$O : \mathcal{F}([0, t_f], \mathbb{N}_0) \to \mathbb{R}^{1 \times n_t}, x \to y = (x(t_1), \ldots, x(t_{n_t})), \tag{24}$$

such that the partial state observation can be compactly described as $y = Ox$. Here, $\mathcal{F}([0, t_f], \mathbb{N}_0)$ denotes the set of all functions from $[0, t_f]$ to $\mathbb{N}_0$. In the example we chose the setting $t_f = 8.5$ and $n_t = 85$. In particular, we do not consider the times and types of the reactions that are fired during the realisation of the stochastic process as amenable to our observations.

Dimerisation Kinetics

A simple reaction system featuring a nonlinear rate function is the dimerisation process [2], [34]. For the species P and P_2 with molecule numbers x_1 and x_2 we consider

$$\begin{aligned} \mathrm{R}_1 &: \ 2\,\mathrm{P} \xrightarrow{q_1} \mathrm{P}_2 \\ \mathrm{R}_2 &: \ \mathrm{P}_2 \xrightarrow{q_2} 2\,\mathrm{P}. \end{aligned}$$

The conservation of the total number x_0 of molecules

$$x_1 + 2x_2 = x_0$$

allows to formulate the stoichiometric matrix S and the rate functions h in terms of P only, i.e., with $X = P$ and $x = x_1 \in \mathbb{N}_0$ we obtain

$$S = \begin{pmatrix} -2 & 2 \end{pmatrix}, \quad h = \begin{pmatrix} h_1(x, q_1) \\ h_2(x, q_2) \end{pmatrix} = \begin{pmatrix} q_1 \frac{x(x-1)}{2} \\ q_2 \frac{x_0 - x}{2} \end{pmatrix}$$

for a then one-dimensional state space. In all simulations of the discrete stochastic dynamics of the model we chose the rate parameters $q^* = (q_1^*, q_2^*)^T = (0.005, 0.03)^T$, the total molecule number $x_0 = 301$ and the initial condition $x(0) = x_0$. With respect to the parameter inference problem we chose the same observation operator O as in (24), now with the setting $t_f = 6.7$ and $n_t = 67$.

$p53$ Signalling System

For testing the practicability of our approach in case of partial state observations we have chosen a model for the $p53$ signalling system which features a feedback loop between the tumor suppressor $p53$, the oncogene $Mdm2$ and its precursor pre_Mdm2. The model was introduced in [46] and also studied in [31]. With $X = (p53, pre\_Mdm2, Mdm2)$ and the associated vector $x = (x_1, x_2, x_3)^T \in \mathbb{N}_0^3$ of molecule copy numbers, its stoichiometric matrix S and rate functions h are given by

$$S = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \quad h(x, q) = \begin{pmatrix} h_1(x, q_1) \\ h_2(x, q_2) \\ h_3(x, q_3) \\ h_4(x, q_4) \\ h_5(x, q_5) \\ h_6(x, q_6) \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 x_1 \\ q_{3a} \frac{x_1 x_3}{x_1 + q_{3b}} \\ q_4 x_1 \\ q_5 x_2 \\ q_6 x_3 \end{pmatrix}.$$

In all simulations of the discrete stochastic dynamics of the model we chose the rate parameters $q^* = [q_1^*, q_2^*, q_{3a}^*, q_{3b}^*, q_4^*, q_5^*, q_6^*]^T = [90, 0.002, 1.7, 1, 1.1, 0.8, 2]^T$ and the initial conditions $x(0) = [10, 10, 80]^T$. With respect to the parameter inference problem we chose the time discretization as in the previous examples and supposed that either the component x_1 or both the components x_1 and x_2 are amenable to observations. The corresponding observation operators are

$$O^1 : \mathcal{F}([0, t_f], \mathbb{N}_0^3) \to \mathbb{R}^{1\times n_t}, x \to y = \begin{pmatrix} x_1(t_1) & \dots & x_1(t_{n_t}) \end{pmatrix}$$

and

$$O^2 : \mathcal{F}([0, t_f], \mathbb{N}_0^3) \to \mathbb{R}^{2\times n_t}, x \to y = \begin{pmatrix} x_1(t_1) & \dots & x_1(t_{n_t}) \\ x_2(t_1) & \dots & x_2(t_{n_t}) \end{pmatrix}.$$

The time observation parameters were chosen as $t_f = 11.1$ and $n_t = 111$.

Data Generation and Sample Moments

All molecular copy numbers used in our tests have been generated by MATLAB [37] simulations of the discrete stochastic model dynamics using the Gillespie algorithm [4]. In general, a single realization of the process allows to mimic a single observation of the system giving rise to an experimental concentration data matrix $\hat{y} \in \mathbb{R}^{d\times n_t}$. Here, n_t is the time discretization parameter and d is the number of observable components of the state variable as defined by the state observation operator

$$O : \mathcal{F}([0, t_f], \mathbb{N}_0^n) \to \mathbb{R}^{d\times n_t}, x \to y.$$

A N-time repitition of the experimental observation of the system (or the computational realization of the stochastic process) then yields the sequence

$$\hat{y}^1, \dots, \hat{y}^N \tag{25}$$

of data matrices. For each discrete time point t_j the data matrix $\hat{y}^j$ allows to calculate sample moments [47], [48]. For instance, the sample mean of the observables at time t_j is given by

$$\hat{\mu}_j^{1,o} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{:,j}^i \in \mathbb{R}^d,$$

while the empirical covariance matrix of the observables at time t_j is given by

$$\hat{\mu}_j^{c,2,o} = \frac{1}{N-1}\sum_{i=1}^{N} \begin{pmatrix} \left(\hat{y}_{1,j}^i - \hat{\mu}_{1,j}^{1,o}\right)\left(\hat{y}_{1,j}^i - \hat{\mu}_{1,j}^{1,o}\right) & \dots & \left(\hat{y}_{1,j}^i - \hat{\mu}_{1,j}^{1,o}\right)\left(\hat{y}_{d,j}^i - \hat{\mu}_{d,j}^{1,o}\right) \\ \vdots & \ddots & \vdots \\ \left(\hat{y}_{d,j}^i - \hat{\mu}_{d,j}^{1,o}\right)\left(\hat{y}_{d,j}^i - \hat{\mu}_{d,j}^{1,o}\right) & \dots & \left(\hat{y}_{d,j}^i - \hat{\mu}_{d,j}^{1,o}\right)\left(\hat{y}_{d,j}^i - \hat{\mu}_{d,j}^{1,o}\right) \end{pmatrix} \in \mathbb{R}^{d\times d},$$

in which $\hat{y}_{s,j}^i$, $\hat{\mu}_{s,j}^{1,o}$ denote the s-th components of the vectors $\hat{y}_j^i$, $\hat{\mu}_j^{1,o}$. An alternative covariance matrix estimate that is more suitable if $N \ll d$ is not satisfied is given in [49]. In general, the data tensor $\hat{y} \in \mathbb{R}^{d\times n_t\times N}$ allows to compute the tupel

$$\hat{\mu}^o = \left(\hat{\mu}_1^o, \dots, \hat{\mu}_{n_t}^o\right)$$

of length n_t, where $\hat{\mu}_j^o$ denotes the sample moments of the observable state components up to the order $\bar{k}$ at time t_j. An example with $\bar{k} = 2$ is

$$\hat{\mu}^o = \left((\hat{\mu}_1^{1,o}, \hat{\mu}_1^{c,2,o}), \dots, (\hat{\mu}_{n_t}^{1,o}, \hat{\mu}_{n_t}^{c,2,o})\right)$$

with sample mean vector $\hat{\mu}_j^{1,o}$ and sample covariance matrix $\hat{\mu}_j^{c,2,o}$ at time t_j.

