PARAMETER ESTIMATION FOR STOCHASTIC BIOCHEMICAL PROCESSES: A COMPARISON OF MOMENT EQUATION AND FINITE STATE PROJECTION

Atefeh Kazeroonian[1], *Jan Hasenauer*[1,2], *and Fabian Theis*[1,2]

[1]Institute of Computational Biology, Helmholtz Center Munich,
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[2]Department of Mathematics, University of Technology Munich,
Boltzmannstr. 3, 85747 Garching, Germany

{atefeh.kazeroonian,jan.hasenauer,fabian.theis}@helmholtz-muenchen.de

ABSTRACT

Many biochemical processes exhibit intrinsic stochastic fluctuations. These intrinsic fluctuations can be modeled using the chemical master equation (CME). The estimation of the parameters of the CME is challenging because the CME is a high or infinite dimensional system.

We compare two approaches currently used to estimate parameters of CMEs from population snapshot data. The first approach relies on a truncation of the CME, the finite state projection, and uses the data directly. The second method relies on moment equations – dynamical systems computing the moments of the CME solution – and merely uses the moments of the data. The second method is computationally more efficient, however, it cannot use all information contained in the data. In this manuscript, we assess the statistical power of the individual approaches and study moment equations of different order. Furthermore, we refine the likelihood function for the moment equation and introduce a novel validation method.

We performed a comparative study of the commonly used 3-stage model of gene expression. Using maximum likelihood estimates and a rigorous uncertainty quantification based on profile likelihoods, we show that the finite state projection approach is statistically more powerful than approaches based on moment equation. Nevertheless, even in case of partial observations, the first and second moments of the CME solution are highly informative and permit parameter identifiability. These findings, in combination with the novel tools for validation and uncertainty analysis, improve the insight into the problem class.

1. INTRODUCTION

In recent years, a multitude of studies have shown that many biochemical processes in prokaryotic and eukaryotic cells exhibit intrinsic stochastic fluctuations [1]. These fluctuations arise from low copy-number effects and are particularly significant for transcription and translation [2]. It is now known that these fluctuations are in many cases required for cellular function, e.g., for robust decision making on the population level [1].

The stochastic dynamics of biological processes can be described using continuous-time discrete-state Markov chains (CTMCs). The statistics of these Markov chains are governed by the chemical master equation (CME). Individual realizations of the process can be obtained via stochastic simulation algorithms (SSAs) [3, 4]. The stochastic process can be studied by analyzing statistics of many such realizations. Alternatively, the CME can be simulated using the finite state projection (FSP) method [5], which relies on truncation of the state space of the CME. While SSAs and the FSP are in principle capable of resolving all details of the dynamics of the CME, they impose a significant computational cost. This computational cost already becomes intractable for many small-scale systems. As an alternative, the method of moments (MM) [6, 7, 8] can be employed to capture the overall statistics of the process, such as mean and variance of individual species as well as covariances.

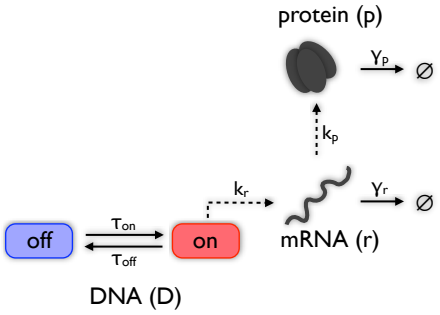
While the SSA, the FSP, and the MM all have advantages and disadvantages, a joint property is that they require accurate parameter values. The models and simulations are only predictive if good estimates of the reaction rates are available. Several estimation methods, relying on different models, were proposed (see, e.g., [9] and references therein), however, in most studies only the optimal parameter estimate has been considered, and the methods have not been compared. In this manuscript, we study the parameter estimates and confidence intervals obtained using FSP and MM. We present the individual likelihood functions and evaluate the informativeness using profile likelihoods. This is done for the widely used 3-stage model of gene expression [2], which is depicted in Figure 1.

2. METHODS

2.1. Modeling and simulation

2.1.1. Chemical master equation

The time evolution of the state $X = (X_1, \dots, X_{n_s})^{\mathrm{T}} \in \mathbb{N}_0^{n_s}$ of stochastic biochemical reaction networks is mostly described using CTMCs. The statistics of CTMCs are



Moment equation (order 1):

$$\dot{\mu}_{\mathrm{D_{off}}} = \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}} - \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}}$$
$$\dot{\mu}_{\mathrm{D_{on}}} = \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}} - \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}}$$
$$\dot{\mu}_{\mathrm{r}} = k_{\mathrm{r}}\mu_{\mathrm{D_{on}}} - \gamma_{\mathrm{r}}\mu_{\mathrm{r}}$$
$$\dot{\mu}_{\mathrm{p}} = k_{\mathrm{p}}\mu_{\mathrm{r}} - \gamma_{\mathrm{p}}\mu_{\mathrm{p}}$$

