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 Global Infectious Disease Analysis, School of Public Health, Imperial College London, London W2 1PG, UK.
3 Department of Computer Science, University of Oxford, Oxford, U.K.
4 Department of Mathematics, Colorado State University, Fort Collins, Colorado, U.S.A.

*ben.c.lambert@gmail.com.

Revision date & time: 2019-06-18 12:48

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 genetically identical 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 compatible with experimental data. Data from laboratory experiments often consist of "snapshots" representing distributions of cellular properties at different points in time, rather than individual cell trajectories. These data are not straightforward to fit using hierarchical Bayesian methods, which require the number of cell population clusters to be chosen *a priori*. Here, we introduce a computational sampling method for estimating mathematical model parameters from snapshot distributions named "Contour Monte Carlo", which is straightforward to implement and does not require cells be assigned to predefined categories. Our method is appropriate for systems where observed variation is attributable mostly to variability in cellular processes rather than experimental measurement error, which may be the case for many systems due to continued improvements in the resolution of laboratory techniques. In this paper, we apply our method to quantify cellular variation for three biological systems of interest and provide Julia code enabling others to use this method.

2 Introduction

Variation, as opposed to 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 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 should evolve cellular phenotypic heterogeneity to allow 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 helps some cancerous tumours to persist [4] and facilitates evolution of resistance to chemotherapies [5]. Identifying and quantifying sources of variation in populations of cells is important for wide ranging applications to discern whether the variability is benign or alternatively requires remedy.

Mathematical models are essential tools for understanding 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 is an ordinary differential equation (ODE)

that lumps individual actors into partitions according to structure or function, and seeks to model the mean behaviour of each partition. Data from population-averaged experimental assays can determine 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 depends on whether averages mask individual differences in behaviour that result in functional consequences [6]. In some cases, differences in cellular protein abundances due to biochemical "noise" are not biologically meaningful [7] and the system is well described by average cell behaviour. In others there are functional consequences. For example, a laboratory study demonstrated 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, many modelling frameworks are available, each posing different challenges for parameter inference. A recent review is presented in [9]. These approaches 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 from 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 distributions of parameter values across cells consistent with observations. A benefit of using HODEs to model cell heterogeneity is these models are computationally straightforward to simulate and, arguably, simpler to parameterise than PBEs. An implicit assumption in using HODEs is that most observed variation comes from differences in biological processes across cells, not inherent stochasticity in biochemical reactions within cells, as in stochastic RDEs.

Inference for HODEs is partly difficult due to the hurdle of generating experimental data of sufficient quality to allow parameter 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 such data 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]. Other experimental techniques, including Western blotting and cytometric fluorescence microscopy, can also generate single cell measurements [14, 15]. These experimental methods are all however destructive, meaning individual cells are sacrificed during measurement, and observations at each time point represent *"snapshots"* of the underlying population [15]. These snapshots are often described by histograms [12] or density functions [9] fit to measurements of each quantity

of interest. Since HODEs assume the state of each cell evolves continuously over time, experimental data tracing individual cell trajectories through time constitutes a richer data resource. The demands of obtaining this data are higher however and typically involve either tracking individual cells through imaging methods [16] or trapping cells in a spatial position where their individual dynamics can be readily monitored [17]. These techniques impose restrictions on experimental practices and are often inapplicable, 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 many applications.

A variety of approaches have been proposed to estimate cellular 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 to solve the exact inverse problem, the underlying ODE system should be simulated for each individual. `SJT: I am not sure what you mean by the ''exact'' inverse problem.` Since the numbers of cells in these experiments typically exceed $\sim 10^4$ [15], exact inference is infeasible due to computational burden and instead the raw snapshot data are approximated by probability densities [12, 15, 18, 19]. Hasenauer et al. (2011) presents a Bayesian approach to inference for HODEs, which models the input parameter space using an ansatz of a mixture of densities of chosen types. The authors then 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 subpopulation structure in snapshot data with multiple-start local optimisation employed to maximise the non-convex likelihood, which they then apply to synthetic and real data from signalling pathway models. Loos et al. (2018) also uses mixture models to represent subpopulation structure and a maximum likelihood approach allowing estimation of within- and between-subpopulation variability which permits fitting to multivariate output distributions with complex correlation structures. Dixit et al. (2018) assigns observations into discrete bins, then uses a maximum entropy approach in a Bayesian framework to estimate cell variability.

Our framework is Bayesian although is distinct from the approach used to fit many dynamic models, since here we assume stochasticity arises solely due to variation in parameters across cells, not due to measurement noise. The approach is hence most suitable when measurement error is not a dominant source of 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 existing methods, CMC is 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 emerge as modes in the posterior parameter distributions. Like [19], CMC can fit multivariate snapshot data and unlike [12], does not require this data to be discretised into bins. As more experimental techniques elucidating single cell behaviour are developed, there will likely be more interest in models which can recapitulate observation snapshots. We argue that due to its simplicity and generality, CMC is a useful addition to the modeller's toolkit and can be used to perform inference on 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 used to generate samples from the posterior parameter distribution. In §4, we use CMC to estimate cell population heterogeneity in three systems of biological interest.

3 Method

In this section, we first develop a probabilistic framework that describes our inverse problem, before introducing the CMC algorithm in pseudocode (Algorithm 1). We also detail the workflow we have found helpful in using CMC to analyse cell snapshot data and suggest practical remedies to issues commonly encountered whilst using this approach (Figure 4). A glossary of variable names used in this paper is included as Table 1.

Experimental methods such as flow cytometry measure single cell characteristics at a given 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).

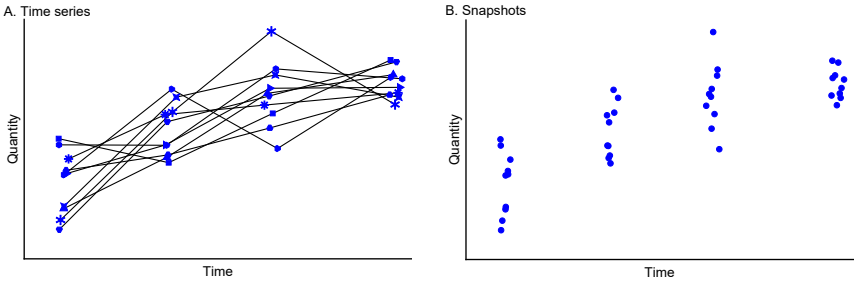

Figure 1: **Time series data (A) versus snapshot data (B) typical of single cell experiments.** In A, note cell identities are retained at each measurement time (indicated by individual plot markers) whereas in the snapshot data in B, either this information is lost or, more often, cells are destroyed 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 typically each element of the system corresponds to the concentration of a particular species. Our initial value problem is

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \boldsymbol{f}(\boldsymbol{x}(t);\boldsymbol{\theta}), \quad \boldsymbol{f}:\mathbb{R}^k\times\mathbb{R}^p\mapsto\mathbb{R}^k, \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0. \end{aligned} \tag{1}$$

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.

3.1 Snapshot data

We assume the variation present in snapshot data arises due to between-cell heterogeneity in the underlying parameters $\boldsymbol{\theta}$. Therefore, the evolution of

the underlying state of cell i is described by an idiosyncratic ODE,

$$\begin{aligned} \frac{d\boldsymbol{x}^{\{i\}}}{dt} &= \boldsymbol{f}\left(\boldsymbol{x}^{\{i\}}(t);\, \boldsymbol{\theta}^{\{i\}}\right), \quad \boldsymbol{f} : \mathbb{R}^k \times \mathbb{R}^p \mapsto \mathbb{R}^k, \\ \boldsymbol{x}^{\{i\}}(0) &= \boldsymbol{x}_0 \end{aligned} \tag{2}$$

where superscript $^{\{i\}}$ indicates the ith cell. The traditional (non-hierarchical) state-space approach to modelling dynamic systems supposes that measurement error introduces a degree of stochastic (random) variation in the output (Figure 2A). Our approach, by contrast, assumes any variation in outputs is solely due to variability in parameter values between cells (Figure 2B). Whether the assumption of "perfect" measurements is reasonable depends on experimental details of the system under investigation, but we argue our method nevertheless provides a useful approximation in cases where the signal to noise ratio is high.

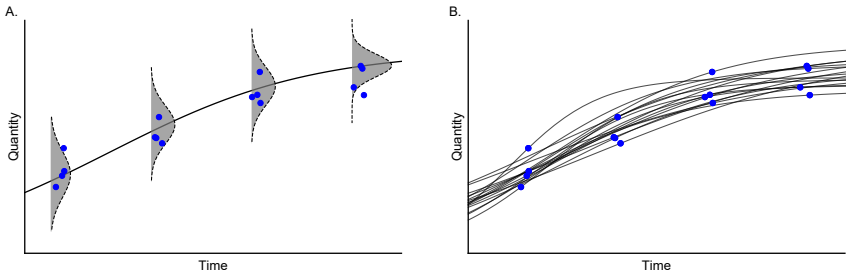

Figure 2: **Two ways to generate variation in measured outputs: the state-space model (A) versus the parameter heterogeneity model (B).** For non-hierarchical state-space models (A), there is assumed to be a single "true" latent state where observations result from a noisy measurement process (grey histograms). 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.