Cost Functions and Adjoint Method

If we split the state variable $x \in \mathbb{N}_0^n$ into the observable part $x^o \in \mathbb{N}_0^d$ and the unobservable part $x^u \in \mathbb{N}_0^{n-d}$ according to $x = [x^o; x^u]$, this separation carries over to the k-th order moments of x approximatively described by the ODE system (21), i.e., $\mu(\cdot; q) = [\mu^o(\cdot; q); \mu^u(\cdot; q)]$. This is to be understood in the sense that there exists a (linear) splitting operator $\mathcal{N}$ with

$$\mu^o(\cdot; q) = \mathcal{N}\mu(\cdot; q).$$

Then, in order to compare the parameter dependent solution component $\mu^o(\cdot; q)$ of the ODE system (21) to the available sample moment tuple $\hat{\mu}^o$ various distance measures

$$d(\hat{\mu}^o, \mathcal{D}\mu^o(q)) \tag{26}$$

may be utilized where the time discretization operator $\mathcal{D}$ simply evaluates the time-dependent $\mu^o(\cdot; q)$ function at the discrete times t_j of (23), i.e.,

$$\mathcal{D} : \mu^o(\cdot; q) \to (\mu^o(t_1; q), \ldots, \mu^o(t_{n_t}; q)).$$

Examples for $\bar{k} = 2$ with $\mu(\cdot; q) = (\mu^1(\cdot; q), \mu^{c,2}(\cdot; q))$ include

$$\begin{aligned}
d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) &= w_1 \sum_{j=1}^{n_t} \|\hat{\mu}_j^{1,o} - \mu^{1,o}(t_j; q)\|_2^2 + w_2 \sum_{j=1}^{n_t} \|\hat{\mu}_j^{c,2,o} - \mu^{c,2,o}(t_j; q)\|_2^2, \\
d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) &= w_1 \sum_{j=1}^{n_t} (\hat{\mu}_j^{1,o} - \mu^{1,o}(t_j; q))^T \mu^{c,2,o}(t_j; q)^{-1} (\hat{\mu}_j^{1,o} - \mu^{1,o}(t_j; q)) \\
&\quad + w_2 \sum_{j=1}^{n_t} \|\hat{\mu}_j^{c,2,o} - \mu^{c,2,o}(t_j; q)\|_2^2, \\
d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) &= w_1 \sum_{j=1}^{n_t} \sum_{s=1}^{d} \frac{1}{\mu_{ss}^{c,2,o}(t_j; q)} (\hat{\mu}_{s,j}^{1,o} - \mu_s^{1,o}(t_j; q)])^2 \\
&\quad + w_2 \sum_{j=1}^{n_t} \|\hat{\mu}_j^{c,2,o} - \mu^{c,2,o}(t_j; q)\|_2^2,
\end{aligned}$$

where, in general, w_k denotes the weight associated to the k-th order moment comparison. As shown in the results section the choice $w_k \neq 0$ for $k > 1$ can make a decisive difference in the parameter inference problem for stochastic biological models. As lower order statistical moments in general are easier to approximate an ordering of the weights according to $w_k > w_{k+1}$ seems reasonable.

With respect to parameter inference the difference function (26) can be utilized in various manners. For instance, it can be minimized by deterministic or stochastic optimization routines, it can be incorporated as a cost function in approximate Bayesian methods or used in building Kalman filters or Luenberger type observers. In the context of gradient based optimization, the gradient information can be efficiently provided by means of the so-called adjoint technique whenever d can be written as a parameter dependent $\langle \cdot, \cdot \rangle_q$ or parameter independent $\langle \cdot, \cdot \rangle$ inner product of the residual $r(q) = \hat{\mu}^o - \mathcal{D}\mathcal{N}\mu(q)$ with itself, see Supporting Information.

Acknowledgments

I have developed the moment fitting concept presented in this paper during my visiting fellowship at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK and kindly acknowledge the opportunities granted.

References

1. Chen WW, Niepel M, Sorger PK (2010) Classic and contemporary approaches to modeling biochemical reactions. Genes & Development 24: 1861-1875.

2. Wilkinson JD (2012) Stochastic modelling for systems biology. Chapman & Hall/CRC, 2nd edition.

3. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. Journal of Chemical Physics 124: 044104.

4. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry 81: 2340-2361.

5. Gillespie D (2000) The chemical langevin equation. Journal of Chemical Physics 113: 297-306.

6. Van Kampen NG (2007) Stochastic Processes in Physics and Chemistry. Amsterdam: North Holland.

7. Engl HW, Hanke M, Neubauer A (1996) Regularization of Inverse Problems. Dordrecht: Kluwer.

8. Kaipio J, Somersalo E (2004) Statistical and Computational Inverse Problems. Dordrecht: Springer.

9. Isakov V (2006) Inverse Problems for Partial Differential Equations, volume 127 of *Applied Mathematical Sciences*. Springer.

10. Bishwal JPN (2008) Parameter Estimation in Stochastic Differential Equations. Springer.

11. Pepperkok R, Ellenberg J (2006) High-throughput fluorescence microscopy for systems biology. Nature Reviews Molecular Cell Biology 7: 690-696.

12. Shen H, Nelson G, Nelson DE, Kennedy S, Spiller DG, et al. (2006) Automated tracking of gene expression in individual cells and cell compartments. Journal of the Royal Society Interface 3: 787-794.

13. Moles CG, Pedro Mendes P, Banga JR (2003) Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Research 13: 2467-2474.

14. Engl HW, Flamm C, Kügler P, Lu J, Müller S, et al. (2009) Inverse problems in systems biology. Inverse Problems 25: 123014.

15. Rodriguez-Fernandez M, Egea JA, Banga JR (2006) Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC Bioinformatics 7.

16. Golightly A, Wilkinson DJ (2008) Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics 61: 781-788.

17. Toni T, Welch D, Strelkowa N, A I, Stumpf MPH (2009) Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface 6: 187-202.

