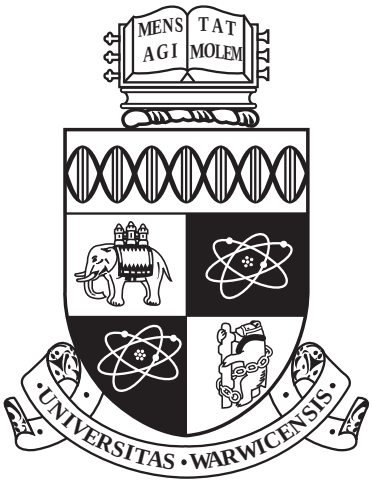


Stochastic models for cell populations undergoing transitions

by

Anas A. Rana

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Physics and Complexity Science

... .

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

Here we set out a description of background material which should prove useful to the reader. This is thesis is multidisciplinary and as such requires an introduction to two distinct areas. Therefore this chapter is split into two separate sections. First the Section 2.1 includes a mathematical background for the main techniques used in the thesis and some important ideas. Section 2.2 contains some background to the main Biological ideas discussed in the thesis.

2.1 Mathematical background

This section starts off by examining Markov chains the main building block of the model introduced in Chapter 3. Then in Section 2.1.2 we introduce the slightly more advanced hidden Markov models before carrying on to a precursor to a model discussed in Chapter 3: the Aggregated Markov chain in discrete time. Then in Section 2.1.4 the estimation procedure is introduced and briefly discussed followed by a discussion on penalisation of estimation in Section 2.1.5. The concept of identifiability of a model and what it means in a context for estimation is outlined in Section 2.1.6. Then we move on to a discussion on model selection and various techniques for distinguishing between models. Finally in Section 2.1.8 we discuss a very useful method for numerical integration used in the second model put forward in Section 6.

2.1.1 Markov Chains

In Physics prior to the advent of Statistical Physics and Quantum Mechanics in the early 20th century the world was modelled as deterministic. Of course we now know that despite many aspects of the observable world being deterministic there is an even larger set of objects which does not lend itself to a deterministic description. Stochastic processes

are often used to describe such systems evolving with a time-dependant stochastic parts. One of the first such attempts was the modelling of Brownian motion by Einstein [1905] which paved the way for further research on this topic. Example applications. Let $X(t)$ be a time dependant random variable and $x_1, x_2 \ldots$ observations at $t_1, t_2, \ldots$ with joint probability $p(x_1, t_1; x_2, t_2; \ldots)$. The conditional probability is written as:

$$p(x_1, t_1; x_2, t_2; \ldots | y_1, \tau_1; y_2, \tau_2, \ldots). \tag{2.1}$$

Here we consider a special case of a stochastic process the Markov chain. We can write the Markov assumption in terms of the conditional probability. If the current state at t_n of the system is x_n and we write the probability of this measurement conditional on all preceding measurements $x_{n-1}, x_{n-2}, \ldots x_1$:

$$p(x_1, t_1 | x_{n-1}, t_{n-1}; x_{n-2} t_{n-2}, \ldots x_1 t_1 \ldots) = p(x_n, t_n | x_{n-1}, t_{n-1}), \tag{2.2}$$

where $t_1 \leq t_1 \leq \ldots t_n$. This means that in a Markov process an observation is only conditionally dependant on the observation immediately preceding it. Using this property it is possible

UNIQUELY DETERMINED (1.1)

CONTINOUS TIME MC

1.29

MASTER EQUATION

STATE OCCUPATION

2.1.2 HMM

SOME TEXT

2.1.3 Aggregated Markov chain

SOME TEXT

2.1.4 Least squares

Parameter estimation in such a model is the next question to be addressed. The most widely spread method for this is to maximise the likelihood function $\mathcal{L}(\theta) = f(X; \theta)$, where X is a random variable and θ is a set of parameters. Often it is more convenient to work with the log-likelihood function with $\ell(\theta) = \log \mathcal{L}(\theta)$. Since the logarithm is a monotone function the maximum of $\ell(\theta)$ is the same as the maximum of $\mathcal{L}(\theta)$. The

advantage of working with the log-likelihood is that it can be easier to work with. In some cases it is possible to obtain the maximum likelihood estimator (MLE) $\hat{\theta}$ that maximises $\mathcal{L}\theta$ and $\ell(\theta)$ in closed form. Often, especially in real world applications, this is not possible and in such cases we need to use a more numerical approach.

Now we present a simple application of these ideas and choose a common statistical model to illustrate the idea. Say there exists a model which predicts variable Y from a set of variables $X_1 \dots X_n$. For simplicity here we choose the simplest possible case, the simple linear regression model:

CHECK IF LOWER OR UPPER CASE

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.3}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and is independent of observations, β_0 and β_1 are parameters

2.1.5 Penalisation

- Why
- Possible penalisation
- Tibshirani

2.1.6 Identifiability

SOME TEXT

2.1.7 Model selection

Any statistical or even mechanistic model includes in its core some assumptions and simplification of the real world problem it is attempting to describe. The aim of a model is to enable description of a complex (sometimes not even fully understood) phenomenon in way tractable by mathematics. Often experiments are sufficient to distinguish between models and identify the one closest to the real world problem. In some cases of course this is not the case and two distinct models appear feasible. This is a problem very often found when employing statistical models and comparing observed data with predictions from models. The ubiquitous problem encountered is that one wishes to compare models to a set of noisy data. Even in cases where the problem itself is identifiable (see above) the existence of noise in observations poses a real difficulty. In such cases the question that one is really trying to answer is one of prediction.

Cross-validation

One method that uses this idea in a data driven fashion is cross-validation. The basic principle is quite simple, the data is split into two independent subsets (the training set and the validation set) and model parameters are estimated on the training set and prediction using these parameters are compared with the validation set resulting in a performance score. A practical approach is called k-fold cross-validation. Here the data set is split into k randomly chosen equally sized subsets, one subset is retained as the validation set and the remaining $k-1$ subsets are used as training data. This step is repeated for each of the k subsets and the performance score is combined giving one score for a model. In some applications as the ones discussed in later chapters of this work it is only feasible to leave out one data point at a time due to limitation in data. This procedure is then repeated for every model that is considered and the optimal model is chosen based on the best score. It is clear that such an approach has drawbacks; When dealing with large data sets computation times can quickly become unfeasible since estimations is performed for k subsets and for all possible models. An additional problem can be that due to the random splits in data the choice of this split influences results. Therefore it is advisable to try multiple splits and compare results.

AIC and BIC

Therefore there are many methods to approximate model selection results based on information obtained when estimating parameters using the whole data set only once. The advantages are obvious since it reduces computation time considerably but it is a further approximation hence one has to be careful interpreting results. In all such approaches the goodness of fit is juxtaposed to model complexity i.e. since more complicated models will perform better we want to penalise these. Such methods are historically referred to as information criteria (IC). Here we will briefly introduce two such models; One of the most widespread is the Akaike information criterion (AIC) formulated by Akaike [1974] and the other is the Bayesian information criterion (BIC) presented by Schwarz [1978]. The derivation of AIC is based on KullbackâĂŞLeibler divergence

-'information-based'??

-AIC KL distence

2.1.8 Monte Carlo integration

SOME TEXT

2.2 Biological background

-disease states and bio cond are character by distinct gene expression (DeRisis 1997; spellman 1998; Eisen and Brown 1999; Brown Botstein 1999)

2.2.1 The cell

- DNA
- RNA
- protein
- multiple timescales
- epigenetics
- state changes without genetic changes

2.2.2 Cancer Biology

- What is a tumor
- Hallmarks of cancer
- Define Oncogene
- Short history of SRC

2.2.3 Stem cells

- Define stem cells
- Reprogramming
- Difference pluripotent and embryonic stem cells

2.2.4 Cell cycle

Finally we come to a biological system underlying all the topics mentioned above the cell cycle.

- Define cell cycle
- Why important in cancer
- Radiation damage

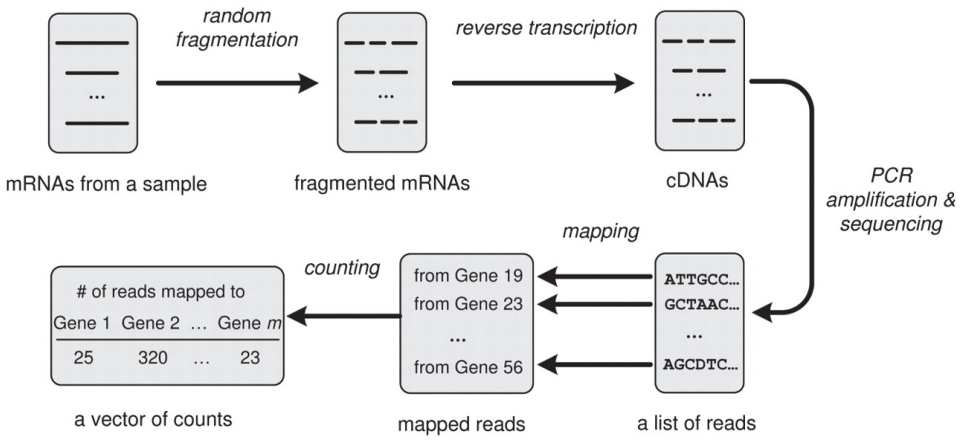


Figure 2.1: Figure from Li et al., Normalization, testing, and false discovery rate estimation for RNA-sequencing data, Biostatistics, 2012, 13(3), by permission of Oxford University Press.

2.2.5 RNA sequencing

Technique

homogenate first The word homogenate was introduced by VR Potter in1941 (J. Biol. Chem 141:775) to refer specifically to suspensions or animal tissues that had been ground in the all-glass "homogenizer"as described by Potter and CA Elvehjem in 1936 (J. Biol. Chem114:495) In the inaugural volume of Methods in Medical Reseach (Vol 1, Van R. Potter, editor-in-chief Yearbook Publishers Chicago:1948) Potter notes that the term had since been used by various investigators to refer to tissue preparations that have been "ground in mortor, with or without sand, or disintegrated in a Waring blendor, or produced by methods not described." (p. 317) Although these preparations are probably no less appropriately called homogenates, Potter says, in line with the stated goals of Methods in Medical Research, "it seems desirable to promote a nomenclature that is as meaningful as possible, and it is suggested that the method of preparation be specified." (p.317) The chief significance of the term, he continues, is that "it serves to distinguish the preparation from slices, minces and extracts....the homogenate technique accepts the fact that the cells in the tissue are no longer living and attempts to obtain surviving groups of enzymes [word "enzymes" emphasized] without loss of in vivo properties."

