Robust calibration of hierarchical population models for heterogeneous cell populations

Carolin Loos[1,2], and Jan Hasenauer[1,2,3,*]

[1]Helmholtz Zentrum München-German Research Center for Environmental Health,
Institute of Computational Biology, Neuherberg 85764, Germany

[2]Chair of Mathematical Modeling of Biological Systems, Center for Mathematics,
Technische Universität München, Garching 85748, Germany

[3]Faculty of Mathematics and Natural Sciences, University of Bonn, 53115 Bonn, Germany,

[*]jan.hasenauer@uni-bonn.de

December 19, 2019

Highlights

- Generalizes hierarchical population model to various distribution assumptions
- Provides framework for efficient calibration of the hierarchical population model
- Simulation study and application to experimental data reveal improved robustness and optimization performance

Abstract

Cellular heterogeneity is known to have important effects on signal processing and cellular decision making. To understand these processes, multiple classes of mathematical models have been introduced. The hierarchical population model builds a novel class which allows for the mechanistic description of heterogeneity and explicitly takes into account subpopulation structures. However, this model requires a parametric distribution assumption for the cell population and, so far, only the normal distribution has been employed. Here, we incorporate alternative distribution assumptions into the model, assess their robustness against outliers and evaluate their influence on the performance of model calibration in a simulation study and a real-world application example. We found that alternative distributions provide reliable parameter estimates even in the presence of outliers, and can in fact increase the convergence of model calibration.

Keywords: Single-cell data; Heterogeneity; Population model; Skew normal distribution; Student's t distribution; Mechanistic Modeling; Dynamic Modeling

1 Introduction

An important goal of systems biology is to obtain insights into the mechanisms and sources of cellular heterogeneity. This heterogeneity is critical for cellular decision making (Balázsi et al., 2011) and studied for various diseases and biological systems. To study heterogeneity, data at the single-cell level, e.g., time-lapse or snapshot data, are collected with experimental techniques such as fluorescent microscopy or flow cytometry.

To mechanistically study single-cell snapshot data, a variety of modeling approaches have been introduced. Ensemble models describe individual cells and model the overall population as a collection of many cells (Henson, 2003; Kuepfer et al., 2007). These models are often computationally expensive. To circumvent this, the cell population can be approximated by its statistical moments. Approximation approaches depend on assumptions about the contribution of intrinsic and extrinsic noise sources. While intrinsic noise is often defined as the stochasticity of gene expression, extrinsic noise is assumed to influence the reaction rates (Swain et al., 2002). Therefore, the statistical moments are obtained using the moment-closure approximation (Engblom, 2006), when intrinsic noise is assumed to be important, or Dirac-mixture approximations (Wang et al., 2019), when heterogeneity occurs mainly due to extrinsic noise. For the case of extrinsic noise, modeling approaches have been developed which infer the distribution of cellular properties using maximum entropy principles (Waldherr et al., 2009; Dixit et al., 2019). However, the aforementioned methods do not explicitly take into account subpopulation structures, which are omnipresent in heterogeneous cell populations (Altschuler and Wu, 2010). We recently introduced the hierarchical population model (Loos et al., 2018). This model describes subpopulation structures using mixture modeling, and ensures computational efficiency by employing approximations for statistical moments of the biological species of individual subpopulations. However, the model relies on a parametric assumption for the distribution of the subpopulations and only the multivariate normal distribution has been employed. This is a substantial limitation as the distribution of properties within subpopulations is often not normal (Pyne et al., 2009; Mar, 2019).

As the measured distribution reflects not only cellular heterogeneity, but also the variability of the measurement process, measurement noise and outliers might result in additional deviations from a normal distribution. In the analysis of ordinary differential equation (ODE) models, it has been shown that distributions with heavier tails than the normal distribution yield more robust parameter estimates in the presence of outliers (Maier et al., 2017). For single-cell snapshot data, the probability of observing outliers in the data is even higher due to the high number of data points (Pyne et al., 2009; Ilicic et al., 2016). Thus, heavy-tailed and skewed distributions are employed when studying flow cytometry (Pyne et al., 2009) or scRNA-seq data (Ding et al., 2019). However, the incorporation and assessment of distribution assumptions for the hierarchical population model is missing. Ideally, the employed distribution should not only provide reliable parameter estimates, but also yield a reasonable performance of model calibration.

In this manuscript, we describe the hierarchical population model and generalize it to different distribution assumptions. We derive the equations for the mathematical formulation of the model with alternative distribution assumptions. These equations, including the likelihood function and its gradient, are required to perform efficient model calibration using gradient-based maximum likelihood estimation. We analyze the influence of the distribution assumptions for simulated data of three biological motifs and three outlier scenarios. These outlier scenarios are motivated by experimental errors that can occur during data generation, e.g., due to a dropout event. Finally, we apply the hierarchical population model with the different distribution assumptions to analyze experimental data of NGF-induced Erk1/2 signaling.

2 Methods

2.1 Hierarchical population model

We consider single-cell snapshot data,

$$\mathcal{D} = \{(\bar{\mathbf{y}}_k^c, t_k, \mathbf{u})\}_{c,k} \ , \tag{1}$$

with indices c for the cell and k for the time point, stimulus vector $\mathbf{u} \in \mathbb{R}^{n_u}$ and a vector of measurements $\bar{\mathbf{y}}_k^c \in \mathbb{R}^{n_y}$ of the cell. The measured quantities might be, e.g., (relative) protein abundances. The measurement follows a distribution, which is a convolution of the measurement noise distribution and an additional distribution for a potential outlier-generating process. For a better readability, we neglected the index for varying stimuli.

In the hierarchical population model as introduced by Loos et al. (2018) the measured property of an individual cell is assumed to be distributed according to

$$\bar{\mathbf{y}} \sim \sum_s w_s \phi(\bar{\mathbf{y}}|\boldsymbol{\varphi}_s) \, , \tag{2}$$

with relative subpopulation sizes w_s of subpopulations $s = 1, \ldots, n_s$ and parametric distribution ϕ which depends on the vector of distribution parameters $\boldsymbol{\varphi}_s$ (Fig. 1A). The density ϕ captures biological variability due to intrinsic or extrinsic noise as well as measurement noise and the outlier distribution. The relative subpopulation sizes w_s are positive and sum up to one. The distribution parameters $\boldsymbol{\varphi}_s$ arise from the single-cell dynamics and models for cell-to-cell variability. Individual single-cell trajectories can be obtained using Markov jump processes (Gillespie, 1977), ordinary (Klipp et al., 2005) or stochastic differential equations (Gillespie, 2000). Instead of simulating the cells individually, which can become time-consuming, we compute the temporal evolution of the statistical moments, i.e., means $\mathbf{m}_s^x$ and covariances $\mathbf{C}_s^x$, of biochemical species $\mathbf{x}$. The temporal evolution is defined by a function g_z, e.g., obtained by moment-closure (Engblom, 2006), sigma-point (van der Merwe, 2004) or Dirac mixture approximations (Wang et al., 2019), depending on the assumptions about intrinsic and extrinsic noise. This yields

$$\dot{\mathbf{z}}_s = g_z\left(\mathbf{z}_s, \boldsymbol{\xi}_s, \mathbf{u}\right), \quad \mathbf{z}_s(0) = \mathbf{z}_0\left(\boldsymbol{\xi}_s, \mathbf{u}\right) \ , \tag{3}$$

with $\mathbf{z}_s = (\mathbf{m}_s^x, \mathbf{C}_s^x)^T$, initial conditions $\mathbf{z}_0$ and subpopulation parameters $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \boldsymbol{D}_s)$, with

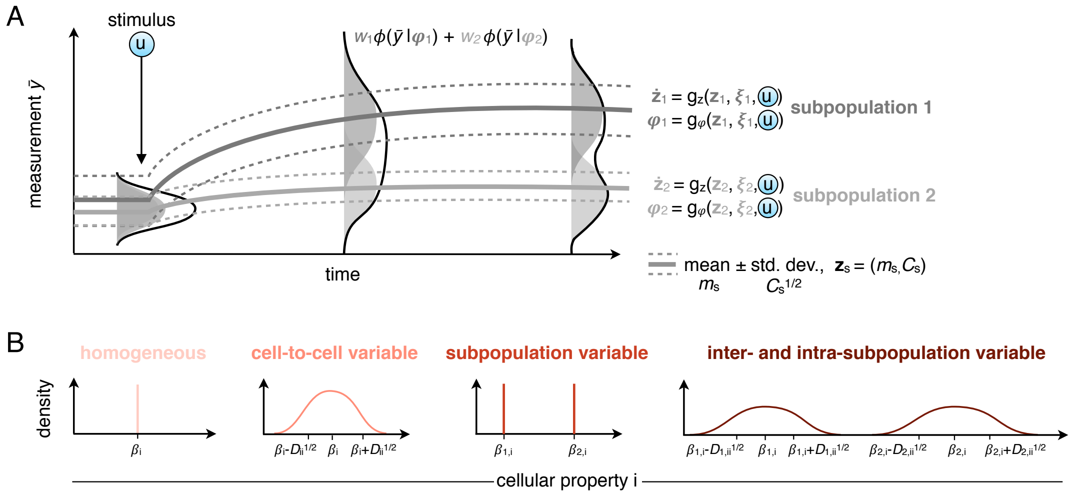

Figure 1: **Illustration of the hierarchical population model**. (A) The cell population comprises two heterogenenous subpopulation which respond differently to stimulation. Means and covariances of the species for each subpopulation are linked to a mixture distribution. The light and dark gray lines show the means of the subpopulations and the black line shows the distribution of the whole cell population. (B) Heterogeneity is captured by assuming each parameter/cellular property to be distributed according to one of the indicated cases.

