THE UNIVERSITY OF
WARWICK

AUTHOR: **Anas A. Rana** DEGREE: **Ph.D.**

TITLE: **Stochastic models for cell populations undergoing transitions**

DATE OF DEPOSIT: .....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.
I agree that the summary of this thesis may be submitted for publication.
I **agree** that the thesis may be photocopied (single copies for study purposes only).
Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

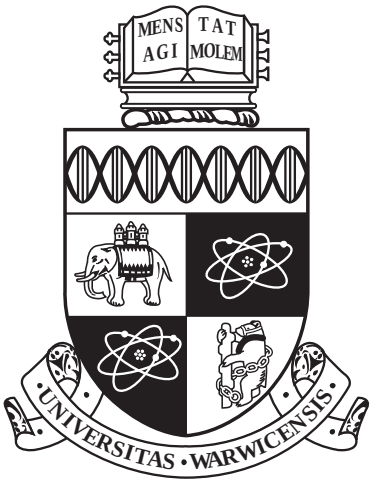
.....

.....

.....

.....

.....



Stochastic models for cell populations undergoing transitions

by

Anas A. Rana

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

.....

... .

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

Declarations

Replace this text with a declaration of the extent of the original work, collaboration, other published material etc. You can use any LaTeX constructs.

Abstract

Chapter 1

Introduction

SOMETHING GOES HERE

Chapter 2

Background

2.1 Maths

2.2 Biology

Chapter 3

MAST

3.1 Model outline

3.1.1 The MAST model

First I provie a self-contained description of the MAST model, following and expanding upon Armond et al. [2013].

MAST defines a latent stochastic process on the single cell level that isn't directly observed. We can obtain a cell-population level likelihood from this. We describe the latent single cell process using a Markov chain with discrete and finite state space and continuous in time. The state space, indexed by $k \in \{1, ..., K\}$, is identified with biological states of the system. We denote transition rates between states k and k' by $\mathbf{w} = \{w_{k,k'}\}$. Assuming that rates of cell death and cell doubling cancel, the probability for any cell to be in state k at any given time t can be obtained by solving Master equation of the Markov chain. The resulting state occupation probability, denoted by $p_k(t; \mathbf{w})$, is a function of time and also the state transitions.

This model can in essence be applied to any type of time-course data, including transcript or protein abundance. Here we will focus the description without loss of generality on gene expression data, unless otherwise stated. Let $x_j(t)$ be the cell-population-averaged gene expression of gene j at time t, obtained from assay such as RNA-seq or microarray expression. We assume that initially all cells in the population occupy the same state, this is often part of the experimental design of investigating changes from an initial homogeneous starting population. At any subsequent time point cells exist in a mixture of states, hence any measurement $x_j(t)$ made on a population level is an average over multiple states. We further assume that there is a mean expression level per gene constant across a state. This is denoted by β_{kj}, the gene expression level for gene $j \in \{1 \dots p\}$ in state k. In the limit of large numbers of cells the fraction of cells in any state k is given by the state

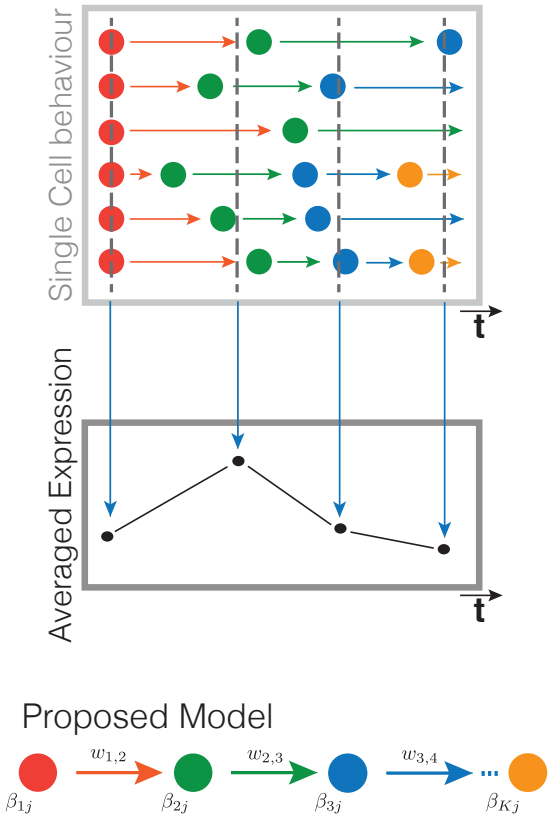


Figure 3.1: SOME TEXT

occupation probability $p_k(t; \mathbf{w})$. We can now write the observed average gene expression, $x_j(t)$ for gene j at time t, as the sum of all occupation probabilities weighted by their respective gene expression signatures. The resulting model for the average gene expression from a latent Markov chain model is written as:

$$x_j(t) = \sum_k p_k(t; (w))\beta_{kj} = P(t; \mathbf{w})\,\beta_j, \tag{3.1}$$

where the right hand side is the vectorized form of the model, with the row vector $P(t; \mathbf{w}) = [p_1(t, \mathbf{w}), \ldots, p_K(t, \mathbf{w})]$ and column vector $\beta_j = [\beta_{1j}, \ldots, \beta_{Kj}]^T$. Assuming an additive Gaussian noise model with gene-specific noise variance σ_j^2 we arrive at the likelihood:

$$\mathcal{L}\left(\mathbf{w}, \beta_j, \sigma_j \mid \{x_j(t)\}\right) = \prod_{t=1}^{T} \mathcal{N}\left(g(x_j(t)) | g\left(P(t; \mathbf{w})\,\beta_j\right),\ \sigma_j^2\right), \tag{3.2}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes a Normal density with mean μ and variance σ^2 and the function g denotes a transformation whose choice depends on the data type under investigation.

When investigating RNA-seq data we use $\mathrm{arcsinh}$ as the transformation [Hoffman et al., 2012; Johnson, 1949], defined as $\mathrm{arcsinh}(x) = \ln(x + \sqrt{(}x^2 + 1))$. RNA-seq data cannot be normalized in the same way as microarray data most importantly because it contains measurements which are exactly zero. The $\mathrm{arcsinh}$ normalization is useful here, because unlike the log transformation it does not have a singularity at zero but has the same variance-normalization properties.
To use the likelihood it is necessary to compute the state occupation probabilities at any time t observations are made.