We suppose m quantities of interest (QOIs) are measured,

$$\boldsymbol{q}^\top = (q_1, q_2, \ldots, q_m) \in \mathbb{R}^m, \tag{3}$$

with n_j observations of each quantity, q_j. Distinct QOIs q_j, may correspond to different functionals of the solution at the same time or the same functional at different times. The observed data for QOI q_j at the corresponding time t_j consists of the n_j cellular measurements,

$$\boldsymbol{y}(t_j)^\top = \left(q_j(x^{\{1\}}(t_j)), q_j(x^{\{2\}}(t_j)), \ldots, q_j(x^{\{n_j\}}(t_j))\right) \in \mathbb{R}^{n_j}. \tag{4}$$

The raw snapshot data $\boldsymbol{X}$ is the collection of all measured QOIs,

$$\boldsymbol{X} = (\boldsymbol{y}(t_1), \boldsymbol{y}(t_2), \ldots, \boldsymbol{y}(t_m)) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \ldots \times \mathbb{R}^{n_m}. \tag{5}$$

The goal of inference is to characterise the probability distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ representing heterogeneity in the cellular processes. The numbers of cells sampled in typical experimental setups is large and following previous work, we choose to represent snapshot data $\boldsymbol{X}$ using probability distributions [12,

15, 18, 19]. In the first step of our workflow (Figure 4(i)), these distributions are approximated by a kernel density model, with support over the space of the QOI vector $\boldsymbol{q} \in \mathbb{R}^m$. We denote $\hat{\Phi}$ as the parameter estimates of the corresponding kernel density model $p(\boldsymbol{q}|\Phi)$ fitted to the raw snapshot data. We assume there is enough observational data that the estimated probability distributions are approximate sufficient statistics of the outputs, meaning $p(\boldsymbol{\theta}|\hat{\Phi}) \approx p(\boldsymbol{\theta}|\boldsymbol{X})$.

Variable	Definition	Dimension
$\boldsymbol{x}(t)$	ODE solution	$\mathbb{R}^k$
$\boldsymbol{\theta}$	ODE parameters	$\mathbb{R}^p$
$\boldsymbol{f}(\boldsymbol{x}(t); \boldsymbol{\theta})$	ODE RHS	$\mathbb{R}^k$
$\boldsymbol{x}^{\{i\}}(t)$	ODE solution for cell i	$\mathbb{R}^k$
$q_j = q_j(\boldsymbol{x}(t_j); \boldsymbol{\theta}) = q_j(\boldsymbol{\theta})$	quantity of interest (QOI) j	$\mathbb{R}^1$
$\boldsymbol{q}^\top = (q_1, \ldots, q_m)$	m distinct QOIs	$\mathbb{R}^m$
$q_j^{\{i\}} = q_j(\boldsymbol{x}^{\{i\}}(t_j))$	QOI j for cell i	$\mathbb{R}^1$
$\boldsymbol{y}_j^\top = \left(q_j^{\{1\}}, \ldots q_j^{\{n_j\}}\right)$	QOI j for cells $1, \ldots, n_j$	$\mathbb{R}^{n_j}$
$\boldsymbol{X} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_m)$	"snapshot" of all QOIs	$\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times ... \times \mathbb{R}^{n_m}$
Φ	parametrises output target distribution $p(\boldsymbol{q}\|\Phi)$	$\mathbb{R}^m$
Ξ	parametrises prior parameter distribution $p(\boldsymbol{\theta}\|\Xi)$	$\mathbb{R}^p$
Ψ	parametrises prior output distribution $p(\boldsymbol{q}\|\Psi)$	$\mathbb{R}^p$
$\hat{a}$	estimates of any quantity a	-
$\Omega(\boldsymbol{z})$	parameter space mapping to $\boldsymbol{q} = \boldsymbol{z}$	$\mathbb{R}^{\leq p}$
$\mathcal{V}(\boldsymbol{z})$	volume of $\Omega(\boldsymbol{z})$	$\mathbb{R}^+$
V	volume of (bounded) parameter space	$\mathbb{R}^+$

Table 1: **Glossary of variable names used in this paper.**

3.2 Theoretical development of CMC

We consider the underdetermined case where $m < p$, so that each specific quantity of interest $\tilde{\boldsymbol{q}}$ (say) can be generated from a non-singular set of parameter values, which we term iso-output contour regions: $\Omega(\tilde{\boldsymbol{q}}) = \{\boldsymbol{\theta} : \boldsymbol{q}(\boldsymbol{\theta}) = \tilde{\boldsymbol{q}}\}$. In general, these contours have "volumes" $\mathcal{V}(\tilde{\boldsymbol{q}})$ which depend on the chosen output value $\tilde{\boldsymbol{q}}$ (Figure 3). Any algorithm which samples parameter values in order to generate a given output target must account for the differential volumes of these sets, otherwise sampling will be biased towards iso-output contours with larger volumes [20]. The problem in most applied problems however is that we do not know *a priori* the volumes of iso-output contours and so they must be estimated. The following analysis provides a brief introduction to a probabilistic formulation of underdetermined inverse problems (see our companion paper [20] for a more comprehensive discussion) and, in doing so, suggests a sampling approach for estimating the volumes of parameter space mapping to each output value, which forms the basis of CMC.

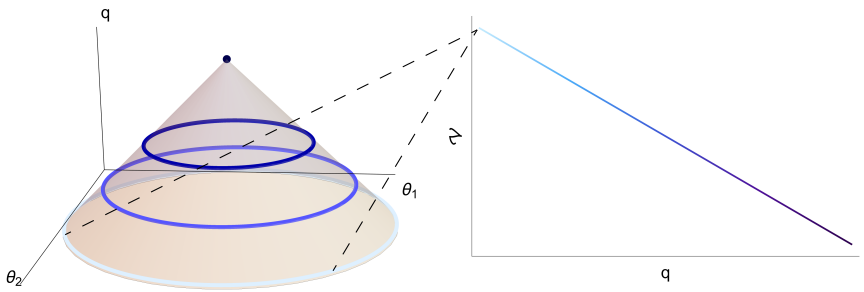

Figure 3: **Left: An example output function $q(\theta_1, \theta_2)$ along with iso-output contours indicated (coloured lines) . Right: The "volume" of output contours as a function of output value.** Note that here, since the input space is two dimensional, the "volume" of each output value corresponds to a length of an iso-output contour.

Solving our inverse problem requires determining the posterior distribution of parameter values $p(\boldsymbol{\theta}|\hat{\Phi})$ which, when used as input to the forward map, results in the target distribution $p(\boldsymbol{q}|\hat{\Phi})$. To derive the posterior parameter distribution, we consider the joint density of parameters and QOIs $p(\boldsymbol{\theta}, \boldsymbol{q}|\hat{\Phi})$. This can be decomposed in two ways,

$$p(\boldsymbol{\theta}, \boldsymbol{q}|\Phi) = p(\boldsymbol{\theta}|\boldsymbol{q}, \Phi) \times p(\boldsymbol{q}|\Phi) = p(\boldsymbol{q}|\boldsymbol{\theta}, \Phi) \times p(\boldsymbol{\theta}|\Phi). \tag{6}$$

Rearranging to obtain the posterior parameter distribution,

$$p(\boldsymbol{\theta}|\Phi) = \frac{p(\boldsymbol{\theta}|\boldsymbol{q}, \Phi) \times p(\boldsymbol{q}|\Phi)}{p(\boldsymbol{q}|\boldsymbol{\theta}, \Phi)}. \tag{7}$$

Since the mapping from parameters to outputs is deterministic, $p(\boldsymbol{q}|\boldsymbol{\theta}, \Phi) = \delta(\boldsymbol{q}(\boldsymbol{\theta}))$, i.e., the Dirac delta function centred at $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{\theta})$. Thus eq. (7) becomes

$$p(\boldsymbol{\theta}|\Phi) = p(\boldsymbol{\theta}|\boldsymbol{q}(\boldsymbol{\theta}), \Phi) \times p(\boldsymbol{q}(\boldsymbol{\theta})|\Phi). \tag{8}$$

