

2 Background

This chapter introduces the key concepts that are needed to understand this thesis. First, we describe the different types of experimental data that are analyzed. Afterwards, the principles of modeling of chemical kinetics are introduced with a focus on the chemical master equation (CME) and its approximations. Finally, we show how experimental data and biological models can be brought together with inference. Inference consists of parameter optimization, identifiability and uncertainty analysis, and model selection.

2.1 Experimental Data

In this thesis, we consider and distinguish two different types of single-cell data \mathcal{D} that provide information about cell-to-cell variability and are frequently collected in biological research.

Single-cell snapshot data $\mathcal{D} = \{\{y_j(t_k)\}_j\}_{k=1}^{n_t}$ provide single-cell measurements for n_t time instances t_k (see Figure 2.1A). Common approaches to generate these data are, e.g., flow cytometry (Davey & Kell, 1996) or single-cell microscopy (Miyashiro & Goulian, 2007). A key advantage of these technologies is the possibility of measuring many genes of plenty of single-cells with low costs. As the cells are not tracked over time, no information about the time-course of an individual cell is available.

To obtain temporal information single-cell time-lapse data $\mathcal{D} = \{\{y_j(t_k)\}_{k=1}^{n_t}\}_j$ (see Figure 2.1B) are required. Single-cell time-lapse data are typically obtained by conducting fluorescent time-lapse microscopy (Muzzey & Oudenaarden, 2009) followed by single-cell tracking (Schroeder, 2011) and image analysis. This approach provides a smaller number of cells than the technologies described before and the generation of single-cell time series is expensive and time-consuming. On the other hand, cells are tracked over time yielding a higher information content of the data.

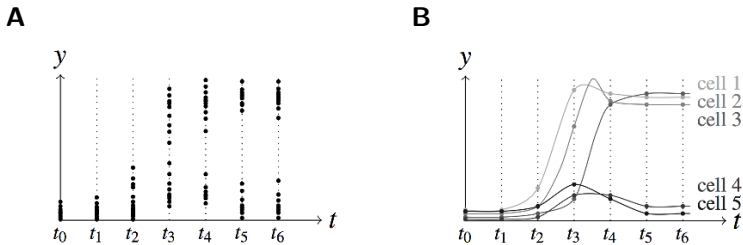


Figure 2.1: Measurement data at the single-cell level adopted from Hasenauer (2013): (A) Illustration of single-cell snapshot data of some measurement y . (B) Illustration of single-cell time-lapse data for five individual cells.

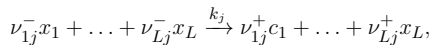
2.2 Modeling Chemical Kinetics

For the detailed analysis of single-cell data, mechanistic mathematical models are used. One possibility is the use of stochastic chemical kinetics, which model biochemical reaction networks as continuous-time discrete-state Markov chains (CTMCs). The time evolution of a CTMC is governed by the CME. A process defined by the CME can either be simulated with the stochastic simulation algorithm (SSA) or its solution can be approximated e.g. with the moment equations (ME). While stochastic modeling is especially important in the case of low-copy numbers, we assume that for high numbers of molecules the system can be described by its average behavior. This can be modeled in a deterministic way by first order ordinary differential equations (ODEs) describing the evolution of concentrations of the species.

2.2.1 Stochastic Chemical Kinetics

Stochastic models are mostly used to describe a biological process, when it is important to consider that molecules only appear in whole numbers (Wilkinson, 2009; Resat *et al.*, 2009). This discreteness yields a stochasticity in the dynamics of the molecules and especially has to be taken into account if only few numbers of molecules are present.