Markov chain and the master equation

Until now we have not placed any restrictions on the latent Markov process in this model and we have formulated the likelihood eqn. (3.2) for a general case. If we let the states be $k \in |1 \dots K|$ and denote transitions between states k and k' as $\mathbf{w} = |w_{k,k'}|$. The topology of the Markov chain has implication on identifiability of the model (further discussion Section 3.1.2). Here we limit ourselves to a pure birth process where $w_{k,k+1} \neq 0$ for all k and zero otherwise. Such a Markov chain also excludes branches. The resulting master equation is written as:

$$\frac{dp_k(t)}{dt} = w_{k-1,k}\, p_{k-1}(t) - w_{k,k+1}\, p_k(t). \tag{3.3}$$

To simplify further calculations the master equation is rewritten in matrix notation. Let $\mathbf{G}(\mathbf{w})$ be a $K \times K$ matrix whose only non-zero entries are on the diagonal, $g_{kk} = -w_{k,k+1}$, and the subdiagonal $g_{k\,k-w} = w_{k-1,k}$. Now we can write the master equation as

$$P(t; \mathbf{w}) \quad = \quad \exp\left(\mathbf{G}(\mathbf{w})\, t\right) P(0) \tag{3.4}$$

In investigating transition processes (such as Section REFERENCE HERE) in general an experimental design is chosen such that the initial cell population is in the same state. Therefore we can set the initial conditions for the state occupation probability, $P(0) = (1, 0, 0, \dots)$. This means all cells are in state $k = 1$ at $t = 0$ just before the cell population is perturbed. This allows us to write the closed form solution for the state occupation probability as:

$$P(t; \mathbf{w}) \quad = \quad \exp\left(\mathbf{G}(\mathbf{w})\, t\right) P(0). \tag{3.5}$$

This expression is also used to evaluate the likelihood (equation (3.2)) of the model for different parameters.

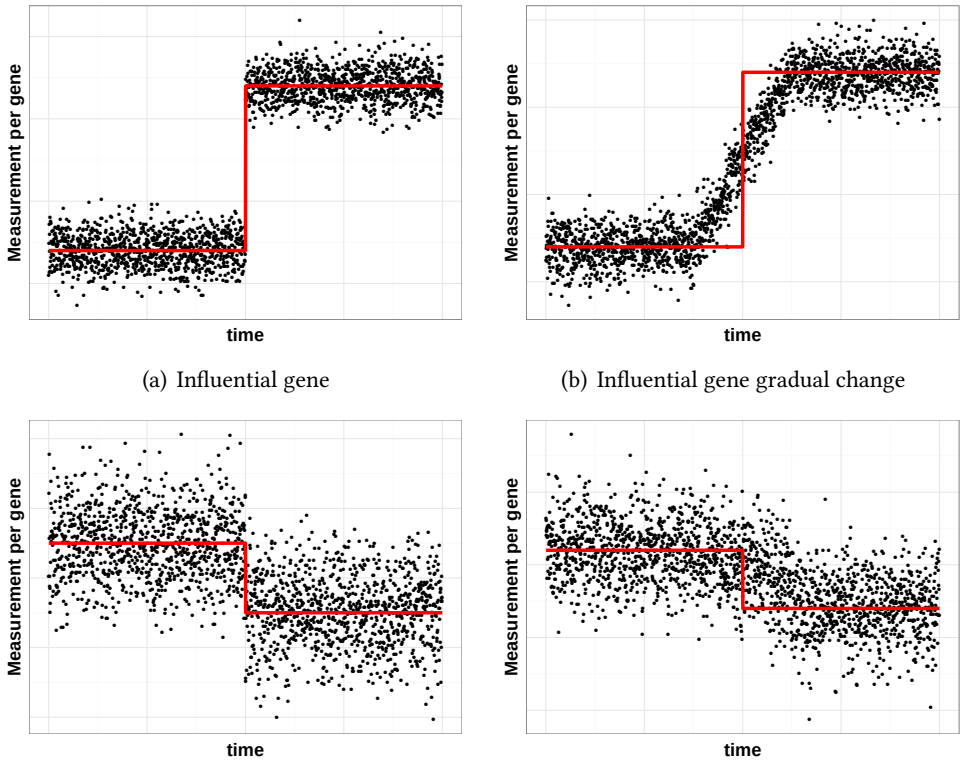
Model Assumptions

We make a number of assumptions in the above model derivation. Here, we focus on some of the key assumptions made regarding the transition process on a single cell level and investigate them further. Ensuring an analytically tractable latent state change model makes these assumptions necessary. In the discussion below we discuss how legitamate these assumptions are, if and how they can be relaxed and how they can be justified.

First, we assume expression of a gene remains constant while it remains inside a given state. The single cell expression of each gene is modeled by a piecewise flat trajectory where expression changes are instantaneous due to a change in state. It also has the effect that the only time-dependence in the likelihood is due to the state occupation probabilities of the Markov chain. In this simple approximation, interaction between genes are ignored; allowing us to formulate a computationally efficient pipeline to estimate parameters for time courses with many genes, see Section 3.2.3 for further details. It is a very strong assumption and apart from the noise that is prevalent in most biological systems this does not hold in general. Temporal changes within a state should be much smaller than the difference between biologically distinct states for genes influential in such a transition. This case is illustrated in Figure 3.2(a) and 3.2(b). Therefore this is still a good first approximation in the case of transition processes.

A second assumptions relates to the topology of the Markov chain. To ensure parameter identifiability (see Discussion Section 3.1.2) we have to restrict the latent process to a linear pure birth process. This restricts the topology of the Markov chain quite drastically, but is arguably defensible when applied to externally driven transition processes. The external drive can take many different forms, in the two examples we investigate it is genetic induction. Of course back transitions are likely, for such cases our model is mis-specified and the forward transitions are only effective values where the back transitions have been absorbed into the model. Consequently estimated forward transitions rates are lower than the real values. On occasion back transitions or topologies of the latent process are of interest. The likelihood eqn. (3.2) is general and does not make any assumptions about the topology of the Markov chain, but additional data or constraints would be required for identifiability of more complex transition topologies. Often the limiting factor is available data hence we focus here on the more useful but special case where only time-course data is available and the latent stochastic process is a linear birth process. In Section 3.3 we include a detailed investigation the impact of breaking this assumption in a simulation.

Finally, we assume rates of cell death and cell duplication cancel each other out and the population therefore remains roughly constant in time. Consequently the fraction of cells in a given state only depends on the transition rates between the states. Especially in the case of oncogenic transformation (Section 4) this is clearly not the case since tumorous



(a) Influential gene

(b) Influential gene gradual change

(c) Gene not influential

(d) Gene not influential gradual change

Figure 3.2: Illustration. First assumption made is that gene expression remains constant for a gene while it remains inside a state. The model Section 3.1.1 describes the single cell measurement of a gene transitioning between states as an instantaneous step change (red line). In reality the measurement will at least fluctuate and transition won't be instantaneous. For influential genes (a) - (b), this assumptions is reasonable whether or not the transition is instantaneous, the points here are meant to represent single measurements. Genes where within state temporal changes are comparable to between state changes, (c) - (d), the approximation is not good. These genes are not influential for the transition process therefore this is not problematic.

cells in general have a much higher proliferation rate. In Section 3.3 we test how well parameters are estimated when this assumption is violated.

