Parameter Estimation for Reaction Rate Equation Constrained Mixture Models

Carolin Loos[1], Anna Fiedler[1], and Jan Hasenauer[1]

[1]Helmholtz Zentrum München-German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany, and Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany

Abstract

The elucidation of sources of heterogeneity in cell populations is crucial to fully understand biological processes. A suitable method to identify causes of heterogeneity is reaction rate equation (RRE) constrained mixture modeling, which enables the analysis of subpopulation structures and dynamics. These mixture models are calibrated using single cell snapshot data to estimate model parameters which are not measured or which cannot be assessed experimentally. In this manuscript, we evaluate different optimization methods for estimating the parameters of RRE constrained mixture models under the normal distribution assumption. We compare gradient-based optimization using sensitivity analysis with two other optimization methods – gradient-based optimization with finite differences and a stochastic optimization method – for simulation examples with artificial data. Furthermore, we compare different numerical schemes for the evaluation of the log-likelihood function. We found that gradient-based optimization using sensitivity analysis outperforms the other optimization methods in terms of convergence and computation time.

1 Introduction

In the past years, methods for studying biological processes on a single cell level have been developed and improved. It is possible to quantify the (relative) abundance of molecular species in single cells using, e.g. flow cytometry [2] or single cell microscopy [11]. With these techniques, it is possible to also detect heterogeneity in expression for cells of a same cell population. This heterogeneity has been shown to play an important role for e.g. cancer cells or neurons [10, 14]. For homogeneous cell populations, dynamic mathematical models are convenient tools to study biological systems [8]. However, they only capture the dynamic of the mean response in the cell population and cannot account for possible subpopulations. To exploit the information available in single cell data, dynamical models that are able to account for subpopulation structures of the cells are needed.

A suitable method to study subpopulation structures of heterogeneous cell populations is the method of RRE constrained mixture modeling introduced by Hasenauer et al. [5]. These models can in principle be fitted to experimental single cell data to estimate unknown parameters of the biological system, such as kinetic rates, initial conditions or subpopulation weights. Subsequently, hypotheses about mechanistic differences between individual subpopulations can be tested. However, it has not yet been discussed how

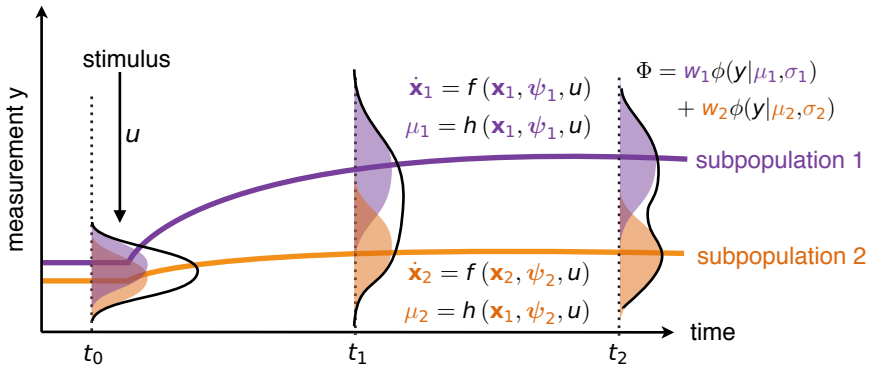

Figure 1: Illustration of RRE constrained mixture modeling for an example of two subpopulations. The means of measurement y for the individual subpopulations are calculated with RREs and plotted as purple and orange lines for the high and low responsive subpopulation, respectively. The overall cell distribution Φ is plotted as black curve and is calculated by a weighted mixture of the individual distributions for the subpopulations (purple and orange areas).

the parameters of RRE constrained mixture models can be estimated in an efficient and accurate way and there is no comparison of methods available.

In this manuscript, we consider maximum likelihood methods for parameter estimation. For this, a likelihood function which provides a measure of how well the data is explained by the current parametrization of the model is maximized. This maximization can be performed using e.g. local deterministic or global stochastic optimization techniques [3, 12, 15]. Most deterministic optimizers employ information about the gradient of the likelihood function. This gradient with respect to the parameters can be approximated by finite differences or, if possible, calculated with sensitivity analysis [12, 13]. An example of a global stochastic optimizer is particle swarm optimization presented in [15]. This optimizer does not rely on information about the gradient and has been shown to outperform other global optimizers [15].

We describe the concept of RRE constrained mixture models and provide the likelihood function and the sensitivity equations for the calculation of its gradient with respect to the parameters. Additionally, we explain the standard and a robust approach for the evaluation of a mixture likelihood. We compare the deterministic optimization using sensitivities to the deterministic method using finite differences and to the stochastic particle swarm optimization algorithm for artificial single cell snapshot data of a one stage and three stage cascade.

2 Methods

In this section, we outline the method of RRE constrained mixture modeling for single cell snapshot data and the corresponding likelihood formulation for the parameter estimation. We establish the gradient of the likelihood with respect to the model parameters and the sensitivity equations. Further, a numerically robust evaluation of the log-likelihood is presented.

2.1 RRE Constrained Mixture Models

RRE modeling provides the temporal evolution for the mean concentrations $\vec{x} = (x_1, \ldots, x_{n_x})$ of n_x chemical species involved in a biological process, which is stimulated by an external stimulus u. These RREs can be written as

$$\dot{\vec{x}} = f(\vec{x}, \boldsymbol{\psi}, u)\,, \qquad \vec{x}(0) = \vec{x}_0(\boldsymbol{\psi}, u)\,, \tag{1}$$

an ODE system with initial conditions $\vec{x}_0(\boldsymbol{\psi}, u)$ and vector field f. The parameter vector $\boldsymbol{\psi}$ comprises e.g. kinetic rates, initial concentrations or observation parameters. Often, the concentrations $\vec{x}$ of the species cannot be measured directly or only a subset of them can be observed. In most experiments, only a single property is assessed. Therefore, we considered an observable

$$y = h(\vec{x}, \boldsymbol{\psi}, u)\,,$$

with h denoting the mapping. The observation process depends on observation parameters included in $\boldsymbol{\psi}$ such as scaling and offset constants.

Mixture models enable the depiction of subpopulations within an overall population. The probability distribution is described by the weighted sum of probability density functions ϕ for individual mixture components, i.e., subpopulations

$$p(y|w_s, \mu_s, \sigma_s) = \sum_{s=1}^{n_s} w_s \phi(y|\mu_s, \sigma_s^2)\,.$$

In this manuscript, we assumed ϕ to be a normal distribution, which is parametrized by its mean μ and variance σ^2.

