MATHEMATICAL AND STATISTICAL METHODS FOR SINGLE CELL DATA

by

WILLIAM THOMSON

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Mathematics
College of Engineering and Physical Sciences
The University of Birmingham
February 2020

Abstract

The availability of single-cell data has increased rapidly in recent years and presents interesting new challenges in the analysis of such data and the modelling of the processes that generate it. In this thesis, we attempt to deal with some of those challenges by developing and exploring mathematical and statistical models for the evolution of population distributions over time, and methods for using aggregated single-cell data from individual patients in predictive diagnostic models of disease. In the first part of the thesis, we explore structured population models – a class of partial differential equations for describing the evolution of individual-level cell properties in a population over time. We begin by analysing an age-structured model of cell growth in which rates of proliferation and cell death are controlled by an external resource. We follow this with a method for extracting properties of a more general class of structured population models directly from single-cell data. In the final part of the thesis, we develop a flexible Bayesian statistical framework for building predictive models from possibly high-dimensional data collected from patients using single-cell technologies and find that the performance is promising compared to a number of existing methods.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Sara Jabbari & David Smith, for their unwavering support and encouragement. I will never understand their apparent faith in me, but I am hugely grateful for it.

I would like to thank the other people I have worked with along the way in my industrial and academic collaborations – the opportunity to experience different problems in different settings has been very valuable.

I am grateful for the support of the EPSRC in funding this research.

Finally, I would like to thank my parents and my brother for being a constant source of comfort, support, stability and fun.

LIST OF PUBLICATIONS

Thomson, W., Jabbari, S., Taylor, A. E., Arlt, W., & Smith, D. J. (2019). *Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior*. Journal of the Royal Society Interface, 16(150).

Kerr, R., Thomson, W. M., & Smith, D. J. (2019). *Mathematical modelling of the vitamin C clock reaction*. Royal Society Open Science, 6(4).

Eley, Y. L., Thomson, W., Greene, S. E., Mandel, I., Edgar, K., Bendle, J. A. & Dunkley Jones, T. (2019). *OPTiMAL: A new machine learning approach for GDGT-based palaeothermometry*. Climate of the Past. (*Under review*).

CONTENTS

LIST OF ABBREVIATIONS

AlloHSCT Allogeneic Haematopoietic Stem Cell Transplantation

CDF Cumulative Distribution Function

DMD Dynamic Mode Decomposition

DP Dirichlet Process

DPGMM Dirichlet Process Gaussian Mixture Model

GAM Generalised Additive Model

GLM Generalised Linear Model

GvHD Graft-versus-Host Disease

HMC Hamiltonian Monte Carlo

HSCT Haematopoietic Stem Cell Transplantation

ILR Isometric Log Ratio

LASSO Least Absolute Shrinkage and Selection Operator

MCMC Markov Chain Monte Carlo

MDC Multiscale Data Condensation

NMF Non-negative Matrix Factorisation

ODE Ordinary Differential Equation

PBA Population Balance Analysis

PBE Population Balance Equation

PCA Principal Component Analysis

PDE Partial Differential Equation

PDF Probability Density Function

PMF Probability Mass Function

POD Proper Orthogonal Decomposition

ROC Receiver Operating Characteristic

SDE Stochastic Differential Equation

SVD Singular Value Decomposition

VDPGMM Variational Dirichlet Process Gaussian Mixture Model

CHAPTER 1

INTRODUCTION

1.1 Single-cell data

Access to data containing the expression levels of genes & proteins at the level of individual cells has rapidly increased in recent years thanks to technological advances and lessening costs; examples of these technologies are single-cell RNA sequencing, single-cell DNA sequencing and flow cytometry. These methods allow the simultaneous processing of thousands to hundreds of thousands of cells. In the case of RNA sequencing, it is possible to extract information on the expression levels of thousands of different RNAs for each individual cell in the sample. The primary advantage of these methods is that they provide the ability to study not only bulk properties of populations of cells but also cellular heterogeneity, specialised functions, and phenotypic evolution.

Such approaches have seen widespread advancement of our knowledge of various diseases and the cellular characteristics that underpin them. An early example is in the characterisation of gene expression profiles in colorectal cancer [2] – we use the data from this study in chapter 4. A summary of available single-cell sequencing techniques

and other notable examples of the successes of single-cell methods are covered in comprehensive reviews [79, 82].

Much of this thesis focuses on the human immune system. In the following subsection we provide an overview of the functions and cell types involved in protecting humans from infection and disease. We structure the discussion in such a way that we hope the explanation serves both as an illustration of the way in which increasing granularity of data at the single cell level leads to enhanced understanding of the processes underlying biological systems, and to provide context to the model in chapter 2.

1.2 Introduction to the immune system and the cell cycle

The human immune system is a complex network of specialised components. The coarsest way of categorising these components is innate vs. adaptive immune cells. Each of these categories can be subdivided many times, leading to a tree-like hierarchy of immune cells of increasing granularity. The adaptive immune system is characteristic of higher organisms and differs from the more evolutionarily ancient innate immune system in two fundamental respects: specialisation and memory. The adaptive immune system, and in particular its T-cells, will be the biological focus of chapter 2 and be have a key involvement in a case study in chapter 4. We now outline some of the basic biology of the human immune system, beginning at the top of the hierarchy and increasing in granularity towards the role of an individual T-cell. We defer mathematical introductions to their respective chapters. Much of the discussion in this chapter follows Alberts [1] and Janeway [55].

1.2.1 Innate & adaptive immunity

All forms of immunity have a common purpose – protecting an organism from death by infection, i.e. the harmful influence of micro-organisms such as bacteria and viruses, generally known as *pathogens*. Vertebrates and invertebrates share a basic form of immunity known as *innate immunity*. Vertebrates have, layered on top of their innate immune system, the more specialised *adaptive* immune system.

Innate immunity can be understood as a 'first line of defence', a fast response to an invading pathogen which does not rely on prior exposure and detects its targets by general mechanisms, such as distinguishing bacterial from non-bacterial RNA fragments. Removal of targets proceeds directly by phagocytosis (literally, cell eating) in which innate immune cells engulf target cells and destroy them. In addition to this direct mechanism, innate immune cells release chemical mediators which both amplify the responses of other innate immune cells and help to activate adaptive immune cells.

The 'first line of defence' is literal in a physical sense – innate immune cells are highly concentrated near sites where pathogens most commonly enter the body, such as the gut, the lungs and the skin. Many of the most common bacterial pathogens encountered at these sites are dealt with almost immediately, without causing disease. However, occasionally this first line of defence is broken – innate immune cells are unable to effectively remove the pathogen. In this case, the adaptive immune system is recruited. The primary means of communication between innate and adaptive immune cells is the *cytokine* network. A cytokine is defined to be any protein which is released by an immune cell and which influences the behaviour of another immune cell. Cytokines are the orchestrators of the immune response and can be inflammatory or anti-inflammatory, with the balance between the two determining the extent of the response. Cytokines are a significant focus of the model of chapter 2.

1.2.2 Molecular characterisation of T-cells

T-cells are cells of the adaptive immune system involved in direct cell-to-cell killing. Their response to infection is orchestrated by innate immune responses and cell-to-cell signalling between various subtypes of T-cells themselves. The most commonly delineated sub-classes of a T-cell population are based on the expression levels of the co-receptors CD4 and CD8. These are surface proteins which interact with the T-cell receptor to influence the subsequent signalling cascade, and therefore the behaviour of the cell during an immune response. Loosely, CD4 T-cells are 'helper' cells and are responsible for directing the immune response via the production of cytokines. CD8 cells are 'killer' cells and are responsible for directly killing infected cells. In building the statistical model of chapter 4, we use as predictors various sub-sub-classes of T-cells, i.e. subdivisions of CD4 and CD8 classes. We now outline loosely the roles of these subclasses. Note that the model of chapter 2 treats the T-cell population as a whole and is not concerned with sub-divisions. Rather the heterogeneity we deal with there is in the form of cellular ages/developmental stages.

In addition to the functional distinction between CD4 and CD8 T-cells, cells can be categorised based on the expression of other surface molecules. In particular, we consider data containing expression levels of CCR7 and CD45RA. Classifying each cell as having either high or low expression levels of each of these surface proteins leads to four distinct cell types (see figure 1.1).

Naive T-cells express high levels of both CCR7 and CD45RA. The other four classes are particular types of memory cell representing different stages of development. One common theory is that T-cells progress anti-clockwise from the top-right of figure 1.1 during the course of an immune response, beginning as naive and proceeding through central memory, effector memory and CD45RA-expressing effector memory (also called terminally differentiated) stages. Central memory cells are the cells which

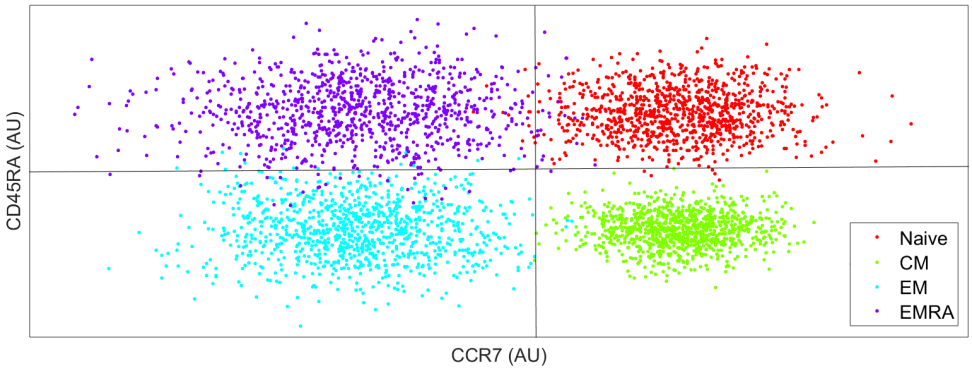


Figure 1.1: Synthetic example of flow cytometry measurements of the expression levels of the cell surface markers CCR7 and CD45RA. Each dot represents a single cell. The legend refers to typical cell subset classifications based on CCR7 and CD45RA measurements. AU: arbitrary units.

go on to survive the contraction phase of the immune response and form the basis of immunological memory, returning to rest predominantly in the secondary lymphoid organs. Uncovering this sort of phenotypic evolution is the focus of chapter 3.

The maintenance of the T-cell repertoire is tightly controlled by both global and local resources. By this we mean that both the overall size and the composition (in terms of clonotypes) of the T-cell repertoire are controlled by different, interacting mechanisms. The global control is achieved by so-called homeostatic cytokines, which provide both survival and replication signals to T-cells. This control is the subject of the modelling in chapter 2. The composition of the T-cell repertoire is controlled by separate homeostatic signals in the form of self-antigen stimulation in which proteins from the hosts own tissues provide weak signals to T-cells through their surface antigen receptors and promote survival. A very interesting question, which we do not deal with in this thesis, concerns the ways in which these signals interact to promote self-tolerance, i.e. the prevention of autoimmunity the misguided mounting of an immune response against healthy host tissue.

Graft-versus-host disease, the subject of a statistical case study in chapter 4, can be thought of loosely as an autoimmune response – transplanted T-cells recognise host

tissues as foreign and mediate their destruction, leading to disease.

1.2.3 The cell cycle

The cell cycle is the process by which a cell reproduces, culminating in cell division and the production of a pair of (up to rare mutations) genetically identical daughter cells. The process occurs in several stages, all of which refer to particular aspects of DNA replication, repair or mechanical preparations for the final division of the cell.

Cells, and in particular T-cells, are not generally in a constant state of division. Survival mechanisms, which we outline in the introduction to chapter 2, mean that cells can remain in a state of quiescence, or rest, for long periods of time, sometimes up to years without dividing [91]. Cells in this quiescent state are referred to as being in phase $G0$ of the cell cycle and simply perform their functions without dividing or dying this is crucial for the retention of immunological memory because high cell turnover favours the survival of recent thymic emigrants (freshly produced naive T-cells) over memory T-cells [91]. Upon stimulation by chemical mediators such as the homeostatic cytokines mentioned above, T-cells may exit the quiescent phase and enter a proliferative, or cycling, state in which they pass through the entire cell cycle and produce two new daughter cells. Upon completion of the cell cycle, cells return to a resting state until they are stimulated to divide again.

1.3 Mathematical models for cellular evolution

Chapters 2 and 3 focus on mathematical models for the distribution of individual cellular characteristics within a population over time. In this section, we provide a brief introduction to the mathematical modelling of dynamics at the single-cell level, the

population level, and a combination of the two.

1.3.1 ODE models for intracellular dynamics

Concentrations of proteins or mRNA within a single cell are typically modelled using ordinary differential equations derived from a collection of reaction rate equations. Much of the exposition here follows [65] and [102].

Suppose that we are modelling the dynamics of k chemical species $\{X_1, \ldots, X_k\}$ (such as intracellular proteins or mRNA) and that we have a list of chemical reactions of the form

$$m_1 X_1 + \cdots + m_k X_k \xrightarrow{r} n_1 X_1 + \cdots + n_k X_k, \tag{1.1}$$

where the m_j are the numbers of molecules of species j involved in the reaction and the n_j are the numbers of molecules of species j produced by the reaction. r denotes the reaction rate.

If one is happy to assume *mass action kinetics* – that the rate at which a reaction occurs is proportional to product of the concentrations of the reactants, which typically is only true for elementary reactions, but is often also appropriate as a coarse-grained model for higher order reactions – the dynamics of the concentration of a single molecular species, $x_j(t)$, can be written as an ordinary differential equation of the form

$$\frac{dx_j(t)}{dt} = \sum_{i=0}^{n} r_i (n_{j,i} - m_{j,i}) \prod_{l=1}^{k} {x_l}^{m_{l,i}}, \tag{1.2}$$

where n is the total number of reactions in the reaction system, $n_{j,i}, m_{j,i}$ are the numbers of molecules lost and gained, and r_i is the reaction rate of the i^{th} reaction in the system.

In general, this leads to a coupled system of nonlinear ODEs which may be analysed

numerically or using methods such as stability analysis or asymptotic expansions.

In some cases, the mass action assumption may be inappropriate and alternative assumptions may be made about the dependence of the reaction rate on the concentrations of chemicals involved – there may be saturation effects involved, in which case a Hill function is a reasonable alternative for the mass action law. The Hill function is used as part of the stochastic model used to generate synthetic data in chapter 3.

This deterministic framework allows one to model bulk properties of populations of individual cells or of concentrations in single cells with large numbers of molecules. We will outline methods by which stochasticity can be incorporated into such models below.

1.3.2 ODE models for population size dynamics

Similar ODE models are commonly used to describe how the numbers of individuals in various populations change over time. One simple example is the Lotka-Volterra predator-prey model, in which the mass-action assumption is again invoked in order to describe two biological populations in which the growth rate of one population is dependent on the availability of another for food.

The system is defined as

$$\dot{x} = x(a - by), \tag{1.3}$$

$$\dot{y} = y(cx - d), \tag{1.4}$$

of course subject to appropriate initial conditions. Here x and y represent the populations of prey and predators respectively.

These equations describe a situation in which the prey reproduce with rate a, the preda-

tors die naturally with rate d, and the reproduction/death of predator/prey depends on the current population of prey/predator.

As with the reaction rate equations of the previous subsection, other functional forms for the birth and death terms may be appropriate for different situations. Also, the dependencies between the populations being modelled are likely to vary significantly from problem to problem. However, the general principle of using the concept of reaction rates to describe population dynamics remain. These principles are relevant to the model in chapter 2 and the general framework of chapter 3.

1.3.3 Introducing stochasticity

The ODE models of the previous 2 subsections are *deterministic* – there is no randomness in the output for a given input and the solution at time t is solely determined by the parameters and initial conditions of the system.

We can think of these ODE models as the small-noise limits of *stochastic* differential equations, whereby the reaction rates are corrupted by time-dependent noise. Collecting the concentrations/population sizes of species of interest in a vector $\boldsymbol{x}(t)$ – $\boldsymbol{x}(t)$ is referred to as the *state* at time t – a stochastic differential equation can be written in Langevin form as

$$\frac{d\boldsymbol{x}(t)}{dt} = \mu(\boldsymbol{x}(t)) + \boldsymbol{\eta}(t), \tag{1.5}$$

where $\boldsymbol{\eta}(t)$ is a continuous time Gaussian process. In the case of chemical reaction systems like those outlined in the foregoing subsections, the variance structure of $\boldsymbol{\eta(t)}$ has a particular form which depends only on the reaction rates and the current concentrations [43] – the variance contributed by each reaction scales like the inverse of the reactant populations. Thus one expects to see significant population heterogeneity when the numbers of reacting molecules within individual cells are small, or when the

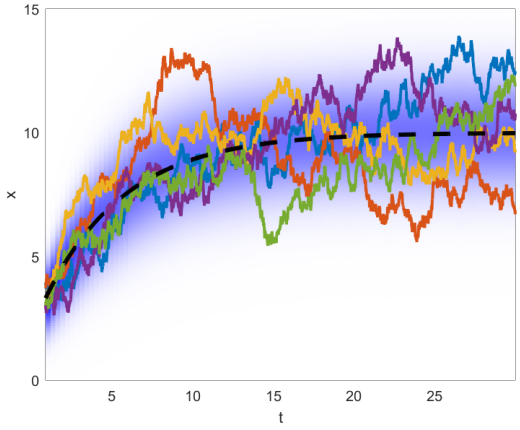


Figure 1.2: Visualisation of the solutions to the three types of equation for an Ornstein-Uhlenbeck process. Black line: ODE; Coloured lines: SDE; blue pseudocolour plot: Fokker-Planck.

numbers of members of competing populations are small.

We make use of the alternative notation

$$d\boldsymbol{x}_t = \boldsymbol{\mu}(\boldsymbol{x}_t) + \boldsymbol{\sigma}(\boldsymbol{x}_t)\boldsymbol{W}_t \tag{1.6}$$

in chapter 3. In this case, $\boldsymbol{W}_t$ is a Wiener process – a special case of the Gaussian process described above in which the elements of $\boldsymbol{W}$ are uncorrelated – the correlation structure is contained in the matrix $\boldsymbol{\sigma}$. The function $\boldsymbol{\mu}$ in both cases describes the deterministic part of the dynamics, which we described in the previous subsections.

Figure 1.2 contains an example of realisations of a particular stochastic differential equation describing the Ornstein-Uhlenbeck process, which can be thought of as a very simple model for stochastic population growth and is given by the SDE

$$dx_t = a - bx_t + \sigma dW_t, \tag{1.7}$$

where $a, b, \sigma > 0$ describe the rates of entry and exit and the influence of the noise term.

In the small-noise limit, this equation becomes the deterministic ordinary differential

equation

$$\frac{dx}{dt} = a - bx. \tag{1.8}$$

In the case that μ is a linear function of x, this deterministic equation is also precisely the expected value of the stochastic process.

1.3.4 PDE models for evolving populations

The Fokker-planck equation describes the evolution of the probability distribution over states $\boldsymbol{x}$ for a population of individuals each evolving according to some SDE. Information about individual trajectories in state-space are lost by describing the population in this way, but the Fokker-Planck equation is often a useful way of describing the bulk properties of a stochastic process.

The Fokker-Planck equation corresponding to the SDE (1.6) is given by [36]

$$\partial_t p(\boldsymbol{x}, t) = -\sum_i \partial_{x_i}[\mu_i(\boldsymbol{x}) p(\boldsymbol{x}, t)] + \frac{1}{2}\sum_i \sum_j \partial_{x_i}\partial_{x_j}\left[\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\sigma}(\boldsymbol{x})^\top p(\boldsymbol{x}, t)\right], \tag{1.9}$$

$$= \mathcal{L} p(\boldsymbol{x}, t), \tag{1.10}$$

where $\mathcal{L}$ is known as the *forward Fokker-Planck operator*.

$\mathcal{L}$ is a linear operator and, due it being self-adjoint, has purely real eigenvalues by the Sturm-Liouville theorem.

The Fokker-Planck equation, however, does not account for possible changes in population size due to reproduction, entry or exit from the system, or death.

A simple way to deal with this limitation is to include a state- and population-dependent

growth term in the Fokker-Planck equation, in which case it becomes

$$\partial_t n(\boldsymbol{x}, t) = \mathcal{L} n(\boldsymbol{x}, t) + r(\boldsymbol{x}, n(\boldsymbol{x}, t)). \tag{1.11}$$

Such models are known as *population balance* equations [75]. (1.11) is a specific form of the general population balance model, of which the age-structured model in chapter 2 is an example.

1.4 Background on statistical concepts used

Chapters 3 and 4 make use of a number of statistical concepts, to which we provide a brief introduction here.

1.4.1 Bayesian inference

The objective of Bayesian statistics is to estimate a probability distribution over the parameters, θ, of a statistical model, conditional on a set of data, $\mathcal{D}$. According to Bayes' theorem

$$\pi(\theta|\mathcal{D}) = \frac{\pi(\mathcal{D}|\theta)\pi(\theta)}{\pi(\mathcal{D})}. \tag{1.12}$$

The distribution $\pi(\mathcal{D}|\theta)$ is known as the *likelihood* and describes the dependence of the data generating process on the parameters $\theta \in \mathbb{R}^p$. The distribution $\pi(\theta)$ is known as the prior distribution and describes our beliefs about plausible values of θ before observing the data $\mathcal{D}$. The constant of proportionality is $\pi(\mathcal{D})$ and is known as the *evidence*.

The main distinguishing element of Bayesian statistics over frequentist statistics is the

use of the prior distribution over model parameters. The prior is simply a distribution chosen to represent our prior beliefs about the model parameters. These prior beliefs may be based on previous experiments, some intuition about the problem domain, or considerations about typical effect sizes in certain classes of problems (so-called default priors, [38]). A typical criticism of Bayesian methods is that the choice of prior is subjective and may exert undue influence on the results of any analysis, and while this is true of the prior, it is also true of the likelihood – the particular model chosen to perform the analysis, the predictors included, and functional forms of relationships between predictors and outcomes are also subjective [37]. With this in mind, careful consideration should be given to the full model specification and a sensitivity analysis of the full gamut of modelling choices should, ideally, be performed.

In the most common cases, the dataset $\mathcal{D}$ consists of a set of scalar outputs $\{y_i\}_{i=1}^n$ and a set of inputs $\{x_i \in \mathbb{R}^d\}_{i=1}^n$.

Bayesian inference then consists of drawing conclusions about the parameters or other quantities of interest under the posterior distribution. Most commonly, these are the *marginal* posterior distributions, $\pi(\theta_i|\mathcal{D})$ or *posterior predictive* distributions over new observations, $\pi\left(y^*|\theta, x^*, \{x_i \in \mathbb{R}^d\}_{i=1}^n\right)$. Both of these quantities, and most other quantities of interest, require integration of the posterior over a subset of its variables. For example, the marginal posterior distributions require integrating the posterior over all other elements of θ. Except in rare cases, the required integrals cannot be computed analytically or in reasonable time using numerical quadrature rules, and one resorts to sampling from the posterior distribution directly instead – techniques for doing so are covered briefly below. Bayesian inference can therefore be broken down into two steps:

- Specify the joint probability distribution $\pi(\mathcal{D}, \theta) = \pi(\mathcal{D}|\theta)\pi(\theta)$ by defining a statistical model (a likelihood) for observations and a prior distribution.

- Approximate the posterior distribution and explore the derived distributions of quantities of interest under the posterior.

The first step is problem specific – chapter 4 contains examples of models specified in this framework.

For the second step, a number of approaches are available for sampling from probability distributions which are known only up to a constant – in this case we only have access to the joint distribution $\pi(\mathcal{D}, \theta)$ but not the normalising constant $\pi(\mathcal{D})$. We cover the two most common, Metropolis-Hastings and Gibbs sampling here. An explanation of an alternative method, Hamiltonian Monte Carlo, is in chapter 4.

Given a probability distribution $p(x)$ with $x \in \mathbb{R}^d$, the Gibbs sampler draws samples from $p(x)$ by sequentially sampling from its conditional marginal distributions. The Gibbs sampler is initialised at some setting of the parameters, x_0. It then cycles through the elements of x, iteratively sampling from the conditional marginal distributions $p(x_k^{t+1}|x_{-k}^t)$. Here, x_{-k}^t is the vector $(x_1^{t+1}, \ldots, x_{k-1}^{t+1}, x_{k+1}^t, \ldots, x_d^t)$ and t is the iteration number [37].

In this way, a sample from $p(x)$ is obtained at each iteration. Often the marginal conditional distributions are available in closed form (if amenable prior distributions are chosen) and so Gibbs sampling is an appropriate method. In cases where this is not the case, the Metropolis-Hastings (MH) algorithm is an alternative.

The MH algorithm is similarly initialised at some initial parameter setting x_0. Rather than sampling from analytically-derived marginal distributions, the MH algorithm proceeds by sampling a new point, x_{prop} from close to the current estimate (according to some proposal distribution which must be specified). In the most common version of the MH algorithm (the Metropolis algorithm, which requires symmetric proposal distributions) x_{prop} is then accepted as a sample from the target distribution $p(x)$ with

probability $\min\left(1, \frac{p(x_{\text{prop}})}{p(x_t)}\right)$. In the more general case of non-symmetric proposals, a correction is required to the ratio in the second argument. The Metropolis scheme can be interpreted as a stochastic optimisation procedure, in which proposals which increase the current probability density are always accepted, while those which do not are only accepted with a probability related to how much worse they are than the current estimate [37].

Intuitively, this idea of taking small steps in random directions is an inefficient way of finding regions of high probability density and a limitation of both Gibbs and Metropolis (-Hastings) sampling that we discuss in chapter 4 is that they are typically only able to make small jumps in parameter space. This leads to behaviour such as chains – the term used for sequences of samples – becoming stuck in single modes of multimodal distributions or converging very slowly to regions of high probability if the initial estimates are far from the mode. More sophisticated strategies such as parallel tempering or Hamiltonian Monte Carlo and related methods are now available for sampling from potentially high-dimensional and multimodal posterior distributions which, again, we discuss in more detail in chapter 4.

1.4.2 Density estimation and mixture modelling

A key aspect of the methodology developed in chapter 3 is *density estimation*.

The goal of density estimation is, given a finite sample of points, $\{x_i\}_{i=1}^n$, estimate the probability distribution, $p(x)$, from which they were sampled.

The simplest option is the histogram, but another common approach is the class of mixture models. Mixture models attempt to estimate the density as a weighted sum of simple probability distributions, such as Gaussians.

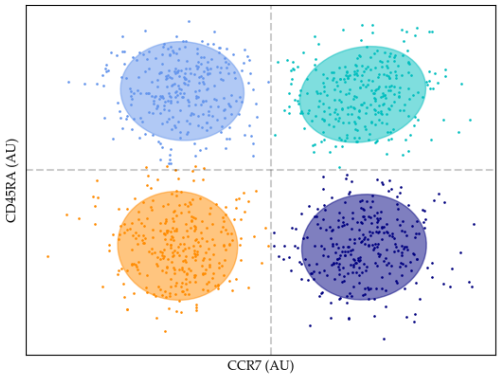


Figure 1.3: A gaussian mixture model fit to synthetic flow cytometry data. Ellipses represent regions containing 75% of the total probability for each component.

A Gaussian mixture model with k mixture components can be written

$$p(x) = \sum_{j=1}^{k} w_j \mathcal{N}(x|\mu_j, \Sigma_j), \tag{1.13}$$

where w_j are weights which sum to 1; $\mathcal{N}(x|\mu_j, \Sigma_j)$ is the (multivariate) Gaussian probability density function evaluated at x and μ_j and Σ_j are the means and covariances of the components. Typically, these three sets of parameters are estimated via the Expectation-Maximisation algorithm. An example of a 4 component Gaussian mixture model fit to synthetic two-dimensional flow cytometry data is shown in figure 1.3. One major disadvantage of the basic Gaussian mixture model, especially in high dimensions, is that the number of components, k, is often not known in advance.

One common option is to set k to be equal to the number of data points, set the weights to be equal for each component, and use the location of each datapoint as the mean for each Gaussian component. The covariance matrix is often set to be constant across observations and is known as the *bandwidth* matrix.

This approach is known as *kernel density estimation*, in which case we estimate the den-

sity $p(x)$ via

$$p(x) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{N}(x|x_j, \Sigma). \tag{1.14}$$

One common choice for the shared covariance matrix Σ is the Silverman matrix [85], a diagonal matrix with entries

$$\Sigma_{ii} = \frac{4}{n(d+2)}^{2/(d+4)} \sigma_i^2,$$

where d is the dimensionality of the data and σ_i is the sample variance along the i^{th} dimension.

The Silverman matrix is a simple rule-of-thumb approach to bandwidth selection, and may be replaced by methods which use the data to estimate appropriate shared covariance matrices. One can use bootstrapping or cross-validation to select appropriate covariance matrices [48,78]. An alternative approach is to attempt to estimate the number of components alongside the means, covariances and weights for each component.

One way of doing this is through a Bayesian nonparametric approach called the *Dirichlet process Gaussian mixture model* [29].

One takes $k \to \infty$ in (1.14) and places a Dirichlet process (DP) prior on the weights w_j. The DP is an infinite dimensional generalisation of the Dirichlet distribution, which is defined on the unit simplex (the space of vectors whose elements sum to one). The parameters of the Dirichlet process are a *base measure* H_0 and a *concentration* parameter α – samples from a Dirichlet process are *atomic* probability distributions, i.e. a countably infinite set of locations, each associated with a weight, where the locations are drawn from H_0 and α controls the typical number of non-zero weights.

In order to use the DP in the setting of a Gaussian mixture model, one can use it as a prior for the weights, means and covariances of the components. In that case H_0

becomes a shared prior distribution on the means and covariances. Inference is made tractable by the fact that, *almost surely* (i.e. with probability 1), the number of non-zero weights is finite. In practice, it is often quite small for typical settings of α.

The full Dirchlet process Gaussian mixture model can be specified as

$$G \sim DP(\alpha H_0), \tag{1.15}$$

$$(\mu_i, \Sigma_i) \sim G, \tag{1.16}$$

$$x_i \sim \mathcal{N}(\mu_i, \Sigma_i). \tag{1.17}$$

Here, observations x_i are drawn from a normal distribution with some mean and covariance μ_i, Σ_i. The mean and covariance are themselves drawn from an atomic distribution G, i.e. a distribution with countably infinite support, and G is itself drawn from a Dirichlet process with base measure H_0.

The procedure we use to infer the means and covariances below is a *variational* Bayesian method, a way to obtain approximate posterior distributions by approximating them with simpler, parametric distributions and optimising those parameters rather than sampling from the distribution [14, 58]. Often, this is much faster than sampling via, say, Gibbs sampling. Full details of the accelerated variational Dirichlet process method are available in the original paper [58] and in the MATLAB code [1] for its implementation in the context of Gaussian mixture models.

1.4.3 Dimensionality reduction

The final piece of background material is an overview of relevant dimensionality reduction methods: principal component analysis, non-negative matrix factorisation and

[1] Available at: https://sites.google.com/site/kenichikurihara/academic-software/variational-dirichlet-process-gaussian-mixture-model

diffusion maps (including how they relate to discrete approximations to Fokker-Planck operators).

Firstly, we cover principal component analysis (PCA) as an introduction to the problem of dimensionality reduction. Suppose we have a matrix of observations $X \in \mathbb{R}^{n \times d}$. PCA aims to find a new set of coordinate axes in which to represent the data, such that the new dimensions are uncorrelated and are ordered by variance.

The principal components can be computed using the *singular value decomposition* (SVD) of X, $X = U\Sigma V^\top$, where Σ is a scaled version of the (diagonal) covariance matrix in the new coordinates. The dimensionality of the data can subsequently be reduced by projecting it onto only the first few principal components (the dimensions with the largest variance).

In certain circumstances, the data contained in X are non-negative and one would like to preserve this non-negativity after dimensionality reduction. For this purpose, non-negative matrix factorisation is appropriate. Similarly to PCA, the objective is to decompose the data matrix X into factors, W, H. These non-negative factors are determined through an alternating algorithm which preserves the non-negativity at each iteration [59].

PCA can be viewed as a way of uncovering the manifold on which the data lie. In the case of PCA, this manifold is assumed to be linear, i.e. it is a hyperplane. A number of methods exist to relax this linearity assumption. We focus on diffusion maps [66] because of their relevance to the discussion in chapter 3.

Diffusion maps work on a matrix of similarities between points, as computed according to some kernel function defined by the user, rather than the data themselves. This similarity matrix, when appropriately normalised, can be interpreted as a transition matrix describing a random walk on the data – the probability that a random walker transitions from data point i to data point j is proportional to the similarity between

the two points. Typically the similarity function is chosen to be a Gaussian function of the Euclidean distance between the points.

In the limit as the number of data points becomes very large, this random walk becomes a diffusion process in continuous space and can be described by a Fokker-Planck equation. It can be shown [66] that the eigenvalues and eigenvectors of the transition matrix are finite-dimensional approximations to the eigenvalues and -vectors of a Fokker-Planck operator describing diffusion in a potential and whose stationary distribution (i.e. the large-time solution) is the probability distribution from which the data were sampled.

CHAPTER 2

MODELLING A POPULATION OF DIVIDING CELLS WITH LIMITED RESOURCES: A PDE SYSTEM WITH AGE-STRUCTURE AND A LIMITED RESOURCE (ASLR)

2.1 Biological Motivation

Understanding quantitatively the proliferation kinetics of T-cells in the body is an important problem for biologists and clinicians, particularly in the context of haematopoietic stem cell transplantation (HSCT), a common treatment for haematologic disorders such as chronic myelogenous leukaemia [101]. The transplant procedure aims to eliminate the malfunctioning haematopoietic system of the host and replace it with a new, healthy system. However, complications arise due to the incomplete nature of the transplant. That is, the graft contains far fewer T-cells than are usually present in the host. Homeostatic mechanisms, in most cases, restore the T-cell population to its equilibrium size. However, during the recovery period the new immune system of the host is severely weakened, leaving patients vulnerable to infection and illness. The dynamics of T-cell reconstitution are also likely to have a considerable influence on other aspects of the response to HSCT, such as graft-versus-host disease, as we discover in

chapter 3. A quantitative understanding of the reconstitution kinetics should aid in the planning of treatment courses for patients undergoing HSCT.