3.1.2 Identifiability

3.2 Estimation

3.2.1 Parameter estimation

We begin by stating the maximum likelihood estimates (MLEs) based on the likelihood eqn. (3.2):

$$(\{\hat{\beta}_j\}, \hat{\mathbf{w}}) = \underset{\{\beta_j\},\mathbf{w}}{\operatorname{argmin}} \sum_{j=1}^{p} \sum_{t=1}^{T} \|g(x_j(t)) - g\left(P(t;\mathbf{w})\,\beta_j\right)\|_2^2, \tag{3.6}$$

where $\|\cdot\|_q$ denotes the ℓ_q norm with respect to its argument. The transformation g is in general non-linear as discussed above (e.g. $\log$ in microarrays or $\operatorname{arcsinh}$ in RNA-seq); for such transformations the MLE (3.6) cannot be obtained in closed form. Genome wide measurements yield readings with number of genes, p, of up to 10^4. Directly optimizing eqn. (3.6) is not practical for a problem with large p. TODO include plot. We adopt a two-step estimation procedure proposed by Armond et al. [2013]. The first step is based on the observation that many genes have similarities in their measured time-courses; this allows us to cluster genes obtaining m clusters describing typical temporal patterns. Details for choosing the parameter m are discussed in Section 3.2.3. The m cluster centroids are used to estimate the transition rates $\mathbf{w}$ via eqn. (3.6), instead of all genes. This approach reduces computation time significantly when $m << p$. The transition rates estimated using cluster centroids, $\hat{\mathbf{w}}$, are fixed and the β values for all remaining genes are estimated. The MLE is now written as:

$$\hat{\beta}_j = \underset{\beta_j}{\operatorname{argmin}} \sum_{t=1}^{T} \|g(x_j(t)) - g\left(P(t;\hat{\mathbf{w}})\,\beta_j\right)\|_2^2 + \lambda\|\beta_j\|_1 \tag{3.7}$$

where the final term is an (optional) ℓ_1 penalty with tuning parameter λ. It is invoked when potential over-fitting needs to be counteracted (choice of λ is discussed in Section 3.2.3).

The optimization eqn. (3.7) greatly simplifies estimation (compared to eqn. (3.6)), since estimation for individual β_j for gene j can be performed independently. This is possible because time-courses between individual genes are only coupled by transition rates $\mathbf{w}$; once they are fixed, individual gene trajectories can be examined independently.

3.2.2 Model selection

The estimation steps described in Section 3.2.1 apply to a model with a fixed number of states K. Here we present a procedure to determine the number of states K that best represent a data set under investigation. Depending on the application K itself can be of scientific interest. In general estimated state-specific expression signatures, β, are influenced by the number of states. Underestimating number of states results in distinct states being merged. Overestimating the number of states introduces artificial states in the transformation. Both scenarios lead to poor estimation of parameters.
In general model selection can be performed using a form of cross-validation (CV) by leaving out part of the data as a validation set. TODO include BIC AIC reference move to supplement. In applications to time-series, cross-validation is often non-trivial due to discrete and irregularly spaced observations. The MAST model has an underlying continuous-time latent process, which allows for prediction of any time points from estimated parameters; therefore comparison between predicted time-points from estimated parameters and the corresponding held-out time-points. In this application due to poor time resolution it is often not possible to include more than one time point in the validation data; this variant is called leave-one-out cross-validation (LOOCV). If t is the held-out time point let the estimated parameters for the remaining subset be rates $\hat{\mathbf{w}}^{-t}$, state specific expression $\{\hat{\beta}_j^{-t}\}$ and gene-specific standard deviation $\{\hat{\sigma}_j^{-t}\}$. State occupation probabilities at the held-out time point $P(t; \hat{\mathbf{w}}^{(-t)})$ are obtained by solving the master equation using estimates derived from the training data. There we can now write a prediction for the expression of gene j at the held-out time point $\hat{x}_j^{\mathrm{CV}}(t) = P(t; \hat{\mathbf{w}}^{(-t)})\, \hat{\beta}_j^{(-t)}$ and the cross-validation mean squared error ($\mathrm{MSE}_{\mathrm{CV}}$) is simply

$$\mathrm{MSE}_{\mathrm{CV}} = \sum_{t=2}^{T} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^{(-t)}} \big(g(\hat{x}_j^{\mathrm{CV}}(t)) - g(x_j(t))\big)^2. \tag{3.8}$$

The strength of this type of model selection in comparison to the Bayesian approach presented in Armond et al. [2013] is twofold. Firstly application of the computationally efficient estimation procedure outlined in Section 3.2.1, allows this cross-validation procedure to be applied to the whole data set efficiently. Secondly it doesn't require parameters to be set by the user except those required for estimation. The Bayesian approach requires a computationally demanding Monte Carlo estimation and has several hyper-parameters which have to be set by the user.

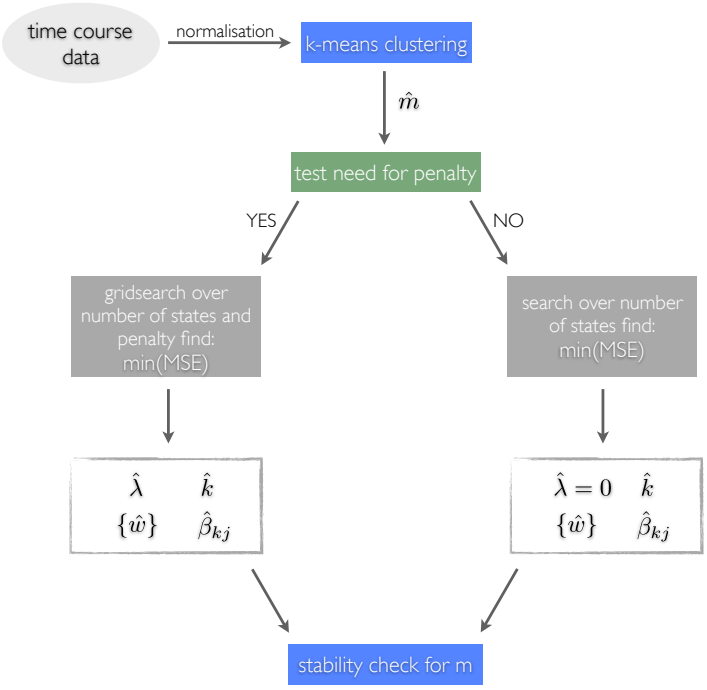


Figure 3.3: Schematic of estimation pipeline.

3.2.3 Estimation pipeline

We now present a computationally efficient pipeline for setting tuning parameters required for estimation. The pipeline is also summarized in Figure 3.3. The required tuning parameters are:

- The number of clusters, m, used in the first step of the two-step estimation.
- The strength of the penalty term, λ, applied in eqn. (3.7), where $\lambda = 0$ is equivalent to no penalty.
- The number of states in the latent Markov chain, K.