Figure 1. **Three-stage gene expression model.** (left) Schematic of the 3-stage gene expression model shows two DNA states (on, off), mRNAs and proteins. Transitions as well as synthesis and degradation reactions are shown as arrows. (right) Moment equations for means and variances of the individual species. The subscripts indicate the dependency, e.g., μ_r is the mean mRNA number.

governed by the CME. For a process with n_r chemical reactions,

$$R_k : \sum_{i=1}^{n_s} \nu_{ik}^- X_i \rightarrow \sum_{i=1}^{n_s} \nu_{ik}^+ X_i,$$

with reaction stoichiometries ν_k^-, ν_k^+, and $\nu_k = \nu_k^+ - \nu_k^-$, and reaction propensities $a_k(X, \theta)$, the CME is

$$\frac{\partial}{\partial t}p(x;t) = \sum_{\substack{k=1 \\ x \geq \nu_k^+}}^{n_r} a_k(x-\nu_k,\theta)p(x-\nu_k;t) - \sum_{k=1}^{n_r} a_k(x,\theta)p(x;t).$$

The solution of the CME depends on the parameters θ, which are for instance reaction rates.

The CME is defined for all reachable states $x \in \Omega \subset \mathbb{N}_0^{n_s}$, where n_s is the number of biochemical species. The set of reachable states Ω is in general very large, or infinite, rendering a direct solution of the full CME infeasible. Fortunately, the set of states with a significant probability mass is often small. This is exploited by the FSP, a direct method for approximating the solution of the CME [5] with pre-specified accuracy. Therefore, a subset Ω^{FSP} of the set of reachable states Ω is chosen. The time evolution of $p(x;t)$ with $x \in \Omega^{\mathrm{FSP}}$ is described by the CME, but influxes from states $x - \nu_k \notin \Omega^{\mathrm{FSP}}$ are removed. Probabilities $p(x;t)$ resulting from the simulation of this truncated system, which can be shown to be a lower bound for the actual probabilities of the CME, converge to the actual probabilities by growing Ω^{FSP} until the pre-specified accuracy is met.

A requirement for the application of the FSP is that the number of states with a significant probability mass is not too large. Novel algorithms can handle some million states [10]. Beyond this, the direct numerical simulation becomes infeasible.

2.1.2. Method of moments

In situations where the FSP is no longer applicable, the method of moments can be employed to approximate the solution of the CME [6]. The MM, also called moment equation, does not reproduce the exact solution of the CME. Instead, it computes the moments of $p(x;t)$, i.e. mean

$$\mu_i(t) = \sum_{x \in \Omega} x_i p(x;t),$$

variance

$$C_{ij}(t) = \sum_{x \in \Omega} (x_i - \mu_i(t))(x_j - \mu_j(t))p(x;t),$$

and higher-order moments [6]. The dynamics of the moments are governed by a set of ordinary differential equations (ODEs). Given that chemical reactions are at most bimolecular, the ODEs for the mean and the variance are

$$\frac{d\mu_i}{dt} = \sum_{k=1}^{n_r} \nu_{ik} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right),$$

$$\frac{dC_{ij}}{dt} = \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_{l} \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{il} + \nu_{jk} \sum_{l} \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{jl} + \nu_{ik}\nu_{jk} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right) \right)$$
$$+ \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{i l_1 l_2} + \nu_{jk} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{j l_1 l_2} \right),$$

in which $C_{il_1l_2}$ and $C_{jl_1l_2}$ are third order moments according to notation used in [6]. The governing equation for arbitrary moment orders can be found in [6, Equation (2.46)]. If all reactions are at most mono-molecular, the moment equation is closed, meaning that the evolution of moments of order m does not depend on moments of order greater than m. In this case, the moment equations are exact. If bimolecular chemical reactions are present, the moment equation ODEs are not closed, and the evaluation of a moment of order m requires the moments of order $m+1$ [6]. Moment closure techniques must be employed [11], and the resulting moments will only be an approximation of the true moments of the solution of the CME.

Moment equations are in general low-dimensional compared to the CME. Thus, they can generally be solved more efficiently. However, a drawback is that a finite number of moments does not allow the reconstruction of the underlying distribution $p(x;t)$. Hence, information is lost.

