A Monte Carlo method to estimate cell population heterogeneity

Ben Lambert[1,2*], David J. Gavaghan[3], Simon Tavener[4].

1 Department of Zoology, University of Oxford, Oxford, Oxfordshire, U.K.
2 MRC Centre for Outbreak Analysis and Modelling, Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK.
3 Department of Computer Science, University of Oxford, Oxford, U.K.
4 Department of Statistics, Colorado State University, Fort Collins, Colorado, U.S.A.

*ben.c.lambert@gmail.com.

Revision date & time: 2019-05-05 17:09

1 Abstract

Variation is characteristic of all living systems. Laboratory techniques such as flow cytometry can probe individual cells and, after decades of experimentation, it is clear that even members of seemingly homogeneous cell populations can exhibit differences. To understand whether this variation is biologically meaningful, it is essential to discern its source. Mathematical models of biological systems are tools that can be used to investigate causes of cell-to-cell variation. From mathematical analysis and simulation of these models, biological hypotheses can be posed and investigated, then parameter inference can determine which of these is most compatible with experimental data. Data from laboratory experiments often takes the form of "snapshots" representing distributions of cellular properties at different points in times, rather than individual cell trajectories. This data is not straightforward to fit using hierarchical Bayesian methods since it requires inferring the identities of the groups to which individual cells belong. Here, we introduce a computational sampling method we call "Contour Monte Carlo" for estimating mathematical model parameters from snapshot distributions which is straightforward to implement and does not require explicitly assigning cells to categories. Our method is most applicable to systems where the dominant source of uncertainty is heterogeneity in cellular processes rather than experimental measurement error which, due to the increasingly finescale resolution of laboratory techniques, may be the case for a wide class of systems. In this paper, we illustrate the use of our method by quantifying cellular variation for two biological systems of interest and provide code in the form of a Julia notebook which allows others to apply this approach to their problem.

2 Introduction

Variation rather than homogeneity is the rule rather than exception in biology. Indeed, without variation, biology as a discipline would not exist, since as evolutionary biologist JBS Haldane wrote, variation is the "raw material" of evolution. The Red Queen Hypothesis asserts that organisms must continually evolve in order to survive when pitted against other - also evolving - organisms [1]. A corollary of this hypothesis is that multicellular organisms may evolve cellular phenotypic heterogeneity to allow for faster adaptation to changing environments, which may explain the observed variation in a range of biological systems [2]. Whilst cell population variation can confer evolutionary advantages, it can also be costly in other circumstances. In biotechnological processes, heterogeneity in cellular function can lead to reduced yields of biochemical products [3]. In human biology, variation across cells can enable pathologies to develop and also prevents effective medical treatment, since medical interventions typically aim to steer modal cellular properties and hence fail to influence key subpopulations. For example, cellular heterogeneity likely contributes to the persistence of some cancerous tumours [4] and may also allow them to evolve resistance to chemotherapies over time [5]. Identifying and quantifying sources of variation in populations of cells is important for a wide range of applications because it allows us to determine whether this variability is benign or alternatively requires remedy.

Mathematical models are essential tools for making sense of cellular systems, whose emergent properties are the result of complex interactions

between various actors. Perhaps the simplest flavour of mathematical model used in biological systems are ordinary differential equations (ODEs) that lump individual actors into partitions according to structure or function, and seek to model the mean behaviour of each partition. Data from population-averaged experimental assays can be a powerful resource to understand whether such models faithfully reproduce system behaviours and can allow quantification of the interactions of various cellular components of complex metabolic, signalling and transcriptional networks. The worth of such models however is determined by whether averages mask differences in behaviour of individual cells that result in functional consequences [6]. In some cases, differences in cellular protein abundances due to biochemical "noise" may not be meaningful biologically [7] and so mean cell behaviour suffices as a description of the system, whereas in others there are functional consequences. For example, a recent study demonstrated that subpopulations of clonally derived hematopoietic progenitor cells with low or high expression of a particular stem cell marker produced different blood lineages [8].

To accommodate cell population heterogeneity in mathematical models, a variety of modelling choices are available, each posing different challenges for parameter inference, and are described in a recent review [9]. These include modelling biochemical processes stochastically, with properties of ensembles of cells represented by probability distributions evolving according to chemical master equations (see [10] for a tutorial on stochastic reaction-diffusion processes; RDEs). Alternatively, population balance equations (PBEs) can be used to dictate the evolution of the "number density" of differing cell types, whose properties are represented as points in $\mathbb{R}^n$ which, in turn, affect their function, including their rate of death and cell division (see [11] for an introduction to PBEs). In a PBE approach, variation in measured quantities results primarily due to differing functional properties of heterogeneous cell types and variable initial densities of each type.

The approach we follow here is similar to that of [12], wherein dynamic cellular variation is generated by describing the evolution of each cell's state using an ODE, but with individual cell differences in the rate parameters of the process. To our knowledge, this flavour of model is unnamed and so, for sake of reference, we term them "heterogenous ODE" models (HODEs). In HODEs, the aim of inference is to estimate the distributions of parameter values across cells consistent with observed distributions of measurements at various timepoints. A benefit of using HODEs to model cell heterogeneity is that these models are computationally straightforward to simulate and, arguably, simpler to parameterise than PBEs. In these models the predominant source of variation is due to differences in biological processes across cells not inherent stochasticity in biochemical reactions within cells, as in stochastic RDEs.

The difficulty of parameter inference for HODEs is partly due to experimental hurdles in generating data of sufficient quality to allow identification. Unlike models which represent a population by a single scalar ODE, since HODEs are individual-based they ideally require individual cell data for estimation. A widely-used method for generating data for individual cells is flow cytometry, where a large number of cells are streamed individually through a laser beam and, for example, abundance measurements are made of proteins labelled with fluorescent markers [13]. Alternatively, experimental techniques such as Western blotting and cytometric fluorescence microscopy can generate single cell measurements [14, 15]. A property of

these experimental methods is that they are destructive, meaning that individual cells are sacrificed as part of the measurement process. This means that the measurements of cell properties conducted at a certain point in time represent what are termed "snapshots" of the underlying population [15]. These snapshots are often described by histograms [12] or density functions [9] fit to the underlying data at different points in time. Since HODEs represent the underlying state of individual cells as evolving continuously through time, corresponding data showing individual cell trajectories constitutes a richer data resource. The demands of obtaining this data are higher however and typically involve either tracking individual cells through imaging methods [citation] or trapping cells in a spatial position where their individual dynamics can be readily monitored [citation]. These techniques impose restrictions on experimental practices meaning that they cannot be realised in all circumstances, including for online monitoring of biotechnological processes or analysis of *in vivo* studies. For this reason, snapshot data continues to play an important role for determining cell level variability in a wide variety of cases.