Stochastic chemical kinetics describe the time evolution of a chemical system consisting of L chemical species x_1, \dots, x_L that interact inside a volume Ω through M reactions R_1, \dots, R_M . A reaction R_j has the form



with stoichiometric coefficients $\nu_{ij}^+, \nu_{ij}^- \in \mathbb{N}_0$ and reaction rate k_j . A state of the system is represented by a vector $\mathbf{x}(t) \in \mathbb{N}_0^L$. Each entry of the vector is the number of molecules of the corresponding species. The stoichiometric matrix $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_M) \in \mathbb{R}^{L \times M}$ is defined by $\{S_{ij}\} = \{\nu_{ij}^+ - \nu_{ij}^-\} := \{\nu_{ij}\}$. Each entry of the matrix describes the change in the number of molecules of species x_i due to a reaction of type j , i.e., the state \mathbf{x} changes to $\mathbf{x} + \mathbf{s}_j$ after reaction R_j took place. The probability that reaction R_j happens in the next infinitesimal time interval $[t, t + dt)$ is $a_j(\mathbf{x})dt$, with propensity function $a_j(\mathbf{x})$.

Several assumptions are typically made when deriving a model of a biological process, e.g. that the system has a constant volume Ω and is well-stirred, i.e., the probability of some molecules of a species being in one particular region is uniform over the volume (Gillespie, 2007). We consider zero-order reactions, which are independent of the number of molecules, unimolecular reactions, in which just a single molecule is necessary to conduct the reaction, and bimolecular reactions, for which two molecules need to collide. Higher order reactions can easily be integrated into the methods proposed in this thesis.

2.2.2 Chemical Master Equation

The CME governs the evolution of the probability that the stochastic process is in a particular state, given by $p(\mathbf{x}, t)$, over time (Gillespie, 1992). The probability $p(\mathbf{x}, t | \mathbf{x}_0, t_0)$ is conditioned on the system being in state \mathbf{x}_0 at time t_0 . To obtain an evolution equation the probability $p(\mathbf{x}, t + dt | \mathbf{x}_0, t_0)$ is first derived in terms of $p(\mathbf{x}, t | \mathbf{x}_0, t_0)$, by assuming dt is small enough that at most one reaction can occur in the time interval $[t, t + dt)$. One possibility for the system being in state \mathbf{x} at time $t + dt$ is that it already has been in this state and no reaction has taken place since time t , which happens with probability $1 - \sum_{j=1}^M a_j(\mathbf{x})dt + \mathcal{O}(dt)$. Another scenario is that the system has been in state $\mathbf{x} - \mathbf{s}_j$ and a reaction of type j occurred with probability $a_j(\mathbf{x} - \mathbf{s}_j)dt$, which yields M more possibilities. After summing up the probabilities and taking the limit $dt \rightarrow 0$, we obtain the CME

$$\frac{dp(\mathbf{x}, t | \mathbf{x}_0, t_0)}{dt} = \sum_{j=1}^M [p(\mathbf{x} - \mathbf{s}_j, t | \mathbf{x}_0, t_0) a_j(\mathbf{x} - \mathbf{s}_j) - a_j(\mathbf{x}) p(\mathbf{x}, t | \mathbf{x}_0, t_0)],$$

with initial condition

$$p(\mathbf{x}, t = t_0 | \mathbf{x}_0, t_0) = \begin{cases} 1, & \mathbf{x} = \mathbf{x}_0 \\ 0, & \mathbf{x} \neq \mathbf{x}_0 \end{cases}.$$

If we neglect \mathbf{x}_0 and t_0 for a simpler notation we obtain

$$\frac{dp(\mathbf{x}, t)}{dt} = \sum_{j=1}^M [p(\mathbf{x} - \mathbf{s}_j, t) a_j(\mathbf{x} - \mathbf{s}_j) - a_j(\mathbf{x}) p(\mathbf{x}, t)],$$

with initial condition $p(\mathbf{x}, t_0) = p_0(\mathbf{x})$. The CME indeed completely determines the probability $p(\mathbf{x}, t | \mathbf{x}_0, t_0)$ and thus totally describes the system. However, it consists of a system of coupled ordinary differential equations (ODEs), with one ODE for every possible state of the system. Since the state space of a biological system is mostly high dimensional or even infinite dimensional, the CME can only be solved analytically or in a feasible numerical way for a few simple cases (e.g. (Jahnke & Huisinga, 2007)).