$\boldsymbol{\beta}_s \in \mathbb{R}^{n_\beta}$ and $\boldsymbol{D}_s \in \mathbb{R}^{n_\beta \times n_\beta}$. These parameters are given by

$$\beta_{s,i} = \begin{cases} \beta_i \\ \beta_i \\ \beta_{s,i} \\ \beta_{s,i} \end{cases} \quad \text{and} \quad D_{s,ii} = \begin{cases} 0 & \text{homogeneous}\,, \\ D_{ii} & \text{cell-to-cell variable}\,, \\ 0 & \text{subpopulation variable}\,, \\ D_{s,ii} & \text{inter- and intra-subpopulation variable}\,. \end{cases} ,$$

in which $\beta_{s,i}$ denotes the ith element of $\boldsymbol{\beta}_s$, and $D_{s,ii}$ denotes the ith diagonal element of $\boldsymbol{D}_s$. The parameter $\beta_{s,i}$ encodes the mean of the cellular property, while the parameter $D_{s,ii}$ encodes its spread within a subpopulation. Homogeneous parameters are assumed to be the same for all cells of the whole cell population; cell-to-cell variable parameters differ between cells of the same subpopulation; subpopulation variable parameters differ between but not within a subpopulation; and inter- and intra-subpopulation variable parameters differ both between and within subpopulations (Fig. 1B).

The moments for a subpopulation are mapped to the distribution parameters,

$$\boldsymbol{\varphi}_s = g_\varphi\left(\mathbf{z}_s, \boldsymbol{\xi}_s, \mathbf{u}\right) , \tag{4}$$

of a distribution ϕ. Thus, g_φ encodes the mapping from the biochemical species to the observables, i.e., the measurable output of the system, as well as the mapping of the observables to the distribution parameters (see Appendix D.1 for an example). So far, the multivariate normal distribution has been employed,

$$\phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = \frac{1}{(2\pi)^{\frac{n_y}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{\mathbf{y}}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}}-\boldsymbol{\mu})} , \tag{5}$$

with distribution parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, comprising mean $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$ and covariance matrix $\boldsymbol{\Sigma} \in$

$\mathbb{R}^{n_y \times n_y}$. This yields

$$\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = g_\varphi((\mathbf{m}_s^x, \mathbf{C}_s^x), \boldsymbol{\xi}_s, \mathbf{u}) = (\mathbf{m}_s^y, \mathbf{C}_s^y + \boldsymbol{\Gamma}) , \tag{6}$$

including measurement noise $\boldsymbol{\Gamma}$, which is generally assumed to be the same for all subpopulations.

2.1.1 Calibration of the hierarchical population model

The parameters of the model, including relative subpopulation sizes w_s, subpopulation parameters $\boldsymbol{\xi}_s$ and parameters for the measurement noise need to be estimated from data. For this, we denote the overall parameter object by $\boldsymbol{\theta}$ (see Appendix D.1 for an example). The model is calibrated by maximizing the likelihood function,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{k,c} \sum_s w_s\left(t_k, \boldsymbol{\theta}, \mathbf{u}\right) \phi\left(\bar{\mathbf{y}}_k^c | \boldsymbol{\varphi}_s(t_k, \boldsymbol{\theta}, \mathbf{u})\right) \\ \text{with} \quad \dot{\mathbf{z}}_s &= g_z\left(\mathbf{z}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right), \quad \mathbf{z}_s(0) = \mathbf{z}_0\left(\boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right) , \\ \boldsymbol{\varphi}_s &= g_\varphi\left(\mathbf{z}_s, \boldsymbol{\xi}_s(\boldsymbol{\theta}), \mathbf{u}\right) . \end{aligned} \tag{7}$$

This can efficiently be done by multi-start local optimization employing the gradient of the likelihood function (Raue et al., 2013; Loos et al., 2016). In the following, we provide the likelihood functions for several distribution assumptions which can be incorporated into the hierarchical population model. For each distribution, we derive the function g_φ introduced in (4), which maps the mean and covariances of the species to the distribution parameters $\boldsymbol{\varphi}_s$.

2.2 Alternative distribution assumptions for the hierarchical population model

We considered two alternatives to the normal distribution: the skew normal and the Student's t distribution (Fig. 2) (see Appendix B for a third distribution, the negative binomial distribution). In the following, we discuss these distributions and provide the equations which are required to incorporate the distributions in the hierarchical population model.

2.2.1 Multivariate skew normal distribution

A challenge in the analysis of single-cell data is that the observed cell population is often skewed (Pyne et al., 2009). Therefore, distributions which account for skewness are often employed in the analysis of, e.g., flow cytometry data (Johnsson et al., 2016). The multivariate skew normal distribution has distribution parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$, with location $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$, covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n_y \times n_y}$ and skew parameter $\boldsymbol{\delta} \in \mathbb{R}^{n_y}$. The probability density function is

$$\phi_{\text{skewnorm}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = 2\phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\Omega})\Phi_{\text{norm}}(\boldsymbol{\alpha}(\bar{\mathbf{y}} - \boldsymbol{\mu})|0, 1) , \tag{8}$$

with $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\delta}\boldsymbol{\delta}^T$, $\boldsymbol{\alpha} = \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}/(1 - \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\delta})^{\frac{1}{2}} \in \mathbb{R}^{1 \times n_y}$ and Φ_{norm} denoting the cumulative distribution function of a univariate standard normal distribution. If $\boldsymbol{\delta} = \mathbf{0}$, the distribution equals a multivariate normal distribution. As provided by Pyne et al. (2009) and Sahu et al. (2003), the mean and covariance matrix of the multivariate skew normal distribution are

$$\begin{aligned} \mathbf{m} &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}\,, \\ \mathbf{C} &= \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\delta}\boldsymbol{\delta}^T\,. \end{aligned} \tag{9}$$

This yields for $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta}), \boldsymbol{\delta}(\boldsymbol{\theta}))$ the relation

$$\begin{aligned} \boldsymbol{\mu}_s(\boldsymbol{\theta}) &= \mathbf{m}_s^y(\boldsymbol{\theta}) - \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}(\boldsymbol{\theta})\,, \\ \boldsymbol{\Sigma}_s(\boldsymbol{\theta}) &= \mathbf{C}_s^y(\boldsymbol{\theta}) - \left(1 - \frac{2}{\pi}\right)\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T + \boldsymbol{\Gamma}(\boldsymbol{\theta})\,, \end{aligned} \tag{10}$$

with measurement noise matrix $\boldsymbol{\Gamma}$. The entries of the skew parameter vector $\boldsymbol{\delta}$ are allowed to differ in each dimension. However, they are restricted in a way that $\boldsymbol{\Sigma}_s$ needs to be positive definite. The derivatives of the probability density (8) and the distribution parameters (10) are provided in Appendix A.2.

2.2.2 Multivariate Student's t distribution

The Student's t distribution is often employed as a robust alternative to the normal distribution in regression (Lange et al., 1989), modeling of population-average data (Maier et al., 2017), and in the analysis of single-cell data (Lo et al., 2008; Pyne et al., 2009; Ding et al., 2019). The tails of the Student's t distribution are heavier than the tails of a normal distribution, and thus the distribution can better cope with outliers in the data. The multivariate Student's t distribution has distribution parameters $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with location $\boldsymbol{\mu} \in \mathbb{R}^{n_y}$, shape matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n_y \times n_y}$ and degree of freedom $\nu \in \mathbb{R}_+$. The probability density function reads

$$\phi_{\text{stud}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = \frac{\Gamma(\frac{\nu+n_y}{2})|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{n_y}{2}}\Gamma(\frac{\nu}{2})\left(1 + \frac{1}{\nu}(\bar{\mathbf{y}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})\right)^{\frac{\nu+n_y}{2}}}\,. \tag{11}$$

For $\nu > 2$, the mean and covariance matrix of the multivariate Student's t distribution are given by

$$\begin{aligned} \mathbf{m} &= \boldsymbol{\mu}\,, \\ \mathbf{C} &= \frac{\nu}{\nu - 2}\boldsymbol{\Sigma}\,. \end{aligned} \tag{12}$$

This yields $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\boldsymbol{\mu}_s(\boldsymbol{\theta}), \boldsymbol{\Sigma}_s(\boldsymbol{\theta}), \nu(\boldsymbol{\theta}))$ with

$$\begin{aligned} \boldsymbol{\mu}_s(\boldsymbol{\theta}) &= \mathbf{m}_s^y(\boldsymbol{\theta})\,, \\ \boldsymbol{\Sigma}_s(\boldsymbol{\theta}) &= \frac{\nu(\boldsymbol{\theta}) - 2}{\nu(\boldsymbol{\theta})}\mathbf{C}_s^y(\boldsymbol{\theta}) + \boldsymbol{\Gamma}(\boldsymbol{\theta})\,. \end{aligned} \tag{13}$$

For $\nu \to \infty$, the Student's t distribution equals a normal distribution. The derivatives of the probability density (11) and the distribution parameters (13) are provided in Appendix A.3.