A variety of approaches have been proposed to estimate cell-to-cell variability by fitting HODE models to snapshot data. In HODEs, parameter values vary across cells according to a to-be-determined probability distribution meaning that in order to solve the exact inverse problem, the underlying ODE system needs to be simulated for each individual. Since the numbers of cells in these experiments are typically $>\sim 10^4$ [15], this usually precludes exact inference due to its computational burden and instead the raw snapshot data is approximated by probability densities [12, 15–17]. Hasenauer et al. (2011) presents a Bayesian approach to inference for HODEs, which models the input parameter space using mixtures of ansatz densities, and use their method to reproduce population substructure on synthetic data generated from a model of tumour necrosis factor stimulus. Hasenauer et al. (2014) uses mixture-models to model the subpopulation structure in the snapshot data and uses multiple-start local optimisation to maximise the non-convex likelihood, which they then apply to a range of synthetic and real reaction data and signalling pathway examples. Loos et al. (2018) uses also uses mixture models to represent subpopulation structure and a maximum likelihood approach that allows for estimation of within- and between-subpopulation variability which also allows fitting to multivariate dependent output distributions. Dixit et al. (2018) discretises cell abundances into bins, then uses a maximum entropy approach as part of a Bayesian framework to fit the distribution representing cell-to-cell variability.

The framework we present here is Bayesian although is distinct from the traditional Bayesian inferential paradigm used to fit dynamic models since the source of stochasticity arises solely due to cell-to-cell parameter variation not measurement noise. Our approach is hence most suitable when measurement error is a minor contributor to observed experimental variability. Our computational method is a two-step Monte Carlo approach which, for reasons described in §3, we term "Contour Monte Carlo" (CMC). Unlike many of the existing methods however CMC is relatively computationally straightforward to implement and does not require extensive computation time. CMC uses MCMC in its second step to sample from the posterior distribution over parameter values and hence does not require specification of ansatz densities. It also does not require *a priori* representation of subpopulation structure using mixture components rather subpopulations appear

naturally as modes in the posterior parameter distributions. Like [17] CMC can fit multivariate snapshot data and unlike [12], does not require this data to be discretised into bins. As more experimental techniques are developed which elucidate single cell behaviour, there is likely to be more interest in methods which can be used to recapitulate the observed snapshots. We argue that due to its simplicity and generality, CMC is a useful addition to the modeller's toolkit, which has a role to play in the analysis of the proliferation of rich single cell data.

Outline of the paper: In §3, we present the details of our methodological framework and detail the CMC algorithm we use to sample from the posterior parameter distribution. In §4, we then use CMC to estimate cell population heterogeneity in three systems of biological interest.

3 Method

In this section, we describe the first describe the probabilistic framework that underlies the CMC algorithm, before introducing CMC in pseudocode (Algorithm 1). We also detail the workflow we have found useful in applying this approach to analyse cell snapshot data and suggest practical remedies to issues we have encountered in using CMC (Figure 4). A glossary of all the variables used in this paper is included as Table 1.

Experimental methods such as flow cytometry can measure single cell characteristics at a given point in time. Cells are typically destroyed by the measurement process and so rather than providing time series for each individual cell, the data consists of cross-sections or "snapshots" of sampled individuals from the population (Figure 1).

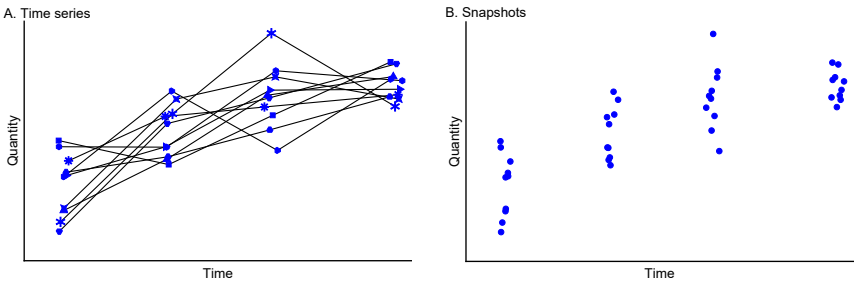

Figure 1: **Time series data (A) versus snapshot data (B) typical of single cell experiments.** In A that the cell identities are retained at each measurement point (indicated by given plot markers) whereas in the snapshot data in B, either this information is lost or, more often, cells are destroyed by the measurement process and so each observation corresponds to a distinct cell.

We model the processes of an individual cell using a system of ordinary differential equations (ODEs), where each element of the system describes the governing dynamics of a particular quantity of interest (for example, protein levels, RNA concentrations and so on),

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t); \boldsymbol{\theta}). \tag{1}$$

Here $\boldsymbol{x}(t) = (x_1(t), x_2(t), ..., x_k(t))$ is a vector of states for each compartment in the system at time t and $f(.)$ is a function of these states and parameters

$\boldsymbol{\theta} \in \mathbb{R}^p$. Note that in most circumstances, the initial state of the system, $\boldsymbol{x}(0)$, is unknown and it is convenient to include these as elements of $\boldsymbol{\theta}$ to be estimated. The solution of eq. (1) is given by $\boldsymbol{x}(t) = g(t; \boldsymbol{\theta})$, where $\boldsymbol{x}(t) \in \mathbb{R}^k$ is a vector of outputs at time t and $g(.)$ is a function that typically won't be analytically-determined; instead approximated via a numerical integration scheme.