In the same way that a single output value can be caused by any member of a set of parameter values, a target output distribution $p(\boldsymbol{q}|\Phi)$ can be caused by any member of a set of parameter distributions. To ensure uniqueness of the "posterior" parameter distributions, our probabilistic framework therefore requires that we specify "prior" distributions for the parameters, as in more traditional Bayesian inference. In what follows, we assume the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{q}, \Phi)$ is independent of the data, i.e., $p(\boldsymbol{\theta}|\boldsymbol{q}, \Phi) = p(\boldsymbol{\theta}|\boldsymbol{q})$, and thus represents a conditional "prior" which can be manipulated using Bayes' rule as,

$$p(\boldsymbol{\theta}|\boldsymbol{q}(\boldsymbol{\theta})) = \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{q}(\boldsymbol{\theta}))}. \tag{9}$$

This results in the form of the posterior parameter distribution targeted by our sampling algorithm,

$$p(\boldsymbol{\theta}|\hat{\Phi}) = \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{q}(\boldsymbol{\theta}))} p(\boldsymbol{q}(\boldsymbol{\theta})|\hat{\Phi}). \tag{10}$$

Again, we defer to our companion piece [20] for detailed explanation of eqs. (9) and (10) 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. (9) is the prior induced on output space by the prior over parameter space. For a uniform prior on parameter values, this is,

$$p(\boldsymbol{\theta}|\boldsymbol{q}(\boldsymbol{\theta})) = \frac{1}{\mathcal{V}(\boldsymbol{q}(\boldsymbol{\theta}))}, \tag{11}$$

where $\mathcal{V}(\boldsymbol{q}(\boldsymbol{\theta}))$ is the volume of parameter space occupied by the iso-output contour $\Omega(\boldsymbol{q}(\boldsymbol{\theta}))$. Therefore a uniform prior over parameter space implies a prior structure where all parameter values producing the same output are given equal weighting.

3.3 Implementation of CMC

Except for some toy examples, the denominator of eq. (9) cannot be calculated, and exact sampling from the posterior parameter distribution of eq. (10) is not, in general, possible. We propose instead a computationally efficient sampling method to estimate $p(\boldsymbol{q}(\boldsymbol{\theta}))$, 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 $\boldsymbol{q}(\boldsymbol{\theta})$. This step involves repeated independent sampling from the prior distribution of parameters $\boldsymbol{\theta}^{\{i\}} \sim p(\boldsymbol{\theta}|\Xi)$, where we condition on Ξ parameterising the prior probability density. Each parameter sample is then mapped to an output value $\boldsymbol{q}^{\{i\}} = \boldsymbol{q}(\boldsymbol{\theta}^{\{i\}})$. The collection of output samples is then fitted using a vine copula kernel density estimator (KDE) [21], $(\boldsymbol{q}^{\{1\}}, ..., \boldsymbol{q}^{\{N_1\}}) \sim p(\boldsymbol{q}|\hat{\Psi})$. Throughout the course of development of CMC, we have tested many KDE methods and 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. (10) with the estimated density $p(\boldsymbol{q}(\boldsymbol{\theta})|\hat{\Psi})$ replacing its corresponding estimand (Algorithm 1; Figure 4(iii)),

$$p(\boldsymbol{\theta}|\hat{\Phi}, \Xi, \hat{\Psi}) = \frac{p(\boldsymbol{\theta}|\Xi)}{p(\boldsymbol{q}(\boldsymbol{\theta})|\hat{\Psi})}\, p(\boldsymbol{q}(\boldsymbol{\theta})|\hat{\Phi}). \tag{12}$$

The final step in CMC is to compare output samples generated by MCMC 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 mapped to the output space, corresponds to the target distribution $p(\boldsymbol{q}|\hat{\Psi})$. In developing CMC, we found that a finite sample of modest size for both steps of CMC results in parameter samples that, when transformed, often represented reasonable approximations of the target. There are however occasions when this is not the case and this final confirmatory step is indispensable since it frequently highlights inadequacies in contour volume estimation or 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.

3.4 Workflow and CMC algorithm

A graphical illustration of the complete CMC workflow is provided in Figure 4. All variables are defined in Table 1. The CMC algorithm is provided in Algorithm 1. For simplicity, in this implementation MCMC sampling is performed via the Random Walk Metropolis algorithm, but for the examples in §4, we use an adaptive MCMC algorithm [22].

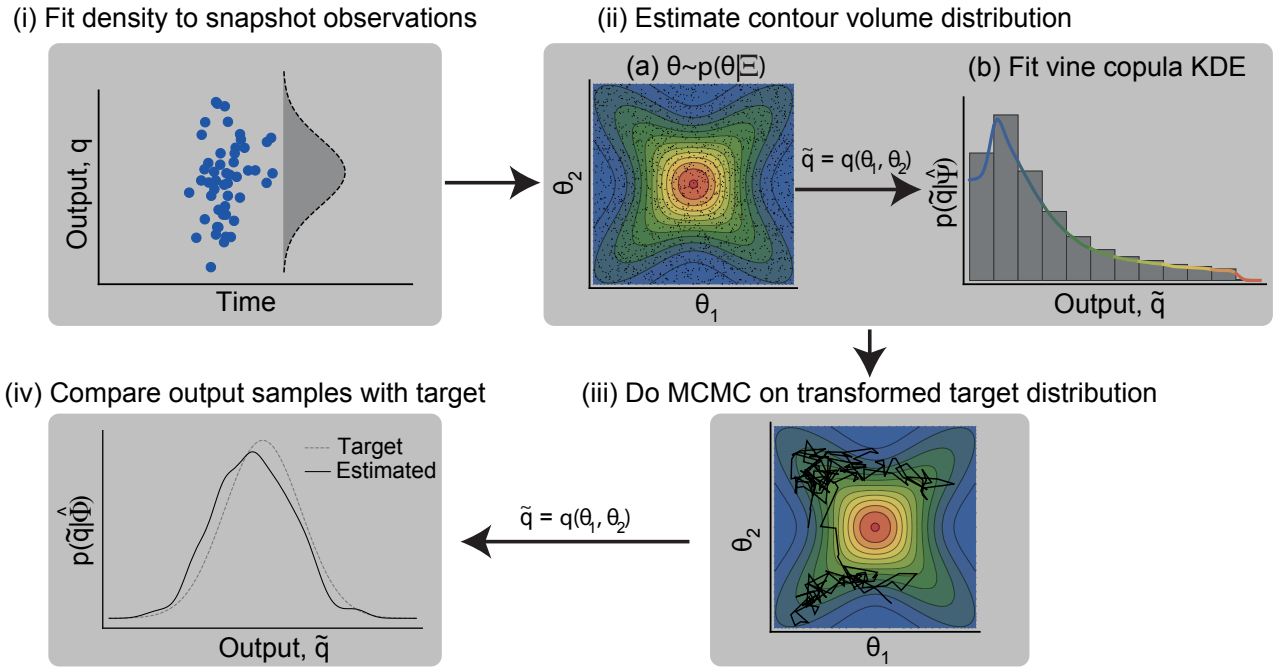

Figure 4: **The workflow for using Contour Monte Carlo to estimate cell population heterogeneity.** The distribution targeted in (iii) is given by eq. (12).

Algorithm 1 Pseudocode for the Contour Monte Carlo algorithm for sampling from the posterior parameter distribution of eq. (12).

procedure CMC($\boldsymbol{X}, \Xi, N_1, N_2$) ▷ Sample from posterior parameter distribution
  $\hat{\Phi} = \text{SnapshotEstimator}(\boldsymbol{X})$
  $\hat{\Psi} = \text{ContourVolumeEstimator}(\Xi, N_1)$
  $\left(\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2\}}\right) = \text{MCMC}(\hat{\Phi}, \Xi, \hat{\Psi}, N_2)$
  converged = $\text{CompareOutputToTarget}((\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2\}}), \hat{\Phi})$
  while converged$\neq$1 **do** ▷ Rerun contour volume estimation (if necessary modify vine copula KDE hyperparmeters) and/or MCMC, with larger sample sizes if required
    $\hat{\Psi} = \text{ContourVolumeEstimator}(\Xi, N_1'),\ N_1' \geq N_1$
    $\left(\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2'\}}\right) = \text{MCMC}(\hat{\Phi}, \Xi, \hat{\Psi}, N_2'),\ N_2' \geq N_2$
    converged = $\text{CompareOutputToTarget}((\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2'\}}), \hat{\Phi})$
    $N_1 \leftarrow N_1',\ N_2 \leftarrow N_2'$
  end while
  return $\left(\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2\}}\right)$
end procedure

procedure SnapshotEstimator($\boldsymbol{X}$) ▷ Fit snapshots with kernel density estimator (KDE)
  $\boldsymbol{X} \sim p(\boldsymbol{q}|\hat{\Phi})$
  return $\hat{\Phi}$
end procedure

procedure ContourVolumeEstimator(Ξ, N_1) ▷ Estimate volume of contours
  for i in $1 : N_1$ **do**
    $\boldsymbol{\theta}^{\{i\}} \sim p(\boldsymbol{\theta}|\Xi)$ ▷ Sample from prior density
    $\boldsymbol{q}^{\{i\}} = \boldsymbol{q}(\boldsymbol{\theta}^{\{i\}})$ ▷ Calculate corresponding output value
  end for
  $\left(\boldsymbol{q}^{\{1\}}, \ldots, \boldsymbol{q}^{\{N_1\}}\right) \sim p(\boldsymbol{q}|\hat{\Psi})$ ▷ Fit vine copula KDE
  return $\hat{\Psi}$
end procedure

