Identifiability analysis for models of the translation kinetics after mRNA transfection

Susanne Pieschner, Jan Hasenauer, Christiane Fuchs

March 2021

1 Introduction

mRNA transfection is the process of introducing mRNA into a living cell. mRNA delivery has become increasingly interesting for biomedical applications because it enables treatment of diseases by means of targeted expression of proteins and it is transient, avoiding the risk of permanently integrating into the genome (see e. g. Sahin et al., 2014). One of the most prominent applications of mRNA transfection at the moment are the mRNA-based vaccine candidates that are currently under investigation to prevent COVID-19 infections (DeFrancesco, 2020). In such a context, it is, of course, very import to have a precise understanding of the dynamics of the underlying processes in order to be able to control them. Yet, many aspects and the determinants of the mRNA delivery process and the translation kinetics are difficult to measure and therefore poorly understood.

We aim at facilitating insights into these aspects through the use of mechanistic modeling and parameter inference for such models from experimental data. The data comes from an mRNA transfection experiment using fluorescence reporters and fluorescence microscopy which is one of the few ways to measure quantities within a living cell over time (i. e. keeping it alive is necessary). Due to the discrete nature of the molecular species within a cell and due to the fact that random fluctuations play a key role (Elowitz et al., 2002, Raj & van Oudenaarden, 2008), a continuous-time, discrete-space Markov process, also called a *Markov jump process (MJP)*, for which the dynamics are described by the so-called chemical master equation (CME), is widely accepted to be an appropriate stochastic description of the biochemical processes within a cell (Gillespie, 1992, Schnoerr et al., 2017). However, parameter inference for MJPs is computationally very demanding and often infeasible (see e. g. Warne et al., 2019). Therefore, several other representations of the biochemical kinetics have been developed. To some extent those can be considered as approximations to the corresponding MJP. The most commonly used representation is the *reaction rate equation (RRE)* which is a system of *ordinary differential equations (ODEs)* and thus provides a deterministic and state-continuous description of the kinetics. One approach that preserves the stochastic nature of the underlying process is the approximation by *Itô diffusion processes*. These are continuous-time, continuous-space stochastic processes described by *Itô-type stochastic differential equations (SDEs)*.

Single-cell fluorescence data from transfection experiments has been analyzed based on ODE modeling in several previous studies e. g. Ligon et al. (2014), Leonhardt et al. (2014), Fröhlich et al. (2018), and Reiser et al. (2019). Yet, several parameters of the considered ODE models for the translation kinetics after mRNA transfection are not identifiable from the experimental data. Moreover, the quality of the parameters estimates, i. e. whether the true kinetic rates are adequately captured, is unclear. Using models that explicitly account for stochasticity can help improve our ability to determine kinetic parameters from experimental data (Munsky et al., 2009). Browning et al. (2020) have recently compared parameter identifiability for ODE and SDE modeling approaches for four example models from different contexts based on simulated data and showed that SDE modeling does improve the identifiability for the studied models.

Here, we use the experimental data from Fröhlich et al. (2018) and study the question whether an SDE model allows to identify more parameters of the model of translation kinetics after mRNA transfection from experimental data compared to the corresponding ODE model. Inference from fluorescence data for SDE models has also been conducted e. g. in Heron et al. (2007), Finkenstädt et al. (2008), and Komorowski et al. (2009), however for an experimental setup that also included the transcription process. Finkenstädt et al. (2008) even considered an SDE and an ODE model in one of their case

studies, but their results did not directly show any differences in the parameter identifiability and the study was not focused on this aspect.

This article is structured as follows ...

2 Experimental data

We consider data that has previously been analyzed (based on ODE modeling) and published in Fröhlich et al. (2018). The data was generated in an experiment where cells (human hepatoma epithelial cell line HuH7) were transfected with mRNA encoding a green fluorescence protein (GFP). The cells were fixed on micro patterned protein arrays and time lapse microscopy images of the cells were taken every 10 minutes over the course of at least 30 hours (i.e. there are at least 180 measurements per cell). For the first hour, mRNA lipoplexes were added. Afterwards, the cells were washed with cell culture medium such that no further lipoplex uptake occurs. The time point at which the lipoplexes were taken up, dissolved and released the mRNA as well as the number of mRNA molecules released are unknown.

The released mRNA was translated into a fluorescence protein which caused the cells to fluoresce. For each image taken during the experiment, the fluorescence intensity is integrated over squares occupied by one cell in order to obtain one value for the mean fluorescence intensity per cell and time point (see Fröhlich et al., 2018, for further details about the image analysis).

The experiment was conducted with two different types of GFP that differ in their protein lifetime: enhanced GFP (eGFP) and a destabilized enhanced GFP (d2eGFP). The data set contains measurements for more than 800 cells for each type of GFP.

Some trajectories of the mean fluorescence intensity are displayed in Figure 1.

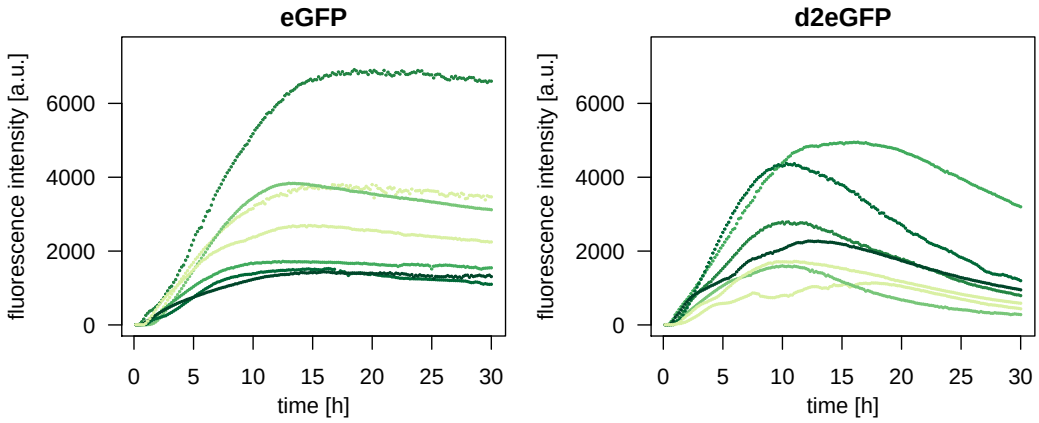


Figure 1: Trajectories of the mean fluorescence intensity for seven cells from the mRNA transfection experiment in Fröhlich et al. (2018) for eGFP and d2eGFP (April 27, 2016), respectively.

As will become clear in the course of this article, ODE models of the translation kinetics of an individual cell are not globally identifiable with the available experimental data as described above. Several of the ODE model parameters cannot be uniquely determined based on one observed fluorescence trajectory. Fröhlich et al. (2018) use a mixed-effect ODE model in order to incorporate the translation kinetics of several cells and data for both different types of GFP (eGFP and d2eGFP). Through this approach, they are able to improve parameter identifiability (by breaking the symmetry between the degradation rate constants); however, their approach is computationally very intense, required conducting the experiment with two types of GFP, and still leaves several parameters non-identifiable. Here, we are interested in the question whether the use of an SDE model can improve the parameter identifiability even when only one fluorescence trajectory is observed.

3 Mathematical models of the translation kinetics

While Fröhlich et al. (2018) use a mixed-effect ODE model in order to incorporate the translation kinetics of several cells, we will focus on modeling the translation kinetics of one cell in order to study parameter identifiability based on one observed fluorescence trajectory.

We consider the basic model configuration that models only the (released) mRNA and the GFP molecules explicitly. Therefore, our model is a dynamic process with two components:

$$\boldsymbol{X}(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} \quad \begin{array}{l} \text{amount of mRNA molecules,} \\ \text{amount of GFP molecules.} \end{array}$$

We assume that all mRNA molecules (within one cell) are released at once at the initial time point denoted by t_0. Before t_0, there are neither mRNA nor GFP molecules, and at t_0, an amount of m_0 mRNA molecules is released, i.e.

$$\boldsymbol{X}(t) \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ for } t \leq t_0 \quad \text{ and } \boldsymbol{X}(t_0) = \begin{pmatrix} m_0 \\ 0 \end{pmatrix}.$$

Conceivable extensions of this basic model configuration are e. g. to include enzymatic degradation of the mRNA and/or the protein, ribosomal binding to the mRNA for translation, and a maturation step of the protein. However, we will only consider the basic configuration as described above.

3.1 Markov Jump Process

Assuming that the matter within the cell is well-stirred and in thermodynamic equilibrium, a MJP is regarded to be the most adequate representation of this system after t_0. In our model configuration, there are three possible reactions:

$$\begin{array}{ll} X_1 \xrightarrow{\theta_1} \emptyset & \text{degradation of mRNA,} \\ X_1 \xrightarrow{\theta_2} X_1 + X_2 & \text{translation,} \\ X_2 \xrightarrow{\theta_3} \emptyset & \text{degradation of GFP.} \end{array}$$

The three reactions change the state of the system in the following way and occur with the following reaction rates:

$$\begin{array}{ll} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \longrightarrow \begin{pmatrix} X_1 - 1 \\ X_2 \end{pmatrix} & \text{with rate } \theta_1 X_1, \\ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \longrightarrow \begin{pmatrix} X_1 \\ X_2 + 1 \end{pmatrix} & \text{with rate } \theta_2 X_1, \\ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \longrightarrow \begin{pmatrix} X_1 \\ X_2 - 1 \end{pmatrix} & \text{with rate } \theta_3 X_2. \end{array}$$

If we denote the probability distribution of the random variable $\boldsymbol{X}(t)$ by

$$P_{i,j}(t) = \mathbb{P}(X_1(t) = i, X_2(t) = j),$$

the corresponding CME reads

$$\frac{\partial P_{i,j}(t)}{\partial t} = \theta_1(i+1)P_{i+1,j}(t) + \theta_2 i P_{i,j-1}(t) + \theta_3(j+1)P_{i,j+1}(t) - (\theta_1 i + \theta_2 i + \theta_3 j)P_{i,j}(t).$$

Although, the system contains only first-order reactions, there is no closed-form solution to the CME. Thus, there is no explicit formula for the transition probability distribution $p(\boldsymbol{X}(t)|\boldsymbol{X}(s), \theta)$ for $s < t$.

3.2 ODE model

The following system of ODEs is a deterministic approximation of the MJP modeling the dynamics as described above:

$$\frac{\mathrm{d}\boldsymbol{X}(t)}{\mathrm{d}t} = \begin{pmatrix} -\theta_1 X_1(t) \\ \theta_2 X_1(t) - \theta_3 X_2(t) \end{pmatrix} \quad \text{for } t \geq t_0. \tag{1}$$

This system admits the solution

$$X_1(t) = m_0 \exp\left(-\theta_1(t-t_0)\right),$$
$$X_2(t) = \begin{cases} \frac{\theta_2 m_0}{\theta_3-\theta_1}\left(e^{-\theta_1(t-t_0)} - e^{-\theta_3(t-t_0)}\right) & \text{, for } \theta_1 \neq \theta_3, \\ \theta_2 m_0 (t-t_0) e^{-\theta_3(t-t_0)} & \text{, for } \theta_1 = \theta_3. \end{cases} \tag{2}$$

Note that the solution for $X_2(t)$ is symmetric in the parameters θ_1 and θ_3.

3.3 SDE model

A stochastic but state-continuous approximation to the MJP in Section 3.1 is given by an Itô diffusion process that is described by the following SDE:

$$\mathrm{d}\boldsymbol{X}(t) = \begin{pmatrix} -\theta_1 X_1(t) \\ \theta_2 X_1(t) - \theta_3 X_2(t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \sqrt{\theta_1 X_1(t)} & 0 \\ 0 & \sqrt{\theta_2 X_1(t) + \theta_3 X_2(t)} \end{pmatrix} \mathrm{d}\boldsymbol{B}(t) \tag{3}$$

for $t \geq t_0$ and where $\boldsymbol{B}(t)$ is a 2-dimensional standard Brownian motion.

Note that for a diffusion approximation (as well as for the ODE approximation), the size of the system can play an important role. However, since the model that we consider here contains only first-order reactions, the size of the system does not affect the interpretation of the kinetic parameters and can simply be assumed to be equal to 1.

4 Model of the observations

In the experiment described in Section 2, neither the amount of mRNA molecules nor that of GFP molecules can be measured over time directly. Instead, a fluorescence signal is observed which is assumed to be a linear transformation of the amount of GFP molecules. Moreover, Fröhlich et al. (2018) state that "Analysis of processed data suggested a constant offset and multiplicative measurement noise in the recorded fluorescence trajectories." Therefore, denoting a trajectory of mean fluorescence intensity observed at time points t_k, for $k = 1, \ldots, K$, by $\{y_k\}_{k=1,\ldots,K}$, we assume that

$$\log(y_k) = \log\left(\mathrm{scale} \cdot X_2(t_k) + \mathrm{offset}\right) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2),$$

where the random variables ϵ_k are independent.

Note that the observations depend only on the amount X_2 of GFP molecules, but not directly on the amount X_1 of mRNA molecules.

Based on the observations $\{y_k\}_{k=1,\ldots,K}$, we aim to infer the following unknown parameters:

- the three kinetic parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ that denote the rate constants for mRNA degradation, translation, and GFP degradation,
- the initial amount m_0 of mRNA molecules and the time point t_0 at which it is released,
- the scaling factor scale and the offset for the fluorescence signal,
- and the standard deviation σ of the measurement errors.

5 Structural identifiability analysis

Our main interest lies in the question which of the model parameters for our two model types (ODE and SDE) can be inferred from the experimental data as described in Sections 2 and 4. Here, we first focus on the parameters $\boldsymbol{\theta}$, m_0, scale, and offset that drive the dynamics of the process and the fluorescence signal. We analyze the structural identifiability which only considers the model equations of the process dynamics and the observation equation (not the actual data) and assumes that we are in a perfect data situation, i. e. we have an infinite amount of data observed without measurement error (Raue et al., 2009). Plainly speaking, structural identifiability analysis answers the question whether different parameter combinations can lead to the same model output. While for ODE models, there are

analytical methods to assess structural identifiability, no such methods exit for SDE models. Therefore, we use several different approaches to heuristically assess structural identifiability for our SDE model. In the following subsections, we consider a transformed version of both model types, we make use of the open source software DAISY as has recently been suggested by Browning et al. (2020), and finally we also study simulations of both model types.

5.1 Transformed models

We can reformulate the differential equations for both model types by setting

$$\boldsymbol{Z}(t) = \begin{pmatrix} Z_1(t) \\ Z_2(t) \end{pmatrix} := \begin{pmatrix} \frac{X_1(t)}{m_0} \\ \text{scale} \cdot X_2(t) + \text{offset} \end{pmatrix},$$

which means that

$$\boldsymbol{Z}(t) \equiv \begin{pmatrix} 0 \\ \text{offset} \end{pmatrix} \text{ for } t \leq t_0, \quad \text{and } \boldsymbol{Z}(t_0) = \begin{pmatrix} 1 \\ \text{offset} \end{pmatrix}.$$

Hence, the second component of the transformed process models the fluorescence signal which we assume to be observed.

Transformed ODE model

For the ODE model in Equation (1), we obtain the transformed model

$$\frac{\mathrm{d}\boldsymbol{Z}(t)}{\mathrm{d}t} = \begin{pmatrix} -\theta_1 Z_1(t) \\ \text{scale}\, \theta_2 m_0 Z_1(t) - \theta_3 (Z_2(t) - \text{offset}) \end{pmatrix} \text{ for } t \geq t_0, \tag{4}$$

and the corresponding solution

$$Z_1(t) = \exp\left(-\theta_1 (t - t_0)\right),$$

$$Z_2(t) = \begin{cases} \frac{\text{scale}\, \theta_2 m_0}{\theta_3 - \theta_1} \left(e^{-\theta_1 (t-t_0)} - e^{-\theta_3 (t-t_0)}\right) + \text{offset} & \text{, for } \theta_1 \neq \theta_3, \\ \text{scale}\, \theta_2 m_0 (t - t_0) e^{-\theta_3 (t-t_0)} + \text{offset} & \text{, for } \theta_1 = \theta_3. \end{cases}$$

The parameters scale, m_0, and θ_2 appear only as a product. Thus, we can already deduce from this equation that at most the product of the three parameters will be identifiable but not the three parameters individually. Moreover, since only $Z_2(t)$ is observed and it is symmetric in the parameters θ_1 and θ_3 (i. e. switching their values will lead to the same model output), these two parameters can at most be locally identifiable.

Transformed SDE model

For the SDE model in Equation (3), we apply the Itô formula (as stated in Section A.1 of the supplementary material) to obtain the transformed model

$$\begin{aligned} \mathrm{d}\boldsymbol{Z}(t) = & \begin{pmatrix} -\theta_1 Z_1(t) \\ \text{scale}\, \theta_2 m_0 Z_1(t) - \theta_3 (Z_2(t) - \text{offset}) \end{pmatrix} \mathrm{d}t \\ & + \begin{pmatrix} \sqrt{\frac{\theta_1}{m_0} Z_1(t)} & 0 \\ 0 & \sqrt{\text{scale}} \sqrt{\text{scale}\, \theta_2 m_0 Z_1(t) + \theta_3 (Z_2(t) - \text{offset})} \end{pmatrix} \mathrm{d}B_t \text{ for } t \geq t_0. \end{aligned} \tag{5}$$

Note that here, the parameters scale and m_0 also appear outside the product $\text{scale}\, \theta_2 m_0$. Therefore, we hope to gain more information about the individual parameters from data for the SDE model than for the ODE model.

5.2 Using a surrogate model and existing software tools

The open source software DAISY (Differential Algebra for Identifiability of SYstems) was introduced by Bellu et al. (2007). It is a software tool that implements a differential algebra algorithm to perform structural identifiability analysis for systems of polynomial or rational ODEs and that also allows to

include unknown initial conditions. Mathematically, the problem translates into checking the solvability of a very large system of nonlinear algebraic equations. However, the use of the DAISY software does not require an in-depth understanding of the underlying theory.

Here, we want to use DAISY to assess the structural identifiability of the parameters in the two models of the translation kinetics. In order to include the parameters scale and offset, we use the transformed models from the previous subsection for the identifiability analysis. For the ODE model in Equation (4), the analysis with DAISY is straight forward since it is intended for the use for ODE models. After applying DAISY, the obtained output shows that when considering the set of parameters $\{\boldsymbol{\theta}, m_0, \mathrm{scale}, \mathrm{offset}\}$, the model is non-identifiable. The DAISY output also reveals that this non-identifiability is due to fact that the parameters θ_1 and θ_3 are only locally identifiable and the parameters θ_2, m_0, and scale are not individually identifiable, but only their product is. This confirms our assertions from the previous subsection. Moreover, we obtain that the remaining parameter offset is structurally identifiable.

