LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Jan Mathias Köhler

Approximate Bayesian parameter estimation in stochastic models of biochemical reactions

Master Thesis
Institut für Statistik
Ludwig-Maximilians-Universität München
Betreuer: Prof. Dr. Dr. Fabian Theis
und: M. Sc. Ivan Kondofersky
angefertigt am: Helmholtz Zentrum München,
AG Computational Modeling in Biology,
Institute of Bioinformatics and Systems Biology
07. Februar 2012

Contents

List of Figures

List of Tables

1 Introduction

In systems biology one of the main interests is to thoroughly understand biological systems. Therefore, a system is modelled into a simplified version, which can easier be analysed, predicted and optimized than the original biological system itself (Chou & Voit 2009).

Research is interested in systems with a small amount of molecules interacting. Often, e.g. in the case of cell fate, molecular events of one cell may have an influence on every subsequent process (Munsky & Khammash 2006). This situation cannot be modelled using deterministic models as they are not able to capture the randomness of the biological processes, thus, stochastic models are required. Boys et al. (2008, p. 125) state that *"conventional deterministic chemical kinetics fail to describe the development of systems of coupled biochemical reactions correctly when both concentrations of reactants and reaction rates are low".*

To understand the biological system and to be able to develop and distinguish among possible systems, information about the parameters of the system, i.e. kinetic rate constants, is necessary. This knowledge is important *"for the end-applications like analysing system properties (e.g. robustness) or predicting the effects of genetic perturbations".* (Poovathingal & Gunawan 2010, p. 1)

This thesis deals with parameter estimation in stochastic systems. As a computation of the likelihood is analytically very complex and often not possible, the likelihood-free methods Approximate Bayesian Computation (ABC) and ABC Sequential Monte Carlo (SMC) are used. Parameter estimation is currently a main research topic and highly important as Poovathingal & Gunawan (2010, p.1) opine: *"Despite the availability of high-throughput cell biology, the estimation of unknown (kinetic) model parameters from experimental data is still considered as the bottleneck in biological model identification, especially for dynamical models".* Boys et al. (2008, p. 125) reckon that *"One of the most important challenges in developing systems-level models of stochastic gene regulatory processes is how to estimate the values of the key rate parameters."*

In this thesis, different distance functions for the ABC (SMC) algorithm are defined. They are based on the mean, standard deviation, negentropy, probability and cumulative distribution function, and on the correlation of the data. In a

computational study, using two distinct gene expression models, they are tested thoroughly with different prior distributions. The aim is to get an understanding, which distance functions result in the best estimation.

Chapter 2 contains theoretical background of modelling chemical reactions and methods for parameter estimation. In section 2.3 the ABC and ABC SMC algorithm are introduced. Two gene expression models, the one-stage and two-stage model, which are used in the computational study, are explained in chapter 2.4. The distance functions and summary statistics, which are used for the ABC (SMC), are explained in chapter 3. Chapter 4 shows how the ABC and ABC SMC algorithm have been implemented, introduces the parameter settings for the computational study and defines a statistic to evaluate the quality of parameter estimation of the posterior distribution. An overview of the conducted simulations and the computational time is additionally given. The results of the computational study for the ABC algorithm is presented in chapter 5. The results for the ABC SMC are discussed in chapter 6. A summary and an outlook is given in chapter 7.

2 Background

This chapter deals with the theoretical background. In the first two parts of the chapter, the modelling of chemical reactions is described and an overview of methods for parameter estimation is given. In section 2.3 the Approximate Bayesian Computation algorithm is presented. The last section explains the gene expression models, which are used throughout the thesis.

2.1 Modelling chemical reactions

There is a vast amount of ways to model genetic regulatory systems and thus biochemical reactions. De Jong (2002) and Karlebach & Shamir (2008) give an overview of different ways. The second authors divide the systems into three categories: I) logical models, including e.g. boolean networks or petri nets, II) continuous models, including, among others, linear models or ODEs and III) single-molecule level models, which contain the later described stochastic simulation algorithm.

In general a system of (bio-)chemical reactions can be specified by reactions written in common chemistry stoichiometric notation of the form

$$c_1 X_1 + ... + c_k X_k \rightarrow p_1 Y_1 + ... + p_l Y_l. \tag{1}$$

The reactants $X_1, ...X_k$ are consumed and transformed into the products $Y_1, ...Y_l$. The coefficients $c_1, ...c_k$ and $p_1, ...p_l$ denote the number of molecules of reactants and products consumed and produced, respectively. Usually this equations show elementary reactions, i.e. the transformation from reactants to products does not involve other intermediate reactions.

In the following the different modelling possibilities are illustrated with a common biochemical model (Thattai & van Oudenaarden 2001), the single gene expression model, also known as two-stage model, which is shown in figure 1. The system consists of three molecule species DNA, mRNA and protein and the fol-

lowing four reactions.[1]

$$DNA \xrightarrow{\theta^{(1)}} DNA + mRNA \qquad \text{transcription} \tag{2a}$$

$$mRNA \xrightarrow{\theta^{(2)}} mRNA + protein \qquad \text{translation} \tag{2b}$$

$$mRNA \xrightarrow{\theta^{(3)}} \emptyset \qquad \text{degradation of mRNA} \tag{2c}$$

$$protein \xrightarrow{\theta^{(4)}} \emptyset \qquad \text{degradation of protein} \tag{2d}$$

In this model the mRNA is a transcript of the DNA, and the mRNA is further translated into protein. Both the protein and mRNA are degraded, thus lowering the existing amount of molecules.

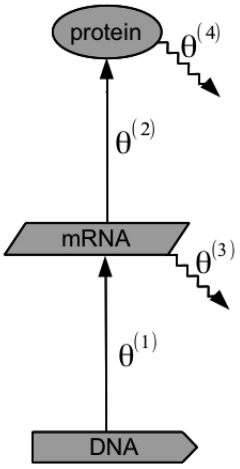


Figure 1: Scheme of the single gene expression model. Solid arrows indicate synthesis, jagged arrows indicate degradation. DNA is transcribed into mRNA with rate $\theta^{(1)}$. mRNA decays with rate $\theta^{(3)}$ and is translated into protein with rate $\theta^{(2)}$. The protein decays with rate $\theta^{(4)}$.

The *rate constants* (sometimes called *reaction rates*) are denoted by $\theta^{(1)}, ..., \theta^{(4)}$ and quantify the speed of the reaction.[2] Note that the rate constants may change due to changes in the parameters which influence the reaction rate, such as volume, temperature and pressure.

[1] The following chapter is partly based on Hayot & Jayaprakash (2008).

[2] From the rate constants $\theta^{(3)}$ and $\theta^{(4)}$ one can derive the half-life of the mRNA $\tau_3 = log(2)/\theta^{(3)}$ and protein $\tau_4 = log(2)/\theta^{(4)}$ respectively (Schwanhäusser et al. 2011, supplementary).

In the following part, possibilities to model chemical reactions are described.

2.1.1 Markov Jump Process (MJP)

This paragraph is partly based on Boys et al. (2008). Taking the rate constants $\theta^{(1)}, ..., \theta^{(4)}$ into account the probability that DNA is transcribed to mRNA in a small time interval $\mathrm{d}t$ is $\theta^{(1)}\mathrm{d}t + o(\mathrm{d}t)$.

In modelling this system as a Markov Jump Process (MJP), the number of molecules of each species is of interest and defines a state. If a reaction occurs, it will change the state of the system, which defines the number of molecules of each species. Reaction 2a from the two-stage model for instance leads to the change from state S_1 to state S_1' as

$$\begin{gathered} S_1 \to S_1' \\ \begin{pmatrix} DNA \\ mRNA \\ protein \end{pmatrix} \to \begin{pmatrix} DNA \\ mRNA + 1 \\ protein \end{pmatrix}. \end{gathered} \tag{3}$$

This state change has the probability

$$P(S_1', t + \mathrm{d}t | S_1, t) = a_1 \mathrm{d}t + o(\mathrm{d}t) \tag{4}$$

with $o(\mathrm{d}t)/\mathrm{d}t \to 1$ as $\mathrm{d}t \to 0$ and $a_1 = \theta^{(1)} d(t)$, where $d(t)$ is the amount of DNA at time t.

For all reactions with states $S_1, S_1', ..., S_4'$ the probabilities of a state change are given by

$$P(S_r', t + \mathrm{d}t | S_r, t) = a_r \mathrm{d}t + o(\mathrm{d}t) \quad \forall r. \tag{5}$$

The propensity functions are[3]

$$a_1 = \theta^{(1)} d(t) \quad a_2 = \theta^{(2)} m(t) \quad a_3 = \theta^{(3)} m(t) \quad a_4 = \theta^{(4)} p(t), \tag{6}$$

with $m(t)$ and $p(t)$ denoting the number of mRNA and protein molecules. The propensity function is a function whose product with dt gives the probability a_r of reaction r occuring in the next infinitesimal time dt.

Because in the model each transition probability only depends on the current state and not on previous states of the system, the system is Markovian of order one and the chemical reactions can be seen as a MJP. Thus the time τ, until the next reaction occurs, is exponentially distributed with $\tau \sim Exp(\sum_r a_r)$, and the probability for an occurring reaction is $a_r / \sum_i a_i$.

2.1.2 Reaction Rate Equations (RRE)

Often the chemical rate equations are formulated as a set of ordinary differential equations (ODEs). Considering all four equations (2a)–(2d), the *rate equations* are

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \theta^{(1)} d(t) - \theta^{(3)} m(t) \tag{7a}$$

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \theta^{(2)} m(t) - \theta^{(4)} p(t) \tag{7b}$$

$$\frac{\mathrm{d}d(t)}{\mathrm{d}t} = 0. \tag{7c}$$

The concentration of DNA does not change over time, as the DNA molecules are not used in the transcription to mRNA, so $d(t) = c \in \mathbb{R}$. Thus, in equation (7c), $d(t)$ stays constant and, therefore, this equation is often not considered. In the steady state it holds that $m(t) = \frac{\theta^{(1)}}{\theta^{(3)}}$ and $p(t) = \frac{\theta^{(1)}\theta^{(2)}}{\theta^{(3)}\theta^{(4)}}$ (see chapter 2.4.2). Considering equations (7a) and (7b), biochemical reactions can be described by a set of coupled ODEs, which are often called the *reaction rate equations* (RRE) (Gillespie 2007).

[3]Note that the $\theta's$ used in calculating the propensity function have the dimension of an inverse time. Regarding the gene expression model, this is the case as can be seen in 7a and 7b. If in a reaction two reactants are involved, the chemical constant θ has dimension of volume/time, and to calculate the propensity functions one needs to divide each θ by the volume V.

These equations state the change in the concentration of each species as a function of the concentration level of some other species. It is assumed that the concentrations are continuous.

Simple ODE systems can be solved analytically and the set of resulting algebraic equations describe how the mean of the concentration evolves over time. For an overview of ODEs in biology with examples one can consider Klipp et al. (2005).

Originally it was proposed by Goodwin (1963) to model regulatory networks with ODEs. Chen et al. (1999), for instance, use ODEs to model biological regulatory networks.

2.1.3 Chemical Master Equation (CME)

Another possibility to model chemical reactions is through a chemical master equation (CME). It describes the evolution over time of a system. The biochemical system is modelled as being in only one of several states at a given time. The changes between states are probabilistic. A derivation can be found in Higham (2008) or in more detail in Gillespie (1992).

For a system of chemical reactions, let $P(\mathbf{x}, t)$ be the probability that $\mathbf{X}(t) = \mathbf{x}$ where $\mathbf{X}$ is an S-dimensional vector representing the number of molecules of each of the S species. A well stirred volume, i.e. the concentrations are homogeneous, is implied. Assume the system had the state $\mathbf{x} - \mathbf{v_r}$ at time t. After reaction r occurred in the interval $[t, t+\mathrm{d}t]$, the system is in state $\mathbf{x}$. So $\mathbf{v_r}$ defines the changes in the number of molecules of the S species through reaction r. The propensity functions are defined by a_r. The Chemical Master Equation is given by

$$\frac{\mathrm{d}P(\mathbf{x}, t)}{\mathrm{d}t} = \sum_{r=1}^{R} a_r(\mathbf{x} - \mathbf{v_r}) P\left(\mathbf{x} - \mathbf{v_r}, \mathbf{t} - a_r(\mathbf{x})\right) P(\mathbf{x}, t). \tag{8}$$

For the simple gene expression model, the CME is (Hayot & Jayaprakash 2008)

$$\begin{aligned}\frac{\partial P(n_P, n_M, t)}{\partial t} &= \theta^{(2)} n_M [P(n_P - 1, n_M, t) - P(n_P, n_M, t)] \\ &+ \theta^{(3)} [(n_M + 1) P(n_P, n_M + 1, t) - n_M P(n_P, n_M, t)] \\ &+ \theta^{(1)} n_D [P(n_P, n_M - 1, t) - P(n_P, n_M, t) \\ &+ \theta^{(4)} [(n_P + 1) P(n_P + 1, n_M, t) - n_P P(n_P, n_M, t)]\end{aligned} \tag{9}$$

where $P(n_P, n_M, t)$ is the probability that n_P proteins and n_M mRNAs are found in the volume at time t. The parameter n_D indicates the number of DNA molecules.

The CME can be solved exactly, which is possible only for systems involving a small number of species. Due to the course of dimensionality, the CME cannot be solved with increasing number of species (Gillespie 1977). Therefore, stochastic simulations are necessary. In the following the stochastic simulation algorithm is introduced.

2.1.4 Stochastic Simulation Algorithm (SSA)

The stochastic simulation algorithm (SSA) takes the assumption that only reactions involving at most two species as reactants occur[4], and that only one reaction can happen at a time.

Biological systems involve many species and many reactions and, therefore, it is often too complicated to solve the CME numerically. Gillespie (1977) introduced an algorithm and showed that it gives the same results as solving the CME of a system. He named the algorithm stochastic simulation algorithm, which is also known as Gillespie's algorithm. A detailed derivation of the algorithm can be found, for instance, in Higham (2008).

The aim of SSA is to simulate trajectories of the species over time. It is based on the assumption that the time to the next reaction is exponentially distributed, and the probability of each reaction is known.

In the following the SSA is described for a system with S species, where the number of molecules is denoted by $X_j, j = 1, ..., S$. There are R reactions with

[4] in nature the probability that three or more species react, i.e. the molecules clash together, is very low and thus negligible

reaction rates $\theta^{(r)}, r = 1, ..., R$ and propensity functions $a_r, r = 1, ..., R$. The time is measured from $t = 0$ until T_{obs}.

Stochastic Simulation Algorithm

1. Initialize values for $X_j, j = 1, ..., S$ and $\theta^{(r)}, r = 1, .., R$ and set $t = 0$.
2. Calculate a_r for all r and $a_0 = \sum_{r=1}^{R} a_r$.
3. Draw the time to the next reaction $\tau \sim Exp(a_0)$.
4. Draw which reaction r occurs where $P(r) = a_r/a_0$.
5. Set $t = t + \tau$ and update the number of molecules according to reaction r.
6. Continue with step 2 until maximum time T_{obs} is reached or maximum number of iterations are executed.

The main advantage of using SSA to model chemical reactions is that SSA is applicable even for small populations. Due to biological variance, small populations show deviation from predictions of deterministic ways of modelling, e.g. modelling by ODEs. Moreover, models such as the RRE consider the amount of molecules as a continuous concentration. This is approximately correct for a large number of molecules. If there are only a few molecules, the discreteness plays an important role, which is not considered by RRE. As it is shown by Tian et al. (2007), parameter estimation and model prediction is poor if deterministic models are used for stochastic data.

2.2 Literature based overview of methods for parameter estimation

In this section a non-exhaustive overview of methods for parameter estimation is given. To the knowledge of the author, there does not exist a publication yet which gives an overview of the different parameter estimation methods. A short overview of methods can be found in the introduction of Lillacci & Khammash (2010).

Poovathingal & Gunawan (2010) describe three ways for estimating the kinetic parameters for models described as a CME and which have a low number of

molecules. The first criterion used is the likelihood function, which can be derived from the CME model. The parameters are estimated by maximizing the likelihood. The other two criteria are based on the probability density function (pdf) and cumulative density function (cdf) of the data. The pdf (cdf) of the experimental data is estimated and the pdf (cdf) for a given set of parameters k is simulated using SSA. For experimental and simulated data the pdf and cdf are estimated using histograms and the cumulative sums of the pdf, respectively. The parameter vector k, which minimizes a defined distance between simulated and experimental pdf (cdf), is calculated. Each method was evaluated and compared on toy data resulting that the maximum likelihood method is applicable with a low number of molecules. The methods based on the density functions, especially the method based on the cdf, are more robust than the likelihood method.

Reinker et al. (2006) propose two methods based on modelling the chemical reactions as a Markov Jump Process. The first method can be used for systems with a low number of reactions in each sampling interval, because one of its assumptions only holds if the sampling rate is relatively high in comparison to the reaction rates. They determine the likelihood and approximate it to be able to evaluate its maximum with numerical optimization algorithms. The second method deals with the data as exact molecule counts and approximates the amount of reactions in each interval. This approximation is done by solving a linear equation. They are able to derive a likelihood which is maximised with numerical optimisation methods. The second estimation method works even for systems with many reactions occurring in each time interval.

Based on modelling the chemical reactions as REE, Lillacci & Khammash (2010) propose an algorithm, which is based on Kalman filtering and extends it, thus being able to handle large parameter spaces and sparse and noisy data. The algorithm can be used for model selection as well. The authors demonstrate their algorithm on two examples.

Munsky et al. (2009) show how to estimate the kinetic parameters by considering the cellular noise which is inherent due to the random movement of reacting molecules. They examine the simple gene expression model and show that, by calculating the first two moments of the number of proteins and mRNA, it is possible to estimate the parameters with a high success rate. The success rate depends on

the number of time measurements and the number of experiments.

Boys et al. (2008) use Bayesian inference to estimate parameters of the Lotka-Volterra model (Lotka (1925),Volterra (1931)). They describe methods how inference can be made depending on the richness of information (complete data trace, discrete data at certain time points and partially observed data). They describe different Markov Chain Monte Carlo (MCMC) algorithms and analyse the performance of the different algorithms with simulated data on the Lotka-Volterra model.

Tian et al. (2007) use a simulated maximum likelihood method for genetic regulatory networks with small molecular numbers. Parameters are estimated for stochastic models which are described by stochastic differential equations or discrete biochemical reactions. They test their method on a one-stage and two-stage model and on a genetic toggle switch. It results that only the method based on discrete biochemical reactions has a robust estimation for small molecule numbers.

2.3 Approximate Bayesian Computation (ABC)

If model parameters are estimated using inference methods based on the likelihood, the likelihood function needs to be computable. If this is not possible because the likelihood does not exist in a closed form or is too costly to evaluate, "likelihood-free" methods can be used to infer the parameters because they simulate from the likelihood instead of evaluating it.

One possibility to not evaluate the likelihood but simulate the data from the associated model is called Approximate Bayesian Computation (ABC). The first to describe explicitly an algorithm for Approximate Bayesian Computation was Pritchard et al. (1999).

The ABC algorithm was developed continuously from the ABC rejection algorithm (Pritchard et al. 1999) and other variations exist, i.e. the ABC MCMC (Marjoram et al. 2003) or ABC SMC (Sisson et al. 2007).

Yet implemented algorithms exist for different purposes. For instance, Lopes et al. (2009) offer an ABC implementation for historical demographic parameters, and a do-it-yourself ABC (DIY ABC) is provided by Cornuet et al. (2008). Liepe et al. (2010) have an open source package called ABC-SysBio that supplies an

implementation of the ABC rejection sampler and an ABC SMC, both for parameter inference and model selection for models written in Systems Biology Markup Language (SBML).

Didelot et al. (2011) give an overview of different ABC methods in the chapter dealing with the background before introducing new approaches for model comparison based on likelihood-free methods.

A genetic regulatory system consists mostly of several reactions, where the velocity is determined by the kinetic rate $\theta^{(r)}$ of reaction r. The following table 1 gives an overview of the notation, which is used in this chapter, whereby the symbols are introduced on their first occurrence. Note that different combinations of indices are possible. This can be seen in the example $\theta_t^{(i,r)}$, which denotes the i-th drawn parameter value (=particle) for the r-th reaction and t-th population.

symbol	explanation	index and range
θ	parameter vector for all reactions	
$\theta^{(\cdot,r)}$	parameter for reaction r	$1 \leq r \leq R$
$\theta^{(i,\cdot)}$	i-th parameter value, called *particle*	$1 \leq i \leq N$
θ_t	parameter vector of population t	$0 \leq t \leq T$
$\{\theta_t^{(i,r)}\}_{1 \leq i \leq N}$	set of parameter values of r-th reaction and t-th population, consisting of N particles	
$w_t^{i,r}$	weight of the i-th particle of reaction r and population t in the ABC SMC	
R	number of reactions in the genetic regulatory system	
N	number of accepted particles	
T	number of populations in the ABC SMC	

Table 1: Notation for the ABC Algorithm.

2.3.1 Basic ABC Algorithm

The aim of the ABC algorithm is to derive a posterior distribution in situations where the likelihood function cannot be calculated in a reasonable time.

Let $\theta = (\theta^{(1)}, ..., \theta^{(r)})$ be the parameter vector of interest. A prior distribution $p(\theta)$ is known and the aim is to approximate the posterior distribution $p(\theta|x)$ for

a given data set x. Particles are drawn from the prior distribution and either accepted or rejected according to a decision rule until N particles are sampled. The posterior distribution consists of the accepted particles.

The Bayes theorem states that (Fahrmeir et al. 2009)

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)\mathrm{d}\theta}, \tag{10}$$

i.e. the density of the posterior distribution $p(\theta|x)$ can be calculated with the knowledge of the likelihood of the data $f(x|\theta)$ and the prior distribution $p(\theta)$ of the parameter and the integral $\int f(x|\theta)p(\theta)\mathrm{d}\theta$. In some cases the evaluation of this integral is too costly or not possible, but considering that it is a constant, it holds that $p(\theta|x) \propto f(x|\theta)p(\theta)$.

Moreover, a data set x^0 is given, which is, for instance, experimental data from the model. In the following it is named *experimental data*.

The basic *ABC algorithm* or *ABC rejection sampler* is described as follows:

1. Draw θ^* from $p(\theta)$.
2. Simulate a dataset x^* from $f(x|\theta^*)$, using for instance the Stochastic Simulation Algorithm (SSA). The conditional probability distribution $f(x|\theta^*)$ describes the genetic model.
3. Calculate the distance between the simulated and the experimental data $d(x^0, x^*)$ and accept θ^* if $d(x^0, x^*) < \epsilon$ for a chosen threshold ϵ. If θ^* is accepted, set $\theta^{(i,\cdot)} = \theta^*$ and $i = i + 1$.
4. While $i < N$ start from 1.

One of the main topics of this thesis is to examine different distances for their ability in the ABC algorithm. The distances are defined in chapter 3.

If the prior distribution is similar to the posterior distribution a large portion of the drawn θ^* are accepted. If the prior distribution is not similar, the distance $d(x^0, x^*)$ is rather high and, therefore, the rate of accepted θ^* is lower thus resulting in a higher number of drawn θ^* and, therefore, a higher computational effort.

An extension tries to avoid the low acceptance rate, the *ABC MCMC algorithm*, which includes the ABC into an MCMC algorithm. It was first proposed by

Marjoram et al. (2003) and is described in Toni et al. (2009) and Didelot et al. (2011).

"Potential disadvantages of the ABC MCMC algorithm are that the correlated nature of samples coupled with the potentially low acceptance probability may result in very long chains and that the chain may get stuck in regions of low probability for long periods of time. The above-mentioned disadvantages of ABC rejection [the acceptance rate is low when the prior distribution is very different from the posterior distribution] and ABC MCMC methods can, at least in part, be avoided in ABC algorithms based on SMC methods [...]" (Toni et al. 2009, p. 188 f.)

The ABC MCMC algorithm is, therefore, not described in detail, as it is not used in the later part of this thesis.

2.3.2 ABC Sequential Monte Carlo (SMC)

Sisson et al. (2007) was the first to propose an ABC approximation within a Sequential Monte Carlo (SMC) sampler, which was proposed by Del Moral et al. (2006). The following *ABC SMC algorithm* was proposed by Toni et al. (2009) and is similar to the algorithm from Sisson et al. (2007). The main difference is in the calculation of the weights. More details are described in the appendix of Toni et al. (2009).

The ABC SMC algorithm constructs the posterior distribution sequentially through intermediate distributions. It can be seen as a sequence of ABC rejection algorithms. In each step a number of parameter values (here called particles) are accepted. They form the posterior distribution, which serves as the prior distribution for the next step. With each step $t = 1, ..., T$, the threshold level ϵ_t decreases, thus resulting in the final posterior distribution $p(\theta|d(x^0, x^*) < \epsilon_T)$. The values for $\epsilon_1 > ... > \epsilon_T$ are defined at the beginning. No rules or remarks could be found how to specifically set these values. The drawn parameter values (particles) forming the posterior distribution are weighted, whereby they receive higher weights if they are more probable in the primary prior distribution. Besides, a sampled particle is perturbed to ensure that the entire particle space is considered effectively.

The algorithm proceeds as follows:

1. Set $t = 0$ and initialize values for the thresholds $\epsilon_1, ..., \epsilon_T$.

2. Set $i = 1$, whereby i is counting the accepted particles.
3. If $t = 0$ draw θ^{**} from $p(\theta)$.
4. If $t > 0$ draw θ^{*} from the population $\{\theta_{t-1}^{(i,\cdot)}\}_i$ with weights $\{w_{t-1}^{(i,\cdot)}\}_i$.
5. Use a perturbation kernel $K_t(\cdot|\cdot)$ to perturb the particle θ^{*} with $\theta^{**} \sim K_t(\theta|\theta^{*})$.
6. If $p(\theta^{**}) = 0$, return to step 4.
7. Simulate a dataset $x^{*} \sim f(x|\theta^{**})$.
8. If $d(x^0, x^{*}) > \epsilon_t$, return to step 4, otherwise set $\theta_t^{(i,\cdot)} = \theta^{**}$
9. The weight for particle $\theta_t^{(i,\cdot)}$ is calculated as follows

$$w_t^{(i,\cdot)} = \begin{cases} 1, & \text{if } t = 0 \\ \frac{p(\theta_t^{(i,\cdot)})}{\sum_{j=1}^{N} w_{t-1}^{(j,\cdot)} K_t(\theta_t^{(i,\cdot)}|\theta_{t-1}^{(j,\cdot)})}, & \text{if } t > 0 \end{cases} \tag{11}$$

10. If $i < N$, set $i = i + 1$ and proceed with step 4.
11. Normalize the weights so $\sum_i w_t^{(i,\cdot)} = 1$.
12. If $t < T$, set $t = t + 1$ and proceed with step 2.

Note that particles drawn from the previous distribution have a single asterisk and particles which are perturbed have a double asterisk. The first step of the ABC SMC for population $t = 0$ is equal to the ABC rejection sampler.

To get a deeper understanding of the weights and perturbation kernels, two examples are offered which are based on Filippi et al. (2011).

Uniform perturbation kernel The first kernel considered is a kernel based on the uniform distribution. The particle θ^{*} which is perturbed in step 5 of the algorithm is, in the following, noted by $\theta_t^{(i,\cdot)}$ to underline that it is the i-th particle in population t. The particle $\theta_t^{(i,\cdot)}$ is perturbed component-wise, i.e. each component of the vector representing a reaction r is perturbed independently. Each of the particles $\theta_t^{(i,r)}$ is perturbed by using a uniform distribution $U[\theta_t^{(i,r)} - \sigma_t^{(r)}, \theta_t^{(i,r)} + \sigma_t^{(r)}]$.

The parameters which define the width of the uniform distribution $\{\sigma_t^{(r)}\}_{1\leq r\leq R}$ can be set at the beginning of the simulation. Alternatively, parameters are used, which depend on the previous population. This is expressed by the index t. For the uniform distribution, $\sigma_t^{(r)}$ is often set to (Filippi et al. 2011)

$$\sigma_t^{(r)} = 0.5\left(\max_{1\leq i\leq N}\{\theta_{t-1}^{(i,r)}\} - \min_{1\leq i\leq N}\{\theta_{t-1}^{(i,r)}\}\right), \tag{12}$$

which presents the width of the previous population.

Calculating the weights of equation (11) for a uniform perturbation kernel with a uniform prior distribution $p(\cdot)$ with constant width results in equal weights $w_t^{(i,r)} = \frac{1}{N} \quad \forall r$. For $t = 0$ this is obvious. Taking $t = 1$ as an example, the numerator $p(\theta_t^{(i,r)}) = p(\theta_1^{(i,r)}) = c_r \in \mathbb{R} \quad \forall r$ is constant as $p(\cdot)$ is a uniform distribution. In the denominator $w_{t-1}^{(j,r)} = w_0^{(j,r)} = \frac{1}{N}\,\forall r$. Therefore, after normalization all weights $w_1^{(i,r)} = \frac{1}{N}\,\forall r$. This continues for $t = 2, 3, ...$ as well.

Gaussian perturbation kernel The second kernel is taken as a component-wise perturbation kernel based on a normal distribution $N(\theta^{(i,r)}, \sigma^{(r)})$ with mean $\theta^{(i,r)}$ and variance $\sigma^{(r)}$, here named *Gaussian kernel*. Beaumont et al. (2009) show that the value

$$\sigma_t^{(r)} = 2Var\left(\{\theta_t^{(i,r)}\}_i\right) \tag{13}$$

minimizes the Kullback-Leibler divergence between true posterior distribution and the estimated posterior distribution.

Using the Gaussian kernel to perturb the particle $\theta_t^{(i,r)}$ as in step 5 of the ABC SMC algorithm, one draws the value of the perturbed particle from the normal distribution $N(\theta_t^{(i,r)}, \sigma_t^{(r)})$.

In calculating the weights in equation (11), the expression $K_t(\theta_t^{(i,\cdot)}|\theta_{t-1}^{(j,\cdot)})$ states the probability of the particle $\theta_t^{(i,\cdot)}$ in the normal distribution $N(\theta_{t-1}^{(j,\cdot)}, \sigma_{t-1}^{(\cdot)})$ with mean $\theta_{t-1}^{(j,\cdot)}$ and standard deviation $\sigma_{t-1}^{(\cdot)}$. As the perturbation kernel perturbs the particle component-wise, i.e. separately for every reaction, a more exact notation in this case would be $K_t(\theta_t^{(i,r)}|\theta_{t-1}^{(j,r)})$.

Filippi et al. (2011) additionally specify other perturbation kernels which are not

component-wise, i.e. a multivariate normal perturbation kernel and a perturbation kernel which is based on the Fisher information.

The last part gave an overview of different perturbation kernels. The perturbation kernels, which are used in the simulation, are $U[0.75 \cdot \theta_t^{(i,r)}, 1.25 \cdot \theta_t^{(i,r)}]$ and $N\left(\theta^{(i,r)}, \frac{0.25}{1.96} \cdot \theta^{(i,r)}\right)$. Their parameter $\sigma_t^{(r)}$ is less than in equation (12) and (13). The choice of the kernels are discussed in detail in chapter 4.4.

2.4 Gene expression models

In this section two gene expression models are described, which are used later for simulation. For each model a network figure is shown. Additionally, the reactions and reaction rate equations are formulated. Several trajectories of the species for different parameter values are presented, and the steady state is derived.

2.4.1 One-stage model

The one-stage model, which is shown in figure 2, is one of the simplest models with regards to the number of reactions and species. It consists of only DNA and protein, where the DNA is processed directly into a protein and the protein decays.

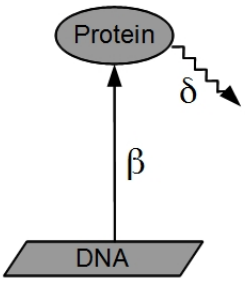


Figure 2: Scheme of the one-stage model. Solid arrows indicate synthesis, jagged arrows indicate degradation. DNA is processed into protein with rate β. The protein decays with rate δ.

Thus the reactions are

$$DNA \xrightarrow{\beta} DNA + protein \tag{14a}$$

$$protein \xrightarrow{\delta} \emptyset. \tag{14b}$$

RRE In a deterministic framework the species concentrations are described by ODEs, which can be inferred from the above reactions. In the one-stage model the ODE is given by

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \beta - \delta p(t) \tag{15a}$$

$$\frac{\mathrm{d}d(t)}{\mathrm{d}t} = 0, \tag{15b}$$

whereby $p(t)$ and $d(t)$ denote the concentration of protein and DNA respectively. As the concentration of DNA does not change due to reaction (14a), $d(t)$ stays constant, and equation (15b) is, therefore, often omitted. Compared with a stochastic representation, the ODE describes the evolution of the mean of the different trajectories over time. A trajectory in a stochastic model represents one possible evolution of the concentration of a species.

Trajectories Figure 3 shows three exemplary trajectories of the number of protein molecules.

The trajectories were simulated using the package *GillespieSSA* (Pineda-Krch 2010), version 0.5-4, in R (R Development Core Team 2010). For simulating the trajectories, the maximum time was set to 23, as this time was used for the simulations in chapter 5, and the initial number of DNA and protein molecules was set to one and zero respectively. The black solid line represents the mean of 1000 realizations. One can see that the mean converges. The exact time of convergence is calculated in the following paragraph.

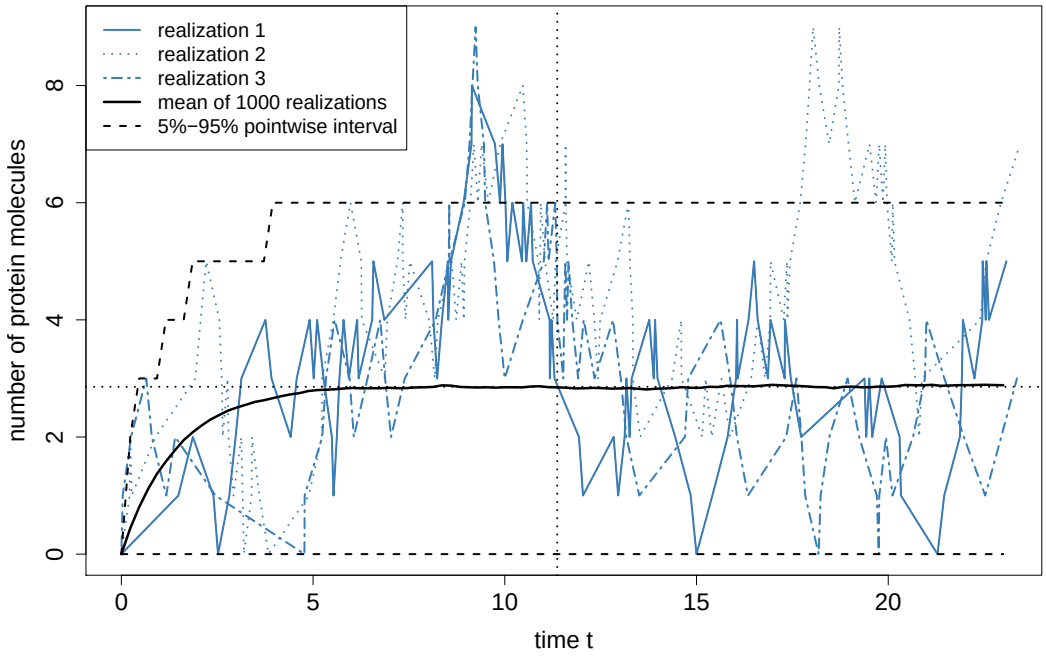


Figure 3: Trajectories of the number of protein molecules for three realizations of the one-stage model with $\beta = 2, \delta = 0.7$ and $d(0) = 1$, $p(0) = 0$ using SSA. Additionally, the mean of 1000 realizations with its 5%-95% point wise interval is plotted. The dotted horizontal line is at the steady state $\beta/\delta \approx 2.86$, and the dotted vertical line is at $t \approx 11.368$.

Steady state To determine the mean amount of protein molecules, one has to solve the ODE (15a), which results with initial condition $p(0) = P_0$ in

$$p(t) = \frac{\beta - (\beta - P_0\delta)\exp(-t\delta)}{\delta} \tag{16}$$

The mean number of molecules in the steady state equals $\frac{\beta}{\delta}$, as $\lim_{t\to\infty} p(t) = \frac{\beta}{\delta}$.

For determining the time until steady state is reached in the simulation, one has to define a bound how close the mean number of molecules should be to their theoretical steady state. Let ϵ be this bound, thus $p(t) \overset{!}{<} \frac{\beta}{\delta} - \epsilon$ which solves for

$\beta - P_0\delta > 0$ to

$$t > -\log\left(\frac{\epsilon \cdot \delta}{\beta - P_0\delta}\right) \cdot \frac{1}{\delta}. \tag{17a}$$

For $\beta - P_0\delta < 0$, i.e. $P_0 > \beta/\delta$, the number of proteins reaches steady state by declining from P_0 to the final amount of β/δ. Therefore, the condition $p(t) \overset{!}{<} \frac{\beta}{\delta} + \epsilon$ solves to

$$t > -\log\left(\frac{\epsilon \cdot \delta}{-\beta + P_0\delta}\right) \cdot \frac{1}{\delta}. \tag{17b}$$

For the case $\beta - P_0\delta = 0$, i.e. $P_0 = \beta/\delta$, equation (16) is $p(t) = \beta/\delta = P_0 \, \forall t$, thus the steady state for the mean number of protein molecules does not change during time.

For the parameter settings which were used in figure 3,$P_0 = 0$ $\beta = 2$ and $\delta = 0.7$, equation (17a) solves for $\epsilon = 10^{-3}$ to $t > 11.368$, and the steady state of protein is $\beta/\delta \approx 2.86$. One can see in figure 3 that the simulation produces a very similar result.

2.4.2 Two-stage model

The two-stage model is also known as the single gene expression model (Thattai & van Oudenaarden 2001) and extends the one-stage model by one species and two reactions. This model was already used as an example in chapter 2.1 and is explained in this chapter in more detail. Figure 4 shows the scheme of the two-stage model. The reactions for the two-stage model are the following

$$DNA \xrightarrow{\alpha} DNA + mRNA \tag{18a}$$

$$mRNA \xrightarrow{\beta} mRNA + protein \tag{18b}$$

$$mRNA \xrightarrow{\gamma} \emptyset \tag{18c}$$

$$protein \xrightarrow{\delta} \emptyset. \tag{18d}$$

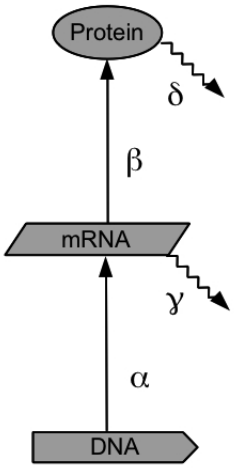


Figure 4: Scheme of the two-stage model. Solid arrows indicate synthesis, jagged arrows indicate degradation. DNA is transcribed into $mRNA$ with rate α. mRNA decays with rate γ and is translated into protein with rate β. The protein decays with rate δ.

RRE Deriving the system of ODEs from equations (18a)–(18d), one obtains the following equations

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \alpha - \gamma m(t) \tag{19a}$$

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = \beta m(t) - \delta p(t) \tag{19b}$$

$$\frac{\mathrm{d}d(t)}{\mathrm{d}t} = 0 \tag{19c}$$

Solution of the ODEs The RREs (19a) and (19b) were solved with the initial conditions $m(0) = M_0$ and $p(0) = P_0$, resulting in the following equations

$$m(t) = \frac{\alpha - (\alpha - M_0\gamma)\exp(-t\gamma)}{\gamma} \tag{20a}$$

$$p(t) = \frac{\exp(-t\gamma)M_0\beta - \exp(-t\delta)(M_0\beta + P_0\gamma - P_0\delta)}{-\gamma + \delta} + \frac{\alpha\beta - \exp(-t\gamma)\alpha\beta}{\gamma(-\gamma + \delta)} - \frac{\alpha\beta - \exp(-t\delta)\alpha\beta}{\delta(-\gamma + \delta)} \tag{20b}$$

For $\gamma = 0$, i.e. the mRNA does not decay, equation (20a) solves to $m(t) = M_0 + \alpha t$. For $\gamma = \delta$, equation (20b) solves to $p(t) = \frac{P_0 - \alpha\beta/\gamma^2 + t\beta(M_0 - \alpha/\gamma)}{\exp(t\gamma)} + \frac{\alpha\beta}{\gamma^2}$.

Steady state Using equations (20a) and (20b), one can derive the mean number of mRNA and protein molecules in the steady state. It holds that

$$\lim_{t\to\infty} m(t) = \frac{\alpha}{\gamma} \tag{21a}$$

$$\lim_{t\to\infty} p(t) = \frac{\alpha\beta}{\gamma\delta}. \tag{21b}$$

Trajectories For the trajectories which are shown in figure 5, the set of parameters for α, β, γ and δ are chosen from Komorowski et al. (2011, supplementary information, chapter 4.1). Table 2 provides an overview of the kinetic rates. The initial value for DNA, mRNA and protein was set to $d(0) = 1, m(0) = 0$ and $p(0) = 0$.

parameter	Set 1	Set 2	Set 3	Set 4
α	100	100	20	20
β	2	2	10	10
γ	1.2	0.7	1.2	0.7
δ	0.7	1.2	0.7	1.2

Table 2: Different sets of kinetic rates for the two-stage model (Komorowski et al. 2011). Sets 1 and 3 have a slow protein degradation rate δ and sets 2 and 4 have a high protein degradation rate. Sets 1 and 2 have a high transcription/translation ratio α/β and sets 3 and 4 have a low ratio.

The time until steady state is reached for the protein and mRNA concentration can be calculated numerically from equations (20a) and (20b). Steady state is defined as being at most 10^{-3} away from the theoretical steady state for $t \to \infty$. Table 3 shows the time until steady state is reached. One can see in figure 5 that the steady state for the mean number of protein is the same for all sets with $\frac{\alpha\beta}{\gamma\delta} \approx 238.1$. The steady state for the mean number of mRNA equals approximately 83.3, 142.9, 16.7, 28.6 number of molecules for set one to four.