In this paper, we assume variation characterised by snapshot data arises due to between-cell heterogeneity in the underlying parameters θ. Therefore, the evolution of the underlying state of cell i is described by an idiosyncratic ODE,

$$\dot{\boldsymbol{x}}^i(t) = f(\boldsymbol{x}^i(t); \boldsymbol{\theta}^i), \tag{2}$$

with solution $\boldsymbol{x}^i(t) = g(t; \boldsymbol{\theta}^i)$. The traditional (non-hierarchical) state-space approach to modelling dynamic systems supposes that measurement randomness generates output variation (Figure 2A). Our approach, by contrast, relies on the assumption that stochasticity in outputs is solely the result of variability in parameter values ($\boldsymbol{\theta}$) between cells (Figure 2B). Whether the assumption of "perfect" measurements is reasonable depends on the experimental details of the system under investigation but we argue that our method nevertheless provides a useful approximation in many cases where the signal to noise ratio is high.

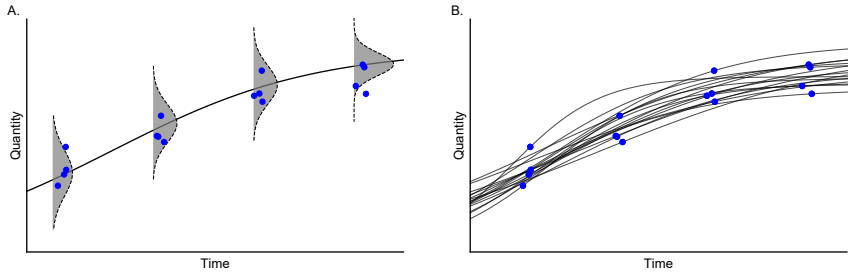

Figure 2: **Two ways to generate randomness in measured outputs: the state-space model (A) versus the parameter heterogeneity model (B).** For non-hierarchical state-space models (A), there is assumed to be a single "true" latent state where observations result from a noisy measurement process (grey histograms). For models with parameter heterogeneity (B), the uncertainty is generated by differences in cellular processes (black lines) between cells. Note that in both cases, individual cells are measured only once in their lifetime.

Our generative model used to produce output observations for a single cell consists of two stages: (i) sample $\boldsymbol{\theta}^i \sim p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is a probability distribution characterising heterogeneity in cellular processes and, (ii) calculate output values using $\boldsymbol{x}^i_j(t') = g(t'; \boldsymbol{\theta})$ where, for each cell, a measurement of a subset j of output states $\boldsymbol{x}^i_j(t') \in \boldsymbol{x}^i$ is made at a single point in time $t' \in (t_1, t_2, ..., t_o)$. The experimental observations for a single time point then consists of the collection of individual cell measurements $\boldsymbol{X}_j(t') = (\boldsymbol{x}^1_j(t'), \boldsymbol{x}^2_j(t'), ..., \boldsymbol{x}^n_j(t'))$, where n is the number of cells measured at time t'. The entire observation dataset is the union of such sets across all measured time points $\boldsymbol{X}(\boldsymbol{t}) = (\boldsymbol{X}_{j_1}(t_1), \boldsymbol{X}_{j_2}(t_2), ..., \boldsymbol{X}_{j_o}(t_o))$, where $\boldsymbol{t} = (t_1, t_2, ..., t_o)$ is the vector of observation times and the subscript a on each j_a allows for measurement of different elements of the system at distinct timepoints.

Variable	Definition	Dimension
$\boldsymbol{x}(t)$	state of modelled system at time t	$\mathbb{R}^k$
$x_j(t) \in \boldsymbol{x}(t)$	individual state j of modelled system at time t	$\mathbb{R}$
$\boldsymbol{\theta}$	parameters of ODE system	$\mathbb{R}^p$
$f(\boldsymbol{x}(t); \boldsymbol{\theta})$	function specifying RHS of ODE system	$\mathbb{R}^k$
$g(t; \boldsymbol{\theta})$	solution of ODE at time t	$\mathbb{R}^k$
$g_j(t; \boldsymbol{\theta})$	solution of ODE for state j at time t	$\mathbb{R}$
$\boldsymbol{x}^i(t)$	state of modelled system of cell i at time t	$\mathbb{R}^k$
$\boldsymbol{x}^i_j(t) \in \boldsymbol{x}^i(t)$	state of subset j of modelled system of cell i at time t	$\mathbb{R}^{\lvert j \rvert}$
$\boldsymbol{X}_j(t) = (\boldsymbol{x}^1_j(t), ..., \boldsymbol{x}^n_j(t))$	collection of n individual cell measurements at time t	$\mathbb{R}^{\lvert j \rvert \times n}$
$\boldsymbol{t} = (t_1, t_2, ..., t_o)$	unique observation times	$\mathbb{R}^o$
$\boldsymbol{X}(\boldsymbol{t}) = (\boldsymbol{X}_{j_1}(t_1), ..., \boldsymbol{X}_{j_o}(t_o))$	all observations collected at times $\boldsymbol{t}$	$\dim(\boldsymbol{X}_{j_1}) \times ... \times \dim(\boldsymbol{X}_{j_o})$
Φ	parameters characterising output target distribution $p(\boldsymbol{x}\lvert\Phi)$	-
$\hat{a}$	estimates of any quantity a	-
$\tilde{\boldsymbol{t}} = (t_1, ..., t_m)$	times when each observable is recorded	$\mathbb{R}^m$
$\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{t}}) = (x_{j_1}(t_1), ..., x_{j_m}(t_m))$	system observables	$\mathbb{R}^m$
$\mathcal{V}(\tilde{\boldsymbol{x}})$	the "volume" of parameter space mapping to an output of value $\boldsymbol{x}$	$\mathbb{R}^+$
$\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}(\tilde{\boldsymbol{t}}; \boldsymbol{\theta}) = (g_{j_1}(t_1; \boldsymbol{\theta}), ..., g_{j_m}(t_m; \boldsymbol{\theta}))$	solution of ODE for each observable at respective times $\boldsymbol{t}$	$\mathbb{R}^m$
V	total volume of parameter space with uniform priors used for all parameters	$\mathbb{R}^+$
$\Omega(\tilde{\boldsymbol{x}}) = \{\boldsymbol{\theta} : \boldsymbol{g}(\boldsymbol{\theta}) = \tilde{\boldsymbol{x}}\}$	region of parameter space mapping to output $\tilde{\boldsymbol{x}}$	$\mathbb{R}^p$
Ψ	parameters characterising output prior distribution $p(\tilde{\boldsymbol{x}}\lvert\Psi)$	-
Ξ	parameters characterising parameter prior distribution $p(\boldsymbol{\theta}\lvert\Xi)$	-

Table 1: **Glossary of variable names used in this paper.** The dimensions of Φ, Ψ and Ξ are listed as "-" since they depend on the form of the density used to represent the process and can be anywhere from $\mathbb{R}^1$ to $\mathbb{R}^\infty$. The variables are listed in the approximate order in which they appear in the text.