For SDE models, Browning et al. (2020) suggest to formulate a surrogate model based on the moment equations of the diffusion process. The moment equations are a system of ODEs, and thus, DAISY can be applied to this system. For the SDE (5), let $m_{ij}(t) = \mathbb{E}\left[(Z_1(t))^i (Z_2(t))^j\right]$ be the (mixed) moment of the diffusion process of order i and j. The moments are obtained by applying the Itô formula (as stated in Section A.1 of the supplementary material) to $(Z_1(t))^i (Z_2(t))^j$ and then taking the expectation. Considering the first and the second moments of the process states results in the following system of ODEs:

$$
\begin{aligned}
\frac{\mathrm{d}m_{10}(t)}{\mathrm{d}t} &= -\theta_1 m_{10}(t), && m_{10}(t_0) = 1,\\
\frac{\mathrm{d}m_{01}(t)}{\mathrm{d}t} &= \mathrm{scale}\,\theta_2 m_0\, m_{10}(t) - \theta_3\, m_{01}(t) + \theta_3 \mathrm{offset}, && m_{01}(t_0) = \mathrm{offset},\\
\frac{\mathrm{d}m_{20}(t)}{\mathrm{d}t} &= \frac{\theta_1}{m_0}\, m_{10}(t) - 2\theta_1\, m_{20}(t), && m_{20}(t_0) = 1,\\
\frac{\mathrm{d}m_{02}(t)}{\mathrm{d}t} &= \mathrm{scale}^2 \theta_2 m_0\, m_{10}(t) + \theta_3(\mathrm{scale} + 2\mathrm{offset})\, m_{01}(t) - 2\theta_3\, m_{02}(t)\\
&\quad + 2\mathrm{scale}\,\theta_2 m_0\, m_{11}(t) - \mathrm{scale}\,\theta_3 \mathrm{offset}, && m_{02}(t_0) = \mathrm{offset}^2,\\
\frac{\mathrm{d}m_{11}(t)}{\mathrm{d}t} &= \theta_3 \mathrm{offset}\, m_{10}(t) + \mathrm{scale}\,\theta_2 m_0\, m_{20}(t) - (\theta_1 + \theta_3)\, m_{11}(t), && m_{11}(t_0) = \mathrm{offset},
\end{aligned}
$$

where the equations for the two first moments m_{10} and m_{01} coincide with the ODE model in Equation (4). Since in the experiment, only the fluorescence signal is observed, we consider the moments that only depend on the second component of the process, i. e. m_{01} and m_{02}, as output states for the identifiability analysis. Using DAISY, we obtain that the surrogate model is globally identifiable, i. e. all six parameter values could be uniquely determined if we were able to observe the moments m_{01} and m_{02} directly, infinitely long over time, and without measurement error. However, this property of structural identifiability (in particular when using a surrogate model) is only a necessary, but not a sufficient condition for practical identifiability. From this result, we cannot conclude that the parameters will be identifiable from the actual experimental data.

5.3 Simulating from the models

Another attempt to assess parameter identifiability is to simulate from both model types for different parameter settings and compare whether we see differences in the simulation output. To obtain simulations from the ODE model, we use its solution in Equation 2. Since the ODE model is deterministic, each parameter setting yields one unique output trajectory while for the SDE model, we simulate several trajectories for each parameter setting using the Euler-Maruyama scheme with a time step of 0.01 hours.

Keeping the product $\mathrm{scale}\,\theta_2 m_0$ constant

As already pointed out in Subsection 5.1, the trajectories of the fluorescence intensity for the ODE model are identical if the product $\mathrm{scale}\,\theta_2 m_0$ and the remaining parameters are fixed, even when the individual factors scale, θ_2, and m_0 vary. Here, we use (approximately) the mean values for the parameters estimated from the data in Fröhlich et al. (2018), and therefore, set $\mathrm{scale}\,\theta_2 m_0 = 350$,

$\theta_1 = 0.11$, $\theta_3 = 0.03$, $\text{offset} = 8.9$, and $t_0 = 0$. For the SDE model, we simulate several trajectories with different values for scale, θ_2, and m_0 while keeping their product constant. For each parameter setting, we set the same random seed at the beginning of the simulation. Figure 2 displays the simulated trajectories.

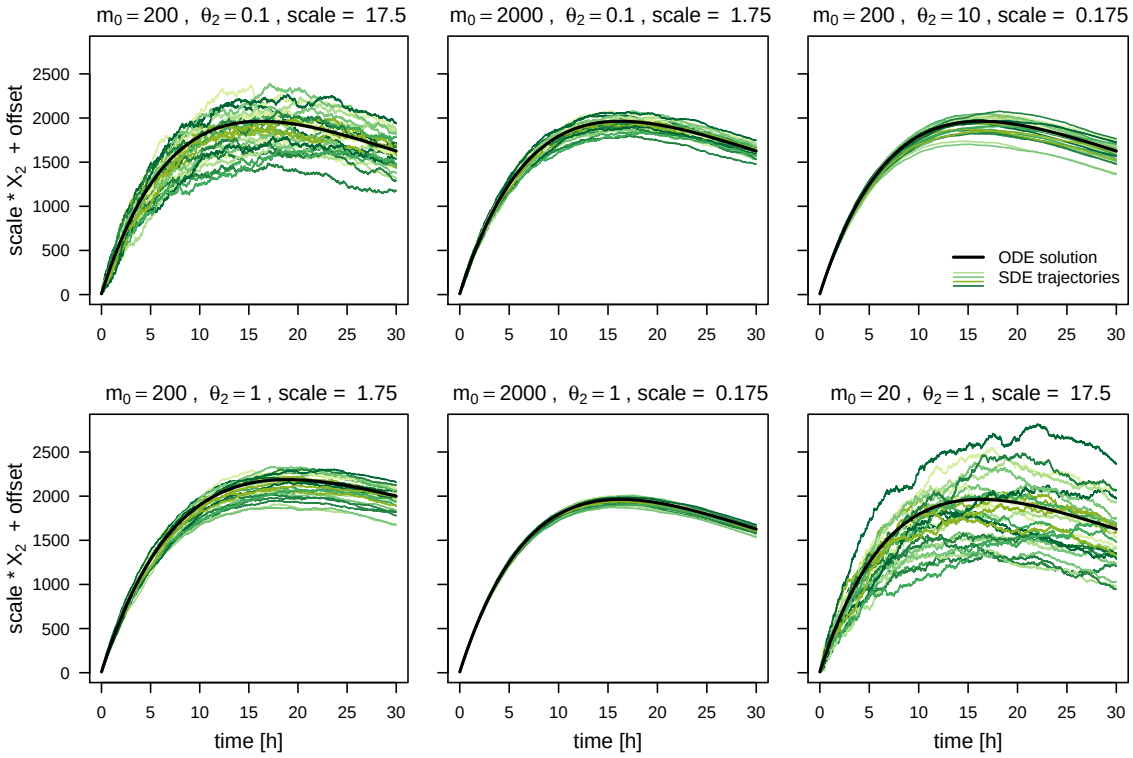


Figure 2: The ODE trajectory and 20 SDE trajectories of the fluorescence intensity simulated for different values of m_0, θ_2, and scale, while keeping their product constant at $\text{scale}\,\theta_2 m_0 = 350$. The remaining parameters are set to $\theta_1 = 0.11$, $\theta_3 = 0.03$, $\text{offset} = 8.9$, and $t_0 = 0$.

It is evident that the SDE trajectories behave differently for different combinations of scale, θ_2, and m_0. For example, when we keep m_0 fixed while increasing scale and decreasing θ_2, the variation between but also within the trajectories increases. When we keep scale fixed while decreasing m_0 and increasing θ_2, especially the variation between trajectories seems to increase. And finally, when we keep θ_2 fixed while decreasing m_0 and increasing scale, the variation between and within the trajectories increases. Our focus is on estimating the parameters from individual observed trajectories. In this context, especially the difference in the variation within the trajectories is relevant.

Swapping the degradation rate constants θ_1 and θ_3

The trajectories of the fluorescence intensity for the ODE model also are identical if the values for θ_1 and θ_3 are swapped while the remaining parameters are fixed. We simulate trajectories for the parameter combinations $(\theta_1, \theta_3) = (0.11, 0.03)$ and $(\theta_1^*, \theta_3^*) = (0.03, 0.11)$, respectively, while setting the remaining parameters to $\text{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\text{offset} = 8.9$, and $t_0 = 0$. For the SDE model, we again simulate several trajectories for both parameter settings and set the same random seed at the beginning of the simulation.

Figure 3 shows the ODE trajectory and several SDE trajectories in one panel for each of the two parameter combinations separately. Whereas, Figure 4 presents one SDE trajectory for each of the two parameter combinations together in one panel. Again, the SDE trajectories do behave differently for the different parameter combinations. While there seems to be only little difference in the variation between the trajectories, the variation within the trajectories is clearly higher for lower θ_1 and higher θ_3. This indicates that it may be possible to uniquely determine the values of θ_1 and θ_3 even when estimating from only one observed trajectory.

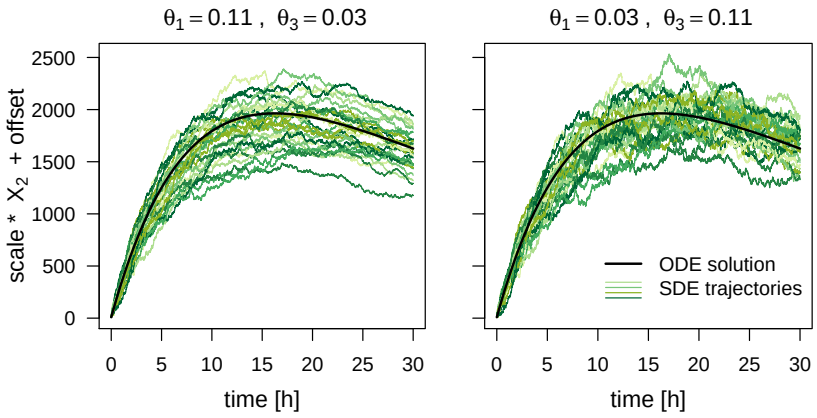


Figure 3: The ODE trajectory and 20 SDE trajectories of the fluorescence intensity simulated for two parameter combinations where the values of θ_1 and θ_3 are swapped. The remaining parameters are set to $\text{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\text{offset} = 8.9$, and $t_0 = 0$.

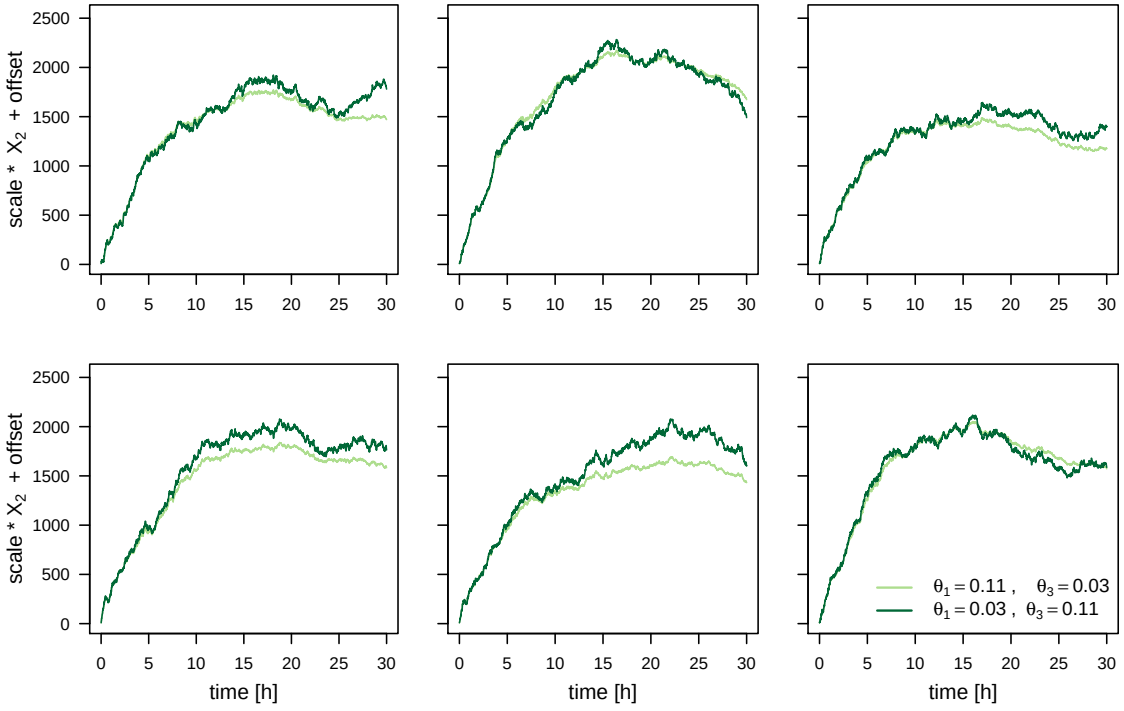


Figure 4: One trajectory of the fluorescence intensity for the SDE model simulated for each of the two parameter combinations where the values of θ_1 and θ_3 are swapped. The remaining parameters are set to $\text{scale} = 17.5$, $\theta_2 = 0.1$, $m_0 = 200$, $\text{offset} = 8.9$, and $t_0 = 0$.

6 Posterior properties and credibility intervals

After having studied the structural properties of the ODE and SDE model in the previous section; next, we would like to assess the practical parameter identifiability by trying to estimate the parameters from data as described in Section 4. We take a Bayesian approach to parameter estimation because it allows for uncertainty assessment of the parameter estimates and also for handling unobserved process components and measurement error by using Markov chain Monte Carlo (MCMC) methods to sample from the parameter posterior distribution. Therefore, in this section, we define the parameter posterior densities for the two model types and study their properties based on MCMC sampling results for simulated data generated without and with (simulated) measurement error.

6.1 Formulation of the inference problem

In Bayesian statistics, we can formulate our assumptions and general knowledge about the model parameters $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^p$ in terms of a prior distribution with probability density $p(\boldsymbol{\xi})$. After having observed data $\mathcal{D}$ about the phenomenon which we are trying to model, we update our knowledge about the parameter and describe it by the posterior distribution with density $\pi(\boldsymbol{\xi} \mid \mathcal{D})$. In our case, we consider data $\mathcal{D} = \{y_k\}_{k=1,\dots,K}$ and the vector of all unknown parameters is $\boldsymbol{\xi} = (\boldsymbol{\theta}, m_0, \mathrm{scale}, \mathrm{offset}, t_0, \sigma)$. The relation between the prior and the posterior density resulting from Bayes' theorem is given by:

$$\pi(\boldsymbol{\xi} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\xi})\, p(\boldsymbol{\xi}),$$

where $p(\mathcal{D} \mid \boldsymbol{\xi})$ denotes the density of the distribution of $\mathcal{D}$ conditioned on $\boldsymbol{\xi}$ and is determined by the considered model. Viewed as a function of the parameter, $\mathcal{L}(\boldsymbol{\xi} \mid \mathcal{D}) := p(\mathcal{D} \mid \boldsymbol{\xi})$ is called the likelihood (function). Comprehensive introductions to Bayesian statistics can be found e. g. in Lee (2012) and Gelman et al. (2013).

To sample from the subsequently formulated posterior densities for the two model types, we use the open source software Stan (Carpenter et al., 2017) which provides an efficient C++ implementation of the Hamiltonian Monte Carlo (HMC) based No-U-Turn Sampler (NUTS). See Section A.3 of the supplementary material for a brief introduction to this topic. We use Stan though its R interface `rstan` (Stan Development Team, 2019).

Posterior of the ODE model

For the ODE model, there is a deterministic relationship between the process values $\boldsymbol{X}(t)$ and the parameters $\boldsymbol{\theta}$, m_0 and t_0 (or between the fluorescence signal and the parameters additionally including scale and offset, respectively).

Define the index $k^* := \min\{k \in \{1, \dots, K\} | t_k \geq t_0\}$ of the first observation time point after the mRNA molecules are released, then the posterior density π from which we would like to sample is proportional to

$$\begin{aligned}
&\pi\left(\theta, m_0, \mathrm{scale}, \mathrm{offset}, \sigma^2, t_0 | \{y_k\}_{k=1,\dots,K}\right) \\
&\quad \propto \left(\prod_{k=k^*}^{K} \phi\left(\log(y_k) \,\middle|\, \log\left(\mathrm{scale} \frac{\theta_2 m_0}{\theta_3 - \theta_1} \left(e^{-\theta_1 (t - t_0)} - e^{-\theta_3 (t - t_0)} \right) + \mathrm{offset} \right), \sigma^2 \right) \right) \\
&\qquad \cdot \left(\prod_{k=1}^{k^*-1} \phi\left(\log(y_k) \,\middle|\, \log(\mathrm{offset}), \sigma^2 \right) \right) \\
&\qquad \cdot p(\theta_1) p(\theta_2) p(\theta_3) p(m_0) p(t_0) p(\mathrm{scale}) p(\mathrm{offset}) p(\sigma^2), \qquad (6)
\end{aligned}$$

where $\phi(\cdot | \mu, \eta^2)$ denotes the density of the normal distribution with mean μ and variance η^2 and the $p(\cdot)$ denote the parameter prior densities which we assume to be independent.

If the priors $p(\theta_1)$ and $p(\theta_3)$ are symmetric to each other, then the posterior is also symmetric with respect to the two degradation rate constants.

The scaling factor scale, the translation rate constant θ_2, and the initial amount of mRNA m_0 appear only as a product in the likelihood function; therefore, as pointed out before, at most their product $\mathrm{scale}\theta_2 m_0$ is identifiable.

Posterior of the SDE model

For the SDE model, the states $\boldsymbol{X}(t_k)$, for $k = 1, \dots, K$, of the process conditioned on the parameters $\boldsymbol{\theta}$, m_0 and t_0 are random numbers (for $t_k \geq t_0$). Hence, we have to marginalize over the process states to obtain the posterior density of the parameters which we want to infer:

$$\begin{aligned}
&\pi\left(\theta, m_0, \mathrm{scale}, \mathrm{offset}, \sigma^2, t_0 | \{y_k\}_{k=1,\dots,K}\right) \\
&\quad = \int_{\mathbb{R}_+^{2 \times K}} \pi\left(\theta, m_0, \mathrm{scale}, \mathrm{offset}, \sigma^2, t_0, \{\boldsymbol{X}(t_k)\}_{k=1,\dots,K} | \{y_k\}_{k=1,\dots,K}\right) \mathrm{d}\boldsymbol{X}(t_1) \dots \mathrm{d}\boldsymbol{X}(t_K).
\end{aligned}$$

Therefore, again defining $k^* := \min\{k \in \{1, \dots, K\} | t_k \geq t_0\}$, we would need to sample from

$$\begin{aligned}
&\pi\left(\theta, m_0, \text{scale}, \text{offset}, \sigma^2, t_0, \{\boldsymbol{X}(t_k)\}_{k=1,\dots,K} | \{y_k\}_{k=1,\dots,K}\right) \\
&\quad \propto \left(\prod_{k=1}^{K} \phi\left(\log(y_k) | \log\left(\text{scale} \cdot X_2(t_k) + \text{offset}\right), \sigma^2\right)\right) \\
&\qquad \cdot \left(\prod_{k=k^*}^{K-1} \pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \theta\right)\right) \pi\left(\boldsymbol{X}(t_{k^*}) | \theta, m_0, t_0\right) \left(\prod_{k=1}^{k^*-1} \delta(\|\boldsymbol{X}(t_k) - (0,0)^T\|)\right) \\
&\qquad \cdot p(\theta) p(m_0) p(t_0) p(\text{scale}) p(\text{offset}) p(\sigma^2),
\end{aligned}$$

where $\phi(\cdot|\mu, \eta^2)$ denotes the density of the normal distribution with mean μ and variance η^2, $\delta(\cdot)$ denotes the Dirac delta function, $\|\cdot\|$ denotes a norm (e. g. the l_2-norm), and the factors $\pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \theta\right)$, $k = k^*, \dots, K-1$, denote the transition probability density of the process. However, the fact that the process $\boldsymbol{X}$ switches from a deterministic regime before t_0 to a stochastic one after t_0 complicates the estimation of t_0 together with the remaining parameters. Therefore, we will assume that t_0 is determined beforehand, e. g. based on the estimates for the ODE model. Consequently, we sample from

$$\begin{aligned}
&\pi\left(\theta, m_0, \text{scale}, \text{offset}, \sigma^2, \{\boldsymbol{X}(t_k)\}_{k=1,\dots,K} | \{y_k\}_{k=1,\dots,K}, t_0\right) \\
&\quad \propto \left(\prod_{k=1}^{K} \phi\left(\log(y_k) | \log\left(\text{scale} \cdot X_2(t_k) + \text{offset}\right), \sigma^2\right)\right) \\
&\qquad \cdot \left(\prod_{k=k^*}^{K-1} \pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \theta\right)\right) \pi\left(\boldsymbol{X}(t_{k^*}) | \theta, m_0, t_0\right) \left(\prod_{k=1}^{k^*-1} \delta(\|\boldsymbol{X}(t_k) - (0,0)^T\|)\right) \\
&\qquad \cdot p(\theta) p(m_0) p(t_0) p(\text{scale}) p(\text{offset}) p(\sigma^2). \qquad (7)
\end{aligned}$$

While for the ODE model, the posterior distribution is only 8-dimensional and can be sampled from directly; for the SDE model, we need to sample from a $(7 + 2K)$-dimensional distribution and then marginalize over the $2K$ dimensions of the process states to obtain the posterior distribution of the parameters of interest. Moreover, there is no explicit exact expression for the transition probability density $\pi\left(\boldsymbol{X}(t_{k+1}) | \boldsymbol{X}(t_k), \theta\right)$; wherefore, it will be approximated by a normal density based on the Euler-Maruyama scheme as explained in Section A.1 of the supplementary material. For this approximation to be appropriate, we have to ensure that the time steps between observations are small enough, we do so in Section A.2 of the supplementary material.