This reconstitution of T-cell numbers is presumed to be driven by the same factors that control the size of the T-cell population in healthy, homeostatic conditions – a homeostatic set point is determined by the availability of limited resources, which become more abundant in the lymphodepleted post-transplant environment. We are particularly concerned with *mature* T-cells, i.e. those that have completed the transition from stem cell to fully formed lymphocyte and are circulating between blood and lymphoid organs. The pool of mature T-cells includes both naïve and memory T-cells. The maintenance of the mature T-cell population in mice and humans is controlled by a variety of survival and growth factors. The most important of these are the 'homeostatic cytokines' IL-7 and IL-15. Signals from these cytokines act in concert with 'tonic' T-cell receptor signals to promote the survival and proliferation of mature T-cells [91]. The central roles of IL-7 and IL-15 in T-cell homeostasis have been revealed via knockout experiments, particularly in mice, and their effects have been investigated previously using ordinary differential equation models, from the elegantly simple [49] to the comprehensive [77].

The model presented below aims to capture the essential ingredients of T-cell homeostasis, which encompasses reconstitution of T-cell counts post-transplant, in the mathematically interesting setting of age-structured modelling. Essentially, the model accounts for variation among cell populations in proliferation and survival statistics, with these processes being controlled by a physiologically produced limited resource (analogous to the homeostatic cytokines and self-peptide complexes that drive T-cell homeostasis and reconstitution *in vivo*). The model, of course, is a simplified representation of the real world phenomena. It can be thought of as an extension of the previous simple model of Hapuarachchi [49] to include intra-population variability. The Hapuarachchi model has recently been used successfully to interpret clinical data

in an AlloHSCT setting [50], illustrating the fact that its simplicity is not necessarily a hindrance.

The chapter is structured as follows: we first provide a brief overview of structured population models and their probabilistic interpretation (section 2.2), followed by the formulation of an age-structured model for T-cell homeostasis (section 2.4). We then examine the equilibrium properties of the model both analytically (section 2.5) and numerically (section 2.6), and conclude with some discussion of the implications and limitations of the model (section 2.7).

2.2 Structured Population Models: the i-level and the p-level

Any biological population is made up of individuals. In some cases these individuals behave very similarly, while in others there is considerable intra-population variability. In a population of cells, for example, individual cells may have varying intracellular concentrations of anti-apoptotic proteins (inhibitors of programmed cell death) which govern their survival behaviour. The survival of the population as a whole is determined by the survival of the individual cells, which in turn is determined by the dynamics of anti-apoptotic protein levels within those cells. A key question is, How can we link knowledge of the individual dynamics with those of the population as a whole? This, incidentally, is the central problem of statistical mechanics, whereby microscopic details are studied in order to describe the macroscopic behaviour of, for example, thermodynamic processes.

A key difference between the modelling of physical processes and biological processes is the incorporation of changes in population size – this complicates the description of a system in terms of probability distributions due to ever-changing normalising con-

stants. However, the main idea of incorporating individual-level detail into population level models is made possible by 'structuring' the population, that is describing the population in terms of the individual characteristics of its constituents. The central advantage of structured population modelling lies in this linking of individual (i) -level dynamics with the behaviour of the overall population (p-level), while allowing for tractable numerical and analytical progress. As an example, we outline below one of the simplest structured population models: the renewal equation, pioneered by Lotka and McKendrick in the early 1900s [63, 83]. This will form the basis for the more detailed model discussed in the following sections. The renewal equation uses age as the 'structuring variable', i.e. a variable describing some quantity of interest within the individuals of the population; examples include age, size and protein concentration.

The renewal equation can be stated in the form of a partial differential equation for the population age density $n(x,t)$, and is supplemented with a non-local flux boundary condition describing the birth process:

$$n_t + n_x = 0, \tag{2.1}$$

$$n(0,t) = \int_0^\infty B(x)n(x,t)\,dx, \tag{2.2}$$

$$n(x,0) = n_0(x). \tag{2.3}$$

The structuring variable x is used here to denote the age of an individual. The boundary condition (2.2) reflects that individuals are born (enter the population at age zero) according to the birth kernel, $B(x)$. Note that system (2.1)-(2.3) is at its heart a linear advection equation – constant solutions propagate along straight line characteristics in the (x,t)-plane with unit velocity, with the constant along each characteristic being determined by the boundary condition (2.2). The total population size $N(t) = \int_0^\infty n(x,t)\,dx$ exhibits exponential growth, and the population settles into an exponentially decaying age distribution (i.e. proportional to $e^{-\lambda x}$) in the large time asymptotic limit [69].

2.2.1 Probabilistic interpretation of renewal equations

We would expect the system (2.1) - (2.3) to exhibit exponential growth in time because it represents a pure birth process; there is no mechanism included to account for death, migration, or any other means of leaving the population. It is straightforward to include such a mechanism. We introduce an age-dependent 'removal rate' $\mu(x)$ and discuss its probabilistic interpretation below. This will form the basis for the formulation presented in section 2.4 and provides a precise means for choosing the age dependence of the rates in the model.

The introduction of the age-dependent removal rate $\mu(x)$ leads to the following version of the renewal equation:

$$n_t + n_x = -\mu(x)n, \tag{2.4}$$

$$n(0,t) = \int_0^\infty B(x)n(x,t)\,dx, \tag{2.5}$$

$$n(x,0) = n_0(x). \tag{2.6}$$

$\mu(x)$ is the instantaneous removal rate of individuals of age x. If we call the probability distribution over removal times $\pi(x)$, and its corresponding cumulative density function $\Pi(x)$. The *survival function* $S(x) = 1 - \Pi(x)$ gives the probability that an individual in an infinitesimal age interval around x has not yet been removed from the population.

We then have the following equality for the distribution of removal ages:

$$\pi(x) = S(x)\mu(x) \tag{2.7}$$

In words, the probability that an individual is removed from the population in the age range $[x, x+dx]$ is the probability that an individual has not yet been removed multiplied by the instantaneous rate at which individuals in the age range $[x, x+dx]$

are removed.

Alternatively, the instantaneous removal rate $\mu(x)$ (also called the hazard rate or hazard function) can be written solely in terms of the PDF $\pi(x)$ as follows:

$$\mu(x) = \frac{\pi(x)}{S(x)} = \frac{\pi(x)}{1 - \int_0^\infty \pi(x')\,dx'}. \tag{2.8}$$

The birth kernel, $B(x)$, uniquely defines the probability distribution over ages at which individuals give birth in a similar way.

As an example, if removal ages are exponentially distributed, $\pi(x) = \lambda e^{-\lambda x}$, we have

$$\mu(x) = \lambda = \text{const.}$$

Using this approach, we can relate intra-population variability in removal rates directly and precisely to the rates in the model. As we discuss below, there may exist more than one mechanism for removal from a population. In this case, the individual hazard rates for each removal process combine additively to produce an overall hazard rate for the population. This overall hazard rate uniquely determines a probability distribution of removal times from the compartment.

2.3 Previous work using structured models in the context of T-cell homeostasis

Age-structured models have been studied in both theoretical and applied settings in the context of T-cell homeostasis. Applied work has focused mainly on the Smith- Martin model of the cell cycle [27,53], in which cells are tracked by the number of divisions they have undergone. The age-structure appears in the model for the resting phase,

while the cycling phase is described by ordinary differential equations. The parameters of these models have been successfully estimated using data on the expression of CFSE, a protein which is distributed evenly between the daughters when a cell divides and which therefore allows measurement of the number of times an individual cell has divided [53, 106]. Our model is in the same spirit, but generalises in two ways – the scope for a non-degenerate distribution of cell-cycle times (i.e. we allow a probability distribution over cycle times rather than assuming that they are of a fixed deterministic length) and the introduction of resource-dependent recruitment and death rates. A resource-dependent DDE model, essentially a special case of the model studied here, has also been applied to in vitro T-cell division experiments [35].

2.4 Model Formulation: the ASLR model

The model for consideration in this section contains four compartments: cycling cells (denoted p), resting cells (denoted q), undivided cells (denoted U) and a cytokine resource (denoted I). We denote the model ASLR, for age-structured with limited resource.

In figure 2.1 we present a schematic representation of the model, illustrating the processes encoded and some example distributions of removal/transfer times. We discuss each of the model compartments and their interactions below. The equations for p and q are age-structured partial differential equations in the manner of (2.4) - (2.6), while the equations for U and I are ordinary differential equations without age-structure.

The inclusion of the undivided cells compartment U ensures that a compatibility condition is satisfied. That is, the initial and boundary conditions for p and q agree at the point $(x, t) = (0, 0)$. An initial discontinuity propagates along the characteristic curve emanating from the origin if this condition is not satisfied, and solutions exist only in

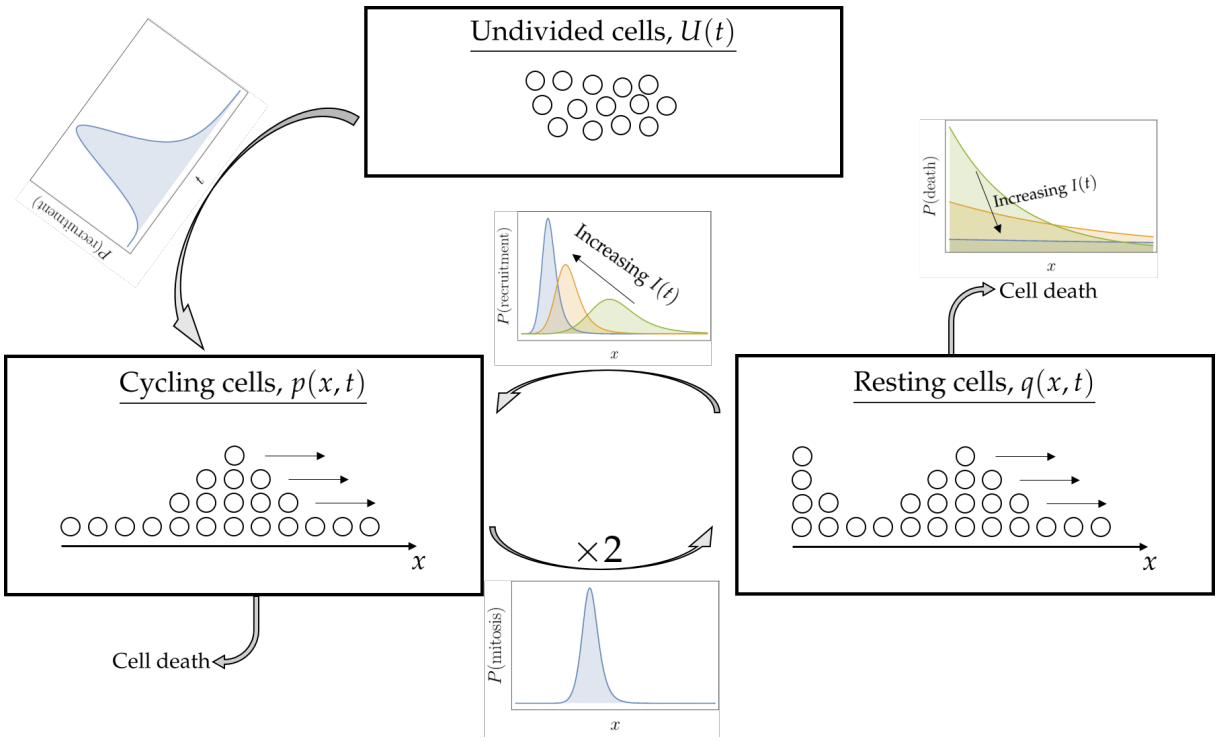


Figure 2.1: Schematic of the model (ASLR). Cells transfer between compartments according to probability distributions. The distributions of death and recruitment times in the resting compartment are under the control of the external resource $I(t)$.

the weak (integral) sense.

The full model, in dimensional form, is presented below with a justification of the terms appearing in each equation.

2.4.1 Undivided cells, $U(t)$

All cells are initially undivided, die at a constant rate d_u and can be recruited into the cell cycle at some rate $g(I(t),t)$. We denote by N_0 the initial size of the cell population. This leads to the following ODE and initial condition for the undivided cell population:

$$\frac{dU}{dt} = -(d_u + g(I(t),t))U, \tag{2.9a}$$

$$U(0) = N_0 \tag{2.9b}$$

2.4.2 Resting, or quiescent, cells, $q(x,t)$

Cells age at unit velocity (i.e. cell age is measured in the same units as experimental time, and increases at the same rate), die at a cytokine-dependent rate $\mu_q(I(t))$ and are recruited into the cell cycle at an age and cytokine-dependent rate $\lambda(I(t),x)$. The boundary condition at $x = 0$ states that the number of newly quiescent cells at time t, i.e. the flux across the zero-age boundary is equal to double (due to cell division) the number of cells completing the cell cycle and dividing at time t (discussed in the p compartment below). We thus have the following PDE with initial and boundary conditions:

$$\frac{\partial q}{\partial t} + \frac{\partial q}{\partial x} = -(\mu_q(I(t)) + \lambda(I(t),x))q, \tag{2.10a}$$

$$q(0,t) = 2\int_0^\infty m(x)p(x,t)\,dx, \tag{2.10b}$$

$$q(x,0) = q_0(x). \tag{2.10c}$$

2.4.3 Cycling, or proliferating, cells, $p(x,t)$

Similarly to the resting compartment, cells age at unit velocity. Cells die at a constant rate μ_p and complete the cell cycle at age-dependent rate $m(x)$, with m standing for mitosis, the juncture of the cell cycle at which the nucleus splits in two and which immediately precedes cell division. The flux through the zero-age boundary is the total number of quiescent cells recruited plus the number of undivided cells recruited at time t. We thus have the following PDE with initial and boundary conditions:

$$\frac{\partial p}{\partial t} + \frac{\partial p}{\partial x} = -(\mu_p + m(x))q, \tag{2.11a}$$

$$p(0,t) = \int_0^\infty \lambda(I(t),x)q(x,t)\,dx + g(I(t),t)U, \tag{2.11b}$$

$$p(x,0) = p_0(x). \tag{2.11c}$$

2.4.4 Cytokine resource, $I(t)$

The dynamics of a trophic cytokine which stimulates both the survival and entry into the cell cycle of resting cells (compartments q and U) are described by the below ODE.

The resource is consumed by both of these sub-populations at a rate $f(Q, U)$ where $Q = \int_0^\infty q(x,t)\,dx$ is the total size of the quiescent cell population. We impose the restrictions that f is a non-negative function and that $f(0,0) = 0$ (i.e. there is no consumption of cytokine if no cells are available to consume it). Additionally, cytokine is produced at a constant rate α_I and is degraded at a constant rate δ_I. We choose a constant rate of cytokine production to reflect that homeostatic cytokines are thought not to be produced by T-cells themselves, rather at a reasonably constant rate by cells lining the secondary lymphoid organs known as fibroblastic reticular cells [91].

We therefore have:

$$\frac{dI}{dt} = \alpha_I - \delta_I I - f(Q,U)I, \tag{2.12a}$$

$$I(0) = I_0 \tag{2.12b}$$

Henceforth, we refer to the full system, (2.9) – (2.12) as (ASLR), for Age-Structured with Limited Resource.

2.5 Steady Distributions and Their Stability

In this section we determine the steady states of (ASLR), that is the large-time age distributions of resting and cycling cells, and the steady concentration of cytokine. We seek solutions $p_s(x)$, $q_s(x)$, U_s, I_s to (ASLR) with vanishing time derivatives. That is, we seek solutions to the following system (obtained via setting time derivatives to

zero), where primes denote x derivatives:

$$p_s'(x) = -(\mu_p + m(x))p_s, \tag{2.13}$$

$$q_s'(x) = -(\mu_q(I_s) + \lambda(I_s, x))q_s, \tag{2.14}$$

$$0 = -(d_u + g(I_s, t))U_s, \tag{2.15}$$

$$0 = \alpha_I - \delta_I I_s - f(Q_s, U_s)I_s. \tag{2.16}$$

satisfying

$$p_s(0) = \int_0^\infty \lambda(I_s, x)q_s(x)\,dx, \tag{2.17}$$

$$q_s(0) = \int_0^\infty m(x)p_s(x)\,dx. \tag{2.18}$$

It is clear that one steady solution is given by

$$\{p_s, q_s, U_s, I_s\} = \left\{0, 0, 0, \frac{\alpha_I}{\delta_I}\right\}. \tag{2.19}$$

We refer to this as the trivial steady solution (despite the non-zero steady cytokine concentration). We now seek other steady solutions, which we refer to as non-trivial.

Equations (2.13) and (2.14) can be integrated directly, yielding:

$$p_s(x) = p_s(0)e^{-\mu_p x}e^{-\int_0^x m(x')\,dx'}, \tag{2.20}$$

$$q_s(x) = q_s(0)e^{-\mu_{q,s} x}e^{-\int_0^x \lambda_s(x')\,dx'}, \tag{2.21}$$

where $\mu_{q,s} = \mu_q(I_s)$ and $\lambda_s(x) = \lambda(I_s, x)$.

We note that the second exponentiated terms in each of equations (2.20) & (2.21) are precisely the survival functions associated with completion of the cell cycle and entry into the cell cycle, respectively, which we denote by $M(x)$ and $\Lambda_s(x)$. We therefore

have the following proportionality relations for the steady age distributions:

$$p_s(x) \propto e^{-\mu_p x} M(x) =: S_p(x), \tag{2.22}$$

$$q_s(x) \propto e^{-\mu_{q,s} x} \Lambda_s(x) =: S_q(x). \tag{2.23}$$

S_p and S_q denote the overall survival functions for cells in states p and q.

Clearly, since $d_u + g(I_s, t) > 0$, $U_s = 0$ is the only solution to (2.15).

It remains to establish the steady cytokine concentration I_s. Clearly, from (2.16), we have:

$$I_s = \frac{\alpha_I}{\delta_I + f(Q_s, 0)}, \tag{2.24}$$

where

$$Q_s = \int_0^\infty q_s(x)\, dx = q_s(0) \int_0^\infty S_q(x)\, dx = \mu_{\text{res}} q_s(0), \tag{2.25}$$

and μ_{res} denotes the mean residence time of cells in the quiescent compartment. The result follows directly from the fact that $\mathbb{E}(X) = \int_0^\infty S(x)\, dx$ for any non-negative random variable X with survival function S.

2.5.1 A necessary condition for the existence of non-trivial steady distributions

The existence of non-trivial steady distributions for classes p and q is dictated by the values $p_s(0)$ and $q_s(0)$. Additionally, the strict positivity of either $p_s(0)$ or $q_s(0)$ is enough to ensure the strict positivity of the other. This can be seen via substitution of, for example, (2.21) into (2.17) – therefore, to show the existence of non-zero steady distributions, it is sufficient to show that either $q_s(0) > 0$ or $p_s(0) > 0$. We show below

that, under certain conditions, $q_s(0) > 0$.

Using boundary conditions (2.17) & (2.18) we obtain the following equality for $q_s(x)$:

$$q_s(x) = 2e^{-\mu_{q,s}x}\Lambda(x)\int_0^\infty m(y)p_s(y)\,dy, \tag{2.26}$$

$$= 2e^{-\mu_{q,s}x}\Lambda(x)\int_0^\infty e^{-\mu_p y_1}m(y_1)M(y_1)\,dy_1\int_0^\infty \lambda_s(y_2)q_s(y_2)\,dy_2, \tag{2.27}$$

$$:= H(x)\int_0^\infty \lambda_s(y_2)q_s(y_2)\,dy_2, \tag{2.28}$$

where we let

$$H(x) = 2e^{-\mu_{q,s}x}\Lambda(x)\int_0^\infty e^{-\mu_p y_1}m(y_1)M(y_1)\,dy_1. \tag{2.29}$$

Equality (2.28) is a homogeneous Fredholm equation of the second kind satisfying:

$$\langle q_s, \lambda_s\rangle = \langle H, \lambda_s\rangle\langle q_s, \lambda_s\rangle, \tag{2.30}$$

where $\langle\cdot,\cdot\rangle$ denotes the L^2 inner product over the non-negative reals. We note also that the solution q_s can be written:

$$q_s(x) = \langle q_s, \lambda_s\rangle H(x). \tag{2.31}$$

We therefore require that $\langle q_s, \lambda_s\rangle \neq 0$ for the existence of a non-trivial steady distribution for q. By (2.30), then, we need $\langle H, \lambda_s\rangle = 1$.

Alternatively, there exists a non-trivial steady distribution $q_s(x)$ if $H(x)\lambda_s(x)$ is a probability density function.

Some manipulation allows the statement of the following condition for the existence of non-trivial steady distributions $q_s(x)$ and $p_s(x)$. Arguing via the monotonicity of G

(as defined below) is a common approach in problems of this type [26].

Theorem. *There exist non-trivial steady solutions $q_s(x)$ and $p_s(x)$ to (2.14) & (2.13) if and only if*

$$\langle S_{\mu_p}, \pi_m \rangle \geq \frac{1}{2} \tag{2.32}$$

where S_{μ_p} is the survival function associated with cell death in the proliferating compartment and π_m is the probability density function describing the distribution of times until mitosis.

Proof. By (2.30), we require $\langle H, \lambda_s \rangle = 1$ so that $\langle q_s, \lambda_s \rangle$ is non-zero and, as a result, $q_s(x)$ and $p_s(x)$ are both non-trivial. Alternatively, we require

$$G(I_s) := 2 \int_0^\infty e^{\mu_p y} \pi_m(y)\, dy \int_0^\infty e^{-\mu_{q,s} x} \pi_m \lambda, s(x)\, dx = 1, \tag{2.33}$$

where the dependence on I_s is due to the fact that $\mu_{q,s} = \mu_q(I_s)$ and $\pi_m \lambda, s = \lambda(I_s m x)\Lambda(I_s, x)$.

We now make use of the following assumed properties of $\mu_q(I)$ and $\pi_m \lambda$:

1. μ_q is a monotone decreasing function of I_s. That is the death rate of resting cells decreases as the amount of survival resource increases.

2. $\mathbb{E}(X_\lambda | I) = \int_0^\infty x \pi(x; I)\, dx$ is a monotone decreasing function of I, where X_λ is a random variable denoting the age of recruitment into the cell cycle. That is, the mean age of recruitment into the cell cycle decreases as the amount of cytokine increases.

Under these assumptions it is clear that G is a monotone increasing function of I_s. We therefore require that

$(C1)$ $\lim_{I_s \to \infty} G(I_s) \geq 1,$

$(C2)$ $G(0) \leq 1$,

in order for a unique value of I_s satisfying 2.33 to exist.

We have that

$$\lim_{I_s \to \infty} G(I_s) = 2 \int_0^\infty e^{-\mu_p y} \pi_m(y)\, dy. \tag{2.34}$$

Therefore, in order to satisfy $(C1)$, we must have:

$$2 \int_0^\infty e^{-\mu_p y} \pi_m(y)\, dy \geq 1, \tag{2.35}$$

which we write compactly as

$$\rho_m = \text{const.} = \langle S_{\mu_p}, \pi_m \rangle \geq \frac{1}{2}. \tag{2.36}$$

Here, ρ_m is the probability that a cycling cell completes the cycle before it dies. We can think of cells in compartment p as being subject to competing risks, that is two distinct ways of leaving the compartment p. These two ways are cell death and transfer to compartment q after dividing. The function $S_{\mu_p} \pi_m$ gives the *cause-specific hazard rate* for division, i.e. the instantaneous rate at which cells in p in some infinitesimal age interval are removed due to division alone, given that they have been subject to the competing risks. The integral over all x of this quantity gives the probability that a cell is removed due to division alone over the whole time it is in compartment p.

We also require that $(C2)$ is satisfied, that is

$$2 \langle S_{\mu_p}, \pi_m \rangle \int_0^\infty e^{-d_u x} \pi_\lambda(x; 0)\, dx \tag{2.37}$$

$$= 2 \langle S_{\mu_p}, \pi_m \rangle \langle S_{\mu_{q,0}}, \pi_\lambda(x; 0) \rangle \leq 1. \tag{2.38}$$

This is trivially satisfied for any reasonable choice of π_λ – the probability that a resting

cell is recruited to cycle before it dies, given by the second inner product above, should be very small in the absence of cytokine. □

Corollary. *A sufficient condition for the existence of a non-trivial steady solution $q_s(x)$ is*

$$\mathbb{E}(X_{mit}) \leq \frac{\log(2)}{\mu_p} \tag{2.39}$$

where X_{mit} is a random variable denoting the time until mitosis, i.e. a random variable distributed according to π_m.

Proof. The result follows from an application of Jensens inequality to (2.36).

We have

$$\langle S_{\mu_p}, \pi_m \rangle = \mathbb{E}(e^{-\mu_p X_{\text{mit}}}) \geq e^{-\mu_p \mathbb{E}(X_{\text{mit}})}, \tag{2.40}$$

so that

$$e^{-\mu_p X_{\text{mit}}} \geq \frac{1}{2} \tag{2.41}$$

guarantees (2.36). □

Remark. *The term on the right hand side of inequality (2.39) is the halving time of the population in the pure death case. In words, condition (2.39) says that the average time to complete the cell cycle should be smaller than this halving time. Alternatively, the average cell should have at least a 50% chance of making it through the cycle.*

In order to complete the proof of the existence of non-trivial steady states when (2.36) holds, there is one more step. It remains to verify that the unique I_s guaranteeing (2.33) also satisfies (2.24).

To this end, we note that I_s is restricted by (2.33) and that I_s in turn restricts $q_s(0)$ via

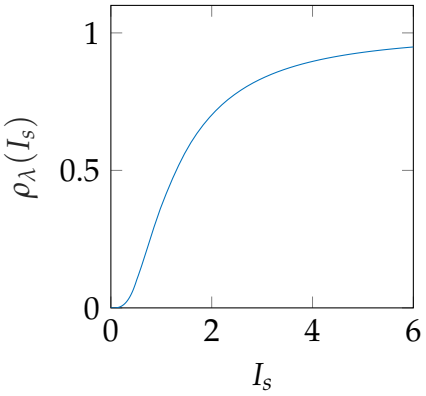


Figure 2.2: Example plot of the function ρ_λ, the probability that a cell successfully survives its time in compartment q as a function of I_s.

(2.24) and (2.25).

Concretely, we have

$$2\rho_m\rho_m\lambda(I_s) = 1 \tag{2.42}$$

so that

$$I_s = \rho_m\lambda^{-1}(\frac{1}{2\rho_m}) \tag{2.43}$$

where $\rho_\lambda(I_s)$ denotes the probability that a resting cell is recruited before it dies at cytokine concentration I_s and is defined similarly to ρ_m in (2.36).

An example plot of the function $\rho_\lambda(I_s)$ is presented in figure 2.2, with the integration being performed numerically in MATHEMATICA; λ is chosen to be the hazard function of a log-logistic distribution. The function is bounded above by 1 and, for physically reasonable restrictions on the I dependence of λ, is an increasing function of I – since the resource promotes survival, the probability of survival should increase as the availability of the resource increases. Hence its inverse exists whenever $\left(\frac{1}{2\rho_m}\right) \leq 1$, i.e. whenever (2.36) is satisfied.

The constant $q_s(0)$ can then be determined via (2.24) & (2.25). For example, making the choice $f(Q,U) = (Q+U)c$ corresponding to a constant rate of cytokine consumption

per cell, we have

$$q_s(0) = \frac{\alpha_I - \delta_I I_s}{c I_s \mu_{\text{res}}(I_s)}$$

from which $p_s(0)$ can be determined via (2.17) & (2.21) and the full steady solution can be determined.

2.5.2 Linear stability of the trivial steady state

To examine the stability of the trivial steady state, we introduce small perturbations and consider their asymptotic behaviour in time. That is, we consider

$$p(x,t) = 0 + \phi_1(x,t), \tag{2.44}$$

$$q(x,t) = 0 + \phi_2(x,t), \tag{2.45}$$

$$U(t) = 0 + \phi_3(t), \tag{2.46}$$

$$I(t) = \frac{\alpha_I}{\delta_I} + \phi_4(t), \tag{2.47}$$

where $\phi_i \ll 1$ for each $i = 1, 2, 3, 4$.

Substituting into system (ASLR), Taylor expanding non-linear terms and keeping only first order terms, we obtain the following linear system:

$$\frac{\partial \phi_1}{\partial t} + \frac{\partial \phi_1}{\partial x} = -(\mu_p + m(x))\phi_1, \tag{2.48}$$

$$\frac{\partial \phi_2}{\partial t} + \frac{\partial \phi_2}{\partial x} = -\left(\mu_q\left(\frac{\alpha_I}{\delta_I}\right) + \lambda\left(\frac{\alpha_I}{\delta_I}, x\right)\right)\phi_2, \tag{2.49}$$

$$\frac{d\phi_3}{dt} = -\left(d_u + g\left(\frac{\alpha_I}{\delta_I}, t\right)\right)\phi_3, \tag{2.50}$$

$$\frac{d\phi_4}{dt} = -\delta_I \phi_4. \tag{2.51}$$

Note that (2.48) and (2.49) will be coupled through their boundary conditions at $x = 0$.

We notice immediately that

$$\phi_4(t) = C_4 e^{-\delta_I t}, \tag{2.52}$$

with C_4 constant, solves (2.51), so that perturbations to the cytokine concentration decay exponentially in time.

We seek separable solutions $\phi_j(x,t) = \alpha_j(x)\beta_j(t), j = 1,2$ to (2.48) and (2.49), so that

$$\frac{\dot{\beta}_1}{\beta_1} = -\frac{\alpha_1'}{\alpha_1} - (\mu_p + m(x)) = \sigma_1, \tag{2.53}$$

$$\frac{\dot{\beta}_2}{\beta_2} = -\frac{\alpha_2'}{\alpha_2} - (\mu_q(I_s) + \lambda(I_s, x)) = \sigma_2, \tag{2.54}$$

where dots denote time derivatives and primes denote age derivatives.

We then have

$$\beta_j(t) = C_j e^{\sigma_j t}, j = 1,2, \tag{2.55}$$

for some constants C_j. Thus, the instability (stability) of the steady distributions $p_s(x) = 0$ and $q_s(x) = 0$ are determined by the positivity (negativity) of the real part of σ.

We are left with the eigenproblem

$$\alpha_1' = -(\mu_p + m(x))\alpha_1 - \sigma_1\alpha_1, \tag{2.56}$$

$$\alpha_2' = -(\mu_q(I_s) + \lambda(I_s, x))\alpha_2 - \sigma_2\alpha_2, \tag{2.57}$$

with the boundary conditions

$$\alpha_1(0) = \int_0^\infty \lambda(I_s, x)\alpha_2(x)\, dx, \tag{2.58}$$

$$\alpha_2(0) = 2\int_0^\infty m(x)\alpha_1(x)\, dx. \tag{2.59}$$

We proceed similarly to subsection 2.5.1.

Integrating (2.56) and (2.57) we obtain

$$\alpha_1(x) = \alpha_1(0)M(x)\exp[-\mu_p x - \sigma_1 x], \tag{2.60}$$

$$\alpha_2(x) = \alpha_2(0)\Lambda(I_s, x)\exp[-\mu_q(I_s)x - \sigma_2 x], \tag{2.61}$$

so that

$$\alpha_2(x) = 2\Lambda(I_s, x)\exp[-\mu_q(I_s)x - \sigma_2 x]\int_0^\infty h_m(y)\exp[-\sigma_1 y]\,dy\int_0^\infty \lambda(I_s, y)\alpha_2(y)\,dy, \tag{2.62}$$

where $h_m(y) = S_p(y)m(y)$ is the probability that a cell is removed from the proliferating compartment via division in some small age interval around y. This h_m is called the *cause-specific* hazard rate [74].

Similarly to section 2.5.1, we obtain an integral equation for α_2:

$$\alpha_2(x) = H_1(x; \sigma_1, \sigma_2)\int_0^\infty \lambda(I_s, y)\alpha_2(y)\,dy, \tag{2.63}$$

with

$$H_1(x; \sigma_1, \sigma_2) = 2\Lambda(I_s, x)\exp[-\mu_q(I_s)x - \sigma_2 x]\int_0^\infty h_m(y)\exp[-\sigma_1 x]. \tag{2.64}$$

Taking the inner product of (2.63) with λ_s yields the relationship

$$\langle \alpha_2, \lambda_s\rangle = \langle H_1, \lambda_s\rangle\langle \alpha_2, \lambda_s\rangle. \tag{2.65}$$

We require $\langle \alpha_2, \lambda_s\rangle \neq 0$, so that $\langle H_1, \lambda_s\rangle = 1$. Now

$$\langle H_1, \lambda_s\rangle = 2\int_0^\infty h_\lambda(x)e^{-\sigma_2 x}\,dx\int_0^\infty h_m(x)e^{-\sigma_1 x}\,dx, := G_1(\sigma_1, \sigma_2), \tag{2.66}$$

with $h_\lambda(x) = S_q(x)\lambda_s(x)$ being the cause-specific hazard rate for recruitment of quiescent cells to the proliferating compartment.

Note that $G_1(0,0) = G(I_s)$, where $G(I_s)$ is given by (2.33).

Now, if there exist non-trivial steady distributions p_s and q_s, we have that $G(I_s) = 1$, so that $\sigma_1, \sigma_2 = 0$ uniquely satisfy $G_1(\sigma_1, \sigma_2) = 1$. Thus, linearisation fails to provide an indication of stability in the case of non-trivial steady distributions existing.

However, if there does not exist I_s satisfying (2.33), we must have that

$$G_1(0,0) = G(I_s) < 0. \tag{2.67}$$

It is clear from the definition that $G_1(\sigma_1, \sigma_2)$ is monotonically decreasing in both σ_1 and σ_2. Thus, there can be no non-negative solutions to $G(\sigma_1, \sigma_2) = 1$ whenever the non-trivial steady state does not exist.

Additionally, G_1 becomes unbounded as both σ_1 and σ_2 tend to $-\infty$, and is continuous, so there exists some pair $\sigma_1, \sigma_2 < 0$ solving $G_1(\sigma_1, \sigma_2) = 1$. Thus, if the non-trivial steady-state does not exist, small perturbations around the trivial steady state decay in time and the trivial steady state is stable.

2.5.3 Stability of the non-trivial steady state

Since the linear stability approach fails for the non-trivial steady-state, due to the eigenvalues having zero real part, we require an alternative approach. One such approach is the so-called relative entropy method, pioneered by DiPerna and Dafermos [22, 25] for conservation laws, which take a similar but not identical form to the equations we are studying, and expanded upon by Perthame for age-structured models [69]. The idea is related strongly to the method of Lyapunov functions commonly used in prov-

ing the stability of equilibria for ODE. Essentially, we construct a functional $H[u]$, with u being the vector of model states, which is zero at the steady state, is everywhere non-negative, and decays along the solutions we start positive and stay positive until we reach the steady state. If such a functional can be found, we prove the stability of the steady state of interest. The relative entropy method often allows for estimates of (commonly upper bounds on) the speed of convergence to the steady state via various enlightening inequalities.