18. Lecca P, Palmisano A, Priami C (2009) Deducing chemical reaction rate constants and their regions of confidence from noisy measurements of time series of concentration. Computer Modelling and Simulation, 2009 UKSIM '09 11th International Conference on : 200-205.

19. Komorowski M, Finkenstädt B, Harper CV, Rand DA (2009) Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinformatics 10.

20. Poovathingal SK, Gunawan R (2010) Global parameter estimation methods for stochastic biochemical systems. BMC Bioinformatics 11.

21. Lillacci G, Khammash M (2010) Parameter estimation and model selection in computational biology. PLoS Computational Biology 6.

22. Sun X, Jin L, Xiong M (2008) Extended kalman filter for estimation of parameters in nonlinear state space models of biochemical networks. PLoS ONE 3.

23. (2008) Parameter estimation in kinetic reaction models using nonlinear observers is facilitated by model extensions, 17th IFAC World Congress.

24. Jaqaman K, Danuser G (2006) Linking data to models: data regression. Nature Reviews Molecular Cell Biology 7: 813-819.

25. Gillespie C (2008). Parameter estimation using moment closure methods. video lecture, available at http://www.videolectures.net.

26. Brown KS, Sethna JP (2003) Statistical mechanical approaches to models with many poorly known parameters. Physical Review E 68.

27. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, et al. (2007) Universally sloppy parameter sensitivities in systems biology models. PLoS Computational Biology 3.

28. Anguelova M, Cedersund G, Johansson M, Franzen CJ, Wennberg B (2007) Conservation laws and unidentifiability of rate expressions in biochemical models. IET Systems Biology 1: 230-237.

29. Audoly S, Bellu G, D'Angio L, Saccomani MP, Cobelli C (2001) Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering 48: 55-65.

30. Ashyraliyev M, Jaeger J, Blom JG (2008) Parameter estimation and determinability analysis applied to drosophila gap gene circuits. BMC Systems Biology 2.

31. Komorowski M, Costab MJ, Rand DA, Stumpf MPH (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. PNAS 108: 8645-8650.

32. Lee CH, Kyeong-Hun Kim KH, Kim P (2009) A moment closure method for stochastic reaction networks. Journal of Chemical Physics 130: 813-819.

33. Engblom S (2006) Computing the moments of high dimensional solutions of the master equation. Applied Mathematics and Computation 180: 498-515.

34. Gillespie CS (2009) Moment-closure approximations for mass-action models. IET Systems Biology 3: 52-58.

35. Milner P, Gillespie CS, Wilkinson DJ (2011) Moment closure approximations for stochastic kinetic models with rational rate laws. Mathematical Biosciences 231: 99-104.

36. Singh A, Hespanha JP (2011) Approximate moment dynamics for chemically reaction systems. IEEE Transactions on Automatic Control 56: 414-418.

37. Mathworks T (2010). Matlab version 7.10.0.499 (r2010a). http://www.mathworks.com.

38. Nocedal J, Wright SJ (2006) Numerical Optimization. Springer.

39. Grima R, Thomas P, Straube AV (2011) How accurate are the nonlinear chemical fokker-planck and chemical langevin equations? The Journal of Chemical Physics 135: 084103.

40. Ross SM (2010) Introduction to Probability Models. Academic Press.

41. Vershynin R (2010). How close is the sample covariance matrix to the actual covariance matrix? unpublished manuscript, available at http://arxiv.org/abs/1004.3484.

42. Cai TT, Zhang CH, Zhou HH (2010) Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics 38: 21182144.

43. Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. Nature Cell Biology 8: 1195-1203.

44. Matis TI, Guardiola IG (2010) Achieving moment closure through cumulant neglect. The Mathematica Journal 12.

45. Bartholomay AF (1958) On the linear birth and death processes of biology as markoff chains. Bulletin of Mathematical Biophysics 20: 97-118.

46. Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, et al. (2006) Oscillations and variability in the p53 system. Molecular Systems Biology 2.

47. Tabachnick BG, Fidell LS (2006) Using Multivariate Statistics. Allyn & Bacon.

48. Hair JF, Black WC, Babin BJ, Anderson RE (2009) Multivariate Data Analysis. Prentice Hall.

49. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4: 99-104.

Figure Legends

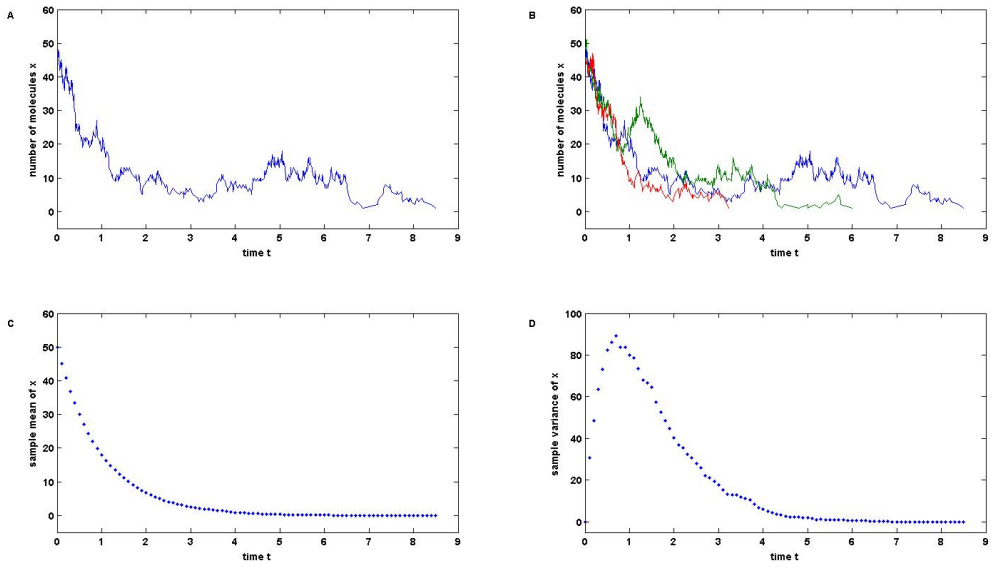

Figure 1. Simulation and data of the linear birth death process. (A) A single realization of the true stochastic process with $x_0 = 50$ and rate parameters $q_1^* = 3$, $q_2^* = 4$. (B) Three (out of $N = 1000$) realizations of the true stochastic process. The process is finished as soon as the case $x = 0$ occurs. (C) Sample mean $\hat{\mu}_j^1$ at the discrete observation times t_j. (D) Sample variance $\hat{\mu}_j^{c,2}$ at the discrete observation times t_j.