Raw snapshot data consists of measurements of individual cells with exact inference requiring simulating the underlying ODE system for each individual. This is cumbersome and impractical for the numbers of cells sampled in typical experimental setups and so, instead, we follow previous work and instead represent snapshot data using probability distributions [12, 15–17]. The snapshots themselves can either be distributions of a single species or multiple species, which can be approximated by univariate and multivariate probability distributions respectively. These probability distributions are characterised by parameter estimates $\hat{\Phi}$ determined by the output observations $\boldsymbol{X}(\boldsymbol{t})$. The dimensionality of these probability distributions depends on the set of m distinct observables $\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{t}}) = (x_{j_1}(t_1), x_{j_2}(t_2), ..., x_{j_m}(t_m))$ recorded by experimental measurements. Note that, $\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{t}})$ corresponds to a particular set of measurements from a hypothetical cell and is distinct from $\boldsymbol{X}(\boldsymbol{t})$, which represents the full set of experimental outputs. The vector $\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{t}})$ is hypothetical because in reality each cell is measured at a single timepoint (although we suppose measurements of different cellular attributes are possible contemporaneously).

The goal of our inference process is to characterise the probability distribution $p(\boldsymbol{\theta}\lvert\boldsymbol{X}(\boldsymbol{t}))$ representing heterogeneity in cellular processes. The first step in our inference workflow is to fit the output distributions using probability distributions (Figure 4(i)). We assume that the volume of observational data means the estimated probability distributions are approximate sufficient statistics of the outputs, meaning $p(\boldsymbol{\theta}\lvert\hat{\Phi}) \approx p(\boldsymbol{\theta}\lvert\boldsymbol{X}(\boldsymbol{t}))$.

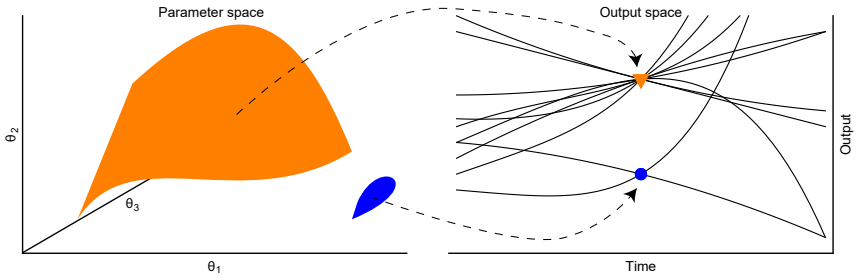

Figure 3: **The non-linear mapping from parameter values (left panel) to outputs (black lines; right panel) means different sized regions of parameter space (orange and blue surfaces; left panel) correspond to distinct output values (orange triangle and blue square; right panel).** In the right panel, each black line represents a distinct model simulation $g(t; \theta_1, \theta_2, \theta_3)$ and the triangle and square indicate outputs at a given point in time.

The models we seek to fit to snapshot data mostly cannot be identified from the observations. This is often because the number of model parameters exceeds the dimensionality of the output distribution (that is, $m > k$) meaning there typically exist non-singular sets of parameter values mapping to a single set of output values. That is, each vector of observed outputs $\tilde{\boldsymbol{x}} \in \mathbb{R}^m$, can often be caused by many combinations of parameters although, due to the non-linearity of the map from parameters to outputs, the "volume" of these regions of parameter space, $\mathcal{V}(\tilde{\boldsymbol{x}})$, is a function of output (Figure 3). In what follows, we make clear the distinction between observables $\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{t}})$ and the vector-valued function representing modelled outputs $\boldsymbol{g}(\tilde{\boldsymbol{t}}; \boldsymbol{\theta}) = (g_{j_1}(t_1; \boldsymbol{\theta}), g_{j_2}(t_2; \boldsymbol{\theta}), ..., g_{j_m}(t_m; \boldsymbol{\theta})) \in \mathbb{R}^m$ since the latter is a function whereas that latter is a numeric value; we also drop the $\tilde{\boldsymbol{t}}$ notation from following expressions to minimise clutter.

A consequence of this non-linear parameter to output geometry is that any target output distribution $p(\tilde{\boldsymbol{x}}|\hat{\Phi})$ does not correspond to a unique parameter distribution. For example, suppose $g(\theta_1, \theta_2) = \theta_1 + \theta_2$: the target distribution $\tilde{\boldsymbol{x}} \sim \mathcal{N}(0, 1)$ can be generated by any member of the set of parameter distributions $\sqrt{\eta}\theta_1 + \sqrt{1-\eta}\theta_2$, where $\eta \in [0, 1]$ and $\theta_1, \theta_2 \sim \mathcal{N}(0, 1)$. This means that in order to ensure uniqueness of the "posterior" parameter distributions, we are required to specify "prior" distributions for the parameters, as in more traditional Bayesian inference. An additional consequence of the degeneracy of the mapping from parameters to outputs is that any sampling algorithm aimed at exploring posterior parameter space must account for the differential volumes of iso-output contours. Whilst we refer the interested reader to our companion paper on this subject [citation for tutorial paper published in Open Science], we provide a quick derivation of the posterior parameter distribution which accounts for the non-linear mapping.