Combining these, every subpopulation is treated as a mixture component for which the mean concentration is simulated using RREs [5]. This yields the following model for the distribution of an observable y for some given parameters $\boldsymbol{\theta}$ at a time point t_k,

$$\begin{aligned} p(y|\boldsymbol{\theta}, t_k) &= \sum_{s=1}^{n_s} w_s(\boldsymbol{\theta})\, \phi\left(y|\mu_s, \sigma_s^2(\boldsymbol{\theta}, t_k)\right) \\ \text{with } \dot{\vec{x}}_s &= f\left(\vec{x}_s, \boldsymbol{\psi}_s(\boldsymbol{\theta}), u\right),\ \vec{x}_s(0) = \vec{x}_0(\boldsymbol{\psi}_s(\boldsymbol{\theta}), u)\,, \\ \mu_s &= h\left(\vec{x}_s, \boldsymbol{\psi}_s(\boldsymbol{\theta}), u\right)\,. \end{aligned}$$

The parameter vector can comprise e.g. $\boldsymbol{\theta} = (\{w_s, \sigma_s, \boldsymbol{\xi}_s\}_{s=1}^{n_s}, \boldsymbol{\xi})$, the subpopulation specific mixture weights w_s, standard deviations σ_s and mechanistic parameters $\boldsymbol{\xi}_s$ as well as mechanistic parameters $\boldsymbol{\xi}$ that are shared across subpopulations. The mean of the mixture distribution is linked to the RREs, while the mixture weights and standard deviations do not depend on the RREs. The parameters for the RREs of an individual subpopulation as defined in (1) are thus given by $\boldsymbol{\psi}_s = (\boldsymbol{\xi}_s, \boldsymbol{\xi})$. The concept of RRE constrained mixture models is illustrated in Figure 1. For a more detailed explanation of these models, we refer to [5].

2.2 Single Cell Snapshot Data

We considered single cell snapshot data

$$\mathcal{D} = \left\{ \left\{ y_j^k \right\}_{j=1}^{n_c} \right\}_{k=1}^{n_t}\,.$$

These data contain the measurements y for n_c cells, indexed by j, at n_t time points, indexed by k. In the case considered, the data captures the dynamics of the population on a single cell level after stimulation with some input u.

2.3 Parameter Estimation for RRE Constrained Mixture of Normal Distributions

To obtain the parameters of a RRE constrained mixture model, the model needs to be fitted to experimental data $\mathcal{D}$. This is done by maximum likelihood estimation. A likelihood function $\mathcal{L}(\boldsymbol{\theta})$ describes the probability of observing the data $\mathcal{D}$ given the parameters $\boldsymbol{\theta}$. For the case of RRE constrained mixture models, this function is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &:= \prod_{k,j} \sum_{s=1}^{n_s} w_s(\boldsymbol{\theta})\, \phi\left(y_j^k | \mu_s, \sigma_s^2(\boldsymbol{\theta}, t_k)\right) \\ \text{with } \dot{\vec{x}}_s &= f\left(\vec{x}_s, \boldsymbol{\psi}_s(\boldsymbol{\theta}), u\right),\ \vec{x}_s(0) = \vec{x}_0(\boldsymbol{\psi}_s(\boldsymbol{\theta}), u)\,, \\ \mu_s &= h\left(\vec{x}_s, \boldsymbol{\psi}_s(\boldsymbol{\theta}), u\right).\end{aligned}$$

The mixture parameters μ_s implicitly depend on the parameter vector $\boldsymbol{\theta}$. A different variance parameter σ_s can be used for every measured time point t_k and subpopulation s. Since the number of parameters increases with the number of measured time points and the number subpopulations, an efficient method for parameter estimation is required. Due to its better numerical properties, we used the negative log-likelihood function

$$\begin{aligned}J(\boldsymbol{\theta}) &= -\log \mathcal{L}\left(\boldsymbol{\theta}\right) \\ &= -\sum_{k,j} \log \sum_{s=1}^{n_s} w_s(\boldsymbol{\theta})\, \phi\left(y_j^k | \mu_s, \sigma_s^2(\boldsymbol{\theta}, t_k)\right)\end{aligned}$$

in the optimization, which has the same extrema as the likelihood function. In the following, we derive the gradient of J with respect to $\boldsymbol{\theta}$, which can be employed by deterministic local optimization methods.

2.3.1 Gradient of Negative Log-likelihood Function.

For a simpler notation, we neglect the arguments of w_s and σ_s. The gradient of the log-likelihood with respect to parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{n_\theta})$, with θ denoting an entry of the vector, is given by

$$\begin{aligned}\frac{dJ}{d\theta} &= -\sum_{k,j} \frac{d}{d\theta} \log\left(\sum_{s=1}^{n_s} w_s\, \phi\left(y_j^k | \mu_s, \sigma_s^2\right)\right) \\ &= -\sum_{k,j} \frac{1}{\sum_{s=1}^{n_s} w_s\, \phi\left(y_j^k | \mu_s, \sigma_s^2\right)} \frac{d}{d\theta} \sum_{s=1}^{n_s} w_s\, \phi\left(y_j^k | \mu_s, \sigma_s^2\right) \\ &= -\sum_{k,j} \frac{1}{\sum_{s=1}^{n_s} w_s\, \phi\left(y_j^k | \mu_s, \sigma_s^2\right)} \sum_{s=1}^{n_s} \left(\frac{dw_s}{d\theta}\, \phi\left(y_j^k | \mu_s, \sigma_s^2\right) + w_s \frac{d\phi\left(y_j^k | \mu_s, \sigma_s^2\right)}{d\theta}\right).\end{aligned}$$

Under the assumption that ϕ is a normal distribution, it holds that

$$\frac{d\phi\left(y_j^k | \mu_s, \sigma_s^2\right)}{d\theta} = \frac{1}{\sigma_s} \phi\left(y_j^k | \mu_s, \sigma_s^2\right) \left(\frac{y_j^k - \mu_s}{\sigma_s} \frac{d\mu_s}{d\theta} + \left(\left(\frac{y_j^k - \mu_s}{\sigma_s}\right)^2 - 1\right) \frac{d\sigma_s}{d\theta}\right),$$

with

$$\frac{d\sigma_s^k}{d\theta} = \begin{cases} 1 & \theta = \sigma_s^k \\ 0 & \text{otherwise} \end{cases}, \qquad \frac{dw_s}{d\theta} = \begin{cases} 1 & \theta = w_s \\ 0 & \text{otherwise} \end{cases}.$$