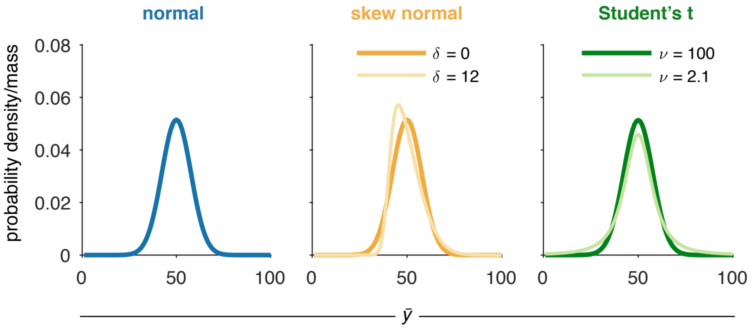

Figure 2: **Distributions assumptions for the hierarchical population model.** The visualized distributions (normal, skew normal and Student's t) for $\bar{y} \in \mathbb{R}_+$ have mean $m = 50$ and variance $C = 60$. The skew normal and Student's t distributions are visualized for different skewness parameters δ and degree of freedom ν, respectively.

3 Results

3.1 Evaluation of influence of distribution assumptions for simulated data

To assess the performance of model calibration under the different distribution assumptions in the hierarchical population model, we simulated data for three different motifs: a conversion process, a two-stage gene expression and a birth-death process (Fig. 3 and Appendix D). The first motif is frequently found in signal transduction networks, and the last two motifs models are commonly used for the description of gene expression. In this study, we assumed that the cell population comprises two subpopulations and that the true underlying difference between the subpopulations is known in the hierarchical population models. However, the hierarchical population model is able to describe more than two subpopulations and the number of subpopulations could be inferred by performing model selection.

All systems were assumed to be in steady state before the stimulus $\mathbf{u}$ was added at time point 0. However, this is not required by the hierarchical population model. For each motif, we chose three parameter vectors, three numbers of time points and four numbers of cells per time point (50, 100, 500, 1000). This yielded 108 data sets which were simulated using the stochastic simulation algorithm (Gillespie, 1977). The differences in the measurements for cells within a subpopulation arose solely due to intrinsic noise and no additional measurement noise was added to the data.

We perturbed the data according to different outlier scenarios (Fig. 4):

- *No outliers*: no outliers were included in the data.
- *Zeros*: the measured concentration at a certain time point t_k is zero, e.g., due to a missing label or entry or a dropout event (Luecken and Theis, 2019). Consequently, we measured $\bar{y}_j = 0$ for randomly chosen cells.
- *Doublets*: one measurement includes the summed information of two cells, e.g., due to wrongly measuring two cells instead of one. As a simplified simulation approach, the measured value of randomly chosen cell was doubled.
- *Uniform*: the measurement does not carry any information. Randomly chosen cells have uniformly distributed values in a defined regime (C.21) instead of the real measurement.

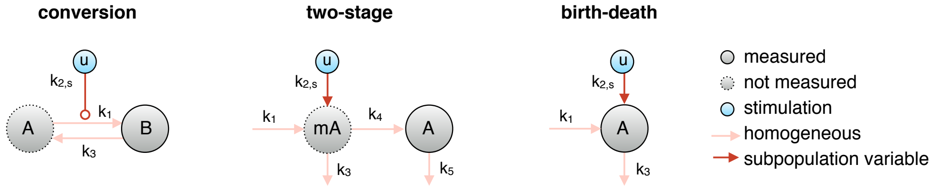

Figure 3: **Three motifs for the simulation study.** We considered a conversion process, a two-stage gene expression and a birth-death process. For each motif, we assumed two subpopulations which differ in their response to stimulus u. All other reaction rate constants were assumed to be homogeneous, i.e., the same for all cells.

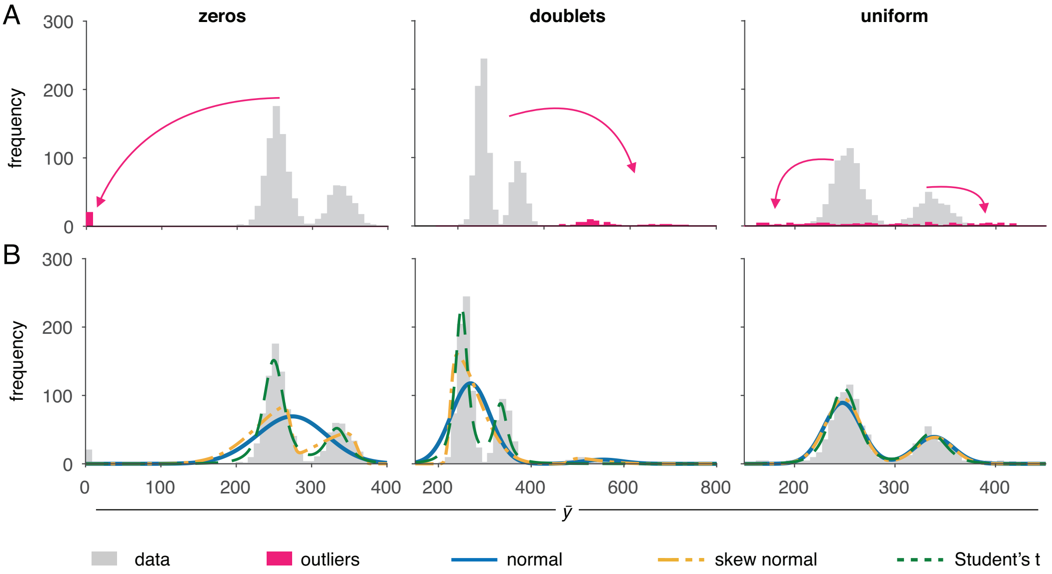

Figure 4: **Outlier scenarios for single-cell snapshot data**. (A) Example data sets of a conversion process with (B) corresponding fits with different distribution assumptions. The arrows in (A) illustrate the outlier-generating mechanisms.

In this simulation study, the deviation between the biological quantity and the mean of the quantity for a subpopulation arises due to intrinsic noise. The discrepancy between the true biological quantity and the measurement only arises due to the outlier-generating process. The amount of outliers in the data has a different influence for the different scenarios. The introduction of *zero* measurements is, for example, a bigger perturbation of the data as the *uniform* scenario. To obtain a comparable perturbation, we used different percentages of outliers for the scenarios and assumed 2%, 5% and 10% of the cells to be outliers for scenarios *zeros*, *doublets*, and *uniform*, respectively.

We calibrated the hierarchical population models based on all data sets for the different distribution assumptions with 30 optimization runs which were started at randomly chosen parameter values. For this simulation study, the measurements were count data and the continuous distributions were evaluated for the untransformed measured counts. The subpopulation sizes were fitted in linear space and the parameters required for the simulation of the statistical moments in $\log_{10}$ space. We assumed the true underlying sources of heterogeneity to be known and allowed for measurement noise. The moments of the subpopulations were obtained using the moment-closure approximation.

The fits for the different distribution assumptions are shown for three example data sets in Fig. 4B.