Number of cluster: In the initial estimation step we cluster gene expression trajectories which results in cluster centroids describing typical trajectories; these permit estimation of transition rates. In empirical results (see Section 3.4.2) we see; if the number of clusters is large enough to capture most of the information in typical trajectories changing m does not have a significant impact on parameter estimation. Therefore we set m using a simple k-means algorithm and inspecting the relative decrease of within-cluster sum of squares objective $J(m)$ as a function of m:

$$\Delta J(m) = \frac{J(m-1) - J(m)}{J(m-1)}$$

We select $\hat{m}$ such that $\hat{m} = \min\{m : \Delta J(m-1) < 0.1\}$, i.e. if the relative decrease in the objective function is smaller than 0.1 for $m-1$ we choose m as the number of clusters. Small fluctuations in the objective function for higher m lead to instabilities in the relative decrease. Once $\hat{m}$ is set we include a post-estimation sensitivity test for the choice of m. In section 3.4.2 we demonstrate with the help of empirical results, how well behaved and computationally efficient this model is. Though it should be noted that the choice of m can be made with any clustering method and the corresponding objective function.

Penalization: The penalty term introduced in eqn. (3.7) is useful in high-dimensional models; even though the data investigated using this model will often be high dimensional, estimation is carried out separately for each gene. Therefore penalization may not be required unless the number of time points is large. Consequently we introduce an additional step to test the need for penalization by comparing estimated expression signatures β with estimates obtained from leaving out individual time points. We specify stability as a Pearson correlation between estimated β values greater than 0.8. If we deem a data set stable under such a test there is no need for penalization and we choose $\lambda = 0$. If the correlation is smaller than 0.8 a penalty term is required (setting the penalty strength is discussed below). In both the simulated data sets and application to oncogenic transformation penalization was not required and $\lambda = 0$ in Sections 3.4 and 4.

Number of states: The final parameter to be set is the number of states in the latent Markov chain. This parameter is set by minimizing the CV score ($\mathrm{MSE}_{\mathrm{CV}}$) for a range of different K. If penalization is required $\mathrm{MSE}_{\mathrm{CV}}$ is minimized by performing a grid search over both λ and K.

All three parameters (m, λ, K) can in principle be set by performing a grid search with respect to $\mathrm{MSE}_{\mathrm{CV}}$, but this is computationally very challenging and would make estimation impractical. Our pipeline make use of heuristic observations to reduce the grid search to one dimension. The observation that estimates are robust to the choice of the number of clusters allows us to remove m from the grid search. Observing that the penalty term is not always needed enables us to exclude λ from the grid search. When choosing m using a clustering method it could be that transition rates haven't converged. Therefore we carry out an additional diagnostic post estimation. We check correlation values between expression levels β estimated using $\hat{m}$ clusters and estimated using larger values $m' > \hat{m}$. Provided results have Pearson correlation above 0.8, the choice of $\hat{m}$ was appropriate, if the correlation is below 0.8 we repeat the pipeline with larger m.

3.3 Simulation setup

To test the validity of the model we need to test it with simulated data where true parameters are known. This will allow both evaluation of strengths and weaknesses in parameter estimation and model selection (choosing number of states). Simulations are performed not at the cell population level of the likelihood eqn. (3.2) but at the single-cell level; allowing for extensive testing of model assumptions. The single cell trajectories are then averaged to obtain homogenate data analogous to RNA-seq data.
Here we describe the step by step simulation procedure for a K state model independent of the number of genes simulated:

State transitions. When setting transition rates between discrete states of the Markov chain we need to keep a few things in mind. Firstly the smallest sampled (observed in real data) time step needs to be smaller than the transition rates. Just like in typical experimental designs for transition processes. Additionally the model won't be able to extract information about a process taking place on a time-scale smaller than gaps between observations. Secondly we are considering transitions processes driven towards an established final state (e.g. oncogenic transformation, pluripotency); so to mimic this behavior in simulated data we need to insure the occupation probability for the final state is higher than the others at the final time point. Of course in realistic experiments even at the final time point the cell population will still be heterogeneous In the discussion that follows in Section 3.4 we use the three transition rates $[1/5, 1/8, 1/15]$ for four state model. In Figure 3.4 we show the state occupation probabilities for these parameters and $k = 4$ has an occupation probability of ≈ 0.64 at the final time point. For every cell in the simulation, state transitions are simulated by drawing jump-times from exponential distribution with parameters given by transition rates as defined for a continuous-time Markov chain.

State-specific expression levels. For all cells, each gene j and state k we set gene expression levels β_{kj}; per gene the expression levels are set to zero with probability $1/K$ otherwise they are sampled uniformly from $(0, \gamma_j]$. Parameter γ_j, chosen from the range $[1, \ldots, 12000]$, effectively sets the scale of gene j [1]. This method ensures simulated trajectories for genes on different scales (see Figure 3.5(a) and the corresponding gene expression signatures Figure 3.5(b)), to emulate real RNA-seq data where a range of five order of magnitude was observed [Wang et al., 2009; Mortazavi et al., 2008]. Gene expression trajectories for single cells are piecewise flat for each gene once β values are sampled. Changes in trajectories only occur at jumps between states and are instantaneous.

[1] It is always chosen from the following $\gamma_j = \{1, 10, 50, 100, 200, 500, 1000, 2000, 4000, 7000, 10000, 12000\}$

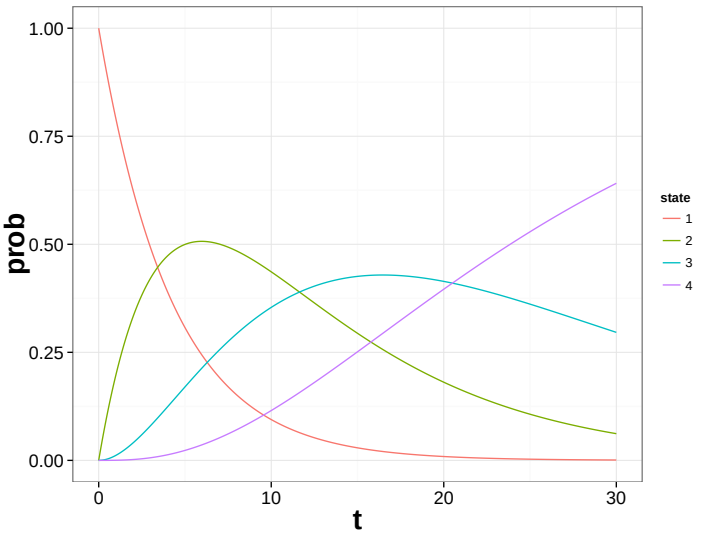


Figure 3.4: Simulation study. State occupation probabilities for a four state model.