The gradient of the objective function comprises $\frac{d\mu_s}{d\theta}$, which can be calculated using sensitivity analysis. The sensitivities $z^{\vec{x}_s} = \left(\frac{\partial x_{s,1}}{\partial\theta}, \ldots, \frac{\partial x_{s,n_x}}{\partial\theta}\right)$ are defined by

$$\begin{aligned} \frac{\partial z^{\vec{x}_s}}{\partial t} &= \frac{\partial f}{\partial \vec{x}_s} z^{\vec{x}_s} + \frac{\partial \vec{x}_s}{\partial \theta}, \quad z^{\vec{x}_s}(0) = \frac{\partial \vec{x}_0}{\partial \theta}, \\ z^{\mu_s} &= \frac{\partial h}{\partial \vec{x}_s} z^{\vec{x}_s} + \frac{\partial h}{\partial \theta}, \end{aligned}$$

with $\frac{\partial f}{\partial \vec{x}_s} = \left(\frac{\partial f_m}{\partial x_{s,l}}\right)_{m,l} \in \mathbb{R}^{n_x \times n_x}$ and $\frac{\partial h}{\partial \vec{x}_s} = \left(\frac{\partial h_m}{\partial x_{s,l}}\right)_{m,l} \in \mathbb{R}^{n_x \times n_y}$. For the case of RRE constrained mixture models, we obtain μ_s and $\frac{d\mu_s}{d\theta} = z^{\mu_s}$ by simulating an ODE system comprising the RREs and sensitivity equations.

2.3.2 Robust Evaluation of the Log-Likelihood Function and Its Gradient.

We explain and tackle the problem occuring when numerically evaluating (log-) likelihood functions of mixture distributions. For this, we formulate the standard and robust approach to evaluate the log-likelihood function following [9]. As already mentioned, rather the log-likelihood than the likelihood function is calculated due to numerical properties. This means, instead of the probability density p, the logarithm $\log(p)$ is evaluated. For the assumption of a normal distribution this circumvents e.g. exponentiation of the difference between measurement and simulation. This is especially advantageous for high differences, since e^{-x} might be numerically evaluated to zero for finite values of x. However, for mixture models, if $n_s > 1$ and $p_s := \phi(y|\mu_s, \sigma_s^2)$, it holds that

$$\log(p) = \log\left(\sum_{s=1}^{n_s} w_s p_s\right) \neq \sum_{s=1}^{n_s} \log\left(w_s p_s\right),$$

i.e., for these cases it is not possible to use the logarithm of the probability density of an individual mixture component directly. This problem also occurs in the calculation of the gradient. We refer to this approach of evaluating the likelihood function as standard approach.

A more robust approach for the log-likelihood calculation is given in the following. With $q_s = \log(p_s)$ and $\hat{s} = \text{argmax}_s q_s$, we reformulate

$$\begin{aligned} \log(p) &= \log\left(\sum_{s=1}^{n_s} w_s e^{q_s}\right) \\ &= \log\left(1 + \sum_{s \neq \hat{s}} \frac{w_s}{w_{\hat{s}}} \left(e^{q_s - q_{\hat{s}}}\right)\right) + \log(w_{\hat{s}}) + q_{\hat{s}}. \end{aligned} \tag{2}$$

Considering p_s to be a normal distribution it follows that

$$log(p_s) = q_s = -\frac{1}{2}\left(\frac{y - \mu_s}{\sigma_s}\right)^2 - \log(\sqrt{2\pi}) - \log(\sigma_s).$$

Regarding the calculation of the gradient it holds that

$$\begin{aligned}\frac{d\log(p)}{d\theta} &= \frac{1}{p}\frac{dp}{d\theta} = \sum_{s=1}^{n_s} \frac{p_s}{\sum_{j=1}^{n_s} w_j p_j} H_s \\ &= \frac{1}{\sum_{j=1}^{n_s} w_j e^{q_j - q_{\hat{s}}}} \sum_{s=1}^{n_s} e^{q_s - q_{\hat{s}}} H_s \,, \end{aligned} \tag{3}$$

with H_s defined by

$$H_s = \frac{1}{p_s}\frac{dw_s p_s}{d\theta} = \frac{dw_s}{d\theta} + \frac{w_s}{p_s}\frac{dp_s}{d\theta} \,.$$

Under the assumption that p_s is a normal distribution this is

$$H_s = \frac{dw_s}{d\theta} + \frac{w_s}{\sigma_s}\left(\frac{y-\mu_s}{\sigma_s}\frac{d\mu_s}{d\theta} + \left(\left(\frac{y-\mu_s}{\sigma_s}\right)^2 - 1\right)\frac{d\sigma_s}{d\theta}\right).$$

The proposed reformulations (2) and (3) are used for the robust evaluation of the log-likelihood function and its gradient. For further details we refer to [9].

2.4 Implementation

The RRE constrained mixture models were implemented in MATLAB. The sensitivity equations were derived and simulated using the toolbox CERENA [7]. For parameter estimation with deterministic optimization, we used the toolbox PESTO,[1] which employs the MATLAB function `fmincon`. For stochastic global optimization we employed a toolbox for the algorithm PSwarm [15].

3 Results

We compared the different optimizers in terms of convergence and computation time for artificial data of a one stage and a three stage cascade.

3.1 One Stage Cascade

For a first comparison of the optimizers we considered a small example of a one stage cascade comprising a conversion between two species A and B.

3.1.1 Model and Artificial Data.

A conversion process describes a reversible reaction between two species, A and B that have the concentrations $[A]$ and $[B]$, respectively. In our example, we assumed that the conversion from A to B takes place with a basal rate $k_2[A]$ and is additionally increased by external stimulus u. Furthermore, B is converted