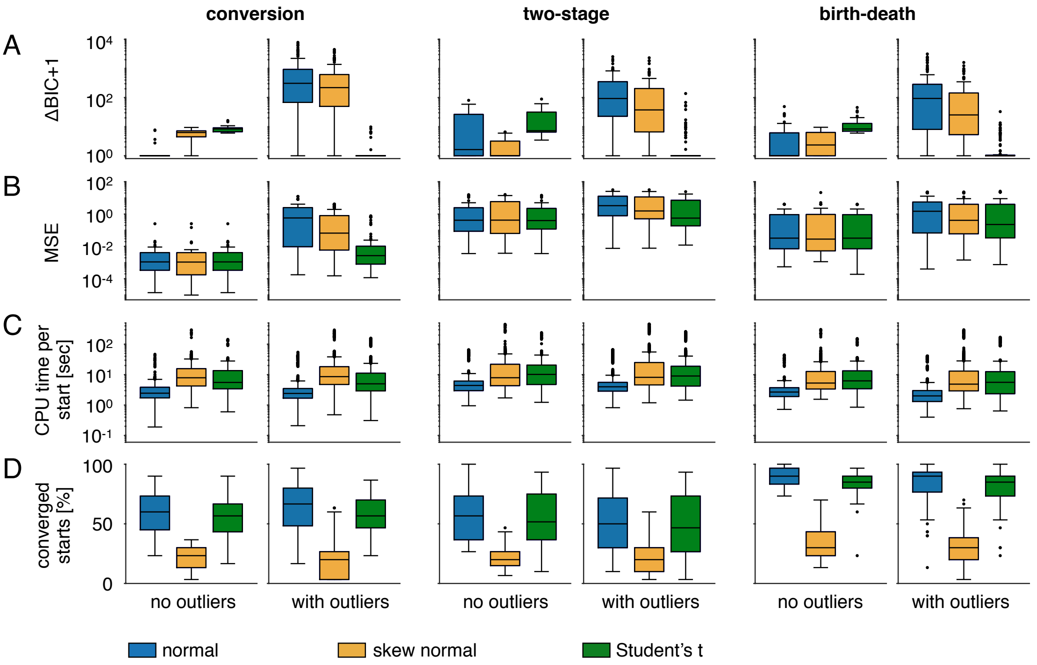

Figure 5: **Results for the simulation study.** Comparison of (A) ΔBIC values, (B) MSE, (C) CPU time per optimization start and (D) number of converged starts for the distribution assumptions for the motifs conversion process, two-stage gene expression and birth-death process. We considered outlier-free and outlier-corrupted data. For *no outliers*, each boxplot in (C) has 1080 points (36 data sets and 30 optimization runs) and the other boxplots (A, B, D) comprise 36 points. The subplots for *with outliers* comprise the results of all three outlier scenarios. For *with outliers*, each boxplot in (C) has 3240 points (3 outlier scenarios, 36 data sets, 30 optimization runs) and the other boxplots (A, B, D) comprise 108 points.

All distributions accurately fitted the data for the *uniform* outlier scenario. However, for the *zeros* and *doublets* scenarios, only the Student's t distribution was not deviated by the outliers, potentially because it has the heaviest tails of the considered distributions.

In the following comparison of the distributions, we only distinguished between the motifs and the absence or presence of outliers. The data sets corresponding to different parameter values, time points, number of cells and outlier scenarios were merged. The comparison of the individual outlier scenarios is displayed in Appendix Fig. D1.

To compare the different models quantitatively, we used the Bayesian Information Criterion (BIC) (Schwarz, 1978),

$$\mathrm{BIC} = -2\log\mathcal{L}(\boldsymbol{\theta}) + \log(n_{\mathcal{D}})n_{\theta}\,, \tag{14}$$

with $n_{\mathcal{D}}$ denoting the number of data points and n_θ denoting the number of parameters. As lower BIC values are preferable (Schwarz, 1978), the BIC rewards high likelihood values and penalizes model complexity. As an alternative, the Akaike Information Criterion (AIC) could be used (Akaike, 1973). We compared the differences in BIC values to the minimal BIC value found for the given data set (=ΔBIC) (Fig. 5A). In the case of outlier-free data, the best distribution assumption differed between the studied motifs. For the conversion process the normal distribution was most appropriate, while for the two-stage gene expression and the birth-death motif the normal and the skew normal distribution achieved low BIC values. As soon as outliers were introduced to the data, the Student's t distribution provided in average the lowest BIC value for all considered motifs and was therefore selected as most suited model.

The mean squared error (MSE) provides a measure for the accuracy of the parameter estimates:

$$\mathrm{MSE}\left[\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^{\mathrm{true}}\right] = \mathbb{E}\left[\sum_i \left(\hat{\theta}_i - \theta_i^{\mathrm{true}}\right)^2\right] . \tag{15}$$

A small (norm of the) MSE indicates a good agreement of the true and estimated parameters. A high MSE indicates a large difference to the true parameters, which in turn might distort model predictions and the capability of the model to provide correct insights into the underlying system of study. We considered the MSE in the logarithmic/linear space in which the parameters were fitted and summed the errors of the parameter estimates to obtain a single value for each data set and distribution. The true parameters which were used to generate the data for the simulation study are known, and we compared the MSEs of the parameter estimates obtained with the different models (Fig. 5B). For the cases of *no-outliers*, the MSE for all distributions were similar. In the presence of outliers, the Student's t distribution still provided low MSEs which were comparable to the those obtained in the outlier-free scenario. The MSE obtained by the normal and skew normal distribution increased in the presence of outliers. The skew normal distribution provided slightly more robust results, i.e., lower MSEs, than the normal distribution. High MSEs do not necessarily result in deviations of the output of the model, e.g., if the output is insensitive to these parameters. Thus, we compared the model outputs for the parameters which were obtained for the outlier-corrupted data to the original, *no-outlier* data (Appendix Fig. D1E). We found that the Student's t distribution consistently outperformed the other distributions.

A further important aspect to consider is the robustness and efficiency of model calibration. Often many models which represent different biological hypotheses need to be calibrated. These hypotheses could include different sources of heterogeneity or different numbers of subpopulations. Therefore, the optimization of each individual model should be fast and robust. To assess the performance of model calibration using the distribution assumptions, we considered the computation time and number of converged starts (Fig. 5C-D). We considered an optimizer start to be converged, if the difference between the obtained log-likelihood value and the best log-likelihood function for this distribution assumption and the considered motif is below 10^{-3}. For most of the here considered motifs and data sets the computation time and convergence were not substantially influenced by the presence of outliers. On average, the normal distribution required the lowest computation time. An explanation for this could be that the optimization problem using this distribution is lower than the problem using the skew normal and the Student's t distribution, since no additional parameters were estimated from the data. In terms of converged starts, we observed some differences between the considered motifs. For all motifs and scenarios the skew normal distribution provided the lowest number of converged starts, while the normal and Student's t distribution did not suffer from convergence problems.

The simulation study showed that the consideration of alternative distribution assumptions was beneficial and the normal distribution was often outperformed by the other distributions. The heavier tails of the distributions allowed for a compensation of the outliers in the data, but did not generate substantial computational overhead in the outlier-free case. Overall, the Student's t distribution provided reliable parameter estimates, the overall closest predictions to the original *no-outlier* data and, at the same time, enabled an efficient calibration of the corresponding model.

3.2 NGF-induced Erk1/2 signaling

To test the distributions in a real application setting, we reanalyzed the data and hierarchical population model studied in (Loos et al., 2018). The model describes binding of NGF to its receptor TrkA, which then induces the phosphorylation of Erk1/2 (Fig. 6A), an important process involved in pain sensitization. The measurements were obtained from primary sensory neurons using fluorescent microscopy. For more details on the model and data, we refer to Loos et al. (2018).

For this analysis, we log-transformed the simulation and data. For the Student's t distribution we assumed that one parameter for the degree of freedom was shared across the dimensions, yielding one additional parameter. For the skew normal distribution, each dimension is allowed to have different skewness parameters, yielding two parameters more than the normal distribution. Calibrating the hierarchical population models, we found that for the univariate data of the pErk1/2 kinetics, the model fits cannot visually be distinguished (Fig. 6B and Fig. D2). The skew normal distribution fitted the bivariate data the best (Fig. 6C,D and Fig. D2). We could not assess the MSE since the true parameters are not known. Visualizing the likelihood waterfall plots (Fig. 6D) and analyzing the performance of the optimization (Fig. 6E), we found that the Student's t distribution substantially outperformed the other distributions in terms of converged optimizer starts (per minute). Interestingly, the skew normal which showed bad convergence for the simulation study even had a higher convergence than the normal distribution for this application problem. The skew normal provided the best likelihood and BIC value. An explanation for why the skew normal distribution was chosen over the other distributions could be a skewed distribution of Erk levels arising from biological variability rather than noise or outliers. However, the hierarchical population model uses a single distribution assumption for the combined influence of cell-to-cell variability, measurement noise and outliers. Accordingly, if the distribution assumptions allow for skewness, the skewness can arise from any of the properties. To summarize, the incorporation of the alternative distribution assumptions into the hierarchical population model demonstrated their robustness and efficiency not only for the simulation study, but also for the considered experimental data set. Thus, the distribution assumptions seem also promising for real experimental data and future work should include the evaluation of the distributions for more experimental data.