The interpretation of the steady state for mRNA is the same as for the protein

time t [h] until steady state	set 1	set 2	set 3	set 4
mRNA	9.45	16.957	8.1	14.657
protein	18.937	18.937	18.937	18.937

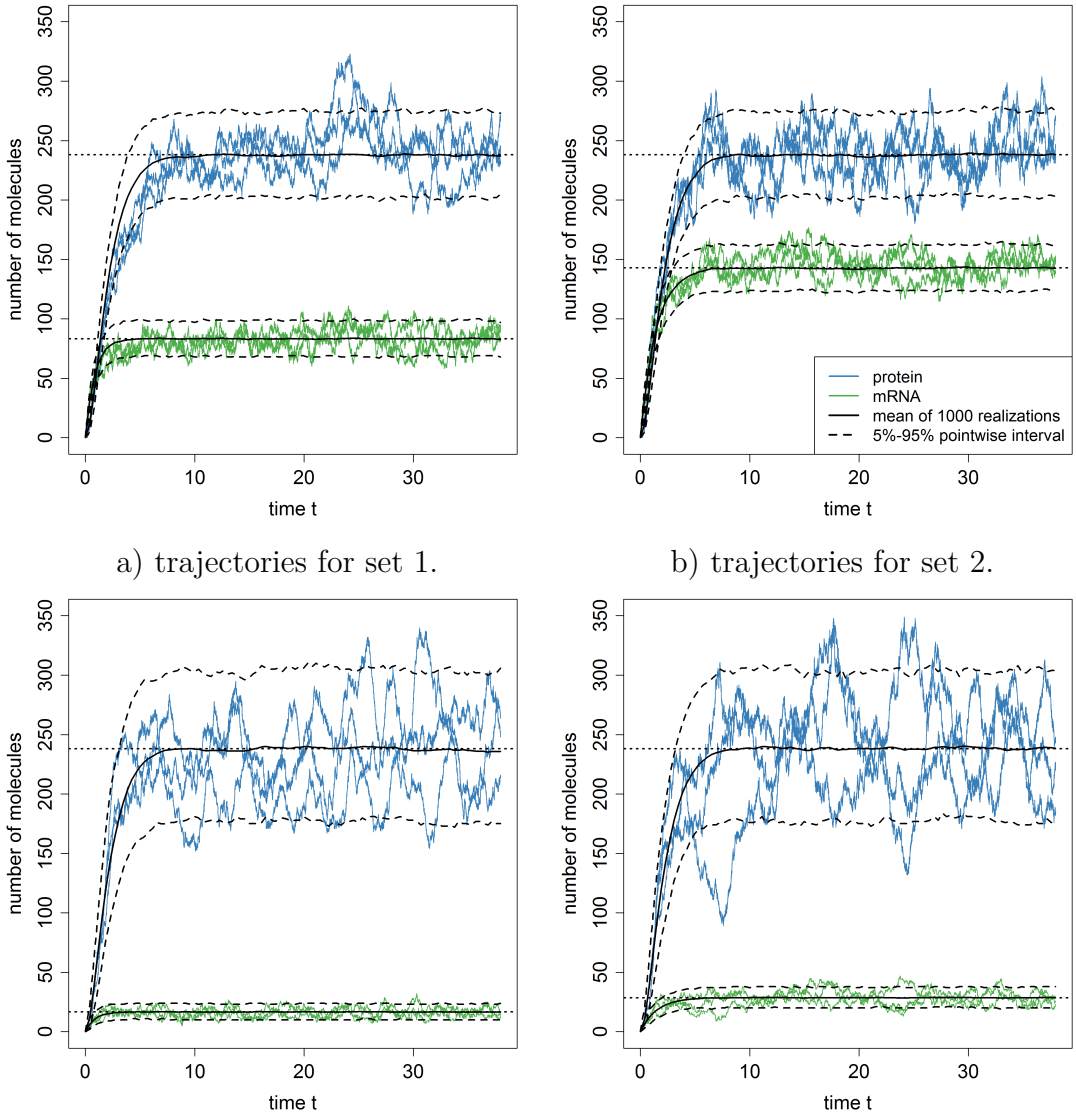
Table 3: Time until steady state is reached for the mean number of mRNA and protein respectively.

in the one-stage model. Thus, for an initial condition $\alpha - M_0\gamma > 0$, i.e. $M_0 < \alpha/\gamma$, the time until steady state is reached, is $t > -\log\left(\frac{\epsilon\cdot\gamma}{\alpha - M_0\gamma}\right)\cdot\frac{1}{\gamma}$.

For $\alpha - M_0\gamma < 0$, i.e. $M_0 > \alpha/\gamma$, it solves to $t > -\log\left(\frac{\epsilon\cdot\gamma}{-\alpha + M_0\gamma}\right)\cdot\frac{1}{\gamma}$.

For $\alpha - M_0\gamma = 0$, i.e. $M_0 = \alpha/\gamma$, it holds that $m(t) = M_0\,\forall t$.

As shown in table 3, the time until steady state is reached for mRNA is lower for sets one and three. The reason for this is the following. For a high degradation rate of mRNA γ, the time t can be low so that the equation $t > -\log\left(\frac{\epsilon\cdot\gamma}{|\alpha - M_0\gamma|}\right)\cdot\frac{1}{\gamma}$ holds.



a) trajectories for set 1.

b) trajectories for set 2.

c) trajectories for set 3.

d) trajectories for set 4.

Figure 5: Trajectories for the number of molecules of mRNA and protein for different sets of kinetic parameters for the two-stage model. The dotted lines indicate the steady state for the mean number of protein and mRNA.

3 Distance functions and summary statistics for the ABC (SMC)

In the ABC and ABC SMC algorithm, described in chapter 2.3, the function $d(x^0, x^*)$ is necessary to calculate the difference between experimental and simulated data. In most cases, the data is first summarized by a summary statistic $S(\cdot)$ before calculating the distance $d(S(x^0), S(x^*))$.

In chapter 3.1 the types of experimental data are described. In chapter 3.2 the different distance functions $d(x^0, x^*)$, which were used in the simulation to determine the distance between the *experimental data* x^0 and the *simulated data* x^*, are described. For all distance functions, except S-Cor and S-CC, the experimental and simulated data is summarized independently from each other, and the distance is measured between the summarized data of $S(x^0)$ and $S(x^*)$. For the functions S-Cor and S-CC the data x^0 and x^* is summarized as the correlation between x^0 and x^*. Based on the correlation, the distance between x^0 and x^* is calculated.

Therefore, in most cases a summary statistic is used. To calculate the distance between the summary statistics, the Euclidean distance is often taken (Nunes & Balding 2010). In this thesis the absolute value is taken due to better interpretation of the results.

The choice of the summary statistic is crucial for the result of the ABC algorithm (Nunes & Balding 2010).[5] Also Fearnhead & Prangle (2012) stress the importance of the summary statistic and provide a method how to construct appropriate summary statistics.

In all distance functions, except S-Cor and S-CC, the important part is the summary statistic because the distance is measured afterwards as the absolute difference of the summary statistics of the experimental and simulated data. To underline this, the distance functions are labelled for instance with S-Mean, with "S" for summary statistic.

In the literature the following approaches are used among others and might be of interest for further research.

[5] The authors also describe algorithms for automatic selection of efficient summary statistics. These algorithms are implemented in R in the ABCME package and available at `http://www.maths.lancs.ac.uk/~nunes/computerstuff/ABC.html`, last retrieved: 09.12.2011.

Tellier et al. (2011) describe different summary statistics, which are based on the joint site-frequency spectrum (JSFS). The JSFS is used for polymorphism data that contain information about parameters in the context of the isolation-migration model. Dealing with a different model type, the summary statistics can, therefore, not be adapted to this thesis.

Sousa et al. (2009) use an ABC without summary statistics as they use the full distribution of the allele in the model. Therefore, the whole data is involved in the ABC approximation.

3.1 Types of experimental data

Data from cells can be measured by methods such as Western blotting, Fluorescence-activated Cell Sorting (FACS) or time-lapse microscopy (Pawley (2006) and Larson et al. (2009)). These methods are able to measure the concentration of a species over time, either coupled to a cell, as the single concentrations of the population for each time point, or as the mean of the entire population for each time point.

Therefore, depending on whether the measurement can be coupled with the cell and the frequency of measurement (discrete data or continuous data) one can distinguish between *population data*, *time point measurements (TP)* and *time series data (TS)*.[6]

Population data This data contains the lowest amount of information with only the mean of species measured. An example is shown in figure 6, where the mean of the experimental and real data is denoted by an asterisk. For detecting proteins, the western blot (Burnette 1981) can be used, for DNA detection the Southern blot and for RNA detection the Northern blot, for instance.

Time point measurements Figure 6 also shows an example for time point measurements, i.e. discretely observed data. For instance, for each time point $t = 1, ..., T_{obs}$ the amount of protein molecules in each of $N = 3$ cells is measured. Thus at each time point there are $N^0 = 3$ data points for the experimental data

[6]The naming of the data types is based on Komorowski et al. (2011).

and $N^* = 4$ data points for the simulated data with $\mathbf{x_t^*}$ denoting the vector of data and $x_{t,j}^*$, $j = 1, ..., N^*$ denoting a single data point.

The data $\mathbf{x_t^*}$ is denoted by an asterisk, the experimental data $\mathbf{x_t^0}$ is denoted by the superscript zero.

To measure protein concentration in single cells FACS is used.

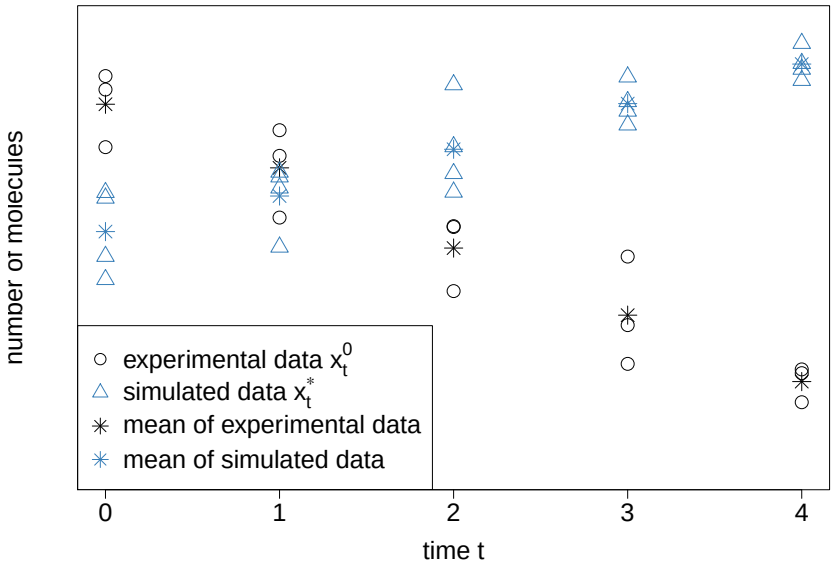


Figure 6: Example for time point measurements and population data. The asterisk represents the mean of the number of species in the population of all measured cells. There are five equidistant measurements.

Time series data For time series data, the development of the number of species for each cell over the time period is tracked. An example is shown in figure 7. The data is, therefore, a function $z(t)$ depending on the time, which represents the number of molecules for a given species at time t. In the experimental setting there are multiple cells measured at discrete time points t. Therefore, the notation is extended to $z_j(t)$, $j = 1, ..., N$ to indicate the value for the j-th cell at time t.

To gain TS data, time-lapse microscopy can be used. An overview of different methods to receive TS data are given in Pawley (2006, ch. 19).

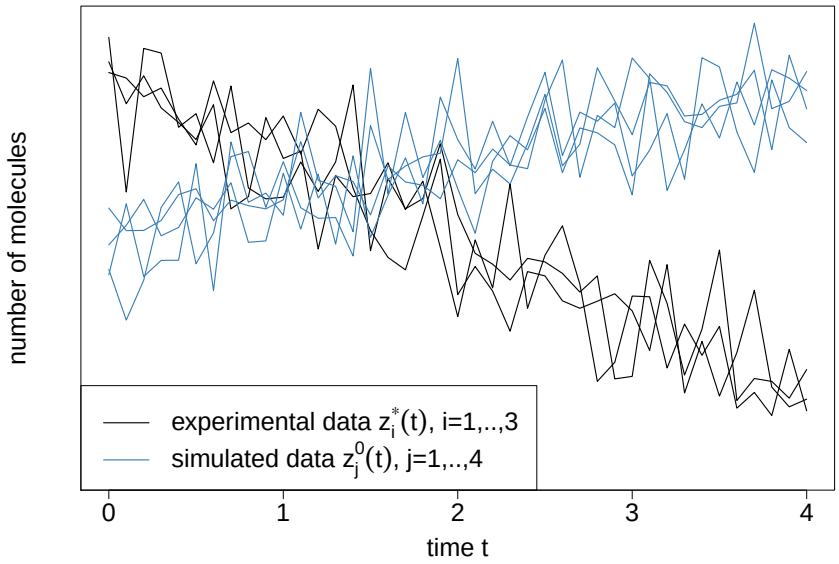


Figure 7: Example for time series data. Each sharp bend is indicating a time point where the data is collected. Although TS data is assumed to be continous, in reality it is only possible to measure it at discrete time points.

3.2 Distance functions

To keep notation simple, the distances in the following are described for experimental and simulated data which have only one dimension of species. For multivariate data, i.e. data consisting of more than one species, the distance is calculated for each species j and then summed over all species, i.e. $d(x^0, x^*) = \sum_j d(x_j^0, x_j^*)$.

S-Mean Munsky et al. (2009) investigate among others the mean in the two-stage model as a statistic for parameter estimation. In this thesis S-Mean is defined as

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| \text{mean}(\mathbf{x_t^*}) - \text{mean}(\mathbf{x_t^0}) \right| \tag{22}$$

with $|\cdot|$ indicating the absolute value of its argument and $\text{mean}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$.

S-Std Munsky et al. (2009) use also the variance (in combination with the mean) for parameter estimation. In this thesis the standard deviation is chosen because it has the same unit as the mean.

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| \text{std}(\mathbf{x_t^*}) - \text{std}(\mathbf{x_t^0}) \right| \tag{23}$$

with $\text{std}(x) = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \text{mean}(x))^2 \right)^{\frac{1}{2}}$.

S-M&Std This distance is a combination of the mean and standard deviation.

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| \text{mean}(\mathbf{x_t^*}) - \text{mean}(\mathbf{x_t^0}) \right| + \sum_{t=1}^{T_{obs}} \left| \text{std}(\mathbf{x_t^*}) - \text{std}(\mathbf{x_t^0}) \right| \tag{24}$$

S-NE Negentropy is used as a measure of normality (Hyvärinen et al. 2001). Thus the idea of the following distance is to determine the degree of normality of the experimental and simulated data at each time point and to take the difference as a measure of deviance from the true reaction rates.

To approximate the negentropy, two ways, among others, are described in Hyvärinen et al. (2001). The approximation in S-NE is cumulant-based, and in S-NE II it is based on nonpolynomial functions.

The distance S-NE is defined as

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| J(\mathbf{x_t^*}) - J(\mathbf{x_t^0}) \right| \tag{25}$$

with $J(\mathbf{x})$ being the negentropy, which, using the cumulant based approximation, can be calculated[7] for a standardized variable $\mathbf{x}$ as (Hyvärinen et al. 2001, p. 115, eq. (5.35))

$$J(\mathbf{x}) = \left(\frac{1}{12} \left[\mathbb{E}\{(\mathbf{x})^{\mathbf{3}}\} \right]^2 + \frac{1}{48} \left[\text{kurt}\{\mathbf{x}\} \right]^2 \right) \tag{26}$$

[7] Equation (26) only holds approximately, as approximations were used in its derivation. For calculation it was used as an exact equation.

where kurt(x) is the bias-corrected kurtosis of x. The kurtosis is estimated by

$$k_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{mean}(x))^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{mean}(x)^2\right)^2} \tag{27a}$$

and the bias corrected kurtosis is given by

$$\text{kurt} = \frac{n-1}{(n-2)(n-3)}\left((n+1)k_1 - 3(n-1)\right) + 3 \tag{27b}$$

and the expected value $\mathbb{E}$ is estimated by the mean(x).

To calculate the negentropy, the data $\mathbf{x_t^*}$ and $\mathbf{x_t^0}$ are standardized first. If var(x) = 0, i.e. only the same number of molecules appear in the sample set, kurt(x) is set to zero.

S-NE II In S-NE II the negentropy $J(\mathbf{x})$ is approximated by non-polynomial functions. The distance is defined as for S-NE with

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| J(\mathbf{x_t^*}) - J(\mathbf{x_t^0}) \right|. \tag{28}$$

Hyvärinen et al. (2001, p. 119, eq. (5.48)) propose the following equation to approximate the negentropy by non-polynomial functions. They state that this approximation[8] is more robust and accurate as the approximation of the former distance function in equation (26):

$$J(x) = k_1\left(\mathbb{E}\{x\exp(-0.5x^2\}\right)^2 + k_2\left(\mathbb{E}\{\exp(-0.5x^2\} - \sqrt{0.5}\right)^2 \tag{29}$$

where $k_1 = \frac{36}{8\sqrt{3}-9}$ and $k_2 = \frac{24}{16\sqrt{3}-27}$ and the data x_t^* and x_t^0 is standardized.

S-pdf Poovathingal & Gunawan (2010) use the pdf and cdf as distance measures between model prediction and experimental data for estimating reaction rates in models, which are represented by a chemical master equation. The idea is to measure the difference between two distribution functions, and, therefore, to determine if the distribution of the simulated data is close to the distribution of

[8] Equation (29) only holds approximately, as approximations were used in its derivation. For calculation it was used as an exact equation.

the experimental data.

The distance is defined as

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left(\int_{-\infty}^{+\infty} |p(x_t) - q(x_t)| \ \mathrm{d}x_t \right) \tag{30}$$

where $p(x_t)$ and $q(x_t)$ are the pdfs of the real and simulated data for time point t respectively. The geometrical interpretation of $d(\mathbf{x^0}, \mathbf{x}^*)$ is the sum over all t of the difference of the area between the probability density functions of $\mathbf{x}_t^*$ and $\mathbf{x}_t^0$.

For computation, the pdfs are approximated by histograms and

$$d(\mathbf{x^0}, \mathbf{x}^*) = \frac{1}{n_\chi} \sum_{t=1}^{T_{obs}} \sum_{x \in \chi_t} \left| P_{\mathbf{x_t^0}}(x) - P_{\mathbf{x_t^*}}(x) \right| \tag{31}$$

with $P_{\mathbf{x_t^0}}(x)$ being the height of the histogram at x. The set χ_t defines the range $[\min(\mathbf{x_t^0}, \mathbf{x_t^*}), \max(\mathbf{x_t^0}, \mathbf{x_t^*})]$, which is divided in n_χ equidistant intervals.[9] To construct the histograms, the number of bins were set to $n_t^0 = \min(\mathrm{unique}(\mathbf{x_t^0}), \lceil \sqrt{N^0} \rceil)$ and $n_t^* = \min\left(\mathrm{unique}(\mathbf{x_t^*}), \lceil \sqrt{N^*} \rceil\right)$. The function unique$(\cdot)$ is the number of unique values.

S-cdf This distance function is defined as

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left(\int_{-\infty}^{+\infty} \left| F^*(x_t) - F^0(x_t) \right| \ \mathrm{d}x_t \right), \tag{32}$$

and the geometrical interpretation of $d(\mathbf{x^0}, \mathbf{x}^*)$ is the sum over all t of the area between the cumulative density functions $F^*(\mathbf{x}_t^*)$ and $F^0(\mathbf{x}_t^0)$.

For computation, the distance is approximated by

$$d(\mathbf{x_t^0}, \mathbf{x_t^*}) = \frac{1}{n_\chi} \sum_{t=1}^{T} \sum_{x \in \chi} \left| \tilde{F}^*(x) - \tilde{F}^0(x) \right|, \tag{33}$$

where $\tilde{F}^*(x)$ and $\tilde{F}^0(x)$ are the empirical cdfs (Fahrmeir et al. 2009) of $\mathbf{x}_t^0$ and $\mathbf{x}_t^*$ respectively. χ is defined analogous to S-pdf.

[9] For simulation n_χ was set to 200.

For TS data the previous mentioned distances would not utilize all information which are available in the data. Therefore, two distances are described, which are based on correlation. The idea is that experimental and simulated data are similar to each other if their trajectories have a high correlation. To measure correlation, both Pearson's correlation and cross-correlation is used.

S-Cor The distance is defined as

$$d(z^*, z^0) = \|\mathbb{D}\|, \tag{34}$$

where $\mathbb{D}$ is the distance matrix between z^* and z^0, which is defined as $\mathbb{1} - \text{COR}$ with $\mathbb{1}$ being a matrix containing only ones. COR is the correlation matrix of z^* and z^0, i.e. an entry $c_{i,j}, i = 1, ..., N^*, j = 1, ..., N^0$ in this matrix is defined as

$$c_{i,j} = \rho(\mathbf{z}_i^*, \mathbf{z}_j^0). \tag{35}$$

$\rho(\cdot, \cdot)$ is the Pearson's correlation coefficient. If $\rho(\mathbf{z}_i^*, \mathbf{z}_j^0)$ is not defined due to $\text{var}(\mathbf{z}_i^*)$ or $\text{var}(\mathbf{z}_i^0)$ being zero, because for instance $\mathbf{z}_i^*$ only consists of the same values, $c_{i,j}$ is set to 0 so that the corresponding distance $d_{i,j}$ in $\mathbb{D}$ is set to one, and, therefore, it is not taken into account when the norm of $\mathbb{D}$ is calculated.

In $\mathbb{D}$ the rows correspond to the simulated data and the columns the experimental data. For the norm $\|\mathbb{D}\|$ the smallest distance from each simulated time trace to all experimental time traces is taken and summed over all simulated data, formally

$$\|\mathbb{D}\| = \sum_{i=1}^{N^*} \min_j d_{ij}. \tag{36}$$

Often the Frobenius norm $\|\cdot\|_F$ is used for measuring the norm of a matrix with $\|\mathbb{D}\|_F = \sum_{i=1}^{N^*} \sum_{j=1}^{N^0} d_{ij}^2$. As non typical time traces can enlarge the distance considerably, the formally explained norm is preferred over the Frobenius norm. In chapter A.3 the result of a simulation performed with the Frobenius norm is illustrated. In summary, the parameter estimation is less accurate as with the norm (36). As shown in chapter 3.3, it is not a norm in the strict mathematical

sense, but it is used here because of its intuitive approach for comparing distances based on correlations between trajectories of experimental and simulated data.

S-CC This distance is similar to S-Cor, and it is based on TS data as well. Instead of Pearson's correlation, it uses the cross-correlation. It is defined as

$$d(z^*, z^0) = \|\mathbb{D}\| \tag{37}$$

with $\mathbb{D} = 1 - \mathrm{COR}$. The entries of the correlation matrix are defined as the maximum of the cross-correlation between z_i^* and z_j^0, i.e.

$$c_{i,j} = \max_{\tau} \left(R_{\mathbf{z}_i^*, \mathbf{z}_j^0}(\tau) \right), \tag{38}$$

where $R_{x,y}(\tau)$ is the cross-correlation between x and y. Consider discrete time points τ. Although TS data is measured, the measurements can only be performed at discrete time points. The not-normalized cross-correlation is calculated as

$$\hat{R}_{x,y}(\tau) = \begin{cases} \sum_{t=0}^{T-\tau-1} x_{t+\tau} y_t & \tau \geq 0 \\ \hat{R}_{y,x}(-\tau) & \tau < 0 \end{cases} \tag{39}$$

Before determining $c_{i,j}$, the cross-correlation $\hat{R}_{x,y}(\tau)$ is normalized, so an auto-correlation with $\tau = 0$ has value 1.0.

3.3 Analysis of the metric characteristics

In this section the defined distance functions are analyzed if they are metrics. For neither the ABC nor ABC SMC algorithm no publication could be found, which states if the distance function has to be a metric or what implications can be derived if $d(\cdot, \cdot)$ is or is not a metric. For example, Toni et al. (2009) and Pritchard et al. (1999) do not mention if $d(\cdot, \cdot)$ should be a metric or not. Joyce & Marjoram (2008) or Liepe et al. (2010), for instance, use the word 'metric' to describe the distance metric but it cannot be derived from the context if it is used in the strict mathematical sense.

A function $d\colon X \times X \to \mathbb{R}$ is called a metric on a set X if for all x, y, z in X the

following conditions are fulfilled (Krause 1986):

M1 $d\left(x,y\right) \geq 0$

M2 $d\left(x,y\right) = 0 \Leftrightarrow x = y$

M3 $d\left(x,y\right) = d(y,x)$

M4 $d\left(x,y\right) \leq d(x,z) + d(z,y)$

M1 is implied by M2-M4 as $2d\left(x,y\right) = d\left(x,y\right) + d\left(x,y\right) = d\left(x,y\right) + d\left(y,x\right) \geq d\left(x,x\right) = 0$ and is, therefore, often not proofed separately.

A special metric is the taxicab metric[10] which defines a metric (Krause 1986). The taxicab metric for two vectors $a, b \in \mathbb{R}^n$ is defined as

$$d(a,b) = \sum_{i=1}^{n} |a_i - b_i| \tag{40}$$

In the following the distances are examined if they are metrics.

S-Mean, S-Std, S-NE and S-NE II The distances S-Mean, S-Std, S-NE and S-NE II can be written as

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \left| (S(\mathbf{x_t^0}) - S(\mathbf{x_t^*}) \right| \tag{41}$$

with $S(\cdot)$ being the respective summary statistic. Thus, these distances would be metrics if the value of $S(\mathbf{x_t^0})$ and $S(\mathbf{x_t^*})$ is different for a different $\mathbf{x_t^0}$ and $\mathbf{x_t^*}$. This is not true, as the mean, standard deviation or negentropy are non-injective functions. Therefore, M2 is violated and the distances are pseudometrics. M1, M3 and M4 still hold. As the sum of pseudometrics is a pseudometric, S-M&Std represents also a pseudometric.

[10] also known as rectilinear distance, L_1 distance, city block distance or Manhattan distance

S-pdf, S-cdf For S-pdf for a certain time point t the distance can be written in the form of the taxicab metric

$$d(\mathbf{x_t^0}, \mathbf{x_t^*}) = \frac{1}{n_\chi} \sum_{x \in \chi} \left| P_{\mathbf{x_t^0}}(x) - P_{\mathbf{x_t^*}}(x) \right| \tag{42}$$

The approximation of the pdf is a non-injective function, therefore, M2 is violated and $d(\mathbf{x_t^0}, \mathbf{x_t^*})$ presents a pseudometric.

The entire distance S-pdf for all time points is $d(\mathbf{x^0}, \mathbf{x^*}) = \sum_{t=1}^{T_{obs}} d(\mathbf{x_t^0}, \mathbf{x_t^*})$ thus the sum of pseudometrics and, therefore, represents a pseudometric as well.

The same argument holds for S-cdf.

S-Cor, S-CC For S-Cor and S-CC, M3 is violated as in $d(z^*, z^0) = \|\mathbb{D}\|$ the rows contain the simulated data and $\|\mathbb{D}\|$ is defined as the sum of the minimum of all columns. For $d(z^0, z^*)$ the rows would contain the experimental data, thus $\|\mathbb{D}\|$ would yield another result.

Taking a norm which is the same for $\mathbb{D}$ and the transpose of $\mathbb{D}$, e.g. the Frobenius norm $\|\cdot\|_F$, this would result in the acceptance of M3.

However, for S-Cor M2 is still violated. Assume w.l.o.g. that only one cell is measured and let X and Y be two random numbers representing these measurements with $X \sim F$ and $Y = X + a$, $a \in \mathbb{R}$ where F can represent any distribution. Their distance is

$$d(X, Y) = \|\mathbb{D}\| = \|1 - \mathrm{COR}(X, Y)\| = \|1 - \mathrm{COR}(X, X)\| = \|1 - 1\| = 0 \tag{43}$$

But as $X \neq Y$, M2 is violated.

A similar argument holds for S-CC, as $\max_\tau (R_{X,Y}(\tau)) = 1$ for $\tau = 0$.

The proof of M4 is omitted, as M2 and M2 are violated.

In summary, all distances are pseudometrics except S-Cor and S-CC.

Investigating the norm of S-Cor and S-CC In this paragraph it is shown whether or not the expression (36), i.e. $\|\mathbb{D}\| = \sum_{i=1}^{N} \min_j d_{ij}$, is a norm for $\mathbb{D}$. To be a norm the following properties must hold (Prugovecki 2006, p. 20). It is defined for vectors in general, here matrices $\mathbb{A}$ and $\mathbb{B}$ are taken.

N1 $\|\mathbb{A}\| = 0 \Rightarrow \mathbb{A} = 0$

N2 $\|\alpha \cdot \mathbb{A}\| = |\alpha| \cdot \|\mathbb{A}\|$

N3 $\|\mathbb{A} + \mathbb{B}\| \leq \|\mathbb{A}\| + \|\mathbb{B}\|$

N1 is violated as any matrix $\mathbb{A}$ which has at least one zero element as the minimum in each row has a norm of zero although it is not the zero matrix.

N2 only holds for $\alpha \geq 0$. Let the minimum of each row be $\widetilde{a_{ij(i)}}$. So, $\|\alpha \cdot \mathbb{A}\| = \sum_{i=1}^{N^*} \min_j \alpha \cdot a_{ij} \overset{\alpha \geq 0}{=} \sum_{i=1}^{N^*} \alpha \cdot \widetilde{a_{ij(i)}} = \alpha \cdot \|\mathbb{A}\|$. For $\alpha < 0$ the maximum of each row of $\mathbb{A}$ multiplied by a negative α becomes the minimum.

N2 would hold if $\|\cdot\|$ is changed to $\|\mathbb{D}\| = \sum_{i=1}^{N} \min_j |d_{ij}|$. This holds for our case even without the absolute value, as all d_{ij} are greater than zero.

N3 does not hold, which is shown by means of an example.

$$\begin{aligned} \|\mathbb{A} + \mathbb{B}\| = \left\| \begin{pmatrix} 0 & 1 & 1 \\ 3 & 0 & 2 \\ 2 & 0 & 3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \right\| = \left\| \begin{pmatrix} 1 & 1 & 3 \\ 3 & 1 & 4 \\ 2 & 1 & 5 \end{pmatrix} \right\| = 3 \nleq \\ \left\| \begin{pmatrix} 0 & 1 & 1 \\ 3 & 0 & 2 \\ 2 & 0 & 3 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \right\| = 0 + 0 = 0 \end{aligned} \tag{44}$$

4 Preface to computational study

For the computational study the ABC algorithm was slightly changed. In section 4.1 the changes are presented, and the choice of the parameters for the algorithm are explained. A statistic, the sum of normalised absolute residuals, which measures the quality of the posterior distribution w.r.t. parameter estimation, is introduced in section 4.2. Moreover, an overview of the conducted simulations is presented in section 4.3. The computational time necessary for simulation is explained in section 4.1.2. In the last section the implementation of the ABC SMC and the choice of its parameters are presented.

The following questions are tried to be answered with the simulations. The results of the simulations are discussed in chapter 5.

- How is the overall performance of the different distances?
- How is the distance $d(x^0, x^*)$ distributed?
- What parameters out of the number of drawn particles N_{all}, the acceptance rate τ and the sampling frequency Δ are affecting the estimation?
- How does the performance of the ABC algorithm depend on the informativeness of the prior distribution?
- Is the ABC SMC performing better than the ABC for the same amount of drawn particles?

4.1 Implementation of the ABC algorithm & choice of parameters

This section explains how the ABC algorithm has been implemented and explains the parameters, which have to be set.

4.1.1 Implementation of the ABC rejection algorithm

The ABC algorithm described in chapter 2.3.1 was adjusted so that all drawn particles θ^* are accepted at first. Then the particles with the lowest distance were

accepted to form the posterior distribution. To determine how many particles are accepted, an acceptance rate τ is introduced, which states the percentage of accepted particles of all drawn particles. The number of drawn particles are denoted by N_{all} and the number of accepted particles are then $N = N_{all} \cdot \tau$. This approach is used for instance in Beaumont et al. (2009) and Tellier et al. (2011).

Thus from step 3 onward, the algorithm, described in chapter 2.3.1, is changed accordingly to

3. Calculate the distance between the simulated and the experimental data $d(x^0, x^*)$ and record θ^* as θ_i^*.

4. While $i < N_{all}$ start from step 1.

5. Accept the $N = N_{all} \cdot \tau$ particles from θ_j^*, $j = 1, ..., N_{all}$ with the lowest distance, which are then recorded as $\theta^{(i,\cdot)}$, $i = 1, .., N$.

The reason for this adjustment was to be able to compare the different distances w.r.t. their ability for estimating the true reaction rates. Without introduction of an acceptance rate, each distance would have a different number of drawn particles and, therefore, a different acceptance rate until N particles were accepted. This would make comparison of the distances unfair.

The trajectories were simulated using a software called StochKit2 (stochastic simulation kit, Sanft et al. (2011)), which implements the SSA in the programming language C. Version 2.0.2 was used, and it was accessed through Matlab with a wrapper for Matlab provided by Michael Strasser[11]. To simulate the trajectories with the SSA, a number of parameters need to be set. These are described in table 4. The values of the parameters are explained in the following paragraphs.

4.1.2 Determining the observation time $\mathbf{T_{obs}}$

The final time, until the evolution of protein was observed, was set to $T = \lceil 2 \cdot T_{ss} \rceil$ with T_{ss} being the time until steady state, as defined in chapter 2.4.1, is reached for all species. For the chosen parameter settings, this resulted for the one-stage model in $T = \lceil 2 \cdot 11.368 \rceil = 23$ and for the two-stage model in $T = \lceil 2 \cdot 18.937 \rceil = 38$.

[11] PhD student at the Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum, München under supervision of Fabian Theis.

parameter	explanation
model	the model for which the trajectories should be simulated, e.g. the one-stage model
$\theta^{(r)}$, $r = 1, ..., R$	the reaction rates of the model
$X_j(t = 0)$	the initial condition of the number of molecules for each species
N_{traj}	the number of simulated trajectories
T_{obs}	the maximal time until the trajectories are simulated
Δ	the number of species are recorded at certain equally spaced time points. The frequency Δ determines the time between two time points. The number of recorded time points are then $T_{obs}/\Delta + 1$

Table 4: Parameters for the SSA in the computational study.

4.1.3 Determining sampling frequency Δ

The sampling frequency Δ is defined as the time between subsequent observations $t_i - t_{i-1}$, assuming equidistant sampling. The sampling frequency is set, based on the determinant of the Fisher information matrix (FIM). Komorowski et al. (2011) propose a method to numerically calculate the FIM without the need for Monte Carlo simulations. To sum their method up, the kinetic model is written using the linear noise approximation (LNA). This is the basis for deriving the likelihood of the experimental data. From the likelihood the FIM can be calculated.

The idea to determine the sampling frequency is because *"the amount of information in a sample does not depend solely on the type of data (TS,TP), but also on other factors [... like] the sampling frequency [...] The amount of information in a sample was understood as the determinant of the FIM."* (Komorowski et al. 2011, p.8648)

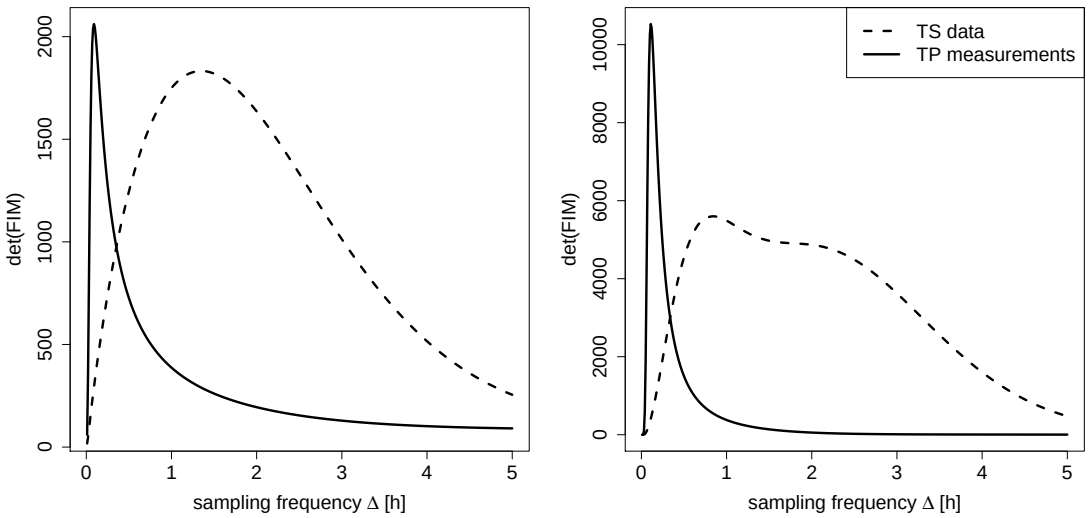
Therefore, to determine the sampling frequency, the determinant of the FIM was calculated using a Matlab package called StochSens.[12]

One aspect, which needs to be considered, is that for calculating the likelihood of the experimental data, estimates of the kinetic rates and the model underlying the data must be given. Thus for model selection, where neither the kinetic rates

[12]This package is available from the website of Michał Komorowski, Imperial College, London at `http://www.theosysbio.bio.ic.ac.uk/resources/stns/`.

nor the underlying model is known, this proposes a challenge.

Figure 8 shows the determinant of the FIM plotted against sampling frequency for the one-stage and two-stage model.



a) one-stage model b) two-stage model

Figure 8: Determinant of the FIM for the one-stage and two-stage model with reaction rates $\alpha = 100, \beta = 2, \gamma = 1.2, \delta = 0.7$. It is assumed that $N_{traj} = 100$. The maximum of det(FIM) for the one-stage model is at a frequency of 0.09 h and 1.35 h for TP and TS data respectively. For the two-stage model it is at 0.11 h and 0.84 h.

4.1.4 Choice of the acceptance rate τ

As can be seen in the results of the computational study in chapter 5, the choice of the acceptance rate τ influences the quality of the posterior distribution w.r.t the parameter estimation. The following publications use different acceptance rates. Tellier et al. (2011) report that 1% of 300,000 simulated data were accepted using an isolation-migration model. Toni et al. (2009) use a threshold ϵ to determine if a particle is accepted. This results in $\tau = 7 \cdot 10^{-5}$ for 14.1 million simulated particles in a Lotka-Volterra model. Beaumont et al. (2009) use $\tau = 0.01$ in a mutation model. Sousa et al. (2009) state that typical values for τ are from 10^{-5} to 10^{-2}.

4.1.5 Choice of the prior distribution

Table 5 gives an overview of exemplary prior distributions, which were used in the literature. The exact model and the meaning of the parameter is described in detail in the corresponding publications, here, only a short explanation can be given. The aim of this overview is to give an insight into the choice of prior distributions. The SNAR value for each prior was calculated as well. The SNAR values are defined in chapter 4.2.

author	model	true parameter	prior	$\text{SNAR}_{p(\theta)}$	explanation
(1)	mutation m.	0.0005	$\text{U}[10^{-4}, 10^{-3}]$	0.456	mutation rate
		1,000	$\text{U}[10, 10^4]$	4.105	time of divergence
		10,000	$\text{U}[10^2, 10^5]$	4.105	effec. population size
(2)	Lotka-Volterra	1	U[-10,10]	5.05	pred.-prey interaction
(3)	neutral IM m.	0.01 to 9	U[0.01, 10]	499.69 - 0.455	divergence time
		0.01 to 9	U[0.01, 10]	499.69 - 0.455	migration rate
(4)	mutation m.	1000	U[200,2000]	0.456	population size
		0.7	U[0.1,0.9]	0.357	admixture rate
		30	U[2,150]	1.71	effec. population size
(5)	cyclic chain	2	$U[10^{-3}, 10^3]$	249.07	kinetic rate
		1.5	$U[10^{-3}, 10^3]$	332.41	kinetic rate
		3.2	$U[10^{-3}, 10^3]$	155.21	kinetic rate

Table 5: Choice of prior distributions in recent papers of (1) Beaumont et al. (2009), (2) Toni et al. (2009), (3) Tellier et al. (2011), (4) Robert et al. (2011) and (5) Müller et al. (2012).

For the computational study a lognormal distribution $\text{LN}(\mu, \sigma)$ was chosen as the prior distribution. The parameters were set so that the mode $\exp(\mu - \sigma^2)$ of the distribution is at the true reaction rate θ. Therefore, $\exp(\mu - \sigma^2) \stackrel{!}{=} \theta$ and thus $\mu = \log(\theta) + \sigma^2$. This setting was chosen because one can regulate the prior distribution by only adjusting the parameter σ. By increasing σ, the prior is becoming less informative.

4.2 Sum of normalised absolute residuals (SNAR)

In the publications which are concerned with ABC sampling, the quality of the posterior distribution is most often checked visually by means of density estimates. Often the mean, median or 95% quantile range is reported as well. But there is no statistic which measures the gain of accuracy in parameter estimation from the prior to the posterior distribution.

Therefore, to measure the accuracy of the posterior distribution with regards to the true reaction rate, the sum of normalized absolute residuals (SNAR) is introduced. For the true reaction rates $\theta = \{\theta^{(1)}, ..., \theta^{(r)}\}$ and the estimated values $\theta^* = \{\theta^{*(\cdot,1)}, ..., \theta^{*(\cdot,r)}\}$ the SNAR is defined as

$$\text{SNAR} = \sum_{r=1}^{R} \frac{\left| \sum_{i=1}^{N} \left(\theta^{*(i,r)} - \theta^{(r)} \right) \right|}{N \cdot \theta^{(r)}} = \sum_{r=1}^{R} \frac{mean\left(\theta^{*(\cdot,r)} \right)}{\theta^{(r)}} - 1. \tag{45}$$

The posterior distribution is estimated from N data points. For each data point the difference to the true reaction rate is measured. For comparison between the different reaction rates, it is normalized with the value of the true reaction rate. For instance a SNAR value of 2.5 for the reaction rate $\theta^{(r)}$ means that on average a particle from the posterior distribution has a distance of $2.5 \cdot \theta^{(r)}$ from $\theta^{(r)}$.

A SNAR value of zero can only be achieved if the posterior distribution consists of only the true reaction rate $\theta^{(r)}$. The width of the posterior distribution is reflected in the SNAR, as narrow distributions have a small SNAR because the difference to the true reaction rate is small.