2.2. Parameter estimation

In this work, we considered population snapshot data $\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{S_k}$, $k = 1, \ldots, N$, obtained by sampling cells $s = 1, \ldots, S_k$ from the cell population and measuring one (or more) properties of these cells, e.g., using flow cytometry or microscopy. For notational simplicity, we assume that one observable, $\bar{Y} = h(X)$, can be measured. The observation function h describes the type of measurement; in the most simple case $h(X) = X_i$. The measurement is assumed to be noise-free as we later want to assess the informativeness of single-cell data vs. the moments.

Given a realization X at a certain time t_k, the probability of observing $\bar{Y}$ at time t_k is $p(y = \bar{Y}; x = X)$. The total probability to observe $\bar{Y}$ at time t_k is obtained by taking into account all possible realizations $X \in \Omega$ of the process. Given that the number of molecules is a discrete variable, this total probability is obtained by marginalizing over the state space Ω,

$$p(y; t_k, \theta) = \sum_{x \in \Omega} p(y; x)\, p(x; t_k, \theta),$$

where $p(x; t_k, \theta)$ is the solution of the CME. Bearing in mind that we do not consider any measurement noise, y is a deterministic function of x, $y = h(x)$, thus

$$p(y|x) = \begin{cases} 1 & \text{if } y = h(x) \\ 0 & \text{otherwise,} \end{cases}$$

so the sum simplifies to

$$p(y; t_k, \theta) = \sum_{\substack{x \in \Omega \\ h(x) = y}} p(x; t_k, \theta).$$

Following the argumentation above, the probability distribution $p(y; t_k, \theta)$ is the distribution from which the observations are drawn. Thus,

$$p(y = \bar{Y}^{(s)}(t_k)) = p(y; t_k, \theta), \quad s = 1, \ldots, S_k.$$

In the following, we compare two classes of likelihood functions for these data, namely an FSP-based likelihood function and a moment-based likelihood function with respect to their statistical power. As mentioned before, we do not consider any measurement noise in this comparison, but the inclusion of noise in the presented procedure would be rather straightforward.

2.2.1. FSP-based estimation

As outlined earlier, for CTMCs with a small effective state space, the FSP can be used to approximate the solution of the CME for a given parameter set θ. Using this approximation of the probability distribution of the hidden state, $p(x; t, \theta)$, and the corresponding approximation of the probability distribution of the observable, $p(y; t, \theta)$, the likelihood of the stochastic process,

$$\mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) = c \prod_{k=1}^{N} \prod_{s=1}^{S_k} p(y = \bar{Y}^{(s)}(t_k); t_k, \theta),$$

can be evaluated. Basically, the probabilities are evaluated and multiplied for all observed states. The constant c depends only on the data and can be neglected for optimization purposes. For a detailed introduction of this FSP-based likelihood function, we refer to [12, 13]. Given the FSP-based likelihood function, the estimation problem can be formulated. The FSP-based maximum likelihood (ML) estimation problem is:

$$\begin{aligned} &\underset{\theta}{\text{maximize}} \ \log \mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) \\ &\text{subject to } \Sigma^{\text{FSP}}(\theta), \end{aligned}$$

in which $\Sigma^{\text{FSP}}(\theta)$ denotes the finite-dimensional ODE model resulting from the FSP of the CME on the subset Ω^{FSP}. To reduce numerical problems, the problem is formulated using the log-likelihood function $\log \mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta)$. Furthermore, we optimize the logarithm of the parameters $\xi = \log_{10}(\theta)$ to ensure positivity and improve the performance of the optimization routines. The optimal solution of the FSP-based ML estimation problem is the parameter vector for which the likelihood of observing the single cell data is maximized. This estimator uses all available information.

2.2.2. *Moment-based estimation*

For many processes the approximation of the CME solution using the FSP is not feasible because the number of states with non-negligible probability is too large. In such cases, the moment equation can be employed to approximate the statistics of the CME solution. To employ moment equations for parameter estimation, the statistics of the snapshots are computed, e.g., mean and variance,

$$\bar{\mu}_y(t_k) = \frac{1}{S_k} \sum_{s=1}^{S_k} \bar{Y}^{(s)}(t_k),$$

$$\bar{C}_{yy}(t_k) = \frac{1}{S_k} \sum_{s=1}^{S_k} \left(\bar{Y}^{(s)}(t_k) - \bar{\mu}_y(t_k) \right)^2.$$