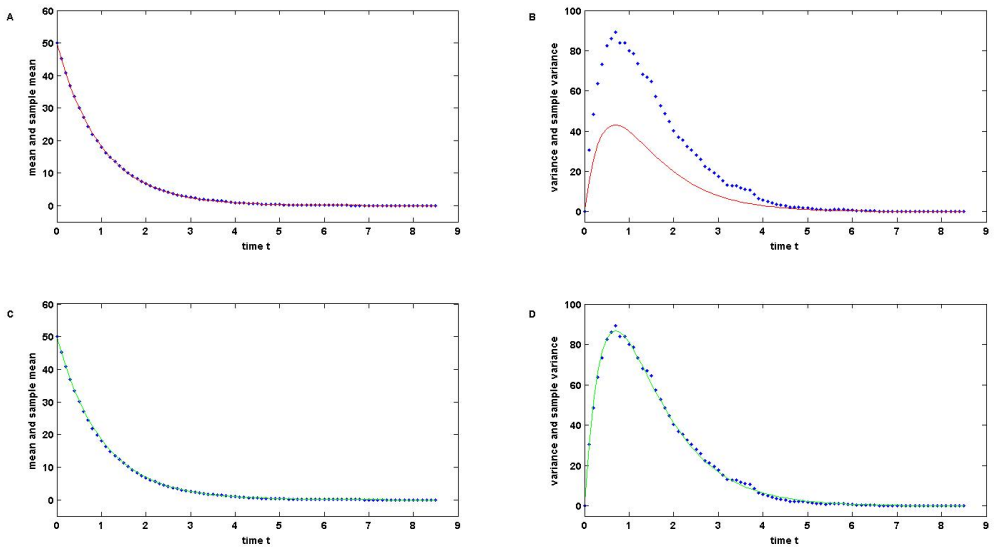

Figure 2. Moment fitting for the linear birth death process. (A) The minimization of the cost function (4) yields a perfect match between the mean data and the first moment expression $\mu^1(t;\tilde{q})$, but a parameter result $\tilde{q} = [1.2392, 2.2487]^T$ with large deviations from the true values $q^* = [3, 4]^T$. (B) As a consequence, the variance data cannot be explained by the second order moment expression $\mu^{c,2}(t;\tilde{q})$. (C) The alternative minimization of (5) once more yields a perfect match between the mean data and $\mu^1(t;\hat{q})$ but in addition the significantly improved parameter estimate $\hat{q} = [2.9492, 3.9415]^T$. (D) As the fitting of the central second order moment has been explicitly taken into account by (5), now also the variance data is reproduced by $\mu^{c,2}(t;\hat{q})$.

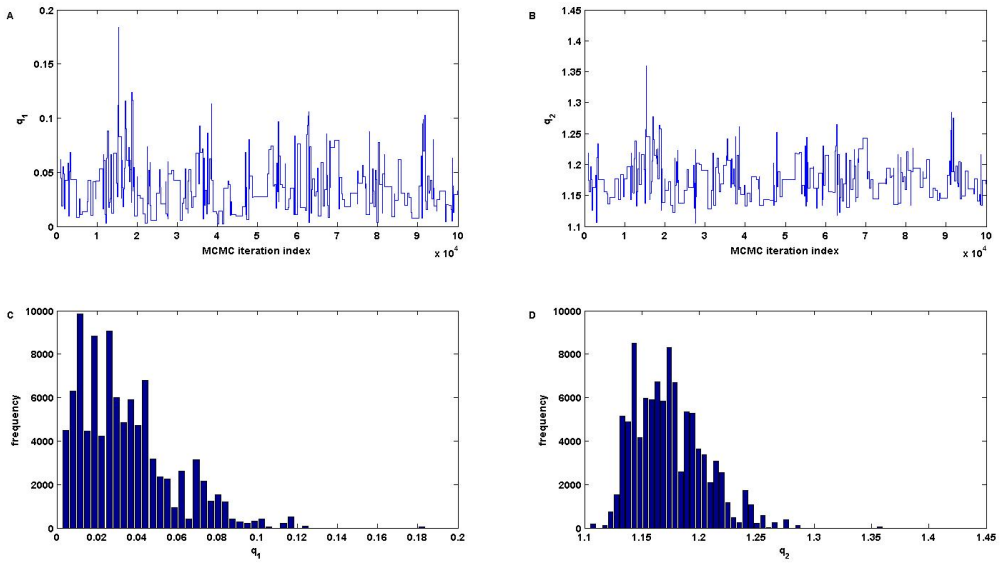

Figure 3. MCMC Metropolis random walk for the linear birth death process. Figures (A)-(F) show the output of a MCMC Metropolis random walk for the inference of the marginal parameter posterior density distributions using the likelihood function (6). The prior parameter distributions were chosen as $q_1 \sim \Gamma(3,1)$ and $q_2 \sim \Gamma(3,1)$, and the candidate parameter vector q^c at stage j was given by $q^c = q^{(j-1)} + s^{(j)}$ with random innovations $s^{(j)}$ drawn from $U(-0.5, 0.5)$. The iteration number of the sampler was set to 100000 and the first 1000 steps were discarded as burn-in and ignored in the monitoring plots (A)-(F). (A,B) Trace plots of the marginal posterior distributions for q_1 and q_2 with only small movement around the mean values $\tilde{q} = [0.0343, 1.1739]^T$, largely deviating from the true values $q^* = [3,4]^T$. (C,D) Frequency histograms with 50 bins corresponding to the trace plots of (A,B).