Ever since the mid to late nineties the importance of expression of individual genes in determining biological condition and likelihood of disease. Until recently the most prevalent method for obtaining

Chapter 3

State transitions using aggregated Markov models

3.1 Introduction

Diverse biological processes have been observed to undergo transitions under influence of a stimuli. These transitions lead to changes on a cellular level between distinct phenotypic states. These changes can be morphological, epigenetic, or on a protein level.

Phenotypic changes occurring on a single cell level are not observed at the single cell level although it is only possible to perform observations on a population level. This restriction is experimental in nature and although sometimes it is possible to make observations on a single cell level there are limitations in these experiments.

Hidden Markov models (HMMs) are widely used to describe latent processes in biological applications and have previously been used to describe cell populations [26] and model the cell cycle [27, 28, 29, 30]. It is interesting to contrast our model with a classical HMM. The key differences are twofold. First, our model involves aggregation of single-cell level Markov chains, thus it deals with states that are not only hidden, but whose connection to population-level observables necessarily involves averaging over multiple instances of the latent process. In contrast, a HMM applied to time-course data from a transition process does not provide a model at the single-cell level. Second, our model operates in continuous time and applies naturally to non- uniformly sampled data. In contrast, in a HMM the underlying Markov process operates in discrete time, such that the probability of a state transition is the same between successive time points regardless of the intervening time period. This assumption generally will not hold at all under uneven time sampling of a heterogenous population. Due to these reasons, in our view HMMs are intrinsically ill-suited to the study of transition processes of the type we consider here.

3.2 Model outline

3.2.1 The STAMM model

First we provide a detailed description of a model using 'State Transitions using Aggregated Markov Models' (STAMM), following and expanding on work done by Armond et al. [2013]. The model attempts to discern single cell parameters from observations performed on a population level.

STAMM defines a latent stochastic process on the single cell level that isn't directly observed. Using the latent stochastic processes and aggregating across cells we can obtain a cell-population level likelihood. The latent single cell process is described using a Markov chain with a discrete and finite state space but it is continuous in time. Biological states in the system are identified with the state space, indexed by $k \in \{1, ..., K\}$, is identified with biological states of the system. Transitions between states k and k' are determined by transition rates between these states denoted by $\mathbf{w} = \{w_{k,k'}\}$. Assuming that cell death and cell doubling compensate each other, i.e. the number of cells is conserved at all time t; 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 for the population $p_k(t; \mathbf{w})$ is a function of time and also the state transitions.

This model can fundamentally be applied to any type of time-course data, including transcript or protein abundance. Here unless otherwise stated we will focus the description, without loss of generality, on gene expression data. 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. When investigating transitions it is prudent to design experiments with an initial state that is reasonably homogeneous, therefore our model assumes 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 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:

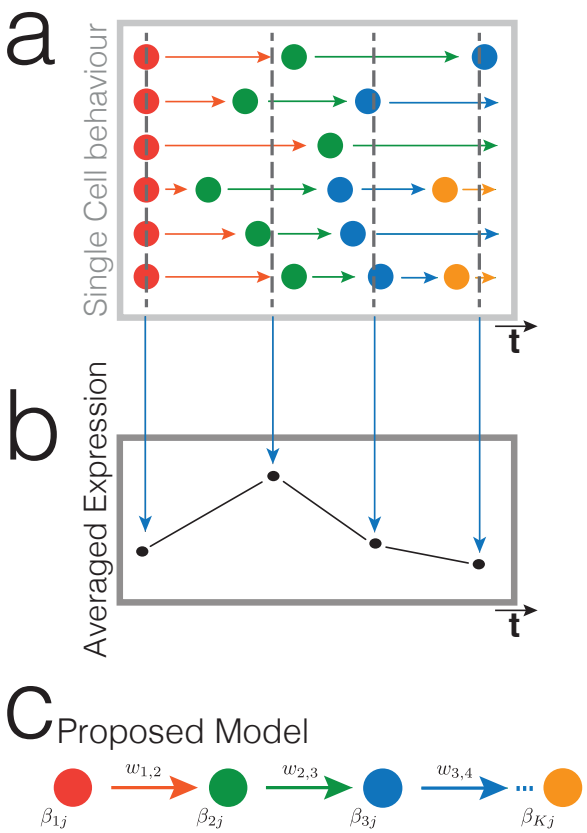


Figure 3.1: Model description. (a) In any biological system undergoing transitions between multiple states where the time of transition is stochastic, cells states are heterogeneous in the population at any given time. (b) Assays performed on homogenates of that cell population will only yield data averaged over sampled sub-populations. (c) We describe this system with 'State Transitions using Aggregated Markov Models' (STAMM) where single cell level processes are described by a latent continuous-time Markov chain which is aggregated over cells to give a likelihood (see Section 3.2.1). The Markov chain has a discrete state space which corresponds to biological states of the system (shown in different colours). Estimation of parameters in STAMM is performed using population level data. (Figure adapted from Armond et al. [2013].)

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

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.

Applied to microarray experiment which use fluorescence intensities to measure expression the transformation g used is $\log_2$. This is of course applied to data which has been normalised for fluorescent intensities Dudoit et al. [2002]. When investigating RNA-seq data we use $\mathrm{arcsinh}$ as the transformation [Hoffman et al., 2012; Johnson, 1949], defined as $\mathrm{arcsinh}(x) = \ln(x + \sqrt{(x^2 + 1)})$ (for more details see Section 4.4). RNA-seq data cannot be normalised in the same way as microarray data most importantly because it contains measurements which are exactly zero. The $\mathrm{arcsinh}$ normalisation is useful here, because unlike the log transformation it does not have a singularity at zero but has the same variance-normalisation 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 $k \in |1 \ldots K|$ 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 3.2.2). 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, $P(0) = (1, 0, 0, \ldots)$. 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:

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

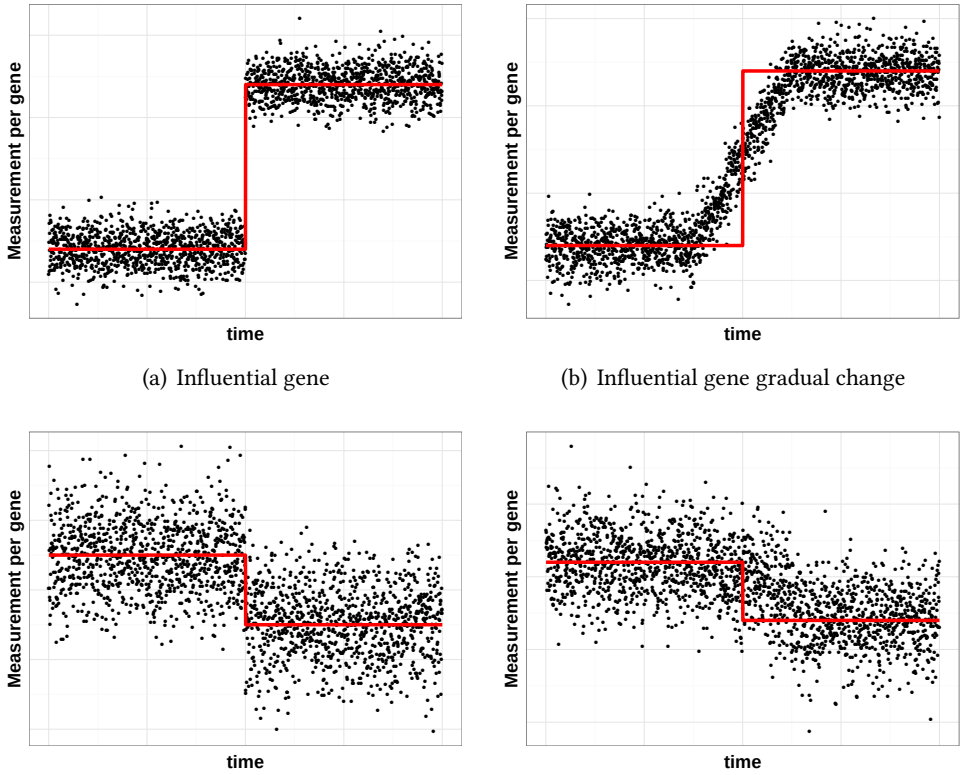
This expression is also used to evaluate the likelihood (equation (3.2)) of the model for different parameters.

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 legitimate 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 3.3.3 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 3.2.2) 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



(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 3.2.1 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.

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 3.4 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 cells in general have a much higher proliferation rate. In Section 3.4 we test how well parameters are estimated when this assumption is violated.

3.2.2 Identifiability

3.3 Estimation

3.3.1 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 optimising 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 3.3.3. 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 3.3.3).

The optimisation 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.

3.3.2 Model selection

The estimation steps described in Section 3.3.1 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 STAMM 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

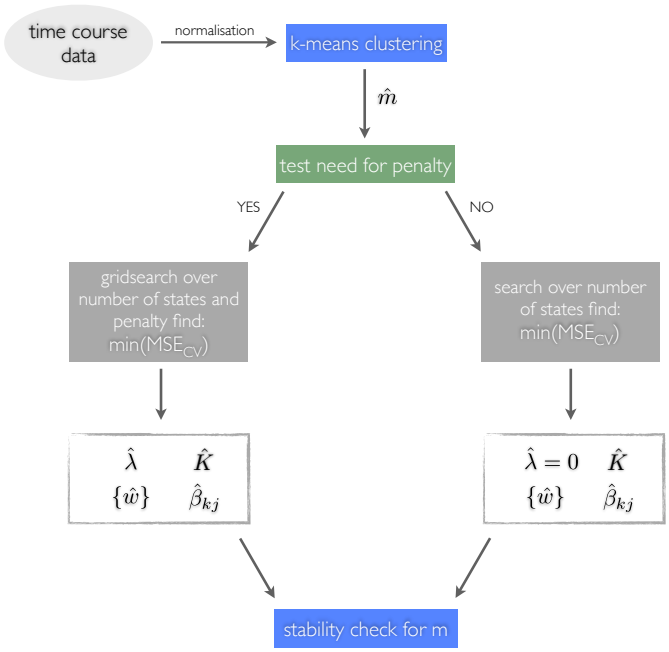


Figure 3.3: Schematic of estimation pipeline.

cross-validation mean squared error ($\mathrm{MSE}_{\mathrm{CV}}$) is simply

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

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 3.3.1, 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.

3.3.3 Estimation pipeline

We now present a computationally efficient pipeline for setting tuning parameters required for estimation. The pipeline is also summarised 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 3.5.2) 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 3.5.2 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.