These measured moments are compared to moments predicted by the model and the observation function $h(x)$. Since the sample sizes S_k are often quite large – for flow cytometry often in the order of 10^4 – it follows from the central limit theorem that the empirical moments, e.g., $\bar{\mu}_y(t_k)$ and $\bar{C}_{yy}(t_k)$, are almost normally distributed around the true moments [14]. Hence, a normal error model is assumed,

$$\mathcal{L}_{\mathcal{D},\mu_y}^{\text{MM}}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(\mu_y(t_k,\theta) | \bar{\mu}_y(t_k), \sigma^2_{\bar{\mu}_y}(t_k) \right),$$

$$\mathcal{L}_{\mathcal{D},C_{yy}}^{\text{MM}}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(C_{yy}(t_k,\theta) | \bar{C}_{yy}(t_k), \sigma^2_{\bar{C}_{yy}}(t_k) \right),$$

where $\mathcal{N}(\cdot|\mu,\sigma^2)$ is the probability density of the normal distribution. Such a likelihood function can be derived for every moment predicted by the model, e.g., also the third and fourth order central moments. Clearly, the consideration of additional, non-redundant moments provides additional information about the model parameters as the individual likelihood functions are multiplied, e.g., if mean and variance are employed then a reasonable likelihood function is

$$\mathcal{L}_{\mathcal{D}}^{\text{MM}}(\theta) = \mathcal{L}_{\mathcal{D},\mu_y}^{\text{MM}}(\theta) \cdot \mathcal{L}_{\mathcal{D},C_{yy}}^{\text{MM}}(\theta).$$

Unfortunately, also the computational complexity of simulating the moment equations increases with each additional moment considered in the model.

The likelihoods $\mathcal{L}_{\mathcal{D},\mu_y}^{\text{MM}}(\theta)$, $\mathcal{L}_{\mathcal{D},C_{yy}}^{\text{MM}}(\theta)$ and those for the higher-order moments require information about the error variance of the respective empirical estimator, e.g., $\sigma^2_{\bar{\mu}_y}$ for $\bar{\mu}_y(t_k)$ and $\sigma^2_{\bar{C}_{yy}}$ for $\bar{C}_{yy}(t_k)$. The variance of the estimators for the first and second order moments can be found in [14]. For third and higher-order moments the calculation of these estimators become increasingly complex, and we did not find respective results in the literature. To circumvent the analytical derivation, we propose to estimate the variance of the empirical estimators using non-parametric bootstrapping [15]. This approach employs a two-step procedure. At first, a sample of size S_k is drawn from $\{\bar{Y}^{(s)}(t_k)\}_{s=1}^{S_k}$ (all $\bar{Y}^{(s)}(t_k)$ have probability $\frac{1}{S_k}$) and the moments of this artificial sample are evaluated. This step is repeated a large number of times, in general more than one thousand times, yielding a large sample for each moment of interest. Therefore, the variance of each moment can easily be computed from the corresponding sample. This sample variance is a reliable measure for the uncertainty, if $S_k \gg 1$. It does not require any distribution assumption for $p(y;t_k,\theta)$ and is easily applicable to any higher-order moments.

Given the likelihood function $\mathcal{L}_{\mathcal{D}}^{\text{MM}}(\theta)$, which is the product of the likelihood functions for the moments of interest, the moment-based ML estimation problem,

$$\begin{aligned} &\underset{\theta \in \mathbb{R}_+^{n_\theta}}{\text{maximize}} \ \log \mathcal{L}_{\mathcal{D}}^{\text{MM}}(\theta) \\ &\text{subject to } \Sigma^{\text{MM}}(\theta), \end{aligned}$$

can be formulated. $\Sigma^{\text{MM}}(\theta)$ is the model used to simulate the moment equations for the moments of interest.

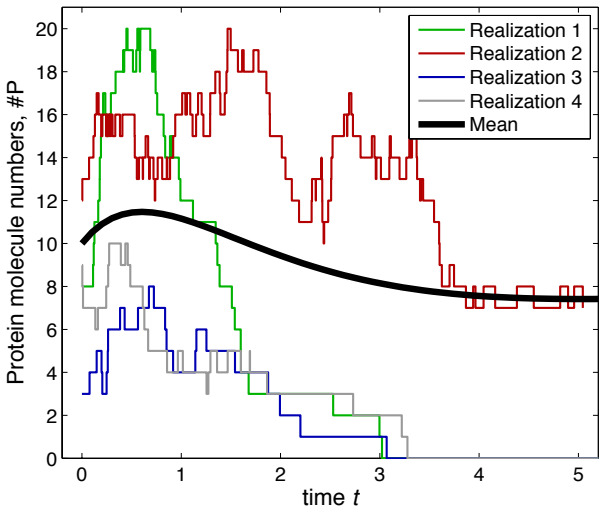
2.2.3. *Identifiability and uncertainty analysis*

As the measurement data are limited and potentially noise corrupted, the parameters can in general not be estimated precisely. To assess the remaining parameter uncertainty and the practical identifiability, we use profile likelihoods [16]. Given the likelihood function $\mathcal{L}_{\mathcal{D}}(\theta)$, the profile likelihood of parameter θ_i is