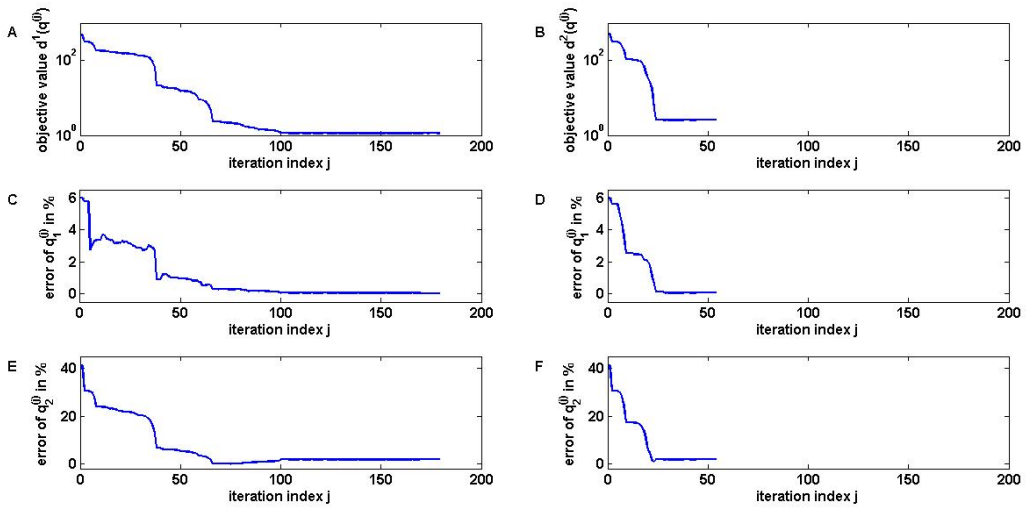

Figure 4. Iterative minimization for inference of the dimerisation process parameters. Iterative minimization of the cost functions (7) and (8) using the MATLAB trust region algorithm with default settings and and initial guess $q^{(0)} = [0.0047, 0.0177]^T$. The gradient information both for (7) and (8) was provided by means of the adjoint method in order to avoid error-prone finite differencing. (A) Plot of the value of the cost function (7) at the iterate $q^{(j)}$. The optimization algorithm terminates after 180 (outer) iteration steps and yields the minimizer $\tilde{q} = [0.005, 0.0294]^T$. (B) Using the cost function (8) instead of (7), the algorithm already terminates after 55 (outer) iteration steps and yields the minizer $\tilde{q} = [0.005, 0.0295]^T$. (C,E) Plots of the relative errors $100 \cdot \frac{|q_i^{(j)} - q_i^*|}{|q_i^*|}$, $i = 1, 2$ show that convergence to the true parameter vector q^* is obtained (up to a negligible error in q_2) if (7) is chosen as objective function. (D,F) Parameter convergence is also obtained if (8) is chosen instead of (7). However, parameter convergence is much faster in this case.

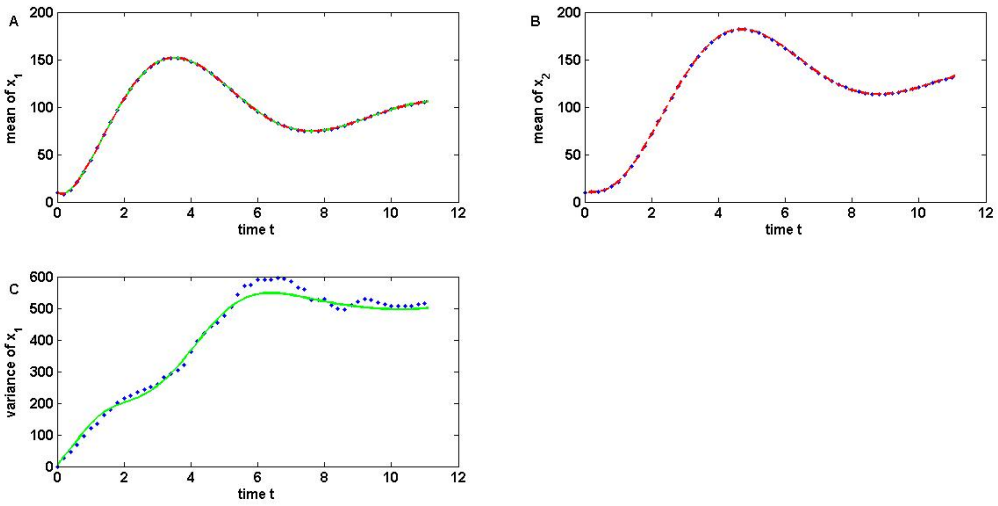

Figure 5. Moment fitting for the $p53$ model. (A, B) The minimization of the cost function (9) yields a perfect match between the sample mean data $\hat{\mu}_1^1$, $\hat{\mu}_2^1$ and the outputs $\hat{\mu}_1^1(\cdot;\tilde{q})$, $\hat{\mu}_2^1(\cdot;\tilde{q})$ (red curves) of the linear noise approximation for the d^1-minimizer $\tilde{q}$. (C) The plot shows the variance data $\hat{\mu}_{11}^{c,2}$ and the model output $\hat{\mu}_{11}^{c,2}(\cdot;\hat{q})$ (green curve) calculated with the minimizer $\hat{q}$ of the cost function (10). Though the approximation errors are more pronounced the consideration of $\hat{\mu}_{11}^{c,2}$ in the inference task yields a improved parameter estimate $\hat{q}$.

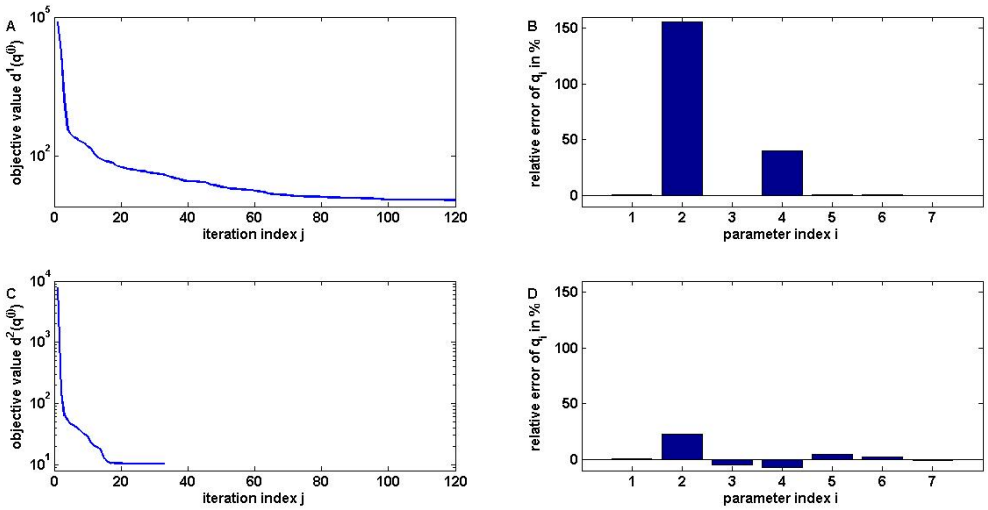

Figure 6. Iterative minimization for inference of the $p53$ model parameters. Iterative minimization of the cost functions (9) and (10) using the MATLAB trust region algorithm with default settings, the initial guess $q_i^{(0)} = (1 + 0.1 \cdot (-1)^i)q_i^*$, $i = 1, \ldots, 7$, and the adjoint method for providing the gradient information. (A) Plot of the value of the cost function (9) at the iterate $q^{(j)}$. The optimization algorithm terminates after 120 (outer) iteration steps and yields the minimizer $\tilde{q}$. (B) Plot of the relative errors of $\tilde{q}$ in % showing a huge deviation from the true parameter q^* in the second and fourth components. (C) Plot of the value of the cost function (10) at the iterate $q^{(j)}$. The optimization algorithm terminates after only 33 (outer) iteration steps and yields the minimizer $\hat{q}$. (D) The quality of the parameter estimate $\hat{q}$ has significantly improved in comparison to $\tilde{q}$.