6.2 Estimation based on simulated data

In order to assess how well we can recover the model parameters for both model types from individually observed trajectories, we first work with simulated data that is generated with Gillespie's algorithm (Gillespie, 1976, 1977).

6.2.1 Sampling results for simulated data without measurement error

For now, we assume the fluorescence intensity to be observed without measurement error. The data was simulated with Gillespie's algorithm with parameters $\theta = (0.2, 0.32, 0.01)$, $m_0 = 240$, $t_0 = 0.96$, $\text{scale} = 1.8$, and $\text{offset} = 8.5$ over a time interval of 30 hours, and we use the fluorescence intensity at 181 equidistant time points as observations to mirror the structure of the experimental data. The simulated fluorescence intensity (without measurement error) is depicted by the blue dotted line on the right hand side of Figure 5. We use Stan to sample from the posterior distributions of the ODE model and the SDE model given the simulated data. Since we assume the data to be observed without measurement error, the parameter offset can be determined directly from the first observation. Therefore, we do not include measurement error (and thus the parameter σ) and the parameter offset in the posterior distribution of the SDE model. Whereas for the ODE model, deviations of the observed data from the deterministic ODE trajectory have to be attributed to measurement error; therefore,

the parameter σ has to be included in the posterior distribution of the ODE model. We also include the parameter offset for the ODE model in order to avoid degeneracy of the posterior. We use the following prior distributions: $\theta_i \sim \mathcal{N}_{\geq 0}(0, 5^2)$ for $i = 1, 2, 3$, $m_0 \sim \mathcal{N}_{\geq 0}(300, 300^2)$, $\mathrm{scale} \sim \mathcal{U}(0, 30)$, where $\mathcal{N}_{\geq a}(\mu, \eta^2)$ denotes the normal distribution truncated from below by a, and additionally for the ODE model, $\mathrm{offset} \sim \mathcal{U}(0, 30)$, $\sigma \sim \mathcal{U}(0.001, 10)$, and $t_0 \sim \mathcal{U}(0, 30)$.

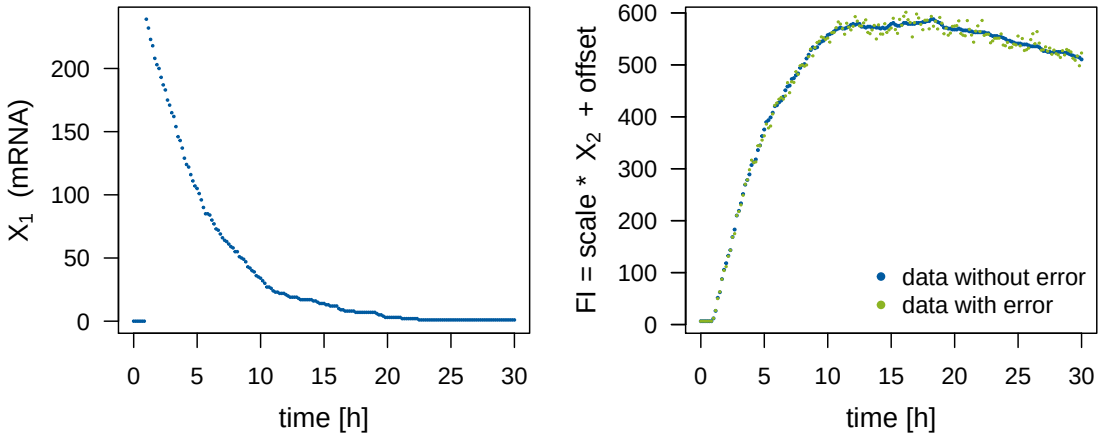


Figure 5: One trajectory used in the simulation study that was simulated with Gillespie's algorithm with parameters $\theta = (0.2, 0.32, 0.01)$, $m_0 = 240$, $t_0 = 0.96$, $\mathrm{scale} = 1.8$, and $\mathrm{offset} = 8.5$, and for the green dotted line, multiplicative measurement error with $\sigma = 0.02$ was added to the fluorescence intensity (FI).

For both model types, we generate 8 HMC chains of 5000 iterations and discard the first half of the iterations as warm-up. Thus, we use a posterior sample of size 20,000 for each model type in the subsequent analysis. Tables 1 and 2 summarize the Stan output of the posterior samples for the ODE and the SDE model, respectively, and also include the true parameter values that were used to simulate the data for comparison. The tables also contain the 2.5%-, 50%-, and 97.5%-quantiles of the samples. We use the interval between the 2.5%- and the 97.5%-quantile as an estimate of the 95%-credible interval (CI). For the ODE model, we see that the parameters offset and t_0 are well estimated since mean and median of the sample correspond to the true value, the CIs are very narrow, the effective sample size (ESS) n_{eff} is high and $\hat{R}$ is equal to 1. As expected, the measurement error parameter σ is estimated to be higher than the true value of zero. Of greater interest are the remaining parameters as we can compare the results for them between the two model types.

We first focus on the two degradation rate constants θ_1 and θ_3. Our analysis in Section 5 already showed that for the ODE model, these two parameters are only locally identifiable and the posterior distribution is symmetric with respect to them in the case of identical priors for both parameters. This is also apparent in the density plots in Figure 6. The density estimates of the posterior sample for the ODE model are clearly bimodal. The reason that the two modes are not exactly symmetric here is that HMC chains usually are only able to explore one mode and in our example 5 out of the 8 chains happen to end up in the mode where θ_1 is higher than θ_3 while only 3 chains converge to the other mode. The fact that each chain only samples from one of the modes is also the reason for the extremely low ESSs and the very high values of $\hat{R}$ for θ_1 and θ_3 in Table 1. Moreover, note that neither of the modes and not even the ranges of all values in the posterior sample cover the true parameter values of θ_1 and θ_3. For the SDE model on the other hand, Figure 6 and Table 2 show that the posterior density is clearly unimodal with respect to θ_1 and θ_3, the 95% CI are narrow and cover the true parameter values, mean and median of the sample are close or equal to the true values, and high ESSs and $\hat{R}$ values equal to 1 are achieved. Thus, we can conclude that the parameters θ_1 and θ_3 are identifiable for the SDE model here.

We have simulated another 99 trajectories with the same parameters and performed Stan sampling in the same way as described in the beginning of this subsection. For each model type and each posterior sample of the different simulated trajectories, we calculate the length of the 95% CI and determine the median and the coefficient of variation (c.v.) over these lengths for each model type. Also, we rescale the lengths of the 95% CI by dividing by the true parameter value and again determine the median

Table 1: Summary of the Stan output for the ODE model given simulated data without measurement error and the true parameter values that were used to simulate the data. c.v. denotes the coefficient of variation and the columns headed by percentages contain the quantiles of the respective percentage value.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.11	0.634	0.02	0.16	0.17	4	26.65
θ_2	0.32	1.52	1.370	0.02	0.64	7.56	12168	1.00
θ_3	0.01	0.07	0.943	0.02	0.02	0.16	4	33.00
m_0	240.00	204.62	1.017	2.26	135.79	724.38	10984	1.00
scale	1.80	7.02	1.137	0.07	3.46	27.28	9806	1.00
offset	6.50	6.50	0.011	6.37	6.50	6.64	17113	1.00
t_0	0.96	0.96	0.002	0.96	0.96	0.97	18718	1.00
σ	0.00	0.03	0.054	0.02	0.03	0.03	16091	1.00
$\theta_2 m_0$	76.80	213.08	2.487	4.57	35.99	1668.06	11087	1.00
$\theta_2\text{scale}$	0.58	6.46	2.875	0.17	0.92	55.00	7704	1.00
$m_0\text{scale}$	432.00	1033.30	2.181	16.48	195.96	7975.22	7899	1.00
$\theta_2 m_0\text{scale}$	138.24	124.67	0.007	122.96	124.66	126.38	13299	1.00

Table 2: Summary of the Stan output for the SDE model given simulated data without measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.19	0.108	0.15	0.19	0.23	3120	1.00
θ_2	0.32	0.39	0.999	0.09	0.26	1.48	206	1.04
θ_3	0.01	0.01	0.167	0.01	0.01	0.02	2514	1.00
m_0	240.00	344.37	0.589	57.21	313.25	800.67	184	1.05
scale	1.80	1.66	0.172	1.19	1.62	2.30	3062	1.00
$\theta_2 m_0$	76.80	82.60	0.178	55.68	81.69	113.53	5296	1.00
$\theta_2\text{scale}$	0.58	0.61	0.923	0.17	0.43	2.21	178	1.05
$m_0\text{scale}$	432.00	576.69	0.634	86.08	511.66	1440.41	232	1.04
$\theta_2 m_0\text{scale}$	138.24	133.18	0.083	112.47	132.74	156.40	5829	1.00

of the normalized quantities. The rescaling is done to transfer the values to a more similar scale. Note, however, that the values are nevertheless not directly comparable between different parameters. Moreover, we check whether the true parameter value that was used to simulate the data is included in the 95% CI. Table 3 shows the aggregated results for the posterior samples of all 100 trajectories and also includes the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distributions. Except for the parameters m_0 and $\theta_2 m_0\text{scale}$, the median length of the 95% CIs for the SDE model is always smaller than for the ODE model. For parameter $\theta_2 m_0\text{scale}$, the CI lengths are a lot smaller for the ODE model; however, the CIs cover the true parameter value only 13 out of 100 times while for the SDE model, the true value is covered 93 times. For the other parameters that we classified as identifiable for the SDE model in the analysis of the individual trajectory (i. e. θ_1, θ_3, scale, and $\theta_2 m_0$), the median length of the 95% CIs is clearly smaller for the SDE model than for the ODE model and the true parameter value is covered at least 91 out 100 times for the SDE model. For parameter m_0, the CI lengths are high for both model types because the parameter is not identifiable for either model type. For the other parameters that we classified as not identifiable for both model types in the analysis of the individual trajectory (i. e. θ_2, $\theta_2\text{scale}$, and $m_0\text{scale}$), the median length of 95% CIs is clearly smaller for the SDE model than for the ODE model, at least by a factor of 4.

The last two columns of Table 3 are visualized in Figure 8 where we plot the median of the rescaled CI lengths against the number of CIs that cover the true parameter value. The desirable region of value combinations is in the bottom right corner of the graph where the number of CIs covering the true value is high and the median rescaled CI length is small. Note that, clearly, more importance should be given to high numbers of CIs covering the true value as it is useless to be very certain about a parameter

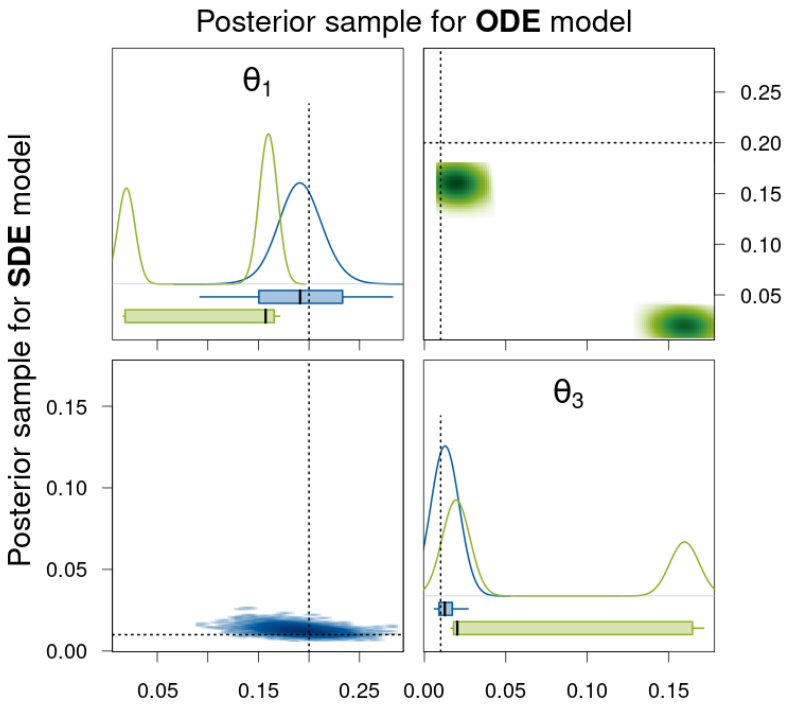


Figure 6: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data without measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.

estimate (indicated by a short CI) while the correct value is not included in the CI. However, even for parameters that are identifiable, we do not expect to obtain a coverage of the true value of 100% since we are considering 95% CIs. Therefore, values of 100 rather tend to hint at non-identifiability. In Figure 8, we can see that for the majority of the parameters, the triangles representing the value combinations for the SDE model are closer to the desirable region. Only for parameter m_0 (which is not identifiable for either model type), the value combinations are almost the same for both model types. And as we already pointed out for the product $\theta_2 m_0 \mathrm{scale}$, the median length of the for the ODE model; however, a lot fewer CIs cover the true parameter value for the ODE model than for the SDE model. Thus, the result obtained for the SDE model is to be preferred.

We provide further Stan-specific diagnostics in Appendix A.5. Those mostly show poorer values for the sampling output for the SDE model than for the ODE model. This is not surprising as we sample from a much higher-dimensional distribution for the SDE model. We do not consider the poor diagnostics as a disadvantage of the procedure as they provide information that we do not even have for other MCMC algorithms and thus cannot compare to them.

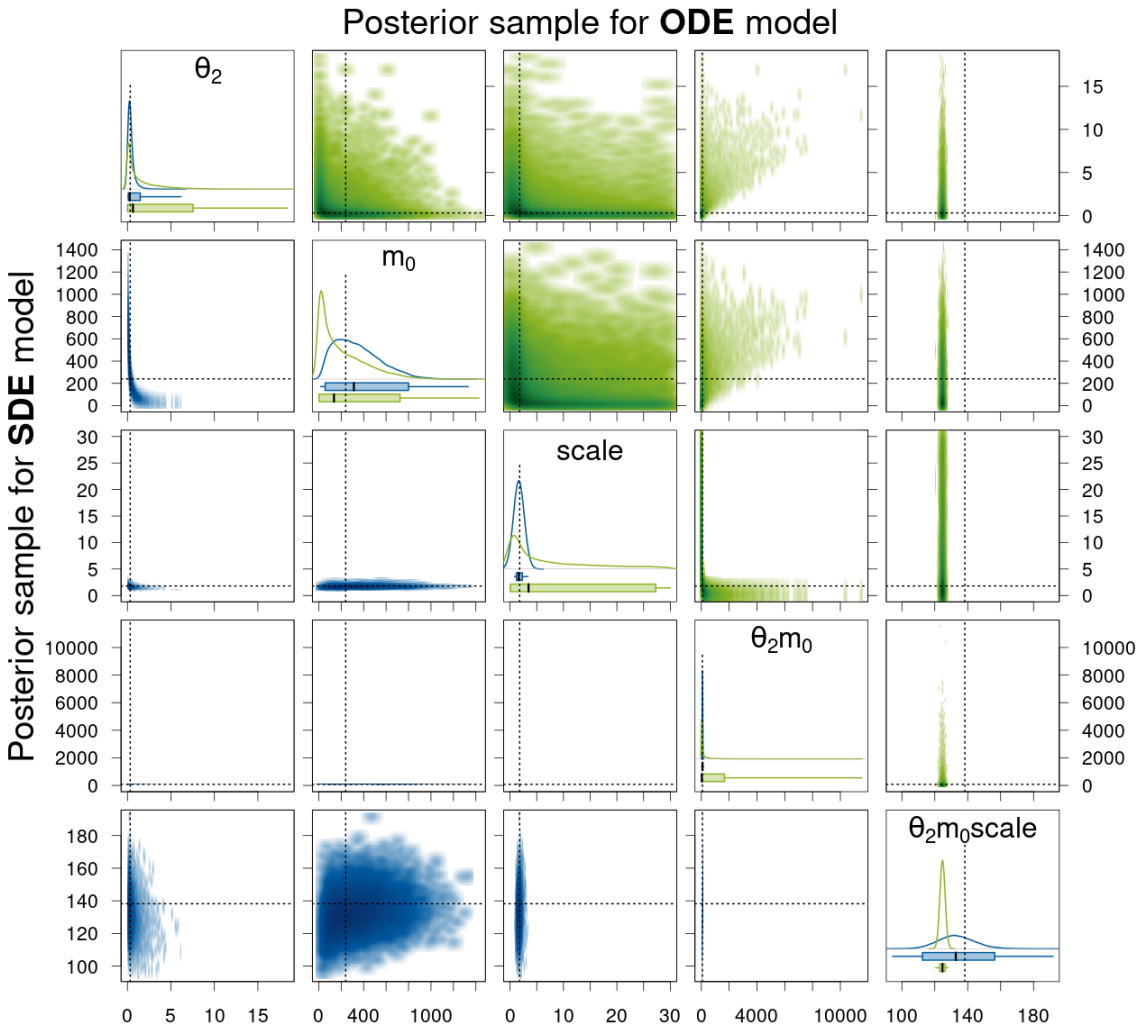


Figure 7: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data without measurement error. For a detailed description of the figure's elements, see Figure 6.

Table 3: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories without measurement error. We also include the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distribution.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs	median of length of CIs rescaled by true value	number of CIs covering true value
θ_1	ODE	11.05	0.20	0.009	1.01	58
	SDE	11.05	0.09	0.002	0.45	93
θ_2	ODE	11.05	7.56	0.002	23.63	100
	SDE	11.05	1.88	0.712	5.87	100
θ_3	ODE	11.05	0.20	0.010	20.00	61
	SDE	11.05	0.01	0.000	0.80	91
m_0	ODE	884.82	730.41	0.057	3.04	100
	SDE	884.82	735.98	9.868	3.07	100
scale	ODE	28.50	27.27	0.001	15.15	100
	SDE	28.50	1.20	0.158	0.67	95
$\theta_2 m_0$	ODE	6056.48	1701.92	2.524	22.16	100
	SDE	6056.48	63.60	2.768	0.83	96
$\theta_2 \mathrm{scale}$	ODE	228.08	55.47	0.164	96.29	100
	SDE	228.08	2.89	0.762	5.01	100
$m_0 \mathrm{scale}$	ODE	19271.13	7923.38	8.603	18.34	100
	SDE	19271.13	1315.63	86.806	3.05	99
$\theta_2 m_0 \mathrm{scale}$	ODE	113232.70	3.90	41.595	0.03	13
	SDE	113232.70	45.28	0.624	0.33	93

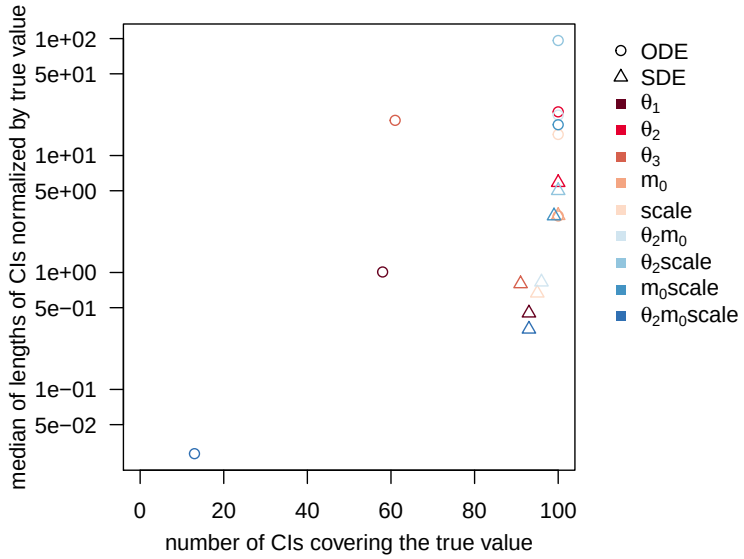


Figure 8: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories without measurement error. The desirable region of value combinations is in the bottom right corner of the graph.