[1] available at `https://github.com/ICB-DCM/PESTO`

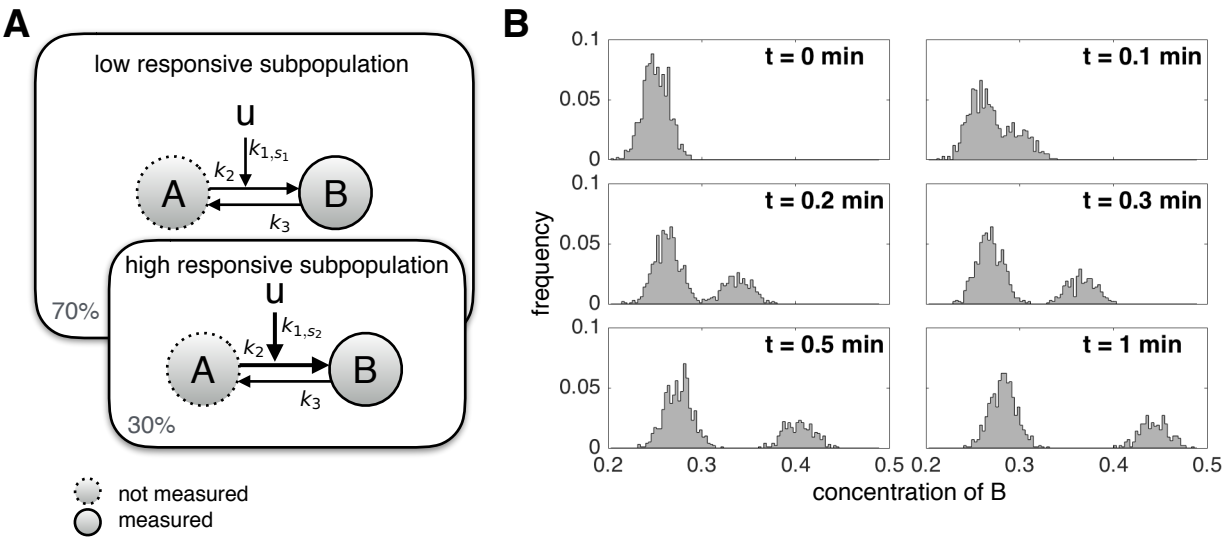

Figure 2: Artificial data of a conversion process. (A) Illustration of a conversion process between chemical species A and B in a cell population. The conversion from A to B is enhanced by a stimulus u. 30% of the cells show a higher response to the external stimulus u than the other cells. Only the concentration of B denoted by $[B]$ is measured. **(B)** Artificial data for the conversion process. The system is stimulated with $u = 0$ for $t < 0$ and $u = 1$ for $t \geq 0$.

back to A with kinetic parameter k_3 yielding the reactions

$$\begin{aligned} \mathrm{R}_1: &\quad \mathrm{A} \to \mathrm{B}, \quad \text{rate} = k_1 u\big[\mathrm{A}\big]\,, \\ \mathrm{R}_2: &\quad \mathrm{A} \to \mathrm{B}, \quad \text{rate} = k_2\big[\mathrm{A}\big]\,, \\ \mathrm{R}_3: &\quad \mathrm{B} \to \mathrm{A}, \quad \text{rate} = k_3\big[\mathrm{B}\big]\,. \end{aligned}$$

We considered that there exist two subpopulations, s_1 and s_2, differing in the stimulus-dependent conversion from A to B. This is described by the kinetic parameter k_1, i.e., the subpopulations share the parameters k_2 and k_3 but have individual parameters k_{1,s_1} and k_{1,s_2} with s_1 and s_2 indicating the kinetic parameters of subpopulation 1 and 2, respectively. The system is in steady state before stimulation ($u = 0$ for $t < 0$). To generate the artificial data we used the parameters $(k_{1,s_1}, k_{1,s_2}, k_2, k_3, w) = (0.1, 0.75, 0.5, 1.5, 0.7)$ and assumed that only the concentration of species B can be measured, yielding the observation model $y = h(\vec{x}, \boldsymbol{\psi}, u) = x_2$, with $\vec{x} = (x_1, x_2)^T = ([A], [B])^T$. An illustration of the system including the subpopulations is given in Figure 2A. This system was simulated using the stochastic simulation algorithm [4], which models random births and deaths of individual molecules. We considered a system size of $\Omega = 1000$ and divided the number of molecules by Ω to obtain the concentration of the species. Moreover, the external stimulus is set to $u = 1$ at $t \geq 0$ and measurements of the concentration of B are recorded at $t = 0, 0.1, 0.2, 0.3, 0.5, 1$ minutes. The data are shown in Figure 2B: For $t = 0$, the system is in steady state and no subpopulation structure is visible, since the subpopulations differ only in the response to stimulation. For $t = 0.1$, the subpopulation structure becomes visible, but the subpopulations still highly overlap. However, for later time points the subpopulations are clearly separated.

The mean of the stochastic single cell trajectories can be described by RREs, i.e., the temporal evolution of x_2 can be described by the ODE

$$\dot{x}_2 = k_1 u + k_2 - (k_1 u + k_2 + k_3)\, x_2\,, \qquad x_2(0) = \frac{k_2}{k_2 + k_3}\,,$$