To consider the influence of the prior distribution in the parameter estimation, the SNAR of the prior, $\text{SNAR}_{p(\theta)}$, is computed as well. Therefore, the estimated values $\theta^{*(i,r)}$ are taken as realisations from the prior distribution. A ratio between $\text{SNAR}_{p(\theta)}$ and the SNAR of the posterior distribution $\text{SNAR}_{p(\theta|x)}$ is calculated, thus

$$\text{SNAR ratio} = \frac{SNAR_{p(\theta)}}{SNAR_{p(\theta|x)}}. \tag{46}$$

A SNAR ratio greater than one means that in average the posterior distribution

is closer to the true reaction rate than the prior distribution.

4.3 Overview of simulations

Table 6 shows the different simulations, which were conducted. Each simulation is identified by an ID to be able to refer to them, and the reaction rates and initial conditions of the one-stage and two-stage model are stated. In all simulations the observed species is protein. The number of simulated trajectories in the SSA is 100. The maximum observed time T is set to twice the length until the mean of all species has reached the steady state, see chapter 4.1.2 for further explanation. The frequency of the observed data is based on the maximum of the determinant of the FIM, see chapter 4.1.3. The number of drawn particles N_{all} is either 10,000 or 100,000. The parameter σ_{LN} means that the log-normal distribution was chosen with σ_{LN}^2 for its second parameter. For simulations 1–4, the sampling frequency was set to the maximum of the FIM for TP data. This was done even for S-Cor and S-CC, which use TS data. The reason was to use the same simulated data for all distances. Simulations with the maximum of the FIM for TS data are handled in simulations 5 and 6. The parameters which change compared to the basic simulations, i.e. ID 1 and ID 3 for the one-stage and two-stage model, are in bold. For the two-stage model, σ_{LN} was set to 1.5 and not to 2 as in the one-stage model. This is due to the following: simulations showed that the data creation by the SSA is time consuming if the kinetic rates are large. Test runs for $\sigma_{LN} = 2$ were aborted after not having simulated enough data even after a couple of days. Other approaches, which replace the SSA, are computationally more efficient. They are mentioned in the outlook, chapter 7.

4.4 Implementation of the ABC SMC algorithm & choice of parameters

This section deals with the implementation of the ABC SMC. Additionally, the parameters which have not been yet introduced for the ABC algorithm are explained.

ID	model	$\theta^{(r)}$	X_j (t=0)	obs. species	N_{traj}	T_{obs}	Δ	N_{all} [10^3]	prior
1	one-stage	$\beta = 1.2$ $\delta = 0.7$	d(0)=1 p(0)=0	prot	100	23	0.09	10	$\sigma_{LN} = 0.2$
2	one-stage	——— as ID 1 ———							$\sigma_{\mathbf{LN}} = \mathbf{2}$
3	two-stage	$\alpha = 100$ $\beta = 1.2$ $\gamma = 2$ $\delta = 0.7$	d(0)=1 m(0)=0 p(0)=0	prot	100	38	0.11	10	$\sigma_{LN} = 0.2$
4	two-stage	——— as ID 3 ———							$\sigma_{\mathbf{LN}} = \mathbf{1.5}$
5	one-stage	——— as ID 1 ———					**1.35**	10	$\sigma_{\mathbf{LN}} = \mathbf{2}$
6	two-stage	——— as ID 3 ———					**0.84**	10	$\sigma_{\mathbf{LN}} = \mathbf{1.5}$
7	one-stage	——— as ID 1 ———						**100**	$\sigma_{\mathbf{LN}} = \mathbf{2}$
8	two-stage	——— as ID 3 ———						**100**	$\sigma_{\mathbf{LN}} = \mathbf{1.5}$

Table 6: Conducted simulations for the ABC. Parameters which change for the respective model compared to ID 1 and ID 3 are in bold.

Implementation of the ABC SMC algorithm In the ABC algorithm an acceptance rate τ is used instead of a threshold ϵ to determine which particles form the posterior distribution. Accordingly for the ABC SMC, the threshold ϵ is not used either. The algorithm, described in chapter 2.3.2, is changed from step 8 onwards to

8. Calculate the distance between the simulated and the experimental data $d(x^0, x^*)$ and record θ^{**} as θ_i^{**}.

9. While $i < N_{all}$ start from step 4.

10. Accept the $N = N_{all} \cdot \tau$ particles from θ_j^{**}, $j = 1, ..., N_{all}$ with the lowest distance, which are then recorded as $\theta^{(i,\cdot)}$, $i = 1, .., N$.

11. For $i = 1, .., N$ calculate the weights for particle $\theta_t^{(i,\cdot)}$ according to equation 11.

12. Normalize the weights so $\sum_i w_t^{(i,\cdot)} = 1$.

13. If $t < T$, set $t = t + 1$ and proceed with step 2.

So, in each population, instead of accepting the particles with a distance below a threshold ϵ_t, the $\tau \cdot N_{all}$ particles with the lowest distance are accepted.

As the threshold ϵ_t decreases with each population t, one could define a different acceptance rate τ_t for each population. In the simulations conducted, τ_t was set to the same value for all populations. This is because the resulting posterior distribution should be formed from the same number of particles N. As the number of drawn particles N_{all} stays constant for each population, τ_t has to stay constant as well.

Choice of parameters for the ABC SMC To perturb the particles, a component-wise perturbation kernel was used, as the prior distribution was component-wise as well. Both an uniform and a gaussian perturbation kernel were used. The uniform perturbation kernel perturbs a particle $\theta_t^{(i,r)}$ by using a uniform distribution $U[\theta_t^{(i,r)} - \sigma_t^{(r)}, \theta_t^{(i,r)} + \sigma_t^{(r)}]$. As discussed in chapter 2.3.2, according to Filippi et al. (2011), often the width of the previous population, equation (12), is used for $\sigma_t^{(r)}$. Due to the time consuming computational runtime for simulating the data, as described in chapter A.2, $\sigma_t^{(r)}$ could not be set as the width of the previous population.[13] Especially for the second population this would result in particles with a large value, which would make the simulation of the data very time consuming.

With the same argument, $\sigma_t^{(r)} = 2Var\left(\{\theta_t^{(i,r)}\}_i\right)$, as described in equation (13), could not be chosen for the gaussian perturbation kernel. It would result in $\sigma_t^{(r)} = 14,498$ for $\sigma_{LN} = 1.5$ and $\theta_t^{(r)} = 1$, $\sigma_t^{(r)} = 144.98 \cdot 10^6$ for $\sigma_{LN} = 1.5$ and $\theta_t^{(r)} = 100$,$\sigma_t^{(r)} = 17.45 \cdot 10^6$ for $\sigma_{LN} = 2$ and $\theta_t^{(r)} = 1$ and $\sigma_t^{(r)} = 174.47 \cdot 10^9$ for $\sigma_{LN} = 2$ and $\theta_t^{(r)} = 100$.

[13] The following population widths $\sigma_t^{(r)}$ were calculated on average from drawing 100 times 10^7 particles from a log-normal distribution with σ_{LN} and reaction rate $\theta_t^{(r)}$. $\sigma_t^{(r)} = 15,497$ for $\sigma_{LN} = 1.5$ and $\theta_t^{(r)} = 1$, $\sigma_t^{(r)} = 1.36 \cdot 10^6$ for $\sigma_{LN} = 1.5$ and $\theta_t^{(r)} = 100$, $\sigma_t^{(r)} = 1.21 \cdot 10^6$ for $\sigma_{LN} = 2$ and $\theta_t^{(r)} = 1$ and $\sigma_t^{(r)} = 1.21 \cdot 10^8$ for $\sigma_{LN} = 2$ and $\theta_t^{(r)} = 100$.

For the simulation the uniform perturbation kernel was set to

$$U[0.75 \cdot \theta_t^{(i,r)}, 1.25 \cdot \theta_t^{(i,r)}] \tag{47}$$

and the gaussian perturbation kernel was set to

$$N\left(\theta^{(i,r)}, \frac{0.25}{1.96} \cdot \theta^{(i,r)}\right). \tag{48}$$

The variance of the gaussian perturbation kernel is chosen to equal 0.25/1.96, meaning that 95% of the perturbed values lie within $[0.75 \cdot \theta^{(i,r)}, 1.25 \cdot \theta^{(i,r)}]$, i.e. the 2.5%-quantile $Q_{2.5\%} = 0.75 \cdot \theta^{(i,r)}$ and $Q_{97.5\%} = 1.25 \cdot \theta^{(i,r)}$.

The other parameters, including the observation time, the sampling frequency and the prior distribution were set as in the ABC algorithm, which is described in chapters 4.1.2, 4.1.3 and 4.1.5.

ID	model	$\theta^{(r)}$	X_j (t=0)	obs. species	N_{traj}	T_{obs}	Δ	N_{all} [10^3]	σ_{LN} of prior	T	τ_t	$K_t(\cdot,\cdot)$
9*	one-stage	———	as ID 1	———				2	0.2	5	0.005	$U[\cdot,\cdot]$
10*	one-stage	———	as ID 1	———				2	2	5	0.005	$U[\cdot,\cdot]$
11*	one-stage	———	as ID 1	———				2	2	5	0.005	$N(\cdot,\cdot)$
12†	two-stage	———	as ID 3	———				2	0.2	5	0.005	$U[\cdot,\cdot]$
13‡	two-stage	———	as ID 3	———				2	1.5	5	0.005	$U[\cdot,\cdot]$
14‡	two-stage	———	as ID 3	———				2	1.5	5	0.005	$N(\cdot,\cdot)$

* S-M&Std, S-NE II, S-cdf and S-MC were used as distance functions.
† S-M&Std, S-NE, S-cdf and S-MC were used as distance functions.
‡ S-M&Std, S-cdf and S-MC were used as distance functions.

Table 7: Conducted simulations for the ABC SMC.

Table 7 summarizes the parameter settings, which were used in the simulations for the ABC SMC.

The acceptance rate τ was set to 0.005 so that from each population the resulting

posterior is built from the ten best particles. In the ABC simulations the best results were achieved for $\tau = 0.001$, which means that the posterior is made of ten particles. Thus, τ has to be set to 0.005 in the SMC to achieve ten particles from a population of 2,000 particles.

5 Computational study for ABC rejection

In this chapter, the results from the computational study are presented. Chapter 5.1 deals with the results from 10,000 drawn particles. The one-stage model is discussed in chapters 5.1.1 and 5.1.2, and the two-stage model is described in chapters 5.1.3 and 5.1.4. For both models two prior distributions with a different degree of informativeness are used. Then chapter 5.2 deals with results for an optimal sampling frequency for TS data, which influences the distances S-Cor and S-CC. Using the information gained from the previous chapters, a new distance function S-MC is defined in chapter 5.3. It is a combination of S-Mean and S-Cor. Then simulations with 100,000 drawn particles for a non-informative prior have been conducted, which are presented in chapter 5.4 for the one-stage and two-stage model. The results from the two-stage model lead to the idea to estimate the kinetic rates in two steps. First γ and δ are estimated, then, based on these results, α and β are estimated. This approach is presented in section 5.5. A summary of the results from the simulations with the ABC algorithm is in section 5.6.

5.1 Results for 10,000 drawn particles

This chapter presents the results when N_{all} is set to 10,000. First, an informative prior distribution is used to analyse the performance of the ABC. Then, a less informative prior is taken to see how the estimation quality changes. Different analyses were conducted to try to gain insights into the distribution of the distance, into the behaviour of the distance functions, and into the importance of the acceptance rate τ and the number of drawn particles N_{all}. All analyses are presented and explained in chapter 5.1.1.

5.1.1 One-stage model with informative prior

In this section the informative prior distribution is used. It has a $\text{SNAR}_{p(\theta)}$ of 0.34 for the two kinetic rates together, i.e. for each kinetic rate the $\text{SNAR}_{p(\theta)}$ is about 0.17.

Table 8 shows the SNAR statistics for simulation 1. Comparing the SNAR ratio across the different acceptance rates, it has its greatest value at $\tau = 0.001$ for

distance	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC
$\text{SNAR}_{p(\theta)}$	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
$\tau = 0.001$									
$\text{SNAR}_{p(\theta\|x)}$	0.19	0.19	0.13	0.3	0.27	0.22	0.19	0.74	0.89
ratio	1.79	1.79	2.57	1.12	1.25	1.57	1.78	0.47	0.39
$\tau = 0.01$									
$\text{SNAR}_{p(\theta\|x)}$	0.2	0.19	0.19	0.37	0.23	0.19	0.19	0.56	0.72
ratio	1.73	1.78	1.8	0.92	1.48	1.77	1.77	0.61	0.48
$\tau = 0.05$									
$\text{SNAR}_{p(\theta\|x)}$	0.22	0.21	0.2	0.36	0.25	0.22	0.2	0.48	0.57
ratio	1.58	1.61	1.68	0.96	1.37	1.58	1.67	0.72	0.6

Table 8: SNAR statistics for simulation 1 for acceptance rates $\tau = 0.001, \tau = 0.01$ and $\tau = 0.05$.

S-Mean, S-Std, S-M&Std and S-cdf. Intuitively this should be the case, as the smaller the τ, the smaller the distance of the particles, which are accepted for the posterior, and, therefore, the closer the value of the particles to the true reaction rate.

Although the SNAR value for the prior is simulated out of 10^7 random samples, the SNAR ratio varies slightly when calculated, and it will vary when the same simulation is conducted again. This is also due to the fact that the SNAR value for the posterior depends on a small number of particles. Therefore, a slight change of the SNAR ratio is not necessarily indicating a change in the estimation quality.

The SNAR ratio is approximately the same for all acceptance rates for S-Mean, S-Std and S-cdf. A considerably decrease of the SNAR ratio is for S-M&Std between $\tau = 0.001$ and $\tau = 0.01$.

For S-Cor and S-CC the SNAR ratio is less than one. This means that sampling from the prior distribution yields a better result for the estimation of the reaction rate than sampling from the posterior distribution. Although S-Cor and S-CC use TS data, i.e. the data containing the most information, this advantage of information is not reflected in the SNAR ratio. For S-NE it is approximately one, meaning that on average there is no improvement from the prior to the posterior distribution.

In figure 9 the prior and posterior distribution for the selected distances S-Mean, S-pdf and S-Cor are illustrated. The parameters are in log-scale, thus the lognormal prior distribution has the shape of a normal distribution. The acceptance rate was set to $\tau = 0.01$ meaning that the posterior distribution is estimated from 100 particles. For $\tau = 0.001$ the posterior would have been narrower for some distances, but the kernel estimation would be based on 10 particles. The plot for all distances can be found in figure A.44.

It becomes clear that the SNAR ratio reflects the behaviour of the posterior distribution, as distances such as S-Mean and S-pdf, which produce a posterior closer to the true value, have a SNAR ratio greater one. The posterior for S-Mean and S-pdf have approximately the same quality of estimating the true value, which is also reflected in a similar SNAR ratio of 1.73 and 1.77 respectively. S-Cor has a SNAR ratio of 0.61, and its posterior distributions is both for β and δ further away from the true value than the prior.

To identify the true reaction rates, the distance around the true reaction rates should be the lowest. Figure 10 shows the logarithm of the distance against the values of each reaction rate sampled from the prior. In an optimal scatterplot the distribution of the distance would be U-shaped with its minimum at θ. This does not hold for any distance function. For S-Mean, S-Std, S-M&Std, S-pdf and S-cdf, the maximum of the distance grows for reaction rates which are further from the true value. S-Cor and S-CC have a linear dependence of parameter value and distance.

For all distance functions, for the same value of reaction rate, a wide distribution of different distances result. The only exception is S-Cor for rate δ. From the former observations it can be assumed that the distance is influenced by both reaction rates. To investigate this, a scatterplot of the distance is shown in figure 11. The distance is colored in nine steps, which represent nine equidistant quantiles, i.e. the quantiles $Q_{0.11}, Q_{0.22}$, etc. A darker orange color implies a larger distance. For β and δ, 0.1% of the largest data points were omitted so that the plots are visually clearer. A desired distribution of the distance would be that the distance is the lowest around the true parameter values, i.e. a bright circle in the center of each plot. The distance functions S-Mean, S-Std, S-M&Std, S-pdf and S-cdf show their lowest distance values along the ratio of δ and β. One should keep in mind

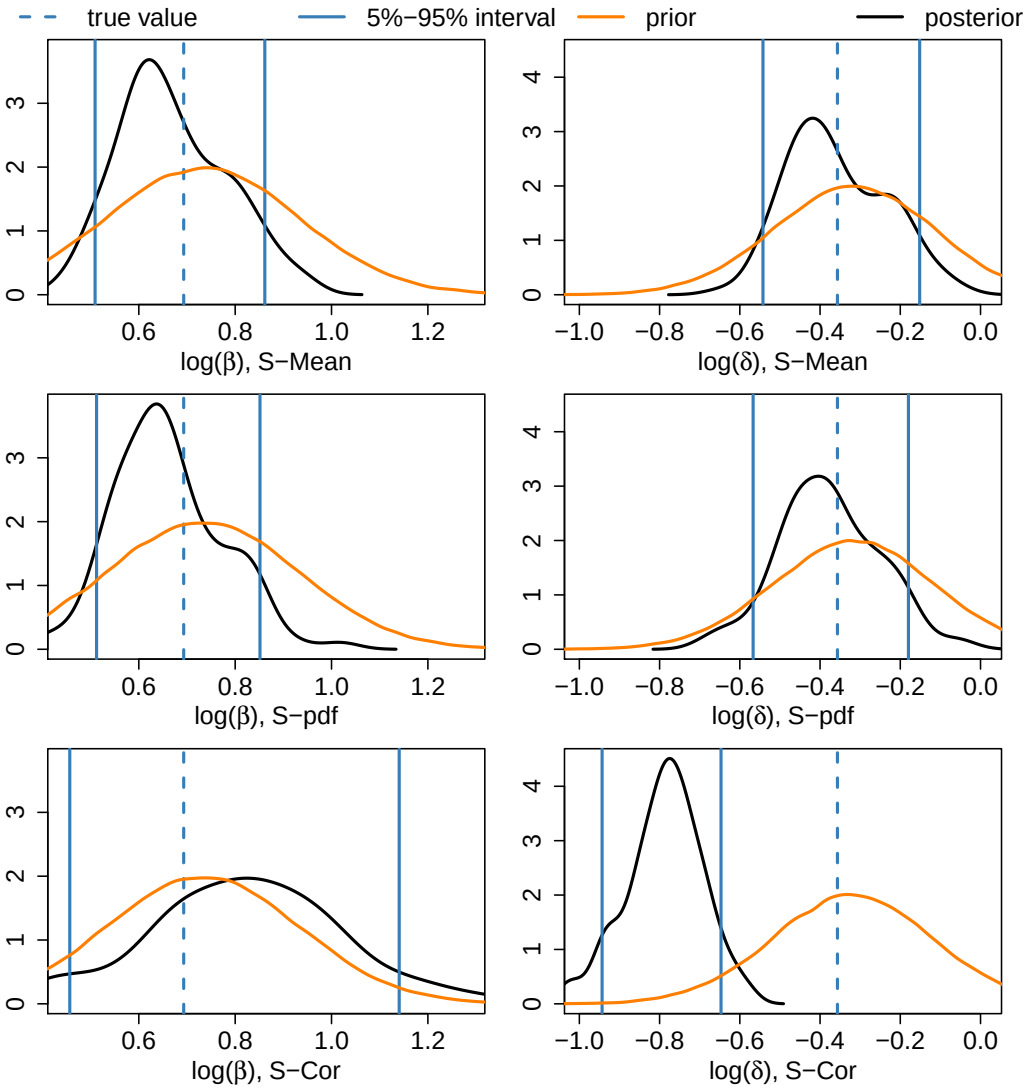
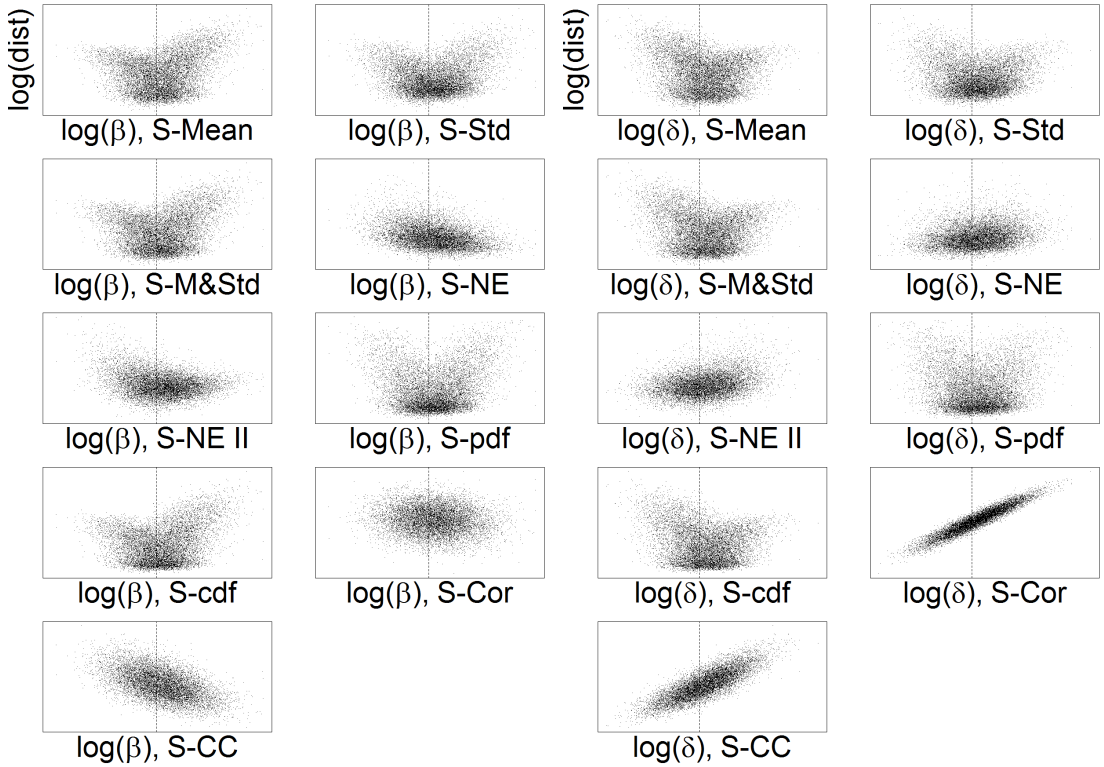


Figure 9: Kernel density estimation of the prior and posterior distribution for S-Mean, S-pdf and S-Cor for the one-stage model (simulation 1), $\tau = 0.01$.

that the steady state in a one-stage model is β/δ. Thus these distance functions can estimate correctly the steady state. The distance functions S-NE and S-NE II do not show this attribute, their distance is somewhat evenly distributed with a tendency to have higher distance values for smaller values of β. The distance functions S-Cor and S-CC have their lowest distances for small values of δ.

The results imply that for certain distance functions the particles with approx-



a) reaction rate β. b) reaction rate δ.

Figure 10: Scatterplot of the logarithm of the distance for nine distance functions for reaction rates β and δ for the one-stage model (simulation 1). The dashed line indicates the logarithm of the true reaction rate.

imately the lowest 11.11% distance values are centered along the ratio of the reaction rates. It is assumed that a smaller acceptance rate than $\tau = 0.1111$ yields parameter estimations closer to the true values. Therefore, the sampled reaction rates with the lowest distance values are plotted. Figure 12 contains the prior distribution of the kinetic rates in cyan, with yellow signalling a higher density. Additionally, the posterior distributions for $\tau = 0.05$, $\tau = 0.01$ and $\tau = 0.001$ are plotted in black, red, and blue respectively. The point size of the posterior distributions is double the size of the points of the prior to be able to differentiate the different posteriors. For the distance functions which result in a low distance along the ratio of the reaction rates, the acceptance rate does not have a major influence on the quality of estimation, as for all τ the posterior has approximately

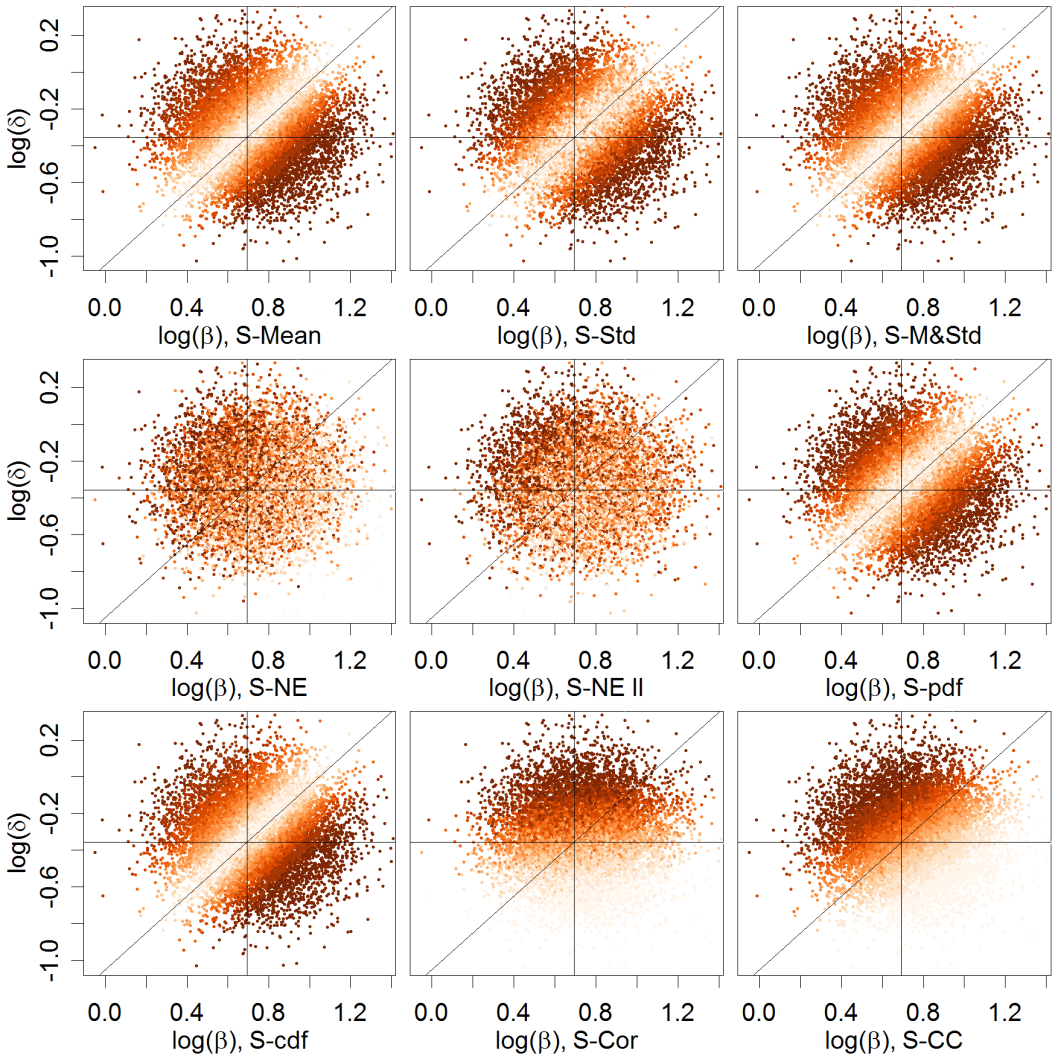


Figure 11: Distance values for the nine distance functions depending on both reaction rates β and δ for the one-stage model (simulation 1). The solid lines indicate the true reaction rates and the ratio δ/β.

the same width of distribution. Only for S-M&Std the blue points lie closer to the true value. This is reflected in a higher SNAR ratio as for $\tau = 0.01$ or $\tau = 0.05$ as well. For S-Cor and S-CC the parameters closer to the right bottom corner, i.e. high β and low δ value, have a smaller distance. The distribution of the distance for S-NE and S-NE II does not seem to change with a different acceptance rate.

To investigate this further, if the quality of the posterior distribution depends on

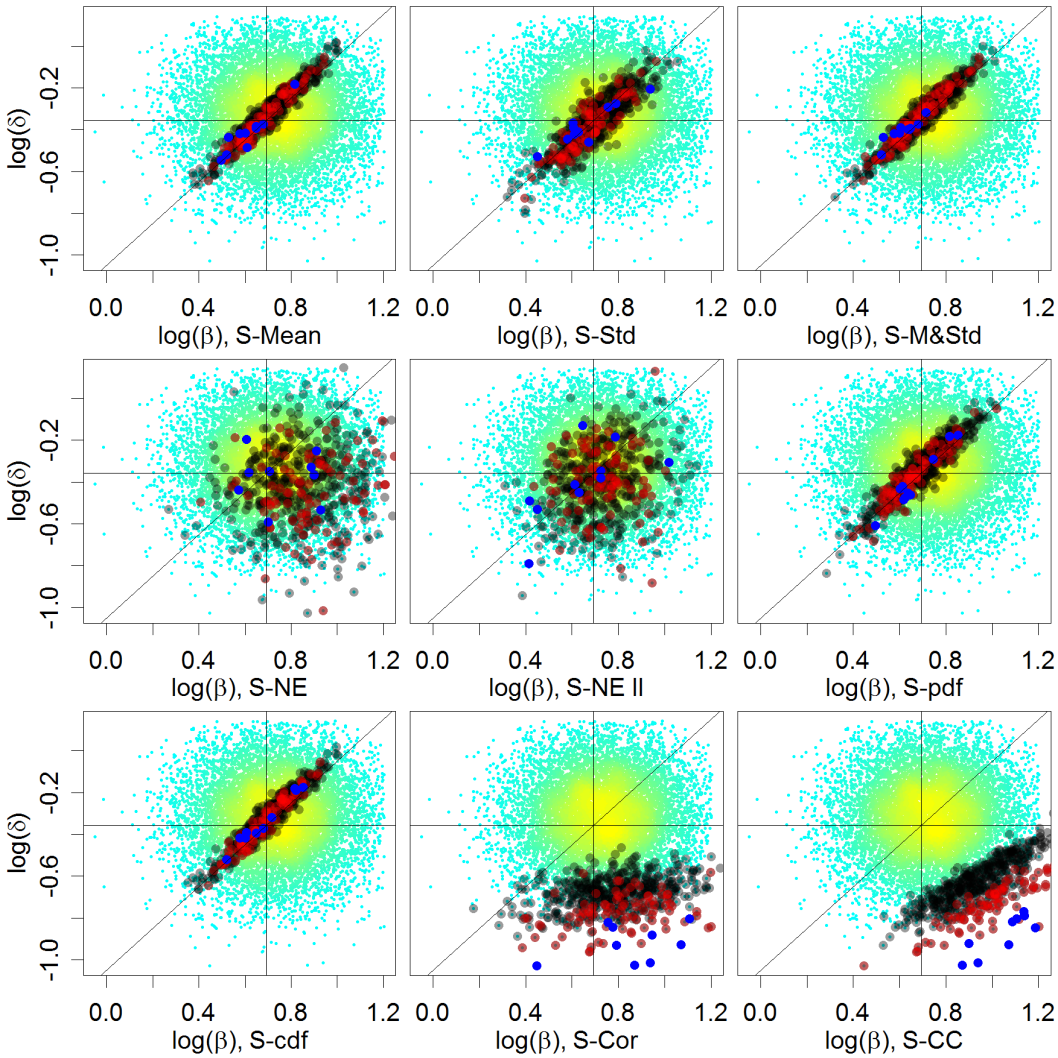


Figure 12: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the one-stage model (simulation 1).

the acceptance rate, the SNAR ratio is plotted in figure 13 against the acceptance rate τ. The plot is for τ up to 0.5 as for ABC sampling an acceptance rate greater 0.5 is rare.

The lowest τ for the calculation in figure 13 was $\tau = 0.001$, i.e. only the ten particles with the lowest distance were accepted.

S-Cor and S-CC have an increasing SNAR ratio for increasing τ. This is the

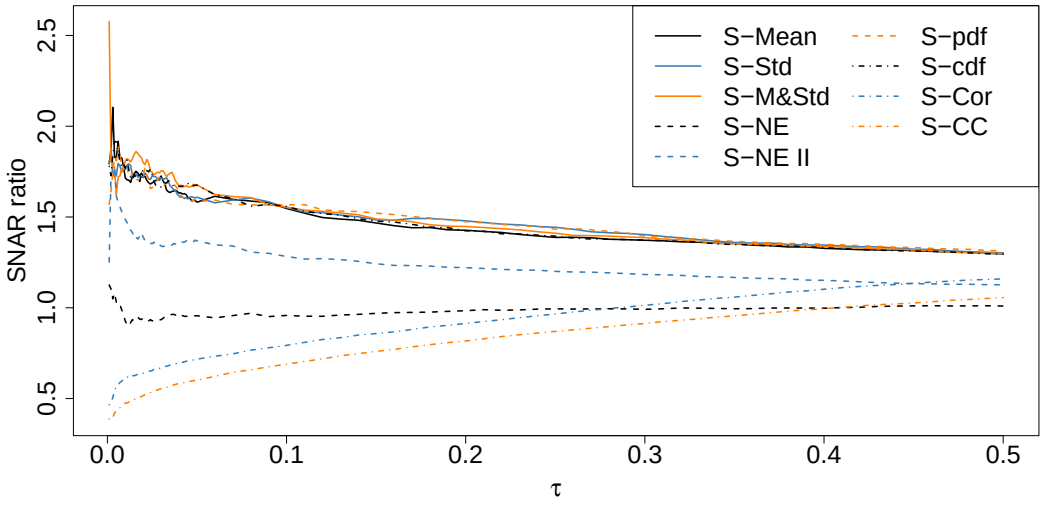


Figure 13: SNAR ratio depending on the acceptance rate τ for all distances for the one-stage model (simulation 1).

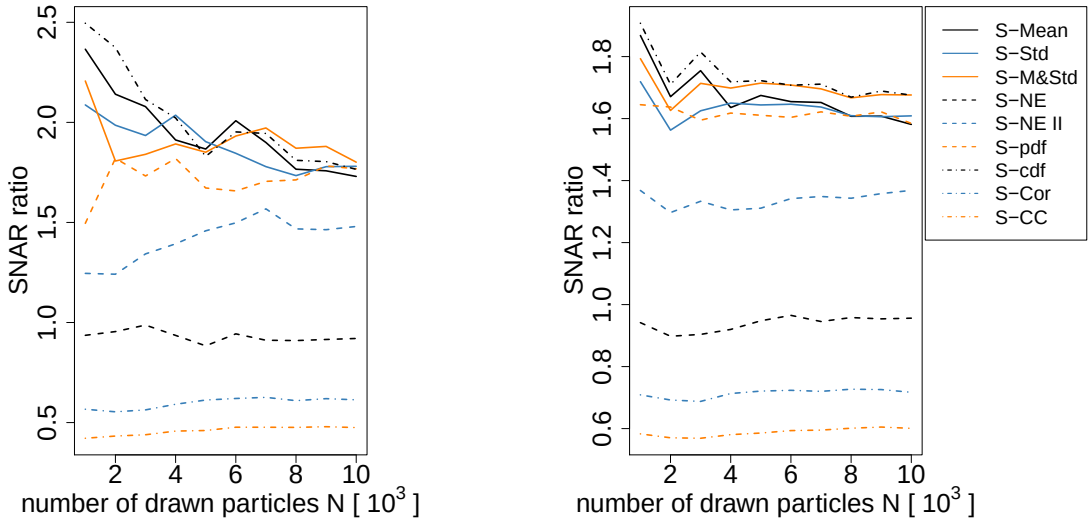
case as for a small τ the particles are far away from the true reaction rate values, and with increasing τ the accepted particles move towards the true reaction rate values.

The SNAR ratio of S-NE tends to be around one. Therefore, this distance does not have the ability to identify the particles which represent a good estimation. Instead, random particles from the prior seem to build the posterior distribution.

S-NE II shows a sharp increase at the beginning up to a maximum of around 1.75 and then decreases. S-pdf shows a similar behaviour with a maximum of about 1.8. The increase at the beginning can be due to the small number of accepted particles. All other distances have an almost exact characteristic, only the maximum value at $\tau = 0.001$ differs. For S-Mean it is approximately 1.79, for S-Std around 1.79, for S-M&Std around 2.57 and for S-cdf about 1.78.

The following figure 14 examines the SNAR ratio against the number of drawn particles for two acceptance rates $\tau = 0.01$ and $\tau = 0.05$. The rate $\tau = 0.001$ was not investigated as for the lowest $N_{all} = 1000$, the posterior would only consist of one particle.

As for $\tau = 0.01$ the fluctuation of the SNAR ratio is still rather large even for large N_{all}, an investigation with a higher amount of drawn particles is necessary

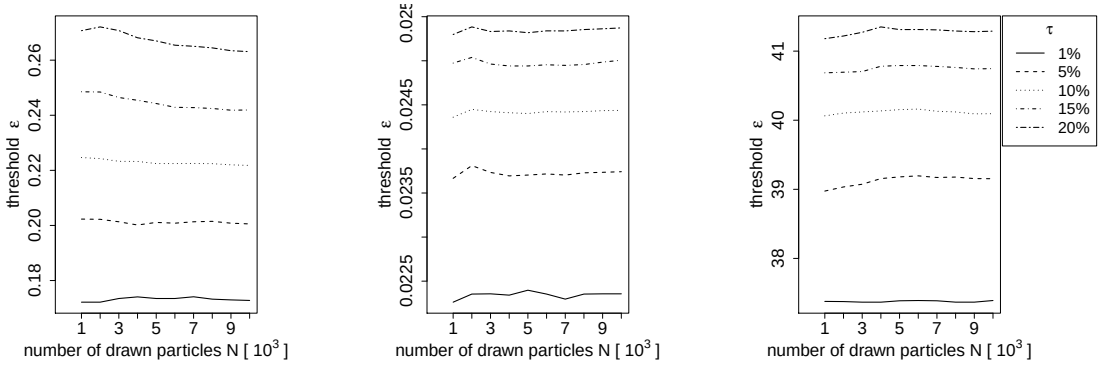


a) $\tau = 0.01$. b) $\tau = 0.05$.

Figure 14: SNAR ratio depending on the number of drawn particles N_{all} for all distances for the one-stage model (simulation 1) and two acceptance rates $\tau = 0.01$ and $\tau = 0.05$.

to see if the ratio stays constant for increasing N_{all}.

Figure 15 shows the theoretical threshold ϵ for different acceptance rates τ and three exemplary distances S-Mean, S-NE II and S-Cor. The threshold for a certain acceptance rate is the largest distance for which a particle is still accepted. The



a) S-Mean b) S-NE II c) S-Cor

Figure 15: Threshold ϵ depending on the number of drawn particles N_{all} for three distance functions and five acceptance rates for the one-stage model (simulation 1).

threshold is approximately constant and does not depend on the number of drawn particles. It only depends on the acceptance rate. Therefore, it can be assumed that, for a given prior distribution, a fixed threshold ϵ always leads to the same expected acceptance rate. A sketch of the proof is in the appendix, chapter A.11.

5.1.2 One-stage model with less informative prior

The difference between this simulation and simulation 1 is the prior distribution, which is less informative. The parameter σ_{LN} was set to $\sigma_{LN} = 2$. Table 9 contains the SNAR values for the prior and posterior distributions. The prior has a SNAR value of $\text{SNAR}_{p(\theta)} \approx 801.9$. Compared with a value of 0.34 in simulation 1, the estimation of the reaction rates is based on much more less information. Compared with simulations conducted in other publications, the prior is considerably less informative. In table 5 the SNAR value of the prior for a single parameter is only in Tellier et al. (2011) for certain settings about 500, for the other papers it is less than 5. As the relative width of the prior is the same for both parameters, the $\text{SNAR}_{p(\theta)}$ divides equally to each of the two parameters, thus resulting in about 402 for each parameter.

distance	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC
$\text{SNAR}_{p(\theta)}$	806.21	806.21	806.21	806.21	806.21	806.21	806.21	806.21	806.21
$\tau = 0.001$									
$\text{SNAR}_{p(\theta\|x)}$	8.19	4.14	8.01	2627	2.14	13050	6.24	1728	2978
ratio	98.39	194.66	100.69	0.31	376.38	0.06	129.2	0.47	0.27
$\tau = 0.01$									
$\text{SNAR}_{p(\theta\|x)}$	181.01	168.18	188.27	1046	143.97	6756	181.04	726.46	531.12
ratio	4.45	4.79	4.28	0.77	5.6	0.12	4.45	1.11	1.52
$\tau = 0.05$									
$\text{SNAR}_{p(\theta\|x)}$	264.34	261.9	267.11	1215	412.86	3072	266.33	354.01	842.67
ratio	3.05	3.08	3.02	0.66	1.95	0.26	3.03	2.28	0.96

Table 9: SNAR statistics for simulation 2 for acceptance rates $\tau = 0.001, \tau = 0.01$ and $\tau = 0.05$.

S-Mean, S-Std, S-M&Std, S-NE II and S-cdf have a large SNAR ratio for $\tau = 0.001$, and its SNAR value increases for larger τ. The distances S-NE, S-pdf and partly S-CC do not achieve a better parameter estimation than the prior

distribution. The SNAR ratio of S-Cor is increasing for larger τ meaning that, similar to simulation 1, the particles with the lowest distance are further away from the true value.

The figures of the prior and posterior distributions for S-NE II, S-pdf and S-CC are illustrated in figure 16. The parameters are in log-scale, thus the lognormal prior distribution has the shape of a normal distribution. As the prior is chosen with the mode being at the true value θ, the normal distribution has the mean $log(\theta) + \sigma^2$ and standard deviation σ, which is reflected in the aspect that the mode of the normal distribution is far to the right of the reaction rate $log(\theta)$.

For $\tau = 0.01$, S-NE II has the best SNAR ratio, which is the case as the posterior is moved to the true values for both reaction rates. S-pdf and S-CC have a better estimation of δ. Their estimation of β, however, is far worse than the prior, which results in a SNAR ratio less than one. The kernel density estimations for all distances are in figure A.49.

The scatterplots showing the distribution of the distance against the single value of a reaction rate are shown in figure A.45 in the appendix.

In figure A.46 the distribution of the distance is plotted against the logarithm of both reaction rates β and δ. As the results are similar to simulation 1 it is shown and commented in the appendix.

The distribution of the prior and the distribution of the resulting posterior for $\tau = 0.001$, $\tau = 0.01$ and $\tau = 0.05$ are shown in figure 17. Similar to simulation 1, S-Mean, S-Std, S-M&Std and S-cdf estimate the ratio of the parameters β and δ very accurate. For $\tau = 0.01$ and $\tau = 0.05$ the posterior seems to have about the same width, however, for $\tau = 0.001$ the accepted particles are closer to the true values. S-Cor and S-CC estimate δ well, but for β the particles are distributed widely. The accepted particles of S-pdf seem to be on a line as well, whereby the line is parallel to the ratio β/δ.