To derive the posterior distribution of parameter values $p(\boldsymbol{\theta}|\hat{\Phi})$, we consider the joint density of parameters and outputs $p(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}|\hat{\Phi})$. This can be decomposed in two ways,

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \hat{\Phi}) \times p(\tilde{\boldsymbol{x}}|\hat{\Phi}) = p(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}|\hat{\Phi}) = p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta}, \hat{\Phi}) \times p(\boldsymbol{\theta}|\hat{\Phi}). \quad (3)$$

The left and right hand sides of eq. (3) can be equated and rearranged to

obtain the posterior parameter distribution,

$$p(\boldsymbol{\theta}|\hat{\Phi}) = \frac{p(\boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \hat{\Phi}) \times p(\tilde{\boldsymbol{x}}|\hat{\Phi})}{p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta}, \hat{\Phi})}. \tag{4}$$

Given parameters $\boldsymbol{\theta}$, the mapping from parameters to outputs is deterministic meaning $p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta}, \hat{\Phi}) = \delta(\boldsymbol{g}(\boldsymbol{\theta}))$ is the Dirac delta function centred at $\tilde{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{\theta})$. In what follows, we assume that the conditional distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \hat{\Phi})$ is independent of the data, meaning it represents a conditional "prior", which can be manipulated by Bayes' rule,

$$p(\boldsymbol{\theta}|\boldsymbol{g}(\boldsymbol{\theta})) = \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{g}(\boldsymbol{\theta}))}, \tag{5}$$

where we have used the Dirac delta function for $p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})$. This results in the form of the posterior parameter distribution targeted by our sampling algorithm,

$$p(\boldsymbol{\theta}|\hat{\Phi}) = \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{g}(\boldsymbol{\theta}))} p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Phi}). \tag{6}$$

Again, we refer to our companion piece [citation] for detailed explanation of eqs. (5) & (6) and instead here provide brief interpretation when considering a uniform prior on parameter space. In this case, $p(\boldsymbol{\theta}) = \frac{1}{V}$, where V is the total volume of parameter space. The denominator term of eq. (5) is the prior induced on output space by the prior over parameter space. For a uniform prior on parameter values, this is just proportion of parameter space where $\boldsymbol{g}(\boldsymbol{\theta}) = \tilde{\boldsymbol{x}}$, meaning,

$$p(\boldsymbol{\theta}|\boldsymbol{g}(\boldsymbol{\theta})) = \frac{1}{\mathcal{V}(\boldsymbol{g}(\boldsymbol{\theta}))}, \tag{7}$$

where $\mathcal{V}(\boldsymbol{g}(\boldsymbol{\theta}))$ is the volume of parameter space occupied by the iso-output region $\Omega(\tilde{\boldsymbol{x}}) = \{\boldsymbol{\theta} : \boldsymbol{g}(\boldsymbol{\theta}) = \tilde{\boldsymbol{x}}\}$. Therefore a uniform prior over parameter space implies a prior structure where all parameter values resulting in the same output $\tilde{\boldsymbol{x}}$ are given equal weighting.

The denominator term of eq. (5) cannot be calculated apart from for some toy examples, meaning that exact sampling from the posterior parameter distribution of eq. (6) is not, in general, possible. We propose instead a computationally efficient sampling method to estimate $p(\boldsymbol{g}(\boldsymbol{\theta}))$, which forms the first step of our so-called "Contour Monte Carlo" (CMC) algorithm (Algorithm 1; Figure 4(ii)), where we estimate the volume of iso-output contours with output value $\boldsymbol{g}(\boldsymbol{\theta})$. This step involves repeated independent sampling from the prior distribution of parameters $\boldsymbol{\theta}^i \sim p(\boldsymbol{\theta}|\Xi)$, where, for completeness, we have conditioned on Ξ parameterising our probability density. Each parameter sample is then converted into an output value $\tilde{\boldsymbol{x}}^i = \boldsymbol{g}(\boldsymbol{\theta}^i)$. The collection of output samples is then fitted using a vine copula kernel density estimator (KDE) [18], $(\tilde{\boldsymbol{x}}^1, \tilde{\boldsymbol{x}}^2, ..., \tilde{\boldsymbol{x}}^{N_1}) \sim p(\tilde{\boldsymbol{x}}|\hat{\Psi})$. Throughout the course of development of CMC, we have tested many forms of KDE and have found vine copula KDE is best suited to approximating the higher dimensional probability distributions required in practice.

The second step in our algorithm then uses Markov chain Monte Carlo (MCMC) to sample from an approximate version of eq. (6) with the estimated density $p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Psi})$ replacing its corresponding estimand (Algorithm 1; Figure 4(iii)),

$$p(\boldsymbol{\theta}|\hat{\Phi}, \Xi, \hat{\Psi}) = \frac{p(\boldsymbol{\theta}|\Xi)}{p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Psi})} p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Phi}). \tag{8}$$

Algorithm 1 Pseudocode for the Contour Monte Carlo algorithm for sampling from the posterior parameter distribution of eq. (8). Here we provide code for the Random Walk Metropolis algorithm for the MCMC sampling but for the examples in §4, we use an adaptive MCMC algorithm [19]. A definition of all variables is provided in Table 1.

procedure CMC($\boldsymbol{X}(\boldsymbol{t}), \Xi, N_1, N_2$) ▷ Sample from posterior parameter distribution
  $\hat{\Phi}$ = SnapshotEstimator($\boldsymbol{X}(\boldsymbol{t})$)
  $\hat{\Psi}$ = ContourVolumeEstimator(Ξ)
  $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2})$ = MCMC($\hat{\Phi}, \Xi, \hat{\Psi}$)
  converged = CompareOutputToTarget($(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2}), \hat{\Psi}$)
  if converged≠1 **then**
    $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2})$=ContourVolumeEstimator($\hat{\Phi}, \Xi, N_1', N_2'$)
    where, $N_1' > N_1$ and/or $N_2' > N_2$ ▷ Rerun contour volume estimation and/or MCMC with larger sample sizes
  end if
  return $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2})$
end procedure