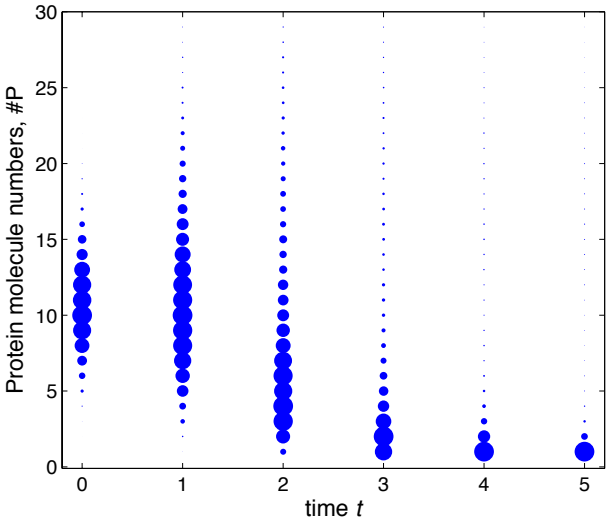
$$\text{PL}(\theta_i) = \max_{\theta_{j \neq i}} \mathcal{L}_{\mathcal{D}}(\theta).$$

This profile likelihood $\text{PL}(\theta_i)$ is the maximal likelihood for a given value of θ_i. Using the profile likelihood, the likelihood ratio $R(\theta_i) = \text{PL}(\theta_i)/\mathcal{L}_{\mathcal{D}}(\hat{\theta})$ can be evaluated, in which $\hat{\theta}$ is the ML estimate. The likelihood ratio R is one at the globally optimal point $\hat{\theta}_i$ and approaches zero for large $|\theta_i - \hat{\theta}_i|$ if the parameter is identifiable. The area under $\text{PL}(\theta_i)$ provides a reasonable measure for the uncertainty of parameter θ_i. For further details, we refer to [16, 17].

In the following, we employ profile likelihoods to assess the information content of the moments of the data in comparison with that of the full distribution of data. More information will result in many identifiable parameters and small parameter uncertainties.



(a) **Four stochastic realizations of the 3-stage model of gene expression.**



(b) **Population snapshot data used for parameter estimation.**

Figure 2. **Dynamics of the 3-stage model of gene expression.** (a) Time-dependent protein number in four representative cells together with the population mean. (b) Population snapshot data obtained by sampling single cell trajectories. The size of the markers in (b) is proportional to the number of observed cells with the corresponding protein number. Due to the long tail of the distribution, the mode of the data seen in (b) differs significantly from the mean of the data depicted in (a).

3. RESULTS AND DISCUSSION

3.1. Parameter estimation for the 3-stage model of gene expression

In this section, we compare the performance of previously mentioned estimation methods, namely, FSP-based and MM-based parameter estimates, using the common 3-stage model of gene expression [2]. A schematic of the process and the corresponding moment equations for mean and variance are shown in Figure 1. The model has six parameters: the transition rate of DNA into the on-state (τ_{on}), the transition rate of DNA into the off-state (τ_{off}), the transcription rate in the on-state (k_{r}), the rate of mRNA degradation (γ_{r}), the translation rate (k_{p}), and the rate of protein degradation (γ_{p}). In the following, we study the problem of estimating these rates from protein measurements. Therefore, we generate artificial data

$$\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{10^5}, \quad k = 1, \ldots, 10,$$

with $t_k = k$ and $\bar{Y}$ being the number of proteins. For the generation of the artificial data, the parameter vector

$$\begin{aligned} \theta^{\text{true}} &= (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}} \\ &= (0.05, 0.05, 5, 1, 4, 1)^{\text{T}} \end{aligned}$$

is used. We refer to this parameter vector θ^{true} as the true parameter vector in the following. Also, no measurement noise is considered in the generation of the data. In the initial state, mRNA and protein numbers follow a Poisson distribution with mean 4 and 10, respectively. The probability to be in the DNA on-state is 0.7. Figure 2 depicts sample paths of the model (Figure 2(a)) as well as the snapshot data (Figure 2(b)) used for parameter estimation. Using these data we estimate $\theta = (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}}$.

For FSP-based and moment-based likelihood functions the maximum likelihood estimates are computed and the parameter uncertainty is evaluated. For the moment-based likelihood function we employed different moment orders. The uncertainty of the moments has been determined using the non-parametric bootstrapping approach introduced before.