2.2.3 Stochastic Simulation Algorithm

Instead of solving the CME, it is possible to simulate samples in form of trajectories and thereby recover the underlying probability distribution. This is motivated by the fact that the chance of a particular trajectory being simulated corresponds to the probability given by the CME. A possibility to obtain trajectories is the SSA (Gillespie, 1977). This algorithm enables an exact simulation of trajectories consistent with the probability distribution and the transition probabilities that are associated with the CME. For the direct method of stochastic simulation we define

- the sum over all reaction propensities $a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x})$,
- the time τ to the next reaction,
- the index j of the next reaction.

It can be shown that τ is exponentially distributed with rate $a_0(\mathbf{x})$ and j has density $\frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}$, which yields the following algorithm:

Algorithm 2.1: Direct method

Input: Initial condition $\mathbf{x}_0 \in \mathbb{N}_0^L$,
 final simulation time t_{end} ,
 reaction propensity functions $a_j(x), j = 1, \dots, M$,
 stoichiometric matrix $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_M) \in \mathbb{Z}^{L \times M}$.

Result: Time trajectory of state vector $\mathbf{x}(t)$.

Set $t \leftarrow 0$ and $\mathbf{x} \leftarrow \mathbf{x}_0$.

while $t < t_{\text{end}}$ **do**

 Evaluate reaction propensity functions $a_j(\mathbf{x})$ and calculate

$$a_0(\mathbf{x}) = \sum_{j=1}^M a_j(x).$$

 Generate two uniformly distributed independent random numbers r_1 and r_2 .

 Calculate the time until the next reaction takes places by $\tau = \frac{1}{a_0(\mathbf{x})} \log(1/r_1)$.

 Find the index j of the next reaction that satisfies $\sum_{j=1}^M a_j(\mathbf{x}) > r_2 a_0(x)$.

 Update the state of the system $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{s}_j$.

 Update the time $t \leftarrow t + \tau$.

end

An example of trajectories obtained by this method is shown in Figure 2.2 for a conversion process (see Section 3.3.2). The computation can be inefficient if lots of events have to be simulated. Therefore, approximations such as τ -leaping have been introduced (for further information see (Gillespie, 2007)).

2.2.4 Method of Moments

A possibility to approximate the solution of the CME and thereby avoid the computational costs of the SSA is the method of moments (Engblom, 2006). This method computes the moments of $p(\mathbf{x}, t)$, i.e., the mean

$$m_i(t) = \sum_{x \in \Omega} x_i p(\mathbf{x}, t), \quad i = 1, \dots, L,$$

of species x_i , and higher order moments such as the covariance

$$C_{ij}(t) = \sum_{x \in \Omega} (x_i - m_i(t))(x_j - m_j(t))p(\mathbf{x}, t), \quad i, j = 1, \dots, L,$$

of species x_i and x_j . The time evolution of the moments is described by a set of ODEs, the so-called moment equations (MEs). If the system comprises bimolecular reactions, the calculation of higher order moments is recursive, i.e., the evolution of a moment of order k depends on moments of order $k + 1$. In this case moment closure techniques must be applied, introducing an approximation error (Lee *et al.*, 2009). Formulas for the first and second order moments of system with at most bimolecular reactions, can be found in (Engblom, 2006, Proposition 2.5.). The first and second order moments, namely mean and variance, of the solution statistics for a conversion process are depicted in Figure 2.2. If a system comprises low- and medium/high-copy species the method of conditional moments (Hasenauer *et al.*, 2014a) can be used. This method conditions the moments of species with medium or higher abundance on the states of species that are only present in low-copy numbers. Therefore, it accounts for the stochasticity of the processes, arising due to the discreteness of the low-abundance species. The method avoids the computational costs arising from a full stochastic description of the system using MEs for the medium and high-copy species.