procedure SnapshotEstimator($\boldsymbol{X}(\boldsymbol{t})$) ▷ Fit density to snapshot observations
  $\boldsymbol{X}(\boldsymbol{t}) \sim p(\tilde{\boldsymbol{x}}|\hat{\Phi})$
  return $\hat{\Phi}$
end procedure

procedure ContourVolumeEstimator(Ξ) ▷ Estimate volume of contours
  for i in $1 : N_1$ **do**
    $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\Xi)$ ▷ Sample from prior density
    $\tilde{\boldsymbol{x}}_i = \boldsymbol{g}(\boldsymbol{\theta}_i)$ ▷ Calculate corresponding output value
  end for
  $(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_{N_1}) \sim p(\tilde{\boldsymbol{x}}|\hat{\Psi})$ ▷ Fit vine copula kernel density estimator to output values.
  return $\hat{\Psi}$
end procedure

procedure MCMC($\hat{\Phi}, \Xi, \hat{\Psi}$)▷ Random Walk Metropolis algorithm targeting posterior parameter distribution.
  $\boldsymbol{\theta}_0 \sim \pi(.)$ ▷ Sample from arbitrary initialisation distribution
  for i in $1 : N_2$ **do**
    $\boldsymbol{\theta}_i' \sim \mathcal{N}(\boldsymbol{\theta}_{i-1}, \boldsymbol{\Sigma})$ ▷ Propose new parameter values for parameters
    $r = \left[p(\boldsymbol{\theta}'|\Xi)p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Psi})p(\boldsymbol{g}(\boldsymbol{\theta}')|\hat{\Phi})\right] / \left[p(\boldsymbol{\theta}|\Xi)p(\boldsymbol{g}(\boldsymbol{\theta}')|\hat{\Psi})p(\boldsymbol{g}(\boldsymbol{\theta})|\hat{\Phi})\right]$ ▷ Metropolis acceptance ratio.
    $u \sim U(0, 1)$ ▷ Sample from uniform distribution
    if $r > u$ **then**
      $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i'$ ▷ Accept proposal
    else
      $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$ ▷ Reject proposal
    end if
  end for
  return $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2})$
end procedure

procedure CompareOutputToTarget($(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_2}), \hat{\Psi}$)▷ Check output distribution close to target
  for i in $1 : N_2$ **do**
    $\tilde{\boldsymbol{x}}_i = \boldsymbol{g}(\boldsymbol{\theta}_i)$ ▷ Compute output for each parameter sample
  end for
  if $(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_{N_2}) \sim p(\tilde{\boldsymbol{x}}|\hat{\Psi})$? **then** ▷ Compare outputs with target
    return 1 ▷ If outputs sufficiently close then converged
  else
    return 0
  end if
end procedure

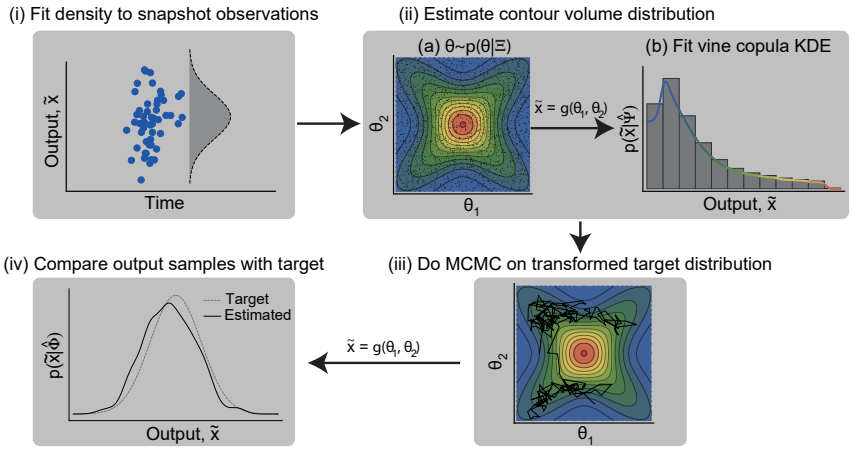

Figure 4: **The workflow for using Contour Monte Carlo to estimate cell population heterogeneity.** In (iii), the distribution targeted is given by eq. (8) but with $p(\tilde{\boldsymbol{x}}|\hat{\Psi}) \to p(g(\boldsymbol{\theta}))$. The variables used in this figure are defined in the text and Table 1.

The final step in CMC is to transform parameter samples from the MCMC into outputs, then compare the sampled outputs with the target distribution (Figure 4(iv)). Asymptotically (in terms of the sample size of both sampling steps), CMC produces a sample of parameter values $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ...)$ which, when transformed to outputs, corresponds to the target distribution $p(\boldsymbol{x}|\hat{\Psi})$. In developing CMC, we have found that a finite sample of modest size for both steps of CMC results in parameter samples that, when transformed, often represent reasonable approximations of the target. There are however occasions when this is not the case and we have found this final confirmatory step indispensable since it frequently highlights inadequacies in the contour volume estimation or the MCMC, meaning more samples from either or both of these steps are required. It may also be necessary to tweak hyperparameters of the KDE to ensure reasonable approximation in the contour volume estimation step. If the target distribution is sensitive to the contour volume estimates, this may also indicate that the target snapshot distribution is incompatible with the model: here, we make no claims on existence of a solution to the inverse problem, only that, if one should exist, Contour Monte Carlo is a pragmatic approach to approximate it by sampling. A useful way to diagnose whether the target distribution can be produced from the model and specified priors is to examine the output values from the contour volume estimation step of CMC. If the majority of probability mass of the target lies outside the bounds of the bulk simulated output values obtained by independent sampling from the prior, then the model and/or chosen prior is unlikely to be invertible to this particular target.

For the contour volume estimation step, we assumed sample sizes were sufficient if the output samples from the MCMC provided a reasonable approximation to the target, although we recognise that future work should refine this process further. For the MCMC step, we use adaptive covariance MCMC (see SOM of [19]) to sample from the target distribution, as we have found that it provides a considerable speed-up over Random Walk Metropolis [20, 21]. We also use the Gelman-Rubin convergence statistic $\hat{R}$ which provides a heuristic measurement of convergence [21, 22], and use a

threshold of $\hat{R} \leq \sim 1.1$ to diagnose convergence.