For the model here, we note that the equation for p can be rewritten, when the nontrivial steady state exists, as:

$$\partial_t \left(\frac{p}{p_s} \right) + \partial_x \left(\frac{p}{p_s} \right) = 0, \tag{2.68}$$

which is in the form of a conservation law for p/p_s. This property is possibly useful as it suggests that p is proportional to p_s along the straight-line characteristics, with the constant of proportionality determined by the boundary conditions. We omit the proof, but the basic building blocks are the fact that the hazard rate can be written as $h(x) = -\partial_x(\log S(x))$ and the fact that the steady state is proportional to the survival function S.

We have not managed to complete the proof that the non-trivial steady state is stable when it exists, however the numerical results suggest strongly that this is the case (see figure 2.3). We see that the L^2 norm itself is not a Lyapunov/entropy functional in general (it is not monotonically decreasing in time), but there is a clear periodic downward trend, so that as $t \to \infty, (p, q) \to (p_s, q_s)$.

A commonly chosen relative entropy functional is

$$H_1[u] = \int_0^\infty u \log(u/u_\infty)\, dx, \tag{2.69}$$

which is clearly zero whenever $u = u_\infty$ and in many cases decays suitably in time.

However, we have been unable to make progress with the choice $H[u] = H_1[p] + H_1[q] + I\log(I/I_s) + U\log(U/U_s)$.

We anticipate that the choice of $H[u]$ will explicitly include the survival functions for each compartment, and may possibly have a biological interpretation such as 'the probability that a cell successfully reached the current average division number'. We can see that whenever the non-trivial steady-state exists, under equilibrium conditions, the probability of completing a full cycle, i.e. successfully passing through both the q and p compartments, is precisely $1/2$. This follows reasonably straightforwardly from the identity $P(A \cap B) = P(A|B)P(B)$, with A, B representing the events of successfully passing through each of the compartments.

Often upper bounds on the rate of entropy dissipation are available via a straightforward application of Gronwall's inequality to the equation for the time-derivative of H. Further, inequalities such as the log-Sobolev inequality appear to be useful for converting between entropy measures and the L^2 norm of the solution, leading to estimates on the rate of L^2 convergence.

The time-dependence of the resource I coupled with the nonlinearity of the model and the fact that the system is open, i.e. total 'mass' is not conserved, pose significant challenges. We do however feel that progress is within reach for this problem in the future.

2.6 Numerical Solutions

In this section we develop a numerical scheme for the solution of the system (ASLR), given by equations (2.9) – (2.12). We use the numerical method of lines, which em-

ploys a discretisation of the age domain to reduce the infinite dimensional system to an approximating finite dimensional system of differential algebraic equations (DAE), which may be solved using established ODE solvers (e.g. ODE15S in MATLAB).

Both $p(x,t)$ and $q(x,t)$ evolve in the domain $D = [0,\infty) \times [0,\infty)$. The idea of the method of lines is to solve the problem on a finite, semi-discrete domain $D_h = \{x_0, x_1, \ldots x_n\} \times [0,T)$, the eponymous lines being $\{x_i\} \times [0,T)$.

To establish a method of lines scheme, we need to make three choices how we will discretise the x-domain; how we will approximate x-derivatives; and how we will approximate the integral boundary conditions. For our purposes, it is sufficient (see figures 2.4 & 2.5) to make simple choices for each of these. Of course, more involved methods are available which are able to improve the computational burden:accuracy ratio, however for our purposes (a general examination of the qualitative model behaviour), computational burden is not of great concern.

We choose a uniform discretisation of the x-domain with step size h and approximate x-derivatives with a simple upwind finite difference. The integral boundary conditions are approximated with the trapezium rule.

This leads to the following semi-discretised system of DAEs:

$$Q_0(t) = h\sum_{i=0}^{n-1}(m_i Q_i(t) + m_{i+1}Q_{i+1}(t)), \tag{2.70}$$

$$P_0(t) = \frac{h}{2}\sum_{i=0}^{n-1}(\lambda_i(I(t))P_i(t) + \lambda_{i+1}(I(t))P_{i+1}(t)) + g(I(t),t)U(t), \tag{2.71}$$

$$\frac{dQ_i}{dt} = -\left(\mu_q(I(t)) + \lambda_i(I(t) + \frac{1}{h})\right)Q_i + \frac{Q_{i-1}}{h}, i \geq 1, \tag{2.72}$$

$$\frac{dP_i}{dt} = -(m_i + \mu_p + \frac{1}{h})P_i + \frac{P_{i-1}}{h}, i \geq 1, \tag{2.73}$$

$$\frac{dU}{dt} = -(d_U + g(I(t),t))U(t), \tag{2.74}$$

$$\frac{dI}{dt} = \alpha_I - \delta_I I - f(\tilde{Q},U)I; \;\; \tilde{Q} = \frac{h}{2}\sum_{i=0}^{n-1}(Q_i + Q_{i+1}). \tag{2.75}$$

The algebraic components (2.70) and (2.71) can be differentiated with respect to time, leading to a system of ordinary differential equations which can be solved with standard ODE solvers.

A way to check the validity of our implementation of scheme (2.70) to (2.75) is to compare the results to a special case in which other standard solvers can be used. In the case that $m(x)$ is a delta function (corresponding to a fixed cell cycle time) and λ is independent of age, the full age-structured system can be reduced to a system of delay differential equations via the method of characteristics, as outlined in [62]. Figure 2.4 shows the numerical solution of these equations compared with the solution of the equivalent system of delay differential equations. The solution to the system was computed using MATLABs `dde23` solver. Clearly, the output of the method of lines scheme agrees very well with the `dde23` solution, at least in terms of relative error, suggesting that the scheme is stable and accurate enough for our purposes.

2.6.1 Numerical results

In this section, we present numerical simulations agreeing with the analysis of the previous sections. Table 2.1 provides a summary of the model parameters used in the simulations. In particular, we qualitatively examine the rates of convergence to equilibria in the p and q compartments (figure 2.3). We find that convergence rates appear to be exponentially bounded, which supports the potential use of a relative entropy approach, using which exponential bounds on L^2-convergence can be obtained. Interestingly, after an initial transient period, the L^2-convergence to the trivial equilibrium appears to be exponential, while the convergence to the non-trivial equilibrium is periodic on the log-scale for both the p and q compartments. This agrees with the notion of a purely imaginary eigenvalue of the linearised problem about the non-trivial steady states.

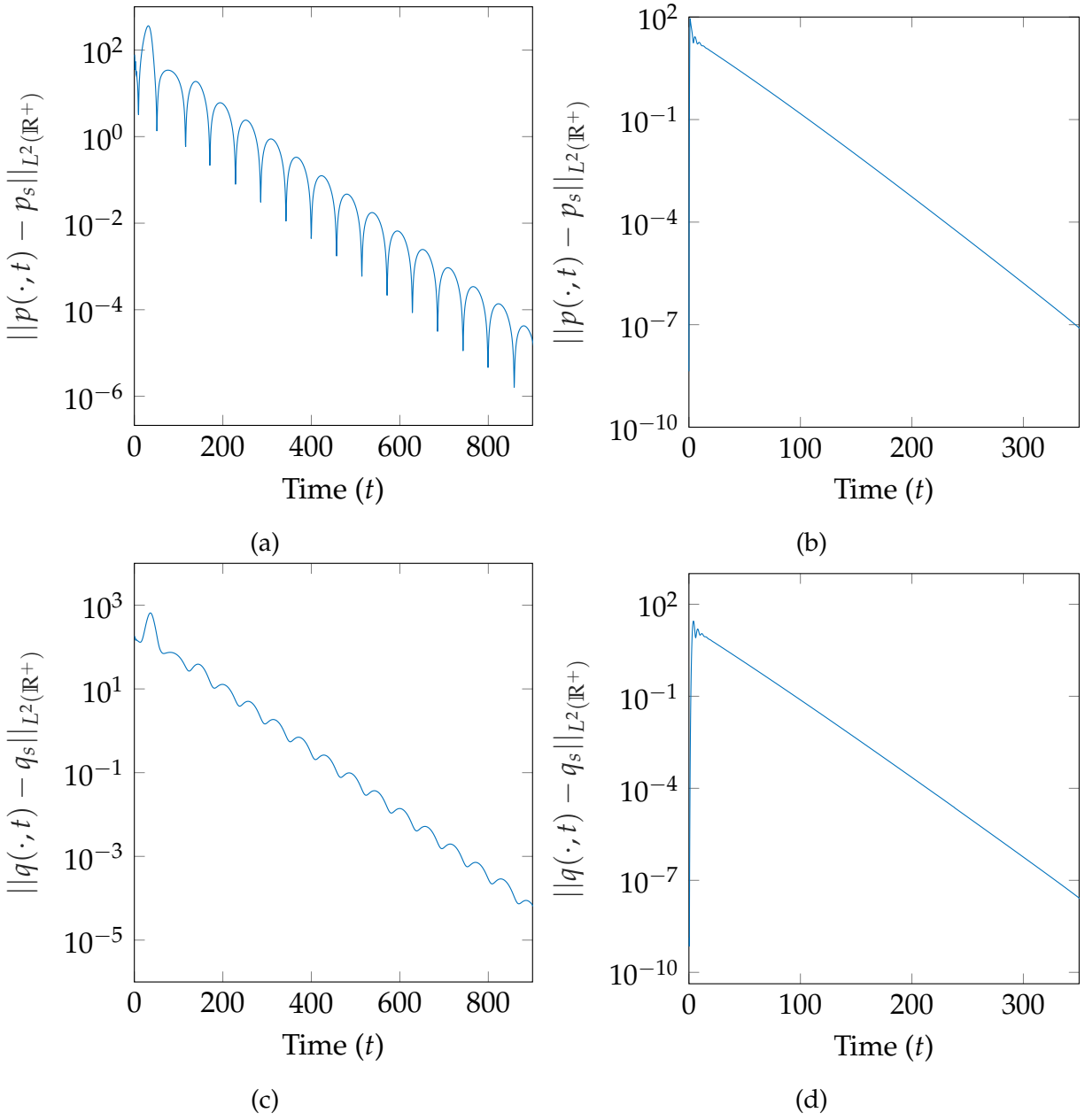
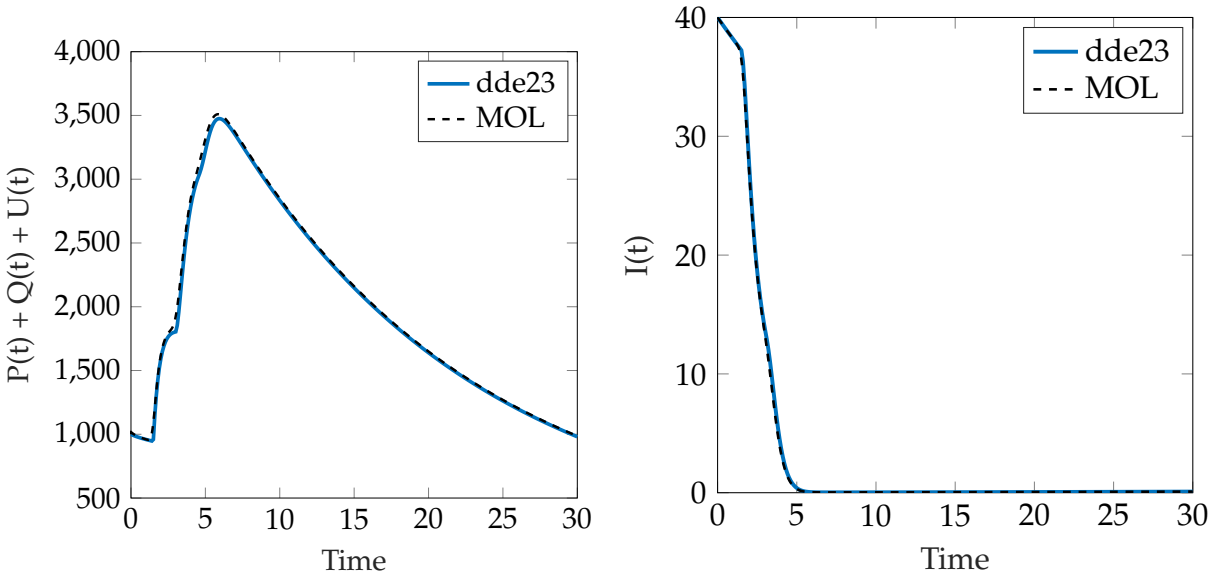


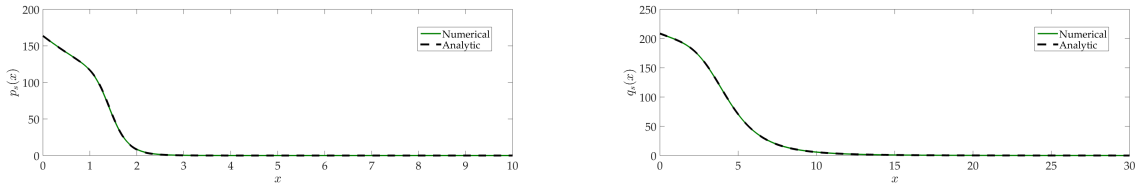
Figure 2.3: Rate of convergence of p and q to their steady distributions in the L^2 norm. Note the log scale on the vertical axis. Parameter sets used: simulation 1 (a and c), simulation 2 (b and d). Refer to table 2.1 for a description of the parameter sets used.

Additionally, we verify that the method of lines scheme agrees with the numerical solution of an equivalent DDE scheme in figure 2.4. There is clearly excellent agreement between the solutions for the total cell population and the cytokine concentration in this model system. Additionally, the system exhibits the expected behaviour – the population consumes the cytokine in order to proliferate, until eventually the resource



(a) (b)

Figure 2.4: Comparison of the numerical solution using the method of lines scheme with an equivalent system solved using `dde23`



(a) (b)

Figure 2.5: Comparison of the numerical steady state with the analytical steady state for the parameters in simulation set 2. Refer to table 2.1 for a description of the parameter sets used.

is completely consumed and the population begins to fall. The delay in the time it takes for cells to complete the cycle leads to an initial decrease in the total population size, followed by a proliferative burst, another less pronounced plateau, and another proliferative burst before the cytokine is depleted.

Figure 2.5 displays a comparison between numerical and analytical steady age distributions. Again, the agreement is excellent, suggesting that the method of lines scheme is suitable for numerically exploring the behaviour of age-structured systems further.

		Simulation set	
Quantity	**Description**	1	2
$U(t)$	Population of cells which have not yet been recruited to the cell cycle	$N_0 = 1000$	$N_0 = 1000$
$q(x,t)$	Cells of age x at time t which have previously divided and are currently quiescent	$q_0(x) = 0$	$q_0(x) = 0$
$p(x,t)$	Cells which are currently cycling at time t and have been doing so for a time x	$p_0(x) = 0$	$p_0(x) = 0$
$I(t)$	Concentration of a representative cytokine resource which drives survival and cell division	$I_0 = 10$	$I_0 = 50$
d_u	Constant death rate of cells in compartment U	0.3	0.3
β_u	Shape parameter of the log-logistic distribution describing recruitment times of cells from U to p	10	10
ρ_u	Parameter controlling the recruitment rate from U to p. The log-logistic distribution is assumed to have median $(\rho_u I)^{-1}$	0.2	0.01
d_q	Maximal death rate of cells in q	0.1	0.1
β_q	Shape parameter of the log-logistic distribution describing recruitment ages of cells from q to p	4	10
ρ_q	Parameter controlling the recruitment rate from q to p. The log-logistic distribution is assumed to have median $(\rho_q I)^{-1}$	0.2	0.01
μ_p	Constant death rate of cells in compartment p	0.3	0.2
β_p	Shape parameter of the log-logistic distribution describing ages at which cycling cells complete cell division	8	20
α_p	Median duration of the cell cycle	1.5	3
c	Rate of consumption of cytokine	10^{-4}	10^{-4}
α_I	Rate of cytokine production	0.1	0.1
δ_I	Rate of cytokine degradation	2×10^{-4}	10^{-4}

Table 2.1: Description of model quantities and the values used for them in the numerical simulations.

2.7 Discussion

In this chapter we have presented and analysed a probabilistic model of lymphocytes proliferating in response to a trophic resource. We have analysed its equilibrium properties both analytically and numerically and derived a simple and biologically intuitive necessary condition for the existence of non-trivial steady cell populations. Similar models have been used to quantify lymphocyte proliferation [27, 35, 53, 106] but have generally focused on simplified models and/or not included resource dynamics explicitly. With the advent of new fluorescence imaging techniques, models with age-structure can be effectively parametrised. The key outcome of such a parametrisation would be the determination of precise distributions for inter-mitotic times, stable age distributions and distributions for time spent in the cell cycle. The latter outcome would be particularly useful for the design of chemotherapeutic treatment strategies for leukaemias. We believe strongly that a movement towards structured population modelling in immunology is essential for making theoretical progress in the field. It is becoming very clear that immune cells cannot be treated as homogeneous populations in the manner of, say, bacteria. Additionally, much of the experimental progress in immunology is focused on individual-level behaviour for example intracellular signalling events & ligand-receptor interactions and structured population models can help to examine the influence of these i-level dynamics at the population level.

The failure of the linear stability approach for the non-trivial distribution is a significant obstacle to further analytic progress, although the numerical results suggest that the problem is amenable to an entropy/Lyapunov functional approach to proving the stability of the non-trivial steady state – a particular advantage of this approach over other methods is the fact that one can obtain not only the desired stability result, but also estimates on the rate of convergence to the steady state in some appropriate norm. This approach has been applied in the context of linear age-structured models, and also in models where a nonlinearity arises due to a density-dependent birth kernel.

Perthame et al. [69] proved an elegant inequality for the basic renewal equation, known as the generalised relative entropy inequality.

We acknowledge that the proof of the existence of a non-trivial steady state is incomplete and only constitutes a necessary, but not sufficient condition. We intend to address this issue in the future. An approach used in ecological contexts is the concept of the basic reproductive number R_0, the expected number of offspring produced by an individual in its lifetime. Proving that the non-trivial equilibrium exists when $R_0 > 1$ has significant precedent in the ecological literature and is likely to be a better approach.

In this chapter we have developed a model of T-cell population dynamics which incorporates resource dynamics explicitly and accounts for population heterogeneity by means of probability distributions describing times until division and recruitment to the cell cycle. We have derived expressions for the steady distributions and made some progress towards stability results. The motivation for the model is the better understanding of T-cell dynamics in the post-transplantation environment of relatively low cell numbers and high homeostatic cytokine concentrations. The primary complication of such allogeneic transplantation procedures is graft-versus-host disease, and in chapter 4 we use clinical T-cell count data, along with clinically relevant patient characteristics, to build a statistical model for GvHD prediction.

In the following chapter, we explore data-driven methods for more general structured population models for single-cell data. The hope is that, in the future, such data-driven methods can be used to effectively parametrise structured population model such as those explored in this chapter.

CHAPTER 3

TIME-RESOLVED POPULATION BALANCE ANALYSIS FOR DATA-DRIVEN APPROXIMATION OF BIOLOGICAL DYNAMICAL SYSTEMS

Single-cell data is now near-ubiquitous in the biological sciences following the introduction of high-throughput technologies such as flow cytometry and RNA sequencing. Such technologies enable the measurement of high-dimensional physical or chemical characteristics of individual cells in large cell populations, leading to potentially large and rich datasets which capture population heterogeneity.

Of particular interest in this chapter is single-cell data collected at numerous time-points in order to gain insight into the dynamical properties of biological processes, i.e. changes in phenotype over time. Our objective in this chapter is to formulate a data-driven method for analysing such time series which is built upon the foundation of a generally applicable underlying model for cellular evolution. We term this method time-resolved population balance analysis (trPBA) after the underlying model and to highlight the fact that the method is designed to deal with time series of single-cell measurements. The work presented in this chapter is a preliminary exploration of the use of equation-free methods for single-cell data and is intended to form the basis of an

extensible framework for tackling general problems in the analysis of single-cell time course data.

3.1 Approaches to learning dynamics from data

3.1.1 Pseudotime

We begin with a review of existing methods for learning about population dynamics from static snapshots. The fundamental idea underlying such methods is that, given a single snapshot of a population *in steady state*, one should be able to learn something about the *dynamical* properties of the population. The guiding principle is that, at equilibrium, the states occupied by the population are equivalent to the states that would be occupied by a single cell evolving in some potential. In other words, we have access to a probability distribution over possible dynamic cellular states. It is then natural to attempt to 'sort' these states according to the order in which they are most likely to be visited, given some initial state. This 'sorting' procedure is referred to as *pseudotime estimation* [95], and has been met with considerable interest from researchers – we outline a number of existing approaches in the following paragraph. Pseudotime estimation is essentially a dimensionality reduction problem in which the objective is to reduce a high-dimensional vector of cellular characteristics to a scalar representing, in some sense, how 'developed' that state is relative to the other possible states.

The pipeline for the majority of pseudotime estimation methods is as follows: perform some form of dimensionality reduction; in the reduced-dimensional space, seek a trajectory through the reduced-dimensional space which optimally connects each cell. For example, three methods which clearly separate these two stages are Embeddr [19], Monocle [95] and Waterfall [84]. Embeddr uses Laplacian eigenmaps [7] for dimen-

sionality reduction and then estimates pseudotime as the arc length along a principal curve through the points in the resulting low-dimensional space [19]. Monocle and Waterfall each use minimum spanning trees to infer trajectories through the reduced dimensional space, while using independent component analysis and principal component analysis for the dimensionality reduction, respectively. As mentioned previously, the majority of pseudotime methods are applicable only to *static* snapshots and do not account for time series data.

There have been a few attempts to incorporate actual capture times into pseudotime estimation procedures in an attempt to more fully capture the dynamics underlying the generation of the single-cell data. Reid & Wernisch [76] use a Gaussian Process Latent Variable Model (GPLVM) to reduce the dimensionality of the data to one. This single dimension is interpreted as pseudotime. Actual capture times are incorporated via the prior over pseudotimes – the prior for the pseudotime associated with each cell is centred on its actual capture time. In contrast with our method (trPBA), there is no overt underlying dynamical model.

A more recent example of a method designed to deal with single-cell time series data is TSEE [3]. The idea is similar – use an appropriate dimensionality reduction algorithm but penalise large deviations from the actual capture time. TSEE incorporates capture times through a penalty on the objective function to be minimised by the elastic embedding [20] dimensionality reduction algorithm. Cells which are far apart in terms of actual capture time are discouraged from being close together in the inferred low-dimensional embedding space. Again, there is no overt link to an underlying dynamical systems model of the biological process.

The two methods most similar in spirit to trPBA are PBA [100] and pseudodynamics [31]. PBA explicitly relates the analysis of single-cell data to the population balance equation [75], a general modelling framework for single-cell data, but is only able to accept single snapshots of steady-state populations as input. Pseudodynamics [31] also

uses a population balance framework for analysing single-cell data, first estimating pseudotimes associated with each cell and inferring the functional forms governing pseudotime dynamics according to the population balance model.

3.1.2 Dynamic mode Decomposition and related methods

Dynamic mode decomposition [80] is a method for extracting information about dynamical systems from data generated by the dynamical system, either through simulation or experiment. We will first outline the basic DMD algorithm and then describe some of its extensions and variants. The key idea from DMD that will carry over to the remaining work in this chapter is the use of data-derived basis functions for formulating approximate Galerkin methods for the solution of PDEs.

Consider a general (discrete-time) dynamical system $\boldsymbol{x}_{t+1} = \boldsymbol{F}(\boldsymbol{x}_t)$. Let

$$X = \begin{pmatrix} | & | & & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_T \\ | & | & & | \end{pmatrix} \tag{3.1}$$

be a matrix formed by snapshots of the solution to the dynamical system at T distinct time points.

DMD begins by taking a linear approximation to $\boldsymbol{F}$, i.e. we approximate $\boldsymbol{F}(\boldsymbol{x})$ by $A\boldsymbol{x}$ for some appropriately sized square matrix A. Under this approximation, the dynamical system becomes $\boldsymbol{x}_{t+1} \approx A\boldsymbol{x}_t$. Alternatively, we can write

$$X_{\cdot,-1} \approx AX_{\cdot,-T},$$

where $X_{\cdot,-j}$ is the matrix X with the j^{th} column removed.

We would like to estimate the matrix A from the data to obtain a linear approximation to the dynamical system. The matrix A minimising the Frobenius norm $||X_{\cdot,-1} - AX_{\cdot,-T}||_F$ is given by

$$A = X_{\cdot,-1}(X_{\cdot,-T})^{+},$$

where $^{+}$ indicates the pseudo-inverse.

Armed with a linear approximation to $\boldsymbol{F}$, the dynamics can be explored by, for example, considering the eigenvalues and eigenvectors of A. Theoretical work [96] has established the link between the eigenelements of A (as computed by DMD) and the Koopman operator for dynamical systems. Extended DMD [103] provides a comprehensive means of data-driven approximation to nonlinear (stochastic) dynamical systems. The most relevant aspect of extended DMD to our work is its interpretation as an approximate data-driven Galerkin scheme for the solution of Fokker-Planck equations. While our approach deviates from extended DMD in terms of the details of its implementation, the underlying idea is similar – attempt to learn basis functions *from the data* in order to represent the solution to a PDE which could have feasibly generated the data.

In the context of stochastic differential equations, the Koopman operator is equivalent to the operator in the backward Kolmogorov equation derived from the SDE. For concreteness, suppose we generate M time series by sampling M trajectories from the multivariate SDE

$$\mathrm{d}\boldsymbol{x}_\mathrm{t} = \boldsymbol{A}(\boldsymbol{x}_\mathrm{t})\mathrm{dt} + \mathrm{B}\mathrm{d}\boldsymbol{W}_\mathrm{t}. \tag{3.2}$$

The corresponding Fokker-Planck equation is

$$\begin{aligned}\partial_t p(\boldsymbol{x}, t) &= -\nabla(\boldsymbol{A}(\boldsymbol{x})p(\boldsymbol{x}, t)) + \frac{1}{2}B\nabla^2 p(\boldsymbol{x}, t), \\ &= \mathcal{P}p,\end{aligned}$$

and the Koopman operator is the adjoint of $\mathcal{P}$, i.e.

$$\mathcal{K}p = \boldsymbol{A}(\boldsymbol{x})\nabla p + \frac{1}{2}B\nabla^2 p(\boldsymbol{x}, t).$$

Given the sampled trajectories and a suitably chosen set of observables (see [103]), extended DMD computes an approximation to the eigenpairs of the Koopman operator via matrix decomposition in a similar way to that described for standard DMD above. These approximate eigenpairs can be used for various downstream tasks such as forecasting via the expansion

$$p(\boldsymbol{x}, t) \approx \sum_{j=1}^{K} \exp(\lambda_j t)\varphi_j(\boldsymbol{x})\boldsymbol{v}_k,$$

where the λ_j, $\varphi_j(\boldsymbol{x})$ and $\boldsymbol{v}_k$ are the Koopman eigenvalues, eigenfunctions and modes, respectively. Each of these quantities is estimated directly by extended DMD from the data. The interpretation of extended DMD as an approximate Galerkin scheme comes from treating the data-derived eigenfunctions φ_j as the basis functions for the scheme. The details of the proof that extended DMD converges to a Galerkin scheme for the underlying model can be found in [103].

Given that SDEs are commonly used to model intracellular compositions [102], extended DMD appears to be a promising way to extract dynamical information from single-cell time series. In a manner similar to the diffusion maps algorithm [66] for dimensionality reduction (which we discuss below), the inferred eigenfunctions and eigenvalues could be fruitfully used for dimensionality reduction/visualisation of relatively high-dimensional single-cell expression data. However, the barrier to the direct application of extended DMD to the single-cell data we consider in this chapter (and the most common form of single-cell timecourse data) is that our samples are not from continuous trajectories through time. In other words, because single-cell technologies such as flow cytometry and RNA-seq are destructive, we are unable to follow indi-

vidual cells *through time*, instead relying on aggregate data. This motivates alternative approaches to inferring the eigenfunctions, eigenvalues and modes of the Koopman operator which are able to deal with aggregate time-series. However, before we outline the trPBA method, we will briefly touch upon the diffusion maps algorithm for dimensionality reduction, which is also based on numerical approximations to eigenpairs of Fokker-Planck operators and provides inspiration for the dimensionality reduction aspect of the trPBA method.

The diffusion maps algorithm [66] is used for nonlinear dimensionality reduction and is another means of approximating the eigenfunctions and eigenvalues of backward Kolmogorov operators or Koopman operators. While diffusion maps can be interpreted as a general tool for nonlinear dimensionality reduction, and may be used to reduce the dimensionality of time-varying data [31], the link to the Fokker-Planck equation relies on the assumption that the data are sampled from the stationary distribution of the corresponding stochastic process. This means that the diffusion maps algorithm cannot be applied directly to data sampled from a dynamic process while retaining the connection to an underlying governing equation – the method developed in this chapter aims to address this point. Nevertheless, it is instructive to outline the algorithm as many of the fundamental ideas are similar to our trPBA method. Also diffusion maps form the basis for the PBA method [100], on which trPBA builds and which will be outlined below.

Diffusion maps are constructed from the eigenpairs of a discrete approximation to the Koopman operator (or backward Kolmogorov operator), which in turn is constructed from data assumed to be *sampled from the stationary distribution* of some SDE. The way in which the discrete approximation to the Koopman operator is constructed is different to that of extended DMD, however, being based on a random walk on a weighted graph with the sampled datapoints as nodes. The details of the method for obtaining the approximate eigenpairs are not especially relevant to the trPBA method, however

their use for dimensionality reduction is relevant. Given approximations to the first m eigenvalues and eigenfunctions $\{\lambda_i, \psi_i\}_{i=1}^m$ of the Koopman operator, a diffusion map is given by

$$\Psi_t(\boldsymbol{x}) = (\lambda_1^t \psi_1(\boldsymbol{x}), \lambda_2^t \psi_2(\boldsymbol{x}), \ldots, \lambda_m^t \psi_m(\boldsymbol{x})).$$

The diffusion coordinates associated with a given datapoint are natural dimensionality reduction coordinates because they encode the dynamics of the underlying SDE at different timescales. For large times, the first few diffusion coordinates dominate the solution. t is typically treated as a tunable parameter of the diffusion map. We will carry these ideas over to our dimensionality reduction applications with slight modifications.

3.2 PDE models for heterogeneous evolving cell populations: population balances

Population balance models [75] are used to describe the dynamics of heterogeneous, possibly growing, populations. They are similar to the Fokker-Planck equation in that they provide a population-level description of stochastic single-entity dynamics, however they are more general in that they allow the size of the population to change over time. A general population balance equation is given by:

$$\partial_t N(\boldsymbol{x}, t) = -\nabla \cdot (\mu(\boldsymbol{x}) N(\boldsymbol{x}, t)) + \nabla^2 (D(\boldsymbol{x}) N(\boldsymbol{x}, t)) + r(\boldsymbol{x}) N(\boldsymbol{x}, t).$$

Here, $\boldsymbol{x}$ is the cellular state – describing intracellular protein concentrations for example – and t is time. $N(\boldsymbol{x}, t)$ is the state distribution function, such that $N(\boldsymbol{x}, t)$ is the number of cells at time t with state lying in an infinitesimal volume around $\boldsymbol{x}$. $\mu(\boldsymbol{x})$ describes the evolution of the state within a single cell in the absence of division, it is analogous to the function $A(\boldsymbol{x})$ in the SDE (3.2). $r(\boldsymbol{x})$ is a state-dependent division rate. The diffusion

function $D(\boldsymbol{x})$ describes the level of stochasticity in the single-cell dynamics of the state $\boldsymbol{x}$. In this form the population balance model is essentially a modified Fokker-Planck equation with additional terms describing the macroscale population dynamics.

Solutions to (3.2) can be expressed in terms of eigenfunctions of the operator $\mathcal{L}$ defined by $\mathcal{L}N(\boldsymbol{x},t) = -\nabla\cdot(\mu(\boldsymbol{x})N(\boldsymbol{x},t)) + \nabla^2(D(\boldsymbol{x})N(\boldsymbol{x},t)) + r(\boldsymbol{x})N(\boldsymbol{x},t)$. Note that $\mathcal{L}$ is self-adjoint and so has purely real eigenvalues.

We define such eigenfunctions as $\phi_i(\boldsymbol{x})$ satisfying $\lambda_i\phi(\boldsymbol{x}) = \mathcal{L}\phi_i(\boldsymbol{x})$ subject to $\phi_i(\boldsymbol{x}) \geq 0$ and $\int_{\mathcal{X}}\phi_i(\boldsymbol{x}) = 1$, i.e. we impose the constraint that the eigenfunctions are probability density functions. For further exploration of such eigenproblems, refer to [69]. By the principle of superposition, solutions $N(\boldsymbol{x},t)$ to (3.2) can be written

$$N(\boldsymbol{x},t) = \sum_{n=1}^{\infty} a_i \exp(\lambda_i t)\phi_i(\boldsymbol{x}). \tag{3.3}$$

Our objective will be to learn approximations of the first few eigenfunctions from aggregate data, which we assume has been generated by a general stochastic process which can be described by a population balance equation of the form (3.2). Further, we intend to do so in a computationally efficient manner and explore the ways in which such representations can be utilised in practice.

Before we outline the trPBA method, we will summarise two other approaches to learning population balance models directly from data.

3.3 Description of PBA and pseudodynamics

The goal of population balance analysis [100] is to estimate the parameters of a simplified population balance model with diffusion from a single, static snapshot of single-

cell data. The population balance equation they consider is of the form

$$\partial_t N = \nabla \cdot (N \nabla V) + \frac{1}{2} D \nabla^2 N + RN, \tag{3.4}$$

which corresponds to stochastic intracellular dynamics driven by a potential V and state-independent growth rate R with no partitioning of material between daughter cells.

The authors draw on graph theory to estimate the eigenfunctions of (3.4) in a similar way to diffusion maps. However, as with diffusion maps, their procedure is only applicable to static snapshots sampled from the stationary distribution of (3.4), which is the Gibbs distribution associated with the potential V. The authors use their data-derived model to extract meaningful, theoretically grounded pseudotimes and fate probabilities for individual cells.

Pseudodynamics [31] is another method designed to learn population balance models directly from data, and which *is* designed to deal with timecourse data. Pseudodynamics first reduces the dimensionality of the data to one, via some form of pseudotemporal ordering. The authors use diffusion pseudotime [47], but in principle any pseudotime estimation method could be used.