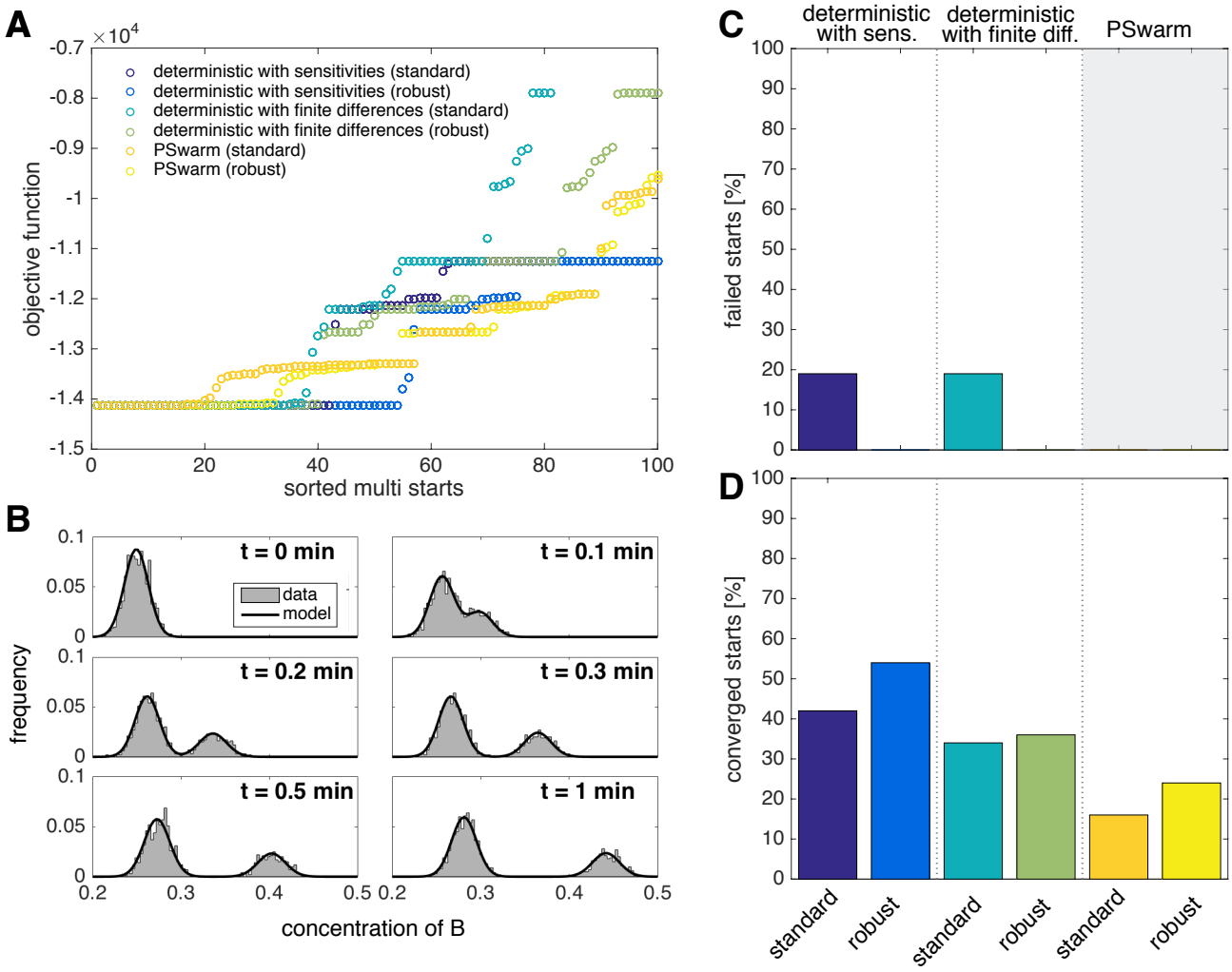

Figure 3: Comparison of optimization methods. **(A)** Convergence plot for the final negative log-likelihood values for 100 starts. The values are sorted from lowest to highest implying a decreasing goodness of fit. **(B)** Data and fit for the optimal value, which was found by all methods. Percentage of starts for which the initial value was ∞ **(C)** and converged starts **(D)**.

using mass conservation, $[A]+[B] = 1$. We then assumed the parameters $\boldsymbol{\theta} = (k_{1,s_1}, k_{1,s_2}, k_2, k_3, w, \{\{\sigma_{s(t_k)}\}_{s=1}^{2}\}_{k=1}^{6})$ to be unknown and estimated them from the data. Since the data comprised six time points and we accounted for two subpopulations, 12 parameters for the standard deviation $\sigma_s(t_k)$ need to be estimated.

3.1.2 Convergence of Optimization Methods.

To evaluate the optimizers, we compared deterministic gradient-based optimization using sensitivities with deterministic gradient-based optimization using finite differences and a stochastic particle swarm algorithm [15]. For all optimizers, the parameter values for the kinetic rates k_i were restricted to the interval $[10^{-6}, 10^4]$, the mixture weight w to $[0, 1]$ and the parameters for the standard deviation of the normal distributions $\sigma_s(t_k)$ to $[10^{-2.5}, 10^{2.5}]$. Each algorithm was started 100 times and the deterministic optimizers were started from the same randomly drawn start points.

The final negative log-likelihood values for every start are sorted with decreasing goodness of fit and shown in Figure 3A. The data and fit, which correspond to the optimal value found by all methods, are shown in Figure 3B. The model shows a good agreement with the data. For a detailed comparison of the results obtained by the different optimization methods, we assessed the percentage of failed starts, i.e., the

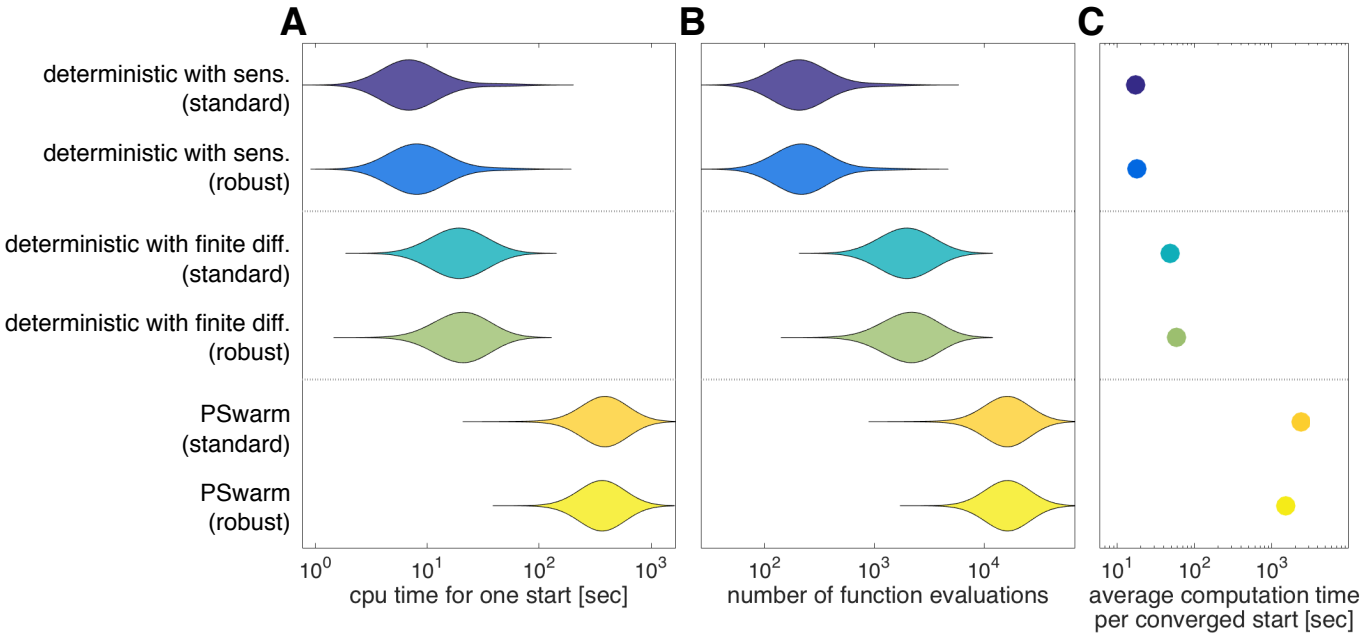

Figure 4: Performance comparison of optimization methods. (A) Time needed for one optimization start. **(B)** Number of objective function evaluations for one optimization start. **(C)** Average computation time needed per converged start.

starts for which the log-likelihood function was infinite at the start point (Figure 3C). For almost 20% of all drawn start points the log-likelihood has an infinite value when using the standard evaluation of the log-likelihood. However, the log-likelihood can be evaluated for all start points when using the robust calculation approach. Since for PSwarm an initial particle population is used instead of a single initial value, there are no failed starts and it is not possible to compare this property with the deterministic optimizers. We expect the percentage of failed log-likelihood evaluations for the initial particle population to be similar to the percentages found for the failed starts in the deterministic optimization using the standard approach. The likelihood was numerically evaluated to zero for all start points. For the log-likelihood, we counted the number of objective function values that are close to the minimal objective function value found, i.e., below a statistical threshold according to a likelihood ratio test [6]. These starts are then likely to have converged to the global optimum. The percentage of converged starts determined for each optimizer is depicted in Figure 3D. Clearly, the best convergence is obtained by deterministic local optimization with an analytical gradient that is calculated with sensitivities. For this optimizer, the robust calculation of the log-likelihood and the gradient yielded better convergence compared to the standard approach. For both approaches, standard and robust evaluation of the log-likelihood function, deterministic local optimiziation with finite difference approximation to the gradient shows less convergence than when using sensitivites. The stochastic optimization with PSwarm has even less converged runs than the deterministic optimization with finite differences.