procedure MCMC($\hat{\Phi}, \Xi, \hat{\Psi}, N_2$) ▷ 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
    ▷ Calculate Metropolis acceptance ratio
    $r = p(\boldsymbol{\theta}^{\{i\}'}|\Xi)\ p(\boldsymbol{q}(\boldsymbol{\theta}^{\{i-1\}})|\hat{\Psi})\ p(\boldsymbol{q}(\boldsymbol{\theta}^{\{i\}'})|\hat{\Phi}) / \left[p(\boldsymbol{\theta}^{\{i-1\}}|\Xi)\ p(\boldsymbol{q}(\boldsymbol{\theta}^{\{i\}'})|\hat{\Psi})\ p(\boldsymbol{q}(\boldsymbol{\theta}^{\{i-1\}})|\hat{\Phi})\right]$
    $u \sim U(0,1)$ ▷ 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 $\left(\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2\}}\right)$
end procedure

procedure CompareOutputToTarget($(\boldsymbol{\theta}^{\{1\}}, ..., \boldsymbol{\theta}^{\{N_2\}}), \hat{\Phi}$) ▷ Check output distribution close to target
  for i in $1 : N_2$ **do**
    $\tilde{\boldsymbol{q}}_i = \boldsymbol{q}(\boldsymbol{\theta}^{\{i\}})$ ▷ Compute QOIs for each parameter sample
  end for
  if $p(\tilde{\boldsymbol{q}}) \approx p(\tilde{\boldsymbol{q}}|\hat{\Phi})$? **then** ▷ Compare sampled output distribution with target
    return 1 ▷ If sufficiently close then converged
  else
    return 0
  end if
end procedure

In generating our results in §4, 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 used adaptive covariance MCMC (see SOM of [22]) to sample from the target distribution, as we have found that it provides a considerable speed-up over Random Walk Metropolis [23, 24]. We also use the Gelman-Rubin convergence statistic $\hat{R}$ which provides a heuristic measurement of convergence [24, 25], and use a threshold of $\hat{R} \leq\sim 1.1$ to diagnose convergence.

To solve the forward model of each differential equation, we used Julia's inbuilt "solve" method for ODE models, which automatically chooses an efficient inbuilt solver [26].

4 Results

In this section, we use CMC to estimate the posterior parameter distribution for three biological systems. In all but one of the examples, 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 mapped to outputs, recapitulates the target density. To accompany the text, we provide the Julia notebook used to generate the results. A table of priors used for each example is provided in Table 3.

4.1 Growth factor model

We first consider the "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 an exogenous ligand L. The governing dynamics are determined by the following system,

$$\frac{dR}{dt} = R_T k_{deg} + k_1 LR(t) + k_{-1} P(t) - k_{deg} R(t) \tag{13}$$

$$\frac{dP}{dt} = k_1 LR(t) - k_{-1} P(t) - k^*_{deg} P(t), \tag{14}$$

with initial conditions,

$$R(0) = 0.0, \qquad P(0) = 0.0,$$

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 these processes. We denote the solution of eq. (14) as $P(t; \boldsymbol{\theta}, L)$ and seek to determine the parameter distribution consistent with an output distribution,

$$\boldsymbol{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} P(10; \boldsymbol{\theta}, 2) \\ P(10; \boldsymbol{\theta}, 10) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 2 \times 10^4 \\ 3 \times 10^4 \end{pmatrix}, \begin{pmatrix} 1 \times 10^5 & 0 \\ 0 & 1 \times 10^5 \end{pmatrix} \right]. \tag{15}$$

4.1.1 Uniform prior

To start, we specify a uniform prior for each of the five parameters, with bounds given in Table 3, and use CMC to estimate the posterior parameter

distribution. In Figure 5A, we show the sampled outputs (blue points) versus the contours of the target distribution (black solid closed curves), illustrating a good correspondence between the sampled and target densities. Above and to the right of the main panel, we also display the marginal target densities (solid black lines) versus kernel density estimator reconstructions of the output marginals from the CMC samples (dashed blue lines), which again highlights the fidelity of the CMC sampled density to the target.

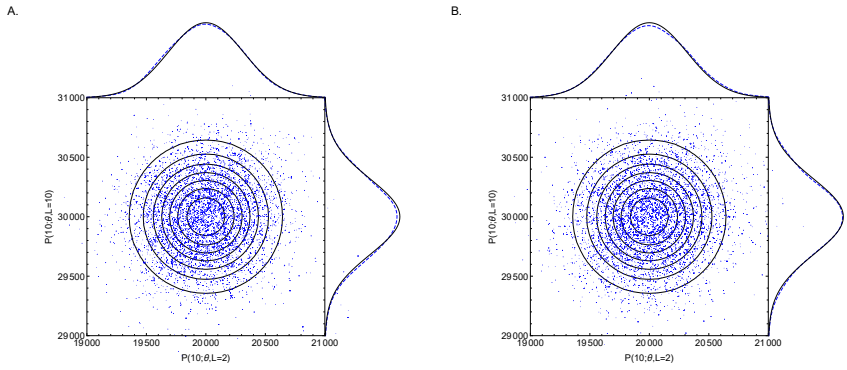

Figure 5: **The target joint output distribution (solid contour lines) and target marginal distributions (solid lines; above and to side of figure) versus outputs sampled by CMC (blue points) and reconstructed marginals (dashed lines) for (A) uniform and (B) Gaussian parameter priors.** In CMC, 100,000 independent samples were used in the "ContourVolumeEstimator" step and 10,000 MCMC samples across each of 4 Markov chains were used in the second step, with the first half of the chains discarded as "warm-up" [24]. For the reconstructed marginal densities in the plots, we use Mathematica's "SmoothKernelDistribution" function specifying bandwidths of 100 with Gaussian kernels [27].

`SJT: Blue dashed lines are barely visible. Please increase font size of axis labels and values.`

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. A given level of bound ligands can be generated in many different ways. Not surprisingly, it is the *ratio* of the forward and reverse reaction rates, k_1 and k_{-1} respectively, that is of greatest importance, and because of this, the distribution representing cell process heterogeneity contains linear positive correlations between these parameters.

In Figure 6B, we show the posterior parameter distribution for k_{deg}, the rate of degradation of ligand-free cell surface receptors and R_T, which dictates the rate of introduction of ligand-free cell surface receptors. This plot shows more concentrated posterior mass than in Figure 6A. Why can we better resolve (k_{deg}, R_T) compared to (k_1, k_{-1}) from our measurements? To answer this, it is useful to calculate the sensitivity of $P(t; \boldsymbol{\theta}, L)$ to changes in each of the parameters. To account for the differing magnitudes of each parameter, we calculate elasticities, the proportional changes in measured output for a proportional change in parameter values, using the forward

sensitivities method described in [28], and these are shown in Figure 7. When the exogenous ligand is set $L = 2$, these indicate the active ligand-bound receptor concentration is most elastic to changes in R_T and k_{deg}, meaning that their range is more restricted by the output measurement than for k_1 and k_{-1}, which have much smaller elasticities at both $t = 2$ and especially at $t = 10$. In Table 2, we show the posterior quantiles for the estimated parameters and, in the last column, indicate the ratio of the 25%-75% posterior interval widths to the uniform prior range for each parameter. These were strongly negatively correlated with the magnitude of the elasticities for each parameter ($\rho = 0.95$, $t = -5.22$, $df = 3$, $p = 0.01$ Pearson's product-moment correlation), indicating the utility of sensitivity analyses for optimal experimental design. We would suggest however that CMC can also be used for this purpose, using synthetic data in place of real measurements.