Aggregation and time-sampling. For each gene j each cell has an associated gene expression trajectory. Similar to RNA-seq experiments where observations are averages of gene expressions over many cells; these trajectories are averaged over a large number of cells to give an average gene expression trajectory. The occupation probability in the model outlined in Section 3.1.1 is derived in the limit of number of cells $\to \infty$, of course in practice the number of cells is finite. We set the number of cells to 1000 which serves as a good test of the limiting assumption.
The simulated time-course is obtained by sampling the simulated trajectories at discrete unevenly distributed time points. Finally Gaussian noise is added to the transformed data (see Section 3.1.1 for details, for RNA-seq $\operatorname{arcsinh}$) with mean zero and standard deviation σ; which we set to $\sigma = 0.2$ unless states otherwise, this provides a reasonable signal to noise ratio for all observations. Similar to the RNA-seq data discussed in Section 4 we choose 15 unevenly spaced time points at $t = \{0, 2, 4, 7, 8, 11, 14, 20, 24, 29, 32, 35, 40, 44, 48\}$. The simulation setup is summarized in pseudocode in Algorithm 1.

3.4 Results

We present results from simulations in two separate phases. For both simulation setup the transition rates are fixed at $[1/5, 1/8, 1/15]$ and we simulate from a model with four states in latent Markov chain. First in a small scale simulation with $p = 9$ genes we perform multiple rounds of direct estimation of the whole data without the need for the initial clustering step. This simpler simulations allows investigation of identifiability and an

Algorithm 1 Pseudocode for single cell simulations

procedure SIMULATION($n.states, n.genes, n.cells, r, p, \tau, dt$)
  $\beta \leftarrow NB(r; p)$
  $jump.t \leftarrow Exp(1/\tau_k)$
  for all $genes, cells$ **do**
    for $t \leftarrow 0, T$ **do**
      while $t < jump.t_{states}$ **do**
        $sim.traject(t) \leftarrow \beta_{j,k}$
      end while
    end for
  end for
  average $sim.traject$ per gene for all cells
  $sim.data \leftarrow sim.traject$ (sampled at discrete time)
  $sim.data \leftarrow sim.data + \mathcal{N}(0, \sigma)$
end procedure

investigation for model selection without considering the two-step estimation procedure outlined above.
Than we consider a larger scale simulation with $p = 120$ genes, where we put the full two-step estimation procedure to the test; including clustering, setting of tuning parameters and finally model selection.

3.4.1 Small scale simulation

Using the small scale simulation we perform three separate tests. One in which we only estimate transition rates and state-specific expression levels; we consider the number of states to be known. Then we consider the model selection problem and finally we investigate estimation under breaking model assumptions.

Number of states known

We simulated 9 genes from a 4 state model as described in Section 3.3. In this small simulation we do not use a penalization term, i.e. $\lambda = 0$. In Figure 3.5(a) we show trajectories for one such realization, here the thicker line represents trajectories from averaging 1000 cells for each gene. The green dots show sampled data with the addition of Gaussian noise to transformed data. In Figure 3.5(b) we show state-specific gene expression signatures for all 9 simulations. The values are shown in pairs of true and estimated. The value on the right is in each case the true value used in simulating the trajectories. The left-hand value is estimated by fitting the 15 time point of the simulation. The corresponding estimated and true transition rates for this realization can be seen in Table 3.1. Using the estimated

transition rates and the expression signatures we can obtain an estimated trajectory, seen as a blue line in Figure 3.5(a).

Transition rates	$w_{1,2}$	$w_{2,3}$	$w_{3,4}$
true mean	0.200	0.125	0.067
estimated	0.236	0.114	0.068

Table 3.1: Transition rates used in the simulation and the estimated values

We repeat fifty such independent simulations at four different noise levels [2], each time β_{kj} are resampled as described above (Section 3.3) while transition rates are shared across simulations. We compute the correlation between estimated and true gene expression signatures for each simulation run $\rho(\beta, \hat{\beta})$. The correlation coefficients for all simulations are summarized in a boxplot, Figure 3.6(a). For all tested noise levels we compute a mean and standard deviation of the correlation coefficient across all fifty runs in Table 3.2; The mean is above 0.9 for all simulations and the highest level for the variance is 0.13. Therefore we can conclude that that state-specific gene expression signatures are recovered well in the simulation. We also introduce a new measure, $s_k = |\hat{w}_{k,k+1} - w_{k,k+1}|$ to test recovery of transition rates. For each simulation we use the mean $\bar{s}$ over the three transition rates as measure for how well transition rates are recovered. In Figure 3.6(b) we show boxplots for the fifty simulations for each of the four noise levels. We find that transition rates are also recovered well, though as expected the estimates become worse with increasing noise levels.

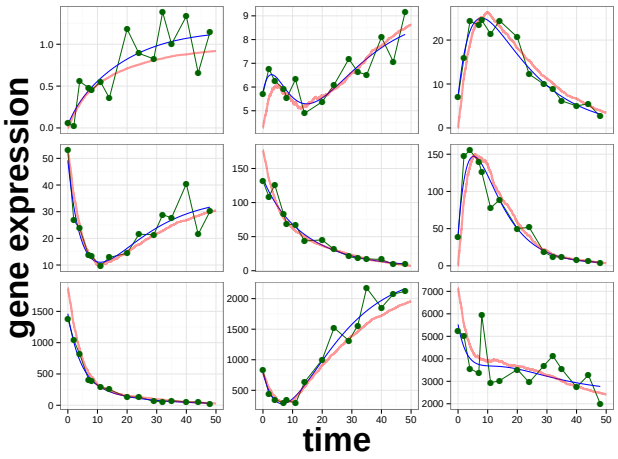
σ	0.05	0.1	0.15	0.2
mean	0.95	0.93	0.94	0.91
std. dev.	0.07	0.09	0.08	0.13

Table 3.2: Correlation between true and estimated gene expression signatures. Mean and standard deviation are estimated across 50 independent simulations.

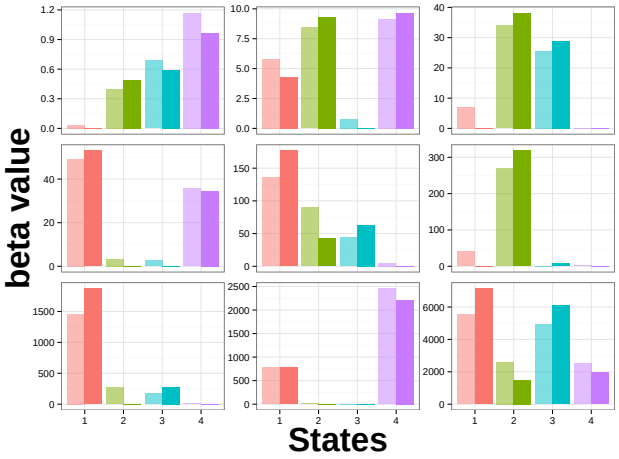
Determine number of states

Next we consider the problem of model selection in this small simulation setup. We simulate data as described above for $p = 9$ genes. In such a model with a latent stochastic process, model selection is a challenging problem especially using noisy data sets. Therefore to test model selection we fifty independent simulations for each of the following noise regimes: $\sigma = \{0.05, 0.1, 0.15, 0.2\}$. We compare models with $K = 1 \ldots 5$ and perform model selection using leave-one-out cross-validation (see Section 3.2.2), for each of