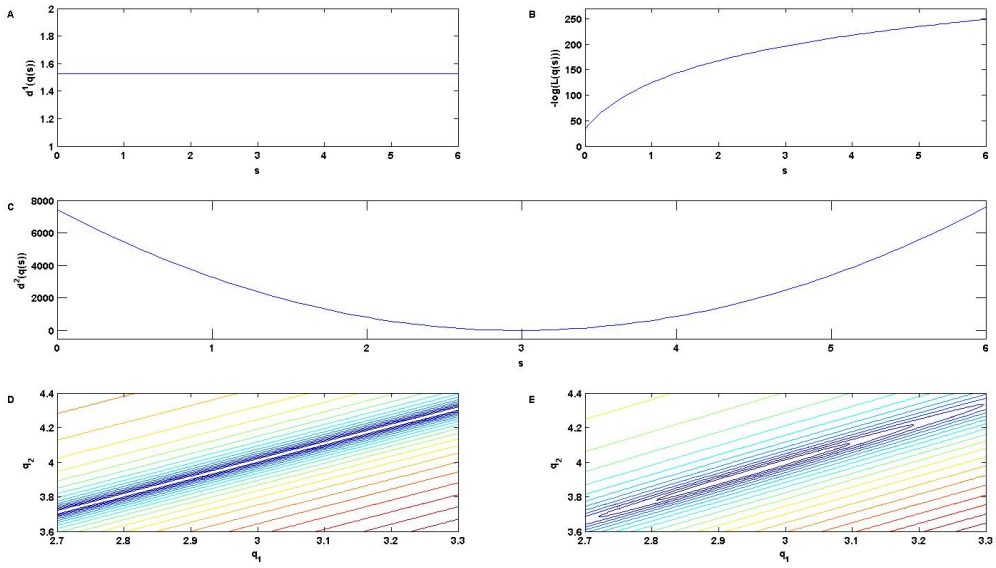

Figure 7. Distance functions for the linear birth death process. The problem of minimizing the least squares error between the sample mean and the analytic mean $\mu^1(q)$ has infinitely many global solutions $q(s) = [0,1]^T + s[1,1]^T$ with $s \in \mathbb{R}_0^+$. (A) Values of the cost function $d^1(q(s))$ from (4) for $s \in [0,6]$. (B) Values of the negative log-likelihood function $-\log(L(q(s)))$ for $s \in [0,6]$ with $L(q)$ as in (6). The likelihood $L(q(s))$ is maximal for $s = 0$ which explains the observation made in Figure 3. (C) Values of the cost function $d^2(q(s))$ from (5) for $s \in [0,6]$. The function also measures the distance between the second order sample and analytic moments and as a consequence admit a unique global minimum at $s \approx 3$ corresponding to the true parameter solution, i.e., $q(3) = [3,4]^T = q^*$. In this example, $s \approx 3$ and $d^1(q(s)) \neq 0$ are due to the finite sampling number N. (D) Level sets of the cost function d^1 indicate extreme parameter sloppiness and infinitely many parameter solutions. (E) Ellipsoidic level sets of the cost function d^2 in the neighbourhood of the true solution q^* indicate its unique identifiability.

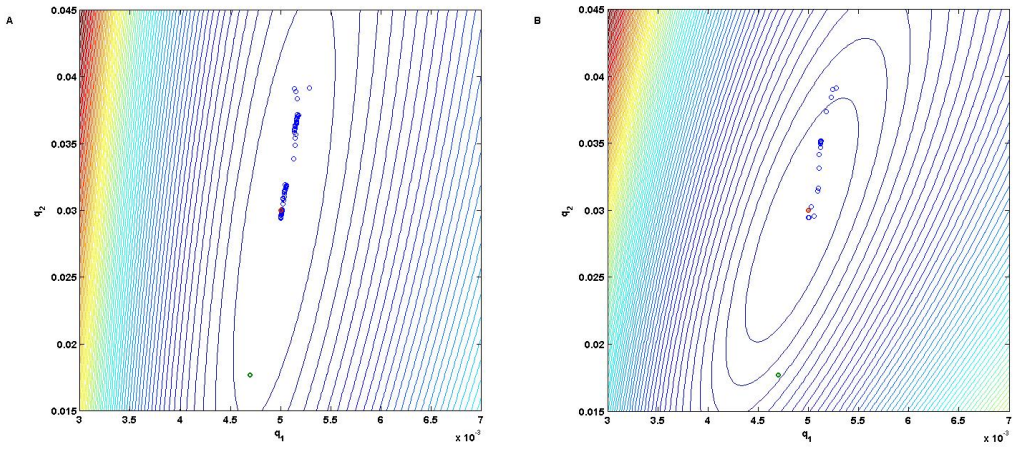

Figure 8. Level sets of the cost functions for the dimerisation process. (A) The level sets of the cost function d^1 reveal an elongated and flat valley in which the iterates of gradient based optimizers may only slowly converge towards the minimizer q^*. (B) The level sets of the cost function d^2 form ellipses with smaller ratio of major axis over minor axis and correspond to a more pronounced trough. As a consequence the iterates converge faster towards q^*.

Supporting Information

Moment Equations for Test Models

Linear Birth Death Process

Based on the Kolmogorov differential equation

$$\frac{\partial \pi}{\partial t}(x,t) = q_1(x-1)\pi(x-1,t) - (q_1+q_2)x\pi(x,t) + q_2(x+1)\pi(x+1,t),$$

the time evolution of the mean

$$\mu^1(t) = E[x(t)] = \sum_{\tilde{x}\in\mathcal{X}} \tilde{x}\pi(\tilde{x},t)$$

is described by

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1(t) &= \sum_{\tilde{x}\in\mathcal{X}} \tilde{x}\frac{\partial \pi}{\partial t}(\tilde{x},t) \\
&= \sum_{\tilde{x}\in\mathcal{X}} \big\{q_1\tilde{x}(\tilde{x}-1)\pi(\tilde{x}-1,t) - (q_1+q_2)\tilde{x}^2\pi(\tilde{x},t) \\
&\qquad +q_2\tilde{x}(\tilde{x}+1)\pi(\tilde{x}+1,t)\big\} \\
&= \sum_{\tilde{x}\in\mathcal{X}} \big\{q_1(\tilde{x}+1)\tilde{x}\pi(\tilde{x},t) - (q_1+q_2)\tilde{x}^2\pi(\tilde{x},t) \\
&\qquad +q_2(\tilde{x}-1)\tilde{x}\pi(\tilde{x},t)\big\} \\
&= (q_1-q_2)\mu^1(t). \qquad (1)
\end{aligned}$$