4 Results

In this section, we use CMC to estimate the posterior parameter distribution for three biological systems of interest targeting synthetic parametric densities. That is, we assume that the first step of CMC ("SnapshotEstimator" within Algorithm 1) has already been undertaken and we are faced with inferring a parameter distribution which, when transformed to outputs, recapitulates the target density. To complement the results, we also provide a Julia notebook which was used to generate the data, which we hope will be of use to others wanting to apply CMC to estimate cell population heterogeneity.

Model	Target density	Parameter	Prior density	Prior parameter 1	Prior parameter 2
Growth factor	two-dimensional normal	R_T	uniform	2.5×10^5	8×10^5
		k_1	uniform	0.25	3.0
		k_{-1}	uniform	2.0	20.0
		k_{deg}	uniform	0.005	0.03
		k^*_{deg}	uniform	0.1	0.5
Growth factor	two-dimensional normal	R_T	normal	5×10^5	1×10^5
		k_1	normal	0.5	0.1
		k_{-1}	normal	3.0	1.0
		k_{deg}	normal	0.02	0.005
		k^*_{deg}	normal	0.3	0.1
Michaelis-Menten kinetics	bimodal normal	k_f	uniform	0.2	15
		k_r	uniform	0.2	2.0
		k_{cat}	uniform	0.5	3.0
Michaelis-Menten kinetics	four-dimensional normal	k_f	uniform	0.2	15
		k_r	uniform	0.2	2.0
		k_{cat}	uniform	0.2	3.0
		E_0	uniform	3.0	5.0
		S_0	uniform	5.0	10.0
		ES_0	uniform	0.0	0.2
		P_0	uniform	0.0	0.2
TNF signalling	bimodal normal	a_1	uniform	0.5	0.7
		a_2	uniform	0.1	0.3
		a_3	uniform	0.1	0.3
		a_4	uniform	0.4	0.6
		b_1	uniform	0.3	0.5
		b_2	uniform	0.6	0.8
		b_3	uniform	0.2	0.4
		b_4	uniform	0.4	0.6
		b_5	uniform	0.3	0.5

Table 2: **The priors used for each problem in §4.**

4.1 Growth factor model

Here we consider the a "growth factor model" introduced by [12], which concerns the dynamics of inactive ligand-free cell surface receptors R and active ligand-bound cell surface receptors P, modulated by a ligand L. The governing dynamics are determined by the following system,

$$\frac{\mathrm{d}R}{\mathrm{d}t} = R_T k_{deg} + k_1 LR + k_{-1} P - k_{deg} R \tag{9}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = k_1 LR - k_{-1} P - k^*_{deg} P, \tag{10}$$

where $\boldsymbol{\theta} = (R_T, k_1, k_{-1}, k_{deg}, k^*_{deg})$ are parameters to be determined. In this example, we use measurements of the active ligand-bound receptors

P to estimate cellular heterogeneity in processes. We denote the solution of eq. (9) as $P(t;\boldsymbol{\theta},L)$ and seek to determine the parameter distribution consistent with an output distribution,

$$\begin{pmatrix} P(10;\boldsymbol{\theta},2) \\ P(10;\boldsymbol{\theta},10) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2\times10^4 \\ 3\times10^4 \end{pmatrix}, \begin{pmatrix} 1\times10^5 & 0 \\ 0 & 1\times10^5 \end{pmatrix}\right]. \tag{11}$$

To start, we specify a uniform prior for each of the five parameters, with bounds given in Table 2. To estimate the posterior parameter distribution, we use CMC, with adaptive covariance MCMC [19] for the second step.

In Figure 5A, we show the target distributions of outputs (black solid lines and contours) versus the sampled distribution (blue dots and dashed lines), illustrating that CMC is able to recapitulate the target densities using modest sample sizes. In Figure 6A, we plot the joint posterior parameter distribution for k_1, the rate of ligand binding to inactive receptors, and k_{-1}, which dictates the rate of the reverse reaction, where the ligands unbind. The outputs we fit to correspond to levels of the bound ligands, which can be generated by low rates of binding and unbinding but also for high rates of both. Because of this, the distribution representing cell process heterogeneity contains linear positive correlations.

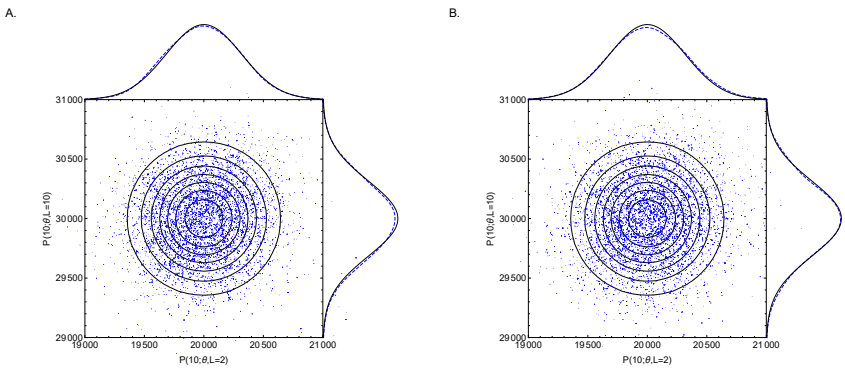

Figure 5: **The target joint output distribution (solid contour lines) and target marginal distributions (solid) versus outputs sampled by CMC (blue dots and dashed lines) for (A) uniform and (B) normal parameter priors.** In the CMC, 100,000 independent samples were used in the "ContourVolumeEstimator" step and 10,000 MCMC samples across each of 4 Markov chains, with the first half of the chains discarded as "warm-up" [21], were used in the "MCMC" step in Algorithm 1. For the reconstructed marginal densities, we use Mathematica's "SmoothKernelDistribution" function with bandwidths of 100 with Gaussian kernels [23].