3.1.3 Computation Time of Optimization Methods.

We compared the performance of the optimizers in terms of computation time (Figure 4A). The best computation time was achieved for the deterministic optimization with sensitivities, while the highest computation time is needed for stochastic optimization. Also regarding the number of function evaluations, the stochastic optimization needed most function evaluation and the deterministic optimization with sensitivities performed best (Figure 4B). Furthermore, regarding the average computation time needed per converged start shown in

Figure 4C, the deterministic optimizer using sensitivities outperforms the other optimizers. However, there were almost no additional computational costs when using the robust approach instead of the standard approach to evaluate the log-likelihood function for all optimizers.

3.2 Three Stage Cascade

To validate the results obtained for the simple conversion process, we studied artificial data of a three stage cascade, namely the Raf/Mek/Erk cascade.

3.2.1 Model and Artificial Data.

The considered pathway comprises the protein kinases Raf, Mek and Erk and their corresponding phosphorylated/active forms pRaf, pMek and pErk. Raf is activated with a stimulus-dependent rate $k_1 u\big[\mathrm{Raf}\big]$ and a basal rate $k_2\big[\mathrm{Raf}\big]$. The activation rate of Mek is proportional to the amount of phosphorylated Raf, while active Mek in turn phosphorylates Erk. These reactions and the dephosphorylation of the active kinases are given by

$$
\begin{aligned}
\mathrm{R}_1: &\quad \mathrm{Raf} \to \mathrm{pRaf}, && \text{rate} = k_1 u\big[\mathrm{Raf}\big]\,,\\
\mathrm{R}_2: &\quad \mathrm{Raf} \to \mathrm{pRaf}, && \text{rate} = k_2\big[\mathrm{Raf}\big]\,,\\
\mathrm{R}_3: &\quad \mathrm{pRaf} \to \mathrm{Raf}, && \text{rate} = k_3\big[\mathrm{pRaf}\big]\,,\\
\mathrm{R}_4: &\quad \mathrm{Mek} \to \mathrm{pMek}, && \text{rate} = k_4\big[\mathrm{pRaf}\big]\big[\mathrm{Mek}\big]\,,\\
\mathrm{R}_5: &\quad \mathrm{pMek} \to \mathrm{Mek}, && \text{rate} = k_5\big[\mathrm{pMek}\big]\,,\\
\mathrm{R}_6: &\quad \mathrm{Erk} \to \mathrm{pErk}, && \text{rate} = k_6\big[\mathrm{pMek}\big]\big[\mathrm{Erk}\big]\,,\\
\mathrm{R}_7: &\quad \mathrm{pErk} \to \mathrm{Erk}, && \text{rate} = k_7\big[\mathrm{pErk}\big]\,,
\end{aligned}
$$

with mass conservation

$$
\begin{aligned}
\big[\mathrm{Raf}\big] + \big[\mathrm{pRaf}\big] &= \big[\mathrm{Raf}\big]_0\,,\\
\big[\mathrm{Mek}\big] + \big[\mathrm{pMek}\big] &= \big[\mathrm{Mek}\big]_0\,,\\
\big[\mathrm{Erk}\big] + \big[\mathrm{pErk}\big] &= \big[\mathrm{Erk}\big]_0\,.
\end{aligned}
$$

For the data generation, we assumed to observe $y = h(\vec{x}, \boldsymbol{\psi}, u) = s\big[\mathrm{pErk}\big]$. To circumvent structural non-identifiabilities, we consider the reformulations

$$
\begin{aligned}
x_1 &= k_4\big[\mathrm{pRaf}\big]\,,\\
x_2 &= k_6\big[\mathrm{pMek}\big]\,,\\
x_3 &= s\big[\mathrm{pErk}\big]\,.
\end{aligned}
$$

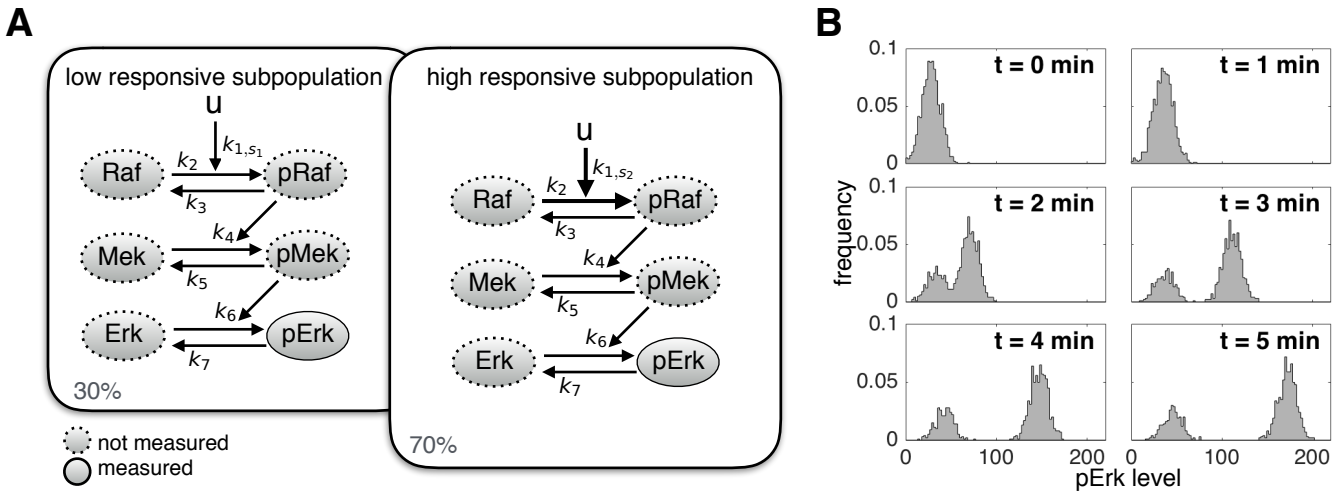

Figure 5: Artificial data of the Raf/Mek/Erk cascade. (A) Illustration of the considered signaling pathway, which comprises the kinases Raf, Mek and Erk and its corresponding actived forms. The model comprises two subpopulations differing in their response to stimulus u. **(B)** Artificially generated data of the Raf/Mek/Erk cascade for measurements of pErk levels.