Once each cell has been characterised by a pseudotime (i.e. developmental status), a population balance model is estimated, taking the pseudotime as a summary of the cell state. More formally, denote pseudotime by τ and actual time by t. Then the pseudodynamics model is given by

$$\partial_t N(\tau, t) = \partial_\tau (D(\tau) \partial_\tau N(\tau, t)) - \partial_\tau (v(\tau, t) N(\tau, t)) + g(\tau, t). \tag{3.5}$$

This is a more general version of the PBA model, allowing state-dependent diffusion, both state- and time-dependence in the deterministic part of the intracellular dynamics,

and state- and time-dependence in the population dynamics.

The authors represent the functions D, v and g with splines and estimate them using maximum likelihood. This is a computationally intensive procedure.

Note that the pseudotemporal ordering takes place prior to the estimation of (3.5) in pseudodynamics.

Below we develop a method which is able to deal with multiple time-points, is theoretically grounded in population balance models, is relatively computationally efficient, and does not require prior dimensionality reduction.

3.4 Developing the trPBA method

As with pseudodynamics and PBA, we use a population balance equation as the foundation. As with (extended) DMD we aim to obtain basis functions directly from the data, in terms of which we can represent the solution to the underlying population balance equation. The basic steps of our trPBA method are as follows:

1. Estimate the probability density over cell states at each available time-point
2. Treat these estimated densities as realisations of the solution to a population balance equation
3. By identifying this solution with an approximate eigenexpansion, extract approximate eigenfunctions and eigenvalues of the underlying population balance operator
4. Use the learned eigenfunctions and eigenvalues for downstream tasks such as dimensionality reduction, pseudotime estimation and back-/forecasting.

We will now walk through each stage of the algorithm, discussing design choices and the details of the implementation. In what follows, we denote the measurement associated with the i^{th} cell at the t^{th} timepoint by $\boldsymbol{x}_{i,t} \in \mathbb{R}^d$. Let T denote the total number of timepoints and let N_t denote the total number of cells measured at the t^{th} timepoint.

3.4.1 Density Estimation

For the density estimation phase, we focus on mixture models. Mixture models estimate densities through weighted sums of simpler density functions, most commonly Gaussians. In particular, we focus on kernel density estimates and Dirichlet Process Gaussian Mixture models.

A (multivariate) kernel density estimate is constructed from a set of points $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ via

$$\hat{p}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} K_H(\boldsymbol{y} - \boldsymbol{y}_i).$$

K_H is known as the kernel and is a non-negative function integrating to 1 and depending on a bandwidth matrix H. Whenever we use a kernel density estimate in the sequel, we choose a Gaussian kernel, i.e. we set

$$K_H(\boldsymbol{y}) = \det(H)^{-1/2}\Phi(H^{-1/2}\boldsymbol{y}),$$

where Φ is the standard multivariate normal density. For simplicity and computational speed, we choose a diagonal bandwidth matrix and set its entries by Silverman's rule of thumb with an additional user-specified smoothness parameter, α:

$$H_{ii} = \alpha\left(\frac{4}{n(d+2)}\right)^{\frac{1}{d+4}}\hat{\sigma}_i$$

where $\hat{\sigma}_i$ is the sample standard deviation along the i^{th} dimension of Y. We add the

smoothness parameter to overcome the fact that the Silverman kernel can often be too diffuse to capture clusters corresponding to individual time points because the standard deviations are taken over the full dataset.

We also experimented with Dirichlet Process Gaussian Mixture Models (DPGMM; [29]) for the density estimation phase. The DPGMM is a Bayesian non-parametric tool for density estimation which does not require the number of mixture components to be specified in advance. Instead, the number of mixture components is estimated jointly with the component parameters. Concretely, the Dirichlet Process Gaussian Mixture Model models the probability distribution generating the observed data Y as the following infinite mixture of Gaussians

$$\hat{p}(\boldsymbol{y}) = \sum_{k=1}^{\infty} w_k \mathcal{N}(\boldsymbol{y}|\mu_k, \Sigma_k).$$

The weights w_k are given a Dirichlet Process prior, an infinite-dimensional generalisation of the Dirichlet distribution over vectors in the unit simplex, i.e. vectors which sum to one. Typically, the component parameters μ_k and Σ_k are given conjugate Normal-Inverse Wishart priors. Note that in the DPGMM, we do not restrict the covariance matrices to be diagonal. At first glance, inference may appear intractable due to the fact that the weight vector has infinitely many components. However in practice, the Dirichlet process ensures that (almost surely) only finitely many of these weights are non-zero. Various characterisations of the Dirichlet Process have been devised to enable tractable inference, such as the stick-breaking process [81]. We omit the details of these characterisations.

As with any Bayesian model, there are a few options when it comes to estimating the model. Sampling-based methods have been developed for DPGMMS [56] but they are computationally intensive and the benefits of full characterisation of uncertainty and theoretical exactness are secondary to our main objective of obtaining a reasonable

density estimate in this case. Alternatively, approximate inference methods have also been developed for DPGMMs [14,58] which carry significantly reduced computational cost. This reduction in computational cost is due to two aspects of the methodology: firstly, the use of a variational approximation to the posterior distribution transforms the problem to one of *optimising* the parameters of the approximate posterior distribution rather than repeatedly sampling from the exact posterior distribution; secondly, the method of [58] leverages the computational advantages of storing data in a kd-tree [33], essentially partitioning the data space to reduce the number of repeated computations for nearby points.

This idea of using kd-trees to speed up the density estimation phase can also be applied to kernel density estimation. Additionally, data condensation methods such as MDC [64] can be applied to reduce the computational cost of constructing a kernel density estimate. We outline both of these approaches and how we use them below.

kd-trees

kd-trees [33] are hierarchical structures for storing data and allow fast retrieval of nearest neighbours. kd-trees work by iteratively partitioning the multidimensional data space into hyperrectangles by splitting along a single dimension at a time. This partitioning proceeds until each data point has been separated from all other data points. Obtaining a list of points within a specified distance of a given query point is then straightforward because the minimum distance between a point in some partition and the query point can be computed easily. This means that entire chunks of the data can potentially be excluded from the search with a single distance computation. Beginning from the top of the tree, we iteratively exclude fewer and fewer points until we are left with a set of points within the required distance of the query point.

This procedure is useful for approximate kernel density estimation because points which only negligibly contribute to the density estimate can be screened prior to eval-

uating the kernel. One such approach to approximate kernel density estimation using kd-trees is as follows [44]. Suppose $X \in \mathbb{R}^{n \times d}$ is a matrix containing d dimensional samples as its rows. To compute an approximate kernel density estimate of the underlying distribution with kernel $K(x_i, x_j) = K(||x_i - x_j||)$, one can implement the following algorithm:

- Construct a kd-tree from the data X, i.e. partition the data into 2^j groups, $\{G_{j,k}\}_{k=1}^{2^j}$ for each level of the tree $j = 1, \ldots, j_{\max}$.
- For each point x_i, traverse the tree from the top (i.e. loop through j beginning at $j = 1$), computing the distance between x_i and the boundary points of $G_{j,k}$ for each j, k.
- For a given (j, k) pair, let $d_{i,j,k}$ denote the minimum distance between x_i and the boundary points of $G_{j,k}$.
- If $K(d_{i,j,k}) < \epsilon$ for some threshold ϵ, exclude all points contained within $G_{j,k}$ from the kernel computation at x_i.

We use the function `rangesearch` from the MATLAB Statistics & Machine Learning toolbox to perform the pruning and we set the threshold ϵ to be machine precision by default, as suggested in the original paper.

Multiscale data condensation

A similar method for reducing the computational burden of kernel density estimation is data condensation. The idea behind data condensation is to find a relatively small set of representative points as a preprocessing step prior to kernel density estimation. Perhaps the simplest version of data condensation is subsampling, i.e. randomly sampling a subset of the data to use for subsequent tasks. While the kd-tree based approach outlined above places a kernel at each point and uses a subset of the points to evaluate the kernel, data condensation based approaches place a kernel at only a subset of

the points and use the remaining points to evaluate the kernel. Essentially, the idea behind the kd-tree based approach is that distant points contribute little to the density estimate so can be ignored, while the idea behind data condensation approaches is that nearby points contribute similarly to the overall density estimate, so only one of them needs to be considered.

One general purpose data condensation algorithm is Multiscale Data Condensation [64]. The algorithm works by iteratively removing points in high-density regions of the data space and replacing them with a single representative from the original dataset. The single tunable hyperparameter in MDC is a number of nearest neighbours k. The algorithm is implemented as follows

1. Let $\mathcal{D}$ denote the active set and initialise $\mathcal{D} = \{x_i\}_{i=1}^n$.
2. For each $x_i \in \mathcal{D}$, compute $r_k(x_i)$, the distance to the k^{th} nearest neighbour in $\mathcal{D}$.
3. For the x_i with the smallest associated $r_k(x_i)$, remove all points within a ball of radius $2r_k(x_i)$ of x_i from the active set $\mathcal{D}$.
4. Repeat steps 2. and 3. until the active set is empty.

This simple procedure produces a dataset of reduced size whose density approximates the density of the original data, i.e. points in dense regions are selected more frequently than points in sparsely populated regions. This is essentially density-dependent subsampling, whereby the probability of a point being selected as a representative is dependent on its local density.

kd-trees can be used to implement the distance computations in MDC and the output can subsequently be fed into the kd-tree based approximate kernel density estimation procedure described above.

In the original paper [64], the authors propose a custom bandwidth estimator for sub-

sequent kernel density estimation using the reduced dataset. Quite effective kernel density estimates can be obtained by deploying some form of data reduction followed by using variable bandwidth kernels centred at the elements of the reduced set. Such an algorithm for fast kernel density estimation is presented in [99]. We use a simple spherical variable bandwidth, equal to the distance to the k^{th} nearest neighbour in the reduced dataset. This idea of using nearest neighbour distances as kernel bandwidths can be traced to [16] and has more recently found applications in self-tuning spectral clustering [105].

Regardless of which of these methods, or combination of methods, we use to perform the density estimation, the outcome is an estimate of the full density of the form

$$\hat{p}(\boldsymbol{x}) = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{x}|\mu_i, \Sigma_i).$$

The density at each experimental timepoint, t_j, can then be expressed in terms of the shared mixture components $\mathcal{N}(\boldsymbol{x}|\mu_i, \Sigma_i)$ as

$$p(\boldsymbol{x}, t_j) \approx \hat{p}_j(\boldsymbol{x}) = \sum_{i=1}^{k} w_{i,j} \mathcal{N}(\boldsymbol{x}|\mu_i, \Sigma_i),$$

where the timepoint-specific weights $w_{i,j}$ are estimated by

$$w_{i,j} = \frac{1}{N_j} \sum_{k=1}^{N_j} \frac{\mathcal{N}(\boldsymbol{x}_{j,k}|\mu_i, \Sigma_i)}{\sum_{l=1}^{k} \mathcal{N}(\boldsymbol{x}_{j,k}|\mu_l, \Sigma_l)},$$

which is the empirical probability that a randomly selected point from timepoint j belongs to component i. Here $\boldsymbol{x}_{j,k}$ is the k^{th} point sampled at timepoint j.

We now discuss how this mixture model with time-varying weights can be identified with a Galerkin scheme for the solution of the population balance model in order to gain insight into the dynamics responsible for generating the data.

3.4.2 Identification with a Galerkin Scheme for the PBE

Recall the population balance equation (3.2)

$$\begin{aligned}\partial_t N(\boldsymbol{x},t) &= -\nabla\cdot(\mu(\boldsymbol{x})N(\boldsymbol{x},t)) + \nabla^2(D(\boldsymbol{x})N(\boldsymbol{x},t)) + r(\boldsymbol{x})N(\boldsymbol{x},t),\\ &\coloneqq \mathcal{L}[N(\boldsymbol{x},t)].\end{aligned}$$

One means of approximating the solution to (3.2) a Galerkin scheme. The aim of Galerkin schemes is to project the solution $N(\boldsymbol{x},t)$ onto a finite dimensional function space spanned by a set of predefined functions. More concretely, suppose the solution $N(\boldsymbol{x},t)$ to (3.2) lies in some (infinite dimensional) function space V, in this case the space of non-negative functions supported on $\mathbb{R}^d \times \mathbb{R}^+$. Galerkin schemes find approximate solutions in some m-dimensional function space V_f, spanned by a set of trial functions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$, such that the approximate solution at time t is the projection of the true solution at time t onto V_f. In practice, this amounts to choosing a set of trial functions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$ and solving a linear system of ODEs to obtain the projected solutions.

Given the trial functions $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$, we seek approximate solutions to (3.2) of the form

$$\tilde{N}(\boldsymbol{x},t) = \sum_{i=1}^m a_i(t)\varphi_i(\boldsymbol{x}).$$

Substituting this into (3.2) yields

$$\sum_{i=1}^m \dot{a}_i\varphi_i(\boldsymbol{x}) = \sum_{i=1}^m a_i(t)\mathcal{L}[\varphi_i(\boldsymbol{x})].$$

Taking the inner product with respect to each of the trial functions yields the following

linear system of ODEs for the time-varying coefficients $a_i(t)$

$$\dot{\boldsymbol{a}} = M^{-1}L\boldsymbol{a}, \tag{3.6}$$

where $M, L \in \mathbb{R}^{m\times m}$ and $M_{ij} = \langle \varphi_i, \varphi_j \rangle_{L^2(\mathbb{R}^d)}$, $L_{ij} = \langle \mathcal{L}\varphi_i, \varphi_j \rangle_{L^2(\mathbb{R}^d)}$.

Rather than predefining the trial functions, we use the estimated mixture components, i.e. we set $\varphi_i(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mu_i, \Sigma_i)$. This essentially leaves us with the problem of estimating the matrix $M^{-1}L$ from the data-derived time-varying mixture weights $w_{i,j}$ in (3.4.1), in turn obtaining a discrete approximation to the operator $\mathcal{L}$.

We now turn our attention to estimating the matrix $M^{-1}L$ in the case where the data are normalised by overall population size, i.e. we have access only to the probability distribution over cell states but not to the total number of cells.

The normalised density over cell states is given by

$$\begin{aligned} n(\boldsymbol{x}, t) &= \frac{N(\boldsymbol{x}, t)}{\int_{\mathcal{X}} N(\boldsymbol{x}, t)}, \\ &\approx \frac{\tilde{N}(\boldsymbol{x}, t)}{\int_{\mathcal{X}} \tilde{N}(\boldsymbol{x}, t)} \\ &= \frac{\sum_{i=1}^m a_i(t)\varphi_i(\boldsymbol{x})}{\sum_{j=1}^m a_j(t)}. \end{aligned}$$

Let

$$g_i(t) = \frac{a_i(t)}{\sum_{j=1}^m a_j(t)}$$

denote the normalised weights and let $\tilde{n}(\boldsymbol{x}, t) = \sum_{i=1}^m g_i(t)\varphi_i(\boldsymbol{x})$ be the approximate probability density function at time t.

Now, by (3.6), the weights $a_i(t)$ satisfy

$$\boldsymbol{a}(t) = \exp((M^{-1}L)t)\boldsymbol{a}_0,$$

where exp is the matrix exponential.

Writing $\exp((M^{-1}L)t)$ in terms of its spectral decomposition, we have

$$\boldsymbol{a}(t) = V \exp(\Lambda t) V^{-1} \boldsymbol{a}_0.$$

Consequently, denoting $V^{-1}\boldsymbol{a}_0$ by $\boldsymbol{b}$,

$$\begin{aligned}
\tilde{N}(\boldsymbol{x}, t) &= \sum_{i=1}^{m} a_i(t)\varphi_i(\boldsymbol{x}), \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{m} v_{ij} \exp(\lambda_j t) b_j \right) \varphi_i(\boldsymbol{x}), \\
&= \sum_{j=1}^{m} b_j \exp(\lambda_j t) \left(\sum_{i=1}^{m} v_{ij} \varphi_i(\boldsymbol{x}) \right), \\
&= \sum_{j=1}^{m} b_j \exp(\lambda_j t) \psi_j(\boldsymbol{x}).
\end{aligned}$$

We have taken the eigenvectors constituting the columns of V to be normalised such that they sum to unity. The modified components $\psi_j(\boldsymbol{x}) := \sum_{i=1}^{m} v_{ij}\varphi_i(\boldsymbol{x})$ each evolve exponentially with rate λ_j. Note that this takes the form of a truncated eigenexpansion of the solution $N(\boldsymbol{x}, t)$ to the full population balance equation.

The normalised weights associated with the modified components are then given by

$$g_i^*(t) = \frac{b_i \exp(\lambda_i t)}{\sum_{j=1}^{m} b_j \exp(\lambda_j t)}. \tag{3.7}$$

Let W be the matrix containing the empirical weights associated with the t^{th} timepoint as its columns, i.e. the matrix with entries w_{ij} as defined in (3.4.1). Let $\tilde{W}$ be the matrix containing the corresponding weights associated with the modified components ψ_j. Then

$$W = V\tilde{W}.$$

Thus, estimating the matrix V, and therefore the approximate eigenfunctions of the population balance operator, becomes a matrix factorisation problem in a similar vein to extended DMD or proper orthogonal decomposition. In the following subsection we discuss possible approaches for computing a suitable factorisation.

3.4.3 Estimating the Eigenfunctions

We now turn our attention to defining a suitably constrained matrix factorisation of the empirical weight matrix W. We have the following desiderata for such a factorisation:

1. The columns of V and $\tilde{W}$ should each sum to unity
2. The columns of V should be orthogonal
3. The elements of $\tilde{W}$ should be non-negative
4. The column of V corresponding to the largest eigenvalue should be non-negative since its weights determine the stationary distribution

There are a number of existing matrix factorisation algorithms which satisfy at least some of these desiderata and which we outline below. They all share the common objective of minimising some matrix norm $||W - V\tilde{W}||$ subject to either hard or soft constraints on the matrices $V, \tilde{W}$.

Perhaps the most commonly used matrix factorisation is the singular value decomposition, i.e. a decomposition of the matrix $W = U_1 \Sigma U_2^\top$ such that U_1, U_2 are orthogonal matrices and Σ is a diagonal matrix with non-negative real entries. One might then take $V = U_1 Z^{-1}$, $\tilde{W} = Z\Sigma U_2^\top$, where Z is the diagonal matrix whose $(i,i)^{th}$ entry is the sum of the i^{th} column of U_1. Such a decomposition satisfies desiderata 1 and 2 but fails to guarantee, or even encourage, desiderata 3 and 4.

A popular alternative factorisation is non-negative matrix factorisation [59]. The objective of NMF is to, as the name suggests, find matrices with non-negative entries whose product is as close as possible in some matrix norm to the matrix being decomposed, i.e. for a general non-negative matrix X, find non-negative matrices F, G such that $X \approx FG^\top$. Such an approach satisfies desiderata 3 and 4, but fails to guarantee or encourage desiderata 1 and 2.

A number of modifications to NMF have been proposed which impose hard or soft orthogonality constraints on F, G, or both (orthogonal NMF, [23]) or relax the non-negativity constraint on F or G (semi-NMF, [24]). F-orthogonal NMF attempts to satisfy all 4 of our desiderata and we describe it in more detail below.

Given a non-negative matrix X, F-orthogonal NMF seeks to solve the following constrained optimisation problem

$$\text{Find} \underset{F,G\geq 0}{\operatorname{argmin}} ||X - FG^\top||_F^2 \text{ subject to } F^\top F = I. \tag{3.8}$$

Note that the constraint imposes not just orthogonality on F but orthonormality. Given such a factorisation, the sum-to-unity constraint can easily be imposed by rescaling the columns of the matrices F, G.

Problem (3.8) can be solved fairly straightforwardly using the iterative update scheme in Algorithm 1.

Algorithm 1 F-orthogonal NMF as per [23]

Input: non-negative matrix $X \in \mathbb{R}^{n\times m}$, rank k
Initialise $F \in \mathbb{R}^{n\times k}$, $G \in \mathbb{R}^{m\times k}$
while not converged **do**

$$G_{ij} \to G_{ij}\frac{(X^\top F)_{ij}}{(GF^\top F)_{ij}}$$

$$F_{ij} \to F_{ij}\left(\frac{(XG)_{ij}}{(FF^\top XG)_{ij}}\right)^{1/2}$$

Output: non-negative matrices F, G

Interestingly, F-orthogonal NMF is theoretically equivalent to k-means clustering on the rows of X. In this case, that corresponds to clustering the mixture components according to the similarity between their weights over time so that components in a cluster are similarly active at similar timepoints.

As long as the initialisation of F is orthonormal, it remains orthonormal at every iteration. We follow the original paper and initialise F using k-means clustering.

Using orthogonal NMF, we obtain a rank k approximation to the matrix of weights $W \approx V^*\tilde{W}^*$, $V^* \in \mathbb{R}^{m\times k}$, $\tilde{W}^* \in \mathbb{R}^{k\times T}$. k corresponds to the dimensionality of the truncated eigenexpansion of the approximate solution to the population balance model. We then normalise the columns of V^* to sum to unity to obtain the final decomposition $W \approx V\tilde{W}$ by forming the diagonal matrix $[D]_{ii} = \sum_j v_{ij}$ and setting $V = V^*D^{-1}$, $\tilde{W} = D\tilde{W}^*$.

The approximate eigenfunctions are then given by

$$\psi_j(\boldsymbol{x}) = \sum_{i=1}^{m} v_{ij}\varphi_i(\boldsymbol{x}),\ j = 1, \ldots, k.$$

3.4.4 Downstream Tasks

We are now in a position to discuss some of the potential applications of trPBA. How can we use these data-derived eigenfunctions to learn about the underlying processes and better understand the biological mechanisms generating the data? In this subsection we outline three common tasks which can be accomplished with the aid of trPBA.

Forecasting

By forecasting, we mean estimating the expression profile, i.e. the probability density over cell states, at unobserved time points. Armed with the approximate eigenfunc-

tions $\psi_j(\boldsymbol{x})$, we use equation (3.7) to estimate the temporal dynamics associated with each $\psi_j(\boldsymbol{x})$. The empirical weights associated with each $\psi_j(\boldsymbol{x})$ are given by the columns of $\tilde{W}$. Thus, by estimating the parameters b_i, λ_i of (3.7) by minimising the distance between the empirical weights and the theoretical weights, we fully specify an approximate solution to the population balance model and can straightforwardly estimate expression profiles at any time point.

The algorithm we use to estimate the functions (3.7) is outlined in algorithm 2.

Algorithm 2 Expression profile forecasting

Input: matrix $\tilde{W}$ obtained via orthogonal NMF; (row) vector $\boldsymbol{t}$ of experimental time points.
$W_{\text{ilr}} \to \text{ilr}(\tilde{W})$
$X \to \begin{bmatrix} \mathbf{1} \\ \boldsymbol{t} \end{bmatrix}$
Initialise parameters $\boldsymbol{\theta}_i,\ i = 1, \ldots, k-1$ via independent linear regressions of $\boldsymbol{t}$ on the rows of W_{ilr}.
for $s \in \mathcal{S}$ **do**
Find $\Theta_s \to \underset{\Theta}{\text{argmin}} \left[||\text{ilr}^{-1}(\Theta X) - \tilde{W}|| + s||\Theta||_F^2 \right]$
Output: Θ_s corresponding to the largest s for which the objective function is within 5% of its minimum value

We utilise the isometric log-ratio transformation [28] to mitigate the identifiability issues that arise due to the normalisation of population sizes. The ilr transformation maps a vector defined on the unit simplex, $\mathbb{S}^d$ to the space of real numbers, $\mathbb{R}^{d-1}$. This reduction of dimensionality is useful because it removes the redundant dimension which arises due to the sum-to-unity constraint. The i^{th} element of the ilr transformation, y_i, of a simplicial vector $\boldsymbol{x}$ is

$$y_i = \sqrt{\frac{i}{i+1}} \log \left(\frac{\text{GM}(x_1, \ldots, x_i)}{x_{i+1}} \right),\ i = 1, \ldots, d-1,$$

where the function GM computes the geometric mean of its arguments.

A particularly useful property of the ilr transformation in our application is the fact that it linearises normalised processes involving exponential growth [28]. Recall that

the functions $\psi_j(\boldsymbol{x})$ are each associated with an exponential temporal evolution with rate λ_j, i.e. each $\psi_j(\boldsymbol{x})$ grows or decays according to $b_j \exp(\lambda_j t)$. We work on the normalised weights $g_i^*(t)$ as defined in (3.7). Letting $\boldsymbol{u}^*(t) = \text{ilr}(\boldsymbol{g}^*)(t)$ and $\boldsymbol{\gamma} = \text{ilr}(\exp(\boldsymbol{\lambda}))$ where the exponential function is applied element-wise, one arrives at the relationship

$$\boldsymbol{u}^*(t) = \boldsymbol{u}^*(0) + \boldsymbol{\gamma} t.$$

In other words, under the ilr transformation, the temporal evolution becomes linear. The original weights $\boldsymbol{g}^*(t)$ can, of course, be recovered by inverting the ilr transformation.

Thus, for the purposes of forecasting, we estimate the ilr-transformed parameters $\boldsymbol{u}^*(0)$ and $\boldsymbol{\gamma}$. To do so, we use penalised maximum likelihood with a grid search over the regularisation hyperparameter (cf Algorithm 2). The quantity $\text{ilr}^{-1}(\Theta X)$ is the matrix of predicted weights under the parameter choice $\Theta = (\boldsymbol{u}^*(0), \boldsymbol{\gamma})$. We add an L_2 penalty on the sizes of the parameters to promote smoothness in the estimated weights and select the size of the regularisation parameter using a grid search over a grid $\mathcal{S}$. We found that the grid $\mathcal{S} = \{10^i | i = -10, \ldots, 5\}$ was sufficient for our purposes, covering a large range of potential smoothing parameters while keeping the number of required function evaluations manageably small. For problems with very different numbers of timepoints or basis functions from the examples in the following sections, a grid encompassing a wider range of values of s may be required because the relative scales of the parameter norm and the residuals may also differ substantially from those in the subsequent sections. We use the simple heuristic in the output section of algorithm 2 to select the hyperparameter and return the corresponding estimate Θ_s. Examples of the forecasting procedure on real data are presented in the subsequent applications.

Pseudotime estimation

We also use the parameters $\boldsymbol{u}^*(0), \boldsymbol{\gamma}$ in the estimation of a measure of the developmen-

tal stage of individual cells. In contrast to common pseudotime estimation methods, our pseudotime is directly linked to the underlying model. We define pseudotime to be the time at which a given cell state is most likely to be observed under the model. In other words, we define the pseudotime associated with a cell state $\boldsymbol{x}$ to be

$$\tau(\boldsymbol{x}) = \operatorname*{argmax}_t \tilde{n}(\boldsymbol{x}, t), \tag{3.9}$$

$$= \sum_{j=1}^{m} \psi_j(\boldsymbol{x}) \text{ilr}^{-1}(\boldsymbol{u}^*(0) + \boldsymbol{\gamma} t). \tag{3.10}$$

In practice, we restrict the pseudotime to lie in some feasible interval to prevent numerical issues arising from cells lying in the tails of the stationary distribution. The stationary distribution, by definition, does not change in time and so, for cells lying in the tails of this distribution, there are many plausible pseudotimes. We simply impose a maximum pseudotime to mitigate this problem, which should be chosen to be large enough that the system has evolved to almost the stationary distribution by that time. As a default, we take the maximum pseudotime to be 1.5 times larger than the largest experimental time and the minimum pseudotime to be 0.

To maximise (3.10) we use Matlab's in-built `fminbnd` function, which performs global optimisation in one dimension on a bounded domain.

Dimensionality reduction

The estimated eigenfunctions (cf the introductory chapter) can be used to reduce the dimensionality of the data for visualisation. We use the temporal weights associated with the first few eigenfunctions as coordinates for dimensionality reduction. In this way each cell is represented by coordinates describing its level of temporal development. We again deploy the ilr transformation to transform these simplicial weights to real space before visualisation.

Let

$$b_{ij} = \frac{\varphi_j(\boldsymbol{x}_i)}{\sum_{l=1}^{m} \varphi_l(\boldsymbol{x}_i)},$$

the probability that cell i belongs to mixture component j. Then the score vector of cell i for dimensionality reduction purposes is given by

$$\boldsymbol{s}_i = (V^\top V)^{-1} V^\top \boldsymbol{b}_i.$$

We assume that the columns of V are ordered by estimated eigenvalue $\text{ilr}^{-1}(\boldsymbol{\gamma})$. Note that $\boldsymbol{s}_i$ defines the projection of the probability density $\sum_{l=1}^{m} b_{li}\varphi_l(\boldsymbol{x}_i)$ onto the space of probability densities spanned by the approximate eigenfunctions ψ_j.

We subsequently ilr transform the simplicial vector $\boldsymbol{s}_i$ and extract the first 2 or 3 elements as lower dimensional coordinates in which to represent the cell.

3.5 Applications

In this section we apply trPBA to three datasets, two containing genuine experimental data and one synthetic dataset derived from a benchmark model used in [100].

3.5.1 Description of datasets

Synthetic Data

We use data generated by a synthetic stochastic gene regulatory network which was used as a benchmark for the original PBA algorithm [100]. Letting X_1, X_2 denote molecules of chemical species 1 and 2 and x_1, x_2 denote the total number of species

1 and 2 molecules in the system, the network is defined by the following reactions:

$$
\begin{aligned}
&X_1 \to \emptyset \text{ at rate } 0.03x_1,\\
&X_2 \to \emptyset \text{ at rate } 0.03x_2,\\
&X_1 \to 2X_1 \text{ at rate } H(0.1x_1, 4) - H(0.03x_2, 4),\\
&X_2 \to 2X_2 \text{ at rate } H(0.1x_2, 4) - H(0.03x_1, 4).
\end{aligned}
$$

$H(a, b)$ denotes the Hill function

$$H(a, b) = \frac{a^b}{1 + a^b}.$$

We initialise the molecular counts x_1, x_2 as uniformly distributed integers between 100 and 150. We then use the Gillespie algorithm [42] to simulate stochastic trajectories of molecular counts for 2000 cells over 1000 time steps.

Given these simulated trajectories, we collect data from individual cells at the eight 'experimental timepoints' $t = 0, 20, 40, 60, 100, 140, 200, 250$. At each experimental timepoint, we randomly select 500 of the 2000 cells and record their molecular counts to obtain a dataset containing 8 temporal snapshots, each containing 500 cells.

Following the original PBA paper [100], we supplement each of the measurements with eight independent, standard log-normally distributed irrelevant count measurements to obtain the final dataset to be fed to trPBA – a set of 500 10-dimensional molecular count measurements at each of 8 experimental timepoints.

Mouse embryonic stem cell data

The second dataset contains timecourse RNA-seq data from mouse embryonic stem cells [31,47,57]. The dataset contains measurements of 24,174 genes in 2,717 cells across

four timepoints: 0, 2, 4 and 7 days after the cells were treated with leukemia inhibitory factor (LIF). We work on a reduced set of 10 genes identified as particularly interesting by Haghverdi et al [47]. The same dataset was previously analysed with pseudodynamics [31].

Auxin data

The third dataset contains timecourse flow cytometry data collected from budding yeast following exposure to auxin [70]. The dataset is available via the R package `flowTime`. The dataset contains 4-dimensional measurements from a total of 24,157 cells over 28 timepoints.

3.5.2 Results

Synthetic data

Figure 3.1 shows the final estimated solution at various timepoints. The solution is constructed using the MDC-based density estimate with $25-$nearest neighbour bandwidth outlined in the density estimation section above and using 12 basis functions to approximate the solution. The estimated evolution is less smooth at early timepoints, undergoing a sharp jump between $t = 0$ and $t = 20$ in particular. This is an inevitable consequence of the relatively sparse coverage of the state space at early timepoints, meaning that the true solution space is not fully spanned by the data-derived basis functions. However, at later timepoints the evolution becomes smoother due to the greater sampling density and larger overlap between timepoints. This smoothness becomes apparent upon inspection of the weights obtained by orthogonal NMF (cf fig 3.2a). There is very little overlap in components at the early timepoints – i.e. certain components are activated only at a single early timepoint. The ridge penalty on the coefficient estimates aims to alleviate this problem somewhat by smoothing the esti-

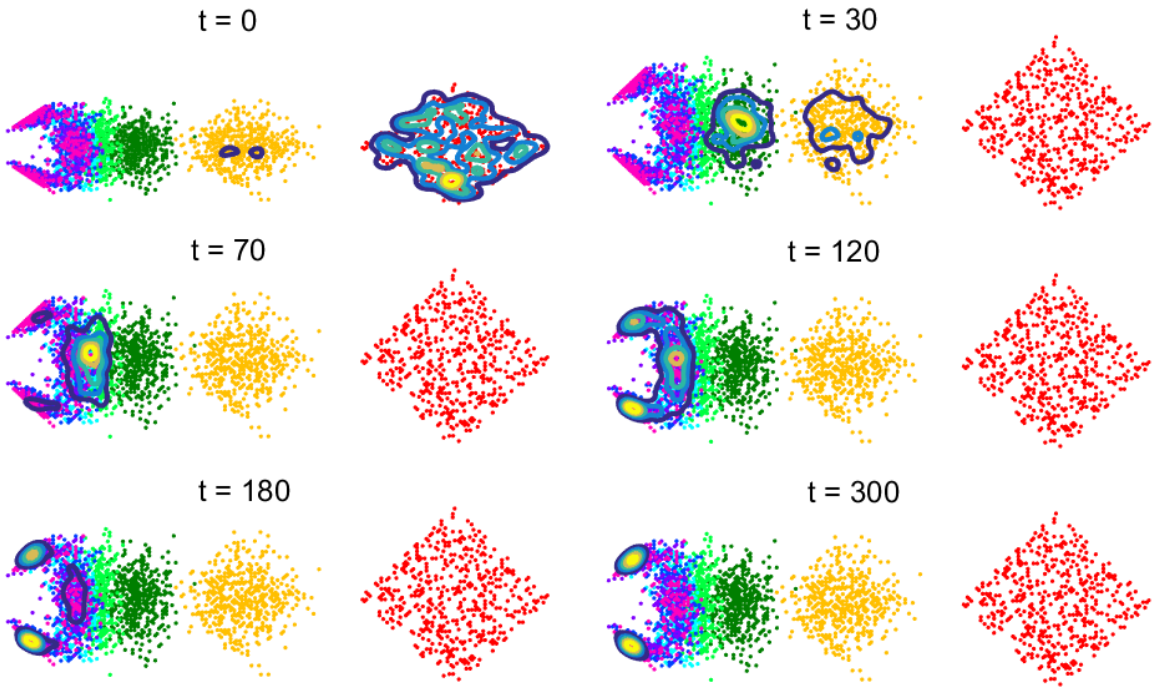
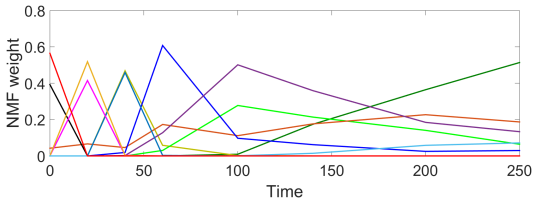


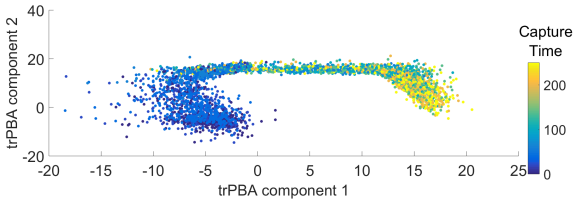
Figure 3.1: Approximate solution at 6 timepoints for the synthetic data. Points represent individual measurements projected onto the first two principal components. Contours are the estimated solutions. Colours of points denote the capture time of the cells: red, 0; gold, 20; dark green, 40; light green, 60; cyan, 100; blue, 140; purple, 200; magenta 250.

mated trajectories. A potential avenue towards addressing this problem more fully is the dependent Dirichlet process, which we discuss later.