6.2.2 Sampling results for simulated data with measurement error

In this section, we use the same simulated data as in the previous section, but for each of the 100 trajectories, we add multiplicative measurement error with parameter $\sigma = 0.02$. Again, we use Stan to sample from the posterior distributions of the ODE model (6) and the SDE model (7) for each of the simulated trajectories and use the same priors as stated in the previous section. We generate 8 HMC chains of 5000 iterations, discard the first half of the iterations as warm-up, and thus use a posterior samples of size 20,000 in the subsequent analysis.

At first, we again focus on the results for one of the trajectories, namely the trajectory represented by the green dotted line in Figure 5. Tables 4 and 5 summarize the Stan output of the posterior samples for the ODE and the SDE model, respectively. The parameter t_0 is estimated very accurately based on the posterior sample for the ODE model. Also, the parameter offset is well estimated for both model types but with a more narrow 95% CI for the SDE model. The parameter σ is accurately determined for the SDE model as well. For the ODE model, σ is again overestimated. Figure 9 visualizes the components of the posterior samples for parameters θ_1, θ_3, offset, and σ. Again, the bimodality of the posterior with respect to θ_1 and θ_3 is apparent for the ODE model and neither the 95% CIs nor the ranges of the sample cover the true parameter values. For the SDE model on the other hand, the distribution is unimodal and the 95% CIs do cover the true parameter values for θ_1 and θ_3. However, their 2-dimensional smoothed scatter plot in Figure 9 is not a simple elliptic shape (as for the simulated data without measurement error) but almost a banana-like shape. This may also be the reason for the deteriorated sampling efficiency discernible from the low ESS and higher $\hat{R}$-values in Table 5.

Table 4: Summary of the Stan output for the ODE model given simulated data with measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.11	0.632	0.02	0.15	0.17	4	20.94
θ_2	0.32	1.54	1.364	0.02	0.65	7.63	11619	1.00
θ_3	0.01	0.07	0.938	0.02	0.02	0.16	4	26.81
m_0	240.00	205.41	1.024	2.25	133.96	738.49	10649	1.00
scale	1.80	6.84	1.152	0.07	3.29	27.06	8778	1.00
offset	6.50	6.50	0.013	6.34	6.50	6.67	14496	1.00
t_0	0.96	0.96	0.003	0.96	0.96	0.97	17129	1.00
σ	0.02	0.03	0.053	0.03	0.03	0.04	13581	1.00
$\theta_2 m_0$	76.80	217.59	2.441	4.56	37.40	1683.06	9479	1.00
$\theta_2\mathrm{scale}$	0.58	6.31	2.841	0.17	0.92	54.94	7209	1.00
$m_0\mathrm{scale}$	432.00	983.07	2.151	16.20	189.75	7514.28	8054	1.00
$\theta_2 m_0\mathrm{scale}$	138.24	123.47	0.009	121.43	123.46	125.58	12035	1.00

Table 5: Summary of the Stan output for the SDE model given simulated data with measurement error and the true parameter values that were used to simulate the data.

	true value	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.16	0.201	0.08	0.17	0.22	304	1.03
θ_2	0.32	0.73	1.448	0.07	0.37	3.93	296	1.02
θ_3	0.01	0.02	0.660	0.01	0.02	0.05	176	1.04
m_0	240.00	274.72	0.742	28.30	224.46	777.96	225	1.04
scale	1.80	1.67	0.430	0.65	1.54	3.36	415	1.02
offset	6.50	6.50	0.007	6.41	6.50	6.60	15599	1.00
σ	0.02	0.02	0.069	0.02	0.02	0.02	2106	1.00
$\theta_2 m_0$	76.80	89.78	0.473	35.58	80.76	191.24	349	1.02
$\theta_2\mathrm{scale}$	0.58	0.93	1.145	0.16	0.56	4.15	247	1.03
$m_0\mathrm{scale}$	432.00	477.62	0.944	29.55	338.17	1697.86	216	1.04
$\theta_2 m_0\mathrm{scale}$	138.24	124.39	0.093	102.85	123.92	148.63	2119	1.01

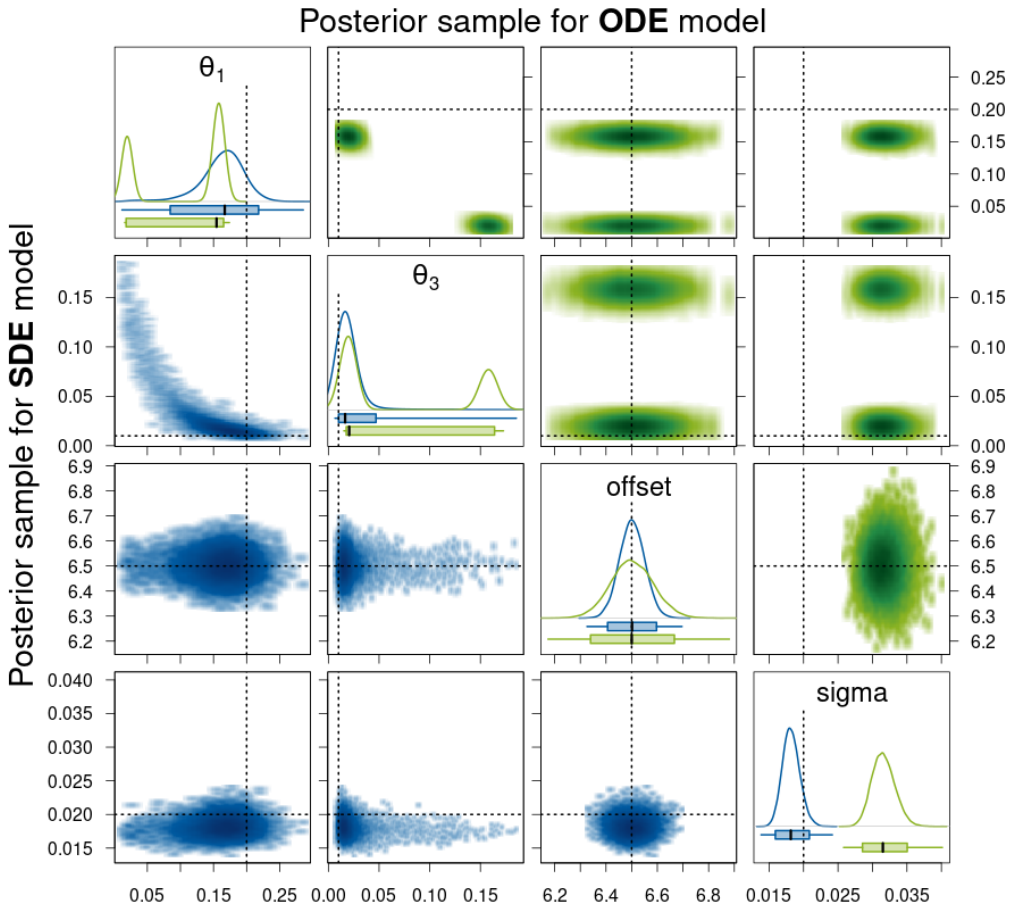


Figure 9: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data with measurement error. For a detailed description of the figure's elements, see Figure 6.

Figure 9 visualizes the components of the posterior samples for parameters θ_2, m_0, scale, and their products. For the ODE model, again only the product $\theta_2 m_0 \mathrm{scale}$ is identifiable in the sense that the corresponding 95% CI is very narrow, the ESS is high, and the $\hat{R}$-value is equal to 1. However, the 95% CI again does not cover the true parameter value. For the SDE model, the 95% CI for $\theta_2 m_0 \mathrm{scale}$ is broader but it does contain the true value. Also, the ESS is high and the $\hat{R}$-value is close to 1. Moreover, the parameters scale and $\theta_2 m_0$ have narrow 95% CIs, high ESSs, and $\hat{R}$-values close to 1 for the SDE model, and thus, we conclude that they are identifiable. Note that also for θ_2, $m_0 \mathrm{scale}$, and $\theta_2 \mathrm{scale}$, the 95% CIs are much narrower for the SDE model than for the ODE model.

In Appendix A.4.1, we include further figures of the sampling output for the trajectory displayed in Figure 5. They present the same posterior samples as used in this and the previous subsection. But instead of comparing the posterior samples between the two model types, the posterior samples are compared between the simulated data without and with measurement error for each model type separately. In summary, we find that for the SDE model, the 95% CIs increase for almost all parameters except m_0 for data with measurement error. Whereas for the ODE model, there is hardly any difference for most of the parameters between the posterior samples for the data without and with measurement error since the majority of the parameters is not identifiable anyway. The marginal posterior samples for the parameters offset, t_0, and $\theta_2 m_0 \mathrm{scale}$ are only slightly affected by the measurement error. Only the marginal posterior sample of the measurement error parameter σ is substantially affected and, as expected, consists of higher values for data with measurement error.

Table 6 and Figure 11 display the statistics of the posterior samples aggregated over the 100 simulated trajectories. Similar to the results for the simulated data without measurement error, the median length of the 95% CIs for the SDE model is always smaller than for the ODE model, except for the parameters m_0 and $\theta_2 m_0 \mathrm{scale}$ and additionally σ (which was not included for the SDE model

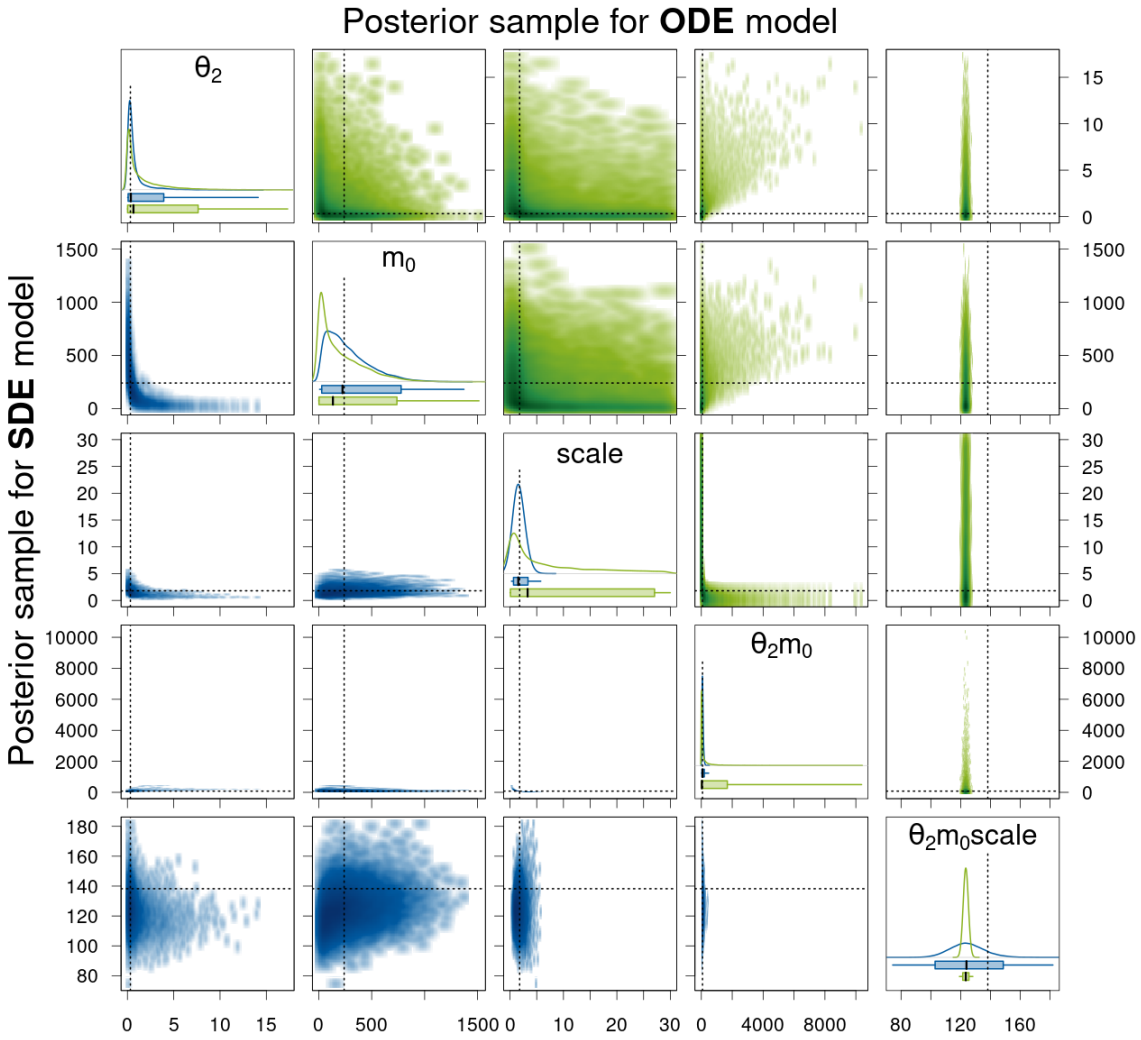


Figure 10: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given simulated data with measurement error. For a detailed description of the figure's elements, see Figure 6.

in the previous subsection). Again, for the majority of the parameters, the results for the SDE model represented by triangles in Figure 11 are closer to the desirable region of value combinations in the bottom right corner of the graph, except for the parameters m_0, $\theta_2 m_0 \mathrm{scale}$, and offset. For the parameter offset, the median CI length is slightly higher for the ODE model than for the SDE model, however, the CIs for the ODE model also contain the true parameter value more often. So for this parameter, the ODE model, for once, shows the preferable result.

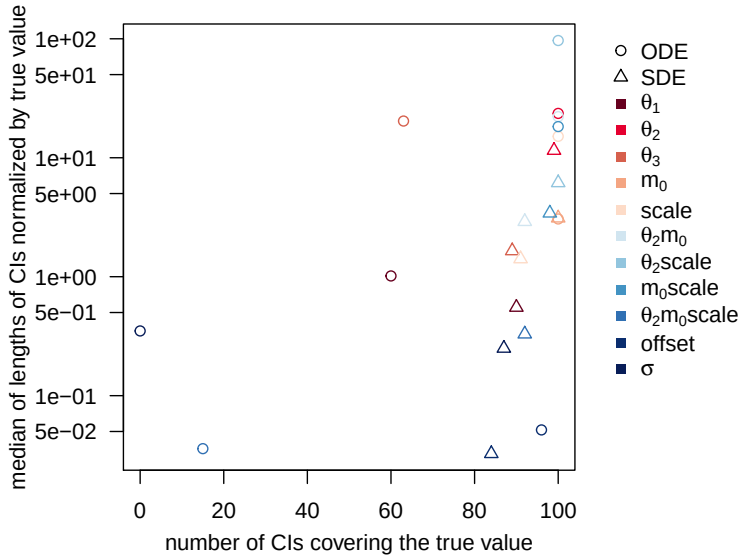


Figure 11: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories with measurement error. The desirable region of value combinations is in the bottom right corner of the graph.

Table 6: Statistics of posterior samples for the two model types aggregated over 100 simulated trajectories with measurement error. We also include the length of the interval between the 2.5%- and the 97.5%-quantile of the prior distribution.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs	median of length of CIs rescaled by true value	number of CIs covering true value
θ_1	ODE	11.05	0.20	0.008	1.01	60
	SDE	11.05	0.11	0.016	0.55	90
θ_2	ODE	11.05	7.55	0.002	23.59	100
	SDE	11.05	3.69	0.996	11.52	99
θ_3	ODE	11.05	0.20	0.008	20.35	63
	SDE	11.05	0.02	0.094	1.65	89
m_0	ODE	884.82	733.85	3.584	3.06	100
	SDE	884.82	746.74	5.909	3.11	100
scale	ODE	28.50	27.25	0.002	15.14	100
	SDE	28.50	2.54	0.255	1.41	91
$\theta_2 m_0$	ODE	6056.48	1702.37	2.714	22.17	100
	SDE	6056.48	223.22	428.350	2.91	92
$\theta_2 \mathrm{scale}$	ODE	228.08	55.70	0.125	96.70	100
	SDE	228.08	3.55	0.639	6.16	100
$m_0 \mathrm{scale}$	ODE	19271.13	7896.07	667.494	18.28	100
	SDE	19271.13	1479.13	253.530	3.42	98
$\theta_2 m_0 \mathrm{scale}$	ODE	113232.70	4.96	38.038	0.04	15
	SDE	113232.70	45.56	1.492	0.33	92
offset	ODE	28.50	0.33	0.987	0.05	96
	SDE	28.50	0.21	0.016	0.03	84
σ	ODE	9.50	0.01	0.325	0.35	0
	SDE	9.50	0.01	0.000	0.25	87

6.3 Estimation based on experimental data

In this section, we use the experimental data published in Fröhlich et al. (2018) and described in Section 2. For each type of GFP (eGFP and d2eGFP), we randomly select 100 observed trajectories for our analysis and again use Stan to sample from the posterior distributions of the ODE model (6) and the SDE model (7) for each of the trajectories using the same priors as stated in Section 6.2.1. We generate 8 HMC chains of 5000 iterations, discard the first half of the iterations as warm-up, and thus use a posterior samples of size 20,000 in the subsequent analysis. For each type of GFP, we first analyze the sampling output for one observed trajectory in detail and then summarize results for all 100 observed trajectories. Here in the main part of the article, we only show the results for the eGFP data and provide the results for the d2eGFP data in Appendix A.4.2. Moreover, we provide further Stan-specific diagnostics in Appendix A.5.

Sampling results for experimental dataset 1 (for eGFP)

Tables 7 and 8 present a summary of the Stan output for the posterior sample of one observed trajectory for the ODE and the SDE model, respectively, and Figures 12 and 13 compare the density estimates of these two posterior samples. The results look qualitatively *very* similar (almost identical) to those obtained for the simulated data with measurement error in Section 6.2.2. Therefore, we do not repeat the detailed description but only point out that the range of values sampled for the parameters θ_1 and θ_3 for the SDE model is slightly smaller for the experimental trajectory here. Thus, we do not see the banana-like shape in the two-dimensional smoothed scatter plot of the two parameters for the SDE model in Figure 12 as for the simulated trajectory in Figure 9 and the sampling efficiency increases as indicated by higher ESSs and lower $\hat{R}$ values for the two parameters in Table 8.