[2] $\sigma = \{0.05, 0.1, 0.15, 0.2\}$d

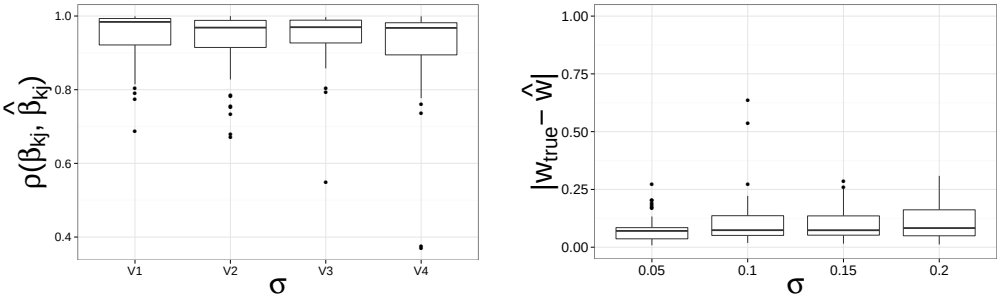


(a) Simulated Trajectories for $p = 9$



(b) Expression signatures for $p = 9$ simulation

Figure 3.5: Simulation study. Small scale simulation for $p = 9$ genes. (a) shows the trajectories for these simulations. The thick red line shows the averaged trajectories over 1000 cells. The green dots show 15 sampled data points with normal noise ($\mathcal{N}(0, \sigma)$, with $\sigma = 0.2$) added to the average data. The blue thin line shows the trajectory from estimated parameters. (b) shows state-specific gene expression signatures for all 9 simulated genes. The true and estimated parameter values are shown next to each other. The lighter color on the left shows estimated parameter values, the solid colors shows true parameter values.



(a) State-specific gene expression

(b) Transition rates

Figure 3.6: Simulation study. Small scale simulation using $p = 9$ genes with 50 independent repeats. Boxplots show results over all repeated simulations at four different noise level $\sigma = \{0.05, 0.1, 0.15, 0.2\}$. (a) Boxplots for correlations between estimated and true gene expression signatures ($\rho(\beta_{\text{true}}, \hat{\beta})$) at four different noise levels. (b) Boxplots for the mean of absolute differences between the estimated and true transition rates $\bar{s}$ for each simulation at four different noise levels.

the fifty simulations. We determine the minimal MSE_{CV} scores (eqn. (3.8)) for different models and juxtapose a comparison between the different models using a simple normalized MSE score for model fit without held-out time points. In each simulation and for each noise regime we determine the model with lowest MSE_{CV} score and lowest MSE score. Then we show the distribution of these minimal scores over the selected number of states in Figure 3.7; the top row shows the distribution for MSE_{CV} and bottom row show the distribution for MSE in different noise regimes.
Here number of parameters increase with number of states, and as a result model fit improves; therefore as expected at all noise levels the maximum number of states ($K = 5$) results in the best fit.

Violating model assumptions

Until now we have considered simulations with a correctly specified model where assumptions underlying the model are not violated. Breaking these assumptions is especially easy in the single cell simulation. We investigate consequences on parameter estimation under violation of a subset of these assumptions. We use three types of plots to investigate parameter estimation for these simulations.

- **Correlation.** For state specific gene expression signatures β_{kj} we compute the Pearson correlation coefficient between true parameters and estimated parameters, $\rho(\beta_{\text{true}}, \hat{\beta})$.

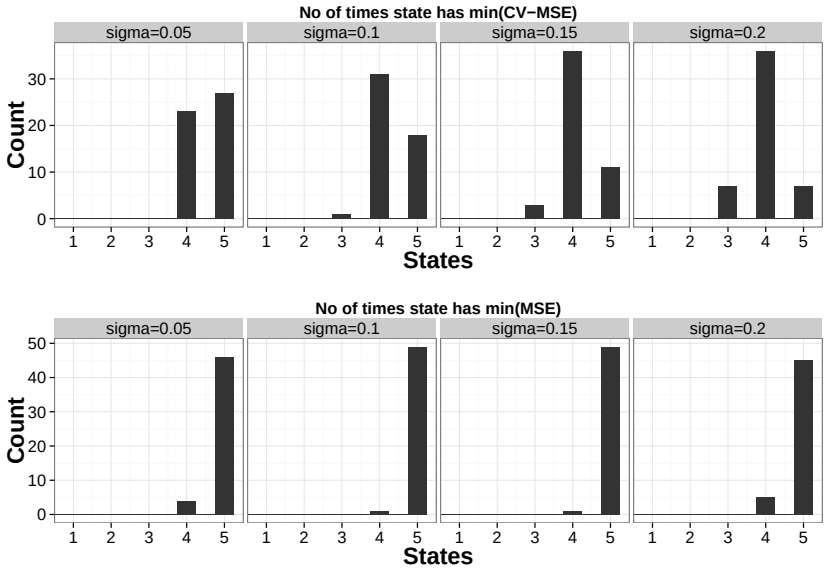


Figure 3.7: Simulation study. We perform fifty independent small scale simulation with $p = 9$.

- **Transition times.** We show boxplots of estimated average transition times for 10 simulations and a horizontal dashed line to represent the true value used in the forward simulation.

- **Probability** In the model itself the transition times do not enter directly they are used to calculate probabilities. We compare the values by calculating a mean difference between probabilities:

$$\langle |\hat{p}_k(t) - p_k(t)| \rangle_{k,t} \,, \tag{3.9}$$

  where k is the number of states, $\hat{p}_k(t)$ is the probability calculated from estimated parameters and $p_k(t)$ is the probability calculated from true values. The average is taken over both the states and time.

Cell death and cell doubling

An assumption we make in MAST is that cell death and cell duplication happens at a constant rate across all states in the transformation process. This is of course not the case in the discussed example of oncogenic transformation, since transformed tumorous cells have a much higher proliferation rate than the initially healthy cells. In the single cell simulation setup we sample a time of death, t_i^d, and a time for cell doubling, t_i^{dup}, from an exponential distribution. If sampled rates for a cell are outside of the time range of

the simulation, the cell remains unchanged. If they are both in the range there are two possible scenarios. Firstly if the death rate is the smaller of the two, cell i is taken out of the simulation $t > t_i^d$. Secondly if $t_i^{dup} < t_i^d$, cell i is taken out of the simulation at $t > t_i^{dup}$ and two new cells are simulated with new sampled state transitions. The simulation and estimate is performed 10 times. Investigating the oncogenic transformation discussed in the paper it was observed that generally cells have a doubling rate of close to 0.05 i.e. doubling of cells is roughly every 20 hours. In Figure 3.8 we fix the doubling rate and since cells in this experiment rarely die, we choose very small death rates. The left panel shows the average as a dark line and the shaded area represents the standard deviation for the 10 repeated simulations. The middle panel shows boxplots for the estimated average transition times. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.
In Figure A.1 we include additional cell doubling rates. In general gene expression signatures are estimated well, but the transition rates are not. During estimation transition rates only enter as probabilities hence badly estimated transition rates don't have a significant negative effect on the estimation of expression signatures. To see what effect the different parameters have on the number of cells simulated at any given time Figure 3.9 show the number of cells as a function of time for different cell death and cell doubling rates.