Figure 3 depicts the model simulation for the ML estimates for the different likelihood functions along with the data. It is clear that for all ML estimates we observe a good agreement with the data used for the estimation. To validate the ML estimates, we employed the higher-order moments of the data, which have not been used for the parameter estimation. We find that all ML estimates, which were obtained using at least the mean and the variance, successfully predict the higher-order moments not used to obtain the ML estimates. Only the ML estimate computed merely from the mean of the data fails. Thus, the information contained in the mean is insufficient. This is confirmed by the profile likelihoods shown in Figure 4, which show that all likelihood functions establish identifiability, except the moment-based likelihood function of order 1. A careful comparison of the profile likelihoods shows that the uncertainty in the estimation of the parameters decreases as more information (more moments) are used. Since the FSP-based likelihood function makes use of all the information, the resulting parameter uncertainties are minimal. If the moment order is increased, the confidence intervals for moment-based likelihood function also become more narrow, however even for moment order 4, the result of the FSP remains superior. Note that for all likelihood functions, the true parameters are con-

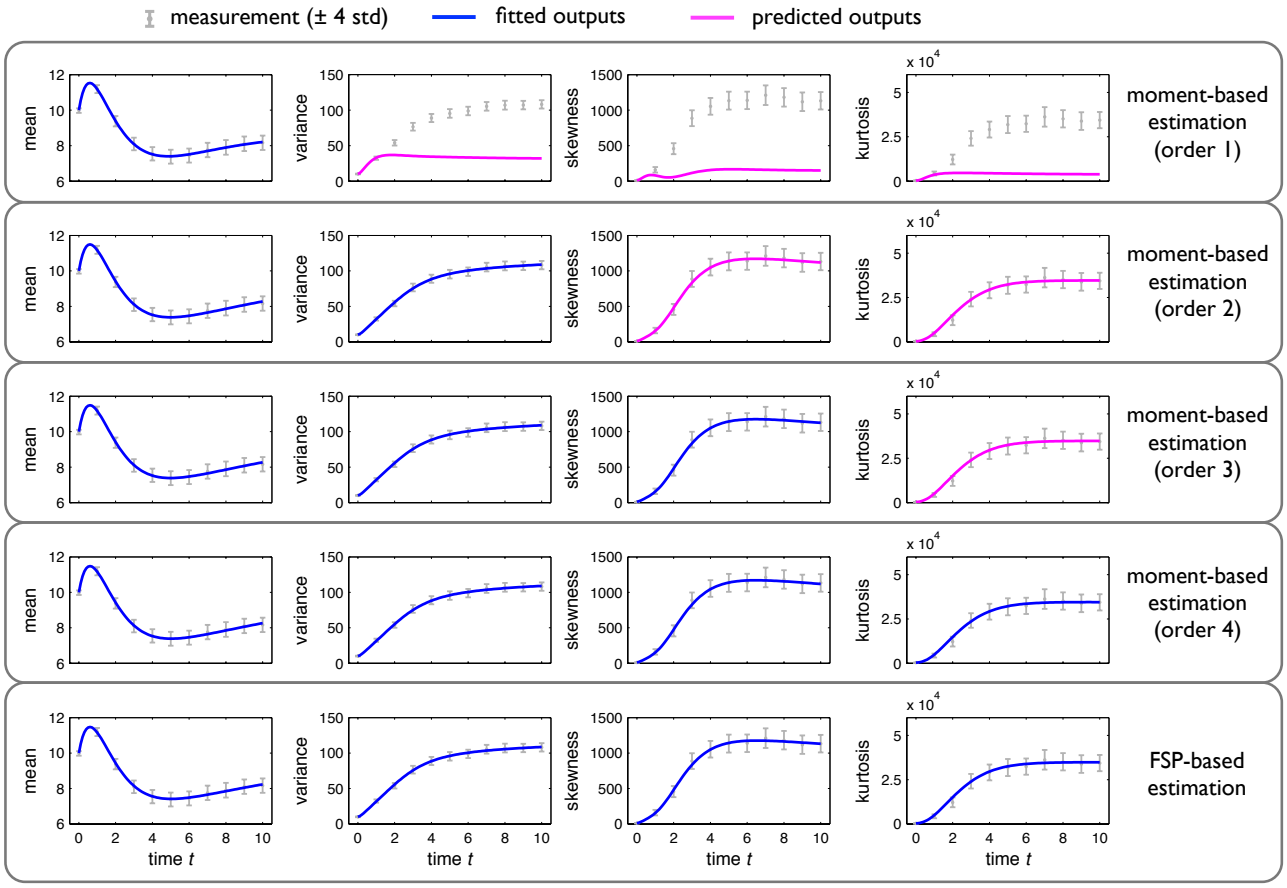


Figure 3. **Model-data comparison for ML estimates obtained using different likelihood functions.** ML estimation has been performed using moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. Gray error bars show the mean and 4-σ intervals ($[\mu-4\sigma, \mu+4\sigma]$) of the measurement data. For the different ML estimates the fit is illustrated by showing the model output (blue lines, —) and the measurement data (grey error bars). All models describe the respective data well. To assess the predictive power of the model, the ML estimates are used to predict the higher-order moments (magenta lines, —) which have not been employed for the parameter estimation. The ML estimate computed using moment-based estimation of order 1 fails to provide good prediction, while already information about mean and variance (order 2) is sufficient to obtain a predictive model.

tained in the 95% confidence intervals constructed from the profile likelihoods (not shown).