The statistics of posterior samples aggregated for 100 experimental trajectories for eGFP in Table 9 are also qualitatively similar to those for the simulated trajectories in Table 6. For the majority of the parameters, the median length of the 95% CI is smaller for the posterior samples for the SDE model than for those for the ODE model. Only for parameters θ_1, θ_2, and $\theta_2 m_0 \mathrm{scale}$, this is not the case. Note in particular that for the parameters $\theta_2 m_0$ and scale, which are non-identifiable for the ODE model (also apparent from the very long CIs here), the median length of the 95% CI for the SDE model is again much narrower compared to that of the ODE and the to the prior. This indicates that these two parameters are identifiable for the SDE model also for the experimental data. That the uncertainty of the parameter estimate for $\theta_2 m_0 \mathrm{scale}$ is greater for the SDE than for the ODE model is consistent with our results for the simulated data. The parameter θ_2 is considered to be not identifiable for either model type and the difference between the median CI lengths is relatively small. Finally, for parameter θ_1, we see that the result is more or less the same as for θ_3 for the ODE model due to the symmetry of the posterior distribution with respect to these two parameters. Whereas for the SDE model there is no symmetry and there is more variance in the posterior samples with respect to θ_1 than to θ_3 (indicated by a greater median CI length). The smaller median CI length of θ_1 for the ODE model compared to the SDE model is due to the fact that for many of the observed trajectories the values of θ_1 and θ_3 seem to be very close together. In this case, the posterior distribution of the ODE model appears to be unimodal and the posterior variance with respect to the two parameters is small (and equal due to the symmetry). Thus, overall this variance is smaller than the posterior variance with respect to θ_1 for the SDE model.

Table 7: Summary of the Stan output for the ODE model given experimental data for eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.11	0.617	0.02	0.16	0.18	4	13.60
θ_2	1.44	1.407	0.01	0.56	7.39	12661	1.00
θ_3	0.08	0.903	0.02	0.03	0.18	4	16.98
m_0	198.99	1.046	1.75	127.89	731.84	12008	1.00
scale	6.64	1.186	0.05	3.01	27.00	10321	1.00
offset	7.18	0.017	6.94	7.18	7.42	17800	1.00
t_0	1.46	0.004	1.44	1.46	1.47	15996	1.00
σ	0.05	0.054	0.04	0.05	0.05	16833	1.00
$\theta_2 m_0$	200.04	2.654	3.17	28.49	1627.75	11121	1.00
$\theta_2$scale	5.53	3.043	0.12	0.67	48.94	8535	1.00
$m_0$scale	942.49	2.307	11.63	153.30	7610.72	9270	1.00
$\theta_2 m_0$scale	85.74	0.014	83.35	85.73	88.19	12577	1.00

Table 8: Summary of the Stan output for the SDE model given experimental data for eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.20	0.152	0.14	0.20	0.26	946	1.01
θ_2	0.33	1.366	0.04	0.19	1.52	305	1.02
θ_3	0.02	0.269	0.01	0.02	0.03	894	1.01
m_0	298.35	0.705	34.46	250.62	809.42	218	1.03
scale	2.12	0.309	1.14	2.02	3.69	772	1.01
offset	7.18	0.012	7.01	7.18	7.35	20589	1.00
σ	0.03	0.058	0.03	0.03	0.04	17224	1.00
$\theta_2 m_0$	47.92	0.327	23.41	46.00	83.31	857	1.01
$\theta_2$scale	0.60	1.182	0.11	0.37	2.54	256	1.02
$m_0$scale	650.64	0.831	57.98	498.32	2049.93	283	1.02
$\theta_2 m_0$scale	92.71	0.115	73.32	92.15	115.16	2711	1.00

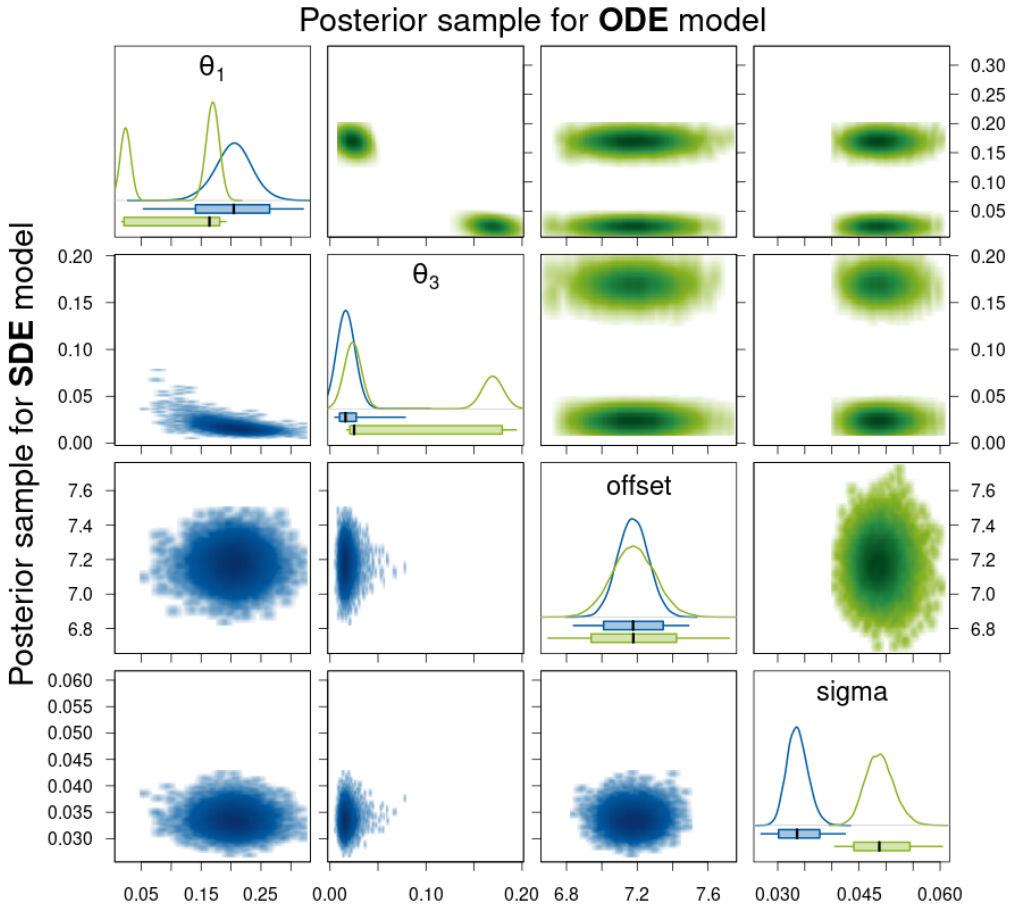


Figure 12: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for eGFP. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values.

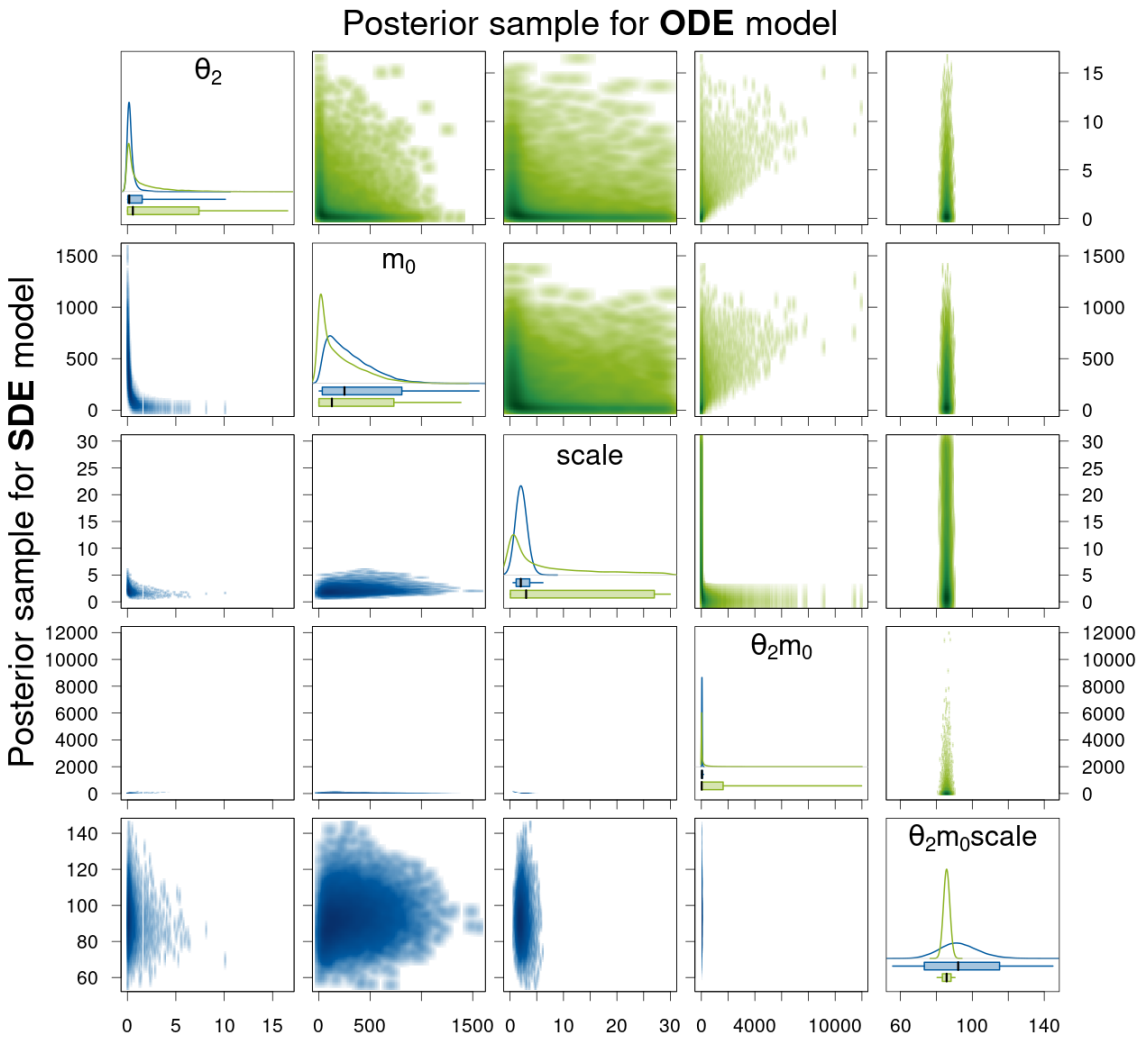


Figure 13: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for eGFP. For a detailed description of the figure's elements, see Figure 12.

Table 9: Statistics of posterior samples aggregated for 100 experimental trajectories for eGFP.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs
θ_1	ODE	11.05	0.11	0.058
	SDE	11.05	0.14	0.012
θ_2	ODE	11.05	7.91	0.016
	SDE	11.05	9.91	0.998
θ_3	ODE	11.05	0.11	0.057
	SDE	11.05	0.06	0.039
m_0	ODE	884.82	747.71	0.212
	SDE	884.82	456.80	172.338
scale	ODE	28.50	27.50	0.004
	SDE	28.50	4.61	5.288
$\theta_2 m_0$	ODE	6056.48	2032.46	23.287
	SDE	6056.48	230.89	219.006
$\theta_2 \mathrm{scale}$	ODE	228.08	68.38	1.879
	SDE	228.08	22.01	33.157
$m_0 \mathrm{scale}$	ODE	19271.13	9093.22	69.292
	SDE	19271.13	1392.46	4601.374
$\theta_2 m_0 \mathrm{scale}$	ODE	113232.70	24.10	57.316
	SDE	113232.70	138.03	78.659
offset	ODE	28.50	0.96	2.661
	SDE	28.50	0.38	1.086
σ	ODE	9.50	0.01	0.244
	SDE	9.50	0.01	0.004

7 Discussion and conclusion

We have modeled the translation kinetics after mRNA transfection using a two-dimensional Itô diffusion process described by an SDE and compared this modeling approach to one using ODEs. We have studied the parameter identifiability for both modeling approaches for the case that we observe a fluorescence signal which we assume to be a linear transformation of the amount of protein molecules (corrupted by multiplicative measurement error). For the ODE model, previous studies had already shown that the degradation rate constants θ_1 and θ_3 for the mRNA and the protein are only locally identifiable, and only the product $\theta_2 m_0 \mathrm{scale}$ of the translation rate constant, the initial amount of mRNA molecules transfected, and the scaling factor of the fluorescence signal is identifiable but the three parameters individually are not identifiable. In order to try to assess structural identifiability of the SDE model, we transformed the model, used the DAISY software, and also simulated from the model. Each of the approaches indicated that the SDE model might lead to better parameter identifiability. The most systematic approach is the one based on the surrogate model and DAISY as suggested by Browning et al. (2020); however, it only provides a necessary condition (even) for structural identifiability of the SDE model parameters. While checking this necessary condition is certainly useful especially e. g. when designing an experiment, it cannot help us *confirm* a difference in the parameter identifiability between the SDE and the ODE model. Especially because we are interested in the parameter identifiability based on one observed trajectory and the DAISY-based approach assumes that we were able to observe the first and the second moment of the fluorescence signal. Even when we take into account that we have several observed trajectories available from the experiment, these do not provide information about the moments because the initial time point t_0 of mRNA release is different for every trajectory and also for the other parameters, in particular for m_0, assuming that they are equal for all observed cells does not seem reasonable. By simulating from the SDE model, we were able to assess the differences in the variation within individual trajectories for different parameter combinations. We saw that the variation within trajectories was clearly higher for lower θ_1 and higher θ_3 which suggest that they are structurally globally identifiable. The variation within trajectories was also higher for higher values of scale and lower values of the product $\theta_2 m_0$. Whereas there did not seem to be much difference in the variation within trajectories when the values of scale and $\theta_2 m_0$ were kept constant and only the individual values of θ_2 and m_0 varied. Therefore, scale and $\theta_2 m_0$ seem to be structurally identifiable, but θ_2 and m_0 do not. While this simple simulation approach worked out well for the model considered here, one of its weak points is, of course, the somewhat subjective visual assessment of the variation within trajectories. A more quantitative approach to this would be to simulate a large number of trajectories (with very small time step) for every considered parameter combination, to approximate the quadratic variation for each trajectory, and then, to compare these values between individual trajectories started with the same seed for different parameter combinations and to compare also the distributions of these values for different parameter combinations. Another drawback of both simulation-based approaches is the fact that the analysis is based on a finite set of parameter combinations that can be considered; and thus, drawing general conclusions for the entire parameter space may be problematic.

Moreover, we have assessed the practical parameter identifiability for both model types by sampling from the parameter posterior distribution given simulated data without and with measurement error and the experimental data published in Fröhlich et al. (2018). We found that the parameters θ_1 and θ_3 are indeed globally identifiable for the SDE model given individual trajectories, unlike for the ODE model. And not only the product $\theta_2 m_0 \mathrm{scale}$ but also the parameter scale and the product $\theta_2 m_0$ are globally identifiable for the SDE model. Moreover, for the simulated data, the 95% CIs for the identifiable parameters for the SDE model covered the true parameter value adequately many times. Whereas for the ODE model, the true parameter values for the parameters θ_1, θ_3, and $\theta_2 m_0 \mathrm{scale}$ were not covered by the 95% CIs for many of the posterior samples and were sometimes not even included in the range of values in the sample. The fact that the parameters θ_1 and θ_3 can be adequately determined using the SDE modeling approach given an individual trajectory renders the multi-experiment approach with different mRNA constructs and the computational intense hierarchical optimization algorithm used in Fröhlich et al. (2018) unnecessary in the case that the determination of these parameters is the main objective. Besides, assuming that a MJP is the most appropriate description of the underlying dynamics, we saw that the estimated parameter values for a single cell trajectory based on the ODE model cannot be trusted even when narrow 95% CIs suggest low uncertainty. While the SDE model is clearly superior in terms of the information that we are able to extract from a single trajectory about the parameters that determine the dynamics of the underlying process, it has nevertheless several disadvantages. First

of all, we were not able to include the estimation of the initial time point t_0 of mRNA release into the Stan sampling procedure. We believe that this is not easily possible due to the fact that for the SDE model, the process switches from a deterministic evolution to a stochastic one at t_0 and including t_0 as a parameter in the posterior distribution leads to non-smoothness of the posterior distribution which cannot be handled by HMC sampling as it makes use of the derivative of the log-posterior. Other sampling approaches such as particle MCMC might alleviate this problem, but to our knowledge, no examples of inferring a random time point for SDE models have been investigated so far and would thus require further work. Another drawback of the SDE model are the higher computational costs as we need to sample from a higher-dimensional distribution (due to the random process values) than for the ODE model. For the SDE model, the sampling in our study takes on average almost 5.5 hours while for the ODE model, it averages at about 20 minutes. In general, estimation procedures for SDE models are more complex and unlike for ODE models, publicly available software tools are rare and usually not generally applicable. There is definitely a need to further develop such tools for SDE models in order to harness their full potential, especially with regard to better identifiability of kinetic parameters. On the other hand, combining both modeling approaches as we have done here by first determining t_0 based on the ODE model and then estimating the kinetic parameters based on the SDE model is clearly also meaningful.

References

Bellu, G., Saccomani, M. P., Audoly, S., & D'Angiò, L. (2007). DAISY: A new software tool to test global identifiability of biological and physiological systems. *Computer Methods and Programs in Biomedicine*, 88(1), 52 – 61.

Betancourt, M. J., Byrne, S., & Girolami, M. (2015). Optimizing the integrator step size for hamiltonian monte carlo. *Preprint*, arXiv:1411.6669.

Braumann, C. A. (2019). *Introduction to Stochastic Differential Equations with Applications to Modelling in Biology and Finance*. John Wiley & Sons, Ltd.

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E., & Simpson, M. J. (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of The Royal Society Interface*, 17(173), 20200652.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1), 1–32.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., & Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in C++. *Preprint*, arXiv:1509.07164.

DeFrancesco, L. (2020). Whither COVID-19 vaccines? *Nature Biotechnology*, 38(10), 1132–1145.

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, 195(2), 216 – 222.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Finkenstädt, B., Heron, E. A., Komorowski, M., Edwards, K., Tang, S., Harper, C. V., Davis, J. R. E., White, M. R. H., Millar, A. J., & Rand, D. A. (2008). Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24), 2901–2907.

Fröhlich, F., Reiser, A., Fink, L., Woschée, D., Ligon, T., Theis, F. J., Rädler, J. O., & Hasenauer, J. (2018). Multi-experiment nonlinear mixed effect modeling of single-cell translation kinetics after transfection. *npj Systems Biology and Applications*, 4(1), 42.

Fuchs, C. (2013). *Inference for Diffusion Processes*. Berlin Heidelberg: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4), 403 – 434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1), 404 – 425.

Heron, E. A., Finkenstädt, B., & Rand, D. A. (2007). Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics*, 23(19), 2596–2603.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47), 1593–1623.

Kloeden, P. E. & Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Berlin Heidelberg: Springer.

Komorowski, M., Finkenstädt, B., Harper, C. V., & Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10(1), 343.

Lee, P. M. (2012). *Bayesian Statistics: An Introduction*. Wiley Publishing, 4th edition.

Leonhardt, C., Schwake, G., Stögbauer, T. R., Rappl, S., Kuhr, J.-T., Ligon, T. S., & Rädler, J. O. (2014). Single-cell mRNA transfection studies: delivery, kinetics and statistics by numbers. *Nanomedicine*, 10(4), 679–688.

Ligon, T. S., Leonhardt, C., & Rädler, J. O. (2014). Multi-level kinetic model of mRNA delivery via transfection of lipoplexes. *PLoS One*, 9(9), e107148.

Munsky, B., Trinh, B., & Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5, 318–318.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. CRC Press.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1), 221–259.

Øksendal, B. K. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Berlin: Springer, 6th edition.

Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2), 216–226.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.

Reiser, A., Woschée, D., Mehrotra, N., Krzysztoń, R., Strey, H. H., & Rädler, J. O. (2019). Correlation of mRNA delivery timing and protein expression in lipid-based transfection. *Integrative Biology*, 11(9), 362–371.

Sahin, U., Karikó, K., & Türeci, Ö. (2014). mRNA-based therapeutics – developing a new class of drugs. *Nature Reviews Drug Discovery*, 13(10), 759–780.

Schnoerr, D., Sanguinetti, G., & Grima, R. (2017). Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9), 093001.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.1, `http://mc-stan.org/`.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC. *Bayesian Analysis*, -1(-1), 1 – 28.

Warne, D. J., Baker, R. E., & Simpson, M. J. (2019). Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *Journal of The Royal Society Interface*, 16(151), 20180943.