For selected distances the SNAR value increases for smaller τ, as can be seen in table 9 as well. A plot with the SNAR ratio depending on the acceptance rate shows the same and is, therefore, in figure A.47 in the appendix.

The SNAR ratio depending on the number of drawn particles is also discussed in the appendix in figure A.48 as the result is similar to simulation 1.

Similar to simulation 1, the threshold ϵ is approximately constant for a given τ

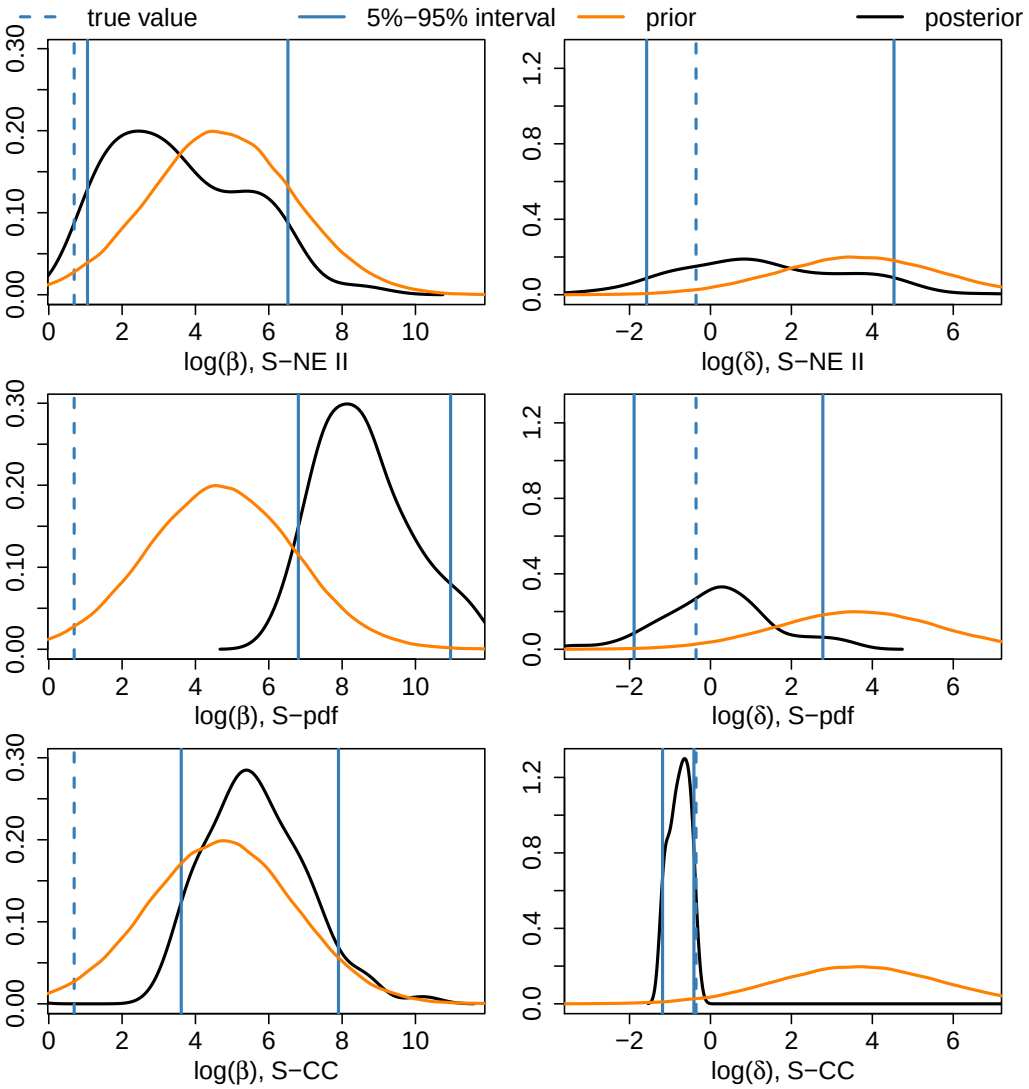


Figure 16: Kernel density estimation of the prior and posterior distribution for S-Mean, S-pdf and S-Cor for the one-stage model (simulation 2), $\tau = 0.01$.

regardless of the number of drawn particles. As it is shown in chapter A.11 that this holds in general, no plots are drawn.

Comparing simulation 1 and 2 Comparing simulation 1 and simulation 2, the question arises if the improvement in the posterior is the larger, the less informative the prior is. It seems that this is the case as most of the distances yielded a higher

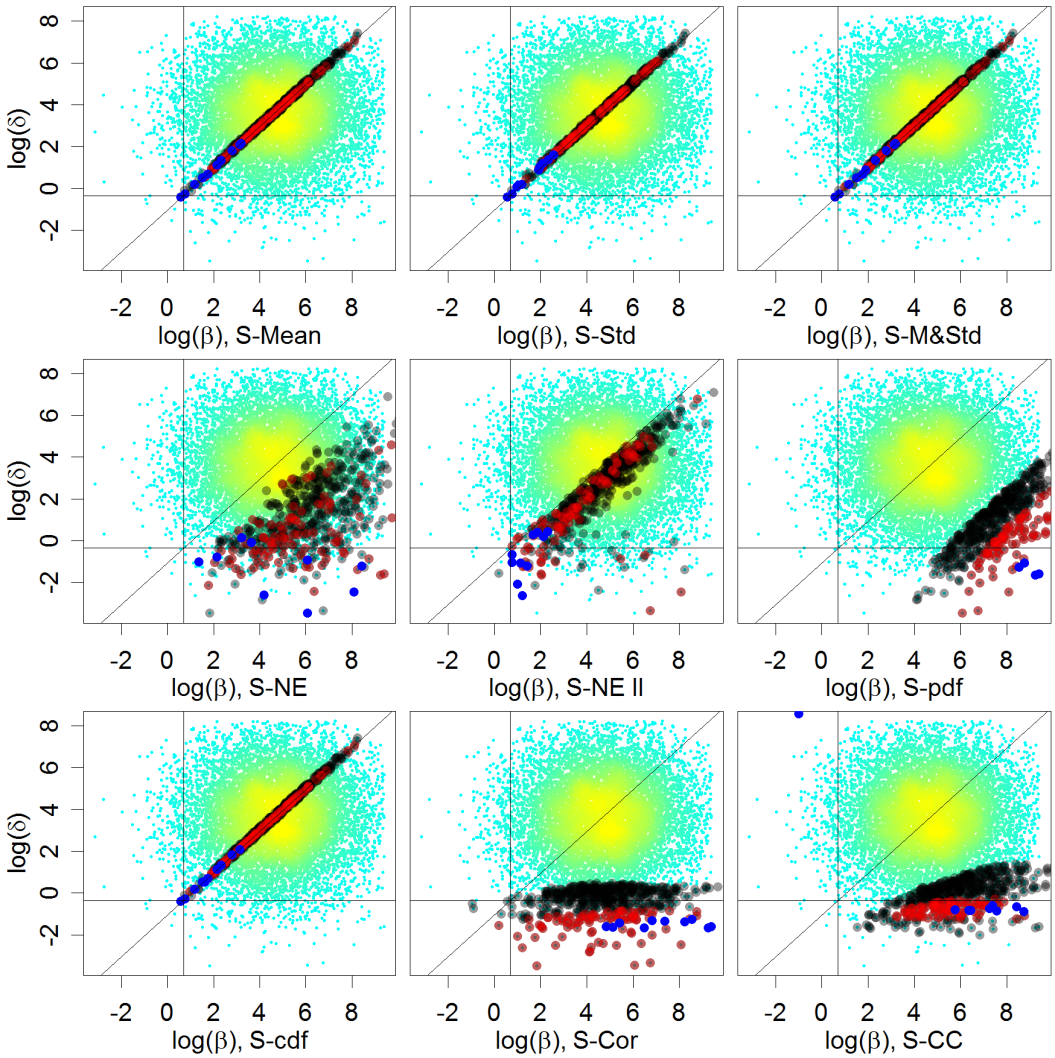
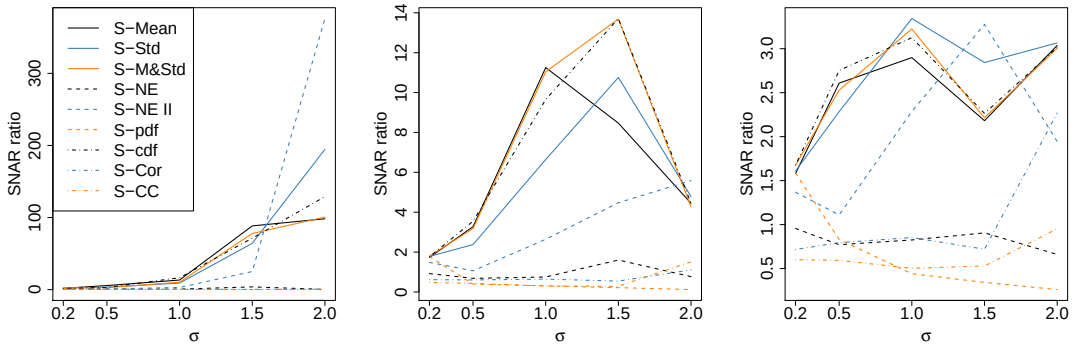


Figure 17: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the one-stage model (simulation 2).

SNAR ratio for a less informative prior. To check this, other simulations were conducted. The result is presented in figure 18.[14] The distances S-NE, S-pdf, S-CC and S-Cor have a SNAR ratio around one or less for almost all σ_{LN} and the three acceptance rates. These distances provide a good estimation of the

[14]For better illustration, the measured SNAR ratios are interpolated with straight lines, although it does not necessarily represent the true evolution of the ratio.



a) $\tau = 0.001$ b) $\tau = 0.01$ c) $\tau = 0.05$

Figure 18: SNAR ratio depending on σ_{LN} of the prior distribution and acceptance rates $\tau = 0.001$, $\tau = 0.01$ and $\tau = 0.05$ for the one-stage model.

parameter δ for the less informative prior, even a better estimation than the other distances, but the estimate of β is less accurate than by the prior. For the prior with $\sigma_{LN} = 0.2$, only S-pdf estimated both parameters correctly. S-NE, S-CC and S-Cor result for both parameters in a worse estimation than the prior.

For S-Mean, S-Std, S-M&Std and S-cdf the SNAR ratio develops roughly the same for different σ_{LN} over all acceptance rates.

NE II has the highest SNAR ratio for $\tau = 0.001$ and $\tau = 0.01$ for the least informative prior with $\sigma_{LN} = 2$, but it does not provide such a good estimation for more informative priors than the former mentioned group of distances.

Summary The following findings and presumptions result from the previous simulations.

- The distances can be grouped in three groups. The first group being S-Mean, S-Std, S-M&Std and S-cdf, which estimate the ratio of β/δ very well, and which yield for different σ_{LN} and different τ approximately the same improvement. The second group being S-NE, S-pdf, S-Cor and S-CC, which have quite an exact estimation of the parameter δ, but cannot estimate β resulting in a poor overall performance. The third group is S-NE II, which does not estimate the ratio β/δ as well as the first group nor one parameter as the second group, but has an acceptable overall performance, especially for $\sigma_{LN} = 2$ being the best distance function there.

- For a smaller τ the distance functions have a better estimation. For group one this is true for $\sigma_{LN} = 0.2$ and $\sigma_{LN} = 2$, for group three especially for $\sigma_{LN} = 2$. For the second group this holds for the estimation of parameter δ and a less informative prior.

- Using a high informative prior, the improvement, which can be achieved in estimating the true kinetic rates, is not as strong as for a less informative prior. But still, for the informative prior an improvement could be achieved, especially with the first group of distance functions.

- The SNAR ratio and, therefore, the quality of the posterior does not seem to depend mainly on the number of drawn particles N_{all}. It seems to depend on the acceptance rate τ, as the SNAR ratio stays roughly constant for a given τ after enough particles have been drawn. This assumption needs to be checked with a higher amount of drawn particles.

- The threshold ϵ stays approximately constant for a given τ after enough particles have been drawn. Thus one can conclude that every acceptance rate τ is equivalent to a threshold ϵ, depending only on the chosen distance.

- The distance depends on all kinetic rates. Analyzing it separately for each kinetic rate does not give all the information gained from observing a combination of kinetic rates.

5.1.3 Two-stage model with informative prior

The two-stage model, its parameters and exemplary trajectories are presented and discussed in chapter 2.4.2. This chapter deals with the results of parameter estimation with the two-stage model. The findings are interpreted in the context of the two-stage model and with comparison to the one-stage model. First, parameter estimation is done with an informative prior, then, in section 5.1.4 a less informative prior is used.

Table 10 contains the SNAR statistics for simulation 3.

The distances S-Mean, S-M&Std and S-cdf have approximately the same SNAR statistics for all acceptance rates and a higher SNAR ratio for lower τ. The ratio

distance	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC
$\text{SNAR}_{p(\theta)}$	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
$\tau = 0.001$									
$\text{SNAR}_{p(\theta\|x)}$	0.42	0.57	0.41	0.56	0.65	1.24	0.39	1.42	1.32
ratio	1.62	1.21	1.67	1.23	1.06	0.55	1.76	0.48	0.52
$\tau = 0.01$									
$\text{SNAR}_{p(\theta\|x)}$	0.5	0.57	0.47	0.64	0.65	0.61	0.47	1.12	1.17
ratio	1.38	1.21	1.45	1.07	1.05	1.12	1.45	0.61	0.58
$\tau = 0.05$									
$\text{SNAR}_{p(\theta\|x)}$	0.58	0.57	0.58	0.67	0.65	0.6	0.57	0.9	0.97
ratio	1.19	1.2	1.19	1.03	1.05	1.15	1.2	0.77	0.71

Table 10: SNAR statistics for simulation 3 for acceptance rates $\tau = 0.001, \tau = 0.01$ and $\tau = 0.05$.

does not change for S-Std and S-NE II for different τ. For S-Cor and S-CC the SNAR ratio is less than one for the different acceptance rates.

The SNAR ratio reflects the overall picture of the posterior distributions of the four kinetic rates. The prior and posterior distributions of selected distances are in figure 19. S-NE II, with a SNAR ratio slightly greater one, has posterior distributions, which are a bit narrower than the prior but do not gain a lot of information in estimating the parameter. For S-cdf, with the highest SNAR ratio, all of its posterior distributions have its mode very close to the true reaction rate. For S-CC the posterior is shifted further from the true value than the prior distribution. It is a similar behaviour as for the one-stage model. Interestingly, for α and β the posterior is shifted to the right of the true value, for γ and δ to the left. A similar attitude has S-Cor. The kernel density estimations of S-Cor and all distances are presented in figure A.52.

The scatterplot of the distance value against the value of each single reaction rate are not shown as there cannot be gained any further insights as from simulation 1.

There are six possible combinations for plotting two of the four reaction rates against each other. Pre analysis showed that the most interesting of these combinations is α versus γ as the ratio of these parameters is the steady state for mRNA. The other combination, for the steady state of protein, is $\alpha\beta$ versus $\gamma\delta$.

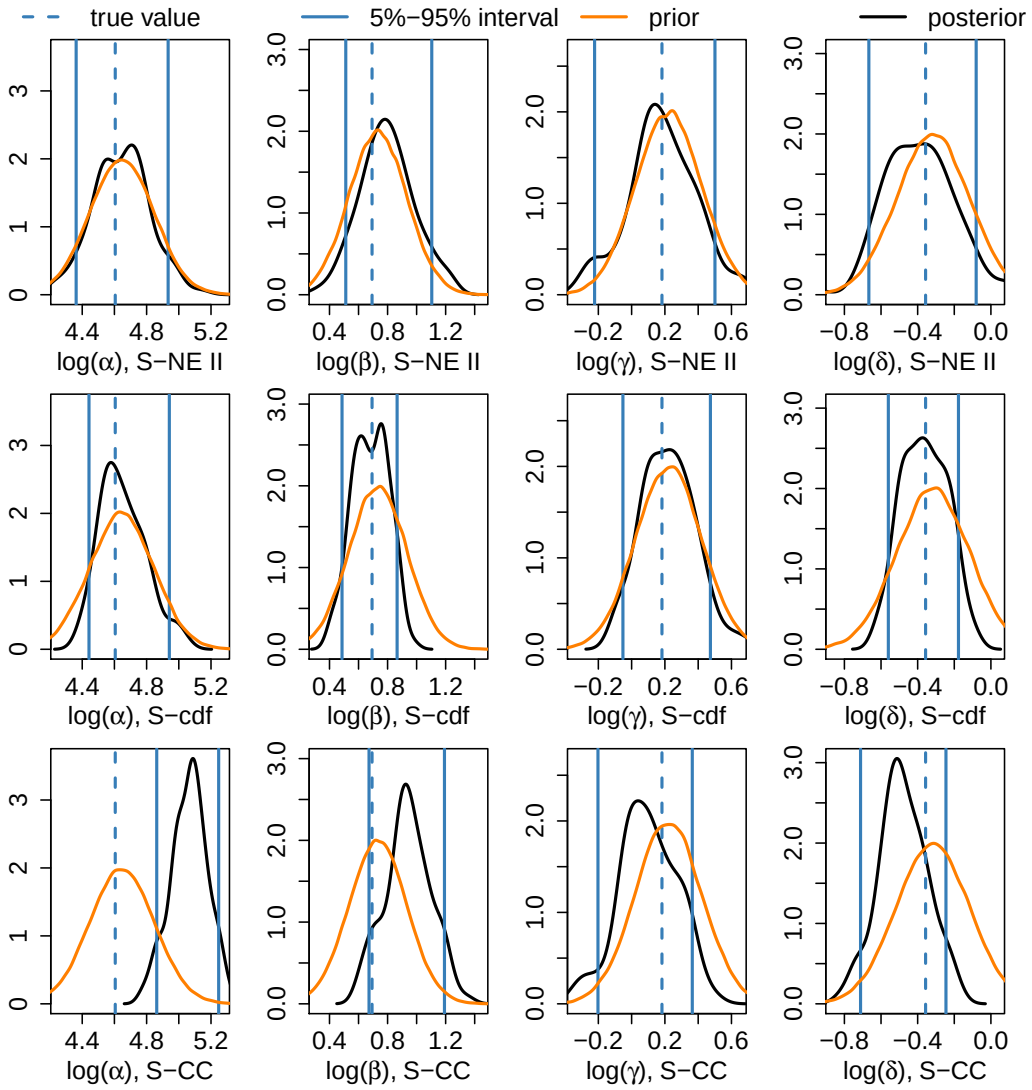


Figure 19: Kernel density estimation of the prior and posterior distribution for S-NE II, S-cdf and S-CC for the two-stage model (simulation 3), $\tau = 0.01$.

The distribution for the distance values for these two combinations are shown in the appendix in figures A.50 and A.51, as the main information can also be seen in the scatterplot of the prior and posterior distributions. Figures 20 and 21 show these plots.[15]

[15]For better visualization the 50 particles being at the extreme of the x- and y-axis have been removed. This results in a compacter illustration for the expense that some points of the posterior distributions are only half visible.

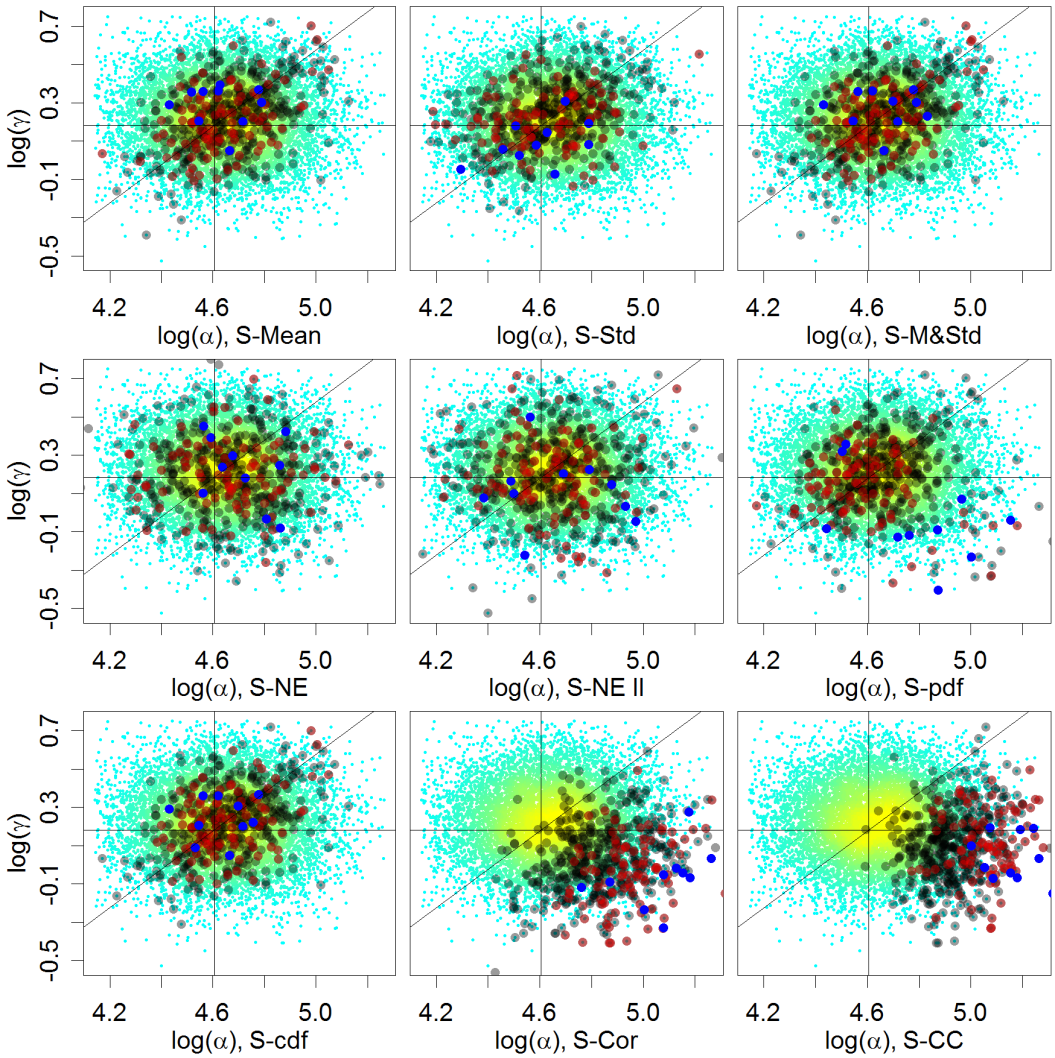


Figure 20: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the two-stage model (simulation 3) for α versus γ.

For figure 20, none of the distances estimates the steady state of mRNA, i.e. the ratio α/γ, correctly. A reason could be that only protein was observed. The posterior of S-Cor is more in the bottom right, meaning higher values for α and lower values for γ than the true reaction rates. For S-CC the posterior overestimates α but is distributed closely around the true reaction rate for γ.

For the steady state of protein, represented in figure 21, S-Mean, S-M&Std and

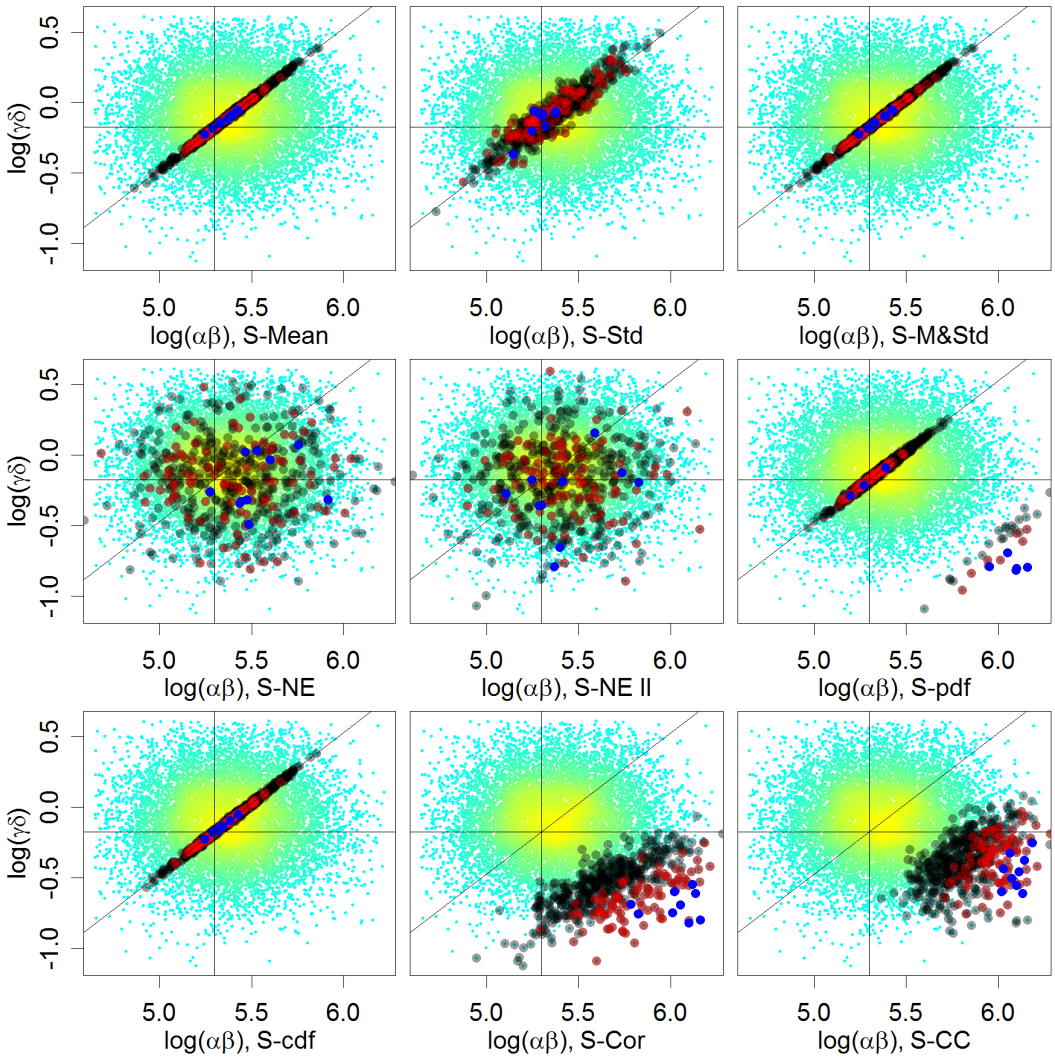


Figure 21: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the two-stage model (simulation 3) for $\alpha\beta$ versus $\gamma\delta$.

S-cdf estimate the ratio very exact. S-Std has a larger variance at estimating the ratio. For S-pdf not all particles, which form the posterior, are centered along the ratio. For S-Cor and S-CC a similar behaviour as in the one-stage model can be seen, as the particles forming the posterior are particles with rather large $\alpha\beta$ value and low $\gamma\delta$ value.

The development of the SNAR ratio depending on the acceptance rate is shown

in figure A.53 in the appendix, as it does not provide new insights compared with the knowledge gained from the one-stage model.

The SNAR ratio against a different number of drawn particles will be analysed in detail for 100,000 drawn particles in chapter 5.4.

5.1.4 Two-stage model with less informative prior

Different to the one-stage model, the parameter σ_{LN} for the less informative prior was set to 1.5 instead of 2. This is due to the high simulation time and is explained in chapter A.2. The $\text{SNAR}_{p(\theta)}$ of all prior distributions is about 113, i.e. for each single kinetic rate it is one fourth of it.

	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC
$\text{SNAR}_{p(\theta)}$	113.19	113.19	113.19	113.19	113.19	113.19	113.19	113.19	113.19
$\tau = 0.001$									
$\text{SNAR}_{p(\theta\|x)}$	42.83	27.28	15.68	132.69	267.53	563.63	26.88	125.66	90.41
ratio	2.64	4.15	7.22	0.85	0.42	0.2	4.21	0.9	1.25
$\tau = 0.01$									
$\text{SNAR}_{p(\theta\|x)}$	124.71	52.47	114.95	82.82	97.88	315.88	119.39	100.72	84.53
ratio	0.91	2.16	0.98	1.37	1.16	0.36	0.95	1.12	1.34
$\tau = 0.05$									
$\text{SNAR}_{p(\theta\|x)}$	90.14	81.08	89.34	92.43	104.63	232.34	89.26	86.92	87.58
ratio	1.26	1.4	1.27	1.22	1.08	0.49	1.27	1.3	1.29

Table 11: SNAR statistics for simulation 4 for acceptance rates $\tau = 0.001, \tau = 0.01$ and $\tau = 0.05$.

S-Mean, S-M&Std and S-cdf have a SNAR ratio greater one for $\tau = 0.001$, which drops below one for $\tau = 0.01$. In simulation 3, these three distances had about the same SNAR values for different τ, for simulation 4 they differ considerably for $\tau = 0.001$. S-Std shows a decreasing SNAR ratio for increasing τ. The distances S-NE II, S-pdf, S-Cor and S-CC perform partly better than in simulation 3 regarding their SNAR values, but their ability for constructing a good posterior for all kinetic rates is low.

Figure 22 shows kernel density estimations for the prior and posterior distributions of selected distances. S-NE II estimates γ and δ well, and for α and β the posterior is almost the same as the prior. The SNAR ratio less than one for S-pdf

is due to its estimation of α and β where the posterior is further away from the true value than the prior. But the estimation of γ and δ is very accurate. S-CC has the best estimation for γ and δ.

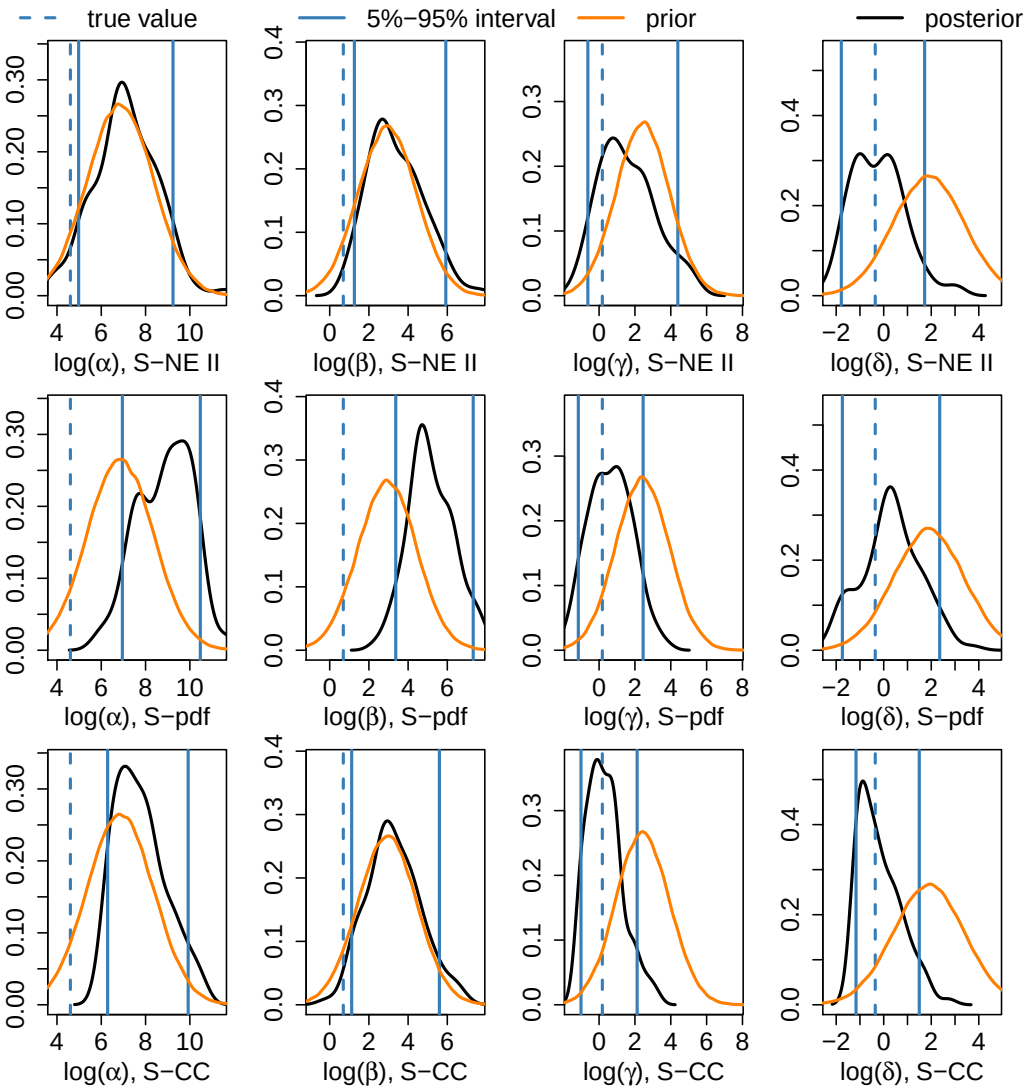


Figure 22: Kernel density estimation of the prior and posterior distribution for S-NE II, S-pdf and S-CC for the two-stage model (simulation 4), $\tau = 0.01$.

The distribution of the distance for parameters α versus γ and $\alpha\beta$ versus $\gamma\delta$ are shown and described in the appendix, figures A.54 and A.55.

Figures 23 and 24 show the prior distribution and the posterior distribution for

two combinations of the reaction rates.

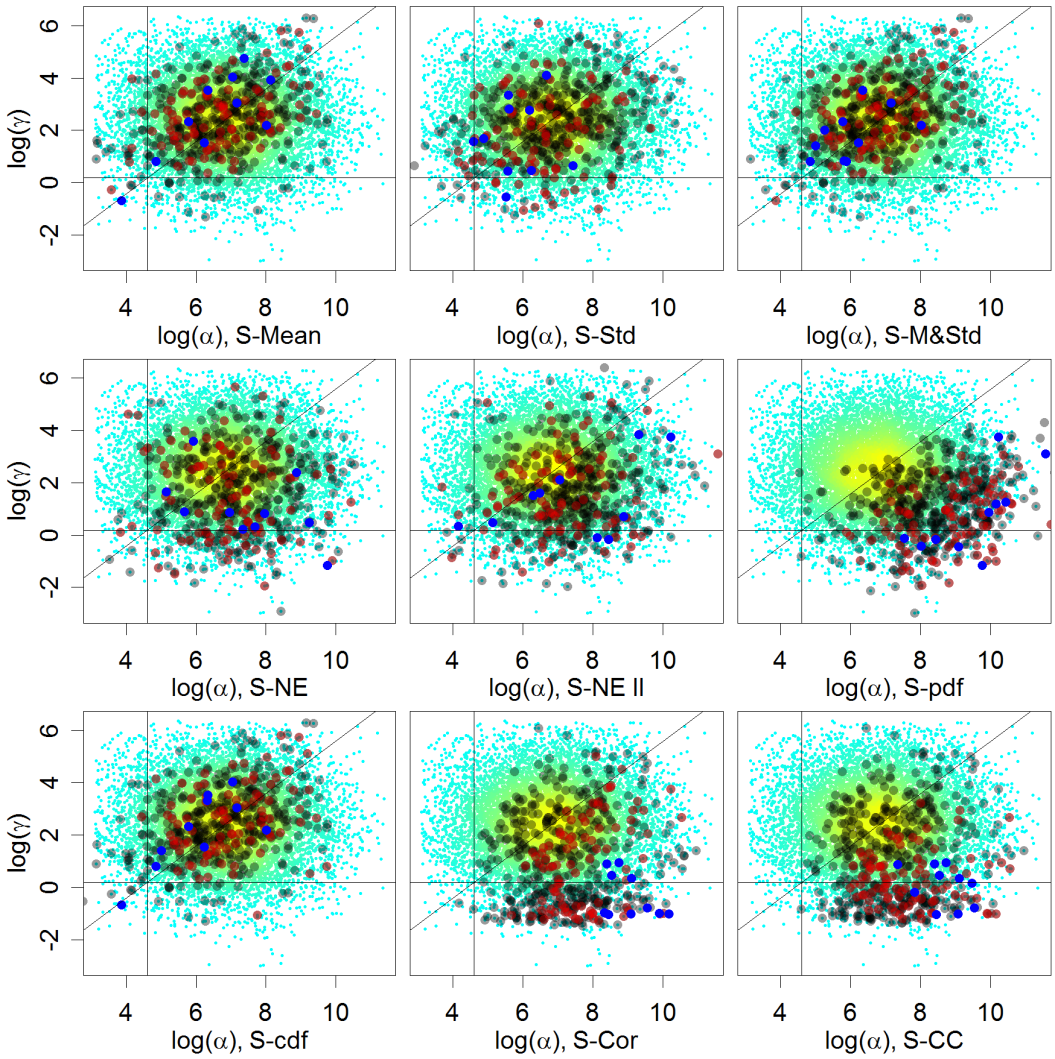


Figure 23: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the two-stage model (simulation 4) for α versus γ.

The distribution of the posteriors in figure 23 is similar to simulation 3 w.r.t. the large variance of the points. They are not centered around the true reaction rates. For $\tau = 0.001$, S-Cor and S-CC estimate γ well. The points of the posterior for S-Mean, S-M&Std and S-cdf are roughly along the ratio line.

For $\alpha\beta$ versus $\gamma\delta$, the distances S-Mean, S-M&Std and S-cdf estimate, as in

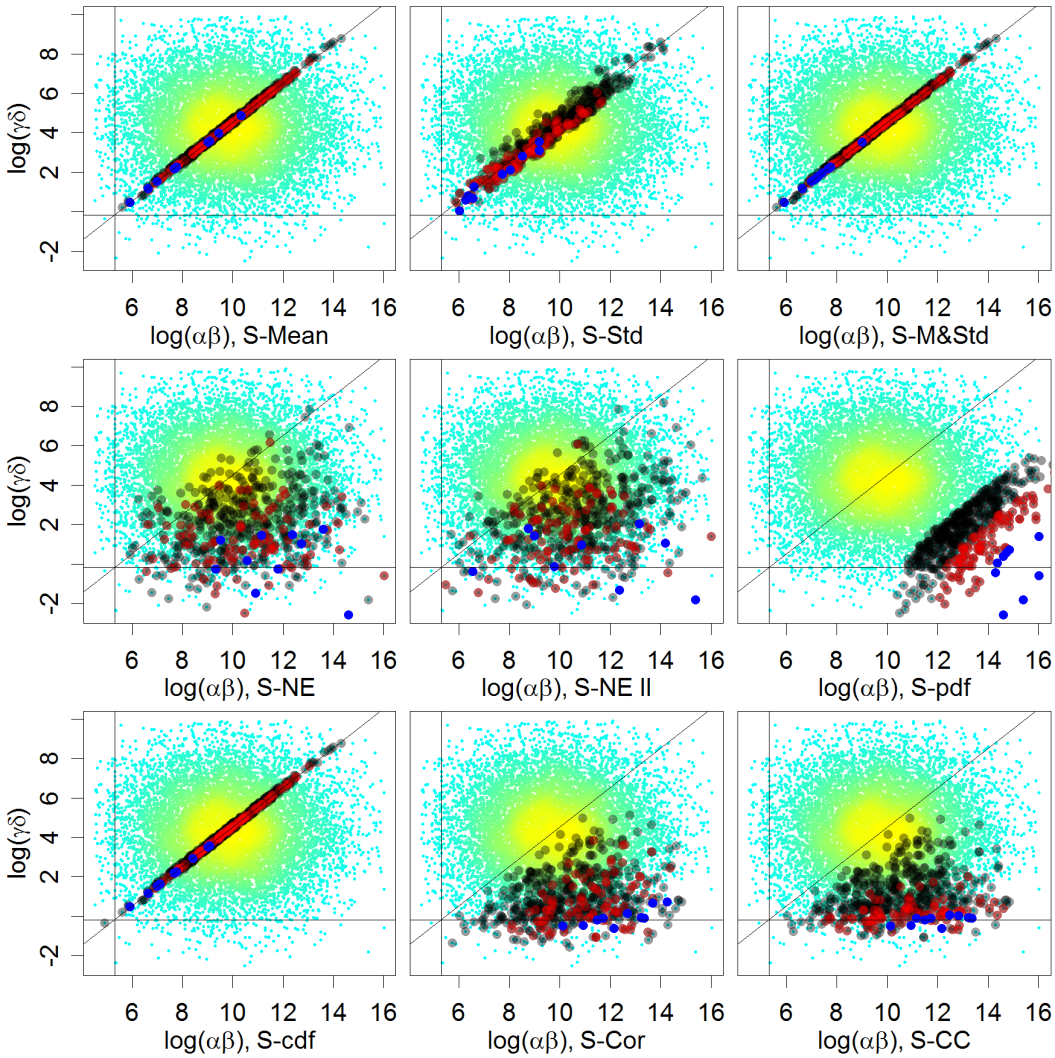


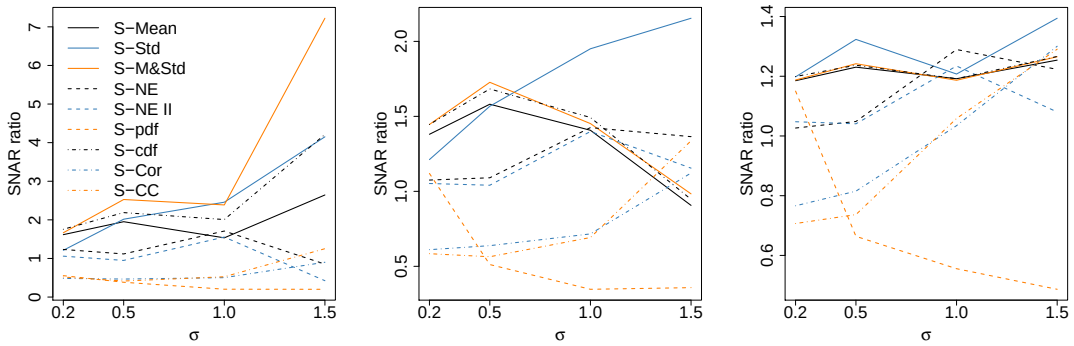
Figure 24: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the two-stage model (simulation 4) for $\alpha\beta$ vs. $\gamma\delta$.

simulation 3, the ratio very well, S-Std estimates it too, but its posterior has a larger variance. The posterior for S-pdf for simulation 3 was mainly on the ratio line. Now it is shifted right parallel to the ratio as in the one-stage model in simulation 2. S-Cor and S-CC estimate very accurate the value of $\gamma\delta$ for $\tau = 0.001$.