3.2. Discussion

The computational complexity of the simulation of CTMCs renders the estimation of their parameters challenging. Different methods have been proposed to circumvent this complexity, among other the moment equations [18, 9, 14]. In this work, we evaluate the information contained in the moments of measurement data with respect to parameter estimation (by employing moment-based likelihood function) and compare it with the complete information contained in population snapshot data (by employing FSP-based likelihood function). The practical identifiability and the uncertainty of the parameter estimates are assessed using profile likelihoods. To the best of our knowledge, this is the first profile likelihood-based uncertainty analysis for stochastic processes, probably because the evaluation of the likelihood function is computationally often infeasible. This is not the case if a moment-based estimation is employed.

As a case study, we consider the widely used 3-stage model of gene expression [2]. For this model, we show that measurements of the mean expression do not in general ensure identifiability, but rather that measurements of the variance are required. This is consistent with results by Munsky *et al.* [18] for the two-stage model of gene expression. Information about third and fourth order moments can decrease the uncertainty further, however this reduction is often insignificant. The full information contained in the data, which is exploited by the FSP-based estimation, remains out of reach for the MM-based estimation approach.

Although the FSP-based likelihood function is statistically more powerful, parameter estimation based on the moment equation is the method of choice for processes,

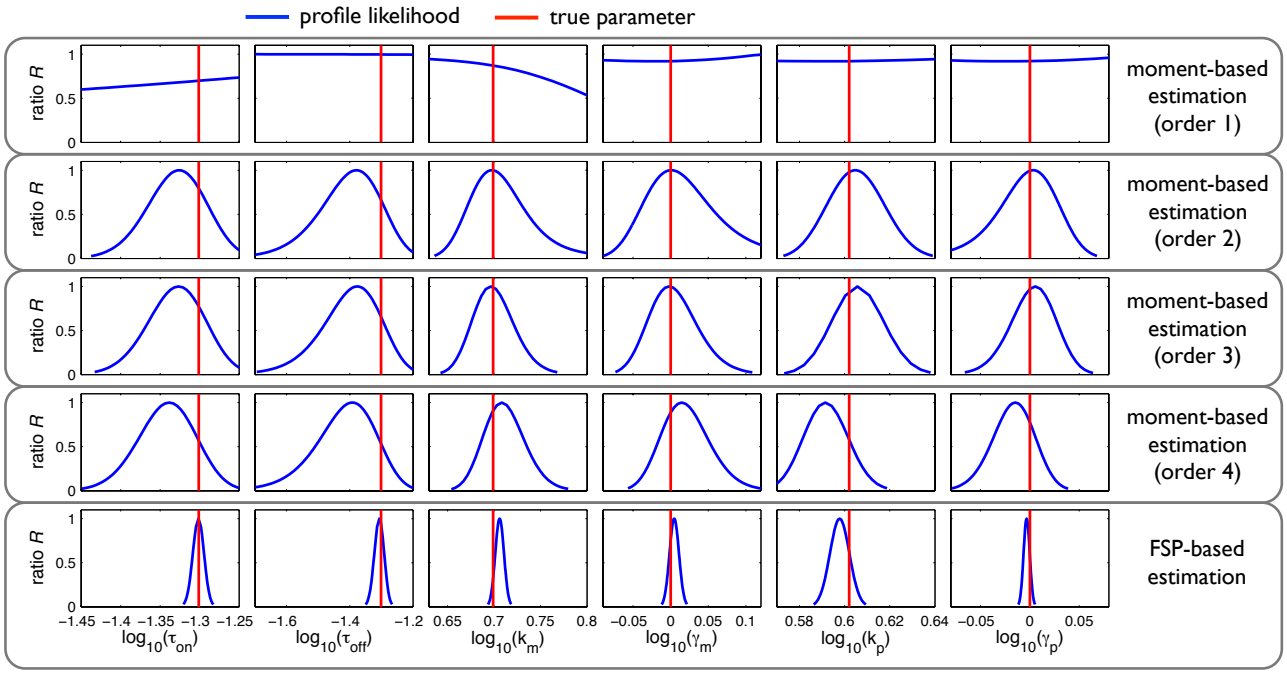


Figure 4. **Parameter uncertainty for different likelihood functions.** The parameter uncertainty and parameter identifiability has been evaluated for moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. The profile likelihoods (blue lines, —) indicate that the measurements of the mean do not carry enough information to identify the parameters. Information about mean and variance ensures identifiability, and the uncertainty is slightly reduced if additional moments are used. The FSP-based likelihood function, which exploits all information contained in the data, yields the smallest uncertainties. All confidence intervals (not shown), derived from likelihood profiles, contain the true parameter values (red lines, —), which indicates consistency.