Furthermore, the time evolution of the second moment

$$\mu^2(t) = \sum_{\tilde{x}\in\mathcal{X}} \tilde{x}^2\pi(\tilde{x},t)$$

is given by

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^2(t) &= \sum_{\tilde{x}\in\mathcal{X}} \tilde{x}^2\frac{\partial}{\partial t}\pi(\tilde{x},t) \\
&= \sum_{\tilde{x}\in\mathcal{X}} \big\{q_1\tilde{x}^2(\tilde{x}-1)\pi(\tilde{x}-1,t) - (q_1+q_2)\tilde{x}^3\pi(\tilde{x},t) \\
&\qquad +q_2\tilde{x}^2(\tilde{x}+1)\pi(\tilde{x}+1,t)\big\} \\
&= \sum_{\tilde{x}\in\mathcal{X}} \big\{q_1(\tilde{x}+1)^2\tilde{x}\pi(\tilde{x},t) - (q_1+q_2)\tilde{x}^3\pi(\tilde{x},t) \\
&\qquad +q_2(\tilde{x}-1)^2\tilde{x}\pi(\tilde{x},t)\big\} \\
&= 2(q_1-q_2)\mu^2(t) + (q_1+q_2)\mu^1(t). \qquad (2)
\end{aligned}$$

Finally, the second central moment (the variance)

$$\mu^{c,2}(t) = \mu^2(t) - \mu^1(t)\cdot\mu^1(t)$$

satisfies

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^{c,2}(t) &= \frac{\partial}{\partial t}\mu^2(t) - 2\mu^1(t)\frac{\partial}{\partial t}\mu^1(t) \\
&= 2(q_1-q_2)\mu^2(t) + (q_1+q_2)\mu^1(t) - 2\mu^1(t)(q_1-q_2)\mu^1(t) \\
&= 2(q_1-q_2)\mu^{c,2}(t) + (q_1+q_2)\mu^1(t). \qquad (3)
\end{aligned}$$

If the diffusion approximation is chosen as modelling approach, the associated Fokker Planck equation reads as

$$\frac{\partial}{\partial t}p(\chi,t) = -\frac{\partial}{\partial \chi}\left\{(q_1-q_2)\chi p(\chi,t)\right\} + \frac{1}{2}\frac{\partial^2}{\partial \chi^2}\left\{(q_1+q_2)\chi p(\chi,t)\right\}.$$

Based on the Fokker Planck equation the time evolution of the first two moments

$$\int_{-\infty}^{\infty} \tilde{\chi}^k p(\tilde{\chi},t)\, d\tilde{\chi},\ k=1,2$$

is described by

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1[\chi(t)] &= E[(q_1-q_2)\chi(t)] = (q_1-q_2)\mu^1[\chi(t)],\\
\frac{\partial}{\partial t}\mu^2[\chi(t)] &= 2(q_1-q_2)\mu^2[\chi(t)] + (q_1+q_2)\mu^1[\chi(t)],
\end{aligned}$$

which is identical to (1), (2). Finally, also the linear noise approximation leads to the ODE system (1), (3) for the description of the time courses of μ^1 and $\mu^{c,2}$.

Dimerisation Process

The Fokker Planck equation of the diffusion modelling approach is given by

$$\frac{\partial}{\partial t}p(\chi,t) = -\frac{\partial}{\partial \chi}\left\{(-q_1\chi(\chi-1)+q_2(\chi_0-\chi))p(\chi,t)\right\} + \frac{1}{2}\frac{\partial^2}{\partial \chi^2}\left\{(2q_1\chi(\chi-1)+2q_2(\chi_0-\chi))p(\chi,t)\right\}$$

and implies

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1[\chi(t)] &= E[(-q_1\chi(\chi-1)+q_2(\chi_0-\chi))]\\
&= -q_1E[\chi^2(t)-\chi(t)] + q_2(\chi_0 - E[\chi(t)])\\
&= q_1\left[\mu^1[\chi(t)] - \mu^2[\chi(t)]\right] + q_2(\chi_0-\mu^1[\chi(t)])\\
&= q_1\left[\mu^1[\chi(t)] - \mu^1[\chi(t)]\cdot\mu^1[\chi(t)]\right] + q_2(\chi_0-\mu^1[\chi(t)]) - q_1\mu^{c,2}[\chi(t)]
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^2[\chi(t)] &= 2E[\chi(t)(-q_1\chi(t)(\chi(t)-1)+q_2(\chi_0-\chi(t)))] + E[2q_1\chi(t)(\chi(t)-1)+2q_2(\chi_0-\chi(t))]\\
&= 2E[-q_1\chi^3(t)+q_1\chi^2(t)+q_2\chi_0\chi(t)-q_2\chi^2(t)] + E[2q_1\chi^2(t)-2q_1\chi(t)+2q_2\chi_0-2q_2\chi(t)]\\
&= q_1\left[-2\mu^3[\chi(t)] + 4\mu^2[\chi(t)] - 2\mu^1[\chi(t)]\right] + q_2\left[2\chi_0(\mu^1[\chi(t)]+1) - 2\mu^2[\chi(t)] - 2\mu^1[\chi(t)]\right].
\end{aligned}$$

This is exactly the same set of moment ODEs if the derivation alternatively is based on the discrete state space and the Kolmogorov equation

$$\frac{\partial \pi}{\partial t}(x,t) = q_1\frac{(x+2)(x+1)}{2}\pi(x+2,t) - \left(q_1\frac{x(x-1)}{2} + q_2\frac{x_0-x}{2}\right)\pi(x,t) + q_2\frac{x_0-x+2}{2}\pi(x-2,t).$$

The corresponding equation for the second central moment (the variance) reads as

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^{c,2}(t) &= \frac{\partial}{\partial t}\mu^2(t) - 2\mu^1(t)\frac{\partial}{\partial t}\mu^1(t)\\
&= q_1\left[-2\mu^3[\chi(t)] + 4\mu^2[\chi(t)] - 2\mu^1[\chi(t)]\right] + q_2\left[2\chi_0(\mu^1[\chi(t)]+1) - 2\mu^2[\chi(t)] - 2\mu^1[\chi(t)]\right]\\
&\quad +2q_1\left[\mu^2[\chi(t)]\cdot\mu^1[\chi(t)] - \mu^1[\chi(t)]\cdot\mu^1[\chi(t)]\right] - 2q_2(\chi_0-\mu^1[\chi(t)])\cdot\mu^1[\chi(t)]\\
&= 2q_1\left[-\mu^3[\chi(t)] + 2\mu^{c,2}[\chi(t)] - \mu^1[\chi(t)] + \mu^1[\chi(t)]\cdot\mu^1[\chi(t)] + \mu^{c,2}[\chi(t)]\cdot\mu^1[\chi(t)]\right]\\
&\quad +2q_1\mu^1[\chi(t)]\cdot\mu^1[\chi(t)]\cdot\mu^1[\chi(t)] + 2q_2\left[\chi_0 - \mu^1[\chi(t)] - \mu^{c,2}[\chi(t)]\right].
\end{aligned}$$

In order to eliminate the dependency of $\mu^1[\chi(t)]$, $\mu^{c,2}[\chi(t)]$ on $\mu^3[\chi(t)]$ one option is the normal closure approximation [1], [2], [3], [4], [5], which sets

$$\mu^3[\chi(t)] = 3\mu^{c,2}[\chi(t)] \cdot \mu^1[\chi(t)] + \mu^1[\chi(t)] \cdot \mu^1[\chi(t)] \cdot \mu^1[\chi(t)].$$