The results of the dimensionality reduction procedure are shown in figure 3.2b. The temporal ordering of the points is clearly apparent in this space. The chosen number of basis functions with which to approximate the dynamics plays an important role in determining the quality of the new coordinates. If there are too many basis functions, the first few eigenvectors contain similar information about large-time behaviour and, taken together, they become less informative about the overall dynamics. Nevertheless, with 5 basis functions, the temporal/developmental ordering of the cells becomes clear in the reduced dimensional space. A potential application of these coordinates is in classifying the developmental stage of previously unseen cells, perhaps taken as a single-timepoint snapshot from a patient. While we have not tested this use case, one



(a) Estimated component weights over time for the synthetic data.



(b) Representation of the data in dimensionality reduction coordinates after trPBA with 5 basis functions.

Figure 3.2

can imagine projecting the patient's cells into this new space in order to establish an idea of the stage of a disease for example.

The Spearman correlation between the pseudotimes estimated by trPBA and the actual capture times is 0.8981. Thus the model is capturing the dynamics of a synthetic cell population undergoing a collective change in phenotype reasonably well. PBA [100] estimates the temporal ordering in a different way – it uses the estimated first passage time from the simulation starting value to order cells temporally. PBA reported a correlation of 0.97 between their temporal orderings and the actual capture times of cells. However, while the process being studied is the same, the context and model requirements are quite different. Notably, PBA is reliant on *a priori* estimates of the entry and exit points of cells in genome space (the genetic profiles of nascent and fully developed cells) and estimates of proliferation, death and diffusion rates. The differences between the objectives and requirements of PBA and trPBA are discussed in the discussion section later.

The correlation is relatively high because the overlap between timepoints is relatively small, especially at the earlier timepoints. This correlation is intended merely as a sanity check. If it were the case that the starting population was already in equilibrium, the pseudotimes and actual capture times would not be expected to correlate well in trPBA, while they may well do for PBA. The high correlation in this example simply serves as a check that the dynamics are being recapitulated reasonably well in a case

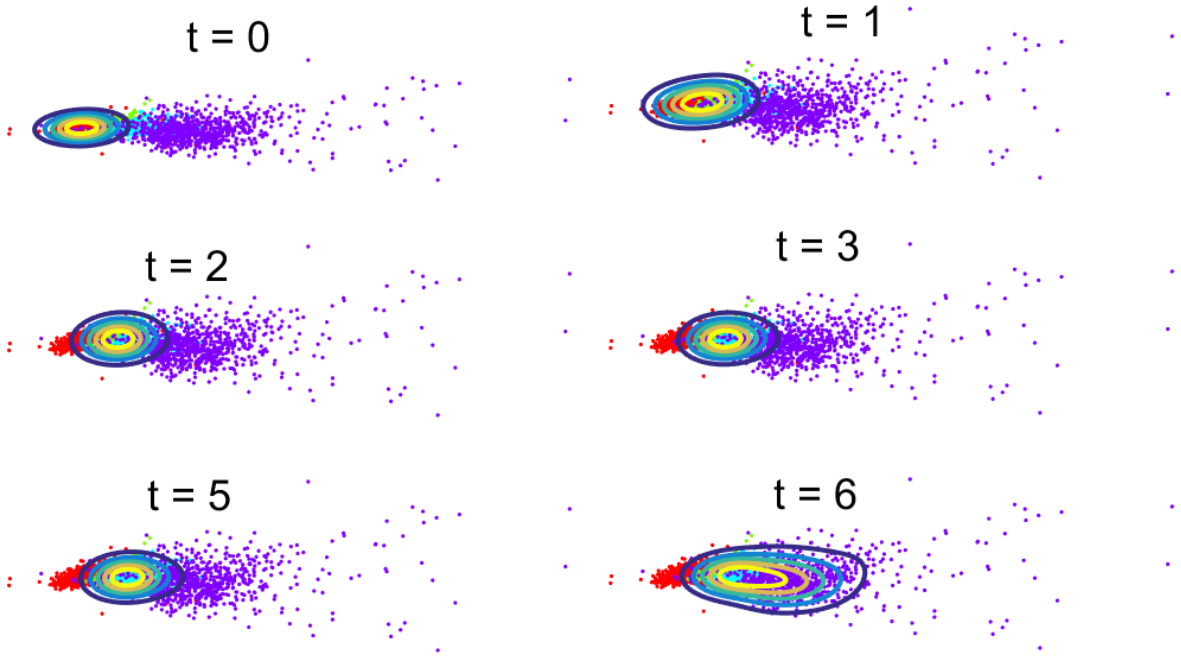


Figure 3.3: Approximate solution at 6 timepoints for the mouse embryonic stem cell data. Points represent individual measurements projected onto the first two principal components. Contours are the estimated solutions. Colours of points denote the capture time of the cells: red, 0; green, 2; cyan, 4; purple, 7.

where we expect pseudotime and actual time to be highly correlated.

Mouse embryonic stem cell data

Figure 3.3 shows the final estimated solution at various timepoints for the mouse embryonic stem cell data. Since the dataset contains only 2717 cells, we do not perform MDC before estimating the trPBA model – the kd-tree based nearest neighbour search performs sufficiently without first selecting a representative subset of points. We use a fixed bandwidth kernel with the smoothness parameter set to 1 (i.e. the default Silverman bandwidth) and 5 basis functions.

The choice of smoothness parameter has a clear influence on the features of the dimensionality reduction. In general, a larger smoothness parameter leads to denser representations, as can be seen in figure 3.4. This behaviour is similar to the effect of

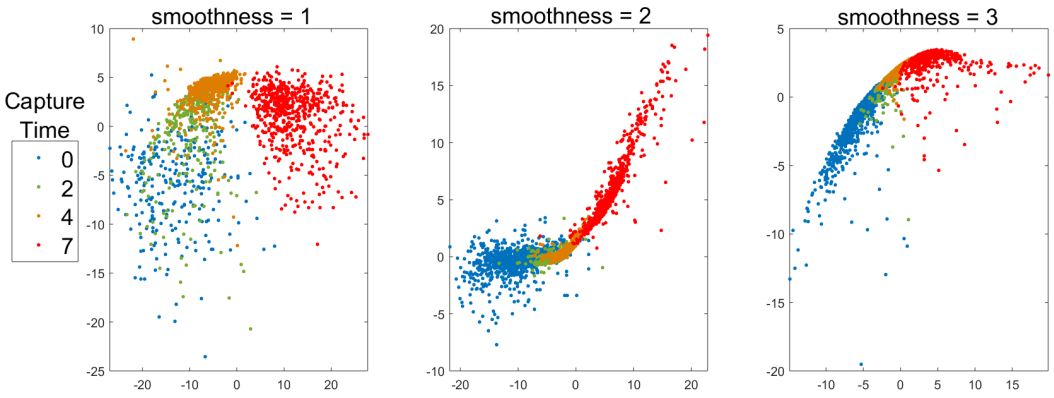


Figure 3.4: Representation of the mouse embryonic stem cell dataset in dimensionality reduction coordinates after trPBA with 5 basis functions & at various settings of the smoothness parameter.

the bandwidth parameter in diffusion maps. We discuss the reasons for this behaviour in the discussion section of this chapter.

The Spearman correlation between the estimated pseudotimes and the actual capture times is 0.9313 for a smoothness parameter of 1 (as in figure 3.3). When the smoothness parameter is 2, the correlation is 0.9127 and for a smoothness parameter of 3, the correlation is 0.929. Again, trPBA is able to recapitulate the dynamics well.

Figure 3.5 shows the actual capture time vs. the estimated pseudotime. Jitter is applied to the capture times for visualisation purposes. One can see that the separation between time points is fairly pronounced, though a number of 'precocious' cells (as defined by [76]) are obvious at the first time point. By the final time point, most cells are deemed to have fully differentiated, with a group of cells which are yet to reach that stage.

We also performed cross-validation to assess the ability of the model to predict the capture time of held-out cells. We estimated the trPBA model (with a smoothness parameter of 1) 50 times on randomly selected training sets of 2500 cells. We then computed the pseudotime for each of the remaining 217 held-out cells and subsequently computed the Spearman correlation between the pseudotimes and the actual capture times

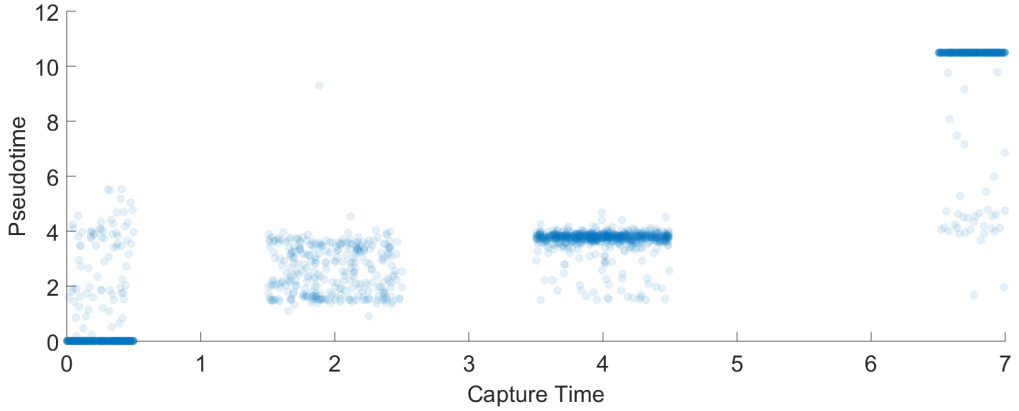


Figure 3.5: Actual capture time vs. pseudotime for the mESC data.

for the held-out cells. The mean Spearman correlation was 0.8876 with a standard deviation of 0.019, a reduction from the 0.9313 for the model trained on the full dataset but still large enough to indicate that the model would be useful for characterising the developmental states of cells not included in the training data.

Auxin data

Figure 3.6 shows the final estimated solution at various timepoints for the auxin data. The dataset contains 24157 individual cell measurements, so we use MDC to select a representative set of points and reduce the computational burden. Again, we use a fixed bandwidth kernel with the smoothness parameter set to 1 (i.e. the default Silverman bandwidth) and 5 basis functions.

Figure 3.7a shows both the empirical weights over time (as estimated by NMF) & the time-varying weights used for forecasting (as estimated by algorithm 2). There is good agreement between the two, suggesting that the assumptions of the model (linear evolution according to the population balance equation) are reasonable. The result of the dimensionality reduction is shown in figure 3.7b. Again, the temporal ordering of the cells is clearly captured in this space.

The time resolution for the auxin data is much finer than for the mESC data, which leads to considerable overlap in the distributions between nearby timepoints. As a

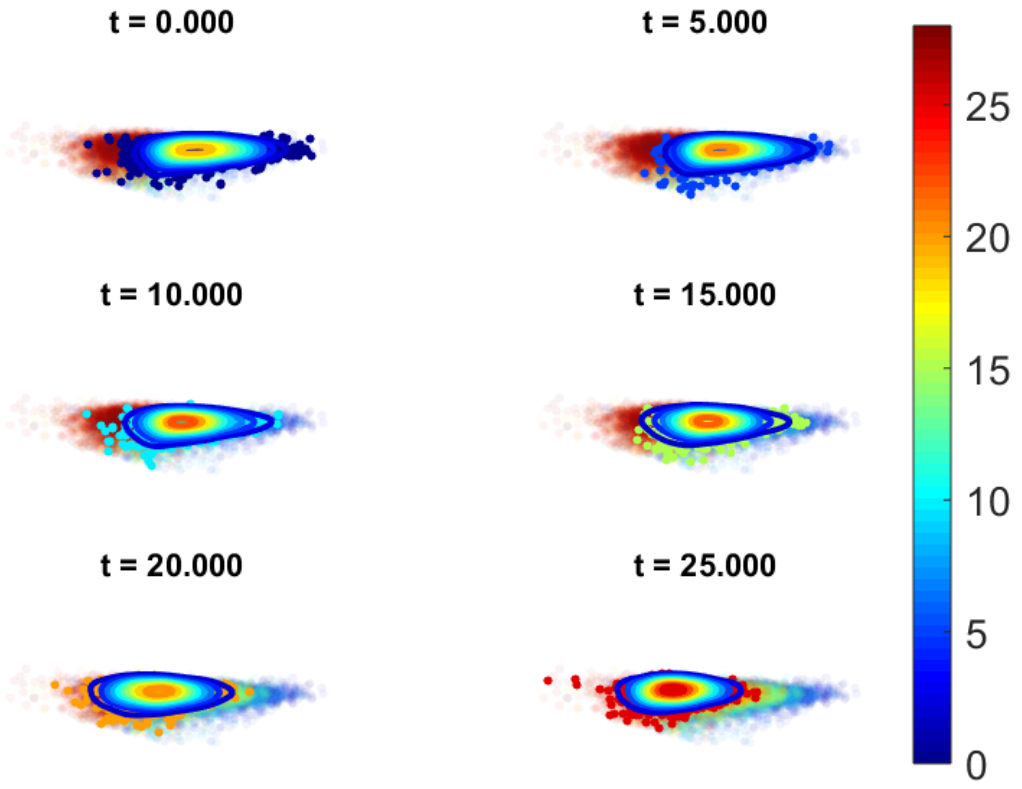
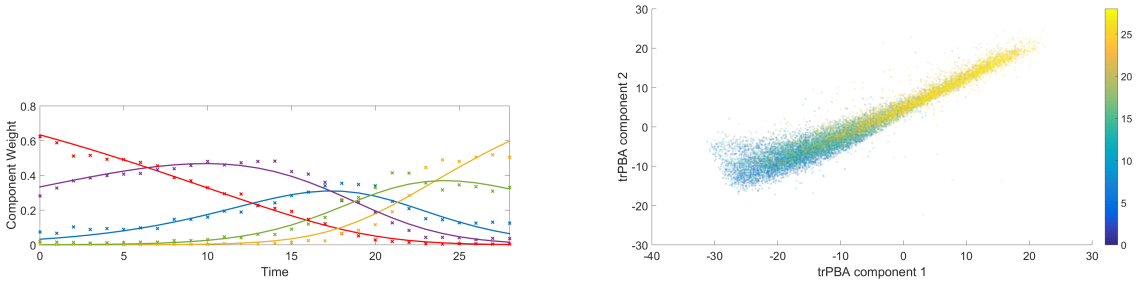


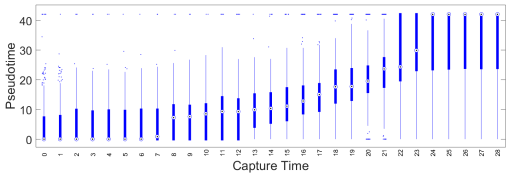
Figure 3.6: Approximate solution at 6 timepoints for the auxin data. Points represent individual measurements projected onto the first two principal components. Contours are the estimated solutions. Colours of points denote the capture time of the cells according to the colour bar on the right.



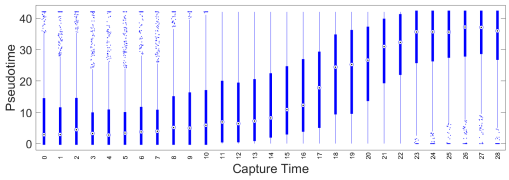
(a) Estimated component weights over time for the auxin data. Crosses are the weights estimated by NMF, lines are the weights $g^*(t)$ as estimated by algorithm 2.

(b) Representation of the auxin data in dimensionality reduction coordinates after trPBA with 5 basis functions. Colour denotes actual capture time.

Figure 3.7



(a) Boxplots of pseudotimes for cells at each timepoint in the auxin data using MDC & kernel density estimation.



(b) Boxplots of pseudotimes for cells at each timepoint in the auxin data using VDPGMM for density estimation.

Figure 3.8

result, the Spearman correlation between pseudotime and capture time decreases to 0.7648. Figure 3.8a shows the distribution of pseudotimes as a function of actual capture time. It is clear that there are cells of varying levels of development at every timepoint, but a general trend towards full differentiation over time.

We also used the variational Dirichlet process mixture model as the density estimation procedure for the auxin data. As one might expect, because it uses fewer mixture components, pseudotimes are less well correlated with actual capture times. The Spearman correlation between true capture times and pseudotimes falls to 0.6. Despite the increased variance in pseudotimes on a per-timepoint basis, the pseudotime trajectory is somewhat smoother.

A considerable benefit of the VDPGMM approach is that the computational burden is much lower. The full trPBA model takes approximately 110 seconds to run using VDPGMM to estimate the density and 250 seconds using MDC & kernel density estimation on a laptop with 8GB of RAM and a 2.3GHz processor.

However, we have found that the results of dimensionality reduction suffer considerably when using VDPGMM. The relatively small number of mixture components is a possible explanation for this – the dimensionality reduction coordinates are formed from the weights associated with each mixture component; when the number of components is small ($\sim$10 vs. $\sim$1000), one would naturally expect the resulting features to be less rich.

3.6 Discussion

In this chapter, we have developed a non-parametric approach to approximating stochastic systems describing cellular evolution. The ideas are similar to those of equation-free modelling, in which features of a dynamical system are learned directly from the data. The application of these ideas to single cell data collected from, for example, flow cytometers leads to a number of interesting and potentially useful outputs: the ability to predict the population profile at unobserved time points, a dimensionality reduction method which can be understood in the context of a general, underlying theoretical model of cellular evolution, and a new way of characterising the developmental status of cells using time-resolved single cell data [76].

Perhaps the two most similar existing approaches are the Gaussian Process Latent Variable model for time course single-cell measurements [76] and the PBA model [100] for single time point, single cell data. Our approach fuses ideas from both of these models to achieve a slightly different set of objectives. trPBA works on experimental time-course data, often obtained as a result of some perturbation experiment, i.e. an experiment in which cells are transferred to an altered environment and their (heterogeneous) adaptation to this new environment is of primary interest. PBA, on the other hand, works on data from a single time point. The objective of the single time point analysis, by contrast, is to attempt to learn something about single-cell dynamics from a system in equilibrium, i.e. the focus is less on the population response to an environmental perturbation and more on the heterogeneity of single cells in a static environment. Because of the lack of time-resolved data, PBA requires estimates of parameters such as proliferation and loss rates, and entry and exit points of cells in gene expression space. While PBA appears to be fairly robust to misspecifcation of these parameters, the use of time-resolved data removes the requirement for the user to specify them altogether.

We also note that visualisation of the results, as in figure 3.6, requires some form of dimensionality reduction to be applied to both the data and the estimated solutions. It should be noted that, if the dimensionality reduction procedure is linear (for example, PCA), the order of dimensionality reduction and trPBA estimation can be interchanged without altering the resultant plots. This dimensionality reduction-first approach is used in pseudodynamics [31]. However, the impact on estimated pseudotimes of performing dimensionality reduction first is unclear and one would lose information about the dependence between individual measured parameters, which may be important for downstream applications.

The linking of principled underlying models and data-driven approaches is a feature that both PBA and trPBA share. While the objectives of trPBA are perhaps more similar to those of [76], their methodology is not explicitly linked to a theoretical model of cellular development. Rather, they use capture time as a prior for pseudotime and train a Gaussian Process Latent Variable model to infer pseudotimes from the time-course data. We believe that the link between our approach and the population balance equation opens up avenues for future research which were not previously available, which we discuss later in this section.

While trPBA does not require the specification of hyperparameters such as proliferation rates (which are likely to be unknown), our smoothness hyperparameter can have a significant influence on the quality of the dimensionality reduction (cf. fig. 3.4). This behaviour is similar to that of diffusion maps [66]. A number of approaches have been developed to deal with the problem of choosing the bandwidth in diffusion maps and some of them may be applicable to trPBA. We chose to primarily use the Silverman bandwidth because of its simplicity and lower computational burden than variable bandwidth kernels [10]. Fixed-bandwidth kernels have been used elsewhere for stationary data arising from unknown stochastic chemical reaction systems [88]. There, the authors develop method of selecting the bandwidth based on the norm of the ker-

nel matrix. If the bandwidth is large and the components are diffuse, the norm of the kernel matrix will be large. As the bandwidth decreases and the components become less diffuse, the norm of the kernel matrix decreases. The authors identify that these norms tend to plateau at small and large bandwidths and propose choosing a bandwidth in between the two plateaus. Such a bandwidth can be selected via a log-log plot for example. Applying a modification of this approach to trPBA could help to mitigate the issues of bandwidth selection in the future.

While we believe that trPBA is useful for describing and learning about the developmental processes of cells from time-course experiments, we also acknowledge that its utility may be better realised when combined with other approaches or used as a sub-element of a larger task. For example, trPBA pseudotime may be a useful as a predictor of disease or a diagnostic marker in cases where there is a characteristic change in phenotype as the disease progresses.

Another interesting avenue for future research is linking the trPBA model to specific parametric choices for the functions in the underlying PBE model. This could lead to efficient procedures for estimating the parameters of PBE models. A basic version of this idea involves simply using trPBA as a pre-processing step in the parameter estimation pipeline. First, one might use trPBA to identify suitable basis functions to use in a Galerkin scheme for solving the underlying PBE. Once this has been done, one can proceed as normal, iteratively improving the parameter estimates using some minimisation algorithm which requires the solution of the PBE under given parameter settings at each iteration. However, it may be possible to exploit the fact that a discrete approximation to the PBE operator is obtained as a byproduct of trPBA (cf equation (3.6)). If a parametric form for the PBE operator is specified, say $\mathcal{L}(\theta)[f]$, it may be possible to estimate the parameters θ by projecting $\mathcal{L}$ onto the function space spanned by the trPBA basis functions and minimising, for example, the Frobenius norm between the matrix L and the projection of $\mathcal{L}(\theta)$.

Along the same lines, one may be able to gain insight into the underlying process by specifying parametric forms for only a few of the terms in the PBE. A simple example is approximating a developmental potential function by assuming a linear proliferation term.

We define the developmental potential function in the same way as [100]. Consider the stochastic single-cell dynamics given by

$$d\boldsymbol{X}_t = -\nabla U(\boldsymbol{X}_t) + \sqrt{2D}d\boldsymbol{W}_t,$$

where $\boldsymbol{X}_t$ is the internal cellular state at time t, U is the developmental potential function, D is the diffusion coefficient, and W_t is a standard Brownian motion.

We assume that, in the absence of cell division or death, this equation adequately describes the evolution of cell states over time. The corresponding Fokker-Planck equation is given by

$$\partial_t p(\boldsymbol{x}, t) = \nabla \cdot (p\nabla U) + D\nabla^2 p$$

and describes the temporal evolution of the probability density function p over cell states.

The potential U can be seen as a quantification of the classical Waddington landscape [98], a popular analogy describing cellular development as a stochastic exploration of a landscape of cell states, with cells eventually coming to rest in the troughs of the landscape. These troughs correspond to mature cellular phenotypes.

In order to estimate a developmental potential function, as is a common objective of models for single cell processes [31,61,100], it is necessary to specify a functional form for the population dynamics part of the underlying population balance equation.

The simplest choice is a state-independent net growth rate, $R \in \mathbb{R}$. This choice leads

to the PBE

$$\begin{aligned}\partial_t N &= \nabla \cdot (N \nabla U) + D \nabla^2 N + rN, \\ &:= \mathcal{L}_1 N + \mathcal{L}_2 N, \\ &:= \mathcal{L} N.\end{aligned}$$

Now, observe that if ψ_i is an eigenfunction of $\mathcal{L}$ with eigenvalue λ_i, ψ_i is also an eigenfunction of $\mathcal{L}_1$ with eigenvalue $\lambda_i - r$: we have

$$\begin{aligned}\lambda_i \psi_i &= \mathcal{L} \psi_i, \\ &= \mathcal{L}_1 \psi_i + \mathcal{L}_2 \psi_i, \\ &= \mathcal{L}_1 \psi_i + R \psi_i,\end{aligned}$$

so that $(\lambda_i - R)\psi_i = \mathcal{L}_1 \psi_i$.

Additionally, the first eigenfunction of $\mathcal{L}_1$ has eigenvalue 0, and is identical to the stationary distribution associated with the Fokker-Planck operator $\mathcal{L}_1$ [36]. The stationary distribution is given by the Gibbs measure

$$p_s(\boldsymbol{x}) = \frac{1}{Z} \exp(-U(\boldsymbol{x})/D),$$

where Z is the normalisation constant which ensures $p_s(\boldsymbol{x})$ integrates to 1.

Consequently, armed with an estimate of the first eigenfunction of $\mathcal{L}$, which is obtained via the non-negative matrix factorisation procedure, we have the relationship

$$U(\boldsymbol{x}) = -D \log(Z \psi_1(\boldsymbol{x})).$$

The diffusion parameter D is is general unknown. However, qualitative aspects such

as the number, location and relative depths of potential wells are unchanged by D. Similarly, the normalising constant Z does not meaningfully alter the key qualitative aspects of the landscape. Thus, for the purposes of visualisation, an informative qualitative appraisal of the single-cell dynamics can be obtained by plotting the quantity $-\log(\psi_1(\boldsymbol{x}))$ projected onto one or two dimensions, be they principal components, diffusion components, or simply dimensions of interest from the original dataset.

Of course, qualitative conclusions drawn from such plots are subject to the assumptions that the intracellular dynamics can be well described by diffusion in a potential (the Waddington assumption) and that the population dynamics are linear and state-independent. These assumptions do not necessarily apply to other conclusions drawn via trPBA.

Estimating a potential under more complex population dynamics is more difficult because the simple relationship between the eigenpairs of the PBE operator $\mathcal{L}$ and the Fokker-Planck operator $\mathcal{L}_1$ becomes less trivial. It is still the case that the leading eigenvalue of $\mathcal{L}$ is equivalent to the leading eigenvalue of $\mathcal{L}_2$, the operator describing the population dynamics. However, the relationship between the eigenfunctions is not straightforward. One could perhaps attempt to exploit the fact that eigenfunctions of Fokker-Planck operators are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_{p_s^{-1}}$ to devise a suitable additive decomposition of the matrix $M^{-1}L$, the Galerkin projection of the operator $\mathcal{L}$. Such extensions are promising avenues for future work.

CHAPTER 4

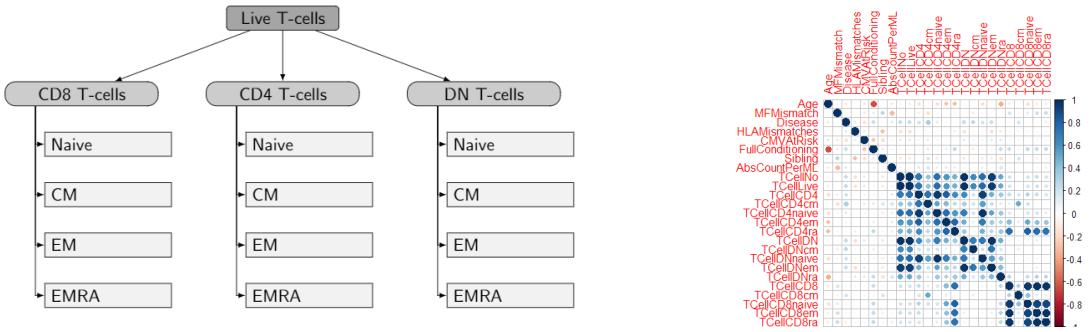
THE LN-CASS PRIOR AND PREDICTION OF GVHD INCIDENCE VIA T-CELL FLOW CYTOMETRY DATA

4.1 Introduction

In this chapter we develop a flexible Bayesian shrinkage method and apply it to four datasets – three real world datasets from the fields of immunology, metabolomics and cancer genomics, and one synthetic dataset. The real world datasets all relate to measurements of potential biomarkers for disease obtained via single-cell technologies and aggregated at the patient level, but the method could equally be applied to prediction problems using non-aggregated single cell data.

We note here that some of the material in sections 4.5 - 4.8 & the discussion section appeared in our paper 'Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior' [93] and is mostly taken directly from that publication.

The primary motivation for the development of the method was to leverage data on the incidence of graft-versus-host disease, the primary complication of allogeneic haematopoi-



(a) (b)

Figure 4.1: (a) the subset structure of the cell count data, (b) a correlation heat map of the raw predictors the cell counts are strongly correlated within subsets.

etic stem-cell transplantation (AlloHSCT). AlloHSCT is used as a treatment for haematologic malignancies such as leukaemia and is generally part of a treatment regimen including a combination of chemotherapy, radiotherapy and immunosuppression.

There is a simple rationale for AlloHSCT-based treatment regimens, but it is complicated by the potential for GvHD. The rationale is that the patients immune system is compromised by the presence of some malignancy – for example, leukaemias are cancers of the immune system. The objective of AlloHSCT is to replace this faulty immune system with a healthy one. First it is destroyed (to varying degrees, this is what the `FullConditioning` predictor variable below refers to) via some myeloablative conditioning procedure such as chemotherapy or radiotherapy. Following this ablation, the seeds of a new immune system are transplanted in the form of haematopoietic (blood-related) stem cells from, for example, the bone marrow of a donor or umbilical cord blood. Ideally, the transplanted stem-cells are able to mature healthily in the patient and provide them with a fully functioning immune system after some time (this can take up to two or three years).

Often, mature immune cells (i.e. cells capable of mounting immune responses) are included in the graft. The reason for this is that in the aftermath of their immune system being depleted, patients are very susceptible to infection. Additionally, disparities in the genetic make-up of the patient and donor mean that residual leukaemic cells that

survived the conditioning procedure are recognised as 'foreign' and destroyed by the transplanted immune cells. This is known as the Graft-versus-Leukaemia (GvL) effect.

However, the GvL effect is coupled (as yet inseparably) to Graft-versus-Host Disease (GvHD), in which the same genetic disparities that are responsible for the GvL effect lead to the transplanted, immunocompetent cells recognising healthy tissue in the patient as foreign and mounting destructive immune responses against it. The gut, skin and lungs are particularly vulnerable to GvHD. Excellent recent reviews of GvHD pathogenesis and treatment are available in [30,92], but we do not go into further detail here.

4.1.1 The GvHD Data

The raw dataset consists of a single binary response variable (`gvhd`) and 25 predictor variables measured for 51 patients. These variables are listed in table 4.1. Among the 51 patients, 16 developed GvHD. Empirical CDFs of the raw predictor variables are shown in figure 4.2 – note that as the subset-granularity increases, i.e. as we move from `TCellNo` to, for example, `TCellDNNaive`, the coverage of the predictor space becomes very sparse. The data were collected and processed in October 2013 by the Moss group at the School of Cancer Sciences, University of Birmingham.

There are a number of complications we need to consider while building a model. Firstly, we notice that the T-cell count predictors are generally strongly concentrated around zero but with a few measurements up to four orders of magnitude larger, meaning that the data are very sparse across the full range of the T-cell count predictors. To get around this issue we transform the data via a continuous non-linear transformation explained below and include an indicator variable for cell counts below the detection limit. Secondly, the number of positive observations of the response variable (GvHD incidence) is smaller than the number of predictors, meaning that including

Variable	Description	Type	Group/Level
Age	Age of patient at time of transplant.	Continuous	NA
MFMismatch	Sex mismatch between patient and donor.	Binary	NA
Disease	Type of leukaemia, myeloid or lymphoma.	Binary	NA
HLAMismatches	Antigenic matching of donor and recipient; good ($> 6/8$), bad ($< 6/8$).	Binary	NA
CMVAtRisk	Cytomegalovirus seropositivity of patient.	Binary	NA
FullConditioning	Extent of conditioning (radiation, chemotherapy) of the patient.	Binary	NA
Sibling	Indicator of donor/recipient sibling relationship.	Binary	NA
TCellLive	Number of live T-cells per ml of blood two weeks post-transplant.	Continuous	Level 0
TCellCD4	Number of live CD4 T-cells per ml of blood two weeks post-transplant.	Continuous	Group 1, level 1
TCellCD4cm	Number of live CD4 central memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 1, level 2
TCellCD4Naive	Number of live CD4 naive T-cells per ml of blood two weeks post-transplant.	Continuous	Group 1, level 2
TCellCD4em	Number of live CD4 effector memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 1, level 2
TCellCD4ra	Number of live CD4 EMRA T-cells per ml of blood two weeks post-transplant.	Continuous	Group 1, level 2
TCellCD8	Number of live CD8 T-cells per ml of blood two weeks post-transplant.	Continuous	Group 2, level 1
TCellCD8cm	Number of live CD8 central memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 2, level 2
TCellCD8Naive	Number of live CD8 naive T-cells per ml of blood two weeks post-transplant.	Continuous	Group 2, level 2
TCellCD8em	Number of live CD8 effector memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 2, level 2
TCellCD8ra	Number of live CD8 EMRA T-cells per ml of blood two weeks post-transplant.	Continuous	Group 2, level 2
TCellDN	Number of live double-negative T-cells per ml of blood two weeks post-transplant.	Continuous	Group 3, level 1
TCellDNcm	Number of live double-negative central memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 3, level 2
TCellDNNaive	Number of live double-negative naive T-cells per ml of blood two weeks post-transplant.	Continuous	Group 3, level 2
TCellDNem	Number of live double-negative effector memory T-cells per ml of blood two weeks post-transplant.	Continuous	Group 3, level 2
TCellDNra	Number of live double-negative EMRA T-cells per ml of blood two weeks post-transplant.	Continuous	Group 3, level 2

Table 4.1: Table of raw predictors. See text for additional details.

all predictors (without penalising the model complexity) would lead to an overparameterised model which perfectly classifies the training data, but probably performs poorly against out of sample data. As a result, we need to regularise the problem in order to stabilise our inferences and improve out of sample predictive performance.