Penalisation: 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 penalisation may not be required unless the number of time points is large. Consequently we introduce an additional step to test the need for penalisation 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 penalisation 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 penalisation was not required and $\lambda = 0$ in Sections 3.5 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 minimising the CV score ($\mathrm{MSE}_{\mathrm{CV}}$) for a range of

different K. If penalisation is required $\mathrm{MSE}_{\mathrm{CV}}$ is minimised 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. The issue we are trying to address is that an increase in m corresponds to an increase in information; this can lead to changes in estimated parameters. If the choice is appropriate parameters estimated increasing the number of clusters should not significantly impact parameters. To this end we compute 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.

3.4 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 homogeneity 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 behaviour in simulated data we need to insure the occupation probability for the final state is higher

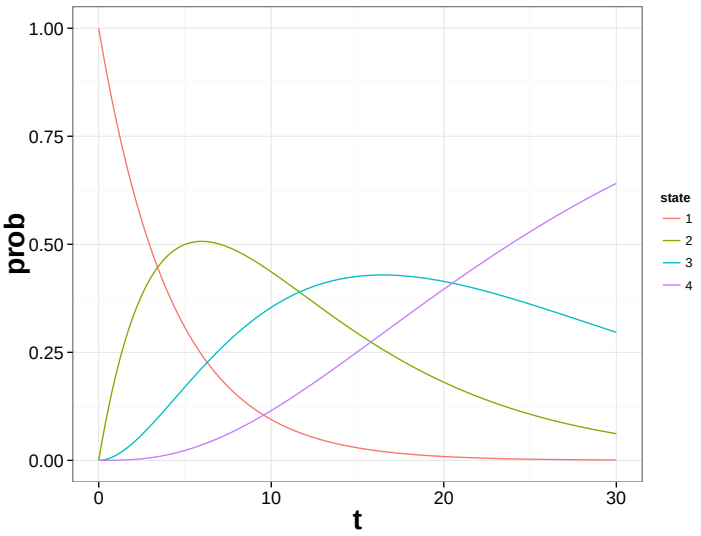


Figure 3.4: Simulation study. State occupation probabilities for a four state model.

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 3.5 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; **?**]. 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.

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

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

of cells to give an average gene expression trajectory. The occupation probability in the model outlined in Section 3.2.1 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 3.2.1 for details, for RNA-seq $\mathrm{arcsinh}$) 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 summarised in pseudocode in Algorithm 1.

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

3.5 Simulation 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 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.

3.5.1 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

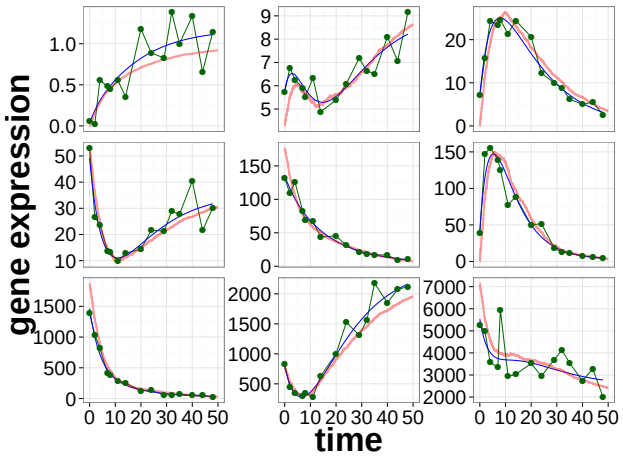
We simulated 9 genes from a 4 state model as described in Section 3.4. In this small simulation we do not use a penalisation term, i.e. $\lambda = 0$. In Figure 3.5(a) we show trajectories for one such realisation, here the thicker line represents trajectories from averaging 1000 cells for each gene. 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 realisation 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

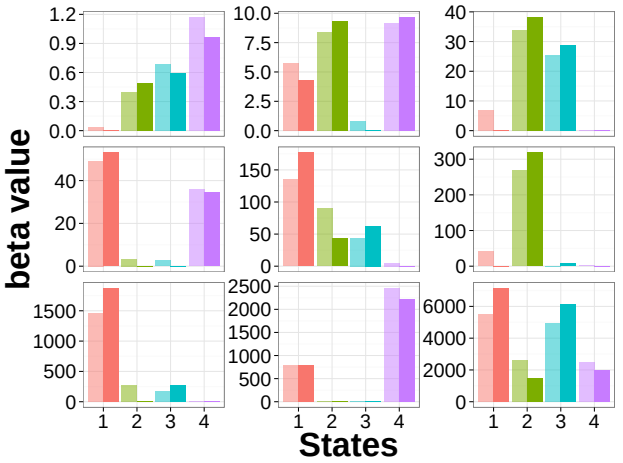
Table 3.1: Transition rates used in the simulation and the estimated values

We repeat fifty such independent simulations at four different noise levels [2], each time β_{kj} are resampled as described above (Section 3.4) 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 summarised 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

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

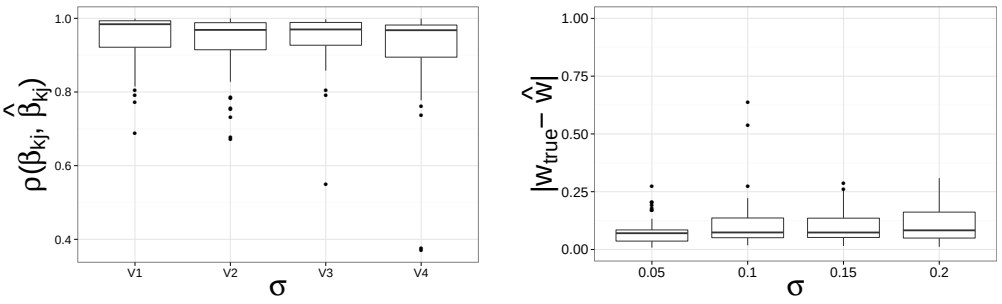


(a) Simulated Trajectories for $p = 9$



(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 colour on the left shows estimated parameter values, the solid colours 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.

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 3.3.2), for each of 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 normalised MSE score for model fit without held-out time points. In each simulation and

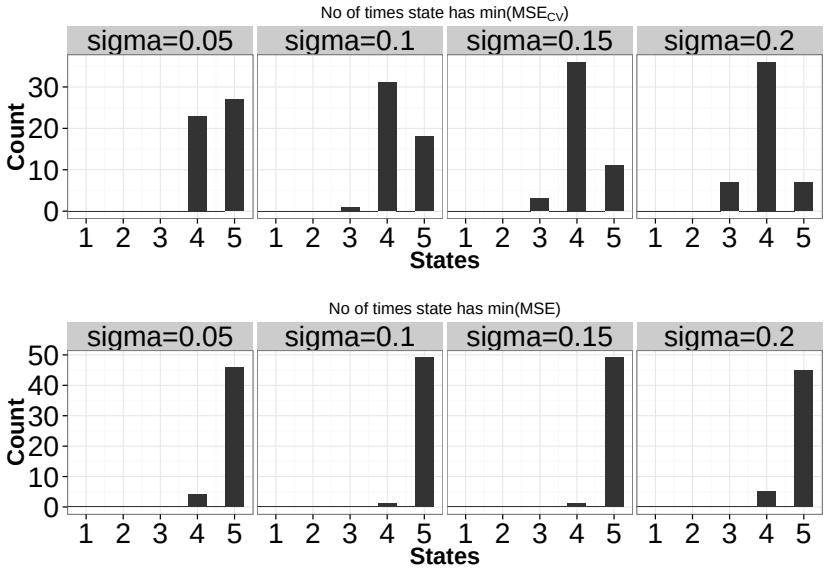


Figure 3.7: Simulation study. We perform fifty independent small scale simulation with $p = 9$.

for each noise regime we determine the model with lowest $\mathrm{MSE_{CV}}$ 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 $\mathrm{MSE_{CV}}$ 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, $\rho(\beta_{\mathrm{true}}, \hat{\beta})$.

- **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 STAMM 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 *tmvtnorm* package in R with degrees of freedom, *df*, as the varied parameter. 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 *df* parameters in *tmvtnorm*. The results are shown in Figure 3.11; 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.

3.5.2 Large scale simulation

Lastly we want to check how well the two-step estimation pipeline outlined in Section 3.3.3 works applied to simulated data. We simulate $p = 120$ genes as described above with $K = 4$ states. To mirror real data where genes are on different scales, we sample 12 scale parameters tau. To get to $p = 120$ genes we sample from all scale parameter 10 sets of β values. All other parameters are as set out in Section 3.4. We follow the procedure set out above and start by clustering simulated trajectories into m clusters. Then we estimated transition rates $\mathbf{w}$ from cluster centroids and keep them fixed for the second step. Next we use these transition rates and estimate expression signatures β_j independently for each

gene j.

3.5.3 Number of states

Applying the first step we cluster the simulated data using a k-means clustering algorithm. We use the relative decrease in the objective function $\Delta J(m)$ to determine the number of clusters and vary m in the range $[2, 30]$, see Figure 3.12. The relative decrease is smaller than 0.1 for $m = 12$ therefore we choose $\hat{m} = 13$. Note that for larger m the objective function $J(m)$ is small and we observe that $\Delta J(m)$ has large fluctuations due to slight deviations in the objective function. Then we test if for this set of data penalisation is necessary using stability of estimated gene expression signatures under deletion of time points. Figure 3.13 shows that for all deletions estimated parameters are stable, therefore we conclude that there is no need for a penalty term. Then we perform a model selection step to determine the number of states K of the latent Markov chain. We compute the $\mathrm{MSE}_{\mathrm{CV}}$ score for $K = \{2, \ldots, 5\}$ states, see Figure 3.14(a); we see a clear minimum for $K = 4$ which is also the correct number of states. In the final step of the pipeline we perform a post-estimation stability test to ensure the number of clusters chosen is not too small (see Section 3.3.3). We compute the correlation for expression parameters estimated with increasing number of clusters, see Figure 3.14(b). We carry out this test for all models $K = \{2 \ldots 5\}$ and the estimated parameters are stable therefore we conclude that the choice of $\hat{m}$ was sensible.

A question that arises from these results is to what extent they are indicative of estimation or if states are only estimated because the model assumes there are states in data. One good way of addressing this question is to permute data and compare $\mathrm{MSE}_{\mathrm{CV}}$ estimated for permuted data and the original data. Here we distinguish between two ways of permuting data: first we perform a coordinated permutation where all simulated trajectories are permuted in the same way. Then we perform a permutation for each trajectory independently. In Figure 3.14(c) we show the results and we can see that with both types of permutations the $\mathrm{MSE}_{\mathrm{CV}}$ values are significantly larger than the original data set.