The SNAR ratio plotted against the acceptance rate does not provide new insights, therefore, it is not shown.

Comparing simulation 3 and 4 Figure 25 shows the SNAR ratio of all distances against σ_{LN} for different acceptance rates. S-Cor and S-CC, as well as S-NE and S-NE II, are very similar. For the one-stage model S-NE and S-NE II differed in their SNAR ratios for different σ_{LN}. The SNAR ratio of S-pdf decreases for a less informative prior.

The distances S-Mean, S-Std, S-M&Std and S-cdf together have a similar development of the SNAR ratio although S-Std deviates from the others for $\tau = 0.01$. For these distance functions the SNAR ratios are higher for a lower acceptance rate.



a) $\tau = 0.001$. b) $\tau = 0.01$. c) $\tau = 0.05$.

Figure 25: SNAR ratio depending on the parameter σ_{LN} of the prior distribution and the acceptance rates $\tau = 0.001$, $\tau = 0.01$ and $\tau = 0.05$ for the two-stage model.

Summary Summing up the simulations from the two-stage model results in the following:

- As in the one-stage model, the distances can be grouped, although the groups differ slightly. The first group consists again of S-Mean, S-Std, S-M&Std and S-cdf. They are able to estimate the ratio of the steady state of protein. The second group is S-pdf, S-Cor and S-CC where only S-pdf has a somewhat good estimation for $\sigma_{LN} = 0.2$. For $\sigma_{LN} = 1.5$, they all show the same behaviour, as their posterior is close to the true value for γ and δ but far from it for α and/or β. The third group being S-NE and S-NE II, which

estimate γ and δ quite well and for which the posterior of α and β is about the same as the prior. This means that S-NE and S-NE II do not improve nor worsen the estimation of these two parameters.

- The ratio α/γ, i.e. the mean number of mRNA at its steady state, is not estimated highly accurate by any distance. This is probably due to mRNA being not observed.

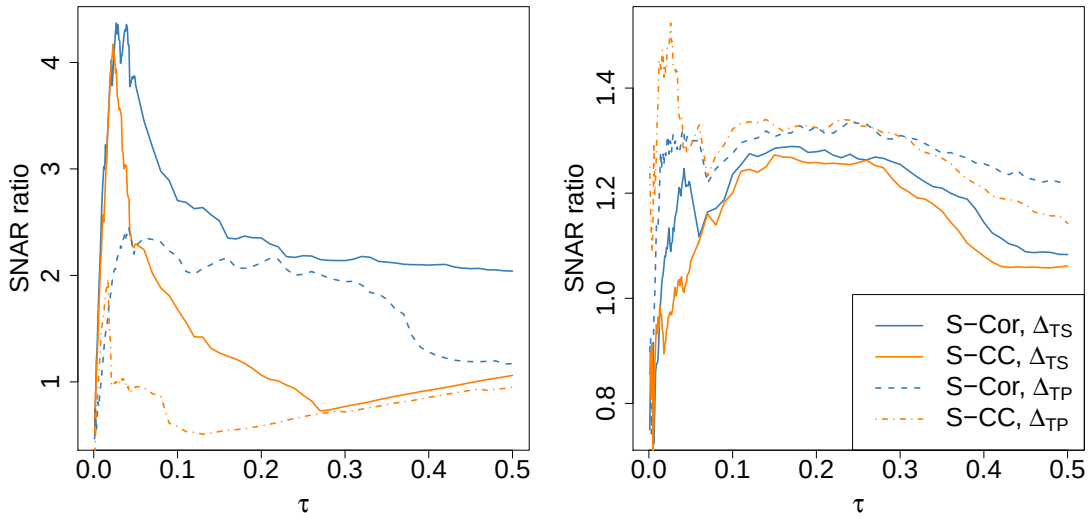
- As well as in the one-stage model a smaller τ results in a better estimation. The distance functions from group one estimate all kinetic rates better than the prior. The distance functions from group two and three only improve the estimation of γ and δ.

5.2 Simulation with optimal frequency for time series data

For simulations 1–4 the frequency was set according to the maximum of det(FIM) for TP data. This frequency was used to test all distances, including S-Cor and S-CC, which use TS data.

In this section the outcome for the one-stage and two-stage model for S-Cor and S-CC, using an optimal frequency based on TS data, is presented. As discussed in section 4.1.3, this frequency is $\Delta = 1.35$h and $\Delta = 0.84$h for the one- and two-stage model respectively. The other parameter settings for these simulations 5 and 6 can be seen in table 6 and are identical as simultions 2 and 4.

Figure 26 shows the result of the SNAR ratio against the acceptance rate τ for the frequency based on an optimum for TS and for TP data. For the one-stage model the frequency Δ_{TS} outperforms the other frequency clearly. Therefore, it is advisable for the one-stage model to take the frequency based on TS data if S-Cor or S-CC is used. For the two-stage model it is the other way round. The SNAR ratio with Δ_{TP} is better than with Δ_{TS}. But one has to take into account that the difference between the SNAR ratios for S-Cor with Δ_{TS} and Δ_{TP} is marginal and can be due to chance. Moreover, the SNAR ratio reflects the estimation of all parameters. It might be that γ and δ is estimated even more accurate using Δ_{TS} but α and β has a worse estimation. To check this, the SNAR values of the posterior distribution for $\tau = 0.001$ for each kinetic rate is presented in table 12.



a) one-stage model b) two-stage model

Figure 26: SNAR ratio depending on the acceptance rate τ for S-Cor and S-CC for the one- and two-stage model (simulation 5 and 6). The optimal frequency is either based on TS data (solid lines) or TP data (dotted lines).

$\text{SNAR}_{p(\theta)}$ of the prior distribution of each kinetic rate is about 28.29.

	S-Cor				S-CC			
rates	α	β	γ	δ	α	β	γ	δ
Δ_{TP}	101.32	21.18	0.67	2.49	68.27	20.6	0.6	0.94
Δ_{TS}	126.62	22.31	0.63	2.26	85.9	43.05	0.83	2.47

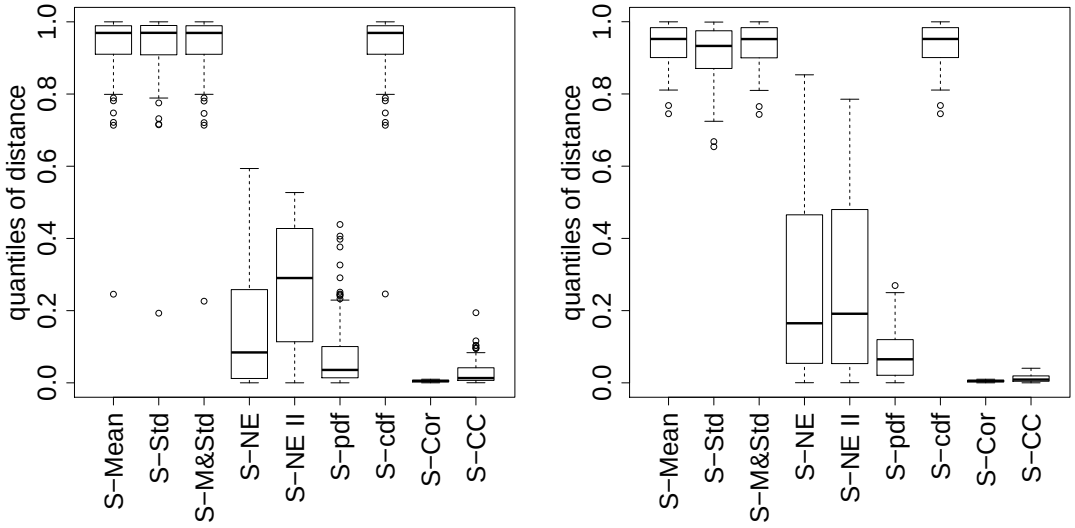
Table 12: $\text{SNAR}_{p(\theta|x)}$ of the posterior distributions for all kinetic rates for the two-stage model using frequency Δ_{TP} (simulation 4) and Δ_{TS} (simulation 6), $\tau = 0.001$.

For S-Cor merely the estimation of α is better using Δ_{TP}. For all other kinetic rates the SNAR value is approximately the same. For S-CC the SNAR value for all parameters with Δ_{TP} is smaller than with Δ_{TS}. For the two-stage model this result is in contrast to the theory that maximisation of det(FIM) results in a better estimation.

5.3 Combination of the distances S-Mean and S-Cor

Although S-Cor and S-CC use TS data, i.e. the data containing the most information, the estimation of the kinetic rates are not as accurate as using other distances. This is true even if an optimal frequency is used, which is discussed in chapter 5.2. One reason could be that the distance $d(x^0, x^*)$ is low for highly correlated data, although the mean of x^0 and x^* might be considerably different. In other words this means that although the trajectories of the simulated and experimental data might be very similar, the trajectories of x^0 and x^* might be somewhat parallel to each other.

To check this, the particles which are chosen for $\tau = 0.01$ from S-Cor and S-CC are taken. For each of these particles the quantile at other distances are calculated. The resulting boxplots are shown in figure 27. As the plots are very similar for S-Cor and S-CC, only S-Cor is presented.



a) one-stage model b) two-stage model

Figure 27: The quantiles of other distances for the best 1% of particles of S-Cor for the one-stage (simulation 2) and two-stage (simulation 4) model.

The boxplots underline the grouping of the distance functions as distance functions being in the same group have a similar boxplot.

The particles with the lowest values for S-Cor have the highest values for S-Mean. That means the data x^* which has a high correlation to x^0 usually has a

large difference of means between x^* and x^0. To adjust for this, a combination of S-Mean and S-Cor is considered. To combine both distance functions, they first need to be standardized because their range is different.

Each distance function $d(\cdot, \cdot)$ can be seen as a function from the set of kinetic rates θ to the set of resulting distances, defined by $d(x^0, x^*(\theta))$ with $x^*(\theta)$ indicating that the simulated data depends on the choice of the kinetic rates θ. In short, this is $\theta \mapsto d(x^0, x^*(\theta))$.

To form the new distance function for each $\theta^{(i,\cdot)}$ it is determined which percentile the resulting distance $d(x^0, x^*(\theta^{(i,\cdot)}))$ holds in S-Mean and S-Cor. This percentiles are then summed. The new distance S-MC is, therefore, defined as

$$d(\cdot, \cdot) = \sum_{i}^{N_{all}} \widehat{F_{Mean}}(\theta^{(i,\cdot)}) + \widehat{F_{Cor}}(\theta^{(i,\cdot)}) \tag{49}$$

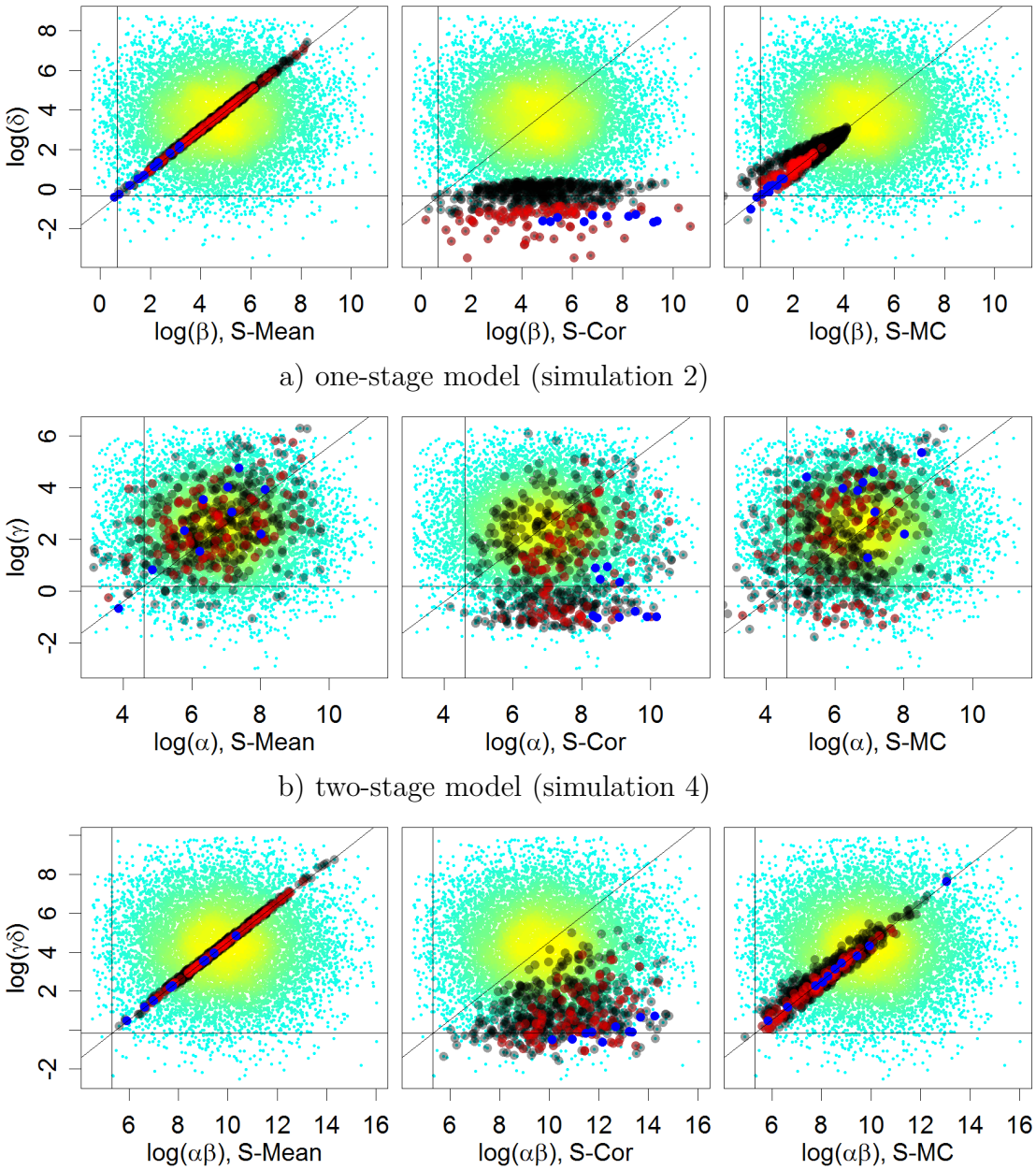
where $\widehat{F_{Mean}}$ and $\widehat{F_{Cor}}$ are the empirical cumulative distribution functions of the distances S-Mean and S-Cor.

To check if the resulting distance S-MC produces good estimations of the kinetic rates, their posterior distribution for different τ is plotted for the one-stage and two-stage model in figure 28.

The estimation of S-MC in the one-stage model, figure 28a), improves the posterior clearly. Both rates β and δ are well estimated, and the width of the posterior of β and δ is much smaller than with S-Mean. For the two-stage model for parameters α and γ, S-MC does not improve the estimation compared with S-Mean or S-Cor. For $\alpha\beta$ and $\gamma\delta$ in figure 28c) S-MC does not estimate $\gamma\delta$ as well as S-Cor, and the ratio is less accurate estimated as by S-Mean. But the width of the posterior for $\tau = 0.01$ and $\tau = 0.05$ is less compared with S-Mean.

Figure 29 shows the kernel estimation of the posterior for S-MC.

For the one-stage model the estimation of β is done best by S-MC compared with all other distance functions, see for comparison figure A.44. The estimation of δ is only better with S-Cor and S-CC. For the two-stage model the rates α and β are estimated well and only S-Std has such a good estimation, too. For comparison see figure A.56. The estimation of γ is severely worse than with S-Cor or S-CC, and the estimation of δ is about the same as with S-Cor and S-CC.



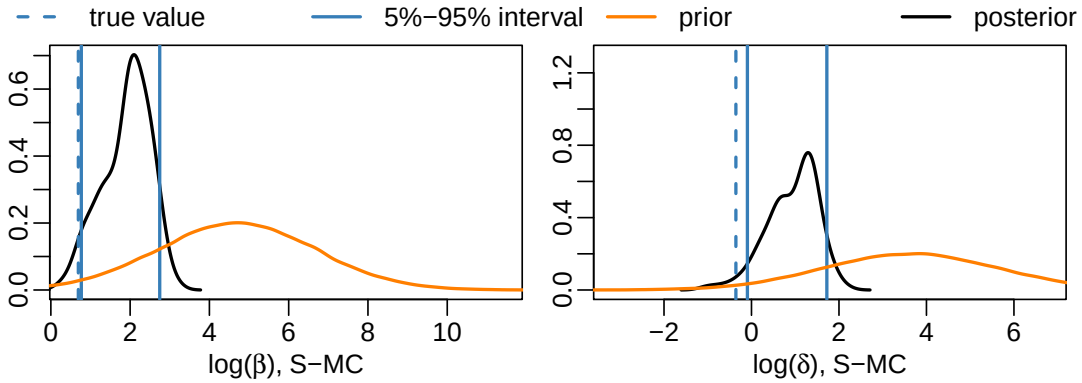
a) one-stage model (simulation 2)

b) two-stage model (simulation 4)

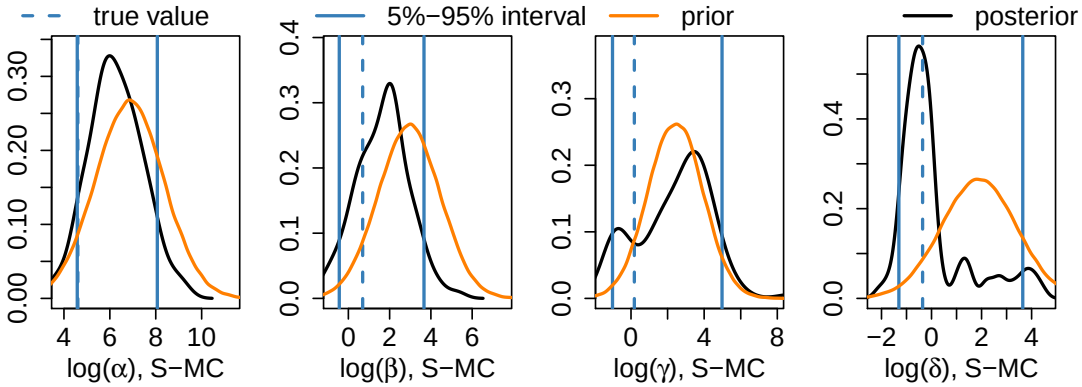
c) two-stage model (simulation 4)

Figure 28: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for the one-stage and two-stage model for S-MC.

In summary, the estimation with the distance S-MC is very accurate for the



a) one-stage model (simulation 2)



b) two-stage model (simulation 4)

Figure 29: Kernel density estimation of the prior and posterior distribution for S-MC for the one-stage and two-stage model, $\tau = 0.01$.

one-stage model. For the two-stage model the estimation of α, β and γ performs well compared with the other distance functions, but the estimation of γ is worse. Therefore, a separate estimation of γ with S-pdf, S-Cor or S-CC should be considered.

This section combined the distance functions S-Mean and S-Cor, using equal weights. Different combinations of distance functions with optimal weights need to be considered as well. This is described in the outlook, chapter 7.

5.4 Results for 100,000 drawn particles

This chapter presents the results for $N_{all} = 100,000$. Both for the one-stage and two-stage model a prior distribution with little informative was chosen.

5.4.1 Results for the one-stage model

In this section the results of simulation 7 and 8 are presented. The aim is to answer questions which have arisen during the first simulations. These are the following:

- Is the estimation of the reaction rates more accurate for a higher number of drawn particles?
- Does the SNAR ratio or the threshold ϵ only depend on τ but not on the number of drawn particles?

To answer the first question for the one-stage model, the SNAR statistics are presented in table 13.

distance	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC	MC
$\text{SNAR}_{p(\theta)}$	806.23	806.23	806.23	806.23	806.23	806.23	806.23	806.23	806.23	806.23
$\tau = 0.0001$										
$\text{SNAR}_{p(\theta\|x)}$	0.44	1.43	0.35	424.87	19.28	24616	0.37	8883	2434	0.6
ratio	1842	562.89	2296	1.9	41.83	0.03	2198	0.09	0.33	1345
$\tau = 0.001$										
$\text{SNAR}_{p(\theta\|x)}$	4.42	13.16	3.68	438.28	82.77	24996	3.57	3972	3344	1.46
ratio	182.25	61.26	219.2	1.84	9.74	0.03	225.73	0.2	0.24	554.01
$\tau = 0.005$										
$\text{SNAR}_{p(\theta\|x)}$	108.9	175.23	127.11	954.41	221.01	11451	96.54	2733	1665	4.12
ratio	7.4	4.6	6.34	0.84	3.65	0.07	8.35	0.3	0.48	195.73

Table 13: SNAR statistics for simulation 7 for acceptance rates $\tau = 0.0001, \tau = 0.001$ and $\tau = 0.005$.

The acceptance rates are chosen so that the number of accepted particles is 10, 100 and 500 which are the same number of accepted particles as with $N_{all} = 10,000$. The overall picture is similar as for simulation 2 with $N_{all} = 10,000$. The first group of distance functions, S-Mean, S-Std, S-M&Std and S-cdf, result in a very good estimation. The second group of distance functions have a poor overall estimation quality and S-NE II as third group does a modest job. The best

estimation is with S-M&Std for $\tau = 0.0001$. The best estimation along different acceptance rates is with S-MC which only performs third best for $\tau = 0.0001$.

These results can be seen in figure 30 in more detail.[16] Especially the good estimation with S-MC is illustrated, as for different acceptance rates the posterior is quite compact around the true values of β and δ. For the second group of distance functions, the behaviour of estimating δ well but not being able to estimate β is visible as well.

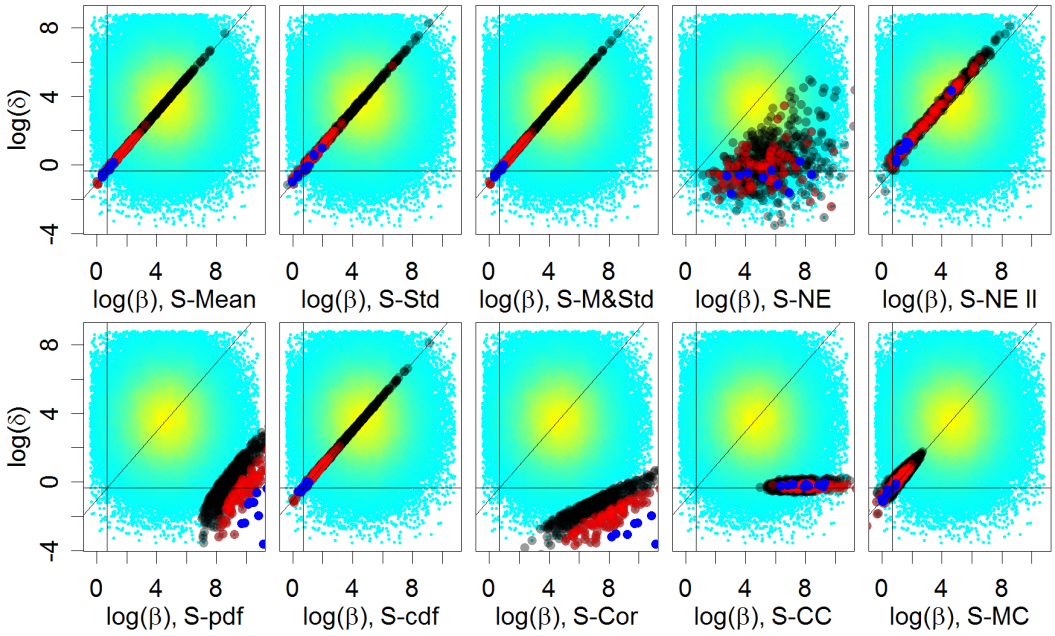
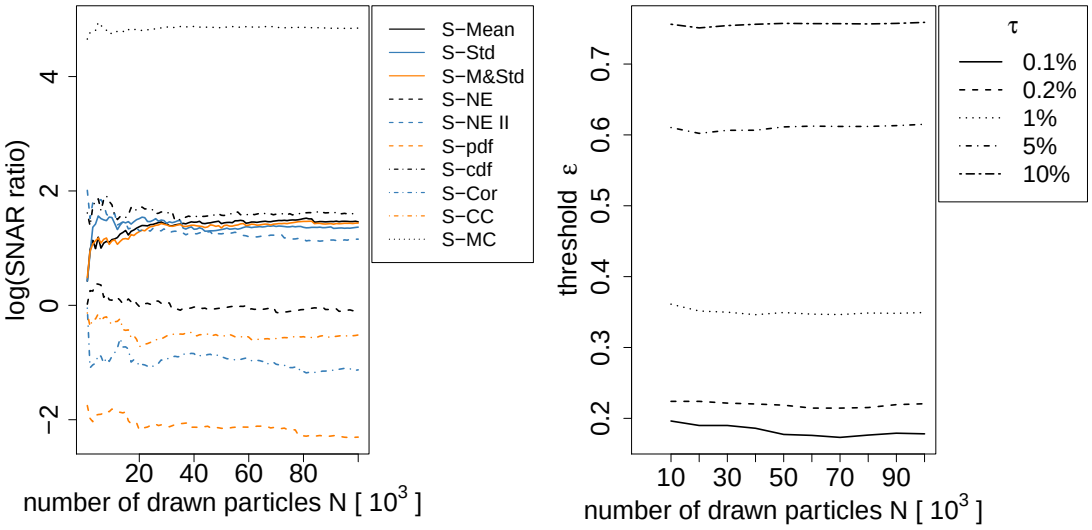


Figure 30: Prior distribution of the kinetic rates (cyan,yellow) and the resulting posterior for $\tau = 0.005$ (black), $\tau = 0.001$ (red) and $\tau = 0.0001$ (blue) for the one-stage model (simulation 7).

Compared with simulation 2, the possibility to have a lower acceptance rate, as more particles have been drawn, results in a better estimation of β and δ by the first group of distance functions. As more particles have been drawn, there are more particles close to the true kinetic rates in the prior distribution. So increasing the number of drawn particles results in a better estimation of the kinetic rates if τ is decreased accordingly.

[16] For better visualization one percent of the data with highest β and δ values have not been plotted.

To answer the second question, figure 31 shows both the SNAR ratio and the threshold ϵ against the number of drawn particles N_{all}.



a) SNAR ratio against N_{all} for $\tau = 0.01$. b) ϵ against number of drawn particles N_{all} for S-MC.

Figure 31: SNAR ratio and threshold ϵ against number of drawn particles N_{all} for the one-stage model, simulation 7.

The SNAR ratio is not constant, even for a large number of drawn particles it still fluctuates slightly. In figure A.47 for $N_{all} = 10,000$ it seemed to converge. But the SNAR ratios in figure 31b) for $N_{all} = 10,000$ fluctuate considerably for larger N_{all}.

Differently, the threshold ϵ stays constant, and it does not depend on the number of drawn particles. The following consideration gives an explanation for both observations. The resulting distances $d(x^0, x^*(\theta^{(i,\cdot)})$ for all particles $\theta^{(i,\cdot)}$ are distributed according to F_D. Drawing a finite number of particles means we approximate the distribution F_D. The threshold ϵ for a certain τ equals, therefore, the $\tau \cdot 100$-percentile of F_D. Drawing a large enough number of particles N_{all}, the approximation of F_D becomes stable and, therefore, a certain percentile has approximately the same value as for another N_{all}. This value is the threshold ϵ. But the particles below this threshold do not seem to form the same posterior distribution. One reason might be that the SSA produces similar x^* for two particles,

although these particles are quite different. Due to the stochastic character of the SSA this is possible.

To sum it up, a larger N_{all} results in a better estimation of the reaction rates as a smaller τ can be applied. But the behaviour of the distance functions stays similar, for instance the second group of distance functions still only estimates δ well.

5.4.2 Results for the two-stage model

The SNAR statistics are presented in table 14. Compared to simulation 4 with

distance	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC	MC
$\text{SNAR}_{p(\theta)}$	113.06	113.06	113.06	113.06	113.06	113.06	113.06	113.06	113.06	113.06
$\tau = 0.0001$										
$\text{SNAR}_{p(\theta\|x)}$	84.75	47.81	64.06	113.08	59.71	1299.81	68.49	189.56	179.34	552.28
ratio	1.33	2.36	1.76	1	1.89	0.09	1.65	0.6	0.63	0.2
$\tau = 0.001$										
$\text{SNAR}_{p(\theta\|x)}$	50.2	51.75	39.21	78.16	96.71	753.83	45.69	174.39	120.8	105.95
ratio	2.25	2.18	2.88	1.45	1.17	0.15	2.47	0.65	0.94	1.07
$\tau = 0.005$										
$\text{SNAR}_{p(\theta\|x)}$	69.28	51.52	55.23	86.44	90.01	459.1	65.4	136.79	109.05	66.88
ratio	1.63	2.19	2.05	1.31	1.26	0.25	1.73	0.83	1.04	1.69

Table 14: SNAR statistics for simulation 8 for acceptance rates $\tau = 0.0001, \tau = 0.001$ and $\tau = 0.005$.

$N_{all} = 10,000$, the SNAR statistics do not indicate a better estimation of the kinetic rates. This is underlined by the posterior distributions, which are shown as a scatterplot in figure A.57 and A.58 in the appendix.

In figure 32 the kernel estimation for 100 accepted particles, i.e. $\tau = 0.001$ for simulation 8, are presented for S-M&Std, S-NE II and S-CC. Compared to simulation 4 with 100 accepted particles, i.e. $\tau = 0.01$, S-NE II estimates α similarly, β is estimated further from the true value, and γ and δ are estimated slightly better, thus resulting overall in a similar SNAR ratio (1.17 for simulation 8 versus 1.16 for simulation 4). For S-CC the rate α is estimated worse, β and γ are estimated about the same, and δ is estimated slightly better. For S-M&Std all kinetic rates are estimated better, especially δ. One observation is that the estimation of γ and δ is becoming better with more particles being drawn. An

explanation for this is given below.

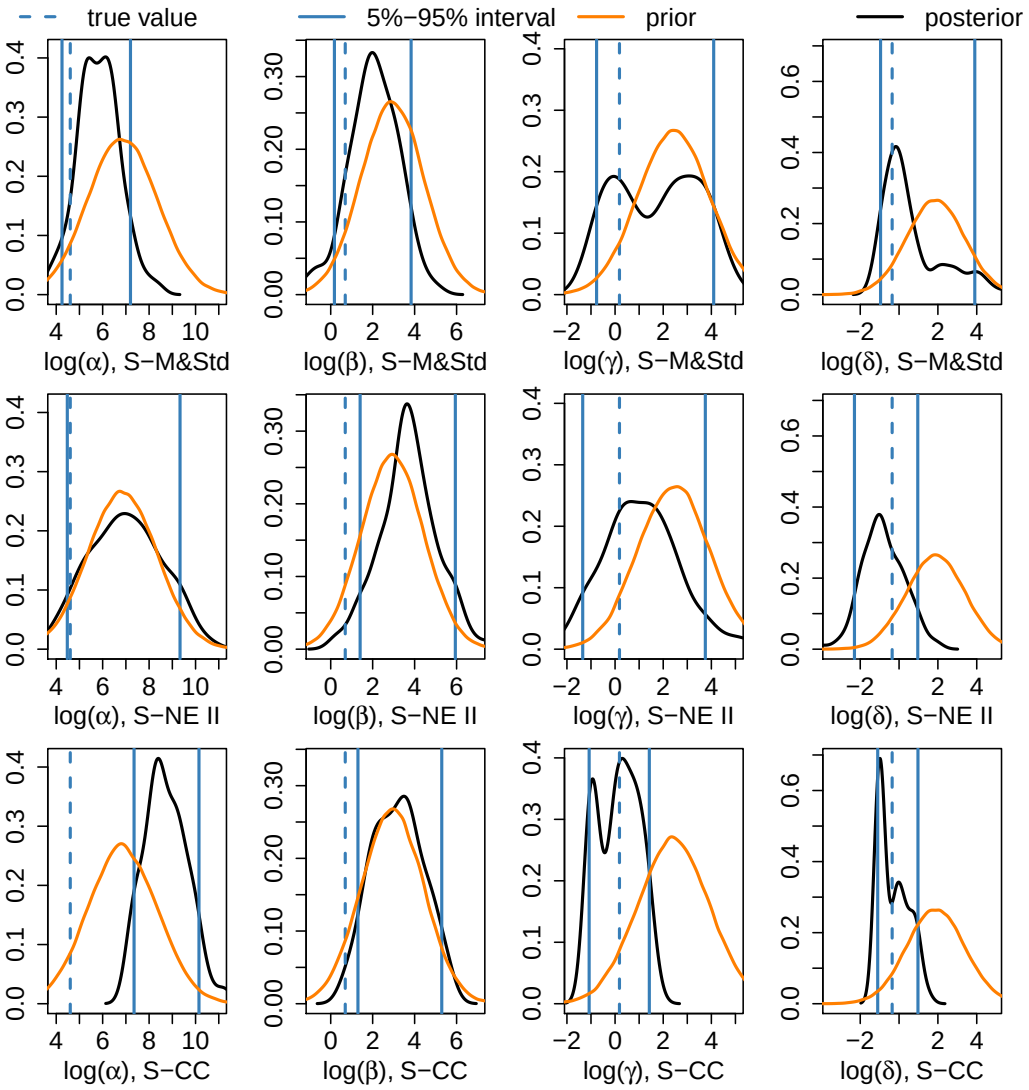


Figure 32: Kernel density estimation of the prior and posterior distribution for S-M&Std, S-NE II and S-CC for the two-stage model (simulation 8), $\tau = 0.001$.

The observation that the estimation for all kinetic rates is not becoming better, although more particles are drawn and a smaller τ is applied, was not expected. A reason might be that due to four rates which have to be estimated there are still not enough numbers of drawn particles which are close to the true values for all reaction rates. Therefore, the particles with the smallest distance might not

be close to the true values and thus even a worse estimation than with a smaller amount of drawn particles result.

To check this, the number of particles from the prior are counted which have an average SNAR ratio less than one. That means for all four kinetic rates the SNAR ratio on average is less than one. For simulation 4 with $N_{all} = 10,000$ there are 26 particles, for simulation 8 with $N_{all} = 100,000$ there are 252 particles. Therefore, there are enough particles which are close to the true kinetic rates.

Another reason might be that the overall distance is mainly dependent on γ or δ. As the estimation of only γ or δ has becoming better for a larger N_{all} this consideration comes up. To control this consideration, a barplot is shown in figure 33. There is one barplot for each kinetic rate α to δ. Each barplot only considers the particles which have a SNAR value of less than one for the specific kinetic rate. The composition of the barplot is explained by means of an example. The barplots for S-Mean and α considers all particles with a SNAR value for α less than one at the distance function S-Mean. From these particles the levels of the barplot indicate the number of particles which have additionally a SNAR value less than one for β, γ or δ. For instance, there are about 5 particles which have an α and β with a SNAR value less than one (the blue shaded part of the barplot). Consider that it might be that these 5 particles have also a SNAR value less than one for γ and/or δ. This information is not contained in the graphic.

The part of the barplot with the same kinetic rate as the barplot counts the number of particles where only this specific kinetic rate has a SNAR value less than one and all other rates have a SNAR value greater one. For instance taking the γ-barplot for S-Mean, the green shaded area means, that there are about 12 particles where only γ has a SNAR value less than one. The SNAR values for α, β and δ for these particles are greater than one.

The barplot can have a total number of more than 100, as it is possible that particles are counted more than once, if three or four kinetic rates in one particle have a SNAR value less than one.

For all distance functions the distance is mainly determined by the value of δ. This can be seen as the purple part of each δ-barplot[17] is large for all distance functions. This means that for these particles the value of δ, being close to the

[17]this counts the particles where only the SNAR value of δ alone is smaller one.

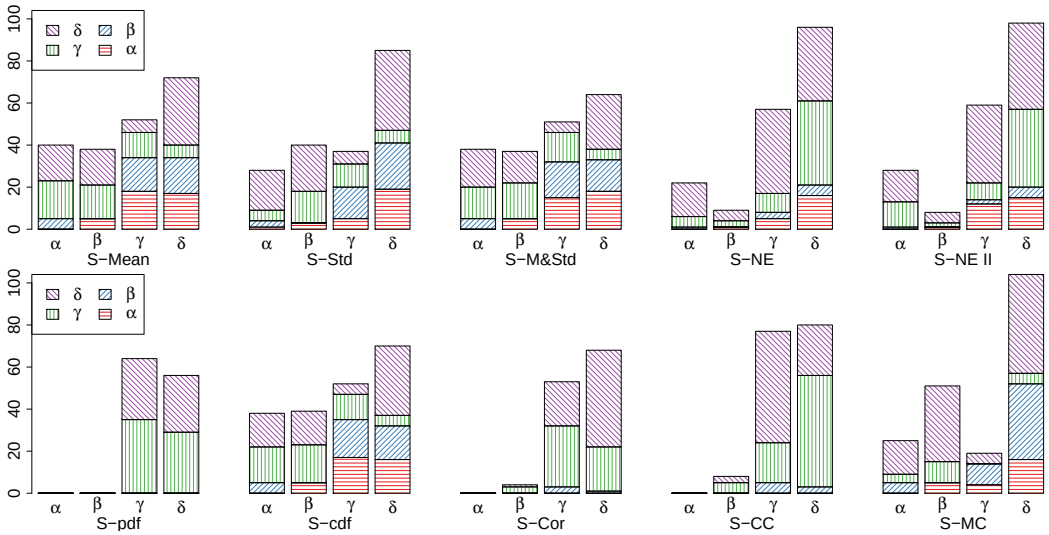


Figure 33: Barplots for simulation 8 for the best 100 particles ($\tau = 0.001$) which have a SNAR value for each particle less than one.

true value, is important for the total distance. In other words, if one particle with a good estimation of δ is drawn, the resulting total distance is in most cases small with the values of α, β or γ having no strong influence. A similar, but weaker effect exists for γ. These effects have a particular strong intensity for S-pdf, S-Cor and S-CC, which represent the second group of distance functions. In this group there are almost none SNAR(α) or SNAR(β) values less than one. This gives another insight for the observation that these distance functions only estimate γ and δ well.

Highly remarkable is the fact that for all particles either SNAR(γ) or SNAR(δ) has a value less than one. From the figure it can be seen that there are no particles with only SNAR(α) or only SNAR(β) being less than one. An additional calculation showed that there is no particle which has a SNAR value of both α or β less than one with neither SNAR(γ) or SNAR(δ) being less than one.

The reason for this observation lies probably in the fact that only protein was observed.

The SNAR ratio and the threshold against the number of drawn particles are shown in figure A.59 in the appendix as the result is similar to the result for the one-stage model.

In short, the result for $N_{all} = 100,000$ show that a larger amount of drawn particles increases the quality of estimation as a lower acceptance rate can be taken. The typical behaviour of each distance function does not change, i.e. a distance function which could only estimate certain kinetic rates for $N_{all} = 10,000$ estimates these rates better for $N_{all} = 100,000$, but it cannot estimate other reaction rates for a larger amount of N_{all}.

For the two-stage model the overall distance is determined strongly by the estimation of δ and γ. A particle which has a good estimation for δ and γ is probable to have a low distance although the estimation of α and β might not be very accurate.

Therefore, another approach is tried, which first estimates γ and δ and based on these results α and β. This is presented in the following chapter.

5.5 Estimation of the kinetic rates in two steps for the two-stage model

The results showed that the distance functions from the second group estimated γ and δ well in the two-stage model, but could not estimate α or β. In the following, first γ and δ are estimated. Their posterior is used as their prior distribution in a new simulation. The new simulation aims to estimate α and β.

For this simulation, the best 10 particles from S-Cor from simulation 4 have been taken to form the new prior for γ and δ. For α and β the same prior as in simulation 4 was used, i.e. a log-normal distribution with $\sigma_{LN} = 1.5$. This simulation is named simulation 4*.

From simulation 4, γ and δ have been well estimated by S-Cor, with a $\text{SNAR}_{p(\theta|x)}$ of 0.67 and 2.49 for the single kinetic rates respectively. The estimations for α and β have not been that accurate for any distance function, though.

Figure 34 shows the posterior distributions for α and β from simulation 4*.

For the first group of distance functions, the estimation of the two kinetic rates is very accurate. Additionally, only S-MC estimates β well. All other distance functions are not able to estimate α and β.

The question comes up if the estimation of γ and δ became worse in comparison to the results from simulation 4. To check this the $\text{SNAR}_{p(\theta|x)}$ values of all four

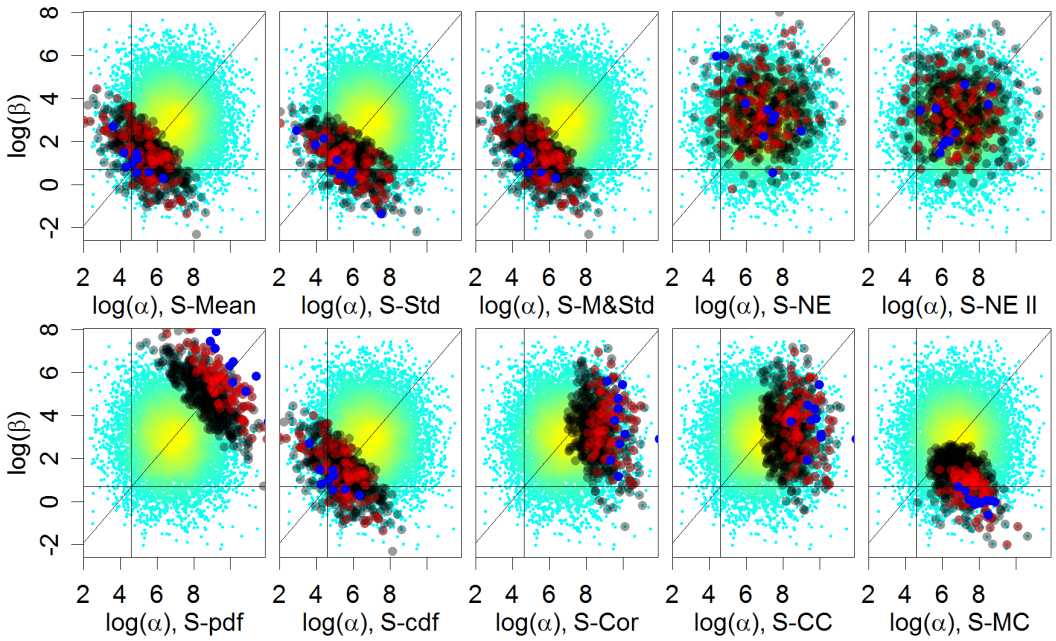


Figure 34: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.05$ (black), $\tau = 0.01$ (red) and $\tau = 0.001$ (blue) for α and β for the two-stage model (simulation 4*).

kinetic rates were calculated for simulation 4 and 4*. The result is presented in table 15.