4 Discussion

The hierarchical population model is a suitable tool for studying cellular heterogeneity, but it requires appropriate distribution assumptions for the cellular subpopulations. Here, we incorporated various distributions for the subpopulations and provided the equations to perform gradient-based model calibration. Gradient-based calibration has previously has been shown to be highly efficient for the considered model class (e.g., Loos et al. (2016)) and, thus, also enhances optimization-based uncertainty analysis using profile likelihoods (Raue et al., 2009). This efficiency of model calibration is especially important when a large number of hierarchical population models, which represent different hypothesis such as differences between or number of subpopulations, needs to be compared. Furthermore, we studied the influence of the choice of distribution on the estimation results and the performance of model calibration for artificial and real experimental data.

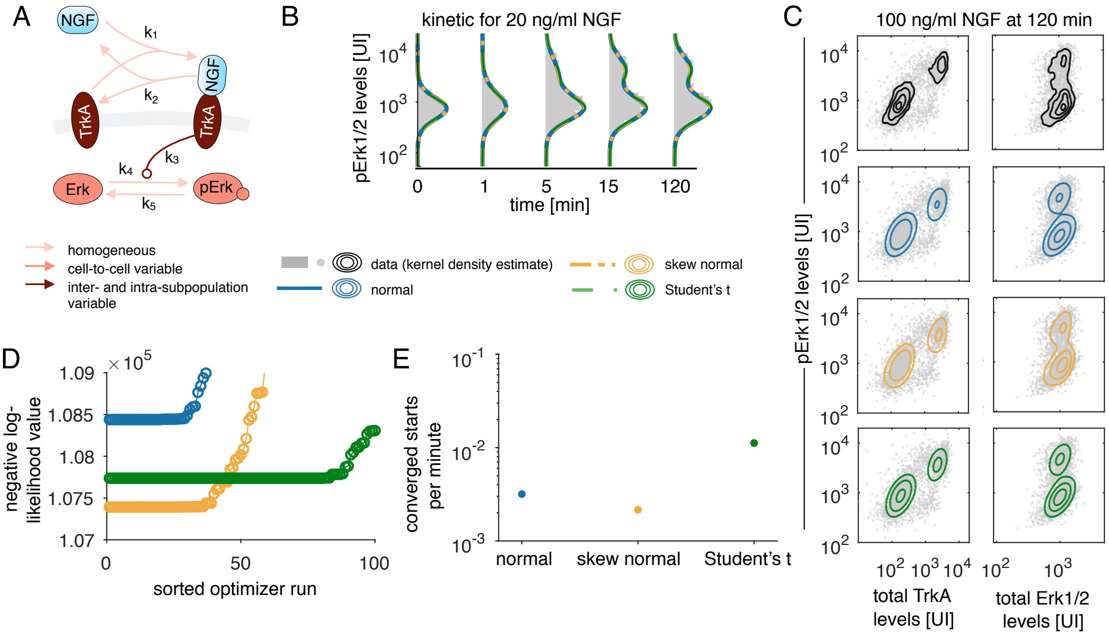

Figure 6: **Robust distributions for NGF-induced Erk1/2 signaling.** (A) Model for NGF-induced Erk1/2 signaling. (B,C) Data and model fits for (A) univariate measurements of pErk1/2 levels and (C) bivariate measurements for pErk/TrkA and pErk/Erk levels. (D) Likelihood waterfall plot for the three different distribution assumptions. The best 80 values are shown and in total 500 optimization runs started at randomly drawn parameter values were performed. (E) The performance of the optimization measured as number of converged starts per minute.

Differences in measurements of single-cell arise due to various factors: biological variability, measurement noise and the outlier-generating mechanisms. The distribution incorporated in the hierarchical population model ideally should capture all these factors. Here, we found that the normal distribution assumption is often appropriate when the biological variability yields a normal distribution of the subpopulations and additionally a limited number of outliers is to be expected. However, the biological variability might not always yield a normal distribution of the subpopulations. This was observed in the experimental data of the primary sensory neurons and suggests that the best choice of distribution highly depends on the particular application problem. This motivates the use and comparison of alternative distributions also when no outliers are present in the data. If the data is outlier-corrupted, alternatives such as the skew normal or Student's t distribution are more reasonable. The Student's t distribution provided reliable results when the data is outlier-free and could be considered as default distribution assumption. If more information is available about the precise type of outliers, e.g., if they arise due to dropout events in scRNA-seq data, computational methods can be adapted accordingly (Pierson and Yau, 2015; Eraslan et al., 2019). While the Student's t distribution suffered from problems of over-fitting in the case of population-average data (Maier et al., 2017), the number of measurements in single-cell data sets is usually much higher and we do not expect to face the same problems as for population-average data.

Other distributions could be incorporated into the hierarchical population model, given that their mean and covariance are finite and an analytical gradient can be provided. To allow for different degrees of freedom in multivariate measurements, a t copula could be employed (Luo and Shevchenko, 2010). Also, a skewed version of the Student's t distribution as, e.g., used by Pyne et al. (2009) could be incorporated. The log-normal distribution could easily be incorporated by transforming the observable. In this case, a constant factor needs to be added to the BIC value when performing model selection.

In summary, we introduced the use of different distribution assumptions for the hierarchical population model and evaluated their performance. Our results on simulation and application examples suggested that these distribution assumptions substantially improve the hierarchical population model and the reliability of its results, and, thus, enhance the study of cellular heterogeneity.

Implementation and code availability

The alternative distribution assumptions are implemented in the MATLAB toolbox ODEMM (Loos et al., 2018) available under `https://github.com/ICB-DCM/ODEMM`. The model calibration was performed using the toolbox PESTO (Stapor et al., 2018). The ODE models were simulated using the toolbox AMICI (Fröhlich et al., 2017). The sigma-point approximation was obtained by the SPToolbox. These toolboxes can be found under `https://github.com/ICB-DCM`. For the moment-closure approximation, we employed the toolbox CERENA (Kazeroonian et al., 2016) available under `https://cerenadevelopers.github.io/CERENA/`. The whole analysis was performed using MATLAB 2017b. The code to reproduce the results of this study is available under `http://doi.org/10.5281/zenodo.3354136`.

Acknowledgements

This work was supported by the German Ministry of Education and Research by the grant SYS-Stomach (01ZX1310B) and the grant INCOME (FKZ 01ZX1705A).

Competing interests

The authors declare no competing interests.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR*, volume 1, pages 267–281. Akademiai Kiado.

Altschuler, S. J. and Wu, L. F. (2010). Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559–563.

Amrhein, L., Harsha, K., and Fuchs, C. (2019). A mechanistic model for the negative binomial distribution of single-cell mRNA counts. *bioRxiv*, 657619.

Balázsi, G., van Oudenaarden, A., and Collins, J. J. (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6):910–925.

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., Kwon, J. Y. H., Barak, B., Ge, W., Kedaigle, A. J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A. K., Villani, A.-C., Regev, A., and Levin, J. Z. (2019). Systematic comparative analysis of single cell rna-sequencing methods. *bioRxiv*, 632216.

Dixit, P. D., Lyashenko, E., Niepel, M., and Vitkup, D. (2019). Maximum entropy framework for inference of cell population heterogeneity in signaling networks. *bioRxiv*, 137513.

Engblom, S. (2006). Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, 180(2):498–515.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, 10(1):390.

Fröhlich, F., Kaltenbacher, B., Theis, F. J., and Hasenauer, J. (2017). Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.*, 13(1):e1005331.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361.

Gillespie, D. T. (2000). The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640.

Henson, M. A. (2003). Dynamic modeling of microbial cell populations. *Curr. Opin. Biotechnol.*, 14(5):460–467.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome biology*, 17(1):29.

Johnsson, K., Wallin, J., and Fontes, M. (2016). BayesFlow: latent modeling of flow cytometry cell populations. *BMC Bioinf.*, 17(1):25.

Kazeroonian, A., Fröhlich, F., Raue, A., Theis, F. J., and Hasenauer, J. (2016). CERENA: Chemical REaction network Analyzer – A toolbox for the simulation and analysis of stochastic chemical kinetics. *PLoS ONE*, 11(1):e0146732.

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

Kuepfer, L., Peter, M., Sauer, U., and Stelling, J. (2007). Ensemble modeling for analysis of cell signaling dynamics. *Nat. Biotechnol.*, 25(9):1001.

Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.*, 84(408):881–896.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A*, 73:321–332.

Loos, C., Fiedler, A., and Hasenauer, J. (2016). Parameter estimation for reaction rate equation constrained mixture models. In Bartocci, E., Lio, P., and Paoletti, N., editors, *Proc. 13th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 186–200. Springer International Publishing.

Loos, C., Moeller, K., Fröhlich, F., Hucho, T., and Hasenauer, J. (2018). A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Syst.*, 6(5):593–603.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: a tutorial. *MSB*, 15(6).

Luo, X. and Shevchenko, P. V. (2010). The t copula with multiple parameters of degrees of freedom: bivariate characteristics and application to risk management. *Quant. Finance*, 10(9):1039–1054.

Maier, C., Loos, C., and Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, 33(5):718–725.