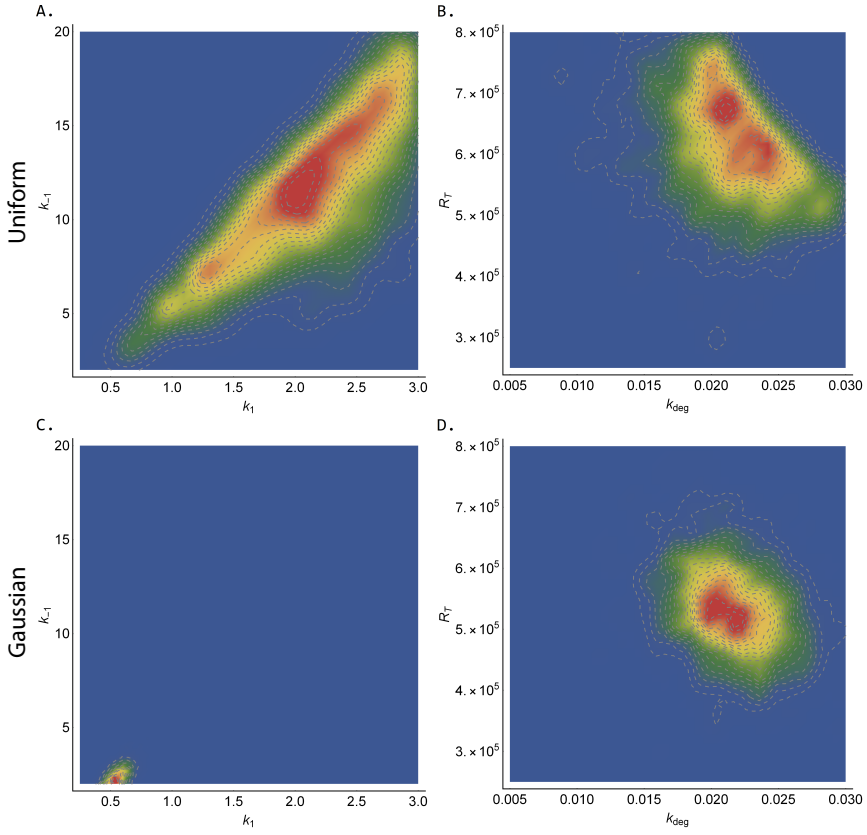

Figure 6: **Left-column: Joint posterior distributions of (k_1, k_{-1}) and right-column: (k_{deg}, R_T) for the growth factor model estimated by CMC sampling using uniform priors (top row) and Gaussian priors (bottom row).** See Figure 5 caption for CMC details and Table 3 for the priors used. Red (blue) indicates areas of relatively high (low) probability density.

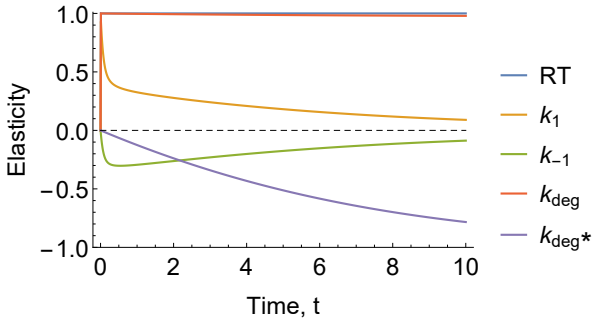

Figure 7: **Elasticities of the level of active ligand-bound receptors P with respect to each parameter against time.** When calculating the elasticities of each parameter, the other parameters were set to their posterior medians given in Table 2 and $L = 2$.

4.1.2 Gaussian prior

For an underdetermined model where the number of QOIs m is less than the number of parameters p, there is typically a non-singular set of parameter distributions 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. It is also preferable to allow priors to influence estimates in studies of cellular heterogeneity, since this allows incorporation of pre-existing biological knowledge with compensatory reductions in estimator variance. CMC accommodates different prior choices, with both the "ContourVolumeEstimation" step and the acceptance ratio in the "MCMC" step (Algorithm 1) being affected in such a way that posterior parameter distribution maps to the same output target.

We now use CMC to estimate the posterior parameter distribution when changing from uniform priors to more concentrated Gaussian priors (prior hyperparameters shown in Table 3). As desired, the target output distribution appears invariant (Figure 5B) although with substantial changes in the posterior parameter distribution (Figure 6C&D). In particular, the posterior distribution obtained from the Gaussian prior are narrower compared to the uniform case (rightmost column of Table 2). The differences in posterior distributions due to changes to priors will be more marked when data provides less information on the underlying process. This is readily apparent in comparing the dramatic change from Figure 5A) to 5C) with the more nuanced change from Figure 5B) to 5D).

4.2 Michaelis-Menten kinetics

In this section, we use CMC to invert output measurements from the Michaelis-Menten model of enzyme kinetics (see, for example, [29]); illustrating the capability of CMC to resolve population substructure from multimodality of the output distribution. The Michaelis-Menten model of enzyme kinetics describes the dynamics of concentrations of an enzyme E,

Parameter	Quantiles					Posterior 25%-75% conc.
	2.5%	25%	50%	75%	97.5%	
Uniform prior						
R_T	441,006	548,275	606,439	677,055	772,484	23%
k_1	0.90	1.69	2.17	2.56	2.95	32%
k_{-1}	4.35	8.35	11.23	14.23	18.71	33%
k_{deg}	0.013	0.019	0.021	0.024	0.029	20%
k^*_{deg}	0.20	0.34	0.40	0.44	0.49	27%
Gaussian prior						
R_T	408,396	487,372	529,558	577,970	678,632	16%
k_1	0.39	0.49	0.54	0.60	0.70	4%
k_{-1}	1.39	1.92	2.26	2.63	3.35	4%
k_{deg}	0.016	0.020	0.022	0.024	0.027	16%
k^*_{deg}	0.22	0.29	0.33	0.38	0.46	21%

Table 2: **Estimated quantiles from CMC samples for the growth factor model with uniform and Gaussian priors.** The last column indicates the proportion of the uniform prior bounds occupied by the 25%-75% posterior interval in each case. The prior hyperparameters used in each case are given in Table 3.

a substrate S, an enzyme-substrate complex C, and a product P,

$$\begin{aligned}
\frac{dE}{dt} &= -k_f E(t)S(t) + k_r C(t) + k_{cat} C(t), \\
\frac{dS}{dt} &= -k_f E(t)S(t) + k_r C(t), \\
\frac{dC}{dt} &= \quad k_f E(t)S(t) - k_r C(t) - k_{cat} C(t), \\
\frac{dP}{dt} &= \quad k_{cat} C(t),
\end{aligned} \tag{16}$$

with initial conditions,

$$E(0) = E_0,\ S(0) = S_0,\ C(0) = C_0,\ P(0) = P_0, \tag{17}$$

where k_f is the rate of the forward reaction $E + S \to C$, k_r is the rate of the reverse reaction $C \to E + S$, and k_{cat} is the catalytic rate at which the product is formed by the reaction $C \to E + P$.

4.2.1 Bimodal output distribution

When subpopulations of cells, each with distinct dynamics, are thought to exist, determining their characteristics - the proportions of cells in each cluster, their distinct parameter values, and so on - is often of key interest [15,19]. Before formal inference occurs, multimodality of the output distribution may signal the existence of fragmented subpopulations of cells and to exemplify this we target a bimodal bivariate Gaussian distribution for measurements of the level of enzyme and substrate at $t = 1$ and $t = 2$

respectively,

$$\boldsymbol{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} E(2.0;\boldsymbol{\theta}) \\ S(1.0;\boldsymbol{\theta}) \end{pmatrix} \sim p(\boldsymbol{q};\boldsymbol{\mu}_1,\Sigma_1,\boldsymbol{\mu}_2,\Sigma_2) \\ = \frac{1}{2}\left(\mathcal{N}(\boldsymbol{q};\boldsymbol{\mu}_1,\Sigma_1) + \mathcal{N}(\boldsymbol{q};\boldsymbol{\mu}_2,\Sigma_2)\right), \tag{18}$$

where $\boldsymbol{\theta} = (k_f, k_r, k_{cat})$. The parameters of the Gaussian mixture components are,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2.2 \\ 1.6 \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} 0.018 & -0.013 \\ -0.013 & 0.010 \end{pmatrix},$$
$$\boldsymbol{\mu}_2 = \begin{pmatrix} 2.8 \\ 1.0 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 0.020 & -0.010 \\ -0.010 & 0.020 \end{pmatrix}.$$

In what follows, we specify uniform priors on each element of $\boldsymbol{\theta}$ (see Table 3). Using a modest number of samples in each step, CMC well-approximates the output target distribution (Figure 8A). Without specifying *a priori* information on the subpopulations of cells, two distinct clusters of cells emerged from application of CMC (orange and blue points in Figure 8B), each corresponding to distinct modes of the output distribution (corresponding coloured points in Figure 8A). It is worth noting however that the issues inherent with MCMC sampling of multimodal distributions similarly apply here and so, whilst adaptive MCMC [22] sufficed to explore this posterior surface, it may be necessary to use MCMC methods more robust to such geometries (for example, population MCMC [30]).