Estimated parameters

One of the strengths of using simulated data is that we can compare estimated and true parameters. Figure 3.15 shows a scatter plot of estimated β parameters against their true values used in the simulation. The parameters are in general well estimated with a Pearson correlation of 0.95; despite that as the plot shows in certain cases the true value of β is exactly zero but estimates are non-zero.

The first clustering step we take is crucial and has a considerable affect on pa-

rameter estimation of transition rates. Hence it is valuable to delve a bit deeper into the first estimation step and its sensitivity to the number of clusters. Figure 3.14(b) shows that estimates expression signatures are very stable when increasing number of clusters. Figure 3.12(b) shows the three transition rates for a model with $K = 4$ as a function of m. The horizontal dashed lines in the plot show true transition rates. We observe that estimated transition rates strongly fluctuate for small m and with increasing number of clusters fluctuations decrease. are highly sensitive to changes in m,

3.6 Summary

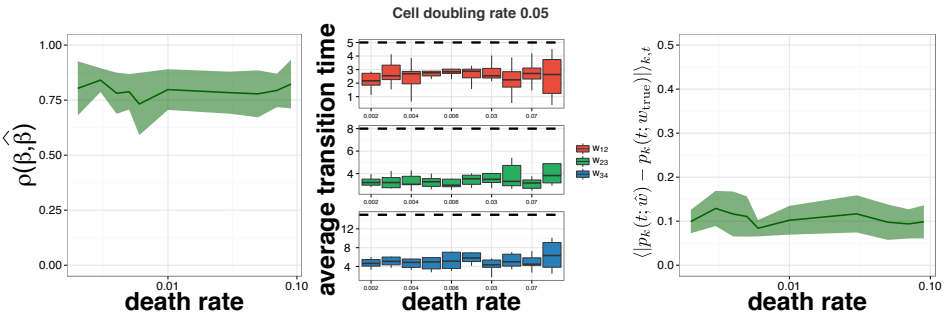


Figure 3.8: Simulation study. Testing assumption about cell death and cell doubling. For each cell a time of death, t_i^d, and a time for cell doubling, t_i^{dup}, is sampled from an exponential distribution with varying average rates. If the sampled rates for a cell are outside of the time range of 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.

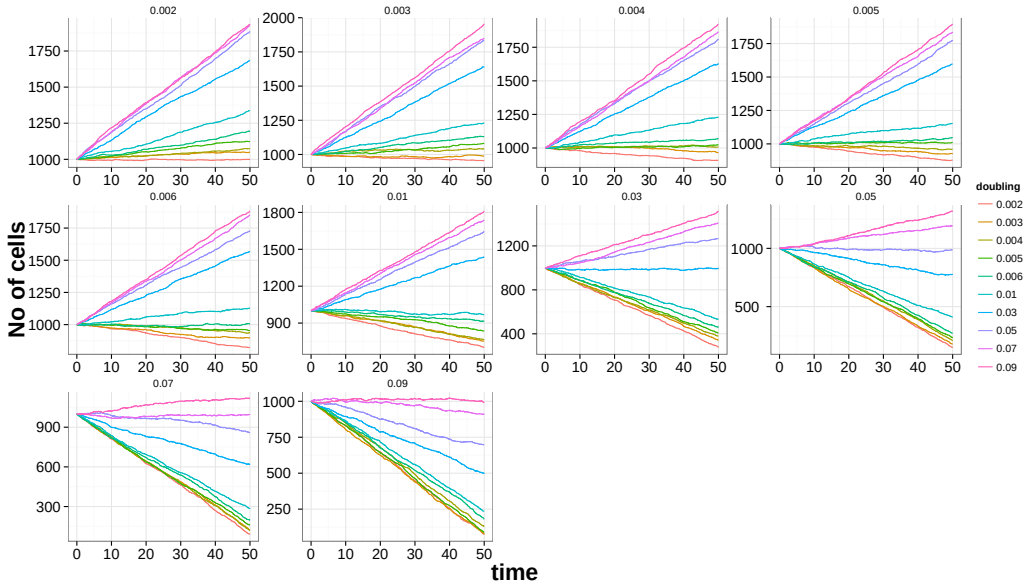


Figure 3.9: Simulation study. Testing assumptions about cell death and cell doubling. Plots show number of cells at different time during the simulation for one of the 10 simulations. Each panel represents different death rates and each colour different doubling rates.

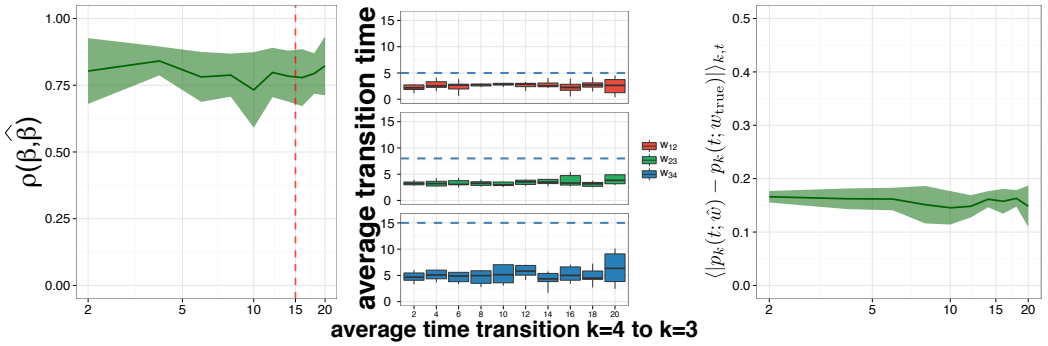


Figure 3.10: Simulation study. To test the affect of back transition on the estimation, we simulate trajectories with back transition at $k = 4$ with different transition times. In the left panel we plot the average correlation for 10 independent runs, between true and estimated β_{kj} parameters for 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.

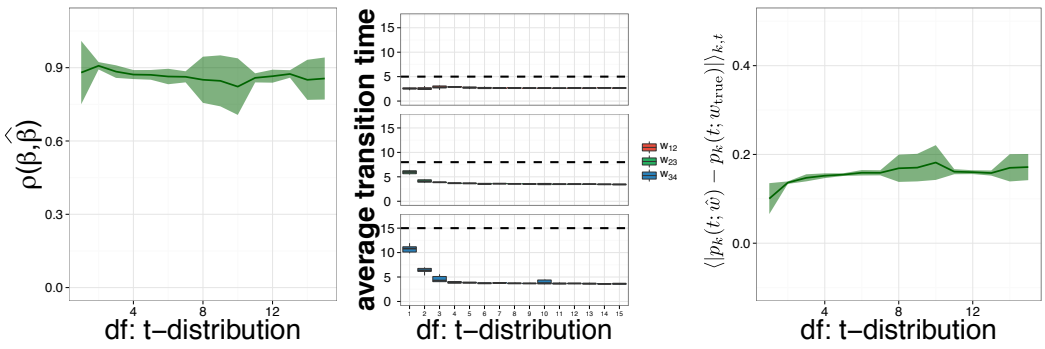
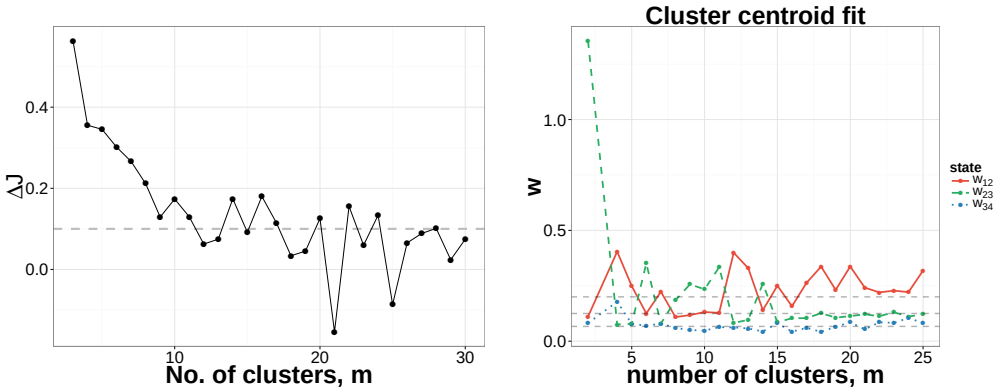


Figure 3.11: Simulation study. We simulate from a non-Markovian system, one where average transition rates are heavy tailed. Here we sample from a truncated Student t-distribution using the *tmvtnorm* package in R. It is truncated at zero since transition rates are always positive. The transition rates are sampled to have means $(1/5, 1/8, 1/15)$ and we vary the degrees of freedom *df* 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.



(a) k-means clustering (b) transition rates

Figure 3.12: Large scale simulation study. Results from clustering. (a) Initial step in the estimation pipeline is use k-means clustering for $p = 120$ genes. The plot shows relative change in the k-means objective function as a function of the number of cluster: $\Delta J(m) = 1 - J(m)/J(m-1)$. We choose the optimum number of clusters $\hat{m}$ such that $\Delta J(m-1) < 0.1$; here $\hat{m} = 13$. For larger m, $J(m)$ is small therefore relative changes have large fluctuations. (b) Estimated transition rates as a function of the number of cluster. Horizontal dashed lines show true values. After large initial fluctuations the transition rates fluctuate around the true value.