Mar, J. C. (2019). The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond. *Biophys. Rev.*, 11(1):89–94.

Pierson, E. and Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16(1):241.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T., Maier, L., Baecher-Allan, C., McLachlan, G., Tamayo, P., Hafler, D., De Jager, P., and Mesirov, J. (2009). Automated high-dimensional flow cytometric data analysis. *Proc. Natl. Acad. Sci. USA*, 106(21):8519–8124.

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

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., Klingmüller, U., and Timmer, J. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Can. J. Stat.*, 31(2):129–150.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. USA*, 105(45):17256–17261.

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insur. Math. Econ.*, 55:18–29.

Stapor, P., Weindl, D., Ballnus, B., Hug, S., Loos, C., Fiedler, A., Krause, S., Hross, S., Fröhlich, F., and Hasenauer, J. (2018). PESTO: Parameter EStimation TOolbox. *Bioinformatics*, 34(4):705–707.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA*, 99(20):12795–12800.

van der Merwe, R. (2004). *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models*. PhD thesis, Oregon Health & Science University.

Waldherr, S., Hasenauer, J., and Allgöwer, F. (2009). Estimation of biochemical network parameter distributions in cell populations. In Walter, E., editor, *Proc. of the 15th IFAC Symp. on Syst. Ident.*, volume 15, pages 1265–1270, Saint-Malo, France.

Wang, D., Stapor, P., and Hasenauer, J. (2019). Dirac mixture distributions for the approximation of mixed effects models. *bioRxiv*, 703850.

A Gradient of the likelihood functions

A.1 Multivariate normal distribution

The probability density function for the multivariate normal distribution is defined in (5). The log-density function is given by

$$\log \phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\Big(n_y \log(2\pi) + \log(\det \boldsymbol{\Sigma}) + (\bar{\mathbf{y}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu})\Big) \,. \tag{A.16}$$

Assuming that the distribution parameters depend on parameter vector $\boldsymbol{\theta}$, the derivative of the multivariate normal density is given by

$$\begin{aligned}\frac{\partial \log \phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial \theta_i} = &-\frac{1}{2}\Bigg(\mathrm{Tr}\left((\boldsymbol{\Sigma}(\boldsymbol{\theta}))^{-1}\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i}\right) + (\boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \\ &+ \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\mathbf{y}}) + (\boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\mathbf{y}})^T \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}}{\partial \theta_i}(\boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\mathbf{y}})\Bigg) \,,\end{aligned} \tag{A.17}$$

with derivatives for the distribution parameters for subpopulation s

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{m}_s^y(\boldsymbol{\theta})}{\partial \theta_i} \,, \\ \frac{\partial \boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{C}_s^y(\boldsymbol{\theta})}{\partial \theta_i} + \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\theta})}{\partial \theta_i} \,.\end{aligned}$$

A.2 Multivariate skew normal distribution

The probability density function for the multivariate skew normal distribution is defined in (8). The log-density function is given by

$$\log \phi_{\text{skewnorm}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = \log(2) + \log \phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\Omega}) + \log(\Phi_{\text{norm}}(\boldsymbol{\alpha}(\bar{\mathbf{y}} - \boldsymbol{\mu})|0,1)) \,.$$

Assuming that the distribution parameters depend on parameter vector $\boldsymbol{\theta}$, the gradient is given by

$$\begin{aligned}\frac{\partial \log \phi_{\text{skewnorm}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial \theta_i} = &\frac{\partial \log \phi_{\text{norm}}(\bar{\mathbf{y}}|\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Omega}(\boldsymbol{\theta}))}{\partial \theta_i} + \\ &\frac{1}{\Phi(\boldsymbol{\alpha}(\boldsymbol{\theta})(\bar{\mathbf{y}}\boldsymbol{\mu}(\boldsymbol{\theta})))}\phi_{\text{norm}}(\boldsymbol{\alpha}(\boldsymbol{\theta})(\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))|0,1) \,\cdot \\ &\left(\boldsymbol{\alpha}(\boldsymbol{\theta})\frac{\partial(\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))}{\partial \theta_i} + (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))\frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \theta_i}\right) \,,\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial \boldsymbol{\alpha}(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left[\boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \underbrace{\left(1-\boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{1}{2}}}_{a(\boldsymbol{\theta})}\right] \\
&= \left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i}\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} + \boldsymbol{\delta}(\boldsymbol{\theta})^T \frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i}\right) a(\boldsymbol{\theta}) + \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \frac{\partial a(\boldsymbol{\theta})}{\partial \theta_i}\,, \\
\frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i} &= -\,\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})}{\partial \theta_i}\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \\
&= -\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \sqrt{\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T}\right)^{-1}\left(\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i} + \frac{1}{2}(\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T)^{-\frac{1}{2}}\cdot\right. \\
&\qquad \left.\left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\boldsymbol{\delta}^T + \boldsymbol{\delta}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i}\right)\right)\cdot\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \sqrt{\boldsymbol{\delta}(\boldsymbol{\theta})\boldsymbol{\delta}(\boldsymbol{\theta})^T}\right)^{-1}\,, \\
\frac{\partial a(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{1}{2}\left(1-\boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{3}{2}} \frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i} \\
&= \frac{1}{2}\left(1-\boldsymbol{\delta}(\boldsymbol{\theta})^T \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\boldsymbol{\delta}(\boldsymbol{\theta})\right)^{-\frac{3}{2}}\left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i}\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\boldsymbol{\delta}(\boldsymbol{\theta}) + \right. \\
&\qquad \left.\boldsymbol{\delta}(\boldsymbol{\theta})^T\left(\frac{\partial \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}}{\partial \theta_i}\boldsymbol{\delta}(\boldsymbol{\theta}) + \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\right)\right).
\end{aligned}$$

The derivatives for the distribution parameters (10) for subpopulation s are given by

$$\begin{aligned}
\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{m}_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \sqrt{\frac{2}{\pi}}\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\,, \\
\frac{\partial \boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{C}_s^y(\boldsymbol{\theta})}{\partial \theta_i} - \left(1-\frac{2}{\pi}\right)\left(\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})}{\partial \theta_i}\boldsymbol{\delta}(\boldsymbol{\theta})^T + \boldsymbol{\delta}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\delta}(\boldsymbol{\theta})^T}{\partial \theta_i}\right) + \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\theta})}{\partial \theta_i}\,.
\end{aligned}$$

A.3 Multivariate Student's t distribution

The probability density function for the multivariate Student's t distribution is defined in (11).

The log-density function is

$$\log \phi_{\text{stud}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}) = \log\Gamma\left(\frac{\nu+n_y}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \log(|\boldsymbol{\Sigma}|^{-\frac{1}{2}}) - \frac{n_y}{2}\log(\pi\nu) - \frac{\nu+n_y}{2}\log\left(1+\frac{1}{\nu}\mathbf{Z}\right)\,,$$

with $\mathbf{Z} = (\bar{\mathbf{y}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})$. The gradient is given by

$$\begin{aligned}
\frac{\partial \log \phi_{\text{stud}}(\bar{\mathbf{y}}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial \theta_i} &= \frac{1}{2}\left(\left(\Psi\left(\frac{\nu(\boldsymbol{\theta})+n_y}{2}\right) - \Psi\left(\frac{\nu(\boldsymbol{\theta})}{2}\right) - \frac{n_y}{\nu(\boldsymbol{\theta})} + \right.\right. \\
&\qquad \left.\frac{\mathbf{Z}(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+n_y) - \nu(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta}))\log(1+\frac{1}{\nu(\boldsymbol{\theta})}\mathbf{Z}(\boldsymbol{\theta}))}{\nu(\boldsymbol{\theta})(\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta}))}\right)\frac{\partial \nu(\boldsymbol{\theta})}{\partial \theta_i} \\
&\qquad \left.-\operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i}\right) - \frac{\nu(\boldsymbol{\theta})+n_y}{\nu(\boldsymbol{\theta})+\mathbf{Z}(\boldsymbol{\theta})}\frac{\partial \mathbf{Z}(\boldsymbol{\theta})}{\partial \theta_i}\right),
\end{aligned}$$

with

$$\begin{aligned}\frac{\partial \mathbf{Z}(\boldsymbol{\theta})}{\partial \theta_i} =&(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right)^T + \left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i}\right)^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta})) + \\ &(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}}{\partial \theta_i}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))\,,\end{aligned}$$

and digamma function denoted by Ψ.