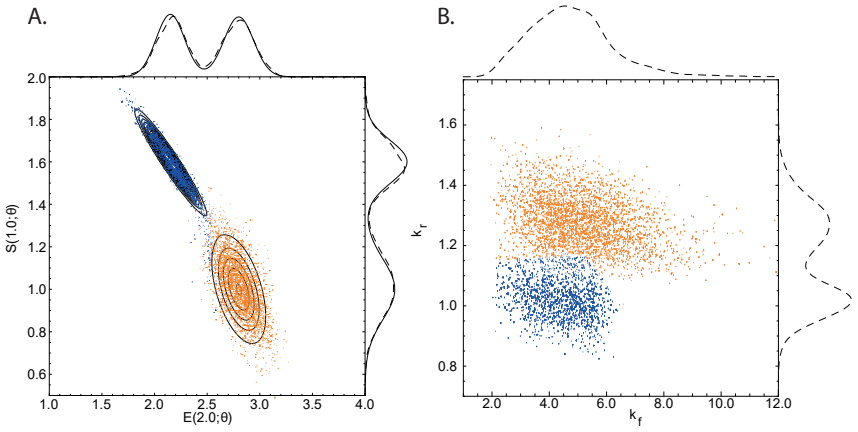

Figure 8: **Michaelis-Menten model CMC results: (A) Bimodal target distribution $\boldsymbol{q}$ (solid contour lines) versus output samples (points) and (B) posterior parameter samples (points).** The solid and dashed lines above and to the side of panel A indicate the target and estimated marginal output distributions, respectively. The orange (blue) points in A were generated by the orange (blue) parameter samples in B. See Figure 5 caption for CMC details. Mathematica's "SmoothKernelDistribution" function [27] with Gaussian kernels was used to construct marginal densities with: (A) default bandwidths, and (B) bandwidths of 0.3 (horizontal axis) and 0.03 (vertical axis). Mathematica's "ClusteringComponents" function [27] was used to identify clusters in B.

4.2.2 Four-dimensional output distribution

Loos et al. (2018) consider a multidimensional output distribution, with correlations between system characteristics that evolve over time. Our approach allows arbitrary covariance structure between measurements, and to exemplify this, we now target a four-dimensional output distribution, with paired measurements of enzyme and substrate at $t = 1$ and $t = 2$,

$$\begin{aligned} \boldsymbol{q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} &= \begin{pmatrix} E(1.0;\boldsymbol{\theta}) \\ S(1.0;\boldsymbol{\theta}) \\ E(2.0;\boldsymbol{\theta}) \\ S(2.0;\boldsymbol{\theta}) \end{pmatrix} \\ &\sim \mathcal{N}\left[\begin{pmatrix} 0.5 \\ 2.8 \\ 0.9 \\ 1.4 \end{pmatrix}, \begin{pmatrix} 0.02 & -0.05 & 0.04 & -0.05 \\ -0.05 & 0.30 & -0.15 & 0.20 \\ 0.04 & -0.15 & 0.12 & -0.17 \\ -0.05 & 0.20 & -0.17 & 0.30 \end{pmatrix} \right]. \end{aligned} \tag{19}$$

Since this target has four QOIs, and the Michaelis-Menten model has three rate parameters (k_f, k_r, k_{cat}), the system is over-identified and so CMC cannot be straightforwardly applied. Instead, we allow the four initial states (E_0, S_0, C_0, P_0) to be uncertain quantities, bringing the total number of parameters to seven.We set uniform priors on all parameters (see Table 3). In order to check that the model and priors were consistent with the output distribution given by eq. (19), we plotted the output measurements used to estimate contour volumes (in the first step of the "ContourVolumeEstimator" method in Algorithm 1) against the target (Figure 9). Since the main support of the densities (black contours) lies within a region of output space reached by independent sampling of the priors (blue points), this indicated the target could feasibly be generated from this model and priors, and we proceeded to estimation by CMC.

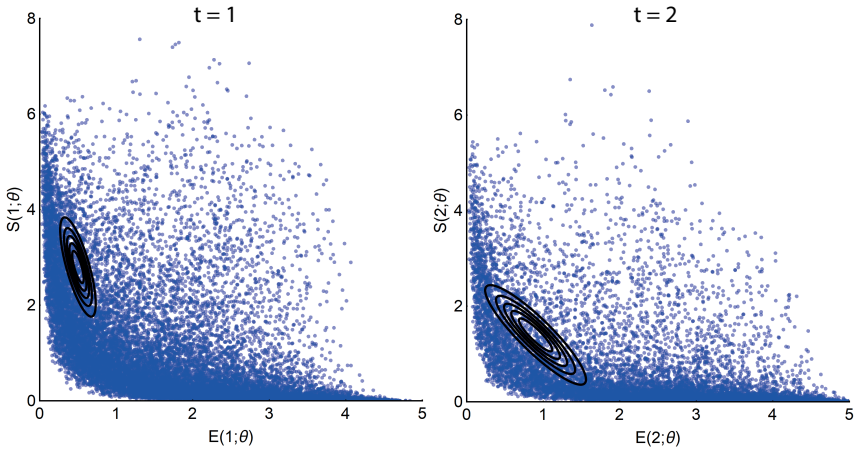

Figure 9: **QOIs shown as blue points for (q_1, q_2) (left panel) and (q_3, q_4) (right panel) obtained by independently sampling the priors $p(\boldsymbol{\theta}|\Xi)$ of the seven parameter Michaelis-Menten model versus the target distribution (black solid contours).** We show 20,000 output samples, where each set of four measurements was obtained from a single sample of all parameters. The output target distribution shown by the contours corresponds to the marginal densities of each pair of enzyme-substrate measurements given by eq. (19).

Figure 10 plots the output samples of enzyme and substrate from the last step of CMC for $t = 1$ (blue points) and $t = 2$ (orange points) versus the contours (black lines) of the joint marginal distributions of eq. (19). The distribution of paired enzyme-substrate samples illustrates that the CMC output sampling distribution approximated the target density, itself representing dynamic evolution of the covariance between enzyme and substrate measurements. The target marginal distributions (solid lines) along with their approximations from kernel density estimation (dashed lines) are also shown above and along the RHS of the main panel of Figure 10, and largely indicate correspondence.

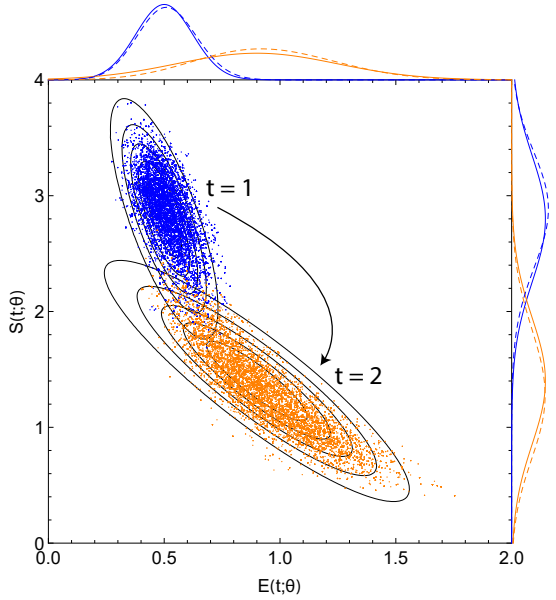

Figure 10: **Posterior output samples from CMC (coloured points) versus the contour plots of the joint marginal distributions of eq. (19) (black solid lines).** Output functionals for (q_1, q_2) and (q_3, q_4) are given by blue and orange points, respectively. Enzyme and substrate measurements are given by the horizontal and vertical axes, respectively. The solid and dashed coloured lines outside of the panels indicate exact target marginals of eq. (19) and those estimated from CMC, respectively. In the "ContourVolumeEstimator" step 200,000 independent samples were used and 10,000 samples across each of 4 Markov chains were used in the MCMC step, with the first half of the chains discarded as "warm-up" [24]. Mathematica's "SmoothKernelDistribution" function with Gaussian kernels [27] of bandwidths varying from 0.1 to 0.4 for the reconstructed marginal densities.

4.3 TNF signalling pathway

We now illustrate how CMC can be applied to an ODE system of larger size, the tumour necrosis factor (TNF) signalling pathway model introduced in [31] and used by [15] to illustrate a Bayesian approach to cell population variability estimation. The model incorporates known activating and inhibitory interactions between four key species within the TNF pathway: active caspase 8, x_1, active caspase 3, x_2, a nuclear transcription factor, x_3

and its inhibitor, x_4, such that

$$\begin{aligned}
\frac{dx_1}{dt} &= -x_1(t) + \frac{1}{2}\left[\beta_4(x_3(t))\alpha_1(u(t)) + \alpha_3(x_2(t))\right] \\
\frac{dx_2}{dt} &= -x_2(t) + \alpha_2(x_1(t))\beta_3(x_3(t)) \\
\frac{dx_3}{dt} &= -x_3(t) + \beta_2(x_2(t))\beta_5(x_4(t)) \\
\frac{dx_4}{dt} &= -x_4(t) + \frac{1}{2}\left[\beta_1(u(t)) + \alpha_4(x_3(t))\right],
\end{aligned} \tag{20}$$

with initial conditions,

$$x_1(0) = 0.0, \quad x_2(0) = 0.0, \quad x_3(0) = 0.29, \quad x_4(0) = 0.625. \tag{21}$$