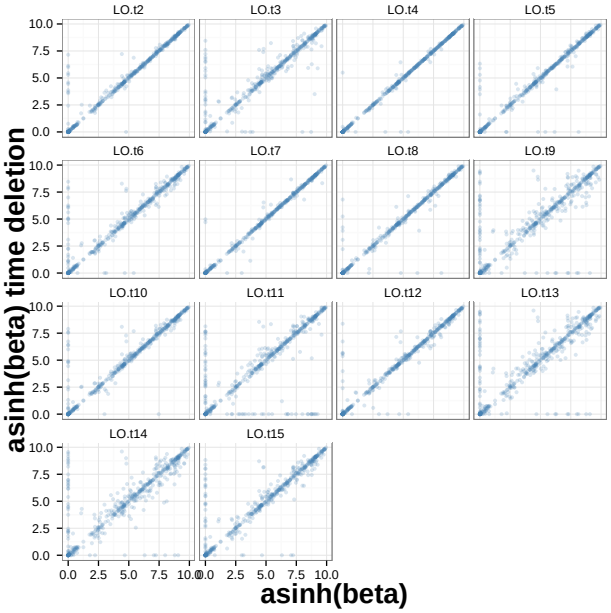
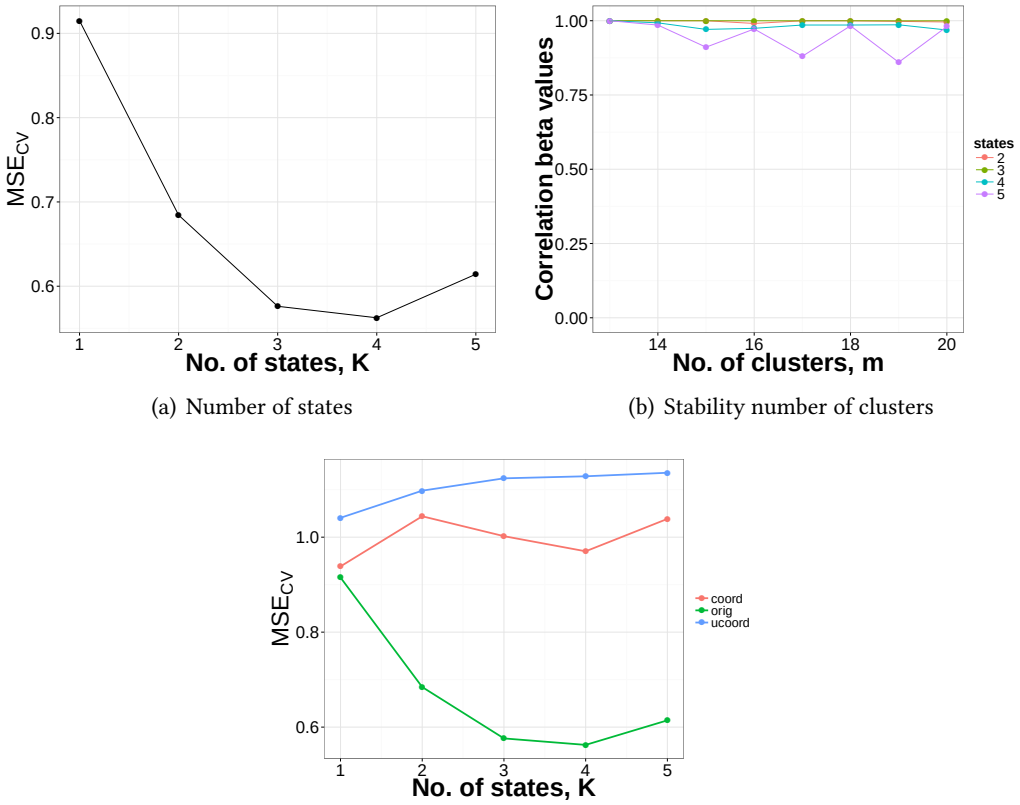


Figure 3.13: Large scale simulation. Stability of estimated expression signatures under time point deletion to determine need for ℓ_1 penalization. We conclude that estimated parameters are stable therefore there is no need for a penalty.



(a) Number of states

(b) Stability number of clusters

(c) Permuted time points

Figure 3.14: Large scale simulation.

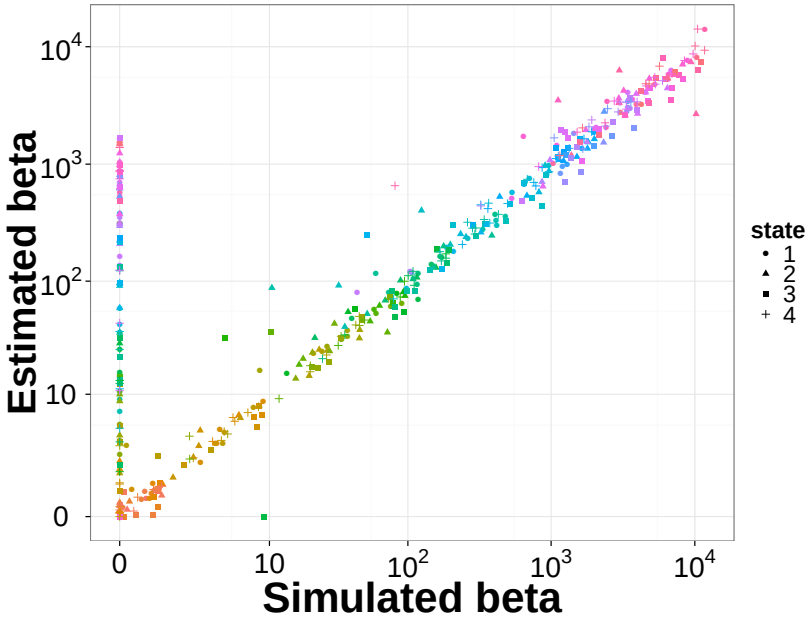


Figure 3.15: Large scale simulation. Estimated gene expression signatures scattered against their true values.

Chapter 4

Oncogenic Transformation

4.1 Introduction

There are a variety of possible applications for the model outlined in chapter 3; examples include stem cell reprogramming (Armond et al. [2013] contributions to which are discussed in Section 5) and estrogen response of Breast cancer cell lines [Casale et al., 2013]. Othere examples include a transitions Here we consider a derivative of the human epithelial MCF10A cell line where the v-Src and estrogen receptor (ER) fusion is integrated; the new cell line is called MCF10A-Er-Src [Hirsch et al., 2010] (for brevity in the discussion that follows we will refere to these as MCF10A). The Src oncogene is activated by by addition of tamoxifen resulting in a rapid transformation of this system. Morphological changes on a cellular level are observed as early as $t = 24$ (between $t = 24 - 36h$) they show the ability to form colonies in soft agar [Hirsch et al., 2010] in this transformed state. Figure 4.1 shows images taken of one realisation of the experiment using a camera attached to a microscope; they are taken at $t = \{0, 24, 48\}$ hours at two different magnification levels ($10x$ and $20x$ as indicated on the figure). The top rows shows images taken after induction of tamoxifen and the bottom row shows a null where no tamoxifen is added (see below for details).

In this chapter we discuss one application of the two-step estimation pipeline of STAMM (see Section 3.3.3). We investigate the oncogenic transformation of an MCF10A cell line using data obtained by performing RNA-Seq measurements.

4.2 Relevance

This is a very interesting system to consider mainly because it consists of only a single perturbation i.e. induction of the classical oncogene v-Src. Additionally it is an exceptionally

fast transformation (between $t = 24 - 36$ hours) and changes are morphological and can be observed under a microscope. Additional the initial state is a cell-line therefore tests or follow up experiments are easy to carry out; which can include verification of estimated parameters.

4.3 Experimental design

The experiments were performed in MCF10A-Er-Src cells, a derivative of mammary epithelial cell line MCF10A containing an integrated fusion of the v-Src oncoprotein with the ligand binding domain of ER [Hirsch et al., 2010; Iliopoulos et al., 2009]. The cells were cultured in DMEM/F12 medium supplemented as described in Debnath et al. [2003]. The Src oncogene was induced by addition of $1\mu M$ tamoxifen to a 70-80% confluent population, and induced and uninduced samples were harvested at the indicated timepoints for RNA isolation. The RNA was isolated by the Trizol method, and prepared for sequencing by the Illumina RNA TruSeq sample protocol.

4.4 Pre-processing data

As described in Section 2.2.5 RNA-Seq data as obtained in this experiment is count data. 'Normalisation' is an important step when comparing samples in different settings change this. Many methods exist to normalise sequencing experiments. We pre-process the data using the edgeR package in R McCarthy et al. [2012]; Robinson et al. [2010]. One important assumption is that most genes are not differentially expressed between samples. To determine genes that are not differentially expressed the procedure uses a robust estimate for the ratio of RNA between samples a weighted trimmed mean of M values (TMM). Two parameters are employed to filter out genes that are differentially expressed; the M-values, log-fold-changes, and the A-value, absolute expression levels. The cut-off set for both the M-value and the A-value is tuneable and the best way to set the tuning parameters is to select a range of cut-off parameters and determining when they stabilise (see Section B.1). It is important to remember edgeR was developed to analyse differential expression Robinson and Oshlack [2010], still the assumptions are also applicable to time course such as the data set for the oncogenic transformation. Note without this step it is not possible to compare data from different samples.

After the first pre-processing step we still can't use the data in our model because the likelihood is based on an additive Gaussian noise model (eqn. (3.2)). For RNA-seq data once it has been 'normalised' the next pre-processing step is to transform the data such that distorting effects at high expression values are reduced. We use a nonlinear transform

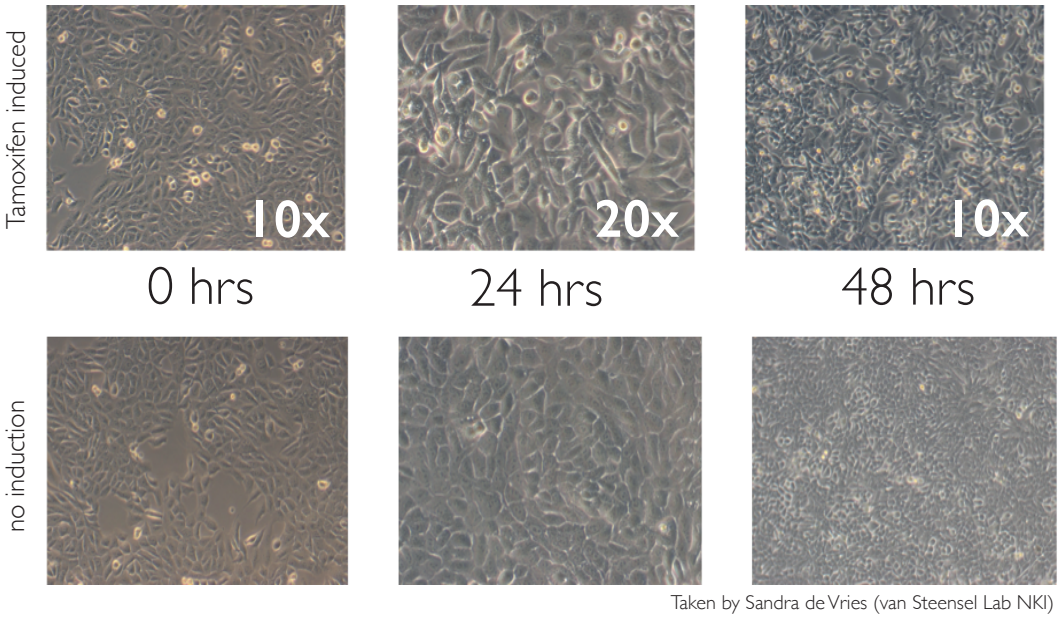


Figure 4.1: During the *in vitro* oncogenic transformation of an MCF10A-Er-Src cell line, morphological transformations can be observed between $24 - 36$ hours [Hirsch et al., 2010]. Here we show images taken of the experiment at three different time points. The initial measurement at $t = 0$ hours and two subsequent measurements at $t = 24$ and $t = 48$ hours. The magnification each image is taken at is indicated on the images. The top row shows images of the experiment where Src is induced by addition of Tamoxifen at $t = 0$; the bottom row shows images taken at the same time points without addition of Tamoxifen. It is possible to see morphological changes especially at $t = 48$ where after addition of Tamoxifen cells are elongated.

proposed by Hoffman et al. [2012] the $\operatorname{arcsinh} x = \ln(x + \sqrt{(}x^2 + 1))$. The advantage of using this transformation compared to the more regularly used $\ln x$ is that it has the same effect at higher values but a much smaller effect at lower values. Now after applying these two pre-processing steps we can use this data in our model.