	S-Mean		S-M&Std		S-Cor		S-MC	
	sim 4	sim 4*	sim 4	sim 4*	sim 4	sim 4*	sim 4	sim 4*
α	11.19	0.87	3.74	0.82	101.32	307.25	15.43	33.11
β	4.41	1.10	4.24	0.64	21.18	39.41	12.06	0.44
γ	24.53	0.58	10.56	0.58	0.67	0.68	550.18	0.74
δ	2.76	4.41	8.74	4.51	2.49	2.03	0.46	2.93
$\text{SNAR}_{p(\theta\|x)}$	42.83	6.96	27.28	6.55	125.66	349.37	578.12	37.22

Table 15: $\text{SNAR}_{p(\theta|x)}$ of the posterior distributions for all kinetic rates for the two-stage model for simulation 4 and 4* for $\tau = 0.001$.

For the distance functions presented in table 15, the estimation for γ is becoming better or stays the same. The estimation for δ is approximately the same for S-Cor. For S-Mean and S-MC it becomes worse, for S-M&Std it gets better.

To sum it up, the best estimation for all kinetic rates consists of two steps. First estimate γ and δ with a distance function from the second group using a non-informative prior for all kinetic rates. Then run the simulation again, taking as prior for γ and δ the posterior of the first simulation.

As final estimations take α, β and γ from the second simulation from a distance function of the first group. The rate δ is best estimated by S-Cor or S-CC, whereby the first and second simulation result in approximately the same estimation.

5.6 Summary

The simulation with the ABC algorithm yielded the following results.

- No single distance function was best for both the one-stage and two-stage model. In the one-stage model the distance functions from the first group, i.e. S-Mean, S-Std, S-M&Std and S-cdf, yielded, additional to S-MC, the best estimation of the kinetic rates. In the two-stage model the distance functions from the first group performed best.

- Increasing the number of drawn particles N_{all} only resulted in a better estimation for the one-stage model. For the two-stage model the SNAR ratio did not improve. This is because in the two-stage model, with only protein being observed, the reaction rates γ and δ are estimated very accurate. If one particle has a value for γ or δ close to the true reaction rates, the particle has a low distance nearly regardless of the values of α and β. Therefore, increasing N_{all} results in a better estimation of γ and δ. But the estimation of α and β is not necessarily improving as particles have a low distance due to a good estimation of γ and δ.

- S-Cor and S-CC, the distances using TS data, do not have the best estimation for all kinetic rates although they use the data with the most information. This holds even if an optimal sampling frequency based on TS data is used. However, the estimation of γ and δ in the two-stage model is very accurate with these distance functions.

- The particles which yielded a low distance with S-Cor and S-CC had a high distance with distance functions from the first group. To adjust for this,

a new distance function S-MC was introduced which corrected the values of S-Cor with S-Mean to consider the difference of the mean of the experimental and simulated data. For the one-stage model this resulted in a very accurate estimation, for the two-stage model, the estimation of α and β and δ improved at the expense of the estimation of γ.

- As there is no best distance function which estimates all kinetic rates of the two-stage model, one possible way of parameter estimation might be the following. Estimate γ and δ with distance functions from the second group, i.e. S-pdf, S-Cor or S-CC or the third group, i.e. S-NE or S-NE II. Use the resulting posterior distribution as prior for γ and δ for another ABC run where α and β are estimated using distance functions from the first group. The estimation from this approach is very accurate. It is discussed in chapter 5.5.

6 Computational study for ABC SMC

For the ABC SMC only certain distances were chosen due to the computational effort for creating the simulated data being otherwise too high. This could be the case if the kinetic rates would move further from the true values with each population. Pre simulations showed that the simulation by the SSA is time consuming if the kinetic rates are large.

Therefore, no distances were taken, which produced for one kinetic rate a posterior worse than the prior in the ABC. This includes the members of the second group S-NE, S-pdf, S-Cor and S-CC for the one-stage model, as the estimation of β is becoming worse. For the two-stage model, with the same argumentation, this is the second group as well, including S-pdf, S-Cor and S-CC, as the estimation of α and β is worse compared to the prior.

For the two-stage model, σ_{LN} was set to 1, as pre simulations with $\sigma_{LN} = 1.5$ showed that for some distance functions the posterior in later populations became worse resulting in a highly large run time for creation of the simulated data.

Thus for the one-stage model S-M&Std was chosen as representative of the first group and S-cdf, also from the first group, but with the difference that it is not based on moments. Additionally, S-NE II was chosen as having the best estimation for $\sigma_{LN} = 2$.

For the two-stage model S-M&Std and S-cdf were chosen for the same reasons as with the one-stage model. From the third group S-NE was selected as having a slightly better performance than S-NE II.

For both models S-MC was selected as well because of its convincing performance at the ABC and because it uses TS data.

For each simulation the ABC SMC was conducted with 5 populations. For each population 2,000 particles are drawn and the best 10 particles are accepted to build the posterior distribution. The number of drawn particles is the same as in simulations 1–4 from the ABC algorithm. Therefore, the results from ABC and ABC SMC can be compared.

6.1 Results for the one-stage model

Simulation 9: $\sigma_{\mathbf{LN}} = 0.2$, uniform perturbation kernel In figure 35 the progress of the SNAR ratio for each population is shown. The SNAR ratio is always calculated with $\text{SNAR}_{p(\theta|x)}$ from the current population and $\text{SNAR}_{p(\theta)}$ from the prior distribution of the first population.

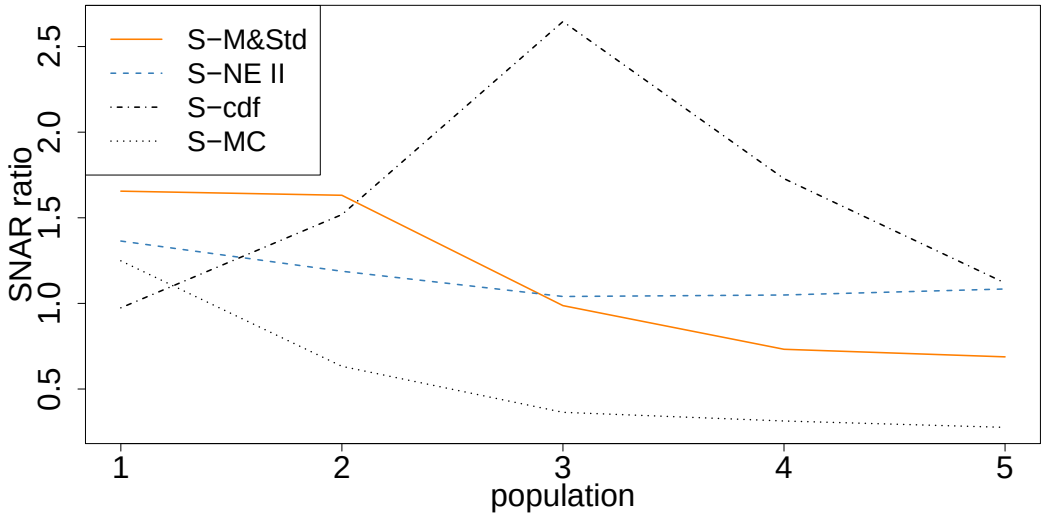


Figure 35: SNAR ratio for each population for ABC SMC for the one-stage model (simulation 9).

Comparing the SNAR ratio from population 5 with population 1, only for S-cdf the SNAR ratio increased slightly. For all other distance functions the SNAR ratio decreases. Therefore, for the chosen settings, the ABC SMC for an informative prior does not yield better results than for the ABC algorithm.

The result for the first population of the ABC SMC is theoretically the same as the result of the ABC algorithm. Comparing the result with simulation 1, though, shows differences. For simulation 9 all SNAR ratios are less than the SNAR ratios which result when one takes the first 2,000 particles from simulation 1 and applies a $\tau = 0.005$. These are 2.31, 1.36 and 2.56 for S-M&Std, S-NE II and S-cdf respectively.

There are two possible reasons for this. First, due to chance, the 2,000 particles drawn for the first population of the ABC SMC are different as for the ABC. This

can have an influence, but out of these 2,000 particles there surely are 10 particles which are close to the true kinetic rates.

The other reason is that the resulting trajectories are simulated stochastically by the SSA. Thus, the distances from similar particles can be different. To check this, two simulations have been conducted with parameter settings as in simulation 1 with the difference that $N_{all} = 2,000$. The drawn values for α and β are the same for both simulations. The first simulation achieved a SNAR ratio of 1.56, 0.64, 1.7, 0.96 for S-M&Std, S-NE II, S-cdf and S-MC respectively. The second simulation achieved a SNAR ratio of 2.56, 0.99, 2.07 and 0.9. For both simulations τ was set to 0.005.

The SNAR ratios for the two simulations are different, although the same values for α and β have been chosen for the simulations. This indicates that there is an influence by the SSA on the resulting distance, as the simulated trajectories differ for the same reaction rates.

In figure 36 the prior distribution and posterior distributions for each population is shown. As there are less particles drawn for the prior distribution (2,000 instead of 10,000) the prior distribution is not as dense as in the simulations of the ABC. The first posterior distribution is in black, the second in violet and the third, fourth and fifth in orange, red and blue.

The posterior for S-M&Std always stays on the ratio of β and δ. However, it moves further from the true rate values. For S-NE II the particles chosen for the posterior distribution seem to be chosen randomly, as no clear direction can be identified. For S-cdf only the first posterior is not centered along the ratio δ/β. The other posterior distributions seem to move further from the true rate values with increasing number of populations. For S-MC the posterior stays approximately along the ratio but moves away from the true values.

Simulation 10: $\sigma_{\mathbf{LN}} = 2$, uniform perturbation kernel For a less informative prior, the estimation improves with the ABC SMC for the analysed distance functions.

The SNAR ratio plotted against the number of populations is in the appendix, figure A.60. The good estimation of the kinetic rates with the last posterior distribution can be seen in figure 37.

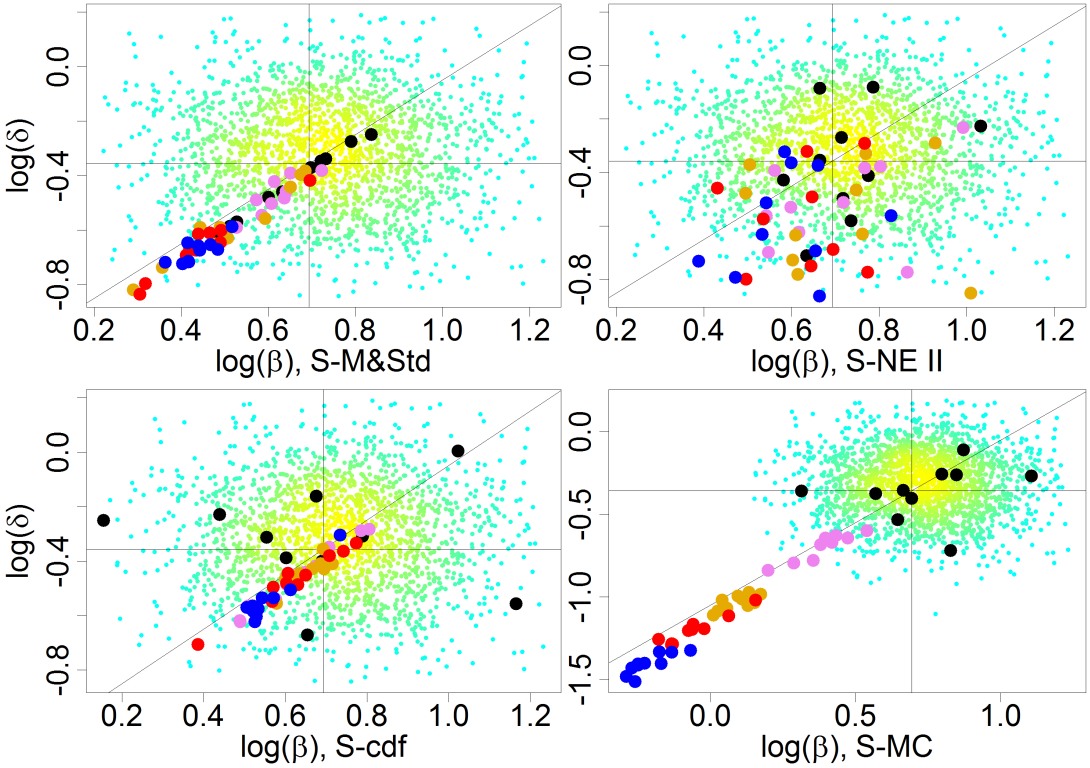


Figure 36: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the one-stage model (simulation 9).

For S-M&Std and S-cdf the third population (green) has a posterior very close to the true kinetic rates. For the fourth and fifth population the good estimation remains. This means that the posterior does not move away from the true reaction rates, as it is the case with S-MC. There, the estimation of the kinetic rates is along the ratio δ/β, but it moves further from the true kinetic rates up to the fifth population. This is probably due to the influence of the distance which is based on the correlation, as S-MC is a combination of S-Mean and S-Cor. As in the results of the ABC, the estimation of the kinetic rates is best with S-MC for the first population.

S-NE II estimates β very good, but underestimates δ slightly. This underestimation seems to stay constant for populations three to five.

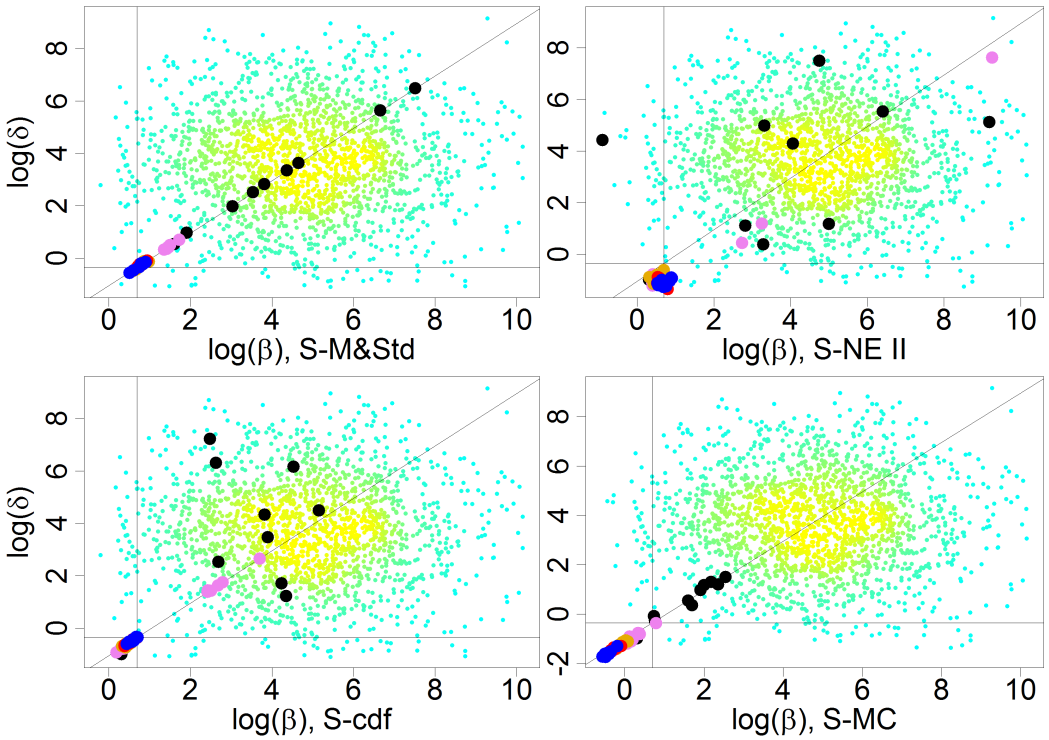


Figure 37: Prior of the first population (cyan, yellow), and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the one-stage model (simulation 10).

Simulation 11: $\sigma_{\mathbf{LN}} = 2$, gaussian perturbation kernel Now a Gaussian perturbation kernel is taken. For this perturbation kernel, no simulation with $\sigma_{LN} = 0.2$ was conducted, as the simulation with a uniform perturbation kernel did not show an improvement in estimation.

The development of the posterior distributions along the populations is shown in figure 38.

S-M&Std, S-cdf and S-MC estimate all along the ratio of δ/β. Compared with simulation 10, the posterior is still too far from the true kinetic rates. With each population, it moves only slowly towards the true rates. This is due to the chosen variance of the gaussian perturbation kernel. 95% of the perturbed particles are within $[0.75\theta^{i,r}, 1.25\theta^{i,r}]$, with the modus of the perturbed particles being at $\theta^{i,r}$. Thus, in most cases the particle $\theta^{i,r}$ is only perturbed slightly. Therefore, each posterior only differs slightly from the previous one.

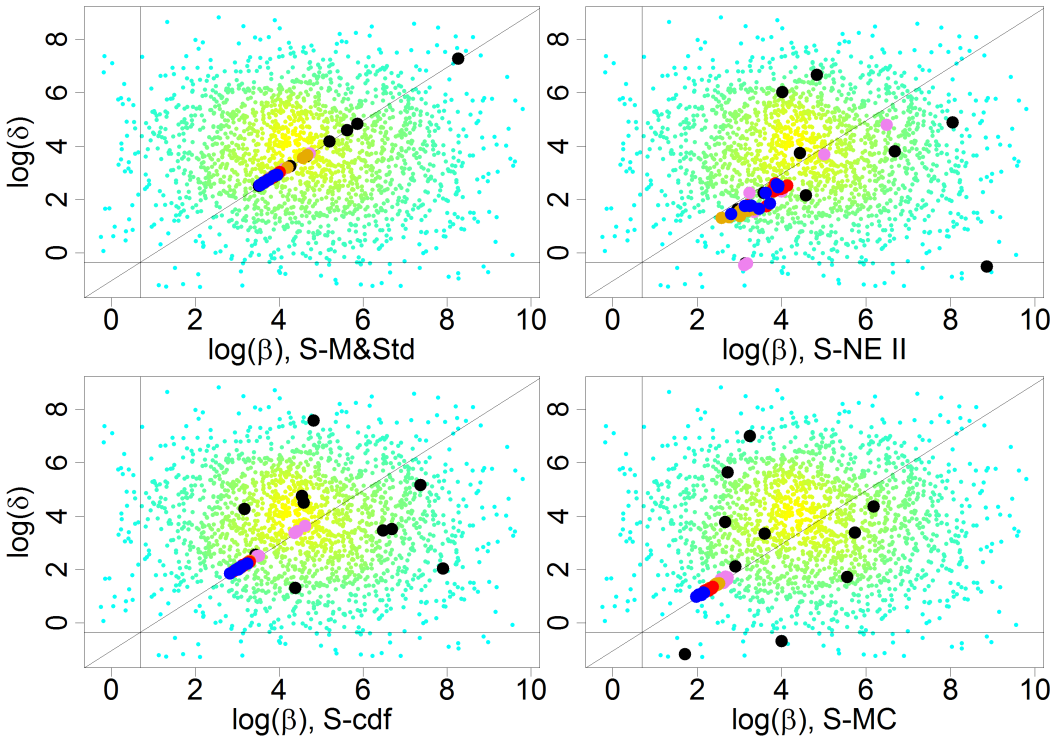


Figure 38: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the one-stage model (simulation 11).

To gain a better estimation, the number of populations or the variance of the gaussian perturbation kernel can be increased.

Summary and comparison to ABC For both the ABC and ABC SMC 10,000 particles have been drawn in total. In the following the results are compared.

- For an informative prior distribution, the estimation with the ABC SMC is not better compared with the ABC. For some distance functions it is even worse.
- For a less informative prior the estimation from the ABC SMC with a uniform perturbation kernel is closer to the true reaction rates than the estimation from the ABC. This is true, except for S-MC. There, the posterior moves away from the true reaction rates with each population.

- In summary, for the one-stage model a distance function from the first group (S-M&Std, S-Std, S-Mean, S-cdf) should be taken within an ABC SMC.

6.2 Results for the two-stage mode

Simulation 12: $\sigma_{\mathbf{LN}} = 0.2$, uniform perturbation kernel For an informative prior, similar to the one-stage model, there is no improvement of the estimation along the populations. The figures are, therefore, shown and shortly commented in the appendix. The posterior distributions for the kinetic rates α and γ is presented in figure A.61 and figure A.62 shows the posterior distributions for $\alpha\beta$ versus $\gamma\delta$.

Simulation 13: $\sigma_{\mathbf{LN}} = 1.5$, uniform perturbation kernel Preliminary simulations showed that the estimation is considerably becoming worse for S-NE, which resulted in a long runtime for the creation of the simulated data by the SSA. Therefore, the simulation for S-NE was aborted after three days and is not presented here.

Figure 39 shows the posterior distributions for the different populations for the kinetic rates α and γ. For S-M&Std and S-cdf, the posterior distributions are not along the ratio, which is similar to the result of the ABC. The final posterior for both distance functions is approximately centered at $\log(\alpha) = 6.5$ and $\log(\gamma) = 3.5$.

For S-MC the final posterior is neither centered around the true kinetic rates.

The posterior distributions for $\alpha\beta$ and $\gamma\delta$, which are shown in figure 39, are along the ratio of the kinetic rates. The final posterior distribution is not located around the section of $\alpha\beta$ and $\gamma\delta$.

Simulation 14: $\sigma_{\mathbf{LN}} = 1.5$, gaussian perturbation kernel Using a gaussian perturbation kernel, the result of the posterior distributions for α and γ is similar to the uniform perturbation kernel. The distributions are not centered along the ratio of the kinetic rates. The final posterior distribution has approximately the same location for the three distance functions.

For the posterior distributions of $\alpha\beta$ and $\gamma\delta$, the result is also similar to simulation 13. The posterior distributions are located along the ratio of the kinetic rates and the distributions of the final population are not estimating the kinetic rates

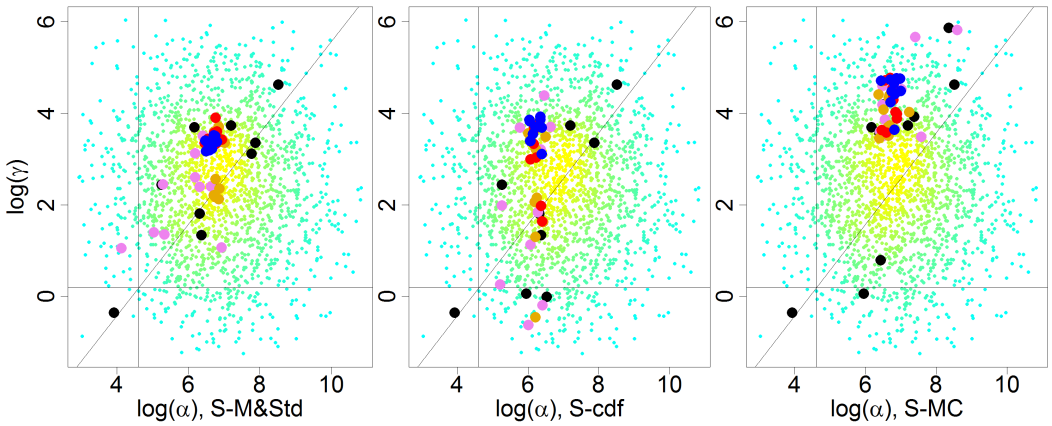


Figure 39: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the two-stage model (simulation 13) for α versus γ.

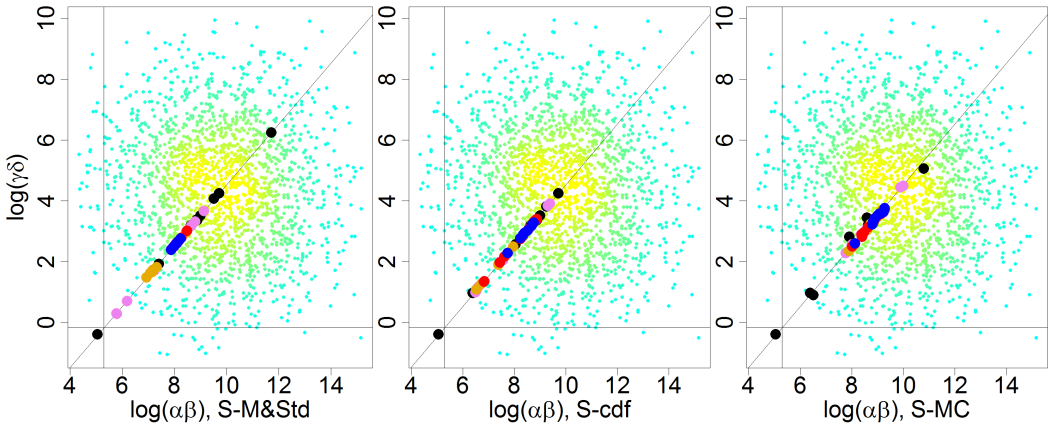


Figure 40: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for $\alpha\beta$ versus $\gamma\delta$ for the two-stage model (simulation 13).

accurately.

Summary In the following a summary is given. Especially, the results are compared with the ABC simulations.

- Compared to the ABC, the estimation of the kinetic rates with the ABC SMC has not the same quality. No simulation could estimate the kinetic rates very accurately, and the estimation did not improve along the populations. This

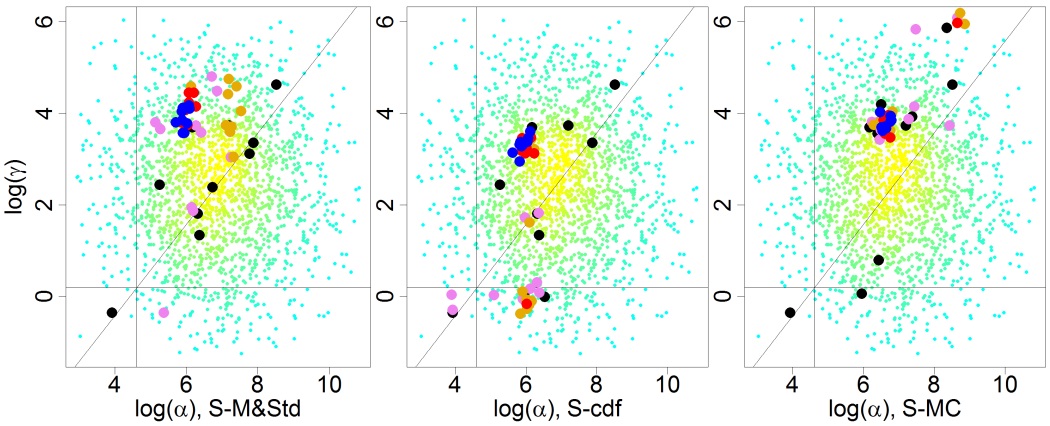


Figure 41: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the two-stage model with gaussian perturbation kernel (simulation 14) for α versus γ.

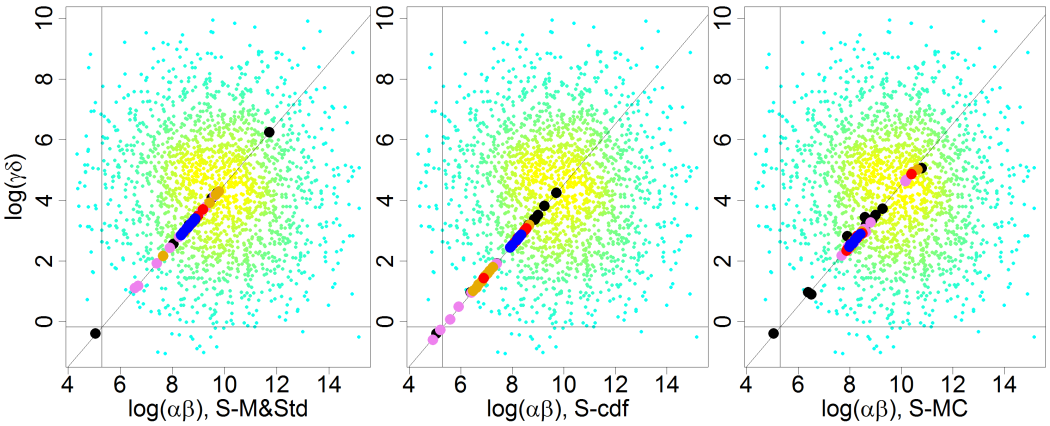


Figure 42: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for $\alpha\beta$ versus $\gamma\delta$ for the two-stage model with gaussian perturbation kernel (simulation 14).

is probably due to the effect which is described in chapter 5.4.2. Therefore, it is proposed to use a similar two-step approach within the ABC SMC as in chapter 5.5.

- There was no large difference in the estimation quality using a uniform or a gaussian perturbation kernel.

7 Summary and outlook

Summary In this thesis, the ABC and ABC SMC algorithms have been investigated with ten different distance functions for the one-stage and two-stage model for gene expression.

The ABC (SMC) algorithm was altered so that the particles are not accepted when their corresponding distance is smaller than a threshold. Instead, a fixed number of particles is drawn and the best are accepted. These two approaches correspond to each other, as a certain threshold always results in approximately the same percentage of accepted particles.

In the one-stage and two-stage model only protein was observed. The observation time, until protein was observed, was set to twice the amount of the time until steady state is reached.

The sampling frequency, i.e. the time between subsequent observations, was determined based on the determinant of the Fisher information matrix. The results indicate that at least for the two-stage model, for the distance functions using time-series data, the sampling frequency did not influence the estimation of the kinetic rates.

For the prior distribution a lognormal distribution $\mathrm{LN}(\mu, \sigma)$ was chosen. The parameters were set so that the mode $\exp(\mu - \sigma^2)$ of the distribution is at the true reaction rate θ. This is different to most publications, which use a uniform distribution as prior.

For the ABC SMC component-wise perturbation kernels were chosen. One is based on the uniform distribution, the other one on the normal distribution.

A statistic, the sum of normalised absolute residuals (SNAR), was introduced. It measures the distance between the average estimation of the kinetic rate and the true kinetic rate. The SNAR has the advantage that the distance to the true rate in the prior distribution can be taken into account. So, different prior distributions can be compared w.r.t their 'informativeness'. Moreover, no publication could be found which explicitly considers the improvement through the ABC (SMC) algorithm based on the amount of information in the prior.

For the one-stage and two-stage model, simulations with a more informative and a less informative prior were conducted for the ABC and ABC SMC. Compared to

other publications, the less informative prior was considerably further away from the true kinetic rates.

For the one-stage model the best estimation could be achieved for the ABC with a distance combined of the mean and the correlation. For the ABC SMC the best distances were based on the mean, standard deviation or the cumulative distribution function. The estimation with the ABC SMC was highly more accurate than with the ABC, using the same number of drawn particles.

For the two-stage model, a very accurate estimation was achieved when γ and δ were first estimated by a distance based on correlation or on the probability density function. Using these results, α and β are then to be estimated in a new simulation, with distance functions based on the mean, standard deviation or the cumulative distribution function. The reason for this is that the distance value, although it depends on all kinetic rates, is mostly determined by the estimation of γ or δ. When one of these rates is estimated well, the values of α or β do not need to be estimated very accurately to achieve a low distance value.

The estimation for the two-stage model, using ABC SMC, did not improve compared to the ABC algorithm. This is due to the fact that all kinetic rates were estimated at once. A better approach would be to first estimate γ and δ and than α and β for each population.

Overall, the steady state of protein, i.e. the ratio of a certain combination of kinetic rates, in the one-stage and two-stage model was estimated very accurate by the distance functions which are based on the mean, standard deviation, the cdf or a combination of mean and correlation. The distance values were lowest along this ratio. For the other distance functions no specific pattern of the distance values could be specified.

Outlook The following list contains ideas for further research:

- The combination of S-Mean and S-Cor showed an improvement, especially for the one-stage model. These two distances were combined using equal weights. One might consider to weight combinations of distance functions differently. Weighting strategies are discussed in Jung & Marjoram (2011). The weights could be a function of time as well.

- For the distance based on the correlation of the simulated and experimental data, Pearson's correlation coefficient and cross-correlation was used. There are other measures to identify a relationship between pairs of variables. Reshef et al. (2011) introduce the maximal information coefficient (MIC), which accounts for both functional and non functional associations.

- A two-step approach was used in chapter 5.5 to estimate the kinetic rates of the two-stage model successfully. The estimation with more complicated models can be divided into several steps. This should increase the estimation quality within the ABC SMC as well.

- The computational study was based on one parameter set with the same initial conditions. The quality of estimation can be studied for different parameter settings and different initial conditions. Moreover, the quality of estimation can be investigated subject to the number of drawn particles.

- It needs a further investigation why the estimation for the two-stage model for S-Cor and S-CC did not improve, although the frequency of the sampled data was changed for simulation 5 and 6 to the optimum of the FIM for TS data.

- Further, the implications of a not equally spaced sampling frequency can be investigated w.r.t. the estimation quality as well.

- Moreover, for calculating the FIM, which is described in chapter 4.1.3, it is necessary to have at least a good estimation of the true kinetic rates and the true underlying model. It can be examined what influence the estimation of the true kinetic rates and the model have on the FIM, on the sampling frequency and on the estimation by the ABC (SMC).

- In the computational study the one-stage and two-stage model were simulated. The quality of estimation can be investigated for more complex models, such as the three-stage model or models including a toggle switch. For the latter one, a toggle switch based on a two-stage model, as discussed in Strasser et al. (2012), might be considered.

- As discussed shortly in chapter 6.1, the quality of estimation is different for different data sets, i.e. different simulated data. The best distance function changed in this small example as well. Nunes & Balding (2010, abstract) found out '*that the optimal set of summary statistics was highly dataset specific, suggesting that more generally there may be no globally-optimal choice, which argues for a new selection for each dataset even if the model and target of inference are unchanged.*' It is interesting if this effect still holds for a large number of drawn particles. Additionally, the influence of the number of simulated trajectories N_{traj} on this aspect can be studied. So the question is how the quality of estimation varies due to the simulated dataset, i.e. how many particles or trajectories need to be simulated to achieve a stable result for the estimation.

- The analysis showed that for the one-stage model, the distance based on mean and correlation had the best estimation for the ABC, i.e. also for the first population of the ABC SMC. For later populations of the ABC SMC other distance functions resulted in a better estimation. Therefore, for different populations different distance functions could be applied to improve the ABC SMC.

- The simulation of the data with the SSA is time consuming. Ramaswamy et al. (2009) propose a new way, which is called the partial-propensity direct method (PDM). It is an exact stochastic simulation algorithm. They state that it outperforms the SSA, especially on strongly coupled reaction networks. The PDM's computational runtime scales linearly with the number of species.

- The SSA is an exact algorithm. Other methods, which are approximations, have a higher computational efficiency. The SSA could be replaced, for instance, by the finite state projection (FSP) method (Munsky & Khammash 2006) or τ leaping algorithms (e.g. Gillespie & Petzold (2003)). Moreover, an approach, which separates the system into slow and fast partitions, where the fast partition is approximated numerically and the slow partition is simulated stochastically can be used. Cao et al. (2005) call this the slow-

scale SSA. Another approach includes the quasi-steady-state approximation (QSSA) (Rao & Arkin 2003).

- Intrinsic noise, i.e. random fluctuations which are due to the discrete and stochastic nature of chemical kinetics for low concentrations can be modelled for instance by the SSA and are considered in this thesis. However, extrinsic noise, which "*is a fairly loosely defined term which essentially represents all noise and heterogeneity in the system not explicitly associated with the intrinsic stochasticity of discrete chemical kinetics*" (Stumpf et al. 2011, p. 365), is not considered. Though, there is no theory yet for modelling of extrinsic noise, Stumpf et al. (2011, ch. 18.3.5) give ideas for modelling extrinsic noise. Extrinsic noise can, in the long run, be included in the estimation process.

References

Beaumont, M., Cornuet, J., Marin, J. & Robert, C. (2009), 'Adaptive approximate Bayesian computation', *Biometrika* **96**(4), 983–990.

Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. L. (2008), 'Bayesian inference for a discretely observed stochastic kinetic model', *Statistics and Computing* **18**(2), 125–135.

Burnette, W. N. (1981), '"Western blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A', *Analytical biochemistry* **112**(2), 195–203.

Cao, Y., Gillespie, D. & Petzold, L. (2005), 'The slow-scale stochastic simulation algorithm', *The Journal of chemical physics* **122**, 014116.

Chen, T., He, H. L. & Church, G. M. (1999), Modeling gene expression with differential equations, *in* 'Pacific Symposium on Biocomputing', pp. 29–40.

Chou, I.-C. & Voit, E. O. (2009), 'Recent developments in parameter estimation and structure identification of biochemical and genomic systems.', *Mathematical biosciences* **219**(2), 57–83.

Cornuet, J.-M., Santos, F., Beaumont, M. a., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T. & Estoup, A. (2008), 'Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation.', *Bioinformatics* **24**(23), 2713–2719.

De Jong, H. (2002), 'Modeling and simulation of genetic regulatory systems: a literature review.', *Journal of computational biology* **9**(1), 67–103.

Del Moral, P., Doucet, A. & Jasra, A. (2006), 'Sequential Monte Carlo samplers', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3), 411–436.

Didelot, X., Everitt, R. G., Johansen, A. M. & Lawson, D. J. (2011), 'Likelihood-free estimation of model evidence', *Bayesian Analysis* **6**(1), 49–76.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2009), *Statistik: Der Weg zur Datenanalyse*, 7th edn, Berlin: Springer.

Fearnhead, P. & Prangle, D. (2012), 'Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation', *Journal of the Royal Statistical Society, Series B* **73**(3), 1–28.

Filippi, S., Barnes, C., Cornebise, J. & Stumpf, M. P. H. (2011), 'On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo', *ArXiv e-prints* . 1106.6280.

Gillespie, D. (1977), 'Exact stochastic simulation of coupled chemical reactions', *The journal of physical chemistry* **81**(25), 2340–2361.

Gillespie, D. (1992), *Markov processes: an introduction for physical scientists*, San Diego: Academic Press.

Gillespie, D. & Petzold, L. (2003), 'Improved leap-size selection for accelerated stochastic simulation', *The Journal of Chemical Physics* **119**, 8229.

Gillespie, D. T. (2007), 'Stochastic simulation of chemical kinetics.', *Annual review of physical chemistry* **58**, 35–55.

Goodwin, B. (1963), *Temporal organization in cells. A dynamic theory of cellular control processes.*, London and New York: Academic Press.

Hayot, F. & Jayaprakash, C. (2008), A tutorial on cellular stochasticity and gillespie's algorithm. Prime Technical Report.

Higham, D. J. (2008), 'Modeling and Simulating Chemical Reactions', *SIAM Review* **50**(2), 347.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001), *Independent component analysis*, New York: Wiley-Interscience.

Joyce, P. & Marjoram, P. (2008), 'Approximately sufficient statistics and bayesian computation.', *Statistical applications in genetics and molecular biology* **7**(1), Article 26.

Jung, H. & Marjoram, P. (2011), 'Choice of Summary Statistic Weights in Approximate Bayesian Computation', *Statistical Applications in Genetics and Molecular Biology* **10**(1).

Karlebach, G. & Shamir, R. (2008), 'Modelling and analysis of gene regulatory networks.', *Nature reviews. Molecular cell biology* **9**(10), 770–80.

Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. (2005), *Systems biology in practice: concepts, implementation and application*, Weinheim: Wiley-VCH.

Komorowski, M., Costa, M. J., Rand, D. A. & Stumpf, M. P. H. (2011), 'Sensitivity, robustness, and identifiability in stochastic chemical kinetics models.', *Proceedings of the National Academy of Sciences of the United States of America* **108**(21), 8645–8650.

Krause, E. (1986), *Taxicab geometry: An adventure in non-Euclidean geometry*, New York: Dover Publications.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Larson, D. R., Singer, R. H. & Zenklusen, D. (2009), 'A single molecule view of gene expression.', *Trends in cell biology* **19**(11), 630–637.

Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T. & Stumpf, M. P. H. (2010), 'ABC-SysBio–approximate Bayesian computation in Python with GPU support.', *Bioinformatics* **26**(14), 1797–1799.

Lillacci, G. & Khammash, M. (2010), 'Parameter estimation and model selection in computational biology.', *PLoS computational biology* **6**(3), e1000696.

Lopes, J. S., Balding, D. & Beaumont, M. A. (2009), 'PopABC: a program to infer historical demographic parameters.', *Bioinformatics* **25**(20), 2747–2749.

Lotka, A. (1925), *Elements of physical biology*, Philadelphia: Williams & Wilkins.

Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. (2003), 'Markov chain Monte Carlo without likelihoods.', *Proceedings of the National Academy of Sciences of the United States of America* **100**(26), 15324–15328.

Müller, C. L., Ramaswamy, R. & Sbalzarini, I. F. (2012), 'Global parameter identification of stochastic reaction networks from single trajectories', *Advances in Systems Biology* **736**(4), 477–498.

Munsky, B. & Khammash, M. (2006), 'The finite state projection algorithm for the solution of the chemical master equation.', *The Journal of chemical physics* **124**(4), 044104.

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

Nunes, M. & Balding, D. (2010), 'On optimal selection of summary statistics for Approximate Bayesian Computation', *Statistical applications in genetics and molecular biology* **9**(1), Article 34.

Pawley, J. (2006), *Handbook of biological confocal microscopy*, 3rd edn, New York: Springer.

Pineda-Krch, M. (2010), *GillespieSSA: Gillespie's Stochastic Simulation Algorithm (SSA)*. R package, version 0.5-4.

Poovathingal, S. K. & Gunawan, R. (2010), 'Global parameter estimation methods for stochastic biochemical systems.', *BMC bioinformatics* **11**(414).

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. (1999), 'Population growth of human Y chromosomes: a study of Y chromosome microsatellites.', *Molecular biology and evolution* **16**(12), 1791–1798.

Prugovecki, E. (2006), *Quantum mechanics in Hilbert space*, 2nd edn, New York: Dover Publications.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ramaswamy, R., González-Segredo, N. & Sbalzarini, I. F. (2009), 'A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks.', *The Journal of chemical physics* **130**(24), 244104.

Rao, C. & Arkin, A. (2003), 'Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm', *The Journal of Chemical Physics* **118**, 4999.