The resulting *approximative* ODE system

$$\mu_t(t) = F(\mu(t), q), \tag{4}$$

for $\mu(t;q) = [\mu^1[\chi(t)], \mu^{c,2}[\chi(t)]]^T$ is given

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1[\chi(t)] &= q_1\mu^1[\chi(t)]\left[1-\mu^1[\chi(t)]\right] + q_2(\chi_0 - \mu^1[\chi(t)]) - q_1\mu^{c,2}[\chi(t)] \\
\frac{\partial}{\partial t}\mu^{c,2}[\chi(t)] &= -2q_1(2\mu^1[\chi(t)]+2)\mu^{c,2}[\chi(t)] - 2q_2\mu^{c,2}[\chi(t)] \\
&\quad +2q_1\mu^1[\chi(t)](\mu^1[\chi(t)]-1) + 2q_2(\chi_0 - \mu^1[\chi(t)]).
\end{aligned}$$

An alternative *approximative* ODE system (4) for $m(t) = [\mu^1(t), \mu^{c,2}(t)]^T$ can be derived from the linear noise approximation which reads as

$$\begin{aligned}
\frac{\partial}{\partial t}\mu^1[x(t)] &= q_1\mu^1[x(t)](1-\mu^1[x(t)]) + q_2(x_0 - \mu^1[x(t)]), \\
\frac{\partial}{\partial t}\mu^{c,2}[x(t)] &= -2q_1(2\mu^1[x(t)]-1)\mu^{c,2}[x(t)] + 2q_2(x_0-1)\mu^{c,2}[x(t)] \\
&\quad +2q_1\mu^1[x(t)](\mu^1[x(t)]-1) + 2q_2(\chi_0 - \mu^1[x(t)]).
\end{aligned}$$

Adjoint Method for Gradient based Optimization

For simplicity, consider a cost function d that can be written as a parameter independent $\langle\cdot,\cdot\rangle$ inner product of the residual $r(q) = \hat{\mu}^o - \mathcal{D}\mathcal{N}\mu(q)$ with itself. Then, the derivative of

$$d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) = \langle r(q), r(q)\rangle = \langle \hat{\mu}^o - \mathcal{D}\mathcal{N}\mu(q), \hat{\mu}^o - \mathcal{D}\mathcal{N}\mu(q)\rangle$$

in direction of the j-th unit vector e^j of $\mathbb{R}^l$ in first place is given by

$$\frac{\partial}{\partial v} d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) = -2\langle r(q), \mathcal{D}\mathcal{N}m^j(q)\rangle,$$

where $m^j(t;q)$ denotes the solution of the linear ODE system

$$\frac{\partial}{\partial t}m^j(t) = F_\mu(\mu(t),q)m^j(t) + F_q(\mu(t),q)e^j,\ m^j(0) = 0 \tag{5}$$

obtained from (4) by linearization. Introducing the associated adjoint system

$$\frac{\partial}{\partial t}u(t) = -F_\mu(\mu(t),q)^T u(t) + \mathcal{N}^*\mathcal{D}^* r(q),\ u(t_f) = 0, \tag{6}$$

where F_μ^T denotes the transposed matrix of the Jacobian F_μ and $\mathcal{N}^*\mathcal{D}^*$ denotes the adoint operator of $\mathcal{D}\mathcal{N}$, the derivative can equivalently be expressed as

$$\frac{\partial}{\partial v} d(\hat{\mu}^o, \mathcal{D}\mathcal{N}\mu(q)) = 2\left(u, F_q(\mu(q),q)e^j\right). \tag{7}$$

This follows from

$$\begin{aligned}
\langle r(q), \mathcal{D}\mathcal{N} m^j(q)\rangle &= \big(\mathcal{N}^*\mathcal{D}^* r(q), m^j(q)\big) \\
&= \left(\frac{\partial}{\partial t}u(q) + F_\mu(\mu(q),q)^T u(q), m^j(q)\right) \\
&= \left(\frac{\partial}{\partial t}u(q), m^j(q)\right) + \big(u(q), F_\mu(\mu(q),q) m^j(q)\big) \\
&= -\left(u(q), \frac{\partial}{\partial t} m^j(q)\right) + \big(u(q), F_\mu(\mu(q),q) m^j(q)\big) \\
&= -\big(u(q), F_q(\mu(q),q) e^j\big)
\end{aligned}$$

where $(\cdot,\cdot)$ denotes the innner product in $L^2(0,t_f)$. Building the gradient information via (5) requires to solve the linearized system l times, i.e., one time for each direction e^j. The computational advantage of the adjoint approach is that only the j-independent linearized system (6) has to be solved (backwards in time) in order to then build (7) for $j = 1,\ldots,l$.

Given function values $f = (f_0,\ldots,f_{n_t})^T$ at discrete time points, the adjoint of $\mathcal{D}$ is defined by

$$\mathcal{D}^* f = \sum_{j=1}^{n_t} f_j \delta(t - t_j),$$

where the δ-function satisfies $\delta(0) = 1$ and $\delta(\tau) = 0$ for $\tau \neq 0$. The adjoint of the embedding operator $\mathcal{N}$ maps moment expressions for observables to moment expressions of the full state space by introducing zeros whenever an unobservable is involved. For instance, with $\bar{k} = 1$ we have

$$\mathcal{N}^* : \mathcal{F}([0,t_f],\mathbb{R}^d) \to \mathcal{F}([0,t_f],\mathbb{R}^n), \mu^{1,o}(\cdot;q) \to [\mu^{1,o}(\cdot;q); 0].$$

The key ingredient to (6) is the Jacobian matrix $F_\mu(\mu(t;q),q)$ which can be optained manually or by symbolic computation tools to be evaluated at the solution $\mu(t;q)$ of (4) for the current parameter guess q. The system (6) can be solved by standard ODE solvers after transformation to the time variable $\tau = t_f - t$ in order to obtain a (well-posed) system forward in time.

References

1. Lee CH, Kyeong-Hun Kim KH, Kim P (2009) A moment closure method for stochastic reaction networks. Journal of Chemical Physics 130: 813-819.

2. Engblom S (2006) Computing the moments of high dimensional solutions of the master equation. Applied Mathematics and Computation 180: 498-515.

3. Gillespie CS (2009) Moment-closure approximations for mass-action models. IET Systems Biology 3: 52-58.

4. Milner P, Gillespie CS, Wilkinson DJ (2011) Moment closure approximations for stochastic kinetic models with rational rate laws. Mathematical Biosciences 231: 99-104.

5. Matis TI, Guardiola IG (2010) Achieving moment closure through cumulant neglect. The Mathematica Journal 12.