The derivatives for the distribution parameters (13) for subpopulation s are

$$\begin{aligned}\frac{\partial \boldsymbol{\mu}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\partial \mathbf{m}_s^y(\boldsymbol{\theta})}{\partial \theta_i}\,,\\ \frac{\partial \boldsymbol{\Sigma}_s(\boldsymbol{\theta})}{\partial \theta_i} &= \frac{\nu(\boldsymbol{\theta})-2}{\nu(\boldsymbol{\theta})}\frac{\partial \mathbf{C}_s^y(\boldsymbol{\theta})}{\partial \theta_i} + \mathbf{C}_s^y(\boldsymbol{\theta})\frac{2}{\nu(\boldsymbol{\theta})^2}\frac{\partial \nu(\boldsymbol{\theta})}{\partial \theta_i} + \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\theta})}{\partial \theta_i}\,.\end{aligned}$$

B Negative binomial distribution

A further distribution assumption which is often employed in the analysis of single-cell data is the negative binomial distribution (Grün et al., 2014; Amrhein et al., 2019), which is a count distribution in contrast to the other distributions. For a two-stage model of gene expression, the protein number in steady state follows a negative binomial distribution if the ratio of mRNA degradation to protein degradation is high (Shahrezaei and Swain, 2008), or if the mRNA molecules are produced in bursts (Amrhein et al., 2019). This distribution has the parameters $\boldsymbol{\varphi} = (\tau, \rho)$ with $\tau > 0$ and $\rho \in [0, 1]$. We considered the univariate case and multivariate data could be modeled by the product distribution which neglects correlations. The probability mass function reads

$$\phi_{\text{nbin}}(\bar{y}|\boldsymbol{\varphi}) = \binom{\bar{y}+\tau-1}{\bar{y}}(1-\rho)^{\bar{y}}\rho^{\tau}\,. \tag{B.18}$$

The mean and variance of the negative binomial distribution are

$$\begin{aligned}m &= \frac{(1-\rho)\tau}{\rho}\,,\\ C &= \frac{(1-\rho)\tau}{\rho^2}\,.\end{aligned} \tag{B.19}$$

Thus, the distribution parameters $\boldsymbol{\varphi}_s(\boldsymbol{\theta}) = (\rho_s(\boldsymbol{\theta}), \tau_s(\boldsymbol{\theta}))$ are mapped to the moments via

$$\begin{aligned}\rho_s(\boldsymbol{\theta}) &= \frac{m_s^y(\boldsymbol{\theta})}{C_s^y(\boldsymbol{\theta})+\Gamma(\boldsymbol{\theta})}\,,\\ \tau_s(\boldsymbol{\theta}) &= \frac{m_s^y(\boldsymbol{\theta})^2}{C_s^y(\boldsymbol{\theta})+\Gamma(\boldsymbol{\theta})-m_s^y(\boldsymbol{\theta})}\,.\end{aligned} \tag{B.20}$$

The derivatives of the probability density (B.18) and the distribution parameters (B.20) are provided in Appendix B. For large τ, the negative binomial distribution approaches a normal distribution.

The log-density function reads

$$\begin{aligned}\log \phi_{\text{nbin}}(\bar{y}|\boldsymbol{\varphi}) &= \log\left(\binom{\bar{y}+\tau-1}{\bar{y}}(1-\rho)^{\bar{y}}\rho^{\tau}\right)\\ &= \log(\Gamma(\bar{y}+\tau)) - \log(\Gamma(\bar{y}+1)) - \log(\Gamma(\tau)) + \bar{y}\log(1-\rho) + \tau\log(\rho)\,.\end{aligned}$$

The derivative of the log-density function is

$$\begin{aligned}\frac{\partial \log \phi_{\text{nbin}}(\bar{y}|\boldsymbol{\varphi}(\boldsymbol{\theta}))}{\partial\theta_i} &= \Psi(\bar{y}+\tau(\boldsymbol{\theta}))\frac{\partial\tau(\boldsymbol{\theta})}{\partial\theta_i} - \Psi(\tau(\boldsymbol{\theta}))\frac{\partial\tau(\boldsymbol{\theta})}{\partial\theta_i} - \\ &\qquad \frac{y}{1-\rho(\boldsymbol{\theta})}\frac{\partial\rho(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\tau(\boldsymbol{\theta})}{\partial\theta_i}\log(\rho(\boldsymbol{\theta})) + \frac{\tau(\boldsymbol{\theta})}{\rho}\frac{\partial\rho(\boldsymbol{\theta})}{\partial\theta_i}\,.\end{aligned}$$

The derivatives of the distribution parameters (B.20) for subpopulation s are

$$\begin{aligned}\frac{\partial\rho_s(\boldsymbol{\theta})}{\partial\theta_i} &= \frac{1}{C_s^y(\boldsymbol{\theta})}\frac{\partial m_s^y(\boldsymbol{\theta})}{\partial\theta_i} - \frac{m_s^y(\boldsymbol{\theta})}{C_s^y(\boldsymbol{\theta})}\frac{\partial C_s^y(\boldsymbol{\theta})}{\partial\theta_i}\,,\\ \frac{\partial\tau_s(\boldsymbol{\theta})}{\partial\theta_i} &= \frac{m_s^y(\boldsymbol{\theta})(2C_s^y(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))}{(C_s^y(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))^2}\frac{\partial m_s^y(\boldsymbol{\theta})}{\partial\theta_i} - \frac{m_s^y(\boldsymbol{\theta})^2}{(C_s^y(\boldsymbol{\theta}) - m_s^y(\boldsymbol{\theta}))^2}\frac{\partial C_s^y(\boldsymbol{\theta})}{\partial\theta_i}\,.\end{aligned}$$

In contrast to the other distributions, the negative binomial distribution can not take into account correlation structures between the dimensions of multivariate data. A multivariate extensions which account for correlations was proposed by Shi and Valdez (2014).

C Uniform outlier scenario

In the *uniform* outlier scenario, the measured values of the outlier cells were assigned to the rounded value of uniformly distributed values on the interval

$$[\max(0, \min_c(\bar{y}_k^c) - 0.25\cdot I), \max_c(\bar{y}_k^c) + 0.25\cdot I] \tag{C.21}$$

with I denoting the length of the interval without outliers.

D Models

D.1 Conversion process

The conversion process is described by the following reactions:

$$\begin{aligned}\mathrm{R}_1: &\quad \mathrm{A}\to\mathrm{B}\,, &\quad \text{rate} = k_1[\mathrm{A}]\,,\\ \mathrm{R}_2: &\quad \mathrm{A}\to\mathrm{B}\,, &\quad \text{rate} = k_{2,s}u[\mathrm{A}]\,,\\ \mathrm{R}_3: &\quad \mathrm{B}\to\mathrm{A}\,, &\quad \text{rate} = k_3[\mathrm{B}]\,.\end{aligned}$$

Reaction R_1 describes the basal conversion from A to B and reaction R_2 the stimulus-dependent conversion. The conversion from B to A, reaction R_3, does not depend on stimulus u. We assumed mass conservation with $[\mathrm{A}] + [\mathrm{B}] = 1000$.

The moment-closure approximation provides the temporal evolution of the moments,

$$\mathbf{z}_s = \begin{pmatrix} m_1^x \\ m_2^x \\ C_{11}^x \\ C_{12}^x \\ C_{11}^x \end{pmatrix},$$

with m_1^x denoting the mean of species A, m_2^x denoting the mean of species B and C_{ij}^x denoting the corresponding entries of the covariance matrix. The measurement noise is defined as $\mathbf{\Gamma} = \sigma_{\text{noise}}$. The subpopulation parameter vector for subpopulation $s = 1, 2$ is given by $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \mathbf{D}_s)$ with

$$\boldsymbol{\beta}_s = \begin{pmatrix} k_1 \\ k_{2,s} \\ k_3 \\ \sigma_{\text{noise}} \end{pmatrix} \begin{array}{l} \text{homogeneous}\,, \\ \text{subpopulation variable}\,, \\ \text{homogeneous}\,, \\ \text{homogeneous}\,, \end{array}$$

and $\mathbf{D}_s = \mathbf{0}$.

The mapping g_φ introduced in (4) then encodes the mapping from the biochemical species A and B to its observable B, as well as the mapping to the distribution parameters as in (6, 10, 13, B.20). The overall parameter object which is estimated from the data is given by

$$\boldsymbol{\theta} = (k_1, k_{2,1}, k_{2,2}, k_3, \sigma_{\text{noise}})\,,$$

for the normal distribution,

$$\boldsymbol{\theta} = (k_1, k_{2,1}, k_{2,2}, k_3, \sigma_{\text{noise}}, \delta)\,,$$

for the skew normal distribution, and

$$\boldsymbol{\theta} = (k_1, k_{2,1}, k_{2,2}, k_3, \sigma_{\text{noise}}, \nu)\,,$$

for the Student's t distribution,

D.2 Two-stage gene expression

The two-stage gene expression is described by the following reactions:

$$\begin{array}{lrl} \mathrm{R}_1: & \emptyset \to \mathrm{mA}\,, & \text{rate} = k_1\,, \\ \mathrm{R}_2: & \emptyset \to \mathrm{mA}\,, & \text{rate} = k_{2,s}u\,, \\ \mathrm{R}_3: & \mathrm{mA} \to \emptyset\,, & \text{rate} = k_3[\mathrm{mA}]\,, \\ \mathrm{R}_4: & \mathrm{mA} \to \mathrm{mA} + \mathrm{A}\,, & \text{rate} = k_4[\mathrm{mA}]\,, \\ \mathrm{R}_5: & \mathrm{A} \to \emptyset\,, & \text{rate} = k_5[\mathrm{A}]\,. \end{array}$$

Reactions R_1 and R_2 describe the stimulus-independent and stimulus-dependent mRNA expression, respectively. Reaction R_3 describes mRNA degradation, reaction R_4 protein expression and reaction R_5 protein degradation.