2.2.5 Reaction Rate Equation

In the limit of large numbers of molecules, the system behaves in a more deterministic way and the importance of considering single molecules vanishes. Therefore, measurements are at a continuous level, in contrast to the discrete state space of stochastic modeling. The evolution of the system is captured by the reaction rate equations (RREs) (Resat *et al.*, 2009; Gillespie, 2007)

$$\frac{d\mathbf{x}(t)}{dt} = \sum_{j=1}^M \mathbf{s}_j a_j(\mathbf{x}(t)).$$

For some simple systems an explicit formula for the solution of the RREs can be derived, but mostly numerical integration is need. Nevertheless, deterministic simulations of a system are generally faster than a stochastic simulation (Szekely & Burrage, 2014).

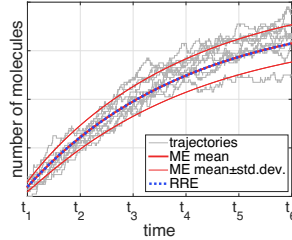


Figure 2.2: Example of trajectories of one species of a conversion process obtained by the SSA (gray), the corresponding approximation with MEs (red) and RREs (blue), where the mean described by ME and the RRE coincident.

2.3 Parameter Inference

The idea of parameter inference is to combine observed data \mathcal{D} and a model \mathcal{M} , which for example has been derived with techniques presented in the previous section. Such a model comprises parameters, for example kinetic rates or initial conditions, and some of these parameters denoted by $\theta \in \mathbb{R}^{n_\theta}$ may be unknown, because either they are not measured or it is impossible to measure them.

2.3.1 Parameter Estimation

A common approach to estimate the parameters of a model is to maximize the likelihood function

$$L(\theta) = p(\mathcal{D}|\theta),$$

which describes the conditional probability of observing \mathcal{D} given θ . Due to better numerical properties for optimization, usually the negative log-likelihood function

$$J(\theta) = -\log L(\theta)$$

is minimized. The parameters θ^{ML} that maximize the likelihood function (or minimize the negative (log-)likelihood function) are called the maximum likelihood estimates (MLE).

In a Bayesian framework we can additionally incorporate prior knowledge about the parameters using the prior distribution $p(\boldsymbol{\theta})$ (Hastie *et al.*, 2009). Applying Bayes' theorem yields the posterior distribution of the parameters

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The parameters $\boldsymbol{\theta}^{\text{MAP}}$ that maximize the posterior distribution are the maximum a posteriori estimate (MAP), the Bayesian counterpart of the MLE. The evaluation of the normalizing constant $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ can be computationally expensive or unfeasible. However, this constant can be neglected for optimization and uncertainty analysis, as it is only needed for model selection based on Bayes factors (Raftery, 1999). The minimization of the negative log-likelihood function can be efficiently performed using multi-start local optimization. For this, the initial values for the optimizer are e.g. obtained by Latin hypercube sampling and then are chosen in a sequential way, such that the corresponding objective function values are decreasing (Raue *et al.*, 2013). For the optimization procedure the calculation of the gradient is of great importance, as the derivative of the objective function is used to determine the next parameter value. For the calculation of the derivatives finite differences or sensitivity analysis can be used (Sengupta *et al.*, 2014). Sensitivity analysis describes the derivatives of the objective function with respect to the parameters. Using them, the gradient can be calculated numerically more robustly. Additionally, we use log-transformed parameters $\boldsymbol{\xi} = \log(\boldsymbol{\theta})$ due to better convergence properties.

If the likelihood cannot be expressed analytically or is computationally too costly to evaluate, so-called likelihood-free parameter estimation methods are required. This class of methods circumvents the calculation of the likelihood function and is also known under the name approximate Bayesian computing (ABC) (Csilléry *et al.*, 2010). We explain these methods in more detail in Section 4.2, as they are the focal point of the work described in Chapter 4.

2.3.2 Identifiability and Uncertainty Analysis

Due to the structure of the examined system and limitations of the available data some parameters can be non-identifiable (Raue *et al.*, 2009), i.e., the parameter can not be

determined from the data. If this is the case even for perfect data, the parameter is structurally non-identifiable. If the parameter can not be identified due to measurement noise or too little data, the parameter is practically non-identifiable. Studying these uncertainties is an important step of parameter inference and explained in the following.