Reinker, S., Altman, R. M. & Timmer, J. (2006), 'Parameter estimation in stochastic biochemical reactions', *Systems Biology* **153**(4), 168 – 178.

Reshef, D. N., Reshef, Y. a., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. (2011), 'Detecting Novel Associations in Large Data Sets', *Science* **334**(6062), 1518–1524.

Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K. & Petzold, L. R. (2011), 'StochKit2: Software for Discrete Stochastic Simulation of Biochemical Systems with Events.', *Bioinformatics* **27**(17), 2457–2458.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. (2011), 'Global quantification of mammalian gene expression control.', *Nature* **473**(7347), 337–42.

Sisson, S. A., Fan, Y. & Tanaka, M. M. (2007), 'Sequential Monte Carlo without likelihoods.', *Proceedings of the National Academy of Sciences of the United States of America* **104**(6), 1760–1765.

Sousa, V. C., Fritz, M., Beaumont, M. A. & Chikhi, L. (2009), 'Approximate bayesian computation without summary statistics: the case of admixture.', *Genetics* **181**(4), 1507–1519.

Strasser, M., Theis, F. J. & Marr, C. (2012), 'Stability and multi-attractor dynamics of a toggle switch based on a two-stage model of gene expression', *Biophysical Journal* **102**(1), 19—-29.

Stumpf, M., Balding, D. & Girolami, M. (2011), *Handbook of Statistical Systems Biology*, Chichester: Wiley.

Tellier, A., Pfaffelhuber, P., Haubold, B., Naduvilezhath, L., Rose, L. E., Städler, T., Stephan, W. & Metzler, D. (2011), 'Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum.', *PloS one* **6**(5), e18155.

Thattai, M. & van Oudenaarden, A. (2001), 'Intrinsic noise in gene regulatory networks.', *Proceedings of the National Academy of Sciences of the United States of America* **98**(15), 8614–8619.

Tian, T., Xu, S., Gao, J. & Burrage, K. (2007), 'Simulated maximum likelihood method for estimating kinetic rates in gene expression.', *Bioinformatics* **23**(1), 84–91.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. (2009), 'Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems', *Journal of The Royal Society Interface* **6**(31), 187–202.

Volterra, V. (1931), *Leçons sur la théorie mathématique de la lutte pour la vie*, Paris: Gauthier-Villars. Reprint 1990 by Jacques Gabay, Paris.

A Appendix

A.1 Distance function based on the Kullback Leibler divergence

One distance based on the Kullback Leibler divergence was formulated and evaluated in preliminary simulations. It is defined in the following

$$d(\mathbf{x^0}, \mathbf{x}^*) = \sum_{t=1}^{T_{obs}} \Big(\mathrm{KL}(\mathbf{x_t^*}, \mathbf{x_t^0}) + \mathrm{KL}(\mathbf{x_t^0}, \mathbf{x_t^*}) \Big) \tag{A.50}$$

where

$$\mathrm{KL}(\mathbf{x_t^*}, \mathbf{x_t^0}) = \int_{-\infty}^{\infty} p(\mathbf{x_t^*}) \cdot \log_2 \frac{p(\mathbf{x_t^*})}{q(\mathbf{x_t^0})} \, \mathrm{d}x \tag{A.51}$$

is the Kullback-Leibler divergence with $p(\cdot)$ and $q(\cdot)$ denoting the probability density functions for $\mathbf{x_t^*}$ and $\mathbf{x_t^0}$ respectively.

For simulation of the Kullback-Leibler divergence, it is calculated as follows:

KS 1 A kernel smoothing density estimate is calculated for $\mathbf{x_t^*}$ and $\mathbf{x_t^0}$, resulting in $P(\mathbf{x_t^*})$ and $Q(\mathbf{x_t^0})$. The densities are approximated with 100 equally spaced data points in the range $[\min(\mathbf{x_t^i} - \epsilon), (\max(\mathbf{x_t^i} + \epsilon]$.[18] In contrast to the standard kernel choice which is the normal kernel, the Epanechnikov kernel is used because it does not result in extensive high density values at the end of the ranges as it would be the case with the Normal kernel. The bandwidth of the kernel smoothing window is estimated by default from $\mathbf{x_t^0}$ and is used for estimating $\mathbf{x_t^*}$ as well.

KS 2 A set of data values χ is defined which is a discrete set of $n = 200$ equally spaced values between $\min(\mathbf{x_t^*}, \mathbf{x_t^0})$ and $\max(\mathbf{x_t^*}, \mathbf{x_t^0})$.

[18] $\pm\epsilon$ is necessary for numerical reasons and it is set to $\epsilon = \pm 10^{-5}$.

The Kullback-Leibler divergence is approximated by

$$\mathrm{KL}(\mathbf{x_t^*}, \mathbf{x_t^0}) = \frac{1}{n} \sum_{x \in \chi} P(x) \cdot \log_2 \frac{P(x)}{Q(x)} \tag{A.52}$$

whereby the possible expression $0\, log_2(0)$ is interpreted as zero. The term $1/n$ is added to receive the mean of the Kullback-Leibler divergence as it would otherwise grow, if the length of the set χ is expanded and is set for simulation to 1/200.

The Kullback Leibler divergence is defined for random variables having the identical support χ (Kullback & Leibler 1951), i.e. if $Q(i) > 0$ for any i such that $P(i) > 0$. This assumption does not hold for the given data, as in most cases the support of $\mathbf{x_t^0}$ and $\mathbf{x_t^*}$ is not identical. Therefore, this distance was not used for the final simulations.

A.2 Computational time for the simulations

Although the results are not yet presented, the time of the simulations should be stated beforehand because it explains why certain simulations have not been conducted. The simulations have run on different devices, partly with other users, resulting in shared memory and different workload. Moreover, all simulations for the ABC have run on a queue, meaning N_{all} was divided into a number of smaller packages, which run separately. The number of packages was different for the two-stage models, as these were split into smaller packages due to their higher simulation time. For these reasons, the time stated in table A.16 for creating the simulated data x^* is only an indication, from which the necessary time can be approximately derived.

model	σ_{LN}	N_{all}	total time [h]
one-stage	0.2	$10,000$	0.67
one-stage	2	$10,000$	3.35
one-stage	2	$100,000$	48.29
two-stage	0.2	$10,000$	4.0
two-stage	1.5	$10,000$	283.36
two-stage	1.5	$100,000$	3110.3

Table A.16: Runtime [h] for creation of x^*.

In table A.16 the runtimes for data creation are all based on splitting the data into packages of 50, 200 or 500. The total time for creation of x^* is the sum of all packages. The standard deviation of the runtime for the models was about 0.001 of the total time for creation, and it is, therefore, not specified explicitly.

In table A.17 only the runtime for calculation of the distance values for the prior with $\sigma_{LN} = 0.2$ is given, as the runtimes for other σ_{LN} are the same within $\pm 10\%$ of this value.

distances	Mean	Std	M&Std	NE	NE II	pdf	cdf	Cor	CC
one-stage	0.06	0.14	0.18	0.98	0.82	3.08	2.83	0.02	12.91
two-stage	0.09	0.17	0.21	0.86	0.82	3.12	3.7	0.2	13.67

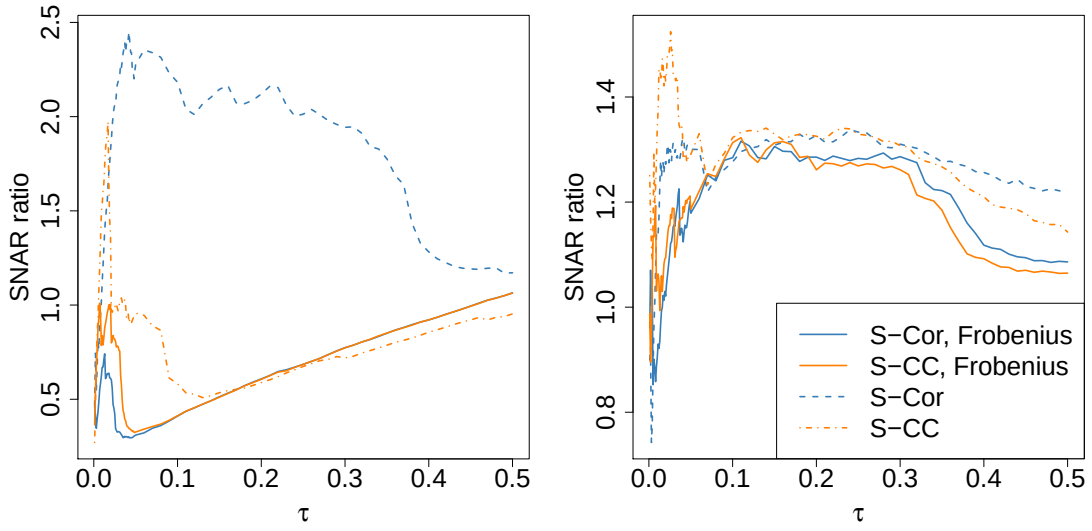
Table A.17: Runtime [h] for calculation of $d(x^0, x^*)$ for $N_{all} = 10,000$.

The long runtime for the two-stage model using $N_{all} = 100,000$ can be explained with the following: from the 500 packages, which were sent to the queue, about five to ten had a runtime of more than three days. The overall standard deviation was 4.94 hours. Due to the long runtime, $\sigma_{LN} = 2$ could not be simulated for the two-stage model, as the creation of the simulated data is too time consuming.

A.3 Using the Frobenius norm instead of norm (36)

This section illustrates the result if the Frobenius norm is used for S-Cor and S-CC instead of the norm (36). In figure A.43 the SNAR ratio is calculated for the one-stage and two-stage model with $N_{all} = 10,000$ and the optimal frequency for TS data, see section 5.2 for further details.

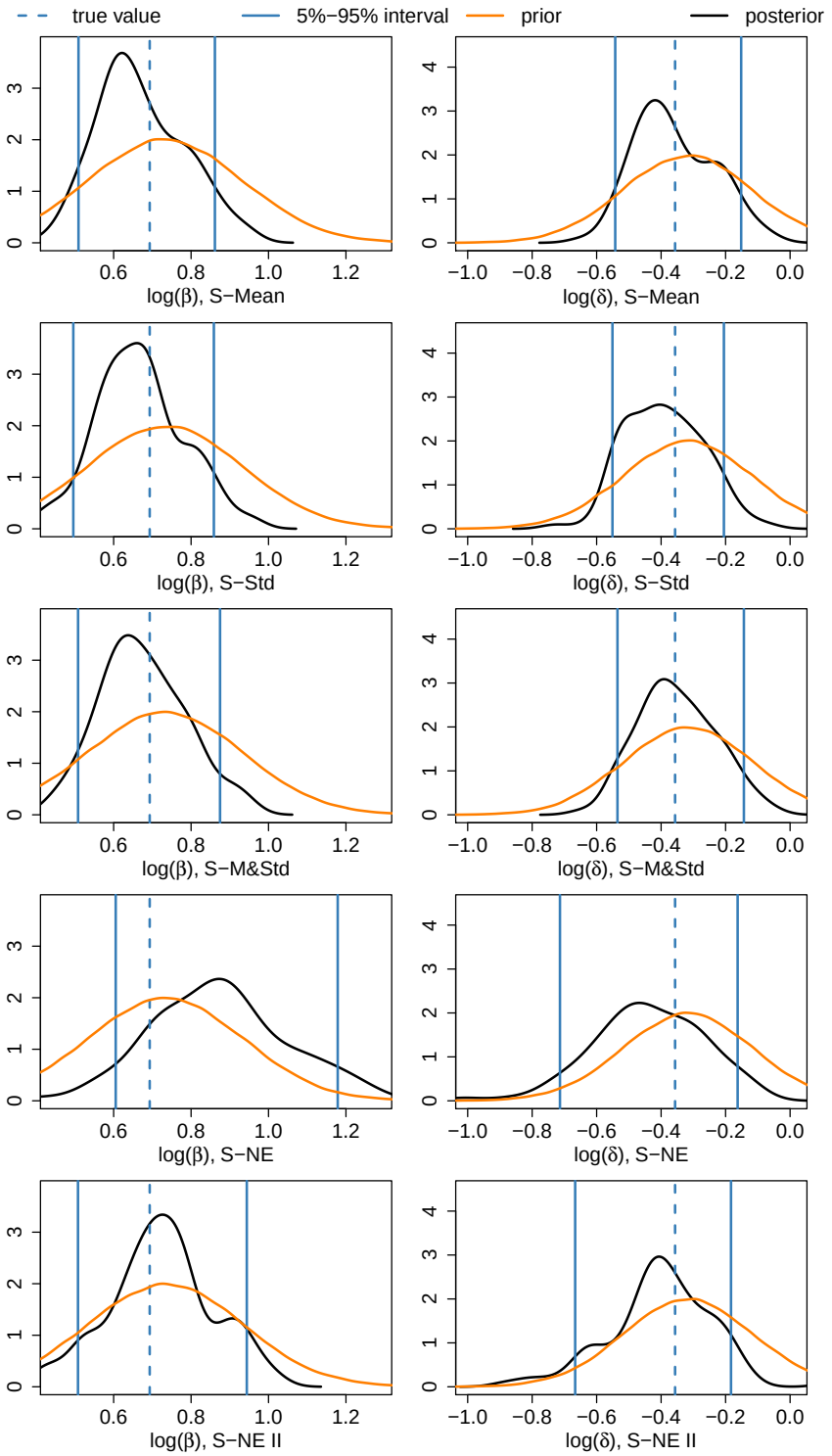
The SNAR ratio using the Frobenius norm is for most τ less than with norm (36). This holds especially for the one-stage model. Therefore, one can assume that the Frobenius norm is not adequate for the distances S-Cor and S-CC.



a) one-stage model b) two-stage model

Figure A.43: SNAR ratio depending on the acceptance rate τ for S-Cor and S-CC using the Frobenius norm and norm (36) for the one-stage model (simulation 5) and two-stage model (simulation 6).

A.4 Appendix for simulation 1



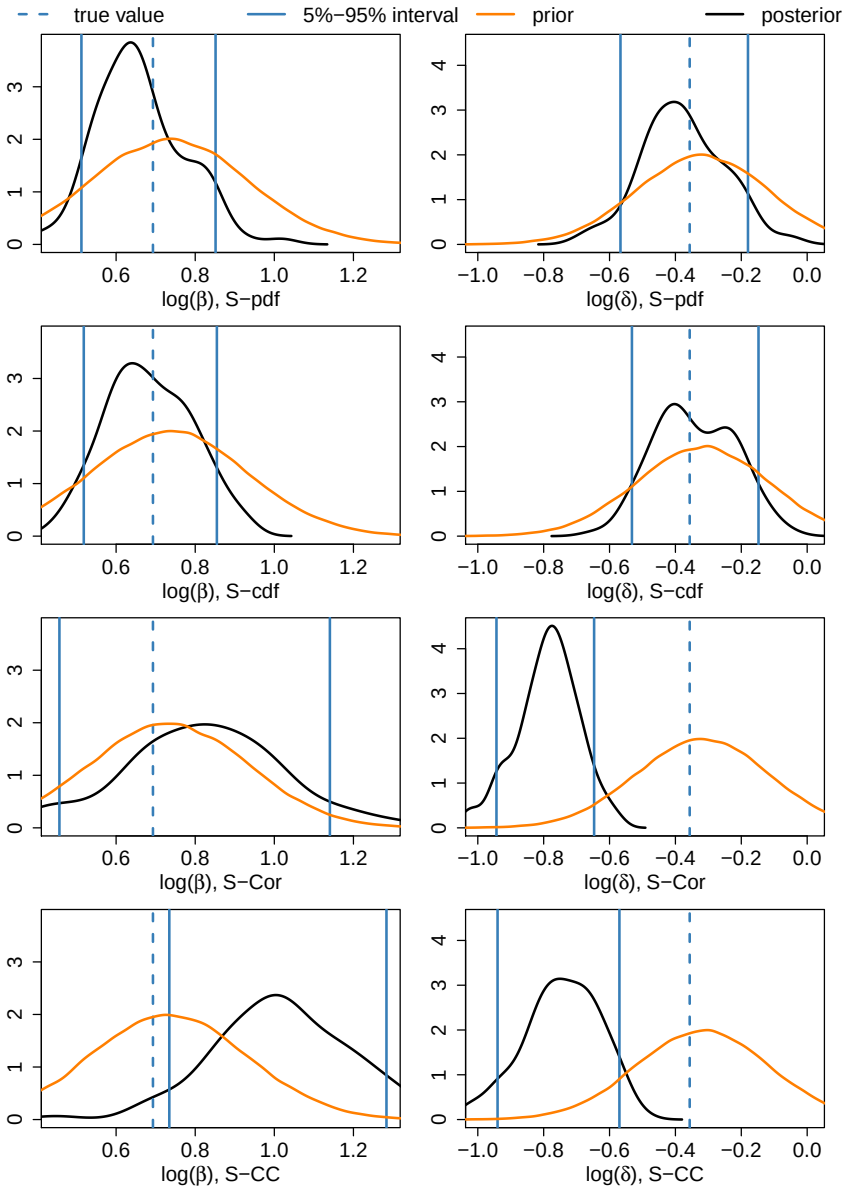
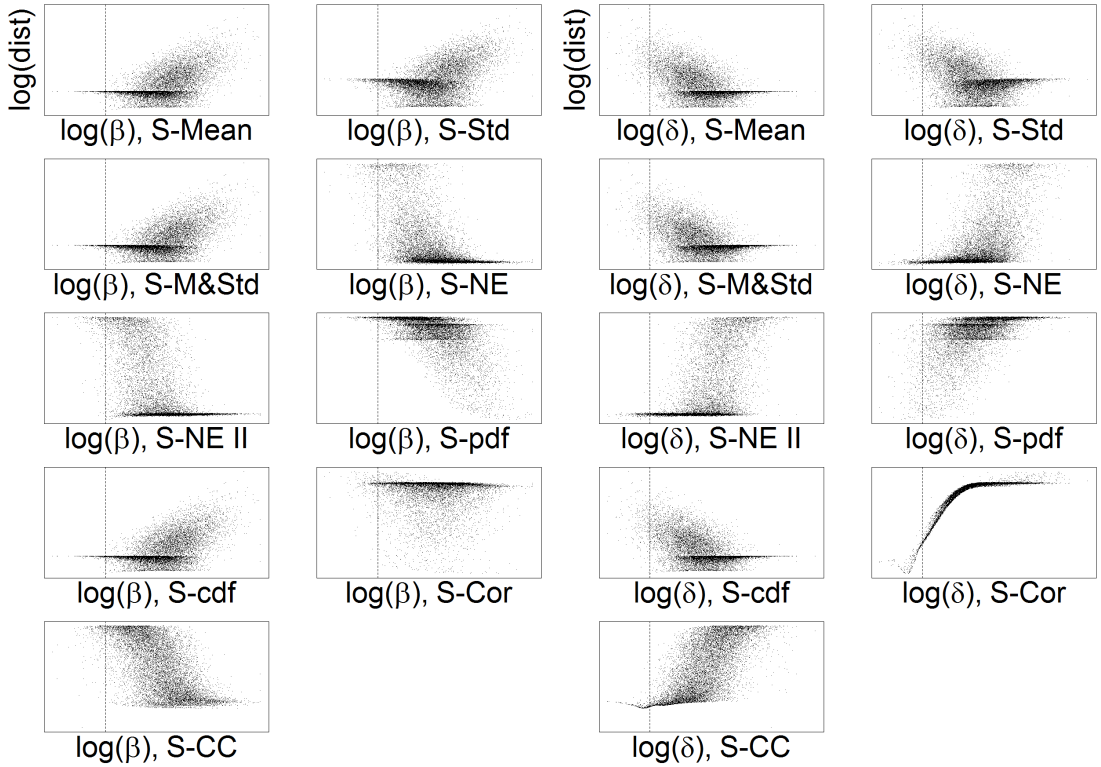


Figure A.44: Kernel density estimation of the prior and posterior distribution for all distances for the one-stage model (simulation 1), $\tau = 0.01$.

A.5 Appendix for simulation 2



a) reaction rate β. b) reaction rate δ.

Figure A.45: Scatterplot of the logarithm of the distance for nine different distance functions for reaction rates β and δ for the one-stage model (simulation 2). The dashed line indicates the logarithm of the true reaction rate.

Description of figure A.46: The distance is lowest for the distance functions S-Mean, S-Std, S-M&Std and S-cdf for the ratio of the reaction rates. S-pdf has a strip of small distance values along the ratio but the bulk of the lowest distances is for large β and small δ. S-NE and S-NE II have their lowest values right of the ratio line. The distribution for S-Cor and S-CC looks roughly the same as in simulation 1. The SNAR ratio is decreasing for larger τ for distances S-Mean, S-Std, S-M&Std, S-cdf and S-NE II. The other distances show either an increasing SNAR ratio for larger τ or stay approximately constant. This is because these distances estimate δ very well but cannot estimate β.

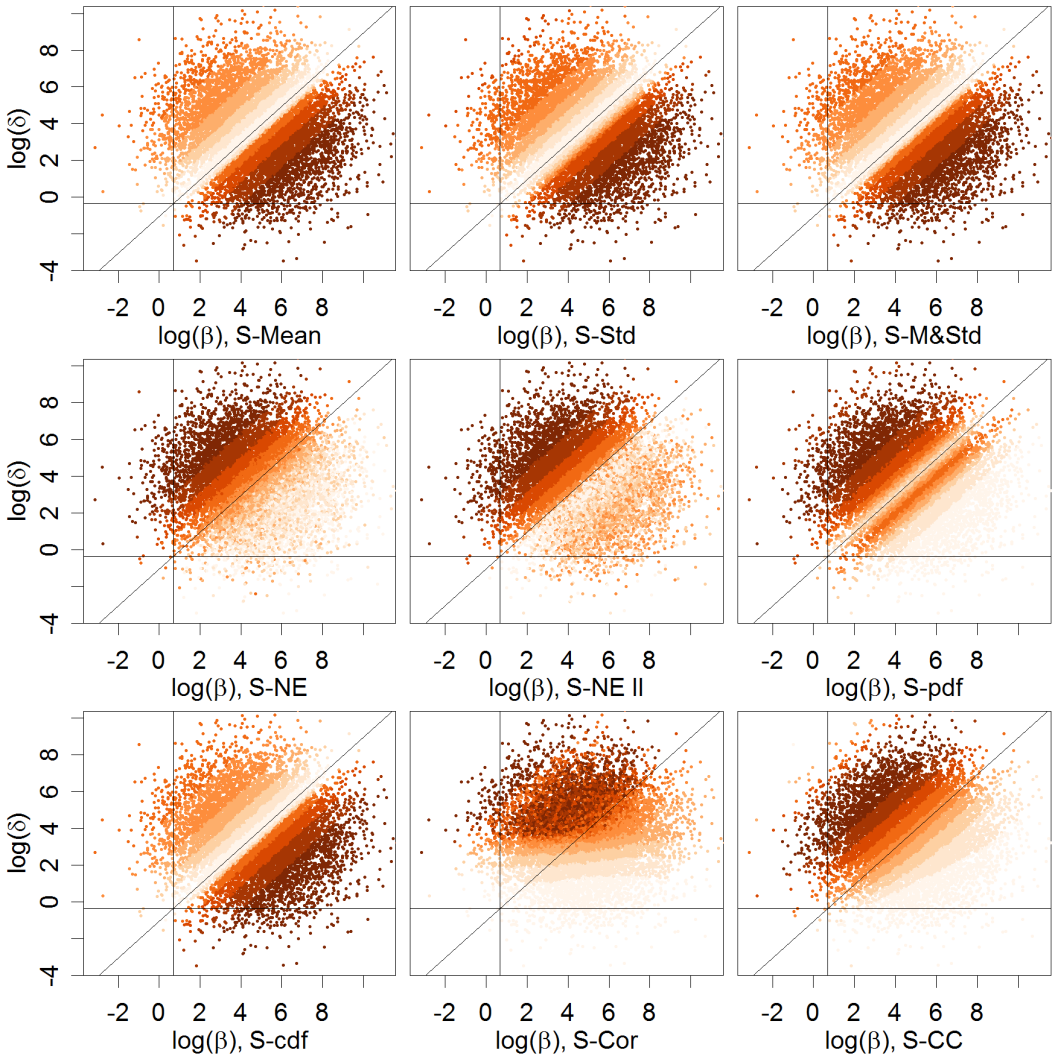


Figure A.46: Distance values for the nine distance functions depending on the reaction rates β and δ for the one-stage model (simulation 2). The solid lines indicate the true reaction rate and the ratio δ/β

Description of figure A.48: The SNAR ratio for different number of drawn particles is shown in figure A.48. As for simulation 1, with an increasing amount of N_{all} the SNAR ratio seems to fluctuate about a certain value. Only the distances S-NE II and partly S-Cor show some anomaly which might diminish for a larger N_{all}.

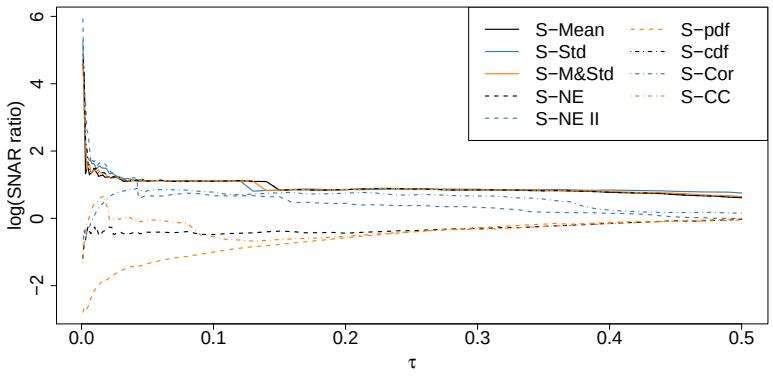
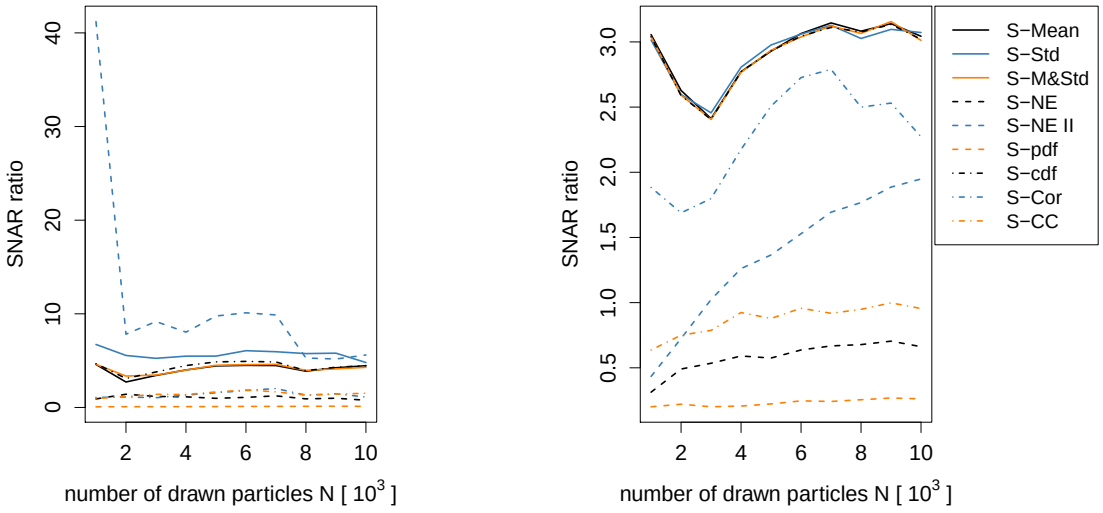
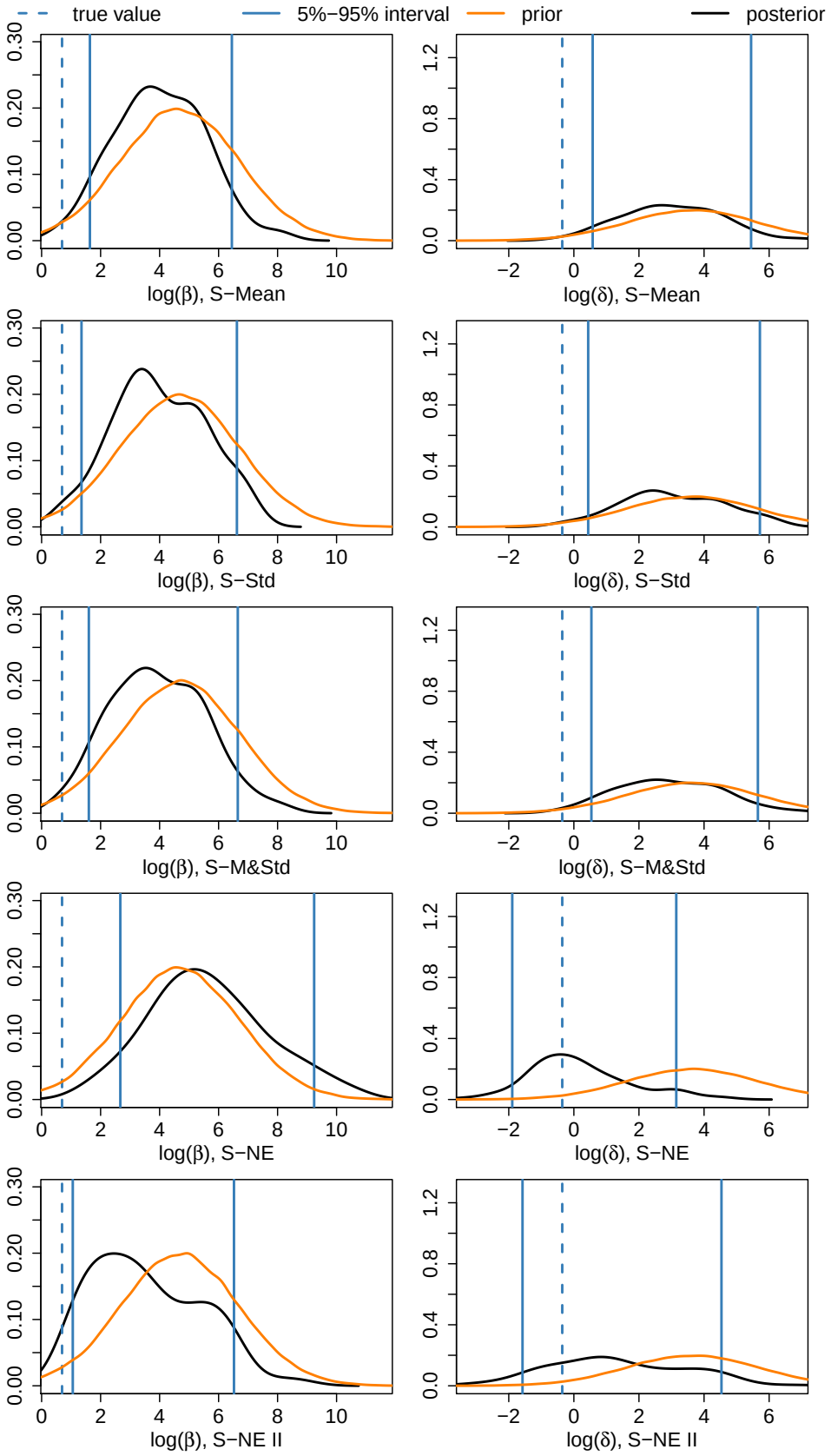


Figure A.47: SNAR ratio depending on the acceptance rate τ for nine distance functions for the one-stage model (simulation 2).



a) $\tau = 0.01$. b) $\tau = 0.05$.

Figure A.48: SNAR ratio depending on the number of drawn particles N_{all} for nine distance functions and two acceptance rates $\tau = 0.01$ and $\tau = 0.05$ for the one-stage model (simulation 2).


true value
5%−95% interval
prior
posterior
log(β), S−Mean
log(δ), S−Mean
log(β), S−Std
log(δ), S−Std
log(β), S−M&Std
log(δ), S−M&Std
log(β), S−NE
log(δ), S−NE
log(β), S−NE II
log(δ), S−NE II

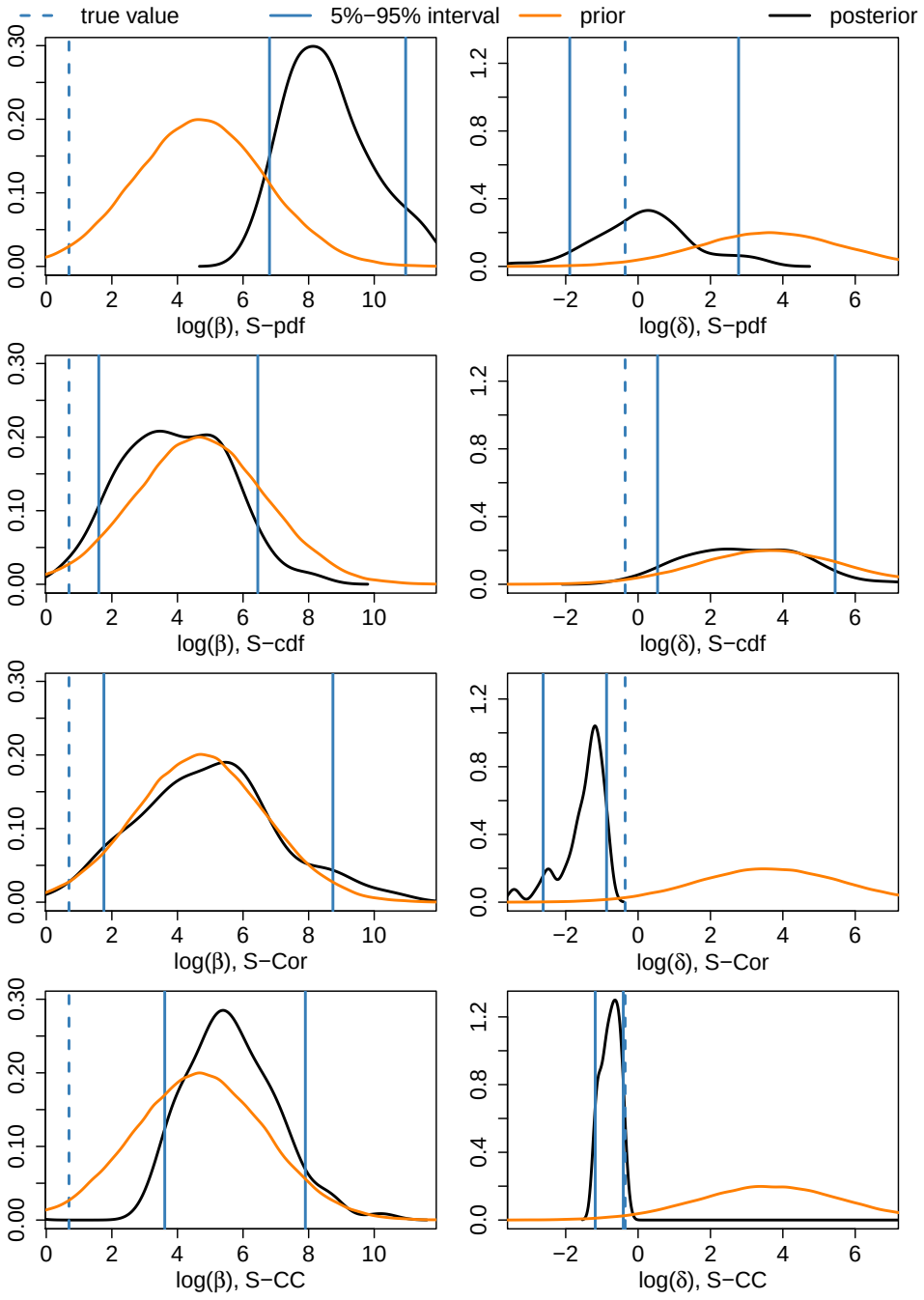


Figure A.49: Kernel density estimation of the prior and posterior distribution for all distances for the one-stage model (simulation 2), $\tau = 0.01$.

A.6 Appendix for simulation 3

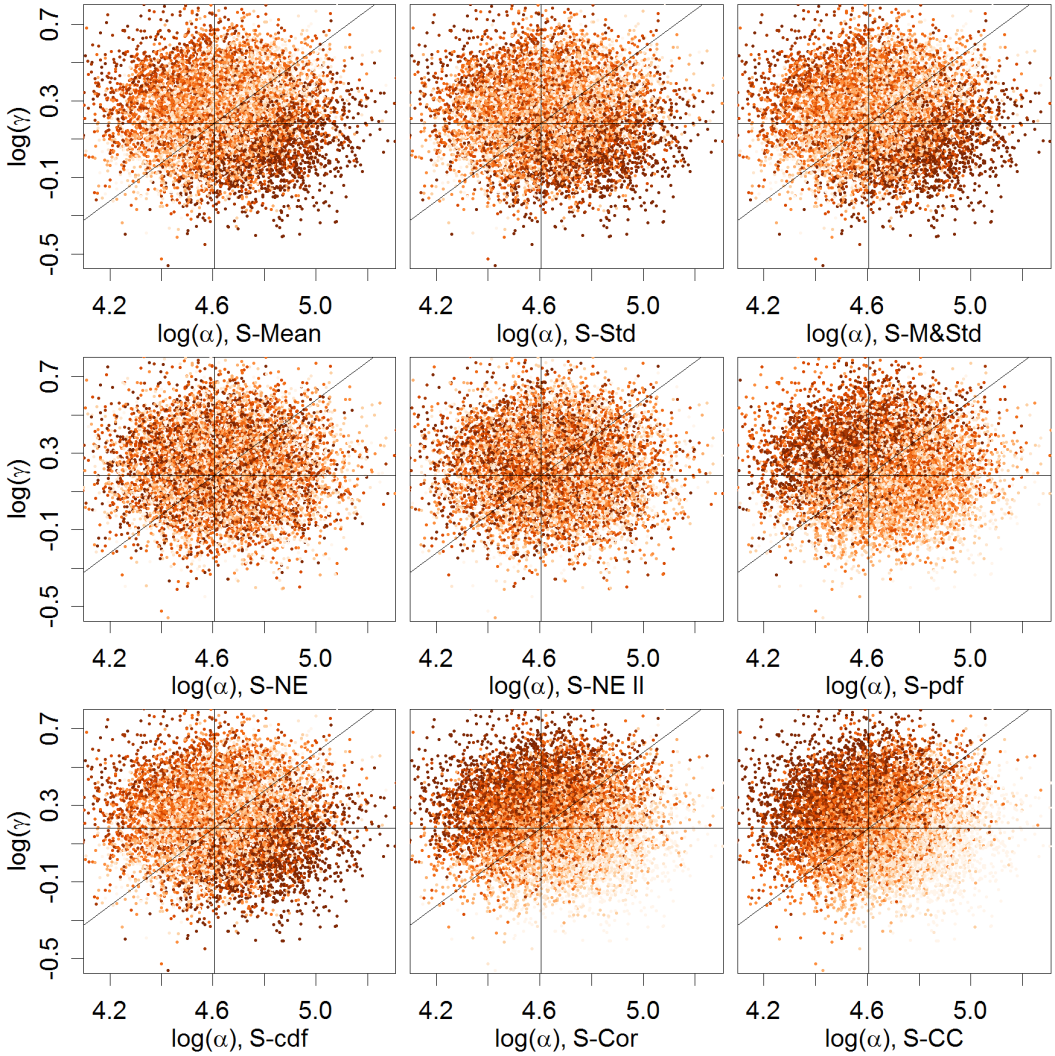


Figure A.50: Distance values for the nine distance functions depending on both reaction rates α and γ for the two-stage model (simulation 3). The solid lines indicate the true reaction rate and the ratio γ/α.