The functions α_i and β_j represent activating and inhibitory interactions respectively,

$$\begin{aligned}
\alpha_i(z) &= \frac{z^2}{a_i^2 + z^2}, \quad i = 1, \ldots, 4, \\
\beta_j(z) &= \frac{b_j^2}{b_j^2 + z^2}, \quad j = 1, \ldots, 5,
\end{aligned} \tag{22}$$

and the parameters a_i for $i \in (1, 2, 3, 4)$ and b_j for $j \in (1, 2, 3, 4, 5)$ represent activation and inhibition thresholds. The function $u(t)$ represents a TNF stimulus represented by a top hat function,

$$u(t) = \begin{cases} 1, & \text{if } t \in [0, 2]. \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

In underdetermined models, a non-singular set of parameters can lead to the same combination of output values. A consequence of this unidentifiability is that we cannot perform "full circle" inference: that is, using a known parameter distribution to generate an output distribution does not result in that parameter distribution being recovered through inference. We illustrate this idea by generating an output distribution by varying a single parameter value between runs of the forward model (20) and performing inference on all nine system parameters, whilst collecting only two output measurements. Specifically, we vary $a_1 \sim \mathcal{N}(0.6, 0.05)$, whilst holding the other parameters constant,

$$(a_2, a_3, a_4, b_1, b_2, b_3, b_4, b_5) = (0.2, 0.2, 0.5, 0.4, 0.7, 0.3, 0.5, 0.4),$$

and measure $q_1 = x_1(2.0)$ and $q_2 = x_2(1.0)$ for each forward model simulation. In doing so, we obtain an output distribution well-approximated by the bivariate Gaussian distribution,

$$\begin{aligned}
\boldsymbol{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} &= \begin{pmatrix} x_1(2.0) \\ x_2(1.0) \end{pmatrix} \\
&\sim \mathcal{N}\left[\begin{pmatrix} 0.26 \\ 0.07 \end{pmatrix}, \begin{pmatrix} 2.1 \times 10^{-4} & 5.9 \times 10^{-5} \\ 5.9 \times 10^{-5} & 1.8 \times 10^{-5} \end{pmatrix} \right].
\end{aligned} \tag{24}$$

We now apply CMC to the target output distribution given by eq. (24) to estimate a posterior distribution over all nine parameters of eq. (20). Apart than for a few cases, the priors for each parameter were chosen to *exclude* the values that were used to generate the output distribution (see

Table 3), to illustrate the non-equivalence between the recovered posterior distribution and the data generating process. In Figure 11A, we plot the actual parameter values (horizontal axis) used in the true data generating process versus the estimated values (vertical axis). This illustrates that apart from a_1, where the estimated parameter values correspond well with the range of values used to generate the data, due to the choice of priors there is a disjunction between the actual and estimated parameter values. Because the model is underdetermined however, despite these differences, the corresponding output distribution well approximates the target (Figure 11B).

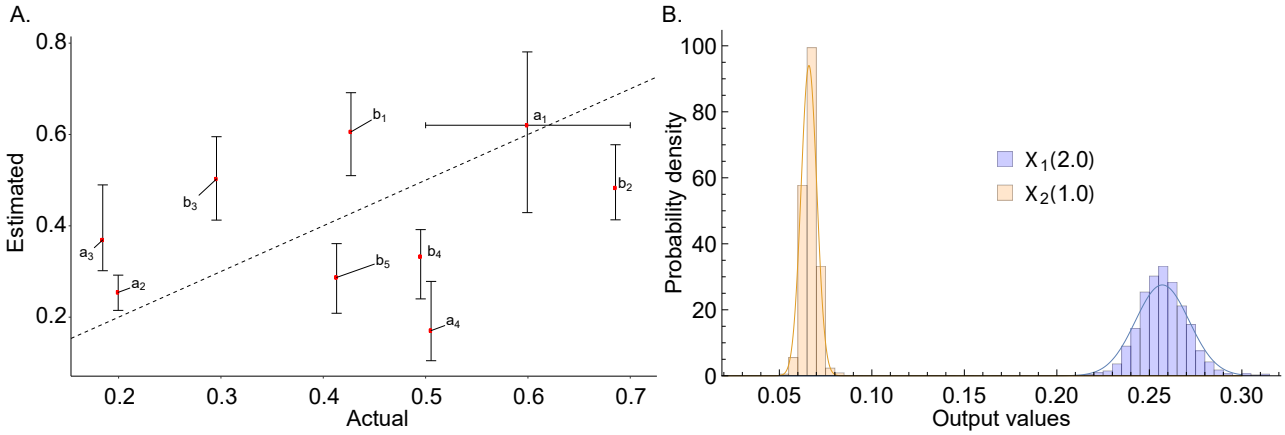

Figure 11: **TNF signalling pathway model: (A) Actual parameter values versus estimated quantiles for the output distribution of eq. (24) and (B) marginal output targets (solid lines) and sampled output distributions (histograms).** In A, in the vertical direction, red points indicate 50% posterior quantiles and upper and lower whiskers indicate 97.5% and 2.5% quantiles, respectively; in the horizontal direction, with the exception of a_1, red points indicate the parameter values used to generate the data; for a_1 the red point indicates the mean of the Gaussian distribution used to generate the data and the whiskers indicate its 95% quantiles. In CMC, 10,000 independent samples were used in the "ContourVolumeEstimator" step and 5,000 MCMC samples across each of 4 Markov chains were used in the second step, with the first half of the chains discarded as "warm-up" [24].

Cell populations may be well described by subpopulations which each evolve along characteristic trajectories over time. We now apply CMC to investigate a bimodal output distribution for the TNF signalling pathway model similar to that investigated by [15]. We aim to estimate the posterior parameter distribution mapping to the following output distribution,

$$\boldsymbol{q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}, \tag{25}$$

where,

$$
\begin{aligned}
q_1 &= \boldsymbol{x}_2(1.0) \sim \mathcal{N}(0.06, 0.01) \\
q_2 &= \boldsymbol{x}_2(2.0) \sim \frac{1}{2}\left(\mathcal{N}(0.1, 0.01) + \mathcal{N}(0.14, 0.01)\right) \\
q_3 &= \boldsymbol{x}_2(4.0) \sim \frac{1}{2}\left(\mathcal{N}(0.1, 0.01) + \mathcal{N}(0.20, 0.01)\right),
\end{aligned}
\tag{26}
$$

where the target distributions for $\boldsymbol{q}_2(2.0)$ and $\boldsymbol{q}_2(4.0)$ indicate mixtures of univariate Gaussians, and the priors used are described in Table 3. This target distribution, along with the unique trajectories obtained by applying the CMC algorithm, are shown in Figure 12. This figure illustrates that the bimodality of the output distribution induces CMC to estimate subpopulation structure in the parameter distribution without us explicitly specifying the number of clusters.

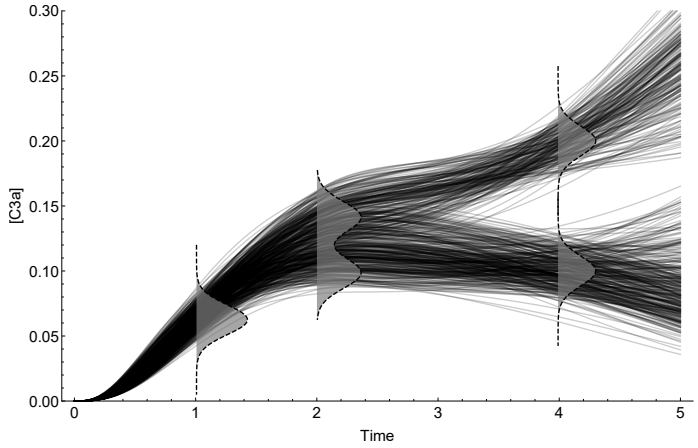

Figure 12: **The target output distribution (dashed plots with grey filling) and unique trajectories (black solid lines) obtained from the posterior parameter distribution.** In CMC, 10,000 independent samples were used in the "ContourVolumeEstimator" step and 5,000 MCMC samples across each of 4 Markov chains were used in the second step, with the first half of the chains discarded as "warm-up" [24].

Model	Target density	Parameter	Prior density	Prior p_1	Prior p_2
Growth factor	2D Gaussian	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	2D Gaussian	R_T	Gaussian	5×10^5	1×10^5
		k_1	Gaussian	0.5	0.1
		k_{-1}	Gaussian	3.0	1.0
		k_{deg}	Gaussian	0.02	0.005
		k^*_{deg}	Gaussian	0.3	0.1
Michaelis-Menten	bimodal Gaussian	k_f	uniform	0.2	15
		k_r	uniform	0.2	2.0
		k_{cat}	uniform	0.5	3.0
Michaelis-Menten	4D Gaussian	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
		C_0	uniform	0.0	0.2
		P_0	uniform	0.0	0.2
TNF signalling	bivariate Gaussian	a_1	uniform	0.4	0.8
		a_2	uniform	0.1	0.7
		a_3	uniform	0.3	0.7
		a_4	uniform	0.1	0.3
		b_1	uniform	0.5	0.7
		b_2	uniform	0.4	0.6
		b_3	uniform	0.4	0.6
		b_4	uniform	0.2	0.4
		b_5	uniform	0.2	0.4
TNF signalling	bimodal Gaussian	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 3: **The priors used for each problem in §4.** The parameters p_1 and p_2 indicate the prior hyperparameters: for uniform priors, these correspond to the lower and upper limits; for Gaussian priors, they correspond to the mean and standard deviation.