We choose to work in a Bayesian framework for the reasons outlined below, and the regularisation is accomplished via prior distributions on the coefficients with strong concentration around zero and heavy tails. Finally, the cell count predictors are strongly correlated, as shown in figure 4.1, as we would expect. We implement a novel Bayesian shrinkage method to explicitly account for the correlations induced by the tree-like structure of the data.

4.1.2 Models for classification: Logit, Probit and Robit

We now provide a brief review of regression methods for classification. The Logit, Probit and Robit models are unified by the common framework of the generalised linear

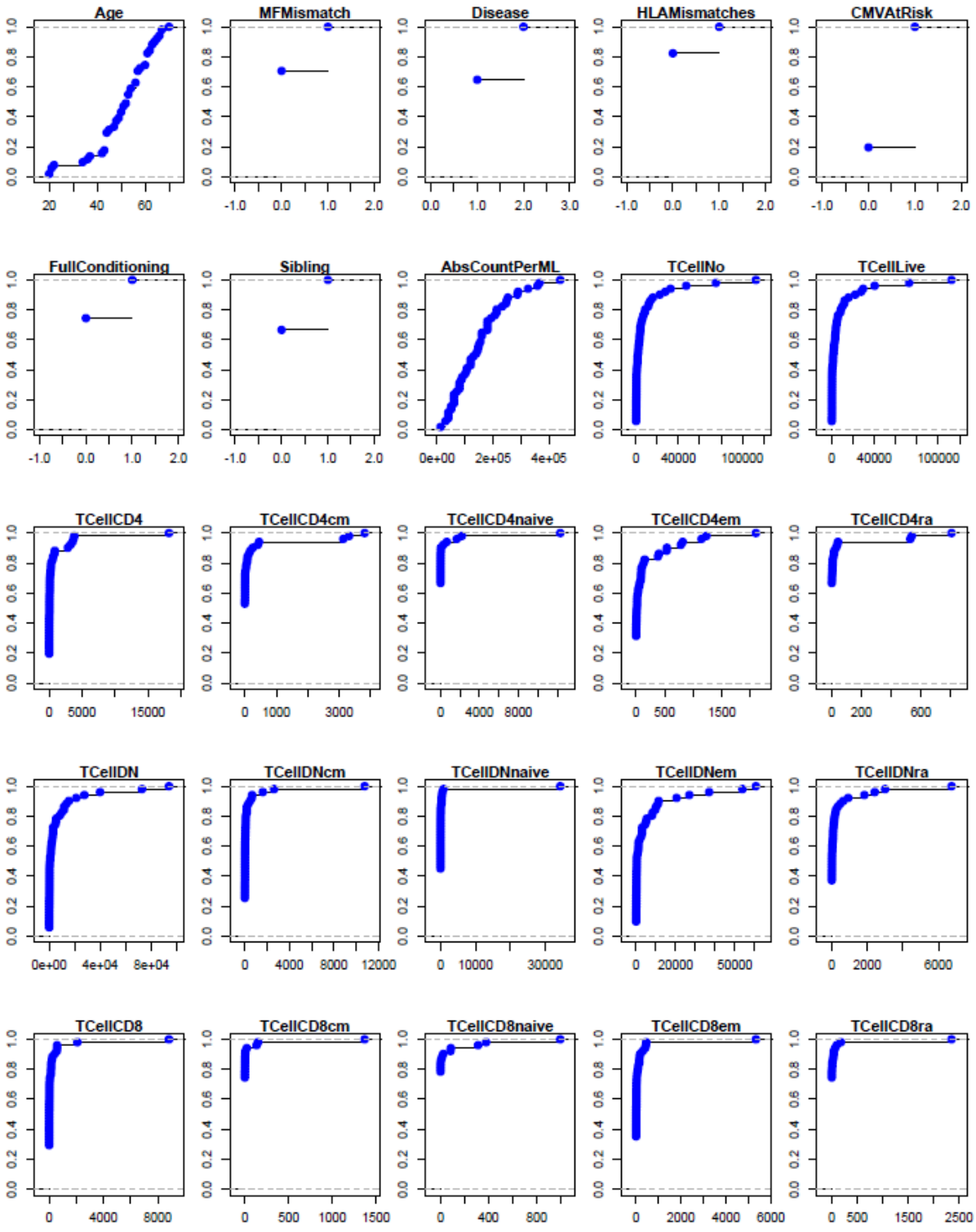


Figure 4.2: Empirical cumulative distribution functions for the raw predictors. Note that many cell counts are zero-inflated and contain large outliers.

model (GLM). GLMs are characterised by the assumption that some transformation of the response (known as the *link function*) can be predicted by a linear combination of the predictors.

The simplest way to distinguish between logit, probit and robit models is through the latent variable formulation of GLMs. Essentially, the three methods use a slightly different sigmoidal cumulative density function to transform the linear predictor, which corresponds to making different distributional assumptions on the error associated with the linear predictor.

The classification model (for binary responses) is formulated as follows (see for example [37])

$$y_i = \begin{cases} 1 \text{ if } \beta_0 + X_i\beta + \varepsilon_i > c, \\ 0 \text{ otherwise,} \end{cases} \tag{4.1}$$

with c being a threshold constant.

Then clearly, assuming ε_i is symmetrically distributed,

$$P(y_i = 1|\beta_0, \beta, X_i) = P(\beta_0 + X_i\beta + \varepsilon_i > c), \tag{4.2}$$

$$= P(\epsilon_i > c - \beta_0 - X_i\beta), \tag{4.3}$$

$$= F_\varepsilon(\beta_0 + X_i\beta - c), \tag{4.4}$$

where F_ε is the CDF of ε.

The logit model follows from choosing ε to be standard logistic, probit from choosing ε to be standard normal, and robit from choosing ε to be standard student-t with any chosen degrees of freedom. Each element of the response vector is then modelled as

being Bernoulli with success probability (4.4). Commonly $c = 0$, but this is varied in, for example, the computation of ROC curves for evaluating classifier performance.

The three methods form a 'robustness hierarchy': the logistic distribution has the lightest tails, meaning that it places the most weight on extreme values of the linear predictor. The tails of the t-distribution become lighter as the degrees of freedom parameter approaches ∞, in which case we recover the normal distribution, i.e. the Probit model. The heavier tail is the reason that the model with t-distributed latent errors is called the robit model, the rob- prefix refers to robustness. Gelman [37] advocates the use of robit regression as a default, and it is particularly appropriate for the GvHD dataset studied below for reasons stated there. Below we choose the t-distribution with one degree of freedom, i.e. the Cauchy distribution, which has very heavy tails. This is also known as a Cauchit model.

4.1.3 Bayesian shrinkage and selection

We begin this subsection with a short philosophical argument that Bayesian inference is the ideal setting for problems in which shrinkage estimation is appropriate, i.e. problems in which we have a small sample size relative to the number of predictors.

Bayesian inference (via Markov Chain Monte Carlo methods) differs from classical statistical methods in two fundamental ways. The first is the introduction of the prior distribution over the parameters. The second is that the touchstone of any inference is the full posterior distribution over the parameters, approximated via sampling. We argue that both of these aspects take on a heightened importance in shrinkage problems and that, as a result, Bayesian inference is more natural than (penalised) maximum likelihood estimation in the shrinkage setting.

Firstly, the prior distribution is most commonly introduced as an encoding of 'prior

belief' or 'prior knowledge' about a parameter, perhaps informed by previous experimental work or theoretical considerations related to the problem at hand. However, the prior distribution is equally interpretable as a function designed to either accentuate or dampen certain aspects of the likelihood in a more general sense. That is, the prior distribution is simply a weighting function for the likelihood. The prior distribution is really an encoding of our initial beliefs about the potential inadequacies of the likelihood in reflecting the 'true' population parameters, or in other words the degree to which we expect that our inferences might not generalise well to the population as a whole. This view of the prior-likelihood relationship is particularly useful in the context of regularisation. A judicious choice of prior acts as a magnifying glass to focus on relevant regions of parameter space and ignore those that may have been artificially amplified as artefacts of the particular sample we happen to have collected. For example, discussing default prior distributions for logistic regression, Gelman et al. [38] make the case that for standardised predictors, logistic regression coefficients larger than 5 are unreasonable in most circumstances, while separation due to only having obtained a finite sample from the population can lead to maximum likelihood estimates that do not even exist. There is a clear balancing act between the data at hand and the population as a whole in problems with small sample sizes relative to the number of predictors, and the prior distribution helps to tip the scales in favour of the population as a whole.

Of course, this modification of the likelihood function is possible through penalised likelihood methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) and ridge estimators [51, 94], in which the loss function to be minimised is modified to more accurately reflect the 'loss' associated with a particular parameter estimate. These methods are discussed further below. A main advantage of Bayesian modelling over penalised likelihood methods is the more all-encompassing evaluation of uncertainty associated with Bayesian models. This is a particular concern in shrinkage problems, where sample sizes are small and asymptotic results that lead to confidence interval

estimates are unlikely to hold. Additionally, in the Bayesian framework, predictions average over all of the posterior uncertainty directly rather than relying on point estimates of parameters. This is a benefit with small sample sizes because distributions over parameters may not be well-described by their modes alone, or may even be multimodal.

A drawback of the Bayesian approach is that it carries an increased computational cost when compared with (penalised) maximum likelihood approaches, which are generally very fast. The disparity grows with the size of the problem and we feel that for small datasets with lots of associated uncertainty, like the ones studied later in this chapter, the modest increase in computational cost is a reasonable price to pay for the reasons stated above.

We now outline the mathematical details of some common shrinkage estimators, both frequentist and Bayesian.

We begin with the Least Absolute Shrinkage and Selection Operator (LASSO, [94]), probably the most commonly used regularisation technique. The LASSO is a special case of the general penalised likelihood estimator, in which the objective is to find

$$\hat{\beta}_{\text{pen}} = \underset{\theta}{\text{argmin}}\{-\ell(\theta) + \lambda\rho(\theta)\}. \tag{4.5}$$

The LASSO corresponds to the choice $\rho(\theta) = ||\theta||_1$, so that we aim to minimise the negative log-likelihood plus a multiple of the ℓ_1 norm of the parameters. This penalty encourages sparse solutions and has a particularly fast estimation routine available in the `glmnet` package in R. Another common penalised likelihood method is the ridge estimator [51], in which the ℓ_2 norm of the parameters is chosen as the penalty function. The ridge encourages shrinkage, but not identically zero parameter estimates. The LASSO and ridge estimators can be combined in a weighted sum to yield the

elastic net [107], which provides both strong and weaker regularisation via the ℓ_1 and ℓ_2 components respectively. A more general class of penalisation methods is class of *bridge* estimators, in which the penalty is the ℓ_q norm. The case $q = 0$ leads to the so-called 'best-subset selection', in which the penalty is simply the number of parameters in the model. Best-subset selection is the gold standard, in much the same way as the spike-and-slab is the gold standard for Bayesian shrinkage/selection, however both approaches are computationally intractable and so relaxed versions such as the LASSO have been developed for practical implementation.

Penalised likelihood methods naturally give rise to Bayesian counterparts, implemented through the choice of prior distributions.

The first such Bayesian adaptation was the Bayesian LASSO [68], and others have been developed since – the Bayesian elastic net [60], the Bayesian bridge [73] and the Bayesian grouped LASSO [104] among them.

The Bayesian versions of these estimators are based on choosing prior distributions which guarantee that the maximum a posteriori estimate is equivalent to the frequentist penalised likelihood estimate.

More concretely, given a prior of the form $\pi(\theta) \propto \exp(-\lambda\rho(\theta))$, we can write the posterior distribution as

$$\pi(\theta|\mathcal{D}) \propto \exp\{\ell(\mathcal{D}|\theta) - \lambda\rho(\theta)\}, \tag{4.6}$$

which is maximised at the β satisfying (4.5), i.e. the maximum a posteriori estimate of the Bayesian formulation is equivalent to the penalised likelihood method. Choosing $\rho(\theta) = ||\theta||_1$ leads to the Bayesian LASSO, in which the parameters are given independent zero-mean Laplace (double-exponential) priors. The ridge estimator is obtained by choosing zero-mean Gaussian priors.

While this connection between Bayesian and frequentist approaches is attractive, in practice better alternatives are available in most situations. The unifying framework of Bayesian shrinkage priors is the global-local shrinkage rule [72], in which priors are assumed to be scale-mixtures of normals [4].

A general global-local shrinkage rule can be stated as follows:

$$\theta_i|\tau, \lambda_i \sim \mathcal{N}(0, \tau^2\lambda_i^2), \tag{4.7}$$

$$\lambda_i \sim L, \tag{4.8}$$

$$\tau \sim G, \tag{4.9}$$

where L and G are appropriate distributions.

The Bayesian LASSO can be obtained by choosing L to be a Rayleigh distribution (so that λ_i^2 has an exponential distribution), while setting G to be the degenerate distribution (corresponding to a constant random variable) centred at some positive value.

A number of shrinkage priors have been developed that fit into this framework. Particular examples are: the horseshoe [21] and the horseshoe-plus [13], which choose G and L to be half-Cauchy distributions, and the normal-exponential-gamma prior/generalised double-Pareto prior [6, 46], which choose λ_i to have an exponential distribution with Gamma-distributed squared rate parameter and rate parameter respectively.

These priors all share the common feature that they are relaxations of the spike-and-slab prior, which can also be formulated as a global-local shrinkage rule with L chosen to be a Bernoulli distribution. The prior we formulate below takes the Logit-Normal PDF as a continuous approximation to the Bernoulli PMF. The reason for this approximation is to try to alleviate the difficulties arising as a result of the combinatorial complexity of the spike-and-slab method. Taking a continuous approximation to the Bernoulli distribution results in a continuous posterior distribution which allows

MCMC sampling procedures to be guided by the posterior curvature and more reliably visit high probability regions of parameter space. This issue is discussed further in the section on the Hamiltonian Monte Carlo method below.

4.2 The LN-CASS prior with group structure

In this section we present a (to our knowledge) novel methodology for Bayesian variable selection, which is applicable to a much wider class of problems than the particular example studied here. We begin by outlining the method in a general setting, and then formulate the particular statistical model we will use for the GvHD data.

We call the prior LN-CASS, which stands for Logit-Normal continuous analogue of the spike-and-slab. As mentioned in the section above, the spike-and-slab prior is the gold standard of Bayesian variable selection. However, effectively exploring the posterior distribution becomes problematic as the number of parameters increases, on account of the combinatorial complexity of the problem. The additional requirement for using Gibbs sampling to deal with the discrete hyperparameters of the spike-and-slab means that the exploration of the posterior distribution is likely to be inefficient when efficiency is of the utmost importance – the algorithm is required to locate regions of high posterior probability in a discrete space of 2^p possible parameter subsets, and estimate posterior distributions within each of those subsets. Relaxing the Bernoulli distribution of the spike-and-slab to a continuous distribution with substantial mass at 0 and 1 is appropriate and has the advantage of ensuring a continuous posterior distribution which can be explored by HMC. While the combinatorial complexity of the problem is still a potential issue, the exploration of the posterior is guided by its curvature and so we should arrive at the 'typical set' – the region of high posterior probability as described by [11] – much more efficiently than with direct spike-and-slab

priors. The group structure is encoded by treating the problem essentially as a mixed-effects model, in which we have both group-level and individual level parameters, with the individual-level parameters being encouraged towards the group-level mean.

We now outline the details of the method. Suppose we are working with a generic (parametric) statistical model of the form

$$y_i \sim \mathcal{F}(\mu = f(X_i; \theta)), \tag{4.10}$$

with $\mathcal{F}(\mu = f(X_i; \theta))$ some problem-specific distribution whose mean is a function of a vector of predictors X_i and and parameters θ.

We additionally assume that there is some group structure associated with the covariates. We denote by X_{G_i} the matrix of covariates in group i, with G_i being the set of indices associated with group i.

The LN-CASS prior without group structure is given by

$$\theta_i | \lambda_i \sim \mathcal{N}(\tilde{\theta}_i, \tau^2 \lambda_i^2), \tag{4.11}$$

$$\lambda_i \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda). \tag{4.12}$$

The prior is thus a global-local shrinkage rule, with global shrinkage parameter τ and local shrinkage parameter λ_i. $\tilde{\theta}_i$ is the value towards which θ_i is shrunk. If a sparse parameter vector is required, we choose $\tilde{\theta}_i = 0$, but other choices may be reasonable, depending on the problem.

The basic LN-CASS prior is used to shrink a single parameter in a non-hierarchical way, i.e. independently of the sizes of other parameters. It is the building block for the hierarchical complexity models used in the remainder of this chapter.

The mixture parameter λ_i is analogous to an inclusion indicator for the variable of

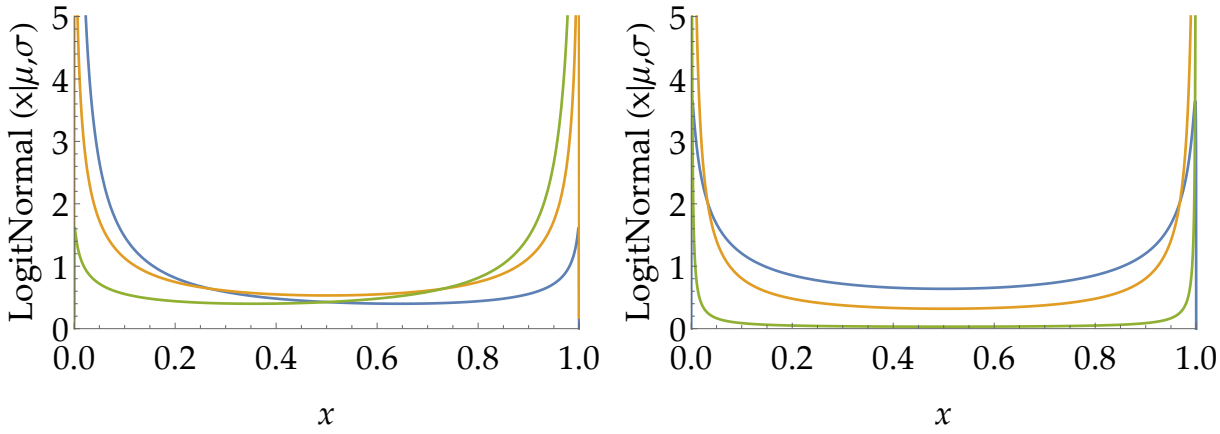


Figure 4.3: Logit-Normal distributions. (a) varying the location parameter μ_λ. (blue, orange, green) = (-2,0,2). (b) varying the scale parameter σ_λ. (blue, orange, green) = (2.5, 5, 50).

interest. Indeed, replacing the Logit-Normal prior with a Bernoulli prior yields a spike-and-slab model with a spike at 0 and a normal slab with variance τ^2, corresponding to a situation in which the i^{th} variable is either completely excluded, or included with a $\mathcal{N}(0, \tau^2)$ prior.

Equations (4.11) & (4.12) can be rewritten in the following way, yielding a model in which inference is performed on a new parameter vector $(\boldsymbol{\theta}, \tilde{\boldsymbol{\lambda}})$ specified solely in terms of a multivariate normal prior with diagonal covariance matrix,

$$\begin{aligned} \theta_i|\tau, \lambda_i &\sim \mathcal{N}(0, (\lambda_i\tau)^2), \\ \tilde{\lambda}_i &\sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \\ \lambda_i &= \text{logit}^{-1}(\tilde{\lambda}_i). \end{aligned}$$

This formulation empirically greatly enhances the performance of the No-U-Turn Markov Chain Monte Carlo sampler (NUTS) used for making posterior inference and results in much improved convergence properties over similar priors (e.g. the horseshoe).

For now we take $\tilde{\theta}_i = 0$ but stress that any finite choice of $\tilde{\theta}_i$ is possible.

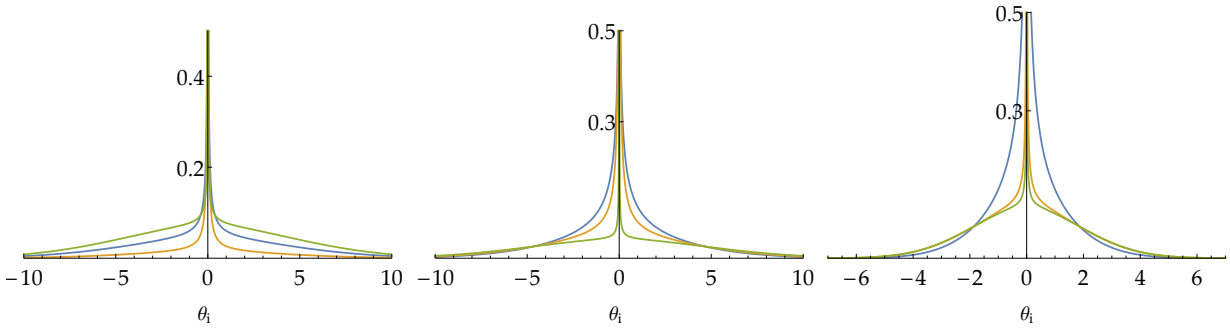
The Logit-Normal distribution has support on $(0, 1)$ and is analogous to the beta distribution, but with less extreme behaviour at the end-points which aids sampling. If X is normally distributed with mean μ and standard deviation σ, the random variable obtained by applying a logistic sigmoid transformation to X is Logit-Normal with 'location' parameter μ and 'scale' parameter σ. We often choose to fix the hyperparameters μ_λ, σ_λ, which encode our prior beliefs about sparsity. μ_λ dictates the relative densities at 0 and 1 and σ_λ dictates the concentration of the prior near the endpoints. The Logit-Normal distribution is plotted in figure 4.3 for various values of the hyperparameters. Choosing $\mu_\lambda = 0$ corresponds to a symmetric prior assigning equal prior probability to inclusion and exclusion of the variable. Positive values assign more probability to inclusion and negative values to exclusion. σ_λ is a 'polarisation' parameter, dictating how harshly we want to enforce binary inclusion/exclusion decisions. A default value of σ_λ seems to work in most situations, if the data are placed on a common $[0, 1]$ scale before analysis. Of course, hyperpriors could be specified for a fully Bayesian analysis. A default hyperprior choice for the global shrinkage parameter τ is the half-Cauchy distribution [21, 38].

For a fixed global shrinkage parameter τ, the prior distribution of θ_i is given by

$$\frac{1}{2\pi\tau\sigma_\lambda}\int_0^1 \frac{1}{\lambda_i^2(1-\lambda_i)}\exp\left(-\frac{(\text{logit}(\lambda_i)-\mu_\lambda)^2}{2\sigma_\lambda^2}-\frac{x^2}{2\tau^2\lambda_i^2}\right)d\,\lambda_i \tag{4.13}$$

We compute this integral numerically via Monte Carlo integration with one million points and present the results in figure 4.4 for a variety of choices of the hyperparameters.

We see that the prior has considerable mass close to zero regardless of the hyperparameter choices, but the sharpness of the prior and the concentration slightly away from zero are affected by the choices. Note that as σ_λ becomes larger, the prior becomes closer to the spike and slab prior as the Logit-Normal distribution becomes more po-



(a) (b) (c)

Figure 4.4: The LN-CASS prior. (a) Varying the location parameter μ_λ. (blue, green, orange) = (0,10,-10) with $\sigma_\lambda = 10$, $\tau = 5$. (b) Varying the scale parameter σ_λ. (blue, orange, green) = (2.5, 5, 50) with $\mu_\lambda = 0$, $\tau = 5$. (c) Varying the scale parameter σ_λ. (blue, orange, green) = (2.5, 20, 50) with $\mu_\lambda = 0$, $\tau = 2$.

larised (figure 4.4 (b)). Varying μ_λ controls the total density allotted to the 'slab', with larger values yielding more slab-like priors (figure 4.4 (a)). We can see more clearly the potential for elastic net-like regularisation, which corresponds to a prior consisting of a product of Laplace and normal distributions, in figure 4.4 (c) where, as σ_μ increases, we obtain a significant spike around zero piercing an approximately Gaussian distribution with standard deviation 2. The prior used below for both the GvHD and simulation studies is the blue curve in figure 4.4 (a), i.e. with $\mu_\lambda = 0$, $\sigma_\lambda = 10$ and $\tau = 5$.

We account for grouped predictors by casting the problem in a mixed-effects framework. We assume that for each $i \in G_i$,

$$\theta_i = \theta_{\text{group}} + \theta_{\text{ind},i}, \tag{4.14}$$

and assign the following LN-CASS priors to both θ_{group} and $\theta_{\text{ind},i}$,

$$\theta_{\text{group}} | \lambda_{\text{group}} \sim \mathcal{N}(0, \lambda_{\text{group}}^2 \tau^2), \tag{4.15}$$

$$\theta_{\text{ind},i} | \lambda_{\text{group}}, \lambda_{\text{ind},i} \sim \mathcal{N}(0, \lambda_{\text{ind},i} \lambda_{\text{group}}^2 \tau^2), \tag{4.16}$$

$$\lambda_{\text{group}} \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda), \tag{4.17}$$

$$\lambda_{\text{ind},i} \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda). \tag{4.18}$$

In doing this we encourage shrinkage at the group-level, i.e. we encourage removal of the group as a whole, as well as at the individual-level, i.e. we encourage the elements of the group to share the parameter θ_{group}.

A typical application domain for sparse, grouped estimators is the analysis of DNA microarray data, in which high-throughput sequencing yields a high-dimensional set of predictors with known groups corresponding to gene regulatory networks [54]. We conduct a simulation study below which indicates that the LN-CASS prior substantially outperforms the sparse-group LASSO method [87] as implemented the authors' R package, and is excellent at recovering the true model parameters, even in a $p > n$ setting. We intend to develop an R package to allow easy implementation of the grouped LN-CASS method. We also show below that the LN-CASS method performs comparably to (in fact, slightly better than) common sparse classification methods, including the horseshoe prior, on a small dataset relating to graft-versus-host-disease, and we conduct two further case studies on problems from metabolomics and cancer genomics, benchmarking against common machine learning methods and the horseshoe prior.

4.3 Application to the GvHD data

In this section we apply the methodology outlined above to the GvHD dataset. The objectives are two-fold: to establish a predictive model of GvHD incidence and to conduct a case study to evaluate the performance of the LN-CASS method for model selection as compared with a number of other common Bayesian and frequentist methods. We now outline the main model we will deploy in this section. Let y denote the binary response vector and X denote the matrix of covariates. We use the grouped LN-CASS prior in the context of a robit regression model as our primary method. Let G_i denote

the set of parameter indices corresponding to the i^{th} group and let $G = \Cup_i G_i$. The full model specification is as follows.

$$y_i|\beta_0, \beta \sim \text{Bernoulli}(F_{\text{Cauchy}}(\beta_0 + X_i\beta)), \tag{4.19}$$

$$\beta_i|\lambda_{G_i}, \lambda_i \sim \mathcal{N}(0, \tau^2\omega_i), \tag{4.20}$$

$$\lambda_{G_i} \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda), \tag{4.21}$$

$$\lambda_i \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda), \tag{4.22}$$

$$\beta_0 \propto 1, \tag{4.23}$$

where

$$\omega_i = \begin{cases} \lambda_{G_i}^2(1 + \lambda_i^2) \text{ for } i \in G, \\ \lambda_i^2 \text{ otherwise,} \end{cases} \tag{4.24}$$

and F_{Cauchy} is the cumulative distribution function of a standard Cauchy random variable. This is a robit regression model, taking the t-distribution with one degree of freedom (i.e. the Cauchy distribution) as the error associated with the linear predictor. The specification of ω_i is explained below.

We fix the hyperparameters τ, μ_λ and σ_λ to the following values. τ represents the standard deviation of the 'slab' component of the prior and we choose to set it to 5, which allows for reasonably large regression coefficients but still provides some further, weaker regularisation in a manner akin to the elastic-net penalty [107]. This additional regularisation is necessary for this problem as the maximum likelihood estimates (via a classical logistic regression) for the parameters are infinite, and could explain the poor performance of the Horseshoe prior, whose tails are too heavy to provide enough regularisation at large parameter values. Similar problems where complete separation in logistic regression impedes the performance of the horseshoe prior have been identified [41, 71]. We fix μ_λ to a default value of zero, which leads to symmetric priors on

the indicator parameters λ_i, λ_{G_i} with equal mass close to zero and one. σ_λ controls the sharpness of the peaks close to 0 and 1 of the Logit-Normal distribution, with larger coefficients corresponding to greater concentration around the endpoints. We choose $\sigma_\lambda = 10$. The orange distribution in figure 4.4 (b) shows the the prior induced on the (group-level) regression coefficients β_i and the orange distribution in figure 4.3 (b) is the corresponding Logit-Normal prior on the relaxed inclusion indicator variable. We can see that the Logit-Normal distribution is similar in appearance to the Beta distribution, but it is more numerically stable due to having finite density near 0 and 1 and the fact that it admits a simple reparameterisation in terms of a standard normal random variable.

We now explain the specification of ω_i. The tree structure of the predictors is taken into account via a mixed-effects type parameterisation at the group level. Each regression coefficient in a particular group is the sum of a group-level coefficient and an individual coefficient. Specifically, we put

$$\beta_i = \beta_{G_i} + \eta_i, \tag{4.25}$$

and shrink both coefficients towards zero, but do so more strongly for the individual level predictor, η_i, by imposing a 'heredity' condition, as introduced in [45]. The following priors on β_{G_i} and η_i lead to the specification of ω_i in (4.24).

$$\beta_{G_i} \sim \mathcal{N}(0, \tau^2 \lambda_{G_i}^2), \tag{4.26}$$

$$\eta_i \sim \mathcal{N}(0, \tau^2 \lambda_{G_i}^2 \lambda_i^2). \tag{4.27}$$

This prior structure corresponds to a hierarchy in which we favour firstly zero coefficients at both the group and individual levels, then the situation in which each member of the group shares the same coefficient, and finally the situation in which each member of the group has a different coefficient. In the study at hand, we have the additional

benefit that the second case allows us to collapse the group to a single predictor (the parent node), since the value of each parent node is the sum of the values of its children. For example, it may be the case that the total CD4 cell count is sufficiently predictive of outcomes and we can shrink the coefficients for the individual subsets to zero.

Essentially this translates to a penalty on the cell subset granularity in the model.

4.3.1 Software and Monte Carlo sampling

All computations in this section are performed in R. MCMC sampling is conducted through the RSTAN package, an R interface to the probabilistic programming language STAN [89], which enables implementation of the No-U-Turn-Sampling (NUTS) variant of Hamiltonian Monte Carlo [52] for posterior sampling.

Hamiltonian Monte Carlo is an alternative to other common MCMC routines, such as Gibbs or Metropolis-Hastings sampling. Its motivation is the fuller exploration of the so-called 'typical set' in parameter space, i.e. the region of parameter space containing the bulk of the posterior probability. Intuitively, the main reason for its improved performance over Metropolis-Hastings or Gibbs methods is the fact that it explicitly accounts for posterior curvature, meaning that the 'guided diffusion' of the MCMC sampling procedure is more strongly directed toward the 'typical set'.

We now provide a brief overview of the mathematical details of HMC, following [17].

The goal of HMC, as with any MCMC procedure, is to draw samples from the posterior distribution over the parameters. To do this, the problem is cast as a physical problem in which the target distribution is represented as a distribution of particle positions, x, in p-dimensional parameter space. An artificial velocity vector, v, is introduced and is responsible for the stochasticity of trajectories as explained below.

The dynamics of the system are described by a Hamiltonian $H(x, v) = U(x) + K(v)$, with U a potential energy and K a kinetic energy, according to the equations

$$\dot{x} = \nabla_v H = \nabla K, \tag{4.28}$$

$$\dot{v} = -\nabla_x H = -\nabla U. \tag{4.29}$$

The Boltzmann distribution describes the probability density function over states (i.e. vectors (x, v)) for the system and is given by

$$P(x, v) = \frac{1}{Z} \exp\left(-\frac{H(x, v)}{T}\right), \tag{4.30}$$

$$= \frac{1}{Z} \exp\left(-\frac{U(x)}{T}\right) \exp\left(-\frac{K(v)}{T}\right) \tag{4.31}$$

with Z a normalising constant, and T a 'temperature' (taken to have units in which the Boltzmann constant is unity) which is usually fixed to 1 in HMC. Note that (4.31) implies independence of x and v.

We take

$$U(x) = -\log(\pi(x)\pi(x|\mathcal{D}), \tag{4.32}$$

$$K(v) = -\log\left(\exp\left(\frac{1}{2} v^\top \Sigma^{-1} v\right)\right), \tag{4.33}$$

which corresponds to taking the particle positions (parameter values) to be distributed according to the posterior probability density $p(x|\mathcal{D}) \propto p(x)p(x|\mathcal{D})$, and the velocities to be distributed multivariate normally with covariance matrix Σ and zero mean.

Pseudocode for the HMC algorithm is given in algorithm 1.

The result is a matrix of accepted particle positions which corresponds to a sample from the posterior distribution. Typically, (**??**) & (4.29) are solved with a leapfrog integrator.

Algorithm 1 Hamiltonian Monte Carlo

Inputs: x_0, iter, T
Initialise: $x_{\text{cur}} \leftarrow x_0$

for $i = 1, \ldots,$ iter **do**

$v_0 \leftarrow$ draw from $\mathcal{N}(\mathbf{0}, \Sigma)$ ▷ draw a new velocity
Initial conditions $\leftarrow (x_{\text{cur}}, v_0)$
$(x, v) \leftarrow$ solve (3.28)–(3.29) on $[0, T]$
$(x_{\text{new}}, v_{\text{new}}) \leftarrow (x(T), v(T))$ ▷ simulate forward for time T to get proposal
$U \leftarrow$ draw from $U(0, 1)$

if $\exp(H(x_{\text{cur}}, v_0) - H(x_{\text{new}}, v_{\text{new}})) > U$ **then** ▷ accept/reject proposal
$x_{\text{cur}} \leftarrow x_{\text{new}}$
end if

return x_{cur}

end for

A number of modified versions of the HMC algorithm have been developed. In particular STAN uses the No-U-Turn Sampler [52], which modifies Algorithm 1 to adaptively set T, the length of the simulated trajectories, at each iteration. The main advantage of HMC over other MCMC schemes is its efficiency. Because the 'random walk' aspect of the sampling procedure is replaced by deterministic trajectories precisely determined by the shape of the posterior distribution (through the potential energy), the posterior can be explored more efficiently and large moves in parameter space have a much higher probability of acceptance. In order to improve the performance of the leapfrog integrator, it is beneficial for estimated parameters to lie on similar scales and to be approximately multivariate normally distributed. One reason for this is that a fixed step size is used for all parameters, meaning that the sampler may struggle to efficiently explore multiple directions in parameter space with very different scales. Another reason is that transformation to multivariate normal prior distributions often helps to reduce the magnitude of the gradient of the log-posterior distribution, which appears on the right hand side of the underlying ODE and can lead to stiffness

which is troublesome for fixed step-size ODE solvers. Further discussion and empirical studies of the impact of reparameterisations on sampling performance can be found in [12,67]. For this reason, we employ a number of reparameterisations using latent variables which are marginally standard normal. For example, if λ_{raw} is standard normal, $\lambda = \text{logit}^{-1}(\mu + \sigma\lambda_{\text{raw}})$ is LogitNormal(μ, σ). We then sample from posterior distributions in λ_{raw} space and transform the posterior appropriately.

For comparison, we also implement a number of other statistical/machine learning techniques, all of which are available in the R package CARET [34].

4.4 Results for GvHD data

In this subsection we present the results of the analysis. Firstly, we focus on using the model (4.20)-(4.23) to generate a predictive model for GvHD incidence. Secondly, we compare the results to a number of common classifiers.

Figure 4.5 shows kernel density plots of the marginal posterior distributions for the regression coefficients. Posterior samples were obtained with STAN. Four Markov chains were run in parallel for 10,000 iterations each with the first half discarded as warm-up samples. Note that the number of iterations required for convergence is substantially reduced for HMC in comparison with other MCMC routines on account of the much larger rate at which HMC accepts proposed moves, however the computational cost associated with each iteration is typically substantially greater than the alternatives, such as the adaptive Metropolis algorithm. Chain convergence and mixing was diagnosed with the Gelman-Rubin $\hat{R}$ statistic [39], which was close to unity for all parameters.

We see that the majority of the marginal posteriors are strongly concentrated around zero. Notably, the effects of all CD8 subsets are the largest in absolute value and CD8 T-cell counts appear to be the strongest predictors of GvHD incidence. This is in agree-

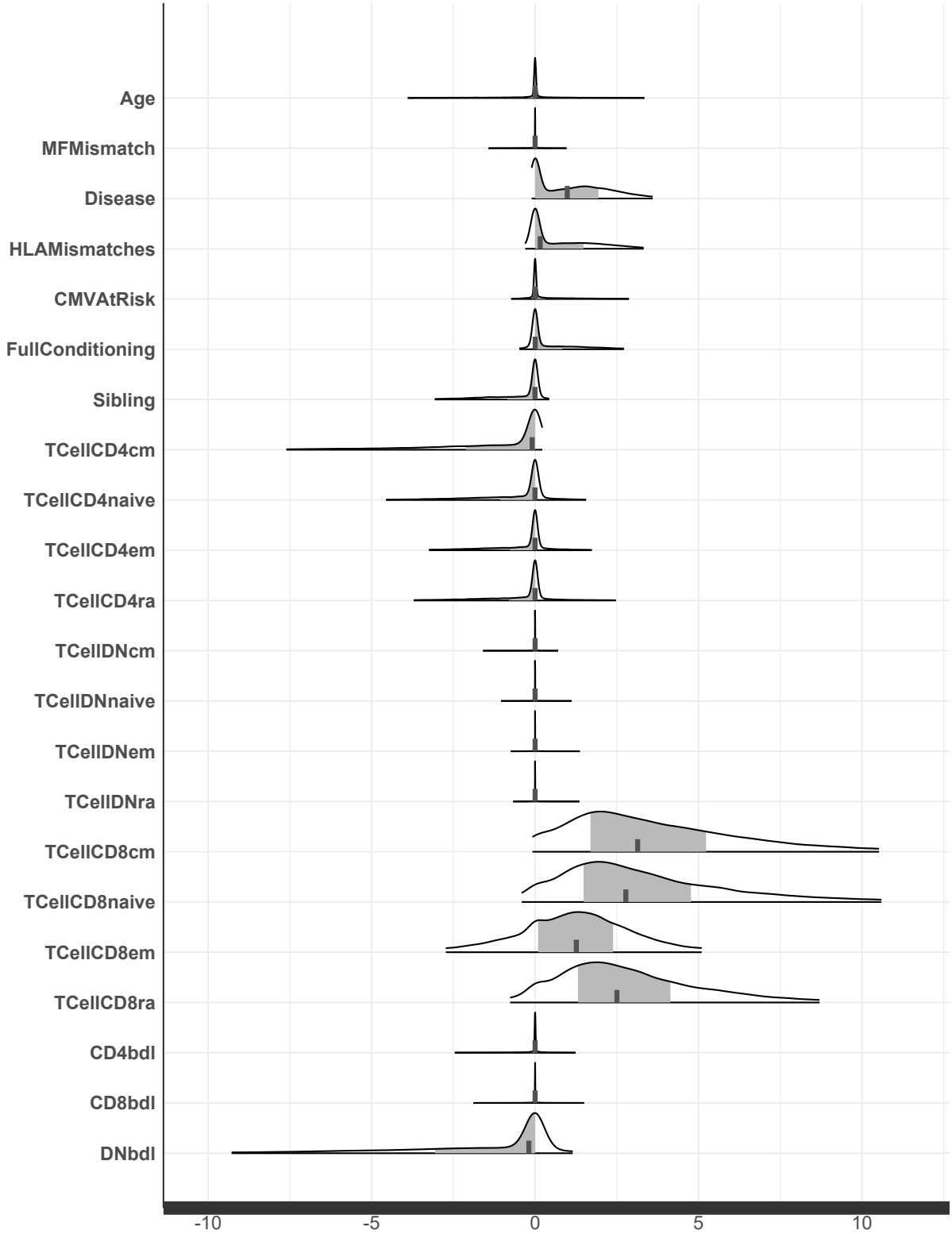


Figure 4.5: Kernel density plots of posterior distributions over the regression coefficients. Grey shaded regions are 50% credible intervals, computed via sample quantiles; tick marks are posterior medians.

ment with biological evidence that CD8 T-cells (also known as 'killer T-cells') are the primary mediators of GvHD [92]. Additionally, the hierarchical structure of the prior has successfully encouraged grouping of the CD8 subsets. There does however seem to be some evidence for a distinct effect of CD8 effector memory cells. The marking out of CD8 effector memory cells is again consistent with current biological knowledge. CD8 effector memory cells are activated CD8 cells, which arise transiently as part of the adaptive immune response which in the case of GvHD is directed towards host organs.

Additionally, there appears to be some evidence that the probability of developing GvHD is increased if the patient is being treated for lymphoma over myeloid leukaemia.

Evidently, the uncertainty in the parameter estimates is large and so the results should be interpreted with caution. Given the very small dataset, this is to be expected.

While one of the commonly cited benefits of the Horseshoe prior is its guarantee of unimodality in the posterior, we find that STANs sampler is able to explore the multimodal distribution resulting from the LN-CASS prior very well, as evidenced by the $\hat{R}$ statistic being close to unity – $\hat{R}$ is a measure of the similarity of multiple MCMC chains and would be far from 1 if chains were sticking in local modes. For higher-dimensional problems, the possibility of multimodal posteriors may cause more substantial sampling difficulties, and this question is worthy of further investigation.

Chain diagnostics and posterior predictive checks

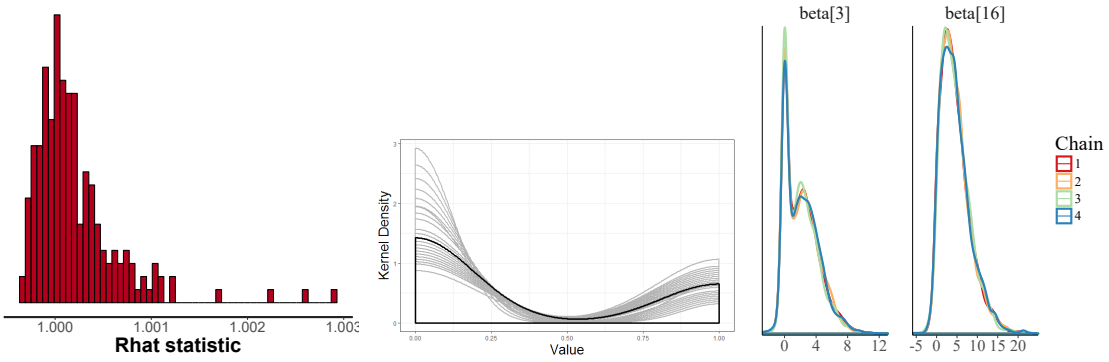
Chain convergence was checked via the Gelman-Rubin $\hat{R}$ statistic. A histogram of the $\hat{R}$ values for each parameter is shown in figure 4.6 (a). Values close to 1 indicate that the 4 chains are well-mixed and are similar in the sense that they represent samples from the same distribution. Figure 4.6 (b) shows that samples from the posterior predictive

distribution are broadly similar to the data used to train the model. This similarity helps to verify that the code is producing the expected output. We also present overlaid kernel density estimates for two of the parameters in figure 4.6 (c), showing that each chain converged to the same distribution. In particular, the multimodality of the posterior distribution for `Disease` (`beta[3]`) is captured by all 4 chains and is not an artefact of the chain sticking in separate regions of parameter space in separate chains.

Estimating out-of-sample predictive performance

For estimating the out-of sample predictive performance of the developed models, we employ repeated 5-fold cross-validation. We use the area under the receiver operating characteristic curve (AUC) as a performance measure due to the imbalance in positive vs. negative cases. The AUC, also known as the c-statistic or concordance index, is interpretable as the probability of correctly distinguishing a pair of cases of different class labels. We choose 32 repeats so that the procedure is neatly parallelised on a 16-core server instance. The cross-validation is stratified so that each fold contains representative proportions of positive and negative samples.

The results are shown in figure 4.7. All methods perform similarly, with the median AUC being approximately equal across methods. The horseshoe is the method with the least variance, while LN-CASS has the best mean performance. Clearly, all methods perform reasonably poorly on average and there is considerable variability in cross-validated performance. All methods perfectly classify at least one holdout set, and perform worse than random guessing a reasonable proportion of the time. This variability is a result of the size of the dataset and the very sparse coverage of the cell-count predictor space and suggests that we should not place too much stock in the parameter inferences or predictive power of the model – more data is required to build a reliable predictive model of GvHD incidence.



(a) (b) (c)

Figure 4.6: Diagnostic plots for the Markov chains. (a) Histogram of $\hat{R}$ values; (b) Samples from the posterior predictive distribution (grey) vs. the true distribution (black); (c) two overlaid posterior kernel density estimates.

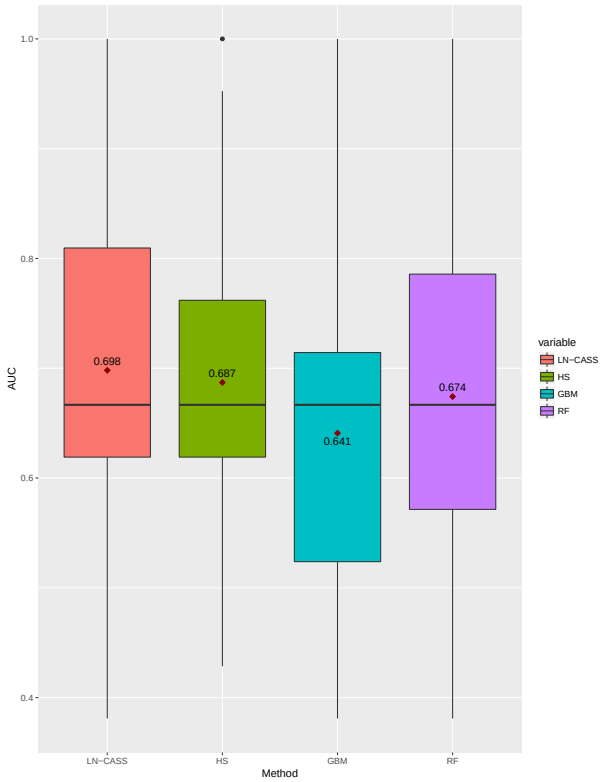


Figure 4.7: AUC for each method tested for 32×5 cross-validated samples. Diamonds indicate means and solid black lines indicate medians.

However, coupled with the results of the simulation study & additional case studies below, we can have reasonable faith that the LN-CASS method performs well in problems with group structure and elsewhere. Data collection is ongoing as part of the project of our collaborators. The results of this section and the simulation study suggest that the LN-CASS method will be a reasonable approach to the problem.

4.5 Simulation study

We now conduct a simulation study to assess the performance of the LN-CASS prior in three linear regression settings. We simulate data from the ordinary linear regression model

$$y = \beta_0 \mathbf{1} + X\beta + \varepsilon, \tag{4.34}$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{4.35}$$

with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$.

We take $n = 100$, and simulate data for $p = 20, 70, 120$. The covariate matrix X is drawn from a unit latin hypercube and the coefficient vector β is composed of $p/5$ groups of 5. Some groups are chosen to have all zero coefficients, some to have constant non-zero coefficients, some to have noisy constant non-zero coefficients, and some to have non-zero coefficients of differing sizes. Table 4.2 contains the details.

We compare the results of the grouped LN-CASS prior to the following shrinkage methods: the horseshoe (HS), the LASSO and the sparse group LASSO (SGL). We also include results from an ordinary least squares fit (OLS), i.e. the unpenalised maximum likelihood estimate, for comparison in the two cases where $p < n$. All computations are performed in R. The horseshoe estimates are obtained with the `rstanarm` package [90];

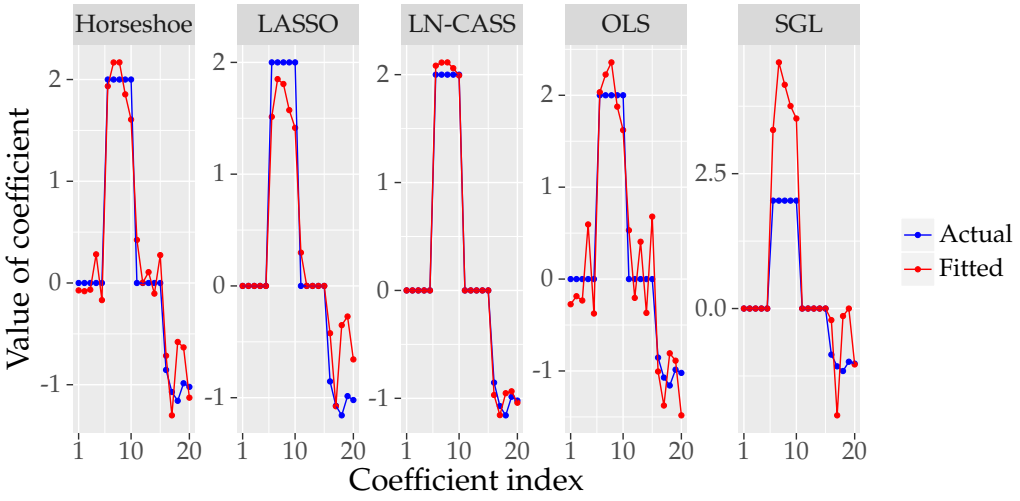
	Zero groups	Constant groups	Noisy groups	Disparate groups
$p = 20$	$\beta_{G_1}, \beta_{G_3} = \mathbf{0}$	$\beta_{G_2} = \mathbf{2}$	$\beta_{G_4} = \mathbf{-1} + \mathcal{N}(0, 0.1^2)$	
$p = 70$	10 zero groups total.	$\beta_{G_4}, \beta_{G_7} = \mathbf{2}$	$\beta_{G_8} = \mathbf{-1} + \mathcal{N}(0, 0.1^2)$	$\beta_{G_{14}} = (-0.5, -0.5, 3, -0.5, -0.5)$
$p = 120$	20 zero groups total.	$\beta_{G_6} = \mathbf{1}$	$\beta_{G_{12}} = \mathbf{2} + \mathcal{N}(0, 0.1^2)$	$\beta_{G_{18}} = (-1, -1, -1, -2, -1)$ $\beta_{G_{19}} = (0.5, 0.5, 0.5, 2, 0.5)$

Table 4.2: True coefficients for each of the simulation experiments. Bold numbers represent constant vectors of length 5 and $\mathcal{N}(0, 0.1^2)$ denotes a vector of 5 samples from a normal random number generator with mean 0 and standard deviation 0.1.

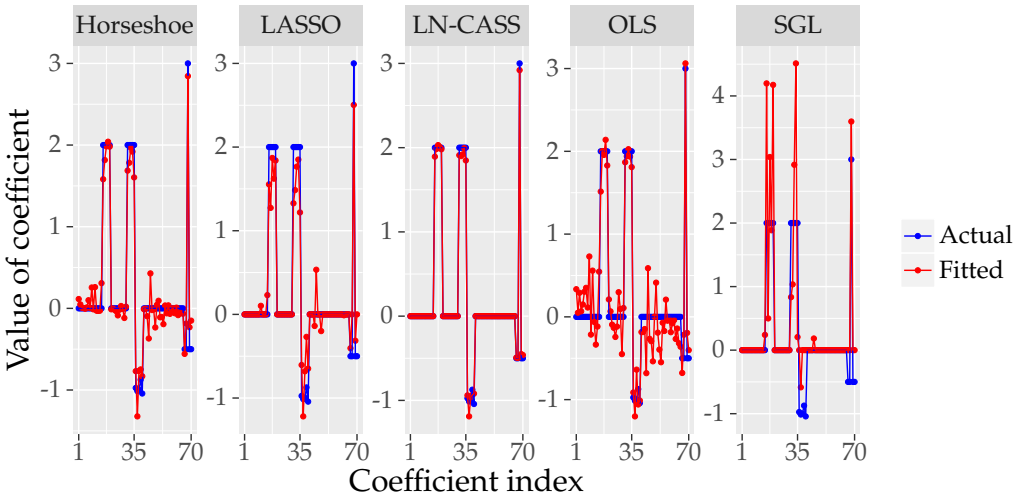
LASSO estimates are obtained with the `glmnet` package [32] and the hyperparameter is estimated via 10-fold cross-validation; SGL estimates are obtained with the SGL package [86], again using 10-fold cross-validation to select hyperparameters. Figures 4.8 and 4.9 plots the true coefficients against those estimated by each of the methods.

For the Bayesian methods, we take the posterior median as a point estimate for the purposes of display, though the full posteriors are, of course, available for subsequent analysis.

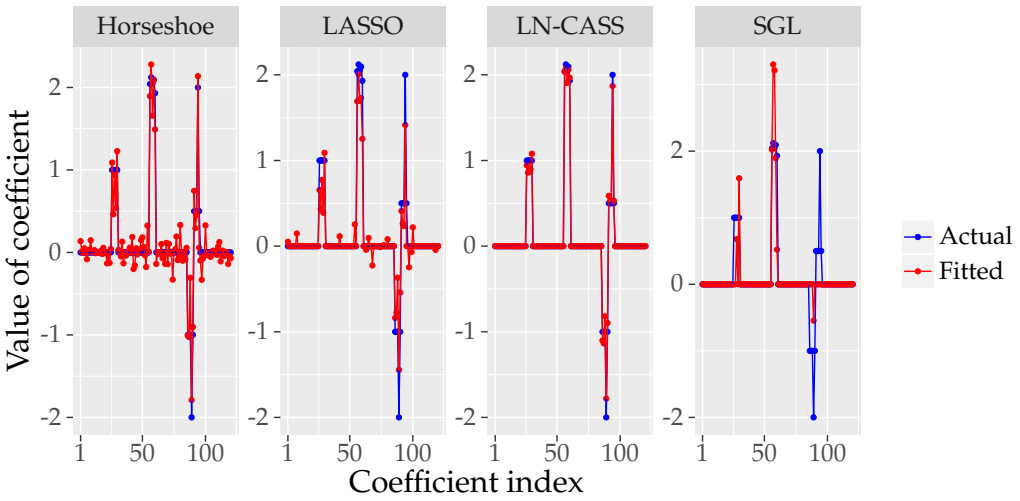
We see that the LN-CASS prior performs very well, correctly identifying every zero-group while simultaneously providing parameter estimates for the non-zero coefficients very close to their true values, even in the $p > n$ example. The horseshoe performs well across settings but does not induce the same level of sparsity, meaning that ad-hoc methods are required for variable selection, for example applying a threshold to the absolute value of parameters or selecting coefficients based on their local-shrinkage parameters [21]. While the LN-CASS prior does not produce estimates of exactly zero, it is much clearer which parameters should be discarded if genuine variable selection is an objective. For the purposes of prediction, we would argue for retaining all predictors with their shrunk estimates. Both the LASSO and the sparse group LASSO provide parameter estimates which are exactly zero but both suffer from the problem that as the penalisation hyperparameter increases, non-zero effects are also shrunk towards zero, meaning that a trade-off is required between sparsity and the accuracy of parameter estimates. This is particularly clear in the case of the SGL, in which non-zero effects



(a)



(b)



(c)

Figure 4.8: Results of the simulation study. (a) $p = 20$, (b) $p = 70$, (c) $p = 120$

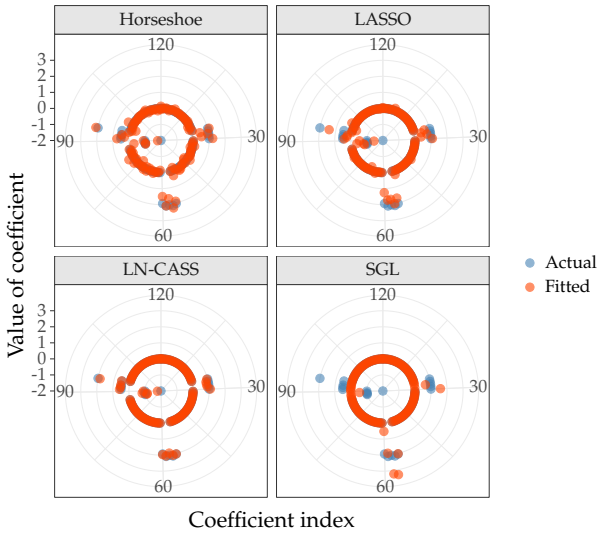


Figure 4.9: Estimated and true parameters for the $p = 120$ case, visualised in polar coordinates.

	HS		LASSO		LN-CASS		OLS		SGL	
	MAE	AUC	MAE	AUC	MAE	AUC	MAE	AUC	MAE	AUC
$p = 20$	0.196	1	0.223	0.99	**0.043**	1	0.317	1	0.643	0.95
$p = 70$	0.111	0.98	0.13	0.941	**0.017**	**1**	0.22	0.94	0.31	0.791
$p = 120$	0.096	0.9755	0.075	0.9702	**0.015**	**1**	NA	NA	0.145	0.7

Table 4.3: Mean Absolute Error and Area Under the ROC Curve (explained in text) for the simulation study. Best performance in bold.

are routinely over- or under-estimated in an attempt to strike the correct balance. The LN-CASS prior appears to strike this balance automatically, allowing large coefficients to be estimated accurately while correctly identifying zero components.

We compare the methods quantitatively in table 4.3 using the mean absolute error (MAE) of the estimated parameters from the true parameters and the area under the ROC curve. In order to assess the quality of the ordering of the coefficients, i.e. the extent to which truly small parameters are estimated to be small and large coefficients to be large, we calculate the area under the ROC curve with the 'response' being a binary vector with an entry of 0 if the true coefficient is 0, and 1 otherwise. The AUC

reflects the trade-off between the false-positive and true-positive rates as the threshold on the absolute values of the parameters for 0/1 classification is varied. We see that the LN-CASS method has an AUC of 1 in every situation, meaning that it perfectly separates zero from non-zero coefficients for some threshold value. Additionally, the mean absolute error of the estimated from the true parameters is considerably smaller for the LN-CASS method compared with any other method tested. Surprisingly, the sparse group LASSO is the worst performing method in each case, while the horseshoe and the LASSO are comparable in performance.

4.6 Case study: Microarray data

This case study focuses on the well-known Colon dataset of Alon et al. [2]. The dataset consists of measurements of the expression levels of 2000 genes in 62 subjects, with the response variable being an indicator of Colon cancer incidence, representing a typical $p \gg n$ problem in the biological/medical sciences. We compare the performance of logistic regression, with LN-CASS priors on the coefficients, to LASSO, Random Forest and Neural Network classifiers. We perform leave-one-out cross-validation (LOOCV) and compute the AUC across the left out samples in order to compare the estimated out-of-sample predictive accuracy of each method. In order to reduce the bias of the AUC estimates, we randomly remove an observation of the opposite class in each fold so that the class proportions are identical across folds.

We preprocess the data by first log-transforming and subsequently standardising (i.e. subtracting the mean and dividing by the standard deviation) the expression level of each gene. We then screen the genes via preliminary univariate Wald tests and select the 500 genes with the largest Z-scores in absolute value, leaving us with a predictor matrix consisting of the expression levels of 500 genes in 62 tissues which act as the

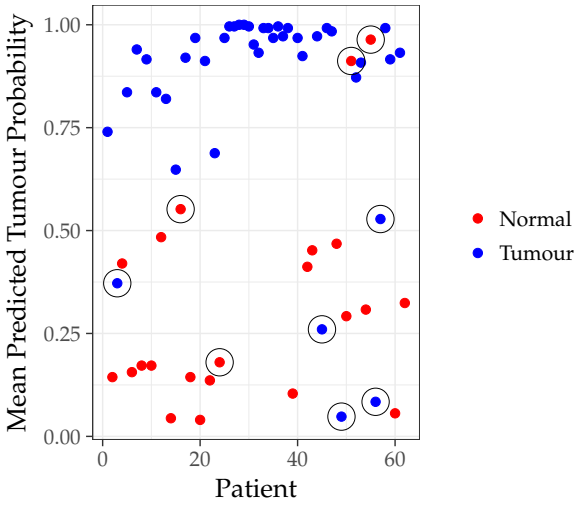


Figure 4.10: Mean predictions and observed outcomes from the LN-CASS model for the microarray data. Circled points have been identified as potentially mislabelled by [2, 15]

input to all subsequent models.

The model applied uses no hierarchical structuring of predictors, it is simply a Bayesian logistic regression with LN-CASS priors on the coefficients.

The pooled LOOCV AUCs for each method are as follows: LN-CASS, 0.904; Neural Network, 0.8898; Random Forest, 0.8892; LASSO, 0.858. LN-CASS performs the best, but again all of the methods perform well and there is not a substantial difference between the estimated out-of-sample performance of each method.

Interestingly, there is some biological evidence for class-mislabelling, i.e. samples being incorrectly marked as either tumour or healthy [2, 15] in the colon dataset. According to Bootkrajang & Kabán [15], there are nine such samples. Figure 4.10 shows the mean posterior prediction for each subject with these 'suspicious' subjects circled. Clearly, there is reasonable agreement based on a visual inspection of the plot between the potentially mislabelled samples and those suggested by visual inspection of the LN-CASS model predictions, especially in the case of false negatives.

4.7 Case study: steroid metabolomics and adrenal tumour malignancy (hierarchical GAM)

In this final case study, we apply a hierarchical version of the LN-CASS prior to clinical data on the concentrations of metabolites in the urine of patients with two different classes of adrenal tumour. The task is to predict the tumour type based on the metabolites, and to do this we used a generalised additive model (GAM) with logit link. The implementation of the prior in a hierarchical fashion here was strongly inspired by a recent paper by Griffin and Brown [45].

The GAM we implement models the effect of each covariate as the sum of linear basis functions. We impose a hierarchy through the LN-CASS prior which favours firstly the complete removal of a covariate from the model, then inclusion of a purely linear effect, and finally allows each of the basis functions to be used to better approximate a non-linear effect. We discuss the details of the model after first outlining the data.

The dataset consists of 158 measurements of 32 covariates [5] collected as part of the EURINE-ACT study, with 45 positive cases (malignant adrenal tumours). All of the covariates are measurements of steroid concentrations in urine samples taken from each of the patients. There is a small proportion of missing data (up to 7% of a covariate's measurements), which we impute using multiple imputation via the `mice()` function in R [18]. We then $\log(1+x)$ transform all of the data because many of the predictors spanned several orders of magnitude. We subsequently scaled all covariates to lie in the interval $[0, 1]$.

The LN-CASS prior was used as part of a hierarchical generalised additive model (GAM). The set up is a logistic regression problem in which we suspect that some covariates may have nonlinear effects, but we wish to 'let the data decide' whether including such effects is worthwhile for the purposes of prediction.

Generalised additive models are extensions of generalised linear models in which the linear predictor is replaced with a sum of nonlinear covariate effects, i.e.

$$g^{-1}(y_i) = f_0 + \sum_{i=1}^{p} f_i(x_i), \tag{4.36}$$

with g being a link function, f_0 a constant, and the f_i functions to be learnt.

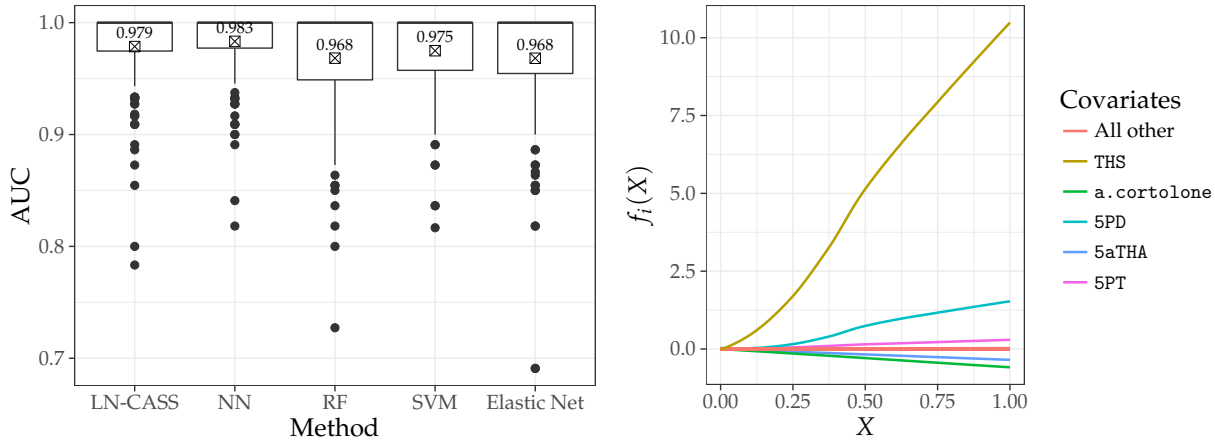
We model the f_i as piecewise linear functions with some pre-specified number of knots, M. Consider, without loss of generality, the covariate space $\chi = [0,1]^p$. The functions

$$\varphi_k(x) = \begin{cases} 0, & x \leq x_k \\ \frac{x - x_k}{1 - x_k}, & x > x_k \end{cases}, \; x \in [0,1], \; k \in 1, \ldots, M, \tag{4.37}$$

form a basis for the space of piecewise linear functions with M knots on $[0,1]$, which are 0 at the origin. We therefore represent each f_i as a linear combination of these basis functions,

$$f_i(x_i) = \sum_{k=1}^{M} \omega_{k,i} \varphi_k(x_i). \tag{4.38}$$

The weights ω_k are the subject of the hierarchical complexity shrinkage procedure, and the structure is very similar to that employed by the grouped LN-CASS prior. The motivation is that, if the ω_{k_i} are all 0 for a given, i, the covariate has no effect, if we allow $\omega_{1,i}$ to be non-zero, we obtain a linear effect, and if other weights are allowed to be non-zero we obtain a piecewise linear effect. This is the complexity hierarchy we wish to impose.



(a) (b)

Figure 4.11: Metabolomics case study. (a) Boxplots of AUCs for each method computed via 16×10-fold cross-validation; (b) estimated mean functions f_i from the LN-CASS hierarchical GAM. Functions have been smoothed for presentation purposes with a LOESS smoother using a small span.

Thus, the prior on the weights is as follows

$$\omega_{1,i}|\lambda_{1,i} \sim \mathcal{N}(0, (\lambda_{1,i}\tau)^2), \tag{4.39}$$

$$\lambda_{1,i} \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda), \tag{4.40}$$

$$\omega_{k,i}|\lambda_{1,i}, \lambda_{k,i} \sim \mathcal{N}(0, (\lambda_{1,i}\lambda_{k,i}\tau)^2), \text{ for each } k = 2, \ldots, M \tag{4.41}$$

$$\lambda_{k,i} \sim \text{LogitNormal}(\mu_\lambda, \sigma_\lambda), \text{ for each } k = 2, \ldots, M. \tag{4.42}$$

Again, by rewriting this in terms of logit transformed normal random variables, we obtain a multivariate normal prior on our parameters of interest.

We compare the classification performance of our hierarchical GAM with the performance of the following methods: Support Vector Machine (SVM), neural network (NN), random forest (RF) and elastic net. Classification performance was measured using the mean AUC over $16 \times 10-$fold cross-validated runs. The results are presented in fig. 4.11 (a).

All of the methods perform comparably in terms of out-of-sample predictive perfor-

mance, with the neural network performing the best and LN-CASS second in terms of both the mean and variability (inter-quartile range) of cross-validated AUCs. We are not aware of an appropriate and well-established statistical test to formalise the comparative performances of each method given the unequal variances, clear non-normality, and obvious dependency between samples for a given method. However, the Kruskal-Wallis test with a *post hoc* Dunn test (and appropriate multiplicity correction) provides a non-parametric test for stochastic dominance (i.e. the tendency of values from one distribution to be larger than values from the other). We used two multiplicity corrections, both of which account for positive dependency (i.e. the tendency of large AUCs to be correlated within cross-validation folds). Using the Benjamini-Hochberg [8] correction, the only null hypotheses to be rejected at 95% significance levels were that the distribution of AUCs for the neural network stochastically dominates those for the Elastic Net and the Random Forest (adjusted p-values 0.0344 and 0.0203, respectively). Using the Benjamini-Yekutieli [9] correction, no null hypotheses were rejected; that is, no significant differences were found between the distributions in terms of stochastic dominance. Note that the Benjamini-Yekutieli correction allows for arbitrary dependencies.

The results suggest that the out-of-sample performance of the hierarchical GAM with LN-CASS prior is comparable with that of state-of-the-art machine learning methods, at least for this problem. We argue that this performance, in conjunction with the accuracy with which LN-CASS recovers 'true' parameters and offers more classically interpretable results make it a valuable addition to the shrinkage and regularisation toolbox for applied scientists.