4.5 Results

The data obtained uses RNA-seq to examine changes in gene expression during this transformation time-course with $T = 13$ time points[1]. According to the assumption in our model all cells are in the initial state at $t = 0$ are in an initial state. Here, by experimental design, the initial state consists of the derivative MCF-10A-Er-Src cell line. More specifically the time point $t = 0$ in the here corresponds to the initial treatment with Tamoxifen an anti-oestrogen drug that binds to ER activating v-Src. Of course this is only the case excluding any unidentified epigenetic heterogeneity in the initial cell culture. Therefore it is reasonable to assume that the cell population is approximately homogenous and comprised mainly of untransformed MCF-10A cells.

To focus on genes that change over time during transformation we filtered genes j with respect to standard deviation σ_j, retaining genes with $\sigma_j > 20$ (on the linear scale, over time). This gave a set of $p = 2809$ gene expression trajectories to which STAMM was applied. Estimation was carried out following the pipeline described in Methods, including an initial test of the need for penalization (penalization was not needed; see Fig 7(b) SI). We used two-stage estimation (see Methods), clustering the data to obtain representative centroids from which to estimate transition rates (Fig 8 SI), and using $\operatorname{arcsinh}$ as the transformation function g. The proposed cross-validation-based model selection score showed a minimum at $K = 4$ states (Fig. **??**(a)). This suggests that oncogenic transformation of the MCF-10A cells occurs via four transcriptionally-distinct states. A representative set of trajectories and corresponding fitted model output are shown in Fig. **??**(b) with corresponding state-specific gene expression parameters β_j shown in Fig. **??**(c). Finally, we carried out a post-estimation diagnostic of sensitivity to the tuning parameter m (number of initial clusters); Fig. **??**(d) shows the correlation between reported β's and those obtained with increasing m; we find that the estimates are not sensitive to choice of m.

The foregoing example illustrates the application of the proposed methods to genome-wide RNA-seq data, including empirical diagnostics. However, investigation and validation of these potentially novel states is beyond the scope of the present paper (see also Discussion below).

[1] $t = \{0, 0.5, 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, 48h\}$, at $t = 0$ and $t = 20$ we have a repeated measurements

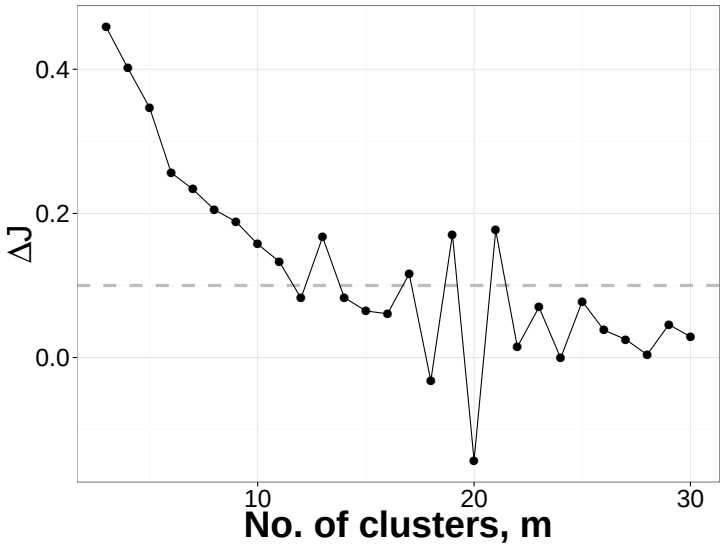


Figure 4.2: K-means clustering of *in vitro* data for the transformation of an MCF10A cell line. Initial k-means clustering is performed to identify representative trajectories. As described in Section 3.3.3 we choose the optimal number of clusters $\hat{m}$ by considering relative changes in the objective function $\Delta J(m) = (J(m-1) - J(m))/J(m-1)$; we set m where $\Delta J(m-1) < 0.1$. The plot shows ΔJ as a function of m and the horizontal dashed line represents our threshold at 0.1. For this example we can see that the $\hat{m} = 13$. Note that fluctuation in ΔJ at higher m are due to small values of $J(m)$.

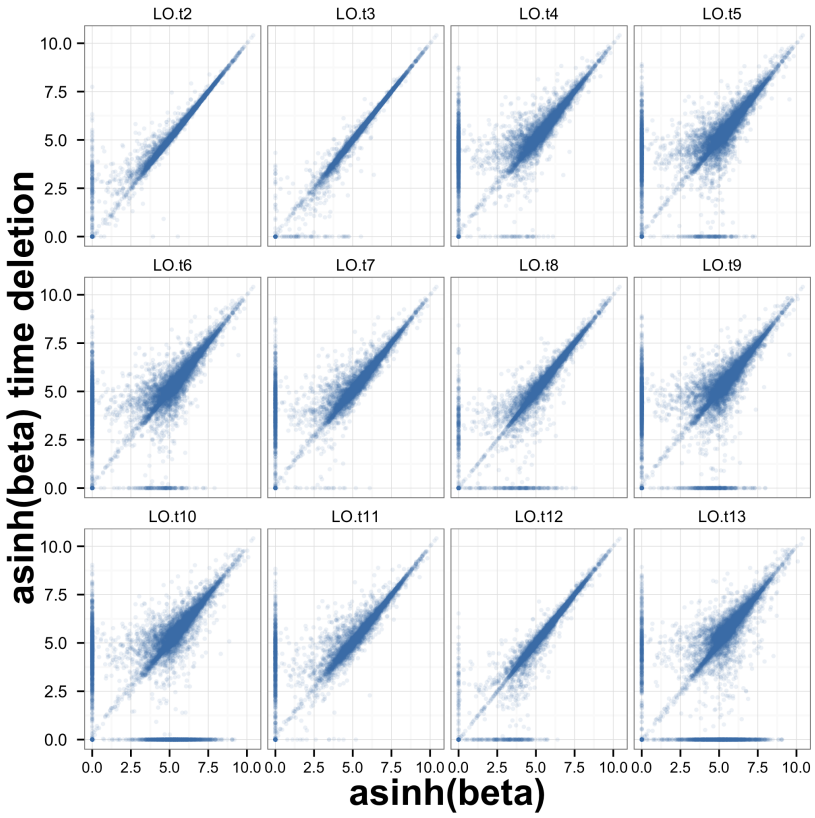
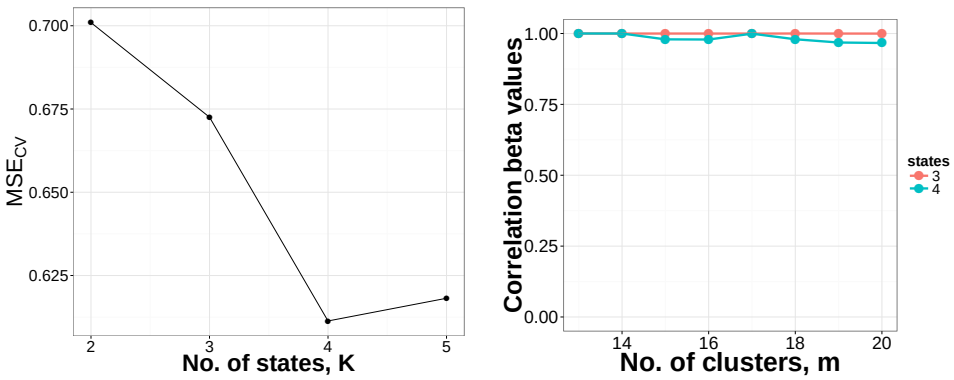


Figure 4.3: The RNA-seq measurements are filtered with respect to standard deviation (we filter out genes with $\sigma_j > 20$ before transformation) because we are interested in genes that change significantly in time. Test stability of estimates to determine need for penalisation.



(a) Number of state K

(b) Stability of number of cluster

Figure 4.4: Determining number of states K and stability test

Chapter 5

Stem cells

5.1 Introduction

Another application of the STAMM model is to reprogramming of somatic cell to a pluripotent state as investigated in Armond et al. [2013]. In the experimental setup a mouse embryonic fibroblasts (MEFs) is used and transformed to a state of pluripotency [Takahashi and Yamanaka, 2006; Jaenisch and Young, 2008] [15]. More specifically we apply the model to the genome-wide microarray gene expression time-course data obtained by Samavarchi-Tehrani et al. [2010]. It is a system that has been extensively studied in recent years and it is suggested that the reprogramming process is inherently stochastic [Hanna et al., 2009]. Progress has also been made on a single-cell investigations of the biological system [Buganim et al., 2012] ([20, 21, 22]). Question still remain on genome-wide level including the number of intermediate state between the initial MEF state and the final pluripotent state.

In this Chapter we start by briefly outlining results obtained Armond et al. [2013] when applying STAMM to a microarray data set in Section 5.2. Then (in Section 5.3) we discuss the main contribution in detail which is a comparative study of parameters obtained from STAMM and single cell experiments performed by Buganim et al. [2012]. The single cell data was obtained by a new kind of experimental technique called a Fluidigm assay. This also illustrates an example of a possible next step in investigating a biological system once parameters from STAMM have been obtained.

5.2 Results from model application

5.2.1 Differences in estimation

The initial step before we can make a comparison to single cell results is to apply STAMM to the microarray time-course; obtaining single cell level parameters and the number of states K. In Armond et al. [2013] there are differences in the estimation pipeline compared to the one outlined in Section 3.3.3.

The main idea of a two-step estimation process is shared. The first difference is that the optimal number of clusters is chosen when increasing the number of clusters does not significantly improve the k-means objective function. The penalty for estimation used to regularise estimation in eqn. (3.7) is set to a small positive number (in this application set to $\lambda = 0.1$). Estimation of transition rates $\{\mathbf{w}\}$ is performed on genes closest to cluster centroids instead of the cluster centroids themselves; then transition rates are fixed and estimation of expression signatures β_{kj} is performed in the same way. Finally estimation of the number of states $\hat{K}$ is performed in two ways. The heuristic approach is to looks at two quantities the model fit, i.e. the residual sum-of-squares (RSS), and the distinctness for individual state signatures quantified by the condition number $C = \max(s_i)/min(s_i)$; where s_i are the singular values of a matrix made up of the expression signatures. The other approach for finding an optimum number of states is employed for genes closest to centroids using a Bayesian model selection approach. Let $\mathbf{y} = \{y_j\}$ denote observed data and M_k the model with k states. The posterior probability is $P(M_n|\mathbf{y}) \propto p(\mathbf{y}|M_k)$ with a flat prior distribution over models. The marginal likelihood $p(\mathbf{y}|M_k)$ accounts for the fit-to-data and model complexity. Writing all model parameters as $\theta = (\beta_{kj}, \{\mathbf{w}\}, \{\sigma_j\})$ the marginal likelihood is:

$$p(\mathbf{y}|M_k) = \int p((y)|\theta, M_k)\, p(\theta|M_k) d\theta. \tag{5.1}$$

We compute the marginal likelihood of the model eqn. (5.1) using annealed importance sampling (AIS) [Neal, 2001], a MCMC method, to compute the marginal likelihood. Hyperparameters for this model are set by hand to reasonable values, see Supplement of Armond et al. [2013] for details. The normalised score is the required posterior probability over the number of states.

5.2.2 Estimation results

The primary data used in Armond et al. [2013] is obtained by reprogramming of a secondary mouse embryonic fibroblast (MEF) where Oct4, Sox2, Klf4, and cMyc are inducibly expressed in the system for 30 days [Samavarchi-Tehrani et al., 2010]. Microarray mea-

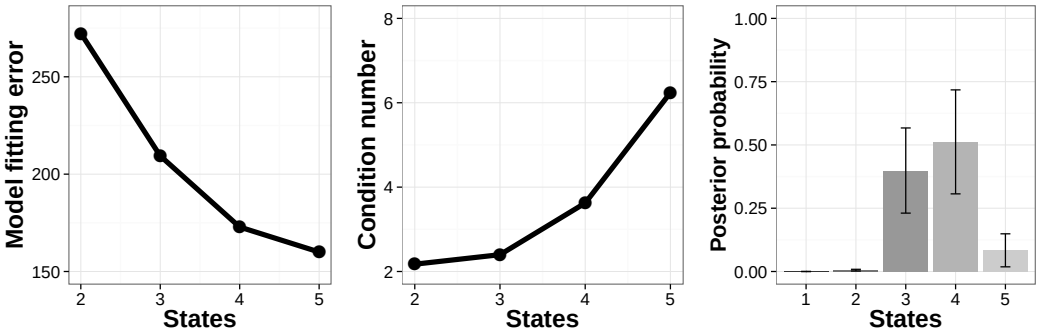
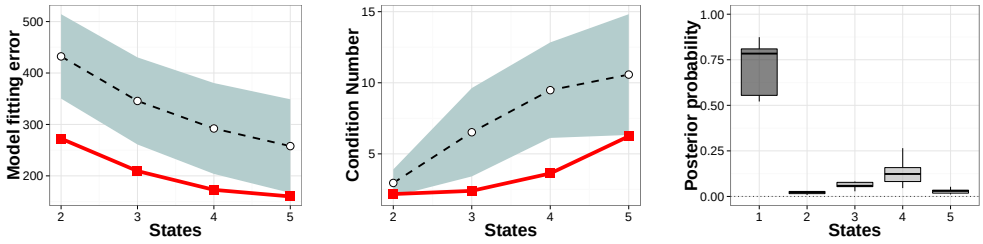


Figure 5.1: Application of STAMM to a microarray time-course (a) Plot of the model fit residual sum of squares (RSS). (b) Plot of the condition number for estimated expression signatures quantifying linear dependence between states. A larger number corresponds to more dependence. (c) Posterior probabilities obtained from Bayesian model selection (see Section 5.2.1 for details).

surements were made at $t = \{0, 2, 5, 8, 11, 16, 21, 30\}$ days after induction of expression factors. The microarray data is standardised per gene such that $y_j(t) = (z_j(t) - \mu_j)/\sigma_j$, where $z_j(t)$ is original $\log_2$ transformed data, μ_j is the mean and σ_j is the standard deviation of the time course data for gene j. A total of 4383 genes are retained out of the whole gene list. Genes are removed if they are expressed at very low levels and therefore would be dominated by noise.

The number of clusters chosen for this data set of 8 time points is $\hat{m} = 7$. As mentioned above the penalty used in this application is $\lambda = 0.1$. With these parameters set the transition rates are estimated from cluster representative genes. Once transition rates are estimated the expression signatures for the remaining genes are estimated. The analysis is carried out for $K = \{2 \ldots 5\}$ and results for model selection are summarised in Figure 5.1. Unsurprisingly the RSS keeps decreasing for increasing K (Figure 5.2(a)) since numbers of model parameters increase. To determine number of states heuristically we compare these results with the condition number (Figure 5.2(b)). We find that the difference in condition number from $K = 4$ to $K = 5$ is larger than the preceding changes. This suggests that decrease in RSS from 4 to 5 states is mostly due to overfitting and the additional state is not distinct. The posterior probabilities form Bayesian model selection for 7 genes closest to the centroid (see above for details), are shown in Figure 5.2(c). Combined these results indicate that a $\hat{K} = 4$ since it strikes a good balance between model fit and distinct expression signatures for states as well as having the highest posterior probability.



(a) RSS (b) Condition number (c) Bayesian model selection

Figure 5.2: Random permutation of time points. a, model fitting error, b, reciprocal condition number and c, Bayesian posterior probability as a func- tion of number of states. We generated a set of randomized data by random reordering of time indices and the model was re-fit for each of the permuted data. Curves (a,b) and box plots (c) shown are over ten samples; in (a,b) dotted lines indicate means and the shaded area standard deviations, whilst the cor- responding result for the correct time ordering is shown in red. Both model fit and distinctness of state signatures are systematically worse under permutation of time indices. Bayesian model selection applied to the randomly permuted data show no evidence of intermediate states (c), in contrast with the original data (Fig. 2c, Main Text). [Dataset from Samavarchi-Tehrani et al. [1]; see SI Text for details.]

5.3 Testing against single cell data

5.3.1 Single cell experiment

Results in Section 5.2.2 are obtained analysing homogenate time-course data; but the transformation process is on a single cell level therefore it is of interest to study behaviour on a single cell level. Comparing results from STAMM to single cell observations also indicates how well the underlying single cell process is modelled. For this purpose we investigate the mRNA single-cell expression performed by Buganim et al. [2012]. They also investigate a secondary MEF system reprogrammed by transduction of Oct4, Sox2, Klf4, and cMyc; obtaining data with the Fluidigm assay, resulting in 96 single-cell measurements with gene expression from 48 genes. Observations are made in populations, starting with MEFs, over cells at $2-6$ days during reprogramming, to the final reprogrammed cells.

5.3.2 Comparing results

The single-cell measurements [Buganim et al., 2012] allow for analysis that has not been possible for population average data. Although important questions about the transformation process such as the number of states and transition rates still remain difficult to track down; this is due to the fact that each time a single-cell measurement is made the

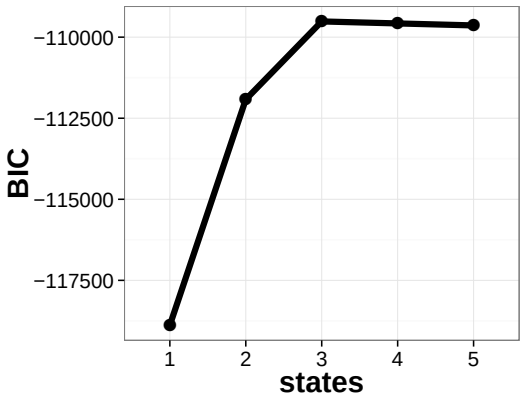


Figure 5.3: Single cell expression levels in different experimental settings from Buganim et al. [2012] are clustering using a standard clustering procedure in R called mclust. We use the Baysian Information Criterion (BIC) to score different cluster sizes. We find the optimal number of clusters to be 3 since the BIC score decreases for larger cluster sizes.

cell has to be destroyed and additional work is necessary to determine distinctive marker for known states for purification.

Given available data we can address interesting question on expression patterns; since we assume that cells belonging to the same states would have a comparable expression patterns across observed genes. This is especially the case since measured genes are deemed important for reprogramming. To this end we cluster the data for all cells in a 48 dimensional gene expression space. To perform the clustering we use a widely available clustering tool in R mclust; it employs a variety of multi-variate clustering methods and scores them using the Bayesian Information Criterion (BIC). The best performing method is shown in Figure 5.3. We find that optimal number of clusters is 3 since the BIC score starts decreasing for larger cluster sizes.