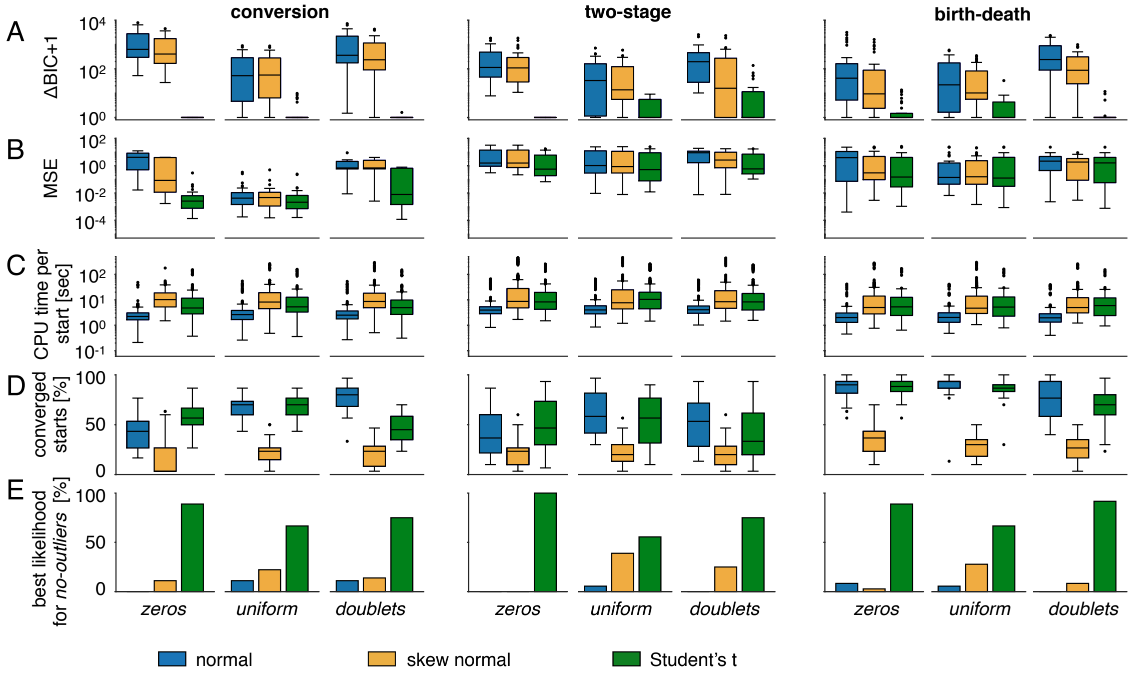

Figure D1: **Results for the simulation study for the individual outlier scenarios.** Comparison of (A) ΔBIC values, (B) MSE, (C) CPU time per optimization start and (D) number of converged starts for the distribution assumptions for the motifs conversion process, two-stage gene expression and birth-death process. Each boxplot in (C) has 1080 points (36 data sets and 30 optimization runs) and the other boxplots (A, B, D) comprise 36 points. (E) The likelihood values were calculated for the *no-outlier* data set using the parameters obtained for the outlier-corrupted data sets. The bars show how often the corresponding distribution provided the best likelihood value. A higher percentage indicates robustness against outliers.

D.3 Birth-death process

The birth-death process is described by the following reactions:

$$\begin{aligned} R_1 :& \quad \emptyset \to A\,, & \text{rate} &= k_1\,, \\ R_2 :& \quad \emptyset \to A\,, & \text{rate} &= k_{2,s}u\,, \\ R_3 :& \quad A \to \emptyset\,, & \text{rate} &= k_3[A]\,. \end{aligned}$$

Reactions R_1 and R_2 describe the stimulus-independent and stimulus-dependent production of A and reaction R_3 its degradation.

D.4 NGF-induced Erk1/2 signaling

We used the model proposed in Loos et al. (2018), assuming cell-to-cell variability in total Erk1/2 levels and inter- and intra-subpopulation variability in cellular TrkA activity. The moments of the system were obtained using the sigma-point approximation. The whole data set with the model fits is shown in Fig. D2.

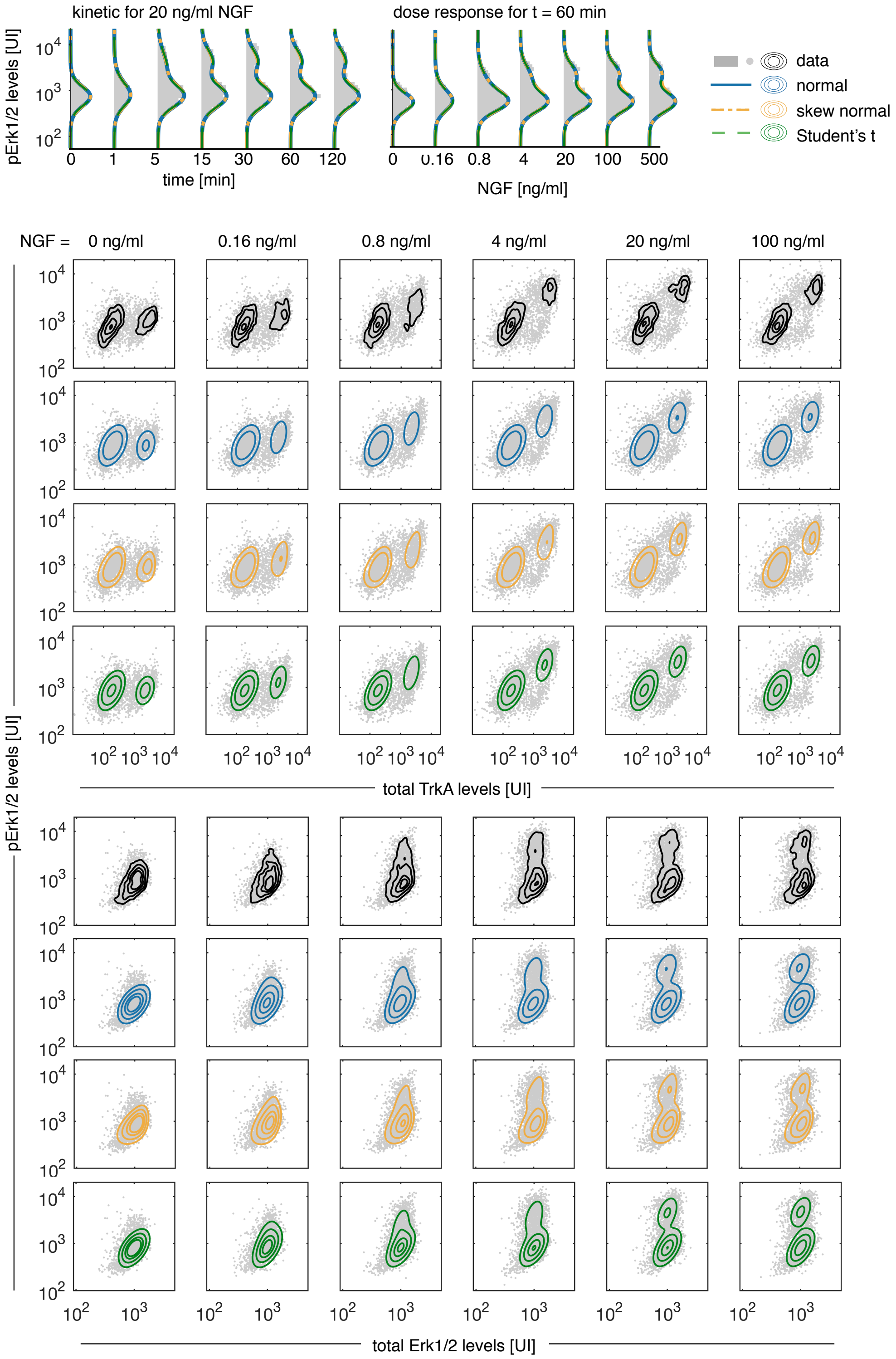

Figure D2: **Data and model fits for NGF-induced Erk1/2 signaling**. pErk1/2 kinetics and dose responses, and multivariate measurements of pErk/TrkA levels and pErk/Erk levels. The upper rows illustrate the data together with a kernel density estimate. The bottom rows visualize the data together with the contour lines of the hierarchical models.