Parameter	Uniform			Normal		
	2.5%	50%	97.5%	2.5%	50%	97.5%
R_T	441010	606440	772480	408400	529560	678630
k_1	0.89	2.16	2.95	0.39	0.54	0.70
k_{-1}	4.35	11.23	18.71	1.39	2.26	3.35
k_{deg}	0.01	0.02	0.03	0.02	0.02	0.03
k^*_{deg}	0.20	0.40	0.49	0.22	0.33	0.46

Table 3: **Estimated quantiles from CMC samples for the growth factor model with uniform and normal priors.** The particular priors used in each case are given in Table 2.

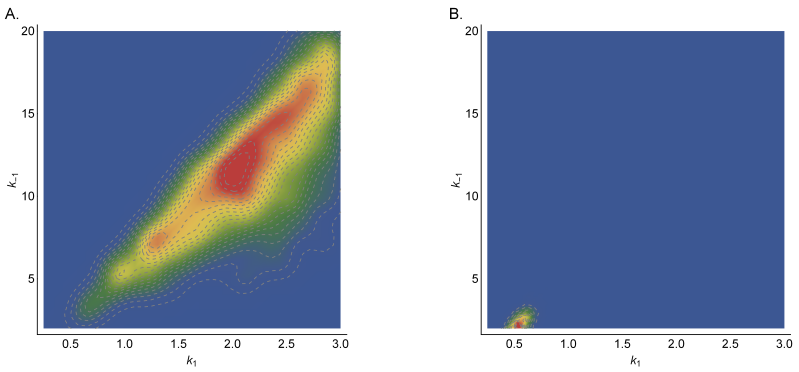

Figure 6: **The joint distribution of (k_1, k_{-1}) for the growth factor model using (A) uniform priors and (B) normal priors.** See Figure 5 caption for CMC details and Table 2 for the priors used.

For an unidentified model, there are typically a multitude of possible probability distributions over parameters which map to the same target output distribution. To reduce the space of posterior parameter distributions to one, it is therefore necessary to specify a prior parameter distribution. CMC accommodates the effect of changes to a prior distribution directly, through changes to both the "ContourVolumeEstimation" step and the acceptance ratio in the "MCMC" step (Algorithm 1), such that the posterior parameter distribution maps to the same output target.

4.2 Michaelis-Menten kinetics

Target a bimodal target then a four-dimensional covariance target. No change in model details apart from target.

4.3 TNF signalling pathway

Target a bimodal target.

5 Discussion

To do:

Make clear that inference is not circular here typically, since we are in an unidentified region.

When struggling to target a given distribution using this method, this indicates a) the contour estimates are not refined enough and b) that the generating model (without measurement uncertainty) is unable to recapitulate the target.

6 Author contributions

BL, DJG and SJT conceived the study. BL carried out the analysis. All authors helped to write and edit the manuscript.

References

[1] Matt Ridley. *The red queen: Sex and the evolution of human nature*. Penguin UK, 1994.

[2] Dawn Fraser and Mads Kaern. A chance at survival: gene expression noise and phenotypic diversification strategies. *Molecular microbiology*, 71(6):1333–1340, 2009.

[3] Frank Delvigne, Quentin Zune, Alvaro R Lara, Waleed Al-Soud, and Søren J Sørensen. Metabolic variability in bioprocessing: implications of microbial phenotypic heterogeneity. *Trends in Biotechnology*, 32(12):608–616, 2014.

[4] RA Gatenby, K Smallbone, PK Maini, F Rose, J Averill, Raymond B Nagle, L Worrall, and RJ Gillies. Cellular adaptations to hypoxia and acidosis during somatic evolution of breast cancer. *British journal of cancer*, 97(5):646, 2007.

[5] Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730, 2015.

[6] Steven J Altschuler and Lani F Wu. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4):559–563, 2010.

[7] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[8] Hannah H Chang, Martin Hemberg, Mauricio Barahona, Donald E Ingber, and Sui Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544, 2008.

[9] Steffen Waldherr. Estimation methods for heterogeneous cell population models in systems biology. *Journal of The Royal Society Interface*, 15(147):20180530, 2018.

[10] Radek Erban, Jonathan Chapman, and Philip Maini. A practical guide to stochastic simulations of reaction-diffusion processes. *arXiv preprint arXiv:0704.1908*, 2007.

[11] Doraiswami Ramkrishna and Meenesh R Singh. Population balance modeling: current status and future prospects. *Annual review of chemical and biomolecular engineering*, 5:123–146, 2014.

[12] Purushottam Dixit, Eugenia Lyashenko, Mario Niepel, and Dennis Vitkup. Maximum entropy framework for inference of cell population heterogeneity in signaling network dynamics. *bioRxiv*, page 137513, 2018.

[13] William G Telford, Teresa Hawley, Fedor Subach, Vladislav Verkhusha, and Robert G Hawley. Flow cytometry of fluorescent proteins. *Methods*, 57(3):318–330, 2012.

[14] Alex J Hughes, Dawn P Spelke, Zhuchen Xu, Chi-Chih Kang, David V Schaffer, and Amy E Herr. Single-cell western blotting. *Nature methods*, 11(7):749, 2014.

[15] Jan Hasenauer, Steffen Waldherr, Malgorzata Doszczak, Nicole Radde, Peter Scheurich, and Frank Allgöwer. Identification of models of heterogeneous cell populations from population snapshot data. *BMC bioinformatics*, 12(1):125, 2011.

[16] Jan Hasenauer, Christine Hasenauer, Tim Hucho, and Fabian J Theis. Ode constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. *PLoS computational biology*, 10(7):e1003686, 2014.

[17] Carolin Loos, Katharina Moeller, Fabian Fröhlich, Tim Hucho, and Jan Hasenauer. A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell systems*, 6(5):593–603, 2018.

[18] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[19] Ross H Johnstone, Eugene TY Chang, Rémi Bardenet, Teun P De Boer, David J Gavaghan, Pras Pathmanathan, Richard H Clayton, and Gary R Mirams. Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models? *Journal of molecular and cellular cardiology*, 96:49–62, 2016.

[20] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[21] Ben Lambert. *A Student's Guide to Bayesian Statistics*. Sage Publications Ltd., 2018.

[22] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

[23] Inc. Wolfram Research. Mathematica 8.0. `https://www.wolfram.com`.