A common approach to analyze uncertainties of the parameters is to calculate confidence intervals, e.g. asymptotic confidence intervals based on the curvature of the likelihood, such as the hessian, or finite sample confidence intervals based on profile likelihoods (for further information see (Raue *et al.*, 2009)). A parameter θ is practically identifiable from the corresponding data, if the corresponding confidence intervals are finite.

In a Bayesian context, in which parameters are treated as random variables, we can get information about the uncertainty of the estimates by considering the whole posterior distribution. Because of a possibly high dimension of the parameter space or the lack of a closed form for the posterior, the use of numerical sampling from the posterior distribution is required. Samples from the posterior distribution can be obtained by Markov chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996).

2.3.3 Model Selection

The last step of parameter inference is to select an optimal model of out a given set of candidate models $\mathcal{M}_1, \dots, \mathcal{M}_l$ that have been derived for some data \mathcal{D} . On the one hand, the chosen model should fit the data very well, which can be easily improved by increasing the number of parameters. On the other hand, the model should be as simple as possible to provide reliable predictions and avoid unnecessary uncertainties. We introduce two existing criteria for model selection that try to solve the trade-off between over- and underfitting of the data. Both criteria consist of a term with the likelihood value of the maximum likelihood estimate and a penalization term for a higher complexity of the model.

The Akaike information criterion (AIC) is based on information theoretical concepts (Akaike, 1998). It gives an estimate for Kullback-Leibler divergence between the densities of the true unknown model and of a candidate model \mathcal{M}_k by

$$\text{AIC}_k = -2 \log(p(\mathcal{D}|\boldsymbol{\theta}^{\text{ML},k})) + 2n_{\theta,k},$$

with $\boldsymbol{\theta}^{\text{ML},k}$ denoting the MLE for model \mathcal{M}_k and $n_{\theta,k}$ denoting the number of parameters of the model. A low value of the AIC indicates that less information has been lost considering the candidate model and therefore a higher reliability. We reject models with $\Delta_{\text{AIC}} = \text{AIC}_k - \text{AIC}_{\min} > 10$ as proposed by Burnham & Anderson (2002).

A Bayesian criterion for model selection can be derived by examining the posterior probability $p(k|\mathcal{D})$ of model \mathcal{M}_k (see (Schwarz *et al.*, 1978) for further information). This criterion is called the Bayesian information criterion (BIC),

$$\text{BIC}_k = -2 \log(p(\mathcal{D}|\boldsymbol{\theta}^{\text{ML},k}) + \log(n_{\mathcal{D}})n_{\theta,k},$$

with $n_{\mathcal{D}}$ denoting the number of data points. As with the AIC, the model with the lowest BIC is chosen and we reject models with $\Delta_{\text{BIC}} = \text{BIC}_k - \text{BIC}_{\min} > 10$ (Raftery, 1999).

In summary, this chapter outlined the key principles that are used in the following chapters of this thesis. We introduced single-cell snapshot data and single-cell time-lapse data, which possess different information contents and number of data points. We discussed different approaches to solve the CME, ranging from exact solutions obtained with the SSA to approximations with MEs and showed the link to deterministic modeling by RREs. Moreover, this chapter contains an introduction to parameter inference, including parameter estimation, identifiability and uncertainty analysis, and model selection. We presented the approach of maximum likelihood estimation using multi-start local optimization, and defined the posterior distribution that is used in a Bayesian context. For identifiability and uncertainty analysis, profile likelihoods and MCMC sampling schemes can be used. Finally, we introduced the AIC and BIC, two criteria used for model selection.

Analysis of Single-Cell Data

ODE Constrained Mixture Modeling and Approximate
Bayesian Computation

Loos, C.

2016, XXI, 92 p. 26 illus., Softcover

ISBN: 978-3-658-13233-0