in particular, if the FSP is infeasible. Furthermore, parameter estimation using the moment equation is more efficient. The parameter estimation using the moment equation of order 2 is roughly 30 times faster than the parameter estimation using the FSP. However, it remains to be studied how moment closures, which are required for systems including bimolecular reactions, influence the parameter estimation. If a bias is introduced, as we expect, it should be analyzed how a refinement of the moment equation, e.g., the conditional moment equation [19], can be used to improve the results.

Beyond the profile likelihood-based evaluation of the information encoded in the moments, we introduced a non-parametric bootstrapping approach to evaluate the uncertainty of the empirical estimates of the moments. This approach allows for the construction of likelihood function without additional distribution assumptions. Furthermore, we illustrated how the higher-order moments, which have not been used for parameter estimation, can be used for model validation. This approach is attractive, as models can basically be fitted and validated on the same dataset.

4. AUTHOR'S CONTRIBUTIONS

AK and JH developed the method and analyzed the 3-stage model of gene expression. JH and FJT devised the project. AK, JH and FJT wrote, read and approved the final manuscript.

5. ACKNOWLEDGEMENTS

The authors acknowledge financial support by the European Union within the ERC grant 'LatentCauses' and the BMBF grant 'Virtual Liver' (grant-nr. 315752). The authors would also like to thank Justin Feigelman and Sabine Hug for proofreading the manuscript.

6. REFERENCES

[1] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nat.*, vol. 467, no. 9, pp. 1–7, Sept. 2010.

[2] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," *Proc. Natl. Acad. Sci. U S A*, vol. 105, no. 45, pp. 17256–17261, Nov. 2008.

[3] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, Dec. 1977.

[4] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie, "Stochastic modelling of gene regulatory networks," *Int. J. Robust Nonlinear Control*, vol. 15, no. 15, pp. 691–711, Oct. 2005.

[5] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *J. Chem. Phys.*, vol. 124, no. 4, pp. 044104, Jan. 2006.

[6] S. Engblom, "Computing the moments of high dimensional solutions of the master equation," *Appl. Math. Comp.*, vol. 180, pp. 498–515, 2006.

[7] J. P. Hespanha, "Modeling and analysis of stochastic hybrid systems," *IEE Proc. Control Theory & Applications,* Special Issue on Hybrid Systems, vol. 153, no. 5, pp. 520–535, 2007.

[8] J. Ruess, A. Milias, S. Summers, and J. Lygeros, "Moment estimation for chemically reacting systems by extended Kalman filtering," *J. Chem. Phys.*, vol. 135, no. 165102, Oct. 2011.

[9] P. Milner, C. S. Gillespie, and D. J. Wilkinson, "Moment closure based parameter inference of stochastic kinetic models," *Stat. Comp.*, 2012.

[10] M. Mateescu, V. Wolf, F. Didier, and T. Henzinger, "Fast adaptive uniformisation of the chemical master equation," *IET. Syst. Biol.*, vol. 4, no. 6, pp. 441–452, 2010.

[11] A. Singh and J. P. Hespanha, "Approximate moment dynamics for chemically reacting systems," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 414–418, Feb. 2011.

[12] J. Hasenauer, N. Radde, M. Doszczak, P. Scheurich, and F. Allgöwer, "Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem," Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland, July 2011.

[13] T. Nüesch, "Finite state projection-based parameter estimation algorithms for stochastic chemical kinetics," Master thesis, Swiss Federal Institute of Technology, Zürich, 2010.

[14] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *Proc. Nati. Acad. Sci. U S A*, vol. 109, no. 21, pp. 8340–8345, May 2012.

[15] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statist. Sci.*, vol. 11, no. 3, pp. 189–228, 1996.

[16] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinf.*, vol. 25, no. 25, pp. 1923–1929, May 2009.

[17] W. Q. Meeker and L. A. Escobar, "Teaching about approximate confidence regions based on maximum likelihood estimation," *Am. Stat.*, vol. 49, no. 1, pp. 48–53, Feb 1995.

[18] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Mol. Syst. Biol.*, vol. 5, no. 318, Oct. 2009.

[19] J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis, "Method of conditional moments (MCM) for the chemical master equation," *submitted to the Journal of Mathematical Biology*, 2012.