Back transitions The second assumption we test is the inclusion of back transitions in the single cell. We simulate trajectories with back transition from $k = 4$ to $k = 3$; they are sampled from exponential distributions with different means. In Figure 3.10 we show comparisons between estimated and true values of parameters as a function of the average back transition time from state $k =$. In the left panel we plot the average correlation for 10 independent runs, between true and estimated β_{kj} parameters as a solid line. The shaded area shows the standard deviation. The vertical dashed red line shows the average forward transition time for $k = 3$. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

Markovian assumption Finally in this section we investigate the latent Markov process and consider a case where jumps are non-Markovian. We want to consider the more realistic case that the transition time is fat tailed; therefore we choose a truncated Student t-distribution with a variety since transition rates are positive. We sample using the *rtmvt*

package in R with degrees of freedom 1, the package also allows specifying the variance of the sample using σ. We perform the simulation as before, but sample transition rates from the t-distribution with means $(1/5, 1/8, 1/15)$ and consider a range of σ parameters in *rtmvt*. The results are shown in Figure 3.11 and the varied parameter is σ for each of the plots. The left panel shows the mean correlation between true and estimated β_{kj} as a solid line and the shaded area constrains the standard deviation. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

3.4.2 Large scale simulation

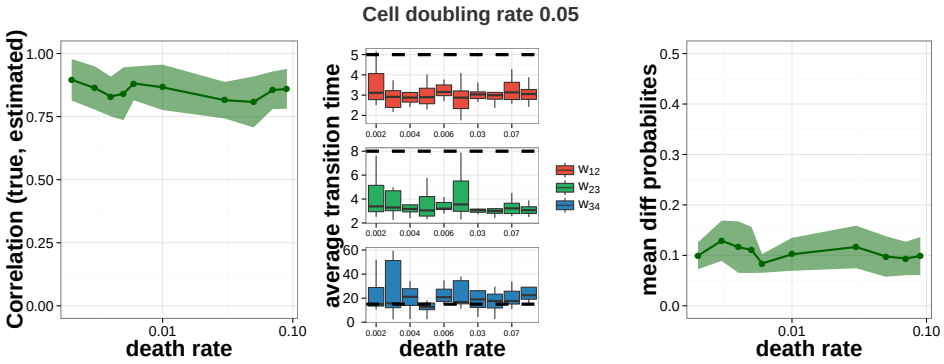


Figure 3.8: Simulation study. Testing assumption about cell death and cell doubling. For each cell a time of death, t_i^d, and a time for cell doubling, t_i^{dup}, is sampled from an exponential distribution with varying average rates. If the sampled rates for a cell are outside of the time range of the simulation, the cell remains unchanged. If they are both in the range there are two options. The first option is that the death rate is the smaller of the two in that case the cell i is taken out of the simulation $t > t_i^d$. If $t_i^{dup} < t_i^d$, cell i is taken out of the simulation at $t > t_i^{dup}$ and two new cells are simulated with new state transitions. The simulation and fit is performed 10 times. In experiments we observe cell doubling time to be roughly 18 hours and very few dead cells. Therefore we simulation with a cell doubling rate of 0.05 and a variety of death rates. The left panel shows the average as a dark line and the shaded area represents the standard deviation for the 10 repeated simulations. The middle panel shows boxplots for the estimated average transition times. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

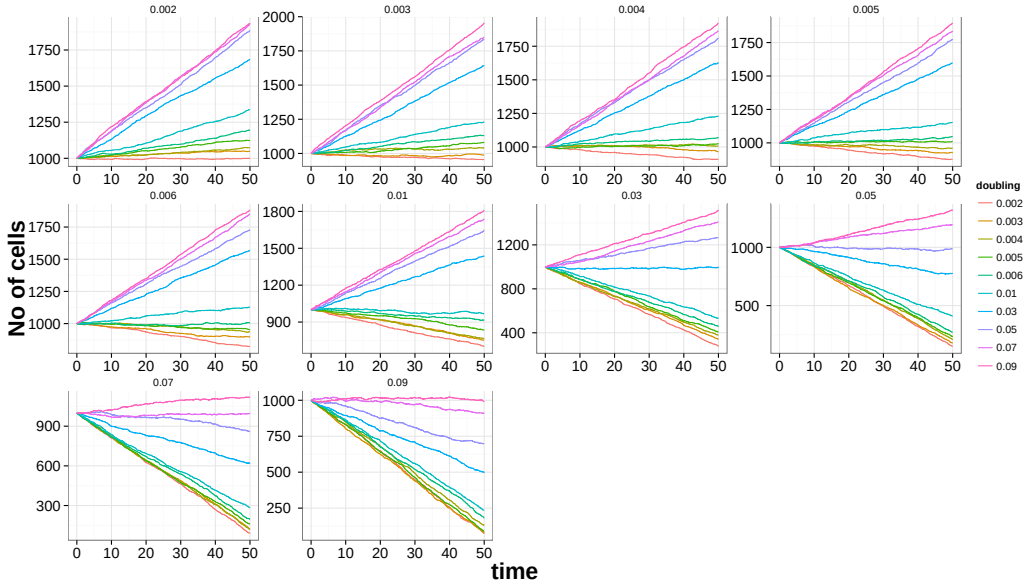


Figure 3.9: Simulation study. Testing assumptions about cell death and cell doubling. Plots show number of cells at different time during the simulation for one of the 10 simulations. Each panel represents different death rates and each coulour different doubling rates.

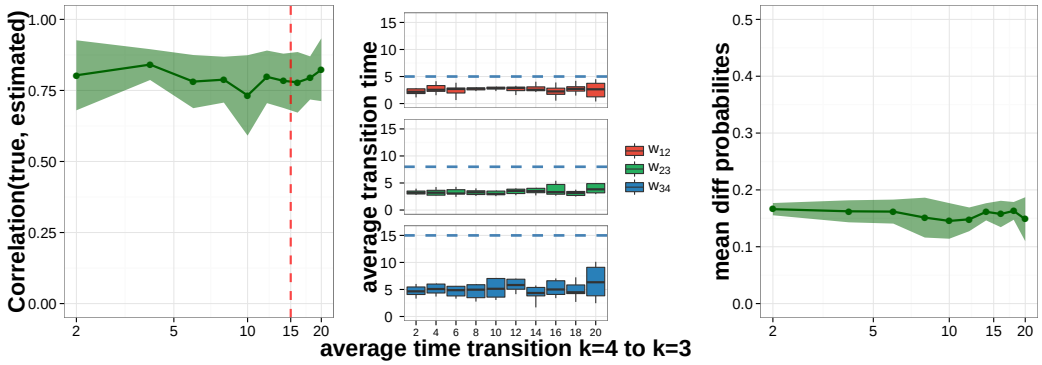


Figure 3.10: Simulation study. To test the affect of back transition on the estimation, we simulate trajectories with back transition at $k = 4$ with different transition times. In the left panel we plot the average correlation for 10 independent runs, between true and estimated β_{kj} parameters for different average time for the back transition as a solid line. The shaded area shows the standard deviation. The vertical dashed red line shows the average forward transition time for $k = 3$. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