Next we try to determine if state specific expression signatures estimated from microarray data can be compared with this new single-cell data. Disregarding conditions for each cells measurement we assign each of the single-cell measurements to the each of the states in the $K = 4$ model. We compute the euclidean distance between gene expression on a single-cell level and estimated gene expression signatures and assign each cell to a specific state. The heatmap in Figure 5.4 shows fractions of cells that are assigned to each state. All MEF conditions have a peak at $K = 1$. Measurements obtained between $t = 2$ and $t = 44$ days are spread over state $K = 1$ and $K = 3$ with very few cells also in the second state. No cells from each of these measurement are close the final state. Measurements for dox-independent and iPS cells occupy only the final two states. These

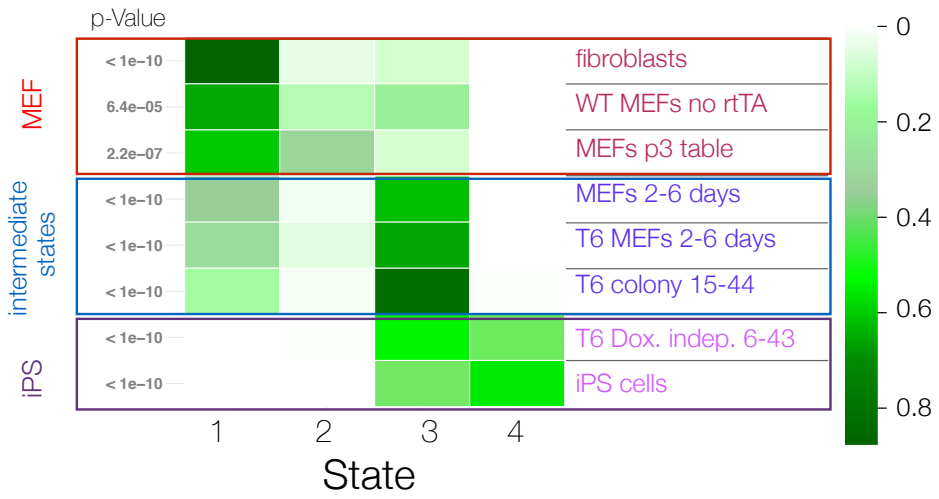


Figure 5.4: Estimated gene expression signatures using STAMM are compared with single-cell measurements performed by Buganim et al. [2012]. Each single-cells is assigned to a state by finding minimum euclidean distance. The heatmap summarises the fraction of cells from each experimental condition assigned to specific states. Different conditions show a clear preference for specific states. The prediction of our model are in line with this observation where initial MEF states undergo a transformation via intermediate states to a final reprogrammed state. As an example all MEF populations (top three entries) have a significantly higher fraction of cells in $K = 1$. Cells measured between $t = 2$ and $t = 44$ days have cells in the spread across the first and third state with very few cells occupying the second state. None of these cells are close to the final state. The two measurements which are reprogrammed cells (iPS cells and Dox. indep.) show similarities to $K = 3$ and $K = 4$, but none of these cells are close to the first two states.

results clearly show a transformation starting with MEF state and undergoing changes across intermediate states before reaching the final reprogrammed state. Of course this is a very small study and a study made on a slightly different system under different conditions therefore even the approximate similarities we find to our estimated parameters are promising.

Chapter 6

Cell cycle

6.1 Introduction

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; Gerlinger et al., 2012]. In this Chapter we outline a model that answers the question: What effect does the initial cell population have on cell fate?

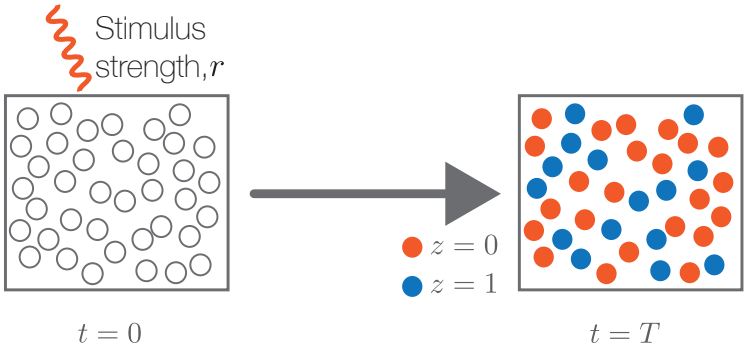


Figure 6.1: Schematic of heterogeneous cell population transforming under stimulus.

An example of such a biological system 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. A schematic of such a system is shown in Figure 6.1. Individual genes are influential in determining cell fate if their expression level is significantly different between cell populations at $t = 0$.

6.2 Formal system description

6.2.1 Concepts

Suppose at time t cell i (out of N cells) occupies state X_{it}; where state here broadly refers to any aspect of the cell's physical configuration. This can include protein profile, transcription, or its chromatin state. Denote ultimate cell fate at $t = T$ for cell i by Z_i. Cell fate Z_i is determined experimentally by enumerating cells in distinguishable states at $t = T$. In the simplest case cells can have two distinct final states; we label the two states $Z_i = 0$ and $Z_i = 1$. We expect the process that determines cell fate to have a stochastic component such that two physically identical cell at $t = 0$ can end up in distinct final states. Hence we assume the probability that cell i is in state $Z_i = 1$ at $t = T$ depends on two things:

- The physical state of cell i at $t = 0$, X_{i0}.
- The dose of the stimulus r.

Since we assume that the fraction of cells that reach the arrested state changes with the strength of the stimulus. The fraction of cells that reach cell fate $Z_i = 1$ is dose dependant and denoted by $\pi(t)$.

6.2.2 Model

Appendices

Appendix A

Additional results MAST

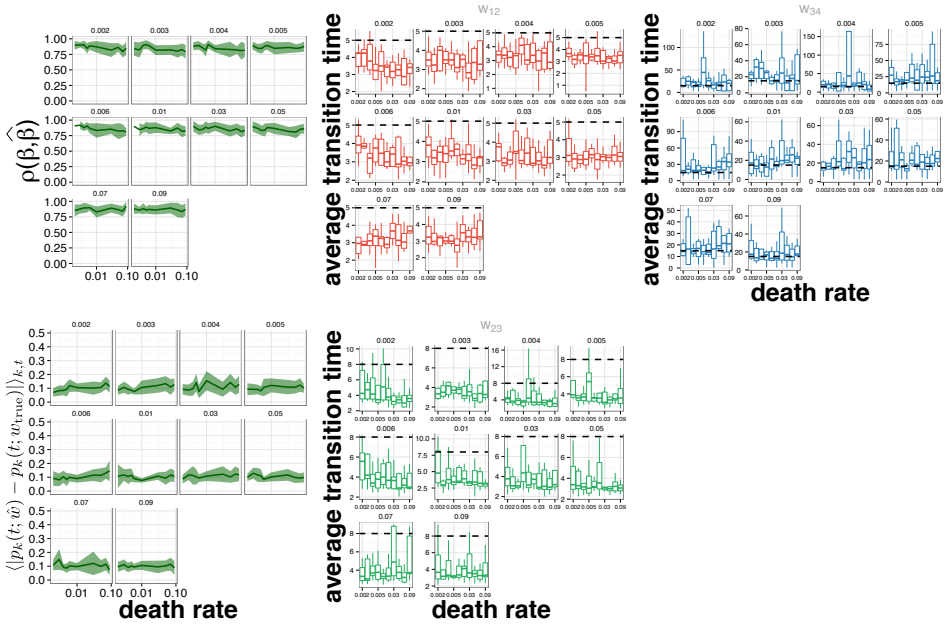


Figure A.1: Simulation study. Extend figure 3.8 to include a wider range of doubling rate. The top left panel shows the correlations between true and estimated β_{kj} as a function of death rates. The bottom left panel shows the mean differences between estimated and true probabilities averaged over 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.

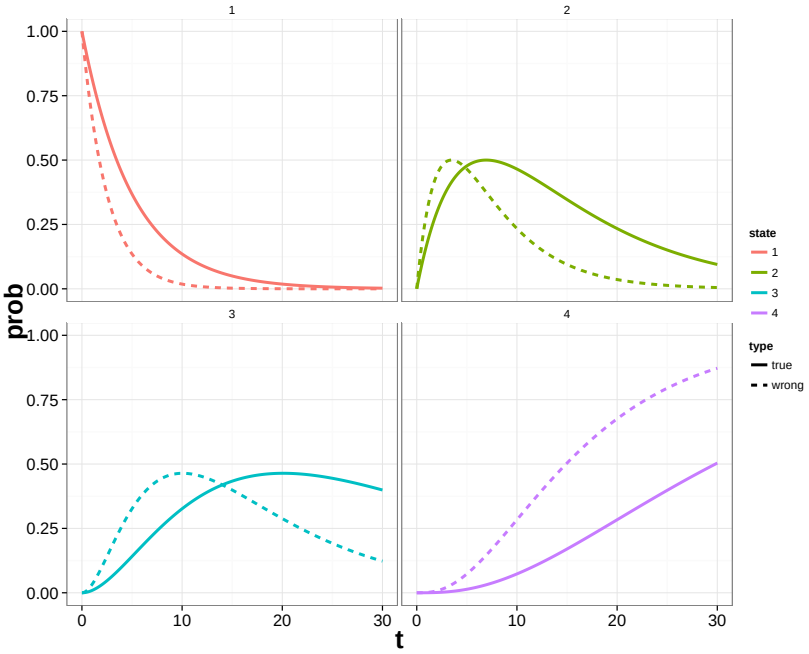


Figure A.2: Simulation study. To highlight the effect of badly estimated

Appendix B

MCF10A results

B.1 Sequencing depth in RNA-seq

Figure B.1: To determine sequencing depth in the measured data we run

Bibliography

H Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

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.

Yosef Buganim, Dina A Faddah, Albert W Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell*, 150(6):1209–1222, September 2012.

FP Casale, G Giurato, G Nassa, JW Armond, CJ Oates, D Corá, A Gamba, S Mukherjee, A Weisz, and M Nicodemi. Single-Cell States in the Estrogen Response of Breast Cancer Cell Lines. *submitted*, 2013.

Jayanta Debnath, Senthil K Muthuswamy, and Joan S Brugge. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods*, 30(3):256–268, July 2003.

S Dudoit, Y H Yang, M J Callow, and T P Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 2002.

Albert Einstein. Über die von der molekularkinetischen theorie der wärme gerforderte Bewegung von in ruhenden . *Annalen der Physik*, 322 (8):549–560, 1905.

M Gerlinger, P A Futreal, L Pusztai, and C Swanton. Intratumor heterogeneity: seeing the wood for the trees. *Sci Transl ...*, 2012.

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.

Heather A Hirsch, Dimitrios Iliopoulos, Amita Joshi, Yong Zhang, Savina A Jaeger, Martha Bulyk, Philip N Tsichlis, X Shirley Liu, and Kevin Struhl. A Transcriptional Signature and Common Gene Networks Link Cancer with Lipid Metabolism and Diverse Human Diseases. *Cancer Cell*, 17(4):348–361, April 2010.

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.

Dimitrios Iliopoulos, Heather A Hirsch, and Kevin Struhl. An Epigenetic Switch Involving NF-kB, Lin28, Let-7 MicroRNA, and IL6 Links Inflammation to Cell Transformation. *Cell*, 139(4):693–706, November 2009.

Rudolf Jaenisch and Richard Young. Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. *Cell*, 132(4):567–582, February 2008.

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

D J McCarthy, Y Chen, and G K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

Payman Samavarchi-Tehrani, Azadeh Golipour, Laurent David, Hoon-Ki Sung, Tobias A Beyer, Alessandro Datti, Knut Woltjen, Andras Nagy, and Jeffrey L Wrana. Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. *Stem Cell*, 7(1):64–77, July 2010.

Gideon Schwarz. Estimating the Dimension of a Model. *Ann. Statist.*, 6(2):461–464, 1978.

Kazutoshi Takahashi and Shinya Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676, August 2006.

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