The recovered effects for each of the metabolites are presented in fig. 4.11 (b), as estimated from the full dataset. Clearly, the dominant predictor is `THS` which is in agreement with the original study, as are the influential roles of both `5PD` and `5PT`. We believe that the ability of the hierarchical GAM to produce plots such as these constitutes a

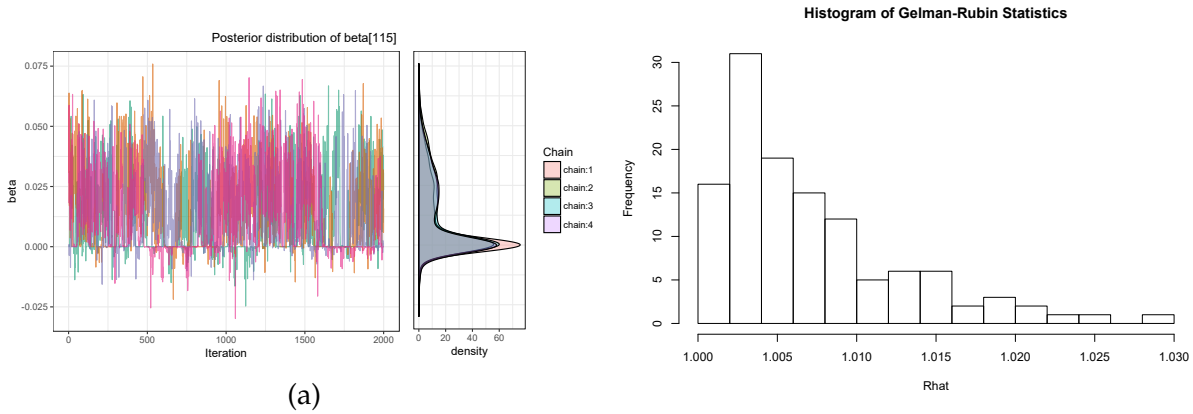
considerable advantage over the machine learning methods tested and highlights the ability of LN-CASS to generate not only strong predictive models, but to be used as an exploratory tool for the generation of hypotheses for future study.

4.8 Further check of mixing and multimodality

Multi-modal posterior distributions are a common complication of spike-and-slab type prior distributions. Most often when multi-modality occurs, the marginal posteriors for some parameters have a mode centred at zero and a second mode elsewhere. This multimodality is particularly prevalent in, for example, linear models with interactions. Random-walk based samplers such as Metropolis-Hastings or Gibbs samplers can experience difficulties effectively exploring multi-modal posteriors. It is often claimed that the No U-Turn Sampler deployed by Stan is much better at exploring multimodal posterior distributions than other common MCMC samplers.

We verify that this is indeed the case in our application with a simulated dataset and a linear model with interactions. The code at `https://github.com/willthomson1/RS-Interface-co` contains the full details of the model.

We use a simple visual check to examine the mixing of the sampler. We run four randomly initialised chains for 4000 iterations each, discarding the first 2000 samples as burn-in. Subsequently, we examine kernel density estimates of the marginal posteriors for each parameter and each chain. Whenever there appears to be a bimodal marginal posterior distribution in any of the chains we check that the other chains had also explored both modes. This was always the case, suggesting that the sampler is able to consistently and robustly escape local modes. Figure 4.12 (a) shows trace plots and kernel density estimates for each chain of a representative multi-modal marginal posterior distribution. The full posterior can be explored using the available code. Figure



(a) (b)

Figure 4.12: (a): Trace plots and kernel density estimates for one of the parameters with a multi-modal posterior distribution for four separate chains. (b): Histogram of the Gelman-Rubin statistic $\hat{R}$ for the parameters subject to the LN-CASS prior.

4.12 (b) shows a histogram of the Gelman-Rubin statistics for the parameters subject to LN-CASS priors. They are all close to one, suggesting that mixing is sufficient.

4.9 Discussion

In this chapter we have developed, tested and applied a Bayesian shrinkage method for data with grouped predictors and situations in which a hierarchical complexity structure is desired. The results in benchmark studies against other methods are encouraging. We have shown that its ability to produce generalisable predictive models is comparable to common machine learning methods on three datasets of biological interest. Additionally, we have demonstrated with a simulation study the ability of our method to recover ground-truth parameters, even when the number of parameters is larger than the number of data points. While this can, in general, not be ensured if the regularisation criteria are not met, the results are encouraging. A study of the efficiency of the method would be a logical next step.

Studying other theoretical properties of the prior is a potentially fruitful avenue for

further research. In particular, it would be interesting to obtain some results that would help to inform the choice of the hyperparameters of the LN-CASS prior, i.e. to inform the precise shape of the Logit-Normal distribution for a given problem. For example, how does the choice of $\mu_{\text{res}}\lambda$ influence the expected number of non-zero coefficients?

Evidently, the estimated out-of-sample predictive performance of all classifiers tested on the current GvHD dataset was relatively poor, generally achieving around 65-70% concordance. One obvious reason for this is that the dataset is relatively small and the predictor space is sparsely covered, meaning that individual observations have too large an influence on the parameter estimation procedure. We alleviated this problem to some degree by implementing a robust form of logistic regression, but the variability in the cross-validated estimates of out-of-sample performance remained high, and this suggests that a larger dataset is required. The Bayesian framework, however, has the advantage that we can clearly see the uncertainty associated with our estimates. While CD8 T-cells were identified as being important predictors, the posterior distributions are very diffuse, suggesting that the data do not really contain enough information to estimate the CD8 T-cell parameters with any reasonable degree of certainty. Additionally, we saw multimodal posteriors for some predictors, which again is a consequence of the small dataset. Small datasets mean flatter likelihoods, and shrinkage priors such as the horseshoe or LN-CASS have the express purpose of amplifying any likelihood near zero considerably, while leaving likelihood further from zero almost unaltered. Despite the predictive performance of the model, the identification of CD8 T-cells as the most important predictors of GvHD is in line with current biological knowledge, as is the differential effect of CD8 effector memory cells.

Our prior requires the choices of three hyperparameters, although we contend that they are more interpretable than those required for other Bayesian shrinkage methods. The three hyperparameters required correspond to, firstly, the standard deviation of the 'slab' component, which we refer to as τ; for standardised predictors, a default

value of $\tau = 5$ has been sufficient for all of our applications because it essentially provides a vague Gaussian prior for non-zero coefficients. Secondly, the parameters of the Logit-Normal distribution (figure 4.3) must be specified; we refer to these parameters as μ_λ and σ_λ. μ_λ can be chosen based on our prior beliefs about the probability of a zero coefficient, and in our experience does not require much tuning; the median of the Logit-Normal distribution is given by $\text{sigm}(\mu_\lambda)$, where $\text{sigm}(\cdot)$ is the logistic sigmoid function. Thus, if we believe *a priori* that each coefficient has a probability p of being non-zero, we simply set $\mu_\lambda = \text{logit}(p)$. σ_λ simply controls the quality of the approximation to the spike-and-slab prior, with larger values corresponding to better approximations. We have used a default value of $\sigma_\lambda = 10$ throughout the paper; results are not sensitive to increases in this value.

The final key advantage of the LN-CASS prior is intuitive nature with which it generalises to problems with a hierarchical complexity structure. This allows finer control of what exactly we mean by a 'complex' model, and what we mean by a desirable model – our example of using a generalised additive model for studying the metabolomics data above illustrates this point. In that case, we imposed a hierarchical complexity structure: no effect $\rightarrow$ linear effect $\rightarrow$ nonlinear effect. In the simulation study, we favoured a complexity structure: no effect $\rightarrow$ shared group effect $\rightarrow$ individual effect. These hierarchies are accomplished simply by propagating the value of the Logit-Normal random variable through each layer and taking its product with a new Logit-Normal random variable.

We note that the prior is particularly amenable to problems in which a hierarchical complexity structure is desired, by which we mean problems in which simpler models are nested within more complex models. The simplest case is the domain of the majority of the shrinkage/regularisation literature: models with fewer parameters are nested within models with more parameters. However, there are other problems with similar properties; linear models are nested within nonlinear models, models with some pre-

dictors sharing coefficients are nested within models in which each predictor has its own coefficient. One possible area of application is in multi-state survival modelling. Multi-state models describe transitions between disease states by distinct hazard functions, which may be difficult to fit with a small sample size. One might expect that the effects of many covariates remain fairly similar regardless of the state, for example age. Thus the LN-CASS prior could be used to introduce a 'soft' constraint, encouraging but not enforcing covariates to share a parameter across hazard functions. This would essentially involve placing a grouped LN-CASS prior on the regression coefficients (as in the simulation study), with the groups corresponding to covariate effects.

As with most Bayesian methods, the main obstacle to the implementation of this methodology is the computational burden of MCMC sampling. Recent developments have made this procedure much more straightforward to implement and much faster [52, 89]. However, for large problems this computational burden is likely to be too large to compete with the much faster frequentist and machine learning methods available. Approximate Bayesian methods offer more computationally tractable alternatives to MCMC sampling, and would be an interesting avenue of future research for this problem and allow its scalability to large problems. In particular, nonparametric variational inference [40] appears to be the most reasonable direction, since it is able to deal both with multimodal posterior distributions and non-conjugate prior distributions.

One concern with spike-and-slab type inference procedures is the presence of multi-modal posterior distributions and the subsequent difficulty of some samplers to sample effectively from the posterior distribution, due to them becoming 'stuck' in local modes. We checked that the sampler we employed was able to effectively explore multi-modal posterior distributions in a linear model with interactions, a problem that is particularly prone to multi-modal posteriors. We found that whenever multi-modal posteriors appeared, they were effectively explored by multiple MCMC chains, and the Gelman-Rubin statistic [39] revealed that samples were consistent across chains.

The LN-CASS method does not inherently provide 'hard' variable selection, i.e. completely removing variables from the model, in the ilk of the LASSO. We advocate using the full model (i.e. including all predictors) for making predictions wherever possible, and using the absolute values of estimated parameters as variable importance measures for identifying the most important predictors for the purposes of hypothesis generation and/or obtaining biological insight. However, particularly in clinical/diagnostic circumstances, hard variable selection is useful to reduce the burden on clinicians/diagnosticians in collecting relevant data for utilising the model at the point of care.

A variety of applicable procedures for hard variable selection in Bayesian shrinkage models are available in an excellent review by Vehtari et al. [97]. One way in which variable selection could be achieved is by recursively eliminating features based on performance against a bootstrapped dataset or a dataset simulated from the posterior predictive distribution, i.e. the features are ordered by the size of their posterior median estimates and are eliminated one-by-one with a loss function calculated at each step. The mean performance of each sub-model is then used as the basis for model selection, perhaps by selecting the model with the smallest number of parameters such that mean performance is at most one standard error from the minimum – this is similar to the strategy employed by Hastie, Tibshirani et al. in the glmnet package for R [32] to estimate the penalisation hyperparameter. This is akin to selecting a threshold value at which to make a zero/non-zero choice for the parameter values as a post-processing step.

The potential multimodality of some of the posterior distributions also suggests that the posterior distribution is capturing a variety of feasible models – some of which include a given predictor and some of which do not. If the posterior distribution contains samples from these different models, it should contain clustered samples in parameter space and so applying data clustering methods to the posterior samples may help to

extract different candidate models for further exploration.

To summarise, we have presented a flexible tool for performing regularised Bayesian regression in a variety of settings, which allows one to construct problem-specific penalties on model complexity. The performance on out-of-sample data is typically at least as good as common machine learning methods, but the prior allows the use of classical statistical models which can be interpreted simply by applied biomedical scientists.

CHAPTER 5

CONCLUSIONS

5.1 Summary of findings and suggestions for future work

In this thesis, we have considered three problems related to single-cell data. In chapter 2, we presented and analysed a model for the evolution of heterogeneous age-structure population of cells undergoing proliferation and death in the presence of a trophic resource which supports both cellular survival and proliferation. In chapter 3, we developed a method for characterising the evolution of cellular phenotypes directly from time course data. In chapter 4, we presented a Bayesian method for regularising parameter estimates in predictive diagnostic models using typical single-cell data.

Chapters 2 and 3 were fundamentally concerned with attempting to characterize population heterogeneity within a population of evolving cells after some perturbation to their environmental conditions. The approach was to focus on the distribution of cellular properties of the population over time rather than, for example, the trajectories of individual cells. Such descriptions fall into the general class of population balance equations.

In chapter 2, the focus was on heterogeneity with respect to cell age. We presented a

model describing the change in the age-distribution of compartments of proliferating and resting cells where both recruitment into the cell cycle and survival in the resting state were dependent on the presence of an external resource. We examined the steady states of the model, deriving a condition for the existence of a non-trivial steady state and characterising the stability of the trivial steady state as a function of the probability distributions associated with ages of recruitment to and completion of the cell cycle. Additionally, we deployed a numerical scheme for the solution of the system and compared the numerical results to our analytic results.

To continue the work of chapter 2, we made initial explorations into the use of relative entropy methods for characterising the non-trivial steady states of the system. These methods offer a potentially fruitful avenue for further work in this area. Of course, the ultimate goal of this sort of modelling is to compare the outputs to experimental data. While methods do exist to estimate the parameters of age-structured models and structured population models more generally, they are computationally expensive and are limited to only one or two dimensional structuring variables.

With this in mind, in chapter 3 we set about developing ways to explore experimental data from single cell time course experiments through the lens of a fairly general population balance model. Borrowing ideas from equation-free modelling and diffusion maps, we first estimate the probability density of the data with a Gaussian mixture model. Subsequently, we treat the estimated mixture components as basis functions in which to expand the solution to a population balance equation. This leads to a way in which we can compare the temporal evolution of each mixture component to the time-varying weights of the mixture model.

The results of this process allow for a variety of useful ways to interrogate the data – a measure of cellular development (*pseudotime*), the ability to predict population distributions at previously unseen timepoints, and a new method of dimensionality reduction which reflects the developmental stages of the individual cells.

We trialled the method on three datasets – one synthetic and two containing real biological data. In all cases, the method produced apparently reasonable pseudotime estimates and captured the dynamics quite well.

As with most methods using Gaussian kernels, we found that the bandwidth of the kernel has a substantial impact on the quality of the dimensionality reduction coordinates in particular. Establishing methods to set the bandwidth automatically is an interesting direction for the future.

Further, establishing ways in which to incorporate more stringent constraints on the types of population balance operators covered by the method is important to allow the incorporation of prior knowledge about the process at hand and allow the effects of intracellular dynamics, stochasticity of expression and population growth/decay to be prised apart.

In the final chapter, we focused on a Bayesian statistical method for building predictive models of disease using aggregated single-cell data relating to the levels of expression of surface proteins in T-cells, gene expression in the colon cancer cae study, and urine metabolites.

The method appears to perform relatively well and produces sparse, interpretable models with uncertainty estimates. We also experimented with ways to incorporate various hierarchical complexity constraints using the prior – building nested models that are automatically encouraged to be as simple as possible, yet flexible enough to allow more complexity if the data support it.

Future work should focus on scalability to larger datasets – while the advantages of full uncertainty estimates and straightforward specification of more complex sparsity structures are useful in applications, the computational expense of sampling-based methods is a potential barrier to the widespread use of such methods in applications.

LIST OF REFERENCES

[1] B. ALBERTS, A. JOHNSON, J. LEWIS, M. RAFF, K. ROBERTS, AND P. WALTER, *Molecular Biology of the Cell, 4th edition*, Garland Science, New York, 2002.

[2] U. ALON, N. BARKAI, D. A. NOTTERMAN, K. GISH, S. YBARRA, D. MACK, AND A. J. LEVINE, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proceedings of the National Academy of Sciences, 96 (1999), pp. 6745–6750.

[3] S. AN, L. MA, AND L. WAN, *Tsee: an elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell rna sequencing data*, BMC genomics, 20 (2019), p. 224.

[4] D. F. ANDREWS AND C. L. MALLOWS, *Scale mixtures of normal distributions*, Journal of the Royal Statistical Society. Series B (Methodological), 36 (1974), pp. 99–102.

[5] W. ARLT, M. BIEHL, A. E. TAYLOR, S. HAHNER, R. LIBE, B. A. HUGHES, P. SCHNEIDER, D. J. SMITH, H. STIEKEMA, N. KRONE, ET AL., *Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors*, The Journal of Clinical Endocrinology & Metabolism, 96 (2011), pp. 3775–3784.

[6] A. ARMAGAN, D. B. DUNSON, AND J. LEE, *Generalized double Pareto shrinkage*, Statistica Sinica, 23 (2013), p. 119.

[7] M. BELKIN AND P. NIYOGI, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural computation, 15 (2003), pp. 1373–1396.

[8] Y. BENJAMINI AND Y. HOCHBERG, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the royal statistical society. Series B (Methodological), (1995), pp. 289–300.

[9] Y. BENJAMINI AND D. YEKUTIELI, *The control of the false discovery rate in multiple testing under dependency*, Annals of statistics, (2001), pp. 1165–1188.

[10] T. BERRY AND J. HARLIM, *Variable bandwidth diffusion kernels*, Applied and Computational Harmonic Analysis, 40 (2016), pp. 68–96.

[11] M. BETANCOURT, *A conceptual introduction to hamiltonian monte carlo*, arXiv preprint arXiv:1701.02434, (2017).

[12] M. BETANCOURT AND M. GIROLAMI, *Hamiltonian monte carlo for hierarchical models*, Current trends in Bayesian methodology with applications, 79 (2015), pp. 2–4.

[13] A. BHADRA, J. DATTA, N. G. POLSON, B. WILLARD, ET AL., *The horseshoe+ estimator of ultra-sparse signals*, Bayesian Analysis, (2016).

[14] D. M. BLEI, M. I. JORDAN, ET AL., *Variational inference for dirichlet process mixtures*, Bayesian analysis, 1 (2006), pp. 121–143.

[15] J. BOOTKRAJANG AND A. KABÁN, *Classification of mislabelled microarrays using robust sparse logistic regression*, Bioinformatics, 29 (2013), pp. 870–877.

[16] L. BREIMAN, W. MEISEL, AND E. PURCELL, *Variable kernel estimates of multivariate densities*, Technometrics, 19 (1977), pp. 135–144.

[17] S. BROOKS, A. GELMAN, G. L. JONES, AND X.-L. MENG, *Handbook of markov chain monte carlo*, Chapman and Hall/CRC, 2011.

[18] S. BUUREN AND K. GROOTHUIS-OUDSHOORN, *mice: Multivariate imputation by chained equations in r*, Journal of statistical software, 45 (2011).

[19] K. CAMPBELL, C. P. PONTING, AND C. WEBBER, *Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell rna-seq profiles*, bioRxiv, (2015), p. 027219.

[20] M. A. CARREIRA-PERPINÁN, *The elastic embedding algorithm for dimensionality reduction.*, in ICML, vol. 10, 2010, pp. 167–174.

[21] C. M. CARVALHO, N. G. POLSON, AND J. G. SCOTT, *The horseshoe estimator for sparse signals*, Biometrika, 97 (2010), pp. 465–480.

[22] C. M. DAFERMOS, *The second law of thermodynamics and stability*, Archive for Rational Mechanics and Analysis, 70 (1979), pp. 167–179.

[23] C. DING, T. LI, W. PENG, AND H. PARK, *Orthogonal nonnegative matrix t-factorizations for clustering*, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 126–135.

[24] C. H. DING, T. LI, AND M. I. JORDAN, *Convex and semi-nonnegative matrix factorizations*, IEEE transactions on pattern analysis and machine intelligence, 32 (2008), pp. 45–55.

[25] R. DIPERNA, *Uniqueness of solutions to hyperbolic conservation laws*, Indiana University Mathematics Journal, 28 (1979), p. 40.

[26] M. DOUMIC-JAUFFRET, P. S. KIM, AND B. PERTHAME, *Stability analysis of a simplified yet complete model for chronic myelogenous leukemia*, Bulletin of mathematical biology, 72 (2010), pp. 1732–1759.

[27] M. R. DOWLING, A. KAN, S. HEINZEL, J. H. ZHOU, J. M. MARCHINGO, C. J. WELLARD, J. F. MARKHAM, AND P. D. HODGKIN, *Stretched cell cycle model for proliferating lymphocytes*, PNAS, 111 (2014), pp. 6377–6382.

[28] J. J. EGOZCUE, V. PAWLOWSKY-GLAHN, G. MATEU-FIGUERAS, AND C. BARCELO-VIDAL, *Isometric logratio transformations for compositional data analysis*, Mathematical Geology, 35 (2003), pp. 279–300.

[29] M. D. ESCOBAR, *Estimating normal means with a dirichlet process prior*, Journal of the American Statistical Association, 89 (1994), pp. 268–277.

[30] J. L. FERRARA, J. E. LEVINE, P. REDDY, AND E. HOLLER, *Graft-versus-host disease*, The Lancet, 373 (2009), pp. 1550–1561.

[31] D. S. FISCHER, A. K. FIEDLER, E. M. KERNFELD, R. M. GENGA, A. BASTIDAS-PONCE, M. BAKHTI, H. LICKERT, J. HASENAUER, R. MAEHR, AND F. J. THEIS, *Inferring population dynamics from single-cell rna-sequencing time series data*, Nature biotechnology, 37 (2019), pp. 461–468.

[32] J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *Regularization paths for generalized linear models via coordinate descent*, Journal of statistical software, 33 (2010), p. 1.

[33] J. H. FRIEDMAN, J. L. BENTLEY, AND R. A. FINKEL, *An algorithm for finding best matches in logarithmic expected time*, ACM Transactions on Mathematical Software (TOMS), 3 (1977), pp. 209–226.

[34] M. K. C. FROM JED WING, S. WESTON, A. WILLIAMS, C. KEEFER, A. ENGELHARDT, T. COOPER, Z. MAYER, B. KENKEL, THE R CORE TEAM, M. BENESTY, R. LESCARBEAU, A. ZIEM, L. SCRUCCA, Y. TANG, C. CANDAN, AND T. HUNT., *caret: Classification and Regression Training*, 2016. R package version 6.0-73.

[35] V. V. GANUSOV, D. MILUTINOVIĆ, AND R. J. DE BOER, *IL-2 regulates expansion of CD4+ T cell populations by affecting cell death: insights from modeling CFSE data*, Journal of Immunology, 179 (2007), pp. 950–957.

[36] C. W. GARDINER ET AL., *Handbook of stochastic methods*, vol. 3, springer Berlin, 1985.

[37] A. GELMAN, J. B. CARLIN, H. S. STERN, AND D. B. RUBIN, *Bayesian data analysis*, vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[38] A. GELMAN, A. JAKULIN, M. G. PITTAU, AND Y.-S. SU, *A weakly informative default prior distribution for logistic and other regression models*, The Annals of Applied Statistics, (2008), pp. 1360–1383.

[39] A. GELMAN, D. B. RUBIN, ET AL., *Inference from iterative simulation using multiple sequences*, Statistical science, 7 (1992), pp. 457–472.

[40] S. GERSHMAN, M. HOFFMAN, AND D. BLEI, *Nonparametric variational inference*, arXiv preprint arXiv:1206.4665, (2012).

[41] J. GHOSH, Y. LI, R. MITRA, ET AL., *On the use of cauchy prior distributions for bayesian logistic regression*, Bayesian Analysis, 13 (2018), pp. 359–383.

[42] D. T. GILLESPIE, *Exact stochastic simulation of coupled chemical reactions*, The journal of physical chemistry, 81 (1977), pp. 2340–2361.

[43] ——, *The chemical langevin equation*, The Journal of Chemical Physics, 113 (2000), pp. 297–306.

[44] A. G. GRAY AND A. W. MOORE, *Nonparametric density estimation: Toward computational tractability*, in Proceedings of the 2003 SIAM International Conference on Data Mining, SIAM, 2003, pp. 203–211.

[45] J. GRIFFIN, P. BROWN, ET AL., *Hierarchical shrinkage priors for regression models*, Bayesian Analysis, 12 (2017), pp. 135–159.

[46] J. E. GRIFFIN AND P. J. BROWN, *Inference with normal-gamma prior distributions in regression problems*, Bayesian Analysis, 5 (2010), pp. 171–188.

[47] L. HAGHVERDI, F. BUETTNER, AND F. J. THEIS, *Diffusion maps for high-dimensional single-cell analysis of differentiation data*, Bioinformatics, 31 (2015), pp. 2989–2998.

[48] P. HALL, *Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems*, Journal of multivariate analysis, 32 (1990), pp. 177–203.

[49] T. HAPUARACHCHI, J. LEWIS, AND R. CALLARD, *A Mechanistic Model for Naive CD4 T Cell Homeostasis in Healthy Adults and Children*, Frontiers in Immunology, 4 (2013), p. 366.

[50] R. L. HOARE, *Modelling Immune Reconstitution following Paediatric Haematopoietic Stem Cell Transplantation and in HIV-Infected Children*, PhD thesis, University College London, October 2015.

[51] A. E. HOERL AND R. W. KENNARD, *Ridge regression: applications to nonorthogonal problems*, Technometrics, 12 (1970), pp. 69–82.

[52] M. D. HOFFMAN AND A. GELMAN, *The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.*, Journal of Machine Learning Research, 15 (2014), pp. 1593–1623.

[53] T. HOGAN, A. SHUVAEV, D. COMMENGES, A. YATES, R. CALLARD, R. THIEBAUT, AND B. SEDDON, *Clonally diverse T cell homeostasis is maintained by a common program of cell-cycle control*, Journal of Immunology, 190 (2013), pp. 3985–3993.

[54] L. JACOB, G. OBOZINSKI, AND J.-P. VERT, *Group lasso with overlap and graph lasso*, in Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 433–440.

[55] C. A. JANEWAY, P. TRAVERS, M. WALPORT, AND M. J. SHLOMCHIK, *Immunobiology, 5th edition*, Garland Science, New York, 2001.

[56] M. KALLI, J. E. GRIFFIN, AND S. G. WALKER, *Slice sampling mixture models*, Statistics and computing, 21 (2011), pp. 93–105.

[57] A. M. KLEIN, L. MAZUTIS, I. AKARTUNA, N. TALLAPRAGADA, A. VERES, V. LI, L. PESHKIN, D. A. WEITZ, AND M. W. KIRSCHNER, *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells*, Cell, 161 (2015), pp. 1187–1201.

[58] K. KURIHARA, M. WELLING, AND N. VLASSIS, *Accelerated variational dirichlet process mixtures*, in Advances in neural information processing systems, 2007, pp. 761–768.

[59] D. D. LEE AND H. S. SEUNG, *Algorithms for non-negative matrix factorization*, in Advances in neural information processing systems, 2001, pp. 556–562.

[60] Q. LI, N. LIN, ET AL., *The bayesian elastic net*, Bayesian Analysis, 5 (2010), pp. 151–170.

[61] Y. LUO, C. L. LIM, J. NICHOLS, A. MARTINEZ-ARIAS, AND L. WERNISCH, *Cell signalling regulates dynamics of nanog distribution in embryonic stem cell populations*, Journal of The Royal Society Interface, 10 (2013).

[62] J. M. MAHAFFY, J. BÉLAIR, AND M. C. MACKEY, *Hematopoietic model with moving boundary condition and state dependent delay: applications in erythropoiesis*, Journal of theoretical biology, 190 (1998), pp. 135–146.

[63] A. MCKENDRICK, *Applications of mathematics to medical problems*, Proceedings of the Edinburgh Mathematical Society, 44 (1926), pp. 98–130.

[64] P. MITRA, C. MURTHY, AND S. K. PAL, *Density-based multiscale data condensation*, IEEE Transactions on pattern analysis and machine intelligence, 24 (2002), pp. 734–747.

[65] J. D. MURRAY, *Mathematical biology: I. An introduction*, vol. 17, Springer Science & Business Media, 2007.

[66] B. NADLER, S. LAFON, I. KEVREKIDIS, AND R. R. COIFMAN, *Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators*, in Advances in neural information processing systems, 2006, pp. 955–962.

[67] O. PAPASPILIOPOULOS, G. O. ROBERTS, AND M. SKÖLD, *A general framework for the parametrization of hierarchical models*, Statistical Science, (2007), pp. 59–73.

[68] T. PARK AND G. CASELLA, *The bayesian lasso*, Journal of the American Statistical Association, 103 (2008), pp. 681–686.

[69] B. PERTHAME, *Transport Equations in Biology*, Birkhauser Basel, Basel, 2007.

[70] E. PIERRE-JEROME, S. S. JANG, K. A. HAVENS, J. L. NEMHAUSER, AND E. KLAVINS, *Recapitulation of the forward nuclear auxin response pathway in yeast*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 9407–9412.

[71] J. PIIRONEN, A. VEHTARI, ET AL., *Sparsity information and regularization in the horseshoe and other shrinkage priors*, Electronic Journal of Statistics, 11 (2017), pp. 5018–5051.

[72] N. G. POLSON AND J. G. SCOTT, *Local shrinkage rules, Lévy processes and regularized regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74 (2012), pp. 287–311.

[73] N. G. POLSON, J. G. SCOTT, AND J. WINDLE, *The bayesian bridge*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76 (2014), pp. 713–733.

[74] H. PUTTER, M. FIOCCO, AND R. B. GESKUS, *Tutorial in biostatistics: competing risks and multi-state models*, Statistics in medicine, 26 (2007), pp. 2389–2430.

[75] D. RAMKRISHNA, *Population balances: Theory and applications to particulate systems in engineering*, Elsevier, 2000.

[76] J. E. REID AND L. WERNISCH, *Pseudotime estimation: deconfounding single cell time series*, Bioinformatics, 32 (2016), pp. 2973–2980.

[77] J. REYNOLDS, M. COLES, G. LYTHE, AND C. MOLINA-PARS, *Mathematical Model of Naive T Cell Division and Survival IL-7 Thresholds*, Frontiers in Immunology, 4 (2013).

[78] S. R. SAIN, K. A. BAGGERLY, AND D. W. SCOTT, *Cross-validation of multivariate densities*, Journal of the American Statistical Association, 89 (1994), pp. 807–817.

[79] R. SANDBERG, *Entering the era of single-cell transcriptomics in biology and medicine*, Nature methods, 11 (2014), pp. 22–24.

[80] P. J. SCHMID, *Dynamic mode decomposition of numerical and experimental data*, Journal of fluid mechanics, 656 (2010), pp. 5–28.

[81] J. SETHURAMAN, *A constructive definition of Dirichlet priors*, Statistica Sinica, 4 (1994), pp. 639–650.

[82] E. SHAPIRO, T. BIEZUNER, AND S. LINNARSSON, *Single-cell sequencing-based technologies will revolutionize whole-organism science*, Nature Reviews Genetics, 14 (2013), pp. 618–630.

[83] F. SHARPE AND A. LOTKA, *A problem in age distribution*, Philosophical Magazine, 21 (1911), pp. 435–438.

[84] J. SHIN, D. A. BERG, Y. ZHU, J. Y. SHIN, J. SONG, M. A. BONAGUIDI, G. ENIKOLOPOV, D. W. NAUEN, K. M. CHRISTIAN, G. LI MING, AND H. SONG, *Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis*, Cell Stem Cell, 17 (2015), pp. 360 – 372.

[85] B. W. SILVERMAN, *Density estimation for statistics and data analysis*, Routledge, 1998.

[86] N. SIMON, J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization*, 2013. R package version 1.1.

[87] N. SIMON, J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *A sparse-group lasso*, Journal of Computational and Graphical Statistics, 22 (2013), pp. 231–245.

[88] A. SINGER, R. ERBAN, I. G. KEVREKIDIS, AND R. R. COIFMAN, *Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 16090–16095.

[89] STAN DEVELOPMENT TEAM, *RStan: the R interface to Stan*, 2016. R package version 2.14.1.

[90] ——, *rstanarm: Bayesian applied regression modeling via Stan.*, 2016. R package version 2.13.1.

[91] C. SURH AND J. SPRENT, *Homeostasis of naive and memory T cells*, Immunity, 29 (2008), pp. 848 – 862.

[92] T. TESHIMA, P. REDDY, AND R. ZEISER, *Acute graft-versus-host disease: novel biological insights*, Biology of Blood and Marrow Transplantation, 22 (2016), pp. 11–16.

[93] W. THOMSON, S. JABBARI, A. TAYLOR, W. ARLT, AND D. SMITH, *Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior*, Journal of the Royal Society Interface, 16 (2019), p. 20180572.

[94] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1996), pp. 267–288.

[95] C. TRAPNELL, D. CACCHIARELLI, J. GRIMSBY, P. POKHAREL, S. LI, M. MORSE, N. J. LENNON, K. J. LIVAK, T. S. MIKKELSEN, AND J. L. RINN, *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells*, Nature biotechnology, 32 (2014), p. 381.

[96] J. H. TU, C. W. ROWLEY, D. M. LUCHTENBURG, S. L. BRUNTON, AND J. N. KUTZ, *On dynamic mode decomposition: Theory and applications*, arXiv preprint arXiv:1312.0041, (2013).

[97] A. VEHTARI, J. OJANEN, ET AL., *A survey of bayesian predictive methods for model assessment, selection and comparison*, Statistics Surveys, 6 (2012), pp. 142–228.

[98] C. H. WADDINGTON, *The strategy of the genes*, Routledge, 1957.

[99] X. WANG, P. TINO, M. A. FARDAL, S. RAYCHAUDHURY, AND A. BABUL, *Fast parzen window density estimator*, in 2009 International Joint Conference on Neural Networks, IEEE, 2009, pp. 3267–3274.

[100] C. WEINREB, S. WOLOCK, B. K. TUSI, M. SOCOLOVSKY, AND A. M. KLEIN, *Fundamental limits on dynamic inference from single-cell snapshots*, Proceedings of the National Academy of Sciences, 115 (2018), pp. E2467–E2476.

[101] L. A. WELNIAK, B. R. BLAZAR, AND W. J. MURPHY, *Immunobiology of allogeneic hematopoietic stem cell transplantation*, Annu. Rev. Immunol., 25 (2007), pp. 139–170.

[102] D. J. WILKINSON, *Stochastic modelling for systems biology*, CRC press, 2011.

[103] M. O. WILLIAMS, I. G. KEVREKIDIS, AND C. W. ROWLEY, *A data–driven approximation of the koopman operator: Extending dynamic mode decomposition*, Journal of Nonlinear Science, 25 (2015), pp. 1307–1346.

[104] X. XU, M. GHOSH, ET AL., *Bayesian variable selection and estimation for group lasso*, Bayesian Analysis, 10 (2015), pp. 909–936.

[105] L. ZELNIK-MANOR AND P. PERONA, *Self-tuning spectral clustering*, in Advances in neural information processing systems, 2005, pp. 1601–1608.

[106] A. ZILMAN, V. V. GANUSOV, AND A. S. PERELSON, *Stochastic models of lymphocyte proliferation and death*, PLoS one, 5 (2010), p. e12775.

[107] H. ZOU AND T. HASTIE, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 301–320.