This yields the ODE system

$$\dot{x}_1 = (k_1 u + k_2)(k_4 \left[\text{Raf}\right]_0 - x_1) - k_3 x_1 \,, \qquad x_1(0) = \frac{k_2 k_4 \left[\text{Raf}\right]_0}{k_3 + k_2} \,,$$

$$\dot{x}_2 = x_1(k_6 \left[\text{Mek}\right]_0 - x_2) - k_5 x_2 \,, \qquad x_2(0) = \frac{x_1(0) k_6 \left[\text{Mek}\right]_0}{x_1(0) + k_5} \,,$$

$$\dot{x}_3 = x_2(s \left[\text{Erk}\right]_0 - x_3) - k_7 x_3 \,, \qquad x_3(0) = \frac{x_2(0) s \left[\text{Erk}\right]_0}{x_2(0) + k_7} \,,$$

with $y = x_3$ and parameters $(k_1, k_2, k_3, k_5, k_7, k_4 \left[\text{Raf}\right]_0, k_6 \left[\text{Mek}\right]_0, s \left[\text{Erk}\right]_0)$. For details regarding the model we refer to [5]. In this example, we considered two subpopulations that differ in their response to stimulus u, captured by parameter k_1 (Figure 5A). We generated measurements of 1000 cells by simulating the ODE system for $\log_{10}(k_{1,s_1}, k_{1,s_2}, k_2, k_3, k_5, k_7, k_4 \left[\text{Raf}\right]_0, k_6 \left[\text{Mek}\right]_0, s \left[\text{Erk}\right]_0)$ = (-2, -1, -2, -0.15, -0.15, -0.15, -2, 2, 3), $w = 0.7$ and normally-distributed measurement noise (Figure 5B). The stimulus u is set to 0 for $t < 0$ and to 1 for $t \geq 0$.

3.2.2 Convergence of Optimization Methods.

For parameter estimation, the intervals for the parameters were set to $[10^{-3}, 10^5]$ for the kinetic parameters, to $[0, 1]$ for the mixture weight and to $[10^{-3}, 10^2]$ for $\sigma_s(t_k)$. The resulting objective function values for 100 runs of the optimization procedures are shown in Figure 6A, and a zoom in of the five best runs in Figure 6B. The optimization with sensitivities and a robust evaluation of the log-likelihood function converged to the optimal value 44 times. This optimal value yields a good fit to the data (Figure 6C). Using deterministic optimization with sensitivities and the standard evaluation of the log-likelihood function the same optimal value as with the robust evaluation was found only once. The other optimizers were not able to find the optimal value at all. For the deterministic optimization and the standard evaluation of the log-likelihood function, only three out of 100 initial parameter values yielded a finite log-likelihood value. Consequently, the remaining runs could not be started. These findings indicate that for higher-dimensional estimation problems, the use of sensitivity-based methods and robust log-likelihood evaluation becomes increasingly

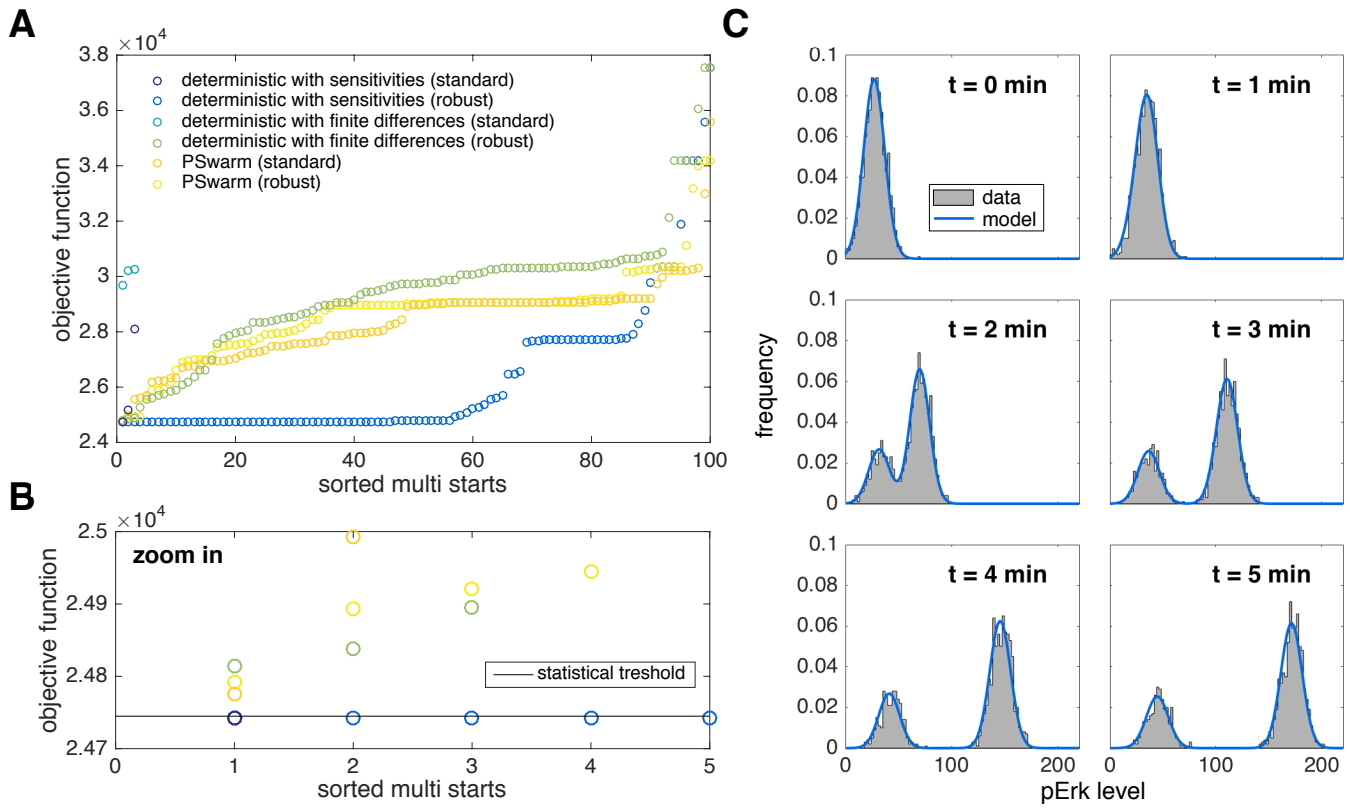

Figure 6: Comparison of optimization methods. **(A)** Final negative log-likelihood values for 100 runs, sorted according to a decreasing goodness of fit. **(B)** Zoom for the five best starts. The black line indicates the statistical threshold according to a likelihood ratio test, which was used to obtain the number of converged starts. **(C)** Data and fit for the optimal parameter value found by deterministic optimization with sensitivities and a robust evaluation of the log-likelihood function.

important.