5 Discussion

Determining the cause of variability in cellular processes is crucial in many applications, ranging from bioengineering to drug development. In this

paper, we introduce a Bayesian method for estimating cellular heterogeneity from "snapshot" measurements of cellular properties, taken at discrete intervals throughout the experimental course. Our approach assumes what we term a "heterogeneous ordinary differential equation" (HODE) framework, in which biochemical processes in individual cells are assumed to follow dynamics governed by a common ODE, although with idiosyncratic differences in parameter values. In this framework, estimating heterogeneity in cellular processes amounts to determining the probability distributions over parameter values of the governing ODE. Our method of estimation is a two-step Monte Carlo sampling process we term "Contour Monte Carlo" (CMC) which does not require *a priori* specification of cell population substructure unlike other approaches. CMC can be used to process high volumes of individual cellular measurements since the framework involves fitting a kernel density estimator to raw experimental data and using these distributions rather than data as the target outcome. CMC also allows for arbitrary multivariate structure in the measurement space, meaning it can capture correlations that occur between the same cellular species at different timepoints or, for example, contemporaneous correlations between different cellular compartments. Being a Bayesian approach, CMC uses prior distributions over parameter values to ensure uniqueness of the posterior distribution, allowing pre-experimental knowledge to be used to improve estimation robustness. The flexible and robust framework that CMC provides means it can be used to perform automatic inference for wide-ranging systems of practical interest.

Our approach also provides a natural way to test that the process is working satisfactorily. Feeding the posterior parameter samples obtained by CMC into forward model simulations, results in a distribution over output values that can be compared to the target. Indeed, we have found this comparison indispensable in applying CMC in practice and include it as the last step in the CMC algorithm (Algorithm 1). Discrepancies between the target output distribution and samples from it by CMC can occur either as a result of poor estimates of the "contour volume distribution" in the first stage of the algorithm or due to insufficient MCMC samples in the second. Either of these issues can often be easily addressed and although kernel density estimation in high dimensional spaces remains an open research problem, we have found vine copula kernel density estimation works well for the dimensionality of output measurements we investigate here [21].

Failure to reproduce a given output distribution can also indicate that the generating model (the priors and the forward model) are incongruent with experimental results. This may either be due to misspecification of the ODE system or, that our assumption the process is deterministic is inappropriate. Our approach currently assumes that output variation is dominated by cellular variation in the parameter values of the underlying ODE, with measurement noise making a negligible contribution. Whether this is a reasonable assumption depends on the system under investigation and, more importantly, on experimental details. We recognise that neglecting measurement noise when it is an important determinant of the observed data means CMC will overstate cellular variation. It may also mean that some output distributions cannot be obtained by our model system (the HODEs without noise). Future work allowing inclusion of a stochastic noise process or, more generally, including stochastic cellular mechanisms is thus likely to be worthwhile.

Whilst we have labelled the approach we follow here as Bayesian, since it

involves explicit estimation of probability distributions and involves priors over parameter values, we recognise that it is not in the form typically utilised by exponents of this framework. This is because rather than aiming to formulate a model that describes output observations, our approach aims to recapitulate output *distributions*. Others [32], (including us [20]), have considered similar problems before; perhaps most notably by Albert Tarantola in his landmark work on inverse problem theory (see, for example, [33]), which has generated considerable interest in areas such as the geosciences [34, 35]. In Tarantola's framework, a joint input parameter and output space is considered, where prior knowledge and experimental theory combine elegantly to produce a posterior distribution whose marginal output distribution is a weighted "conjunction" of various sources of measurement.

The natural world is rife with variation. Mathematical models represent frameworks for understanding the causes of such variation. Typically, the state of biological knowledge is such that one effect, a given pattern of variation, has many possible causes, and observational or experimental data are necessary to apportion weight to each of them, in a process which amounts to solving an inverse problem. The approach we describe here follows the Bayesian paradigm of inverse problem solving whereby uncertainty in potential causes is reflected by probability distributions. Here, we illustrate the utility of our method by applying it to estimate cellular heterogeneity in biochemical processes however, it could equally be used to understand the inversion of systems modeled by an undeterdetermined input to output map in the form of an algebraic map, or a system of odes or pdes, arising in other areas. Contour Monte Carlo provides an automatic framework for performing inference on such underdetermined systems and the use of priors allows for robust and precise parameter estimation unattainable through the data alone.

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] M Ridley. *The red queen: sex and the evolution of human nature*. Penguin UK, 1994.

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

[3] F Delvigne, Q Zune, AR Lara, W Al-Soud, and SJ 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] PM Altrock, LL Liu, and F Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12):730, 2015.

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

[7] MB Elowitz, AJ Levine, ED Siggia, and PS Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[8] HH Chang, M Hemberg, M Barahona, DE Ingber, and S Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544, 2008.

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

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

[11] D Ramkrishna and MR Singh. Population balance modeling: current status and future prospects. *Annual Review of Chemical and Biomolecular Engineering*, 5:123–146, 2014.

[12] P Dixit, E Lyashenko, M Niepel, and D Vitkup. Maximum entropy framework for inference of cell population heterogeneity in signaling network dynamics. *bioRxiv*, page 137513, 2018.

[13] WG Telford, T Hawley, F Subach, V Verkhusha, and RG Hawley. Flow cytometry of fluorescent proteins. *Methods*, 57(3):318–330, 2012.

[14] AJ Hughes, DP Spelke, Z Xu, CC Kang, DV Schaffer, and AE Herr. Single-cell western blotting. *Nature Methods*, 11(7):749, 2014.

[15] J Hasenauer, S Waldherr, M Doszczak, N Radde, P Scheurich, and F Allgöwer. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12(1):125, 2011.

[16] O Hilsenbeck, M Schwarzfischer, S Skylaki, B Schauberger, PS Hoppe, D Loeffler, KD Kokkaliaris, S Hastreiter, E Skylaki, A Filipczyk, et al. Software tools for single-cell tracking and quantification of cellular and molecular properties. *Nature Biotechnology*, 34(7):703, 2016.

[17] FSO Fritzsch, C Dusny, O Frick, and A Schmid. Single-cell analysis in biotechnology, systems biology, and biocatalysis. *Annual Review of Chemical and Biomolecular Engineering*, 3:129–155, 2012.

[18] J Hasenauer, C Hasenauer, T Hucho, and FJ Theis. ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. *PLOS Computational Biology*, 10(7):e1003686, 2014.

[19] C Loos, K Moeller, F Fröhlich, T Hucho, and J 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.

[20] B Lambert, D Gavaghan, and SJ Tavener. Inverse sensitivity analysis of mathematical models avoiding the curse of dimensionality. *BioRxiv*, page 432393, 2018.

[21] T Nagler and C Czado. Evading the curse of dimensionality in non-parametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[22] RH Johnstone, ETY Chang, R Bardenet, TP De Boer, DJ Gavaghan, P Pathmanathan, RH Clayton, and GR 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.

[23] N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, and E Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

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

[25] A Gelman and DB Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

[26] J Bezanson, A Edelman, S Karpinski, and VB Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

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

[28] AC Daly, DJ Gavaghan, J Cooper, and SJ Tavener. Inference-based assessment of parameter identifiability in nonlinear biological models. *Journal of The Royal Society Interface*, 15, 2018.

[29] JD Murray. *Mathematical biology: I. An Introduction (interdisciplinary applied mathematics)(Pt. 1)*. New York, Springer, 2007.

[30] A Jasra, DA Stephens, and CC Holmes. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.

[31] M Chaves, T Eissing, and F Allgower. Bistable biological systems: a characterization through local compact input-to-state stability. *IEEE Transactions on Automatic Control*, 53(Special Issue):87–100, 2008.

[32] T Butler, J Jakeman, and T Wildey. Combining push forward measures and baye's rule to construct consistent solutions to stochastic inverse problems. *SIAM J. Sci. Comput.*, 40(2):A984–A1011, 2018.

[33] A Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. SIAM, 2005.

[34] K Mosegaard and A Tarantola. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447, 1995.

[35] T Vukicevic and D Posselt. Analysis of the impact of model nonlinearities in inverse problem solving. *Journal of the Atmospheric Sciences*, 65(9):2803–2823, 2008.