A Supplementary material

A.1 General Itô diffusion processes

In the main article, we consider an example of a parametric time-homogeneous Itô diffusion process. In general, a d-dimensional *time-homogeneous Itô diffusion process* $(\boldsymbol{X}_t)_{t\geq 0}$ is a stochastic process that fulfills the following SDE:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{\mu}\left(\boldsymbol{X}_t, \boldsymbol{\theta}\right) \mathrm{d}t + \boldsymbol{\sigma}\left(\boldsymbol{X}_t, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 = \boldsymbol{x}_0, \tag{8}$$

with state space $\mathcal{X} \subseteq \mathbb{R}^d$, starting value $\boldsymbol{x}_0 \in \mathcal{X}$, and an r-dimensional Brownian motion $(\boldsymbol{B}_t)_{t\geq 0}$. The model parameter $\boldsymbol{\theta} \in \Theta$ is from an open set $\Theta \subseteq \mathbb{R}^p$. The function $\boldsymbol{\mu} : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ is usually called the drift coefficient and $\boldsymbol{\sigma} : \mathbb{R}^d \times \Theta \to \mathbb{R}^{d\times r}$ the diffusion coefficient. Equation (8) is a symbolic way of writing the stochastic integral equation

$$\boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \boldsymbol{\mu}\left(X_s, \boldsymbol{\theta}\right) \mathrm{d}s + \int_0^t \boldsymbol{\sigma}\left(\boldsymbol{X}_s, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{B}_s \quad \text{for all } t \geq 0 \quad \mathbb{P}\text{-almost surely,}$$

where the first integral is an ordinary Riemann integral and the second integral is a stochastic integral in the Itô sense. In the remainder of this section, we omit the dependence of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ on the parameter $\boldsymbol{\theta}$ and briefly state two important tools for handling SDE models of this type. Elaborate and general introductions to SDEs can be found e. g. in Øksendal (2003), Fuchs (2013), and Braumann (2019).

The Itô integral and thus also Itô diffusion processes do not adhere to the rules of classical calculus. Instead, the following theorem states the stochastic counterpart of the chain rule from classical calculus which is known as *Itô formula*. The formulation of the Itô formula specific for Itô diffusion processes as we state it here follows directly from the general Itô formula as stated in Øksendal (2003, Chapter 4.2).

Theorem A.1 (Itô formula)**.** *Let $\boldsymbol{X}_t$ be a d-dimensional Itô diffusion process described by an SDE as in (8). Let $g(t, \boldsymbol{x}) = (g_1(t, \boldsymbol{x}), \ldots, g_q(t, \boldsymbol{x}))$ be a map from $[0, T] \times \mathbb{R}^d$ into $\mathbb{R}^q$ with continuous first-order partial derivatives in t and continuous first- and second-order partial derivatives in $\boldsymbol{x}$. Then the process*

$$\boldsymbol{Y}(t, \omega) = g(t, \boldsymbol{X}_t)$$

is an Itô process whose k^{th} component $\boldsymbol{Y}^{(k)}$ is given by

$$\begin{aligned}
\mathrm{d}\boldsymbol{Y}^{(k)} &= \frac{\partial g_k}{\partial t}(t, \boldsymbol{X})\,\mathrm{d}t + \sum_{i=1}^{q} \frac{\partial g_k}{\partial x^{(i)}}(t, \boldsymbol{X})\,\mathrm{d}\boldsymbol{X}^{(i)} + \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q} \frac{\partial^2 g_k}{\partial x^{(i)} \partial x^{(j)}}(t, \boldsymbol{X})\,\mathrm{d}\boldsymbol{X}^{(i)} \cdot \mathrm{d}\boldsymbol{X}^{(j)}, \\
&= \left(\frac{\partial g_k}{\partial t}(t, \boldsymbol{X}) + \boldsymbol{\mu}(\boldsymbol{X})^{\mathsf{Tr}} \nabla g_k(t, \boldsymbol{X}) + \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\sigma}(\boldsymbol{X})\boldsymbol{\sigma}(\boldsymbol{X})^{\mathsf{Tr}} \nabla\left(\nabla g_k(t, \boldsymbol{X})\right)\right) \right) \mathrm{d}t \\
&\quad + \left(\nabla g_k(t, \boldsymbol{X})\right)^{\mathsf{Tr}} \boldsymbol{\sigma}(\boldsymbol{X})\,\mathrm{d}\boldsymbol{B}_t,
\end{aligned} \tag{9}$$

where ∇g_k denotes the gradient of g_k with respect to the components of $\boldsymbol{x}$ and $\mathrm{d}\boldsymbol{X}^{(i)} \cdot \mathrm{d}\boldsymbol{X}^{(j)}$ is computed according to the rules $\mathrm{d}\boldsymbol{B}^{(i)} \cdot \mathrm{d}t = \mathrm{d}t \cdot \mathrm{d}\boldsymbol{B}^{(j)} = (\mathrm{d}t)^2 = 0$ and $\mathrm{d}\boldsymbol{B}^{(i)} \cdot \mathrm{d}\boldsymbol{B}^{(j)} = \delta_{ij}\,\mathrm{d}t$ with δ_{ij} denoting the Kronecker delta.

Most SDEs do not have an analytical solution and their transition densities are not explicitly known. Instead, numerical approximation schemes are used for the solution of the SDEs. Kloeden & Platen (1992) provide a detailed description of these methods. The most commonly used approximation is the *Euler(-Maruyama) scheme*. It can be conveniently written in vector notation and approximates the d-dimensional solution $(\boldsymbol{X}_t)_{t\geq 0}$ of an SDE by setting $\boldsymbol{Y}_0 = \boldsymbol{x}_0$ and, then, successively calculating the following:

$$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k + \boldsymbol{\mu}\left(\boldsymbol{Y}_k\right) \Delta t_k + \boldsymbol{\sigma}\left(\boldsymbol{Y}_k\right) \Delta \boldsymbol{B}_k, \tag{10}$$

where $\Delta t_k = t_{k+1} - t_k$, $\Delta \boldsymbol{B}_k = \boldsymbol{B}_{t_{k+1}} - \boldsymbol{B}_{t_k}$, and $\boldsymbol{Y}_k$ is the approximation of $\boldsymbol{X}_{t_k}$ for $k = 0, 1, 2, \ldots$. Since the Euler scheme is a linear transformation of the normally-distributed increments $\Delta \boldsymbol{B}_k \sim \mathcal{N}\left(\boldsymbol{0}, \Delta t_k \boldsymbol{I}_r\right)$ of the Brownian motion, where $\boldsymbol{I}_r$ denotes the r-dimensional identity matrix, the process state $\boldsymbol{Y}_{k+1}$ conditioned on $\boldsymbol{Y}_k$ is also normally-distributed with

$$\boldsymbol{Y}_{k+1} \,|\, \boldsymbol{Y}_k \sim \mathcal{N}\left(\boldsymbol{Y}_k + \boldsymbol{\mu}\left(\boldsymbol{Y}_k\right) \Delta t_k\,,\, \boldsymbol{\sigma}\left(\boldsymbol{Y}_k\right) \boldsymbol{\sigma}^{\mathsf{Tr}}\left(\boldsymbol{Y}_k\right) \Delta t_k\right),$$

where $\mathcal{N}\left(\boldsymbol{a}, \boldsymbol{b}\right)$ denotes the multivariate normal distribution with mean vector $\boldsymbol{a} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{b} \in \mathbb{R}^{d\times d}$.

A.2 Investigating the need for data augmentation for the SDE model

In this section, we focus on the inference problem for the SDE model and investigate whether the amount of data that we have available ($K = 181$ observations per cell with time step $\Delta t = 1/6$ hours) is sufficient for the Euler approximation to be appropriate, i. e. whether the step size between observations is small enough. We simulate one trajectory of the MJP described in Section 3.1 with parameters $\theta_1 = 0.11$, $\theta_2 = 0.3$, $\theta_3 = 0.09$, and $m_0 = 200$ on the time interval $[0, 30]$ using Gillespie's algorithm and use observations at 181 equidistant time points. We assume for now that the amount X_2 of GFP is directly observed without error and that for the amount X_1 of mRNA, we only observe the initial value $m_0 = 200$. All observations are without measurement error and we assume $t_0 = 0$ to be known. Thus, we only estimate the kinetic parameters $\boldsymbol{\theta}$ for the SDE model, and to this end, use Stan and Bayesian data augmentation with different numbers of inter-observation intervals which means that we impute additional (artificial) data points between every two observations and these points are treated as additional parameters in the estimation procedure (for a detailed description of Bayesian data augmentation see Fuchs, 2013). A number of inter-observation intervals of 1 means that we do not impute any points between observations. A number of 2 inter-observation intervals means that we impute one point between every two observations and so on. We generated 4 HMC chains with 1000 iterations after warm-up each. Figure 14 shows the median of the obtained posterior sample as the point estimates and the CIs for the three kinetic parameters and for different numbers of inter-observation intervals. Evidently, the estimation results do not improve when increasing the number of inter-observation intervals. Therefore, we conclude that data augmentation is not necessary and do not make use of data augmentation in the main part of the article.

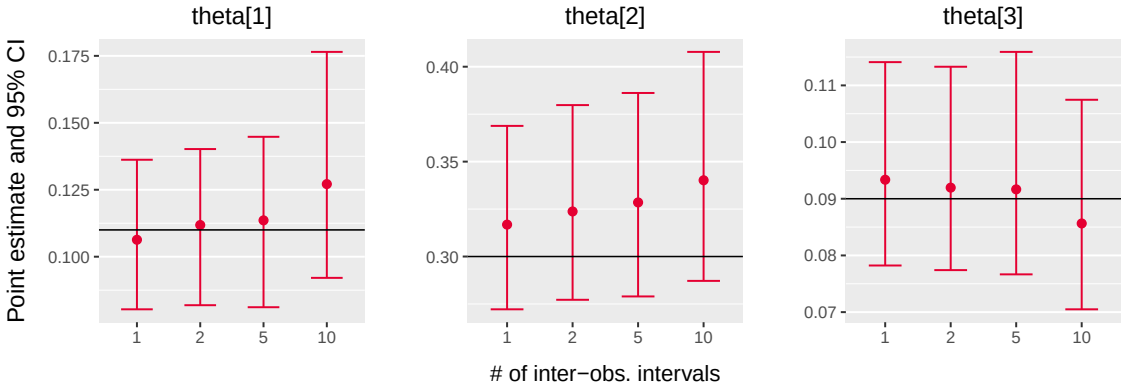


Figure 14: Point estimates (median of the posterior sample) and 95% CIs for the kinetic parameters estimated with Stan and Bayesian data augmentation for different numbers of inter-observations intervals. The black line represents the true parameter values with which the data was generated.

A.3 Hamiltonian Monte Carlo (HMC) methods and Stan

A.3.1 Brief introduction to the algorithm

To sample from the posterior densities of the two model types (ODE and SDE) as formulated in Section 6 of the main article, we use the open source software Stan (Carpenter et al., 2017). Stan provides an implementation of the Hamiltonian Monte Carlo (HMC) based No-U-Turn Sampler (NUTS) to which we give a very brief introduction here which mainly draws from the description in Gelman et al. (2013). Neal (2011) gives a more detailed account. HMC methods (originally called *hybrid* Monte Carlo methods by Duane et al. (1987)) are a class of MCMC methods. The computational cost in each iteration for HMC methods is higher than for other MCMC methods such as Gibbs sampling or Metropolis-Hastings algorithms because HMC makes use of the derivative of the target distribution. But by that, transitions between the chain states can be generated that efficiently span the (with respect to the target distribution) important regions of the state space. By taking into account the information of the gradient, HMC avoids the random walk behavior and difficulties caused by distributions with high correlations that other MCMC methods exhibit.

Assume we want to sample from the p-dimensional distribution $\pi(\boldsymbol{\theta})$ for parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. Motivated by the physical concept of Hamiltonian dynamics, HMC introduces an auxiliary momentum variables $\boldsymbol{\rho} \in \mathbb{R}^p$ and draws from a joint density $p(\boldsymbol{\theta}, \boldsymbol{\rho}) = p(\boldsymbol{\rho} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The joint density defines the so-called Hamiltonian

$$H(\boldsymbol{\theta}, \boldsymbol{\rho}) = -\log p(\boldsymbol{\theta}, \boldsymbol{\rho}) = -\log p(\boldsymbol{\rho} \,|\, \boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}) = K(\boldsymbol{\theta}, \boldsymbol{\rho}) + V(\boldsymbol{\theta}) \tag{11}$$

that describes the total energy of the system and is equal to the sum of the kinetic energy K and the potential energy V. In HMC, the distribution of $\boldsymbol{\rho}$ is usually chosen to be independent of $\boldsymbol{\theta}$. A common choice is $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$, where $\mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$ denotes the multivariate normal distribution with mean vector $\mathbf{0}_p$ and covariance matrix $\boldsymbol{M} \in \mathbb{R}^{p \times p}$, and $\boldsymbol{M}$ is called the *design* or (by analogy to the physical model) *mass matrix* and often chosen to be a diagonal matrix. Thus, the kinetic energy becomes

$$K(\boldsymbol{\rho}) = \boldsymbol{\rho}^{\mathsf{Tr}} \boldsymbol{M}^{-1} \boldsymbol{\rho} / 2, \tag{12}$$

where $\boldsymbol{M}^{-1}$ denotes the inverse matrix of $\boldsymbol{M}$.

In each iteration of the HMC algorithm, a momentum $\boldsymbol{\rho}$ is sampled (e. g. from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$) and then by analogy to the physical model of the frictionless movement of a marble with position $\boldsymbol{\theta}$ and momentum $\boldsymbol{\rho}$ (describing the marble's mass and velocity) across a surface, the dynamics, i. e. the changes in position and momentum, that preserve the total energy are described by the Hamiltonian equations

$$\begin{aligned} \frac{d\rho_i}{dt} &= -\frac{\partial H}{\partial \theta_i}, \\ \frac{d\theta_i}{dt} &= \frac{\partial H}{\partial \rho_i} \end{aligned}$$

for $i = 1, \ldots, p$. With the choice of H, K, and V as in Equations (11) and (12), we have

$$\begin{aligned} \frac{d\boldsymbol{\rho}}{dt} &= -\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) \quad &&= \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}), \\ \frac{d\boldsymbol{\theta}}{dt} &= \nabla_{\boldsymbol{\rho}} K(\boldsymbol{\rho}) \quad &&= \boldsymbol{M}^{-1} \boldsymbol{\rho}, \end{aligned} \tag{13}$$

where $\nabla_{\boldsymbol{x}}$ denotes the gradient with respect to $\boldsymbol{x}$. In each iteration, Equations (13) are numerically integrated to obtain proposals $\boldsymbol{\theta}^*$ and $\boldsymbol{\rho}^*$. A common choice of the numerical integrator is the leap-frog method. Then, an accept-reject step is performed analogously to the Metropolis-Hastings algorithm. We summarize the HMC steps in Algorithm 1.

Algorithm 1: Hamiltonian Monte Carlo algorithm (with leap-frog integrator)

Input: *A target density $\pi(\cdot)$, an initial state $\boldsymbol{\theta}^{(0)}$, number of iterations n, mass matrix $\boldsymbol{M}$, and step size ϵ and number L of steps for numerical integration.*

In each iteration $i = 1, \ldots, n$:

Step 1 Generate $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{M})$ and set $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}^{(i-1)}$ and $\boldsymbol{\rho}^* \leftarrow \boldsymbol{\rho}$.

Step 2 Repeat L leap-frog steps by setting:

$$\begin{aligned} \boldsymbol{\rho}_{\frac{1}{2}} &\leftarrow \boldsymbol{\rho}^* + \frac{1}{2}\epsilon \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^*) \\ \boldsymbol{\theta}^* &\leftarrow \boldsymbol{\theta}^* + \epsilon M^{-1} \boldsymbol{\rho}_{\frac{1}{2}} \\ \boldsymbol{\rho}^* &\leftarrow \boldsymbol{\rho}_{\frac{1}{2}} + \frac{1}{2}\epsilon \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^*) \end{aligned}$$

Step 3 Accept $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^{(i)}$ with probability

$$\alpha(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}, \boldsymbol{\theta}^*, \boldsymbol{\rho}^*) = \min\left[1, \exp\left(H(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\rho}) - H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)\right)\right],$$

if $\boldsymbol{\theta}^*$ is rejected $\boldsymbol{\theta}^{(i)} := \boldsymbol{\theta}^{(i-1)}$.

Output: *A sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}\}$ approximately distributed according to $\pi(\cdot)$.*

Two of the limitations of this general HMC algorithm are on the one hand that due to the use of the derivative with respect parameter, it is only suitable for continuous distributions, and on the other

hand, the choice of the tuning parameters is of crucial importance to the performance of the algorithm and can be cumbersome. The tuning parameters include the mass matrix M, and step size ϵ and number L of steps for numerical integration.

An extension of HMC, the *NUTS*, introduced by Hoffman & Gelman (2014) includes a way to automatically determine the number L of steps for numerical integration using an recursive algorithm that grows a binary tree representing leap-frog steps forward and backward in time which is stopped as soon as further steps do no longer increase the distance between a newly explored point and the original starting point (i. e. as soon as the steps start to make a U-turn).

The open-source Bayesian inference package *Stan* which we make use of through its R interface `rstan` (Stan Development Team, 2019) provides an efficient C++ implementation of NUTS. In Stan, the gradient of the log-posterior distribution is calculated (exactly) by reverse-mode automatic differentiation (Carpenter et al., 2015). Moreover, Stan can automatically optimize the step size ϵ to match a (user-defined) acceptance-rate target based on dual averaging as proposed by Nesterov (2009) and it also estimates the mass matrix M during a warm-up phase consisting of several stages.

A.3.2 Evaluating (general) MCMC output

While in theory, any MCMC method (for which convergence of the transition kernel is ensured) will give a sample from the target distribution if infinitely many iterations are executed; in practice, the sample size can only be finite which makes it necessary to carefully evaluate the MCMC output.

A quantity that can be used to quantify the degree of convergence when several chains have been simulated is the $\hat{R}$ value. The $\hat{R}$ convergence (or rather stationarity) diagnostic compares the between- and within-chain variance for individual model parameters and other univariate quantities of interest. Assume we are considering the scalar parameter ψ for which we have simulations $\psi_{i,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and for m chains (after discarding the warm-up iterations and then splitting each simulated chain in half) of length n. Let

$$\widehat{\mathrm{var}}^{+}(\psi \,|\, \mathcal{D}) = \frac{n-1}{n} W + \frac{1}{n} B \tag{14}$$

be an estimate for the marginal posterior variance of ψ, where the *within-sequence variance* W is defined by

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \qquad \text{with} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi_{\cdot j}})^2,$$

and the *between-sequence variance* B is defined by

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi_{\cdot j}} - \bar{\psi_{\cdot\cdot}})^2 \qquad \text{with} \quad \bar{\psi_{\cdot j}} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \quad \text{and} \quad \bar{\psi_{\cdot\cdot}} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi_{\cdot j}}.$$

Then, $\hat{R}$ is defined as

$$\hat{R} = \sqrt{\frac{\hat{\mathrm{var}}^{+}(\psi \,|\, \mathcal{D})}{W}}.$$

Due to the splitting of chains in half, $\hat{R}$ calculated in this way is also known as split-$\hat{R}$ and was suggested in Gelman et al. (2013). The value can be interpreted as the factor by which the scale of the distribution of the current simulations for ψ can be reduced by continuing the number of iterations to infinity. If chains have mixed well, $\hat{R}$ is close to 1. Gelman et al. (2013) state that values up to 1.1 are acceptable. The $\hat{R}$ reported by Stan is calculated as the maximum of a so-called rank-normalized split-$\hat{R}$ and a rank-normalized folded-split-$\hat{R}$ which was recently suggested by Vehtari et al. (2021).