3.2.3 Performance of Optimization Methods.

We compared the computation times and needed function evaluations of the different optimization methods (Figure 7). Since only the deterministic optimization with sensitivities and robust evaluation reached a sufficient number of converged starts, we did not compare the optimizers in terms of average computation time per converged starts. The analysis for the deterministic optimization with standard evaluations is only based on three starts that have not failed and is therefore not meaningful for the comparison. Among the optimizers for which 100 starts could be analyzed in terms of their computation time and number of function evaluations, the optimization with sensitivities and the robust evaluation of the log-likelihood function performs best. The proposed approach therefore yields better optimization results and is also more efficient than the other optimizers.

4 Conclusion

In this manuscript, we summarized the concept of RRE constrained mixture modeling and studied the calibration of those models to experimental data under the normal distribution assumption. An often used approach to estimate the parameters of mixture models in general is the Expectation-Maximization (EM) algorithm (see e.g. [1]). This algorithm highly depends on the initialization of the mixture components,

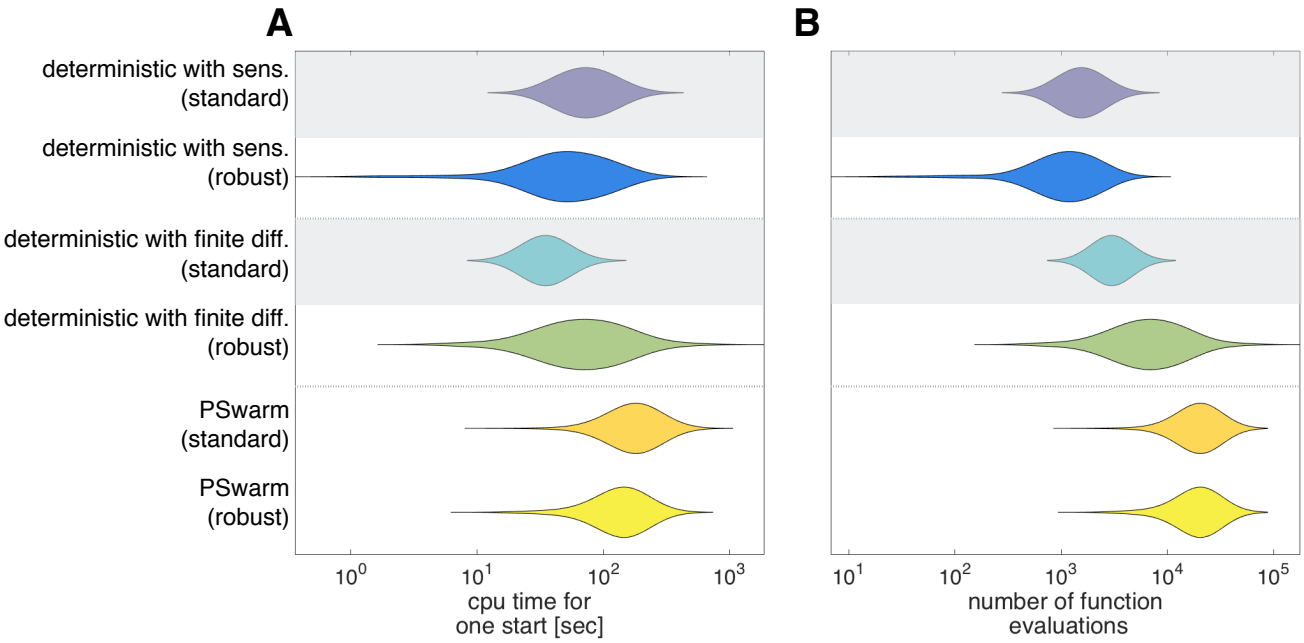

Figure 7: Performance of optimization methods. (A) CPU time needed for one optimization start. **(B)** Number of objective function evaluations for one optimization start. The representation is based on three starts for deterministic optimization with the standard approach to evaluate the log-likelihood (grey shaded), while it is based on 100 starts for the other optimizers.

which is challenging for RRE constrained mixture models since the components depend on the dynamic parameters of the model. In preliminary studies the EM algorithm showed poor convergence. Therefore, we did not consider the EM algorithm in this manuscript and focused on a maximum likelihood approach.

We derived the log-likelihood function and its gradient, which can be used to perform gradient-based deterministic optimization. Additionally, a robust approach of numerically evaluating these terms has been provided. We compared three optimization schemes, two deterministic gradient-based methods, one using the analytical gradient and one using an approximation of the gradient by finite differences, and a stochastic particle swarm algorithm. For each optimizer, we assessed performance and convergence for the standard and robust approach to evaluate the log-likelihood function. The comparison was carried out for examples of artificial single cell snapshot data of a one stage and a three stage cascade. We found that deterministic gradient-based optimization with sensitivities and robust calculation of the mixture probability outperformed all other methods in terms of robustness and convergence. This is especially important, since the complexity of RRE constrained mixture models increases with the number of measured time points. For the example of the three stage cascade only gradient-based optimization with sensitivites and a robust evaluation of the log-likelihood function yielded a reasonable calibration of RRE constrained mixture models to the data. We expect this also to hold when considering even more complicated systems. Accordingly, the proposed approach facilitates a robust and efficient calibration of RRE constrained mixture models to elucidate the sources of heterogeneity.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[2] H. M. Davey and D. B. Kell. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. *Microbiological Reviews*, 60(4):641–696, 1996.

[3] A. Gábor and J. R. Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Syst Biol*, 9:74, 2015.

[4] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[5] J. Hasenauer, C. Hasenauer, T. Hucho, and F. J. Theis. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. *PLoS Computational Biology*, 10(7):e1003686, 2014.

[6] S. Hross, F. J. Theis, and J. Hasenauer. Quantification of CFSE time-series data using division-, age- and label-structure population models. *Bioinformatics*, 2016.

[7] A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, and J. Hasenauer. CERENA: ChEmical REaction Network Analyzer - a toolbox for the simulation and analysis of stochastic chemical kinetics. *PloS CB*, 2016.

[8] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[9] C. Loos. *Analysis of single-cell data: ODE-constrained mixture modeling and approximate Bayesian computation*. Best Masters. Springer, 2016.

[10] F. Michor and K. Polyak. The origins and implications of intratumor heterogeneity. *Cancer Prevention Research*, 3(11):1361–1364, 2010.

[11] T. Miyashiro and M. Goulian. Single-cell analysis of gene expression by fluorescence microscopy. *Methods in Enzymology*, 423:458–475., 2007.

[12] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelker, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, et al. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335, 2013.

[13] B. Sengupta, K. Friston, and W. Penny. Efficient gradient computation for dynamical models. *NeuroImage*, 98:521–527, 2014.

[14] M.-E. Torres-Padilla and I. Chambers. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *Development*, 141(11):2173–2181, 2014.

[15] A. I. F. Vaz and L. N. Vicente. A particle swarm pattern search method for bound constrained global optimization. *Journal of Global Optimization*, 39(2):197–219, 2007.