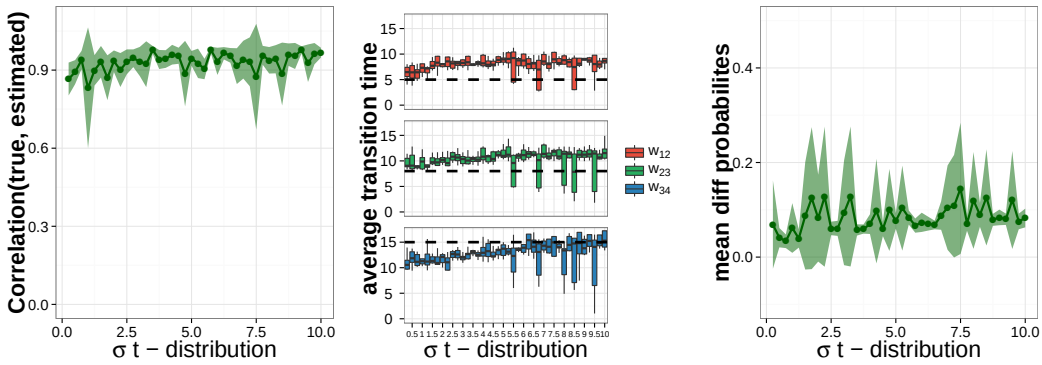


Figure 3.11: Simulation study. We simulate from a non-markovian system, one where average transition rates are heavy tailed. Here we sample from a truncated Student t-distribution using the *rtmvt* package in R. It is truncated at zero since transition rates are always positive. The transition rates are sampled to have means $(1/5, 1/8, 1/15)$ and we vary the σ parameter in the package used. In the left panel we show the average correlation between true and estimated β_{kj}, the mean is shown as a solid line and the standard deviation as a shaded area. The middle panel shows boxplots for the for the average transition time estimated from the model for a system with $K = 4$ states. The right panel shows the mean differences between estimated and true state occupation probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

Chapter 4

Oncogenic Transformation

Chapter 5

Stem cells

Chapter 6

Cell cycle

In the model outlined in Chapter 3 and applied to two biological systems in Chapters 4 and 5 the initial cell population is assumed to be homogeneous. In the two applications discussed before this assumption is warranted due to experimental design. In the case of the oncogenic transformation the experiment is started from a cell line ensuring homogeneity. In the case of stem cell reprogramming the technique outlined by Hanna et al. [2009] tries to ensure initial homogeneity by using a secondary MEF cells.
Recently it has been shown that even seemingly homogeneous cell populations have an inherent mixture be it at an epigenetic level [Heng et al., 2009; **?**]. In this Chapter we setup a model that answers the question: What effect does the initial cell population have on cell fate.

Figure 6.1: Schematic of heterogeneous cell population transforming under stimulus.

An example of such a biological system (schematic Figure 6.1) is one with an initial heterogeneous cell population made up of two types of cells, with an indistinguishable phenotype. At time $t = 0$ the cells receive a stimulus leading to a transformation such that at $t = T$ it is possible to distinguish cells in their final cell fate. Now it is possible to count the fraction of cells that reach each of those final cell fates. The interesting case here is when the strength of the stimulus has an affect on the fraction of cells in each cell fate. We are interested in genes whose expression varies

6.1 Formalization

To formalize the model we start with

Appendices

Appendix A

Additional results MAST

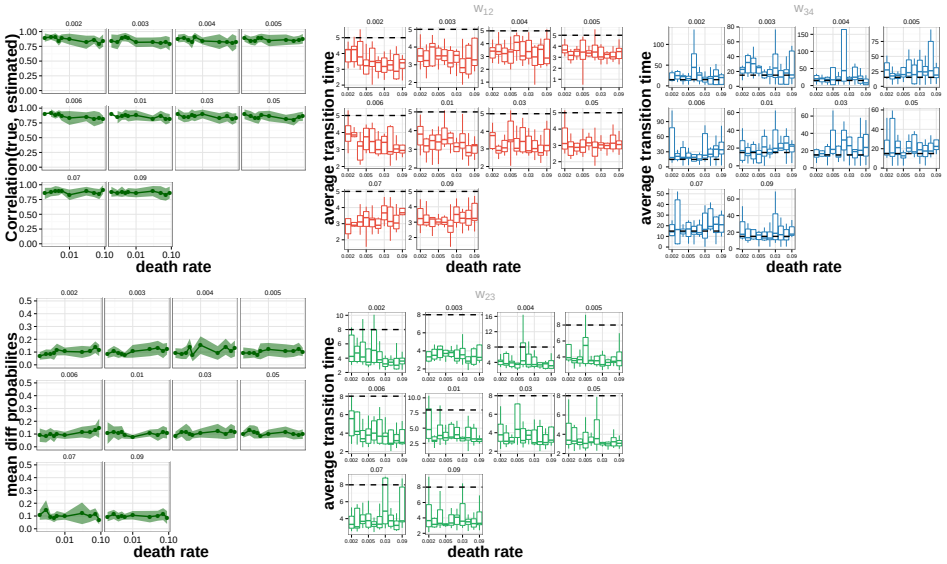


Figure A.1: Simulation study. Extend figure 3.8 to include a wider range of doubling rate. The top left panel shows the correlations between true and estimated β_{kj} as a function of death rates. The bottom left panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.Different panels show a range of doubling rates. The remaining panels show boxplots for the estimated average transition times the horizontal dashed line shows the true values used in estimation.

Bibliography

JW Armond, K Saha, AA Rana, CJ Oates, R Jaenisch, M Nicodemi, and S Mukherjee. A stochastic model dissects cell states in biological transition processes. *submitted*, 2013.

Jacob Hanna, Krishanu Saha, Bernardo Pando, Jeroen van Zon, Christopher J Lengner, Menno P Creyghton, Alexander van Oudenaarden, and Rudolf Jaenisch. Direct cell reprogramming is a stochastic process amenable to acceleration . *Nature*, 462(7273):595–601, March 2009.

Henry H Q Heng, Steven W Bremer, Joshua B Stevens, Karen J Ye, Guo Liu, and Christine J Ye. Genetic and epigenetic heterogeneity in cancer: A genomeâĂŘcentric perspective. *J. Cell. Physiol.*, 220(3):538–547, 2009.

Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Meth*, 9(5):473–476, 2012.

N L Johnson. Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 36(1/2):149–176, January 1949.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628, July 2008.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10:57–63, 2009.