Another issue in MCMC sampling is the fact that the draws are not independent but may even be highly correlated. It is important to keep in mind that such a correlated sample from the parameter posterior distribution does not contain the same amount of information as an independent and identically distributed sample. This issue is addressed by the notion of the effective sample size (ESS). The ESS of a sample of correlated draws quantifies the size of a corresponding independent and identically distributed sample that contains the same amount information.

The ESS for a sample of scalar parameter ψ consisting of m chains each of length n (again after discarding warm-up iterations but without splitting of the chains) can be defined as

$$n_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t},$$

where ρ_t is the autocorrelation of the sequence ψ at lag t. This quantity can be approximated in different ways. Here, we give the approximation that is presented in Gelman et al. (2013) and implemented in `rstan`. The estimated autocorrelations $\hat{\rho}_t$ are computed as

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{\text{var}}^{+}(\psi \mid \mathcal{D})}$$

for $t = 1, \ldots, T$ and, where the estimate $\widehat{\text{var}}^{+}$ for the marginal posterior variance is calculated as in (14) and the variogram V_t at lag t is calculated as

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2.$$

The maximal considered lag T is chosen to be the first odd positive integer for which $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative and finally, the ESS is approximated by

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t}.$$

Gelman et al. (2013) recommend that a minimum ESS of 10 per simulated chain is achieved. The between-chain information is taken into account in the calculation of $\hat{n}_{\text{eff}}$ by including the term $\widehat{\text{var}}^{+}(\psi \mid \mathcal{D})$. Thus, the ESS is affected when we try to sample from multi-modal distributions. In fact, in the case of well-separated modes and each chain sampling only from one of these modes, the ESS roughly equals to the number chains divided by the number of modes.

A.3.3 Further diagnostics of MCMC output specific to HMC and NUTS

In addition to the quality indicators for MCMC output mentioned in the previous section, Stan reports further quantities that are specific to HMC and NUTS and are of interest to assess sampling efficiency. These include the number of divergent transitions, the tree depth, and the (energy) Bayesian fraction of missing (BFMI) which we briefly describe below. See the Stan reference manual for more detailed explanations (Stan Development Team, 2019).

Integrating the Hamiltonian equations (13) in Section A.3.1 analytically would preserve the value of the Hamiltonian $H(\boldsymbol{\theta}, \boldsymbol{\rho})$; however, since analytical integration is not possible for most problems of interest, the equations are numerically integrated which leads to numerical errors. If the difference between $H(\boldsymbol{\theta}, \boldsymbol{\rho})$ of the starting point and $H(\boldsymbol{\theta}^*, \boldsymbol{\rho}^*)$ of the proposed point at the end of the simulated Hamiltonian trajectory becomes too large (where the default threshold is 10^3), Stan will classify the starting point as one of a *divergent transition*. If many of such starting points of divergent transitions are concentrated within a region of parameter space, this may be an indication that the curvature of the posterior is very high in this region and that the step size ϵ is too large to adequately explore this region.

As briefly mentioned in Section A.3.1, NUTS builds up a binary tree when determining the number L of leapfrog steps to take before a U-turn would occur. Stan records the depth of this tree for each iteration and thus also the corresponding starting point. Moreover, the user can specify a maximum tree depth d to avoid long execution times due too many steps; as at most 2^{d-1} leapfrog steps are taken in each iteration. The default value is $d = 10$. Hitting this maximum means that NUTS is terminated prematurely (i. e. more steps would have been possible before a U-turn) and Stan counts how many times this occurs. Reasons for having to take many steps may be a too small step size due to poor adaptation to a posterior of varying curvature or targeting a very high acceptance rate.

According to Betancourt et al. (2015), the *BFMI* indicates how well the energy sets of the Hamiltonian are explored. Let $E = H(\boldsymbol{\theta}, \boldsymbol{\rho})$ be the total energy, $\pi(E|\rho)$ the energy transition distribution, and $\pi(E)$ the marginal energy distribution. If $\pi(E|\rho)$ is substantially more narrow than $\pi(E)$, then a HMC

chain may not be able to completely explore the tails of the target distribution. The BFMI quantifies the mismatch between the two distributions and is defined and approximated by

$$BFMI := \frac{\mathbb{E}_\pi\left[Var_{\pi_{E|\rho}}[E|\rho]\right]}{Var_{\pi_E}[E]} \approx \frac{\sum_{n=1}^{N}(E_n - E_{n-1})^2}{\sum_{n=0}^{N}(E_n - \bar{E})^2} =: \widehat{BFMI}.$$

The Stan development team recommends to ensure that the value of $\widehat{BFMI}$ is greater than 0.2.

A.4 Additional results for the posterior sampling

A.4.1 Sampling results for simulated data

Figures 15, 16, 17, and 18 show the same sampling output (the four posterior samples for the two simulated data sets depicted in Figure 5) as Figures 6, 7, 9, and 10 in Section 6.2; however here, the results are not compared between the ODE and the SDE model but between simulated data with and without measurement error.

For the SDE, we see in Figure 15 that the occurrence of measurement error substantially impacts the distribution of the posterior sample with respect to the parameters θ_1 and θ_3. The shape of the two dimensional projection changes from an elliptic shape to a banana-like shape. Especially for θ_3, the 95% CI and the range of values in the posterior sample increase a lot and the true parameter value is only barely covered by the 95% CI for simulated data with measurement error.

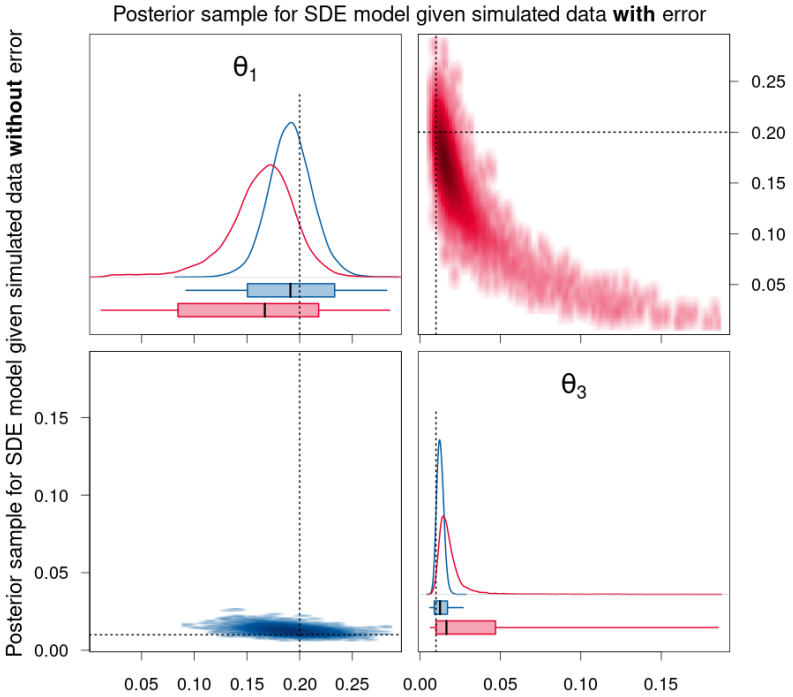


Figure 15: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the SDE model given simulated data without (blue, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.

Similarly for the parameters θ_2, m_0, scale and their products, Figure 16 shows that there is quite a difference between the distributions of the posterior samples for the simulated data without and with measurement error. In particular for the parameters scale and $\theta_2 m_0$ which we consider to be identifiable, the 95% CIs increase substantially for data with measurement error, and also the appearance of the two-dimensional projections with respect to these two parameters changes a lot, from a slightly bent

ellipse to a clear banana shape. For the product $\theta_2 m_0 \text{scale}$, the dispersion of the posterior samples changes only slightly which is apparent from the similar lengths of the 95% CIs in Figure 16 and also from the similar c.v. in Tables 2 and 5 (0.083 for data without measurement error and 0.093 for data with measurement error). The location of the sample measured e. g. by the median slightly shifts away from the true parameter value for the data with measurement error; however, the true value is still included in the 95% CIs. Only for parameter m_0 for which we also did not see much difference in the posterior samples for the ODE vs. SDE model, the occurrence of measurement error does not seem to affect the posterior sample much. For the remaining parameters θ_2, $\theta_2 \text{scale}$, and $m_0 \text{scale}$ which we do not consider to be identifiable but for which the 95% CIs of the posterior samples for the SDE model were clearly more narrow than the 95% CIs of the corresponding posterior sample for the ODE model, the 95% CIs and ranges of values of the posterior sample for the SDE model for data with measurement error are broader than for data without measurement error.

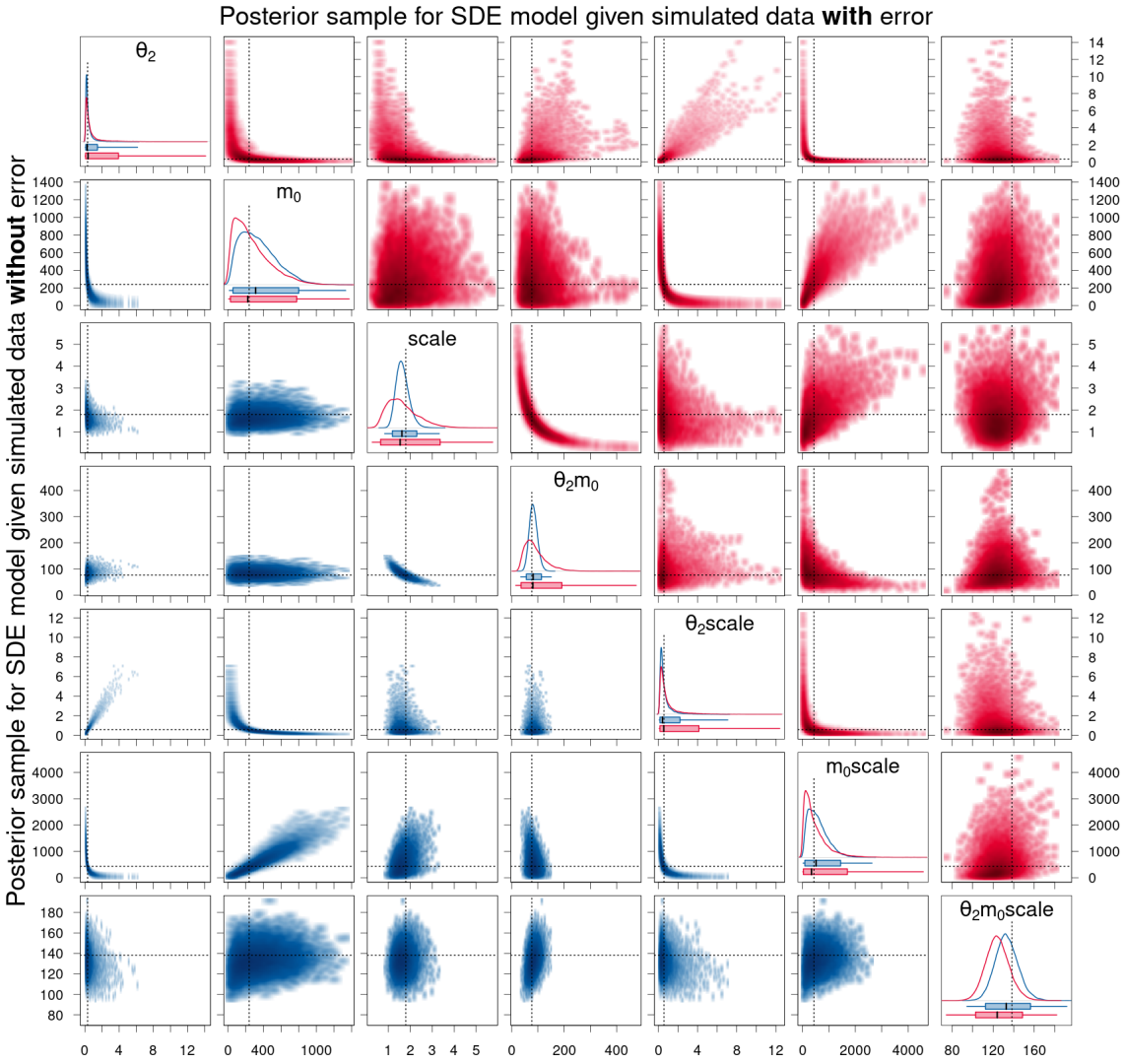


Figure 16: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE model given simulated data without (blue, lower triangle) and with (red, upper triangle) measurement error. For a detailed description of the figure's elements, see Figure 15.

For the ODE model, Figures 17 and 18 show that there is hardly any difference for most of the parameters between the posterior sample for the data without and with measurement error since the majority of the parameters are not identifiable anyway. For the parameters offset and t_0, there is a slight difference. For the measurement error parameter σ, the posterior sample consists of higher values for data with measurement error as expected. Note that for both simulated datasets, the range of the posterior sample does not include the true parameter value for σ. Finally for the product $\theta_2 m_0 \text{scale}$,

the dispersion of the posterior sample increases only slightly for data with measurement error and the location of the sample shifts away from the true parameter value. Also for this parameter, the range of the posterior sample does not include the true parameter value for both simulated datasets.

Figure 19 shows the statistics of the posterior samples for the simulated data without and with measurement error aggregated over 100 simulated trajectories. It visualizes the last two columns of Tables 3 and 6 and compares the results of the posterior samples for the simulated data without to those with measurement error separately for the SDE and the ODE model within each plot, instead of comparing the two model types separately for each kind of data as in Figures 8 and 11.

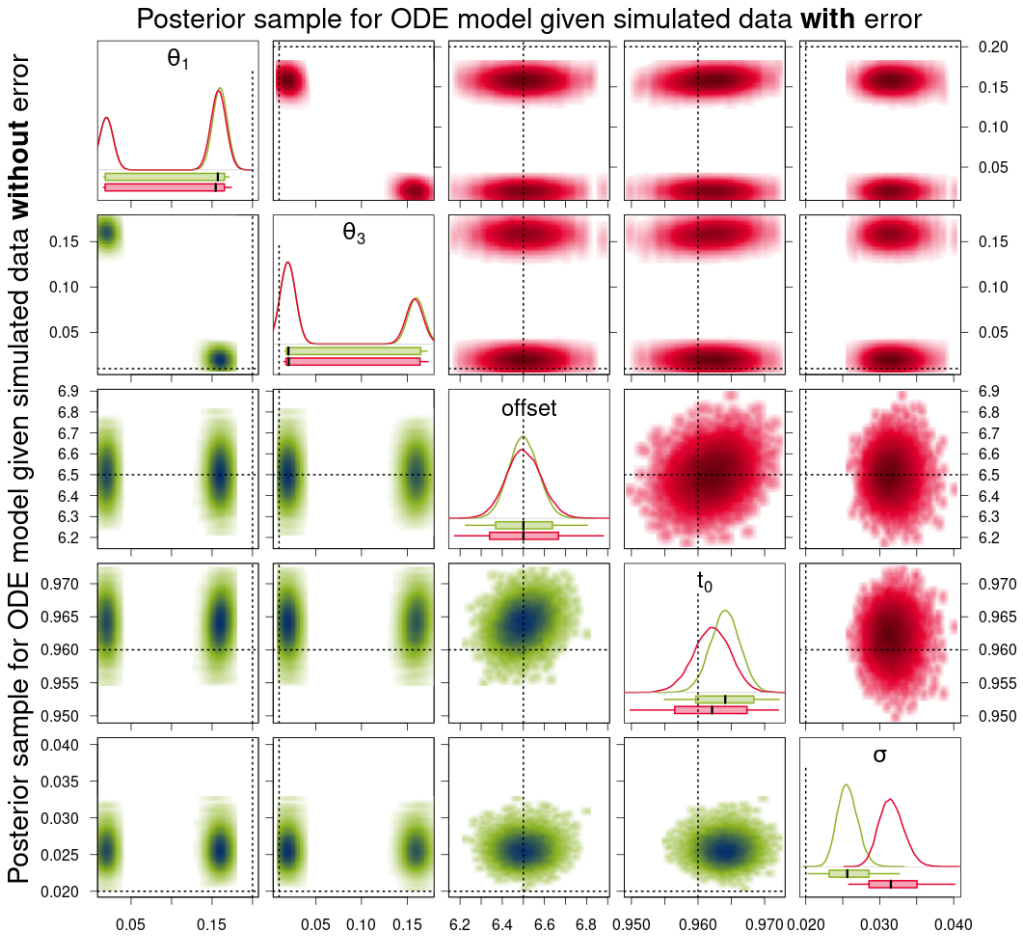


Figure 17: Density estimates of the posterior samples for parameters θ_1 and θ_3 for the ODE model given simulated data without (green, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data. For the parameter σ, the dotted line only represents the true value for the data with measurement error. For the data without measurement error, σ is equal to 0.

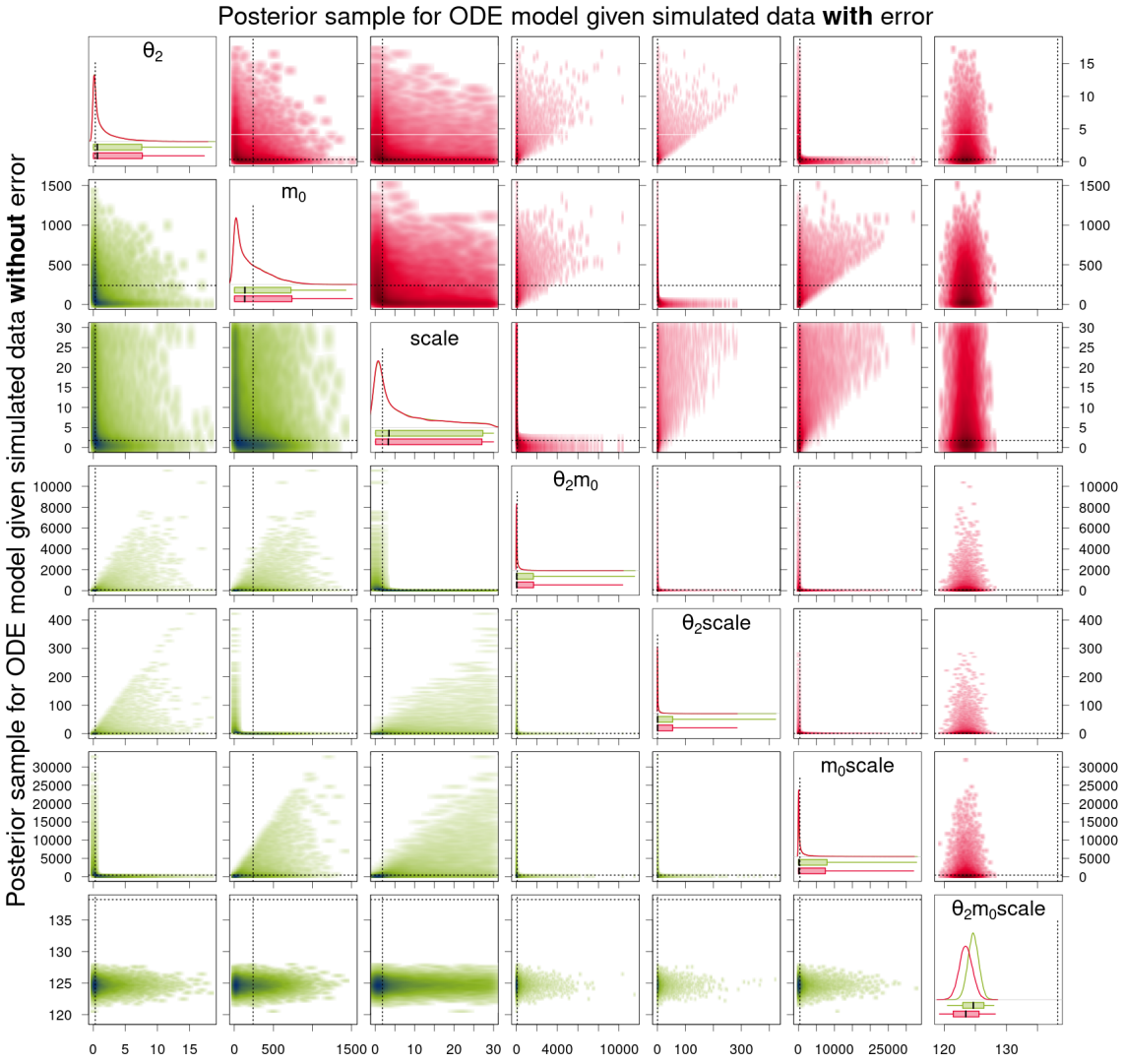
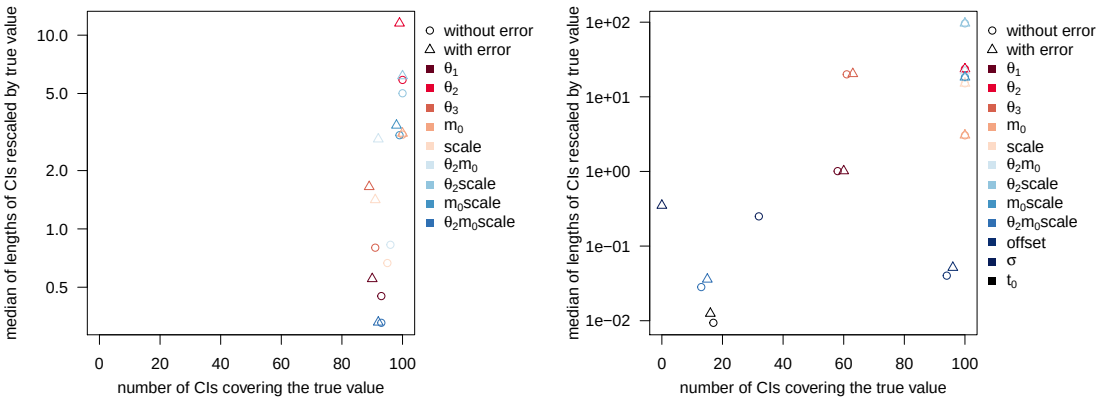


Figure 18: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the ODE model given simulated data without (green, lower triangle) and with (red, upper triangle) measurement error. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values. The dotted lines represent the true parameter values that were used to simulate the data.



(a) SDE results

(b) ODE results

Figure 19: Statistics of posterior samples for the simulated data without and with measurement error aggregated over 100 simulated trajectories. The desirable region of value combinations is in the bottom right corner of each graph.

A.4.2 Sampling results for experimental dataset 2 (for d2eGFP)

Tables 10 and 11 present a summary of the Stan output for the posterior sample of one observed trajectory for d2eGFP for the ODE and the SDE model, respectively, and Figures 20 and 21 compare the density estimates of these two posterior samples. Here, while of course still being symmetric, the posterior sample for the ODE model seems to be unimodal with respect to the parameters θ_1 and θ_3. This is due to the fact that the values of the two parameters are likely to be quite close to each other for this trajectory as can also be seen from the overlapping 95% CIs and the similar mean and median estimates for the SDE model. For the parameter offset, the mean and median estimates from the posterior samples are very similar for the ODE and SDE model, but the 95% CI is a lot wider for the ODE model. For the measurement error parameter σ, the 95% CI for the SDE model is a lot narrower than that for the ODE model and the locations of the samples are quite far apart with a difference in the median estimates of 0.16.

Table 10: Summary of the Stan output for the ODE model given experimental data for d2eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.09	0.079	0.08	0.09	0.11	11585	1.00
θ_2	2.03	1.114	0.08	1.17	8.33	12371	1.00
θ_3	0.09	0.078	0.08	0.09	0.11	11200	1.00
m_0	244.67	0.868	9.82	187.79	761.08	12127	1.00
scale	9.21	0.901	0.35	6.44	28.11	9845	1.00
offset	8.72	0.073	7.52	8.69	10.04	17557	1.00
t_0	0.94	0.011	0.92	0.94	0.96	15806	1.00
σ	0.17	0.053	0.15	0.17	0.18	16365	1.00
$\theta_2 m_0$	367.06	1.893	27.97	121.84	2279.87	11547	1.00
$\theta_2\mathrm{scale}$	12.37	1.907	1.03	4.19	80.33	7681	1.00
$m_0\mathrm{scale}$	1756.24	1.569	94.74	672.69	10085.29	8815	1.00
$\theta_2 m_0\mathrm{scale}$	786.93	0.026	746.72	786.79	828.01	22688	1.00

Table 11: Summary of the Stan output for the SDE model given experimental data for d2eGFP.

	mean	c.v.	2.5%	50%	97.5%	n_{eff}	$\hat{R}$
θ_1	0.11	0.244	0.06	0.10	0.17	1494	1.01
θ_2	10.36	0.292	5.14	10.18	16.77	1226	1.01
θ_3	0.09	0.095	0.08	0.09	0.11	674	1.02
m_0	13.45	0.317	7.06	12.73	23.72	954	1.01
scale	4.93	0.212	3.21	4.83	7.27	785	1.01
offset	8.65	0.005	8.57	8.65	8.74	22392	1.00
σ	0.01	0.067	0.01	0.01	0.01	13124	1.00
$\theta_2 m_0$	130.52	0.229	80.35	127.18	196.70	897	1.01
$\theta_2\mathrm{scale}$	49.67	0.282	27.16	48.09	82.52	838	1.01
$m_0\mathrm{scale}$	65.47	0.363	34.66	60.26	125.01	1343	1.01
$\theta_2 m_0\mathrm{scale}$	615.77	0.092	509.06	614.02	733.51	17424	1.00

For the parameters θ_2, m_0, scale, and their products, the results look somewhat different from those for the eGFP trajectory and those for the simulated data. For the product $\theta_2 m_0\mathrm{scale}$, the 95% CI for the SDE model is again a lot wider than for the ODE model, but here, the CIs do not overlap. For the parameters scale and $\theta_2 m_0$, the 95% CI for the SDE model are again a lot narrower than for the ODE model, and we consider them as practically identifiable for the SDE model but not the ODE model. But here, also for the parameters m_0, $\theta_2 m_0$, and $m_0\mathrm{scale}$, the 95% CI for the SDE model are much narrower than for the ODE model, and the parameters seem to be practically identifiable. For parameter θ_2 the 95% CI for the SDE model is slightly wider than for the ODE model, however, the distribution looks different.

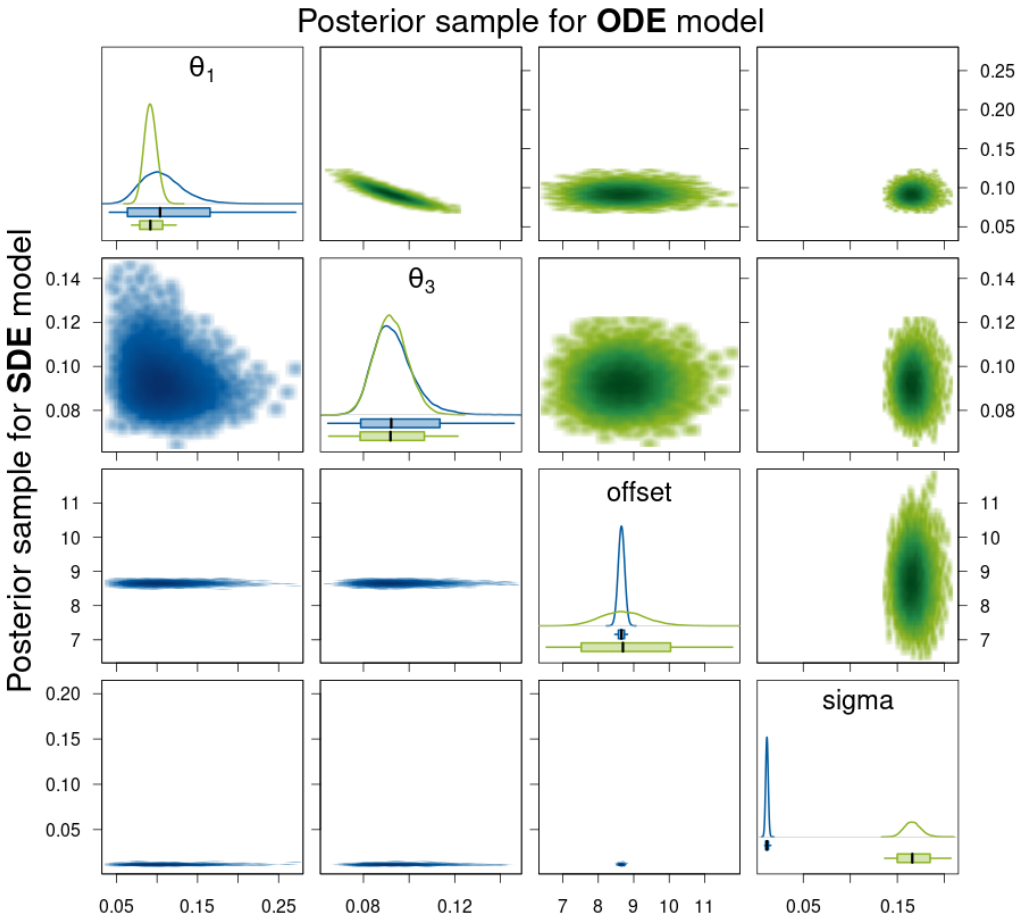


Figure 20: Density estimates of the posterior samples for parameters θ_1, θ_3, offset, and σ for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for d2eGFP. *Diagonal panels:* Marginal densities for the respective parameter and boxplots showing the 95% CI as box, the range of the sample as whiskers, and the median as thick black line. *Off-diagonal panels:* Smoothed scatter plots of the two-dimensional projections of the samples where darker hues signify higher density values.

The statistics of posterior samples aggregated for 100 experimental trajectories for d2eGFP in Table 12 are qualitatively very similar to those for eGFP in Table 12 in the main article. Therefore, we do not repeat the detailed description. We only point out that again unlike for the ODE model, the parameters scale and $\theta_2 m_0$ are identifiable for the SDE model which is indicated by the much narrower median length of the 95% CIs. We also want to mention that here, the median CI lengths for both degradation rate constants θ_1 *and* θ_3 are smaller for the ODE model than those for the SDE model. This is again due to the fact that for the majority of the observed trajectories the parameter values seem to be very close to each other; and therefore, the two modes of the ODE posterior distribution with respect to these parameters simply overlap. This leads to very narrow CIs which is consistent with our results for the simulated data if we consider the width of the individual modes there. However, we would like to remind the reader that the simulated data also showed that often neither of the modes (and sometimes not even the range of sampled values) covered the true parameter. So assuming that a MJP is the most appropriate description for the generating process of the experimental data, the low uncertainty suggested by narrow CIs for the ODE model might be misleading.

A.5 Stan specific diagnostics for the sampling output

Here, we summarize the Stan specific diagnostics described in A.3.3 for the HMC output from Sections 6.2 and 6.3. Tables 13 and 14 present the statistics of the number of divergent transition, Tables 15 and 16 the statistics of the number of times that the user-specified maximal tree depth was

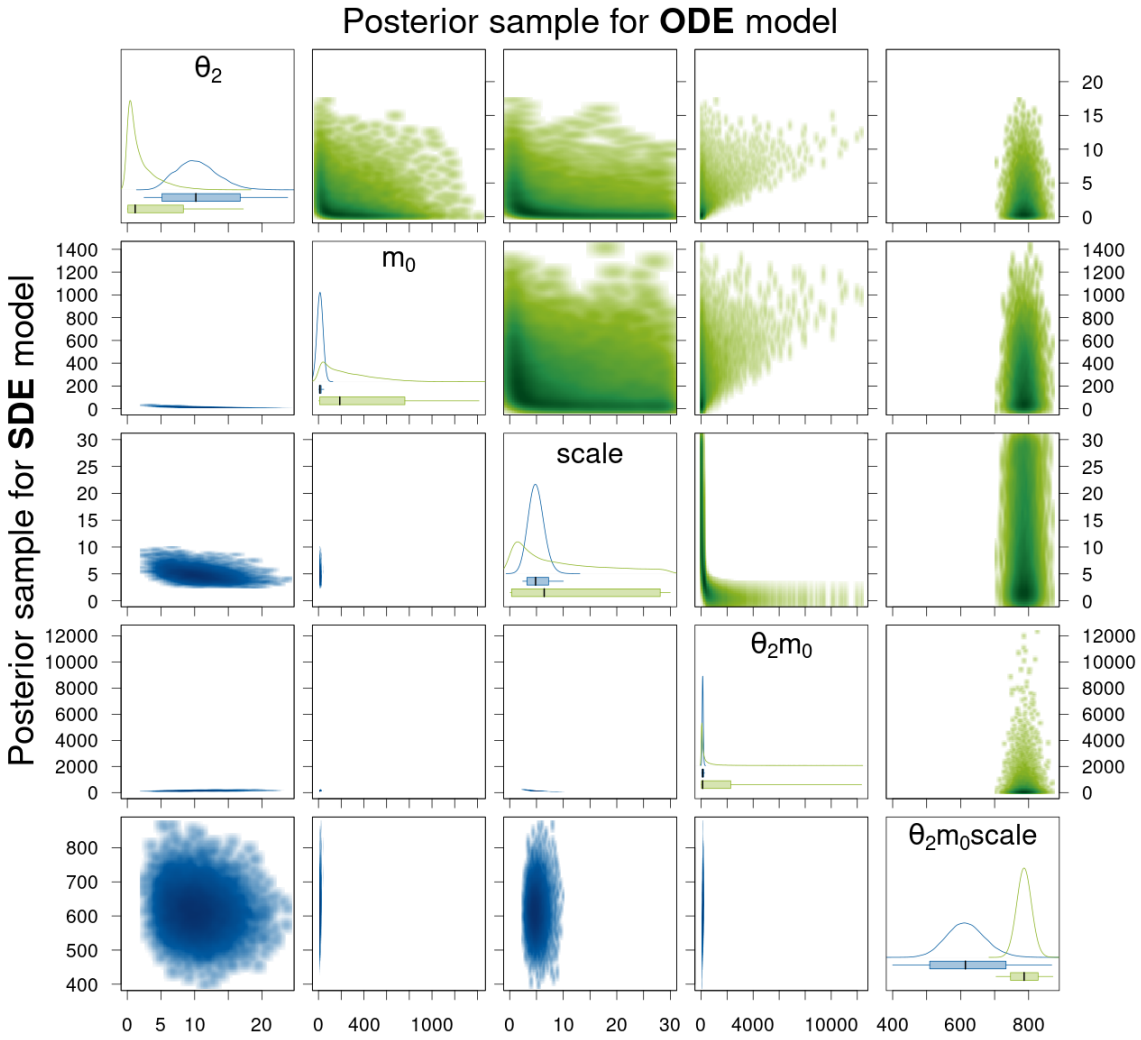


Figure 21: Density estimates of the posterior samples for parameters θ_2, m_0, scale, and their products for the SDE (blue, lower triangle) and ODE (green, upper triangle) model given experimental data for d2eGFP. For a detailed description of the figure's elements, see Figure 20.

exceeded, and Tables 17 and 18 that statistics of the BFMI.

Overall, all three diagnostics show poorer values for the sampling output for the SDE model than for the ODE model. This is not surprising as we sample from a much higher-dimensional distribution for the SDE model. We do not consider the poor diagnostics as a disadvantage of the procedure as they provide information that we do not even have for other MCMC algorithms and thus cannot compare to them.

Table 12: Statistics of posterior samples aggregated for 100 experimental trajectories for d2eGFP.

		length of prior 95% center interval	median length of 95% CIs	c.v. of lengths of 95% CIs
θ_1	ODE	11.05	0.03	0.061
	SDE	11.05	0.12	0.012
θ_2	ODE	11.05	7.88	0.006
	SDE	11.05	11.61	0.269
θ_3	ODE	11.05	0.03	0.062
	SDE	11.05	0.07	0.025
m_0	ODE	884.82	749.92	0.181
	SDE	884.82	76.56	224.650
scale	ODE	28.50	27.55	0.001
	SDE	28.50	5.51	5.550
$\theta_2 m_0$	ODE	6056.48	2048.33	15.537
	SDE	6056.48	172.02	173.641
$\theta_2 \mathrm{scale}$	ODE	228.08	67.95	0.843
	SDE	228.08	34.96	44.862
$m_0 \mathrm{scale}$	ODE	19271.13	9085.56	56.326
	SDE	19271.13	482.87	3408.541
$\theta_2 m_0 \mathrm{scale}$	ODE	113232.70	40.80	30.553
	SDE	113232.70	145.18	83.283
offset	ODE	28.50	2.02	1.751
	SDE	28.50	0.79	1.029
σ	ODE	9.50	0.03	0.006
	SDE	9.50	0.01	0.005

Table 13: Statistics for the Stan diagnostic of the number of divergent transitions for the SDE model. The 100 sampling outputs per dataset are categorized by the number of divergent transitions that occurred after warm-up, i. e. during a total of 20,000 iterations. Hence, the values in columns 1 to 4 sum to 100. Column 5 gives the maximum number of divergent transitions that occurred after warm-up for one sampling output.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	37	10	25	28	1644
simulated data with error	88	4	5	3	568
experimental data for eGFP	93	4	3	0	39
experimental data for d2eGFP	90	3	6	1	540

Table 14: Statistics for the Stan diagnostic of the number of divergent transitions for the ODE model. See Table 13 for a detailed description.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	100	0	0	0	0
simulated data with error	100	0	0	0	0
experimental data for eGFP	99	1	0	0	1
experimental data for d2eGFP	92	8	0	0	2

Table 15: Statistics for the Stan diagnostic of the number of times that the maximal tree depth was exceeded for the SDE model. The user-defined maximal tree depth was set to a value of 15 prior to sampling. The 100 sampling outputs per dataset are categorized by the number of times that the maximal tree depth was exceeded after warm-up, i. e. during a total of 20,000 iterations. Hence, the values in columns 1 to 4 sum to 100. Column 5 gives the maximum number of times that the maximal tree depth was exceeded after warm-up for one sampling output.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	99	0	1	0	11
simulated data with error	10	26	21	43	7126
experimental data for eGFP	25	19	31	25	1976
experimental data for d2eGFP	95	2	3	0	59

Table 16: Statistics for the Stan diagnostic of the number of times that the maximal tree depth was exceeded for the ODE model. See Table 15 for a detailed description.

dataset	none	$1-10$	$11-100$	>100	maximum
simulated data without error	91	6	0	3	2500
simulated data with error	96	0	0	4	2500
experimental data for eGFP	97	0	0	3	2500
experimental data for d2eGFP	100	0	0	0	0

Table 17: Statistics for the Stan diagnostic $\widehat{BFMI}$ for the SDE model. Each of the 100 sampling outputs per dataset consists of 8 HMC chains for each of which $\widehat{BFMI}$ is calculated. Then, we determine the minimum and the mean over the 8 chains. The table presents the mean and the standard deviation (s.d.) of these minima and means aggregated over the 100 sampling outputs per dataset.

dataset	mean of minima	s.d. of minima	mean of means	s.d. of means
simulated data without error	0.03	0.01	0.05	0.01
simulated data with error	0.05	0.02	0.07	0.01
experimental data for eGFP	0.05	0.04	0.08	0.04
experimental data for d2eGFP	0.07	0.05	0.09	0.05

Table 18: Statistics for the Stan diagnostic $\widehat{BFMI}$ for the ODE model. See Table 17 for a detailed description.

dataset	mean of minima	s.d. of minima	mean of means	s.d. of means
simulated data without error	0.95	0.19	1.03	0.06
simulated data with error	0.95	0.15	1.03	0.06
experimental data for eGFP	0.94	0.19	1.03	0.05
experimental data for d2eGFP	0.90	0.23	1.02	0.05