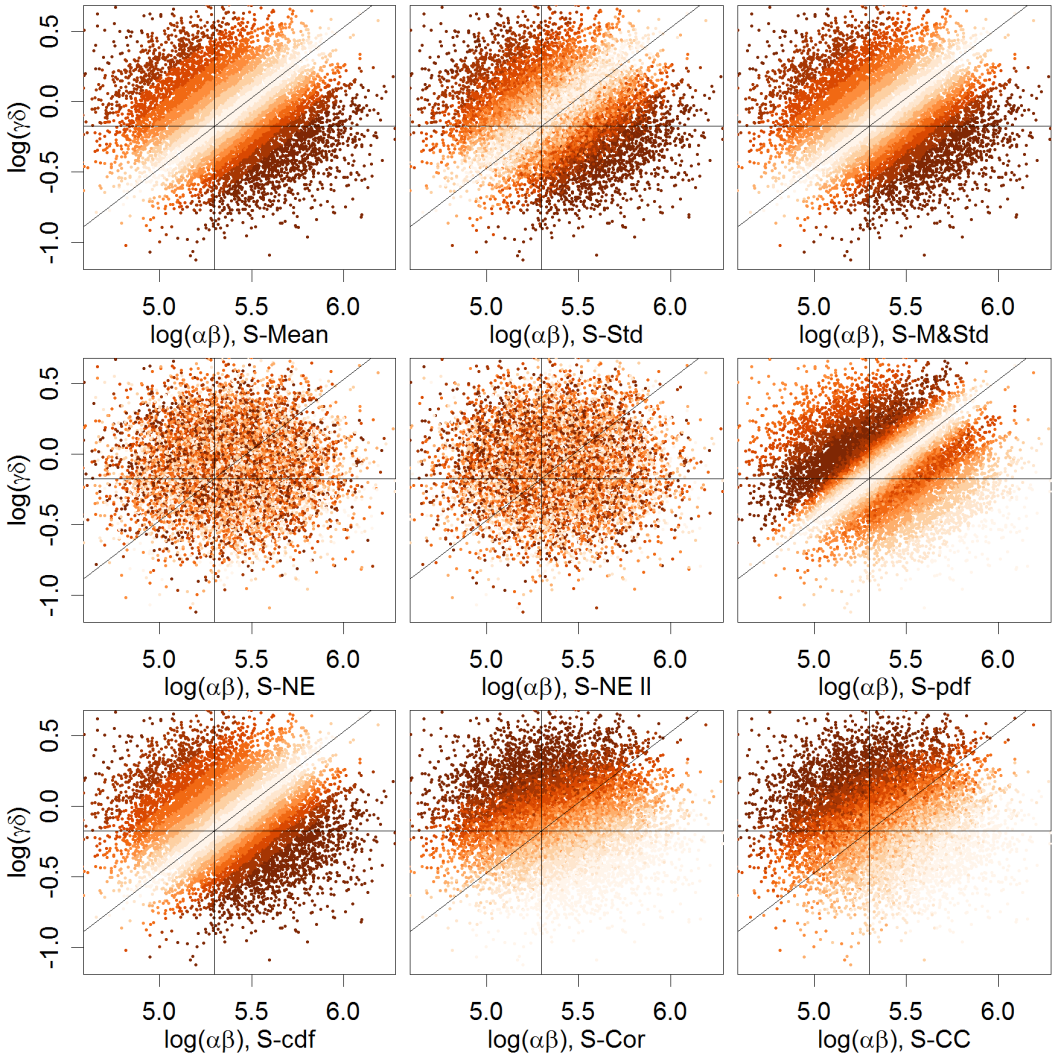
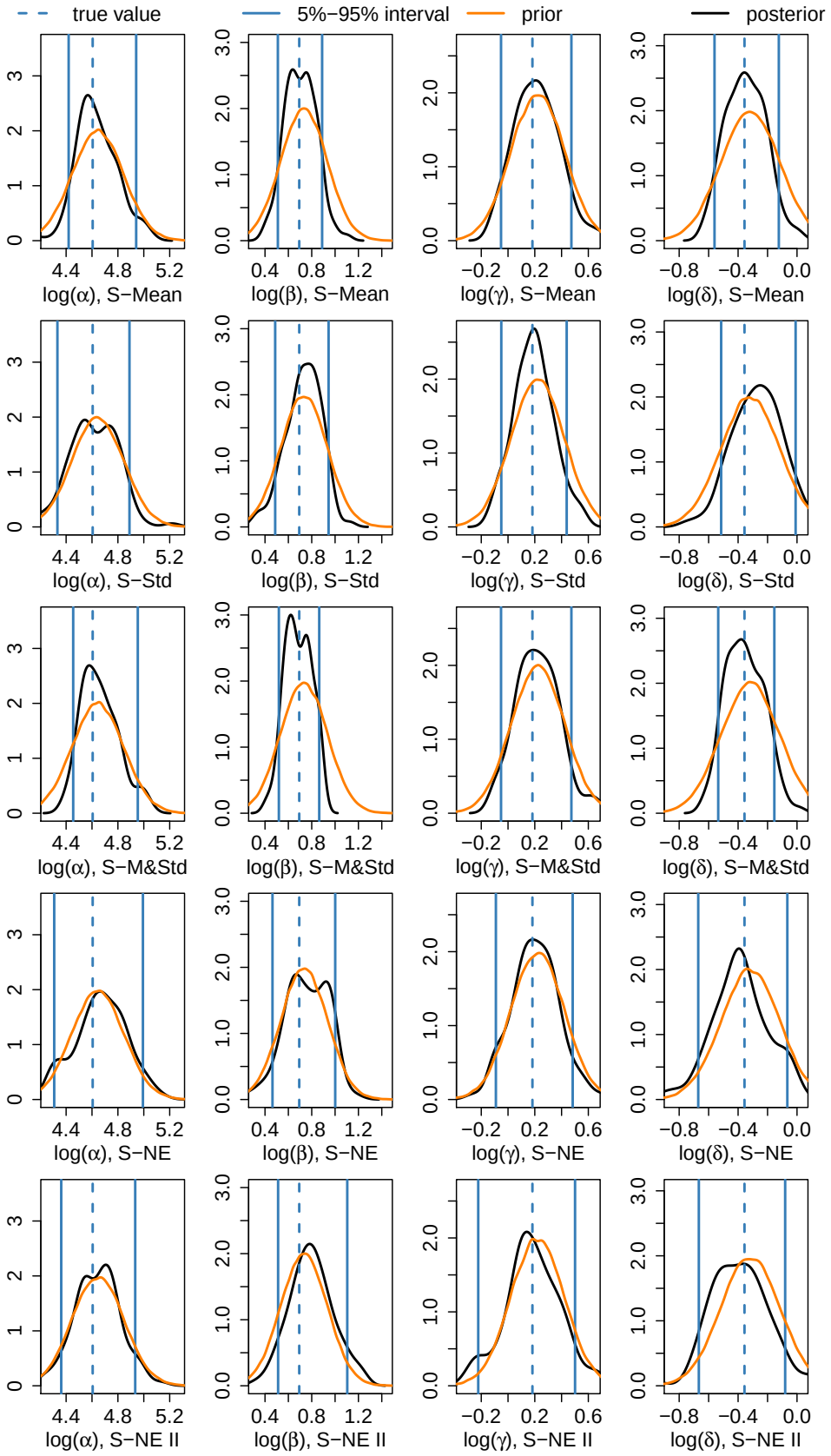


Figure A.51: Distance values for the nine distances depending on the reaction rates $\alpha\beta$ and $\gamma \cdot \delta$ for the two-stage model (simulation 3). The solid lines indicate the true reaction rate and the ratio $(\gamma\delta)/(\alpha\beta)$


true value
5%−95% interval
prior
posterior
log(α), S−Mean
log(β), S−Mean
log(γ), S−Mean
log(δ), S−Mean
log(α), S−Std
log(β), S−Std
log(γ), S−Std
log(δ), S−Std
log(α), S−M&Std
log(β), S−M&Std
log(γ), S−M&Std
log(δ), S−M&Std
log(α), S−NE
log(β), S−NE
log(γ), S−NE
log(δ), S−NE
log(α), S−NE II
log(β), S−NE II
log(γ), S−NE II
log(δ), S−NE II

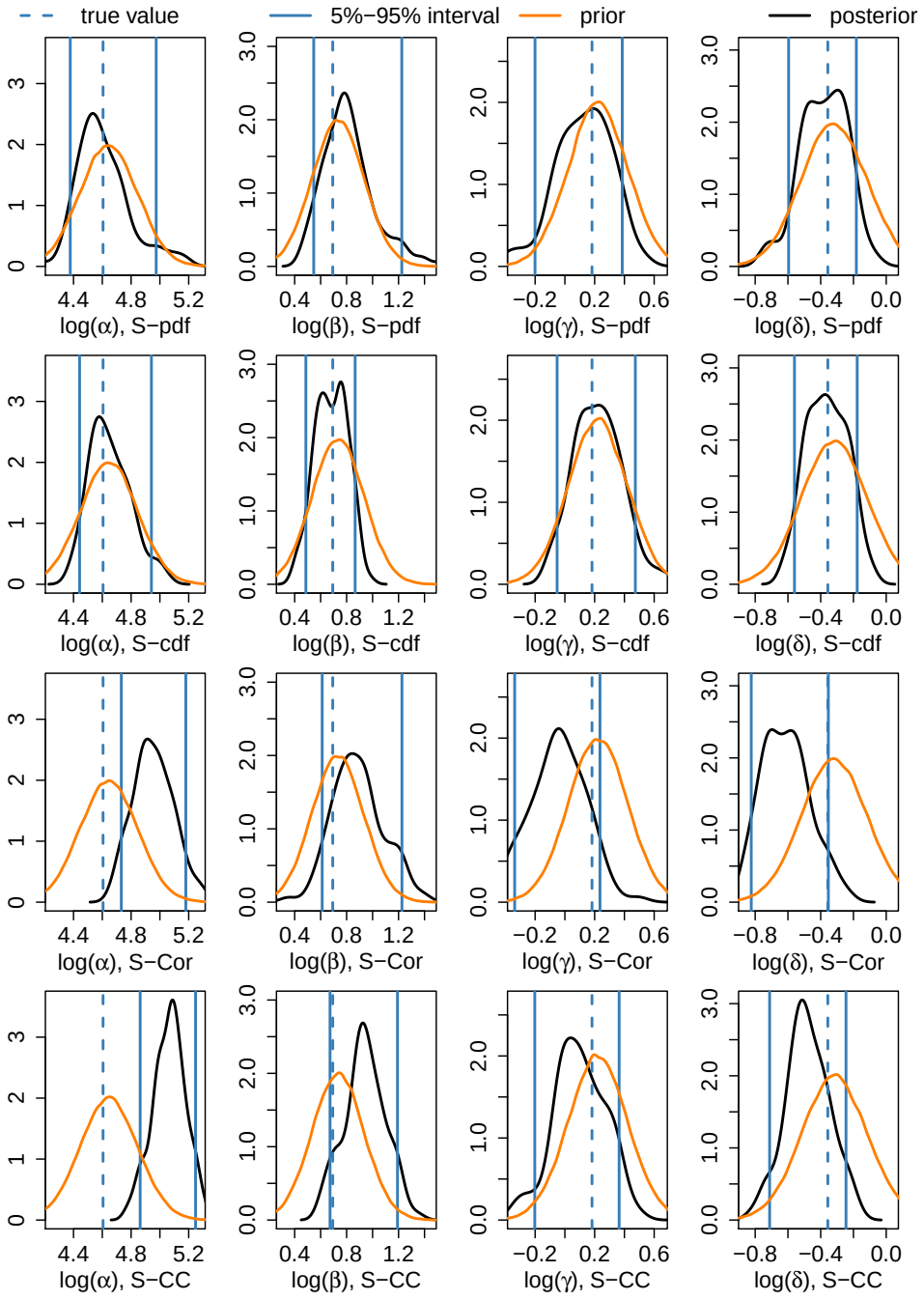


Figure A.52: Kernel density estimation of the prior and posterior distribution for all distances for the two-stage model (simulation 3), $\tau = 0.01$.

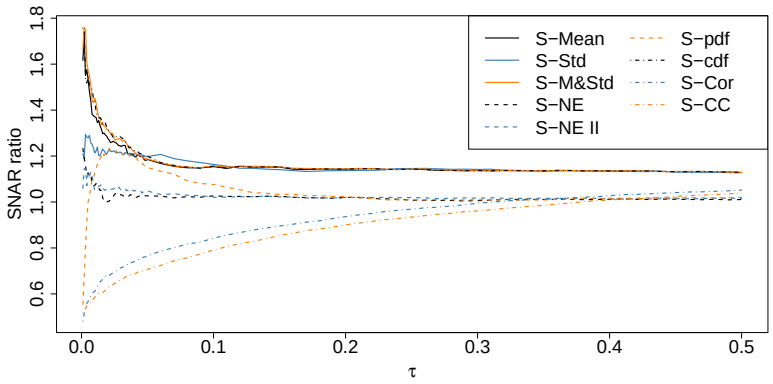


Figure A.53: SNAR ratio depending on the acceptance rate τ for all distances for the two-stage model (simulation 3).

A.7 Appendix for simulation 4

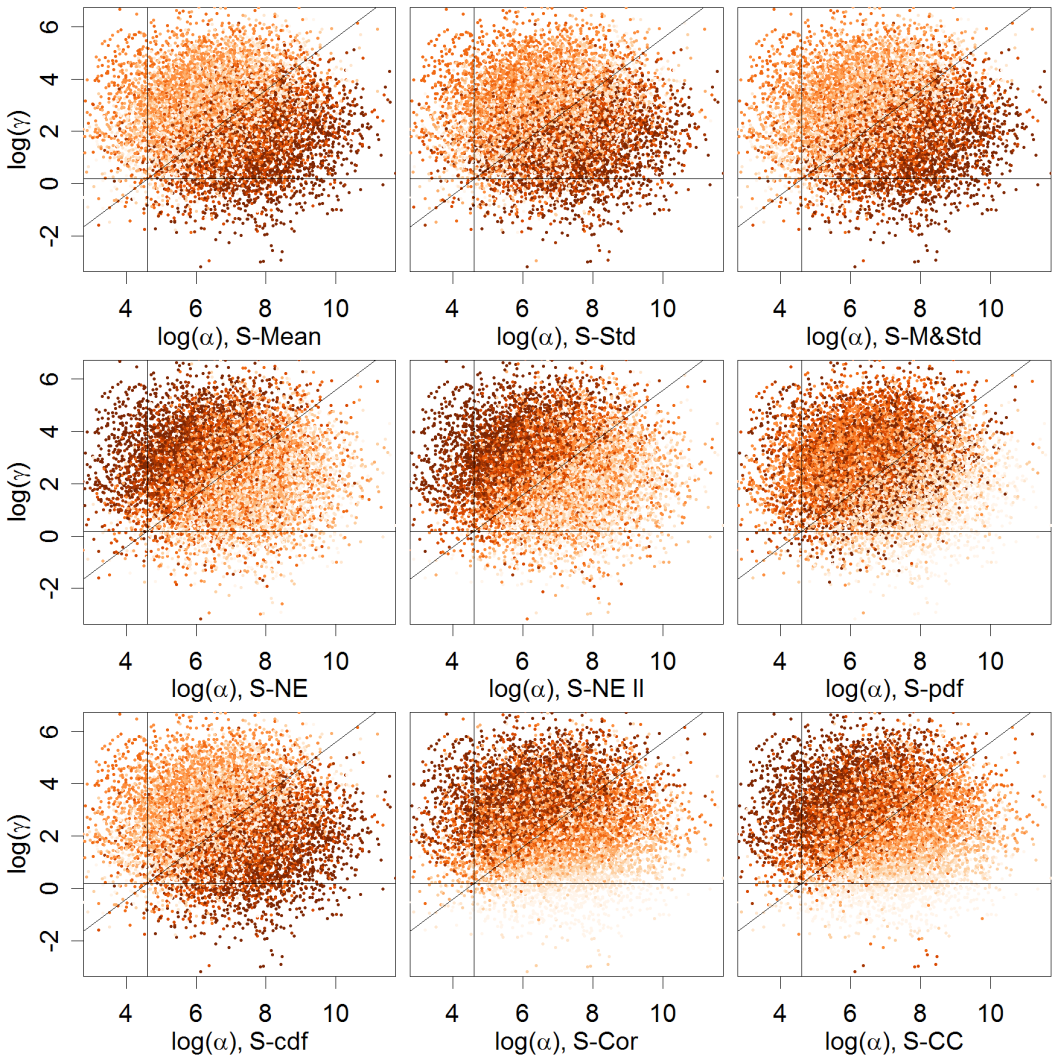


Figure A.54: Distance values for the nine distance functions depending on the reaction rates α and γ for the two-stage model (simulation 4). The solid lines indicate the true reaction rate and the ratio γ/α

For α versus γ (figure A.54) S-Mean, S-Std, S-M&Std and S-cdf have low distance values left of the ratio line. For S-NE, S-NE II and S-pdf the low values are more to the right of the line, for S-Cor and S-CC it is for small γ values.

Considering $\alpha\beta$ versus $\gamma\delta$ (figure A.55) the ratio can be estimated well by S-Mean, S-Std, S-M&Std and S-cdf. S-Cor and S-CC have their low distance values

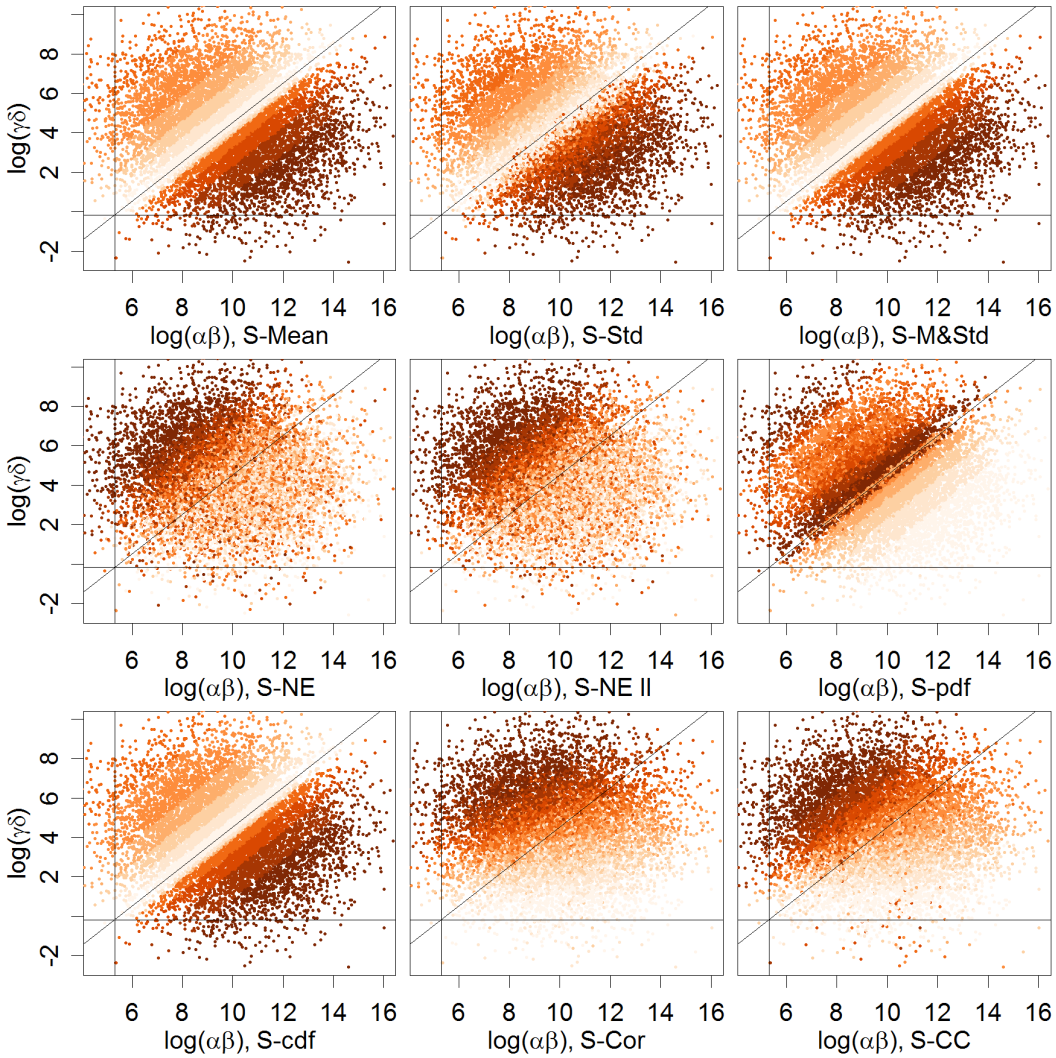
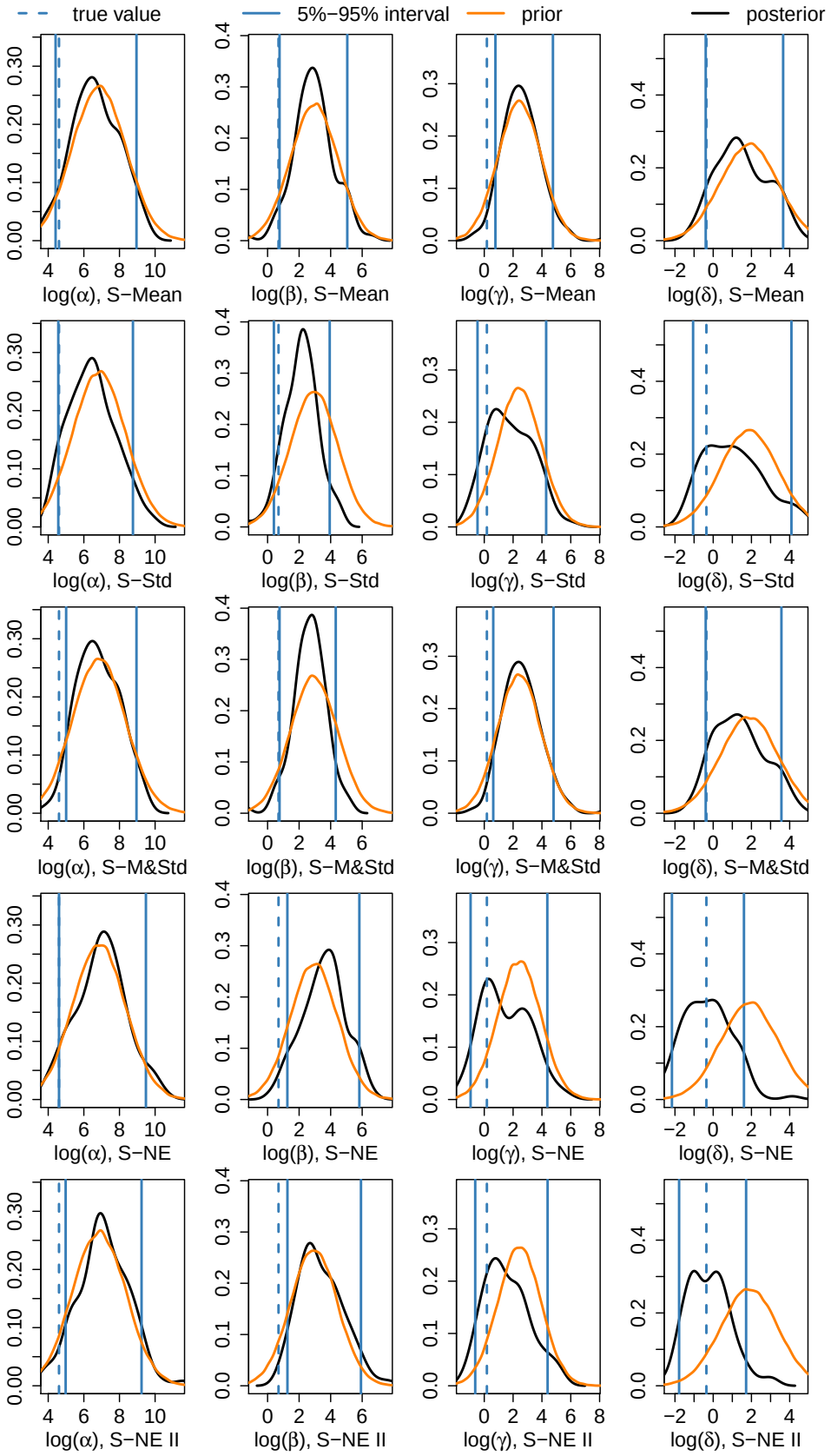


Figure A.55: Distance values for the nine distance functions depending on the reaction rates $\alpha\beta$ and $\gamma\delta$ for the two-stage model (simulation 4). The solid lines indicate the true reaction rate and the ratio $(\gamma\delta)/(\alpha\beta)$

around the true value of $\gamma\delta$.


true value
5%−95% interval
prior
posterior
log(α), S−Mean
log(β), S−Mean
log(γ), S−Mean
log(δ), S−Mean
log(α), S−Std
log(β), S−Std
log(γ), S−Std
log(δ), S−Std
log(α), S−M&Std
log(β), S−M&Std
log(γ), S−M&Std
log(δ), S−M&Std
log(α), S−NE
log(β), S−NE
log(γ), S−NE
log(δ), S−NE
log(α), S−NE II
log(β), S−NE II
log(γ), S−NE II
log(δ), S−NE II

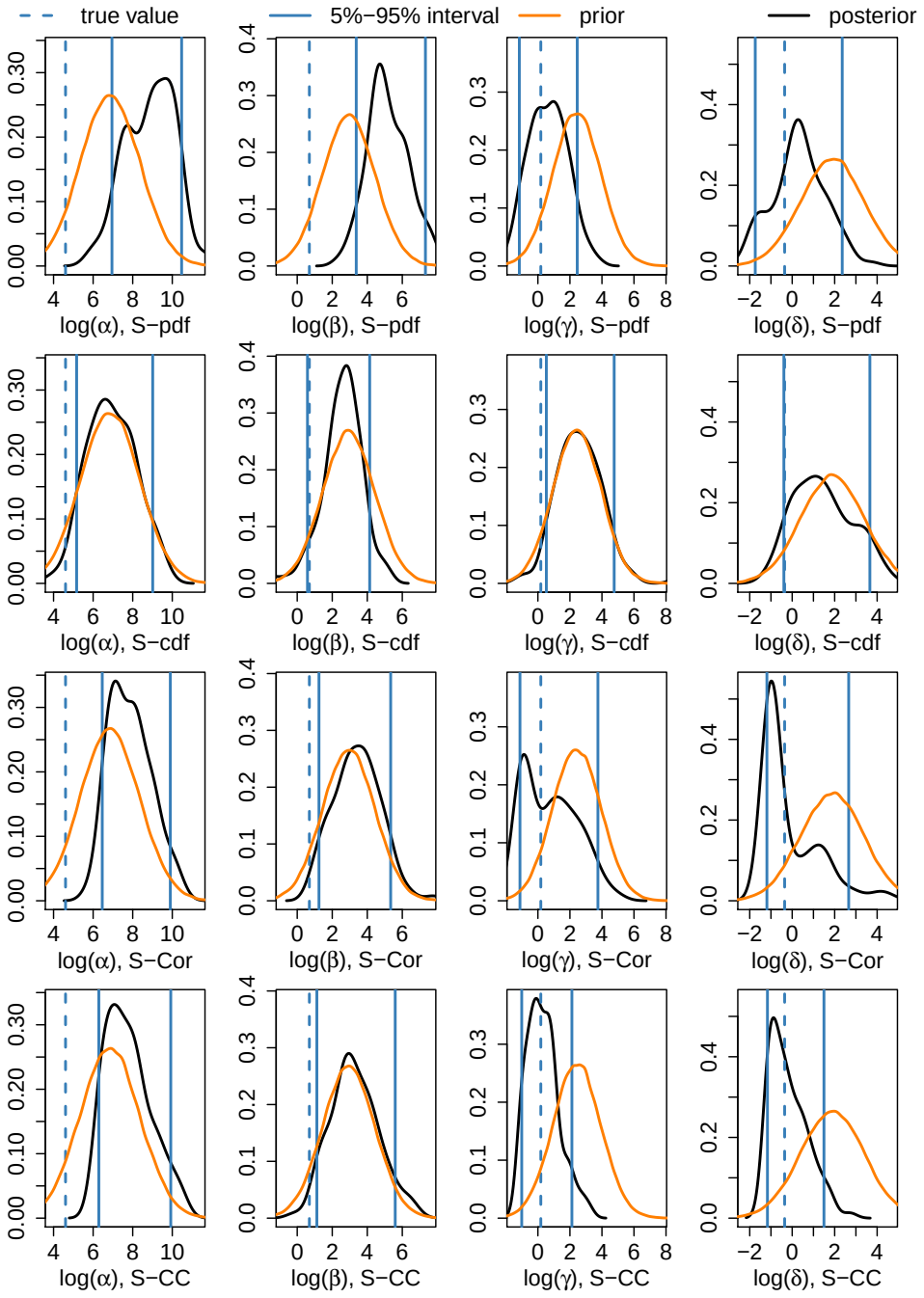


Figure A.56: Kernel density estimation of the prior and posterior distribution for all distances for the two-stage model (simulation 4), $\tau = 0.01$.

A.8 Appendix for simulation 8

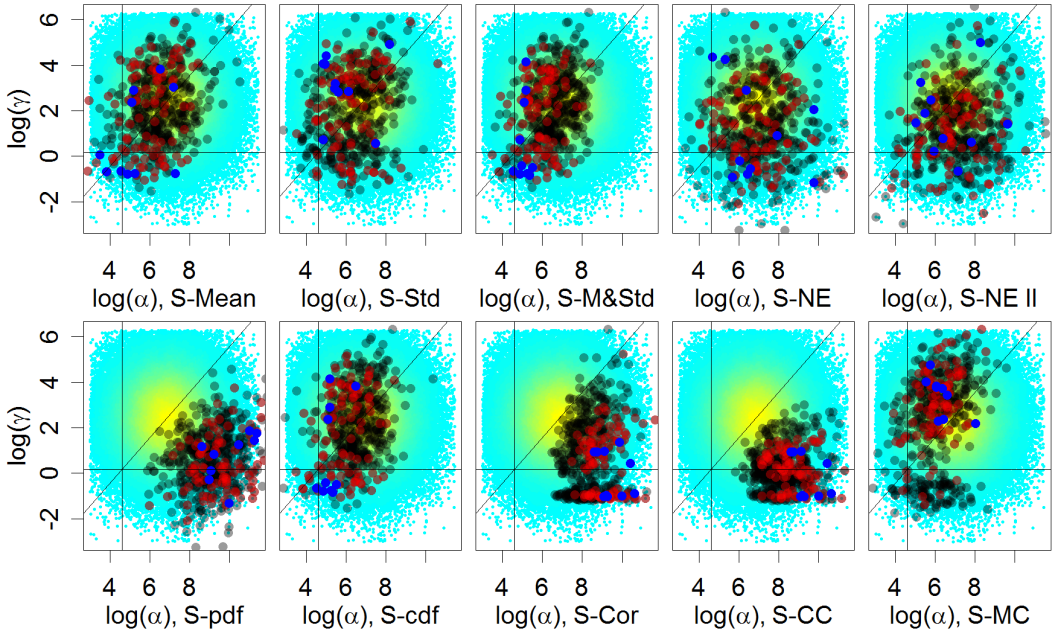


Figure A.57: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.005$ (black), $\tau = 0.001$ (red) and $\tau = 0.0001$ (blue) for the two-stage model for α versus γ (simulation 8).

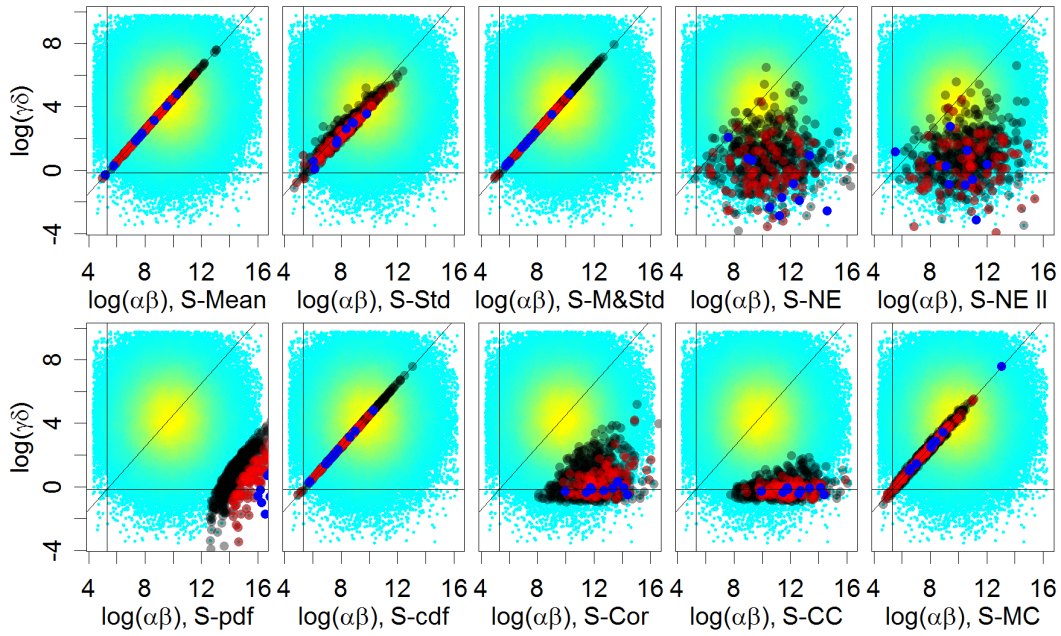
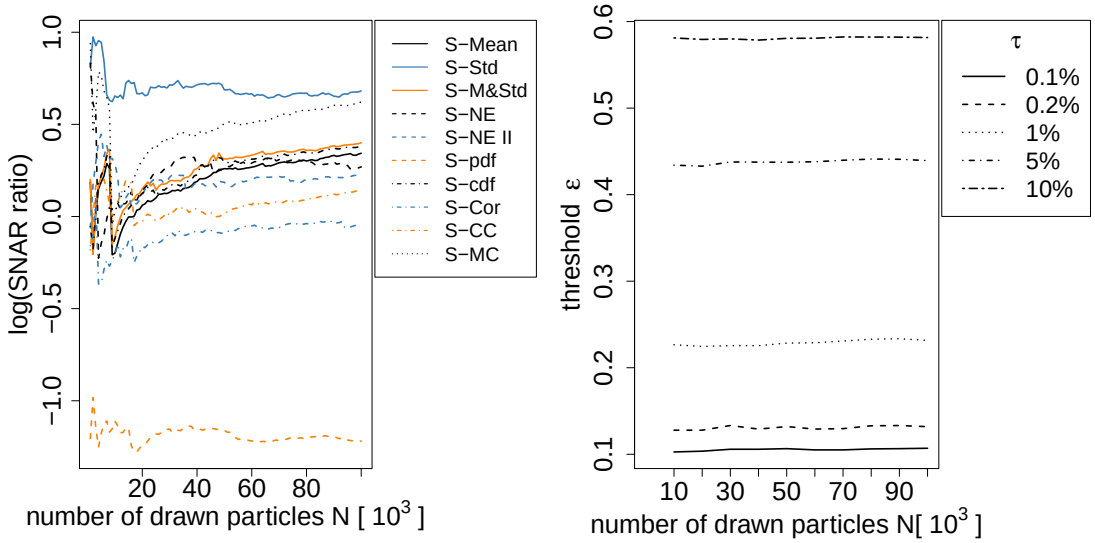


Figure A.58: Prior distribution of the kinetic rates (cyan, yellow) and the resulting posterior for $\tau = 0.005$ (black), $\tau = 0.001$ (red) and $\tau = 0.0001$ (blue) for the two-stage model for $\alpha\beta$ versus $\gamma\delta$ (simulation 8).



a) SNAR ratio against N_{all} for $\tau = 0.01$. b) ϵ against number of drawn particles N_{all} for S-MC.

Figure A.59: SNAR ratio and threshold ϵ against number of drawn particles N_{all} for simulation 8.

A.9 Appendix for simulation 10

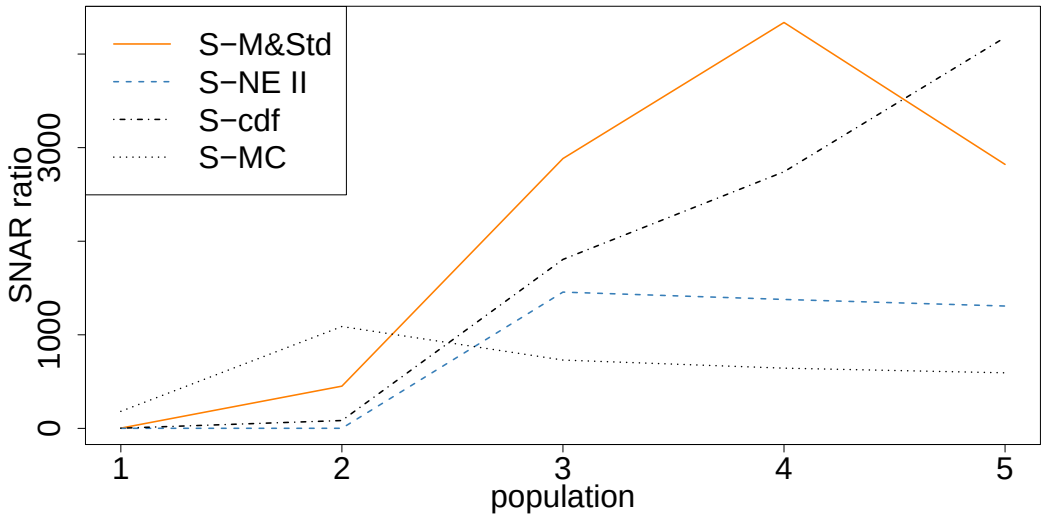


Figure A.60: SNAR ratio for each population for ABC SMC for the one-stage model (simulation 10).

A.10 Appendix for simulation 12

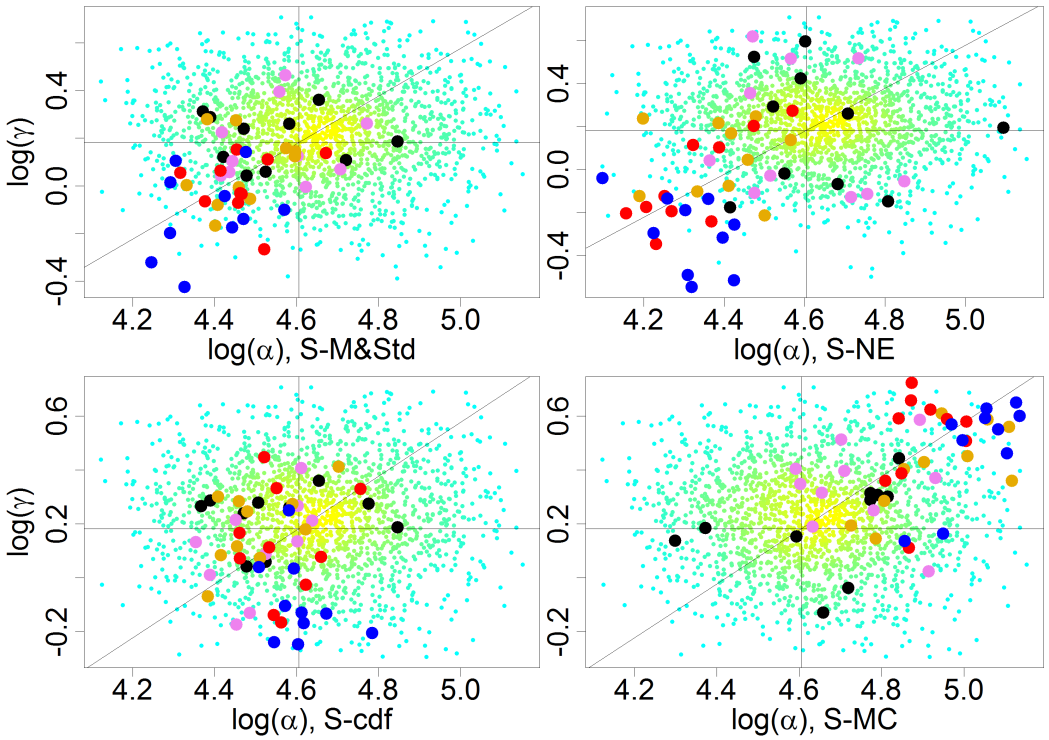


Figure A.61: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for α versus γ for the two-stage model (simulation 12).

Description of figure A.61: Compared to the first posterior distribution, the improvement of the estimation for both rates is low. For S-cdf, the last posterior estimates α well.

Description of figure A.62: For S-M&Std, S-cdf and S-MC, the posterior distributions are along the ratio of the kinetic rates. For S-M&Std and S-cdf it seems to be closely to the true rate values. But as the first posterior (black points) is covered entirely, the posterior distributions are as good as or slightly worse than the first one. This is confirmed by considering the evolution of the SNAR ratio along the populations (This is not shown).

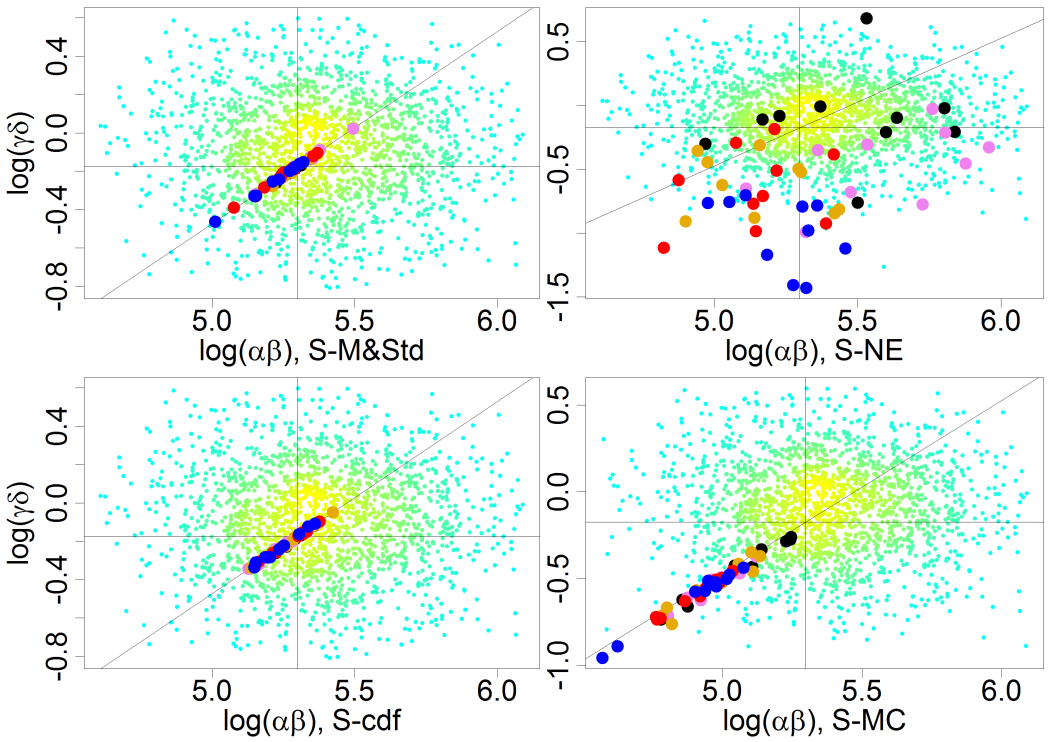


Figure A.62: Prior of the first population (cyan, yellow) and the posterior distribution for each of the five populations (black, violet, orange, red, blue) for the two-stage model (simulation 12) for $\alpha\beta$ versus $\gamma\delta$.

A.11 Outline of the relation between the threshold and the acceptance rate

For a reaction rate θ, one can simulate the data $x^*(\theta)$. Either $x^*(\theta)$ is deterministic, i.e. the data $x^*(\theta)$ being identical for identical θ. For the case where $x^*(\theta)$ is simulated using SSA, this does not hold, as for the same θ different results for $x^*(\theta)$ might be obtained due to the stochastic character of the SSA. But if the number of simulated trajectories N_{traj} approaches ∞, we assume that for each θ the same $x^*(\theta)$ results (this needs to be proofed).

Therefore, D is a random variable measuring the distance

$$\begin{aligned} D : \Theta &\to \mathbb{R} \\ \theta &\mapsto d(x^0, x^*(\theta)) \end{aligned}$$

with Θ as the set of all particles θ.

D is distributed according to F_D. Its quantile Q_τ gives the value for which $P(D < Q_\tau) = \tau$ with $P(\cdot)$ being the probability measure. That means, with probability τ the distance D is less than Q_τ.

Substituting Q_τ by ϵ_τ means that the distance D is smaller than (the threshold) ϵ_τ with probability τ. Therefore, it results in the same distance values, if either ϵ_τ is set and the distances $D < Q_\tau$ are taken or if τ is set and a corresponding ϵ_τ is determined so that $D < \epsilon_\tau$.

This holds only for a distribution F_D. But if we draw N realizations from the distribution F_D, we obtain the resulting frequency distribution $\hat{F_D}$, and it holds that

$$\hat{F_D} \to F_D \text{ for } N \to \infty.$$

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 07.02.2012

(Jan Mathias Köhler)