

A Monte Carlo method to estimate cell population heterogeneity from cell snapshot data

Ben Lambert^{1,2*}, David J. Gavaghan³, Simon Tavener⁴.

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: 2020-07-15 08:22

1 Abstract

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

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 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*. Nor are they amenable to standard nonlinear mixed effect methods, since a single observation per cell is typically too few to estimate parameter variability. Here, we introduce a computational sampling method named “Contour Monte Carlo” (CMC) for estimating mathematical model parameters from snapshot distributions, which is straightforward to implement and does not require cells be assigned to predefined categories. The CMC algorithm fits to snapshot probability distributions rather than raw data, which means its computational burden does not, like existing approaches, increase with the number of cells observed. Our method is appropriate for underdetermined systems, where there are fewer distinct types of observations than parameters to be determined, and where observed variation is mostly due to variability in cellular processes rather than experimental measurement error. This may be the case for many systems due to continued improvements in 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

32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

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 be costly in other circumstances. In biotechnological processes, heterogeneity in cellular function can reduce yields of biochemical products [3]. In human biology, variation across cells can enable pathologies to develop; it can also frustrate treatment of illness because key subpopulations are missed by medical interventions that target “average” cell properties. For example, cellular heterogeneity helps some cancerous tumours to persist [4] and can make tumours more likely to evolve resistance to chemotherapies [5]. To discern whether observed variation is

benign or requires remedy, methods of analysis are needed that can quantify
and help to understand its source.

Mathematical models are essential tools for understanding cellular systems, whose emergent properties are the result of a nexus of interactions between actors. Perhaps the simplest flavour of mathematical model used in biological systems is an ordinary differential equation (ODE) that aggregates individual actors into compartments according to structure or function, and seeks to model the mean behaviour of each compartment. Data from population-averaged experimental assays can determine whether such models faithfully reproduce system behaviours and can be used to understand the structure of complex metabolic, signalling and transcriptional networks. The worth of such “population average” ODE models depends on whether averages mask substantial differences in individual behaviour [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 that subpopulations of clonally-derived hematopoietic progenitor cells with low expression of a stem cell marker, diverged into a separate blood lineage from those with high expression [8].

Many modelling frameworks are available to describe cell population heterogeneity, with each posing different challenges for parameter inference. A recent review is presented in [9]. These approaches include modelling biochemical processes stochastically, where properties of ensembles of cells are represented by probability distributions that evolve according to chemical master equations. See [10] for a tutorial on stochastic simulation of reaction diffusion processes. Alternatively, population balance equations (PBEs) are typically partial integro-differential equations that determine the dynamics of the “number density” of differing cell types. In PBEs, cell properties are represented as points in \mathbb{R}^n , with each dimension corresponding to a different attribute. These attributes include parameters controlling cell life - for example, their rate of death and division, which vary according to a cell’s location in this “attribute” space. These functional differences control the rate at which cells progress through life, which is represented by a “flow” of cells from certain areas of attribute space to others - like chemicals diffusing down a concentration gradient. With PBEs, observed variation at a point in time is due to the initial spread of cells across attribute space coupled with the differing dynamics of cells in different areas of this space. See [11] for an introduction to PBEs.

Here, we suppose heterogeneity in quantities of interest across cells is generated by idiosyncratic variation in the rates of cellular processes. The modelling approach we follow is similar to that of [12] and is based on an ODE framework. In our model, each cell evolves according to an ODE, with its progression directed by parameters whose value varies between cells. To our knowledge, this flavour of model is unnamed, so, for sake of reference, we call 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 is that these models are computationally straightforward to simulate and, arguably, simpler to parameterise than PBEs. By using HODEs, we assume that most observed variation comes from differences in biological processes across cells, not inherent stochasticity in biochemical reactions within cells as is assumed when employing stochastic simulations algorithms.

Inference for HODEs is problematic due, partly, to the experimental hurdles involved with generating data of sufficient standard. 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, the concentrations of fluorescently-labelled proteins are measured [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 hence represent “snapshots” of the underlying population [15]. These snapshots can be described by histograms [12] or density functions [9] fit to measurements of quantities 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. Fluorescent Recovery After Photo-bleaching (FRAP) is one such method, which follows the time-dependent response of cells after an initial bleaching [16]. Methods exists, broadly under the banner of “nonlinear mixed effects models”, which use cell trajectories – individual time series of cellular quantities – to estimate both cellular variation and qualities of measurement noise. See, for example, [16–18]. The demands of obtaining such data are, however, higher and typically involve either tracking individual cells through imaging methods [19], or trapping cells in a spatial position where they can be monitored over time [20]. These techniques impose severe restrictions on experimental practices meaning they cannot be used in many circumstances, including for online monitoring of biotechnological processes or analysis of *in vivo* studies. “Snapshot” data continues to play an important role for determining cell level variability in many applications and in this paper we restrict analysis to only such data.

By fitting HODEs to snapshot data, cellular variability can be estimated, and a number of approaches have been proposed for doing so. In HODEs, parameter values vary across cells according to a to-be-determined probability distribution, and the solution to the inverse problem requires solving the cell-specific ODE system many times for each individual. The count of cells in experiments typically exceeds $\sim 10^4$ [15], so approaches where the computational burden scales with this count are usually infeasible. There are two current approaches for dealing with this burden, and both involve dimensionality reduction. In other words, both approaches require preprocessing raw data before analysis, so result in a degree of information loss. The first involves using population average data – mean values of measurements at different points in time – yet, explicitly model how this mean represents a mixture across different subpopulations [21]. Chan et al. (2016) follow this approach to analyse population substructure in immune cells, which allows them to employ standard Bayesian approaches to fitting. The alternative approach is to fit probability densities to raw snapshot data and use these densities, rather than raw data, for estimation [12, 15, 22, 23]. We follow this approach here as it is likely that more information about the underlying data is retained than in the “population average” one.

We now briefly describe the existing approaches for using HODE models to estimate cell population heterogeneity. Hasenauer et al. (2011) present 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) use mixture models to model subpopulation struc-
ture 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 use
mixture models to represent subpopulation structure and use maximum
likelihood to estimate both within- and between-subpopulation variability,
which permits fitting to multivariate output distributions with complex
correlation structures. Dixit et al. (2018) assign observations into discrete
bins, then choose likelihood distributions according to the maximum entropy
criterion, which they then use to estimate cell variability within a Bayesian
framework.

Our framework is Bayesian although it is distinct from the approach used
to fit many dynamic models, since we assume output variation arises from
parameter heterogeneity across cells, with no contribution from measurement
noise. The approach is, hence, most suitable when measurement error is
minimal. Additionally, our approach is suitable only for underdetermined
models – which we define as the case where there are fewer output quantities
of interest than parameters. Since the generation of snapshots is expensive,
it is often the case that fewer observables are recorded than parameters.
Hence, we believe that this restriction does not present particular issue
to the generalisability of our approach. Our method is a two-step Monte
Carlo approach, which, for reasons described in §3, we call “Contour Monte
Carlo” (CMC). Unlike many existing methods, CMC is straightforward
to implement and does not require extensive computation time. In CMC,
prior probability distributions are used in place of ansatz densities. It also
does not require the number of cell clusters be chosen beforehand, rather,
subpopulations emerge as modes in the posterior parameter distributions.
Like [23], CMC can fit multivariate snapshot data and unlike [12], does not
use discrete bins to model continuous data. As more experimental techniques
elucidating single cell behaviour are developed, interest in models describing
measurement snapshots should follow. We argue that due to its simplicity
and generality, CMC can be used to perform inference on the proliferation
of rich single cell data and, thus, is a useful addition to the modeller’s
toolkit.

Outline of the paper: In §3, we describe our probabilistic model of the
inverse problem and detail the CMC algorithm for generating 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 (Figure 4), and suggest practical
remedies to issues commonly encountered while using this approach. A
glossary of variable names used in this paper is included as Table 1.

Experimental methods such as flow cytometry measure single cell char-
acteristics at a given time. Cells are typically destroyed by the measurement
process, so the data consists of cross-sections or “snapshots” of sampled

individuals from the population, rather than providing time series for each individual cell. The contrast between these two very different scenarios is highlighted in Figure 1. The cells at time t_k are not the same as those at time t_{k-1} and even if they are, there is no way of associating a particular cell at time t_k with the same cell at time t_{k-1} . In other words, there is no sense of a “trajectory” of a specific cell, or of multiple observations assigned to a single cell.

206
207
208
209
210
211
212

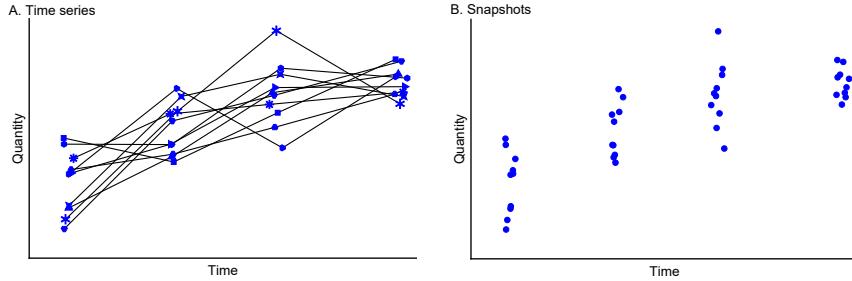


Figure 1: **Data typical of single cell experiments.** **(A) Time series data.** **(B) Snapshot data.** In A, note cell identities are retained at each measurement time (indicated by individual plot markers), whereas in the snapshot data in B, either this information is lost, or more often, cells are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

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

Note that in most circumstances, the initial state of the system, $\mathbf{x}(0)$, is unknown, and it can be convenient to include these as elements of $\boldsymbol{\theta}$ to be estimated.

213
214
215
216

217
218
219

3.1 Snapshot data

We assume the variation in snapshots arises due to heterogeneity in the underlying parameters, $\boldsymbol{\theta}$, across cells. Therefore, the evolution of the underlying state of cell i is described by an idiosyncratic ODE,

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

where superscript $\{i\}$ indicates the i th cell. The collection of such idiosyncratic ODEs across all cells is then referred to as the “HODE model”.

The traditional (non-hierarchical) state-space approach to modelling dynamic systems supposes that measurement error introduces stochastic variation in the output (Figure 2A). Our approach, by contrast, assumes any variation in outputs is solely due to variation in parameter values between

224
225
226
227
228
229

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. Once again we emphasize that we are considering distributions of quantities of interest with no sense of specific individual trajectories, making a mixed effects modelling approach problematic.

230
231
232
233
234
235
236

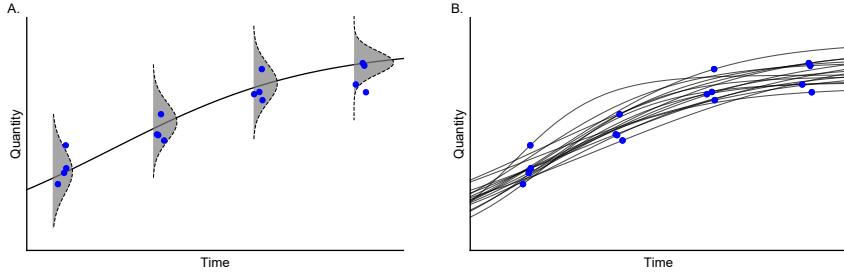


Figure 2: **Models of variation in observed outputs.** (A) **State-space model.** (B) **Parameter heterogeneity model.** (A) For non-hierarchical state-space models , there is a single “true” latent state, and observations result from an imperfect measurement process (grey histograms). (B) For models with parameter heterogeneity, 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.

In an experiment, quantities of interest (QOIs) are measured. Examples of QOIs include concentrations of compounds at different points in time, peak voltages across cell membranes during an action potential, or measurements of cell volume. Here, we suppose $m \geq 1$ QOIs are measured,

$$\mathbf{q}^\top = (q_1, q_2, \dots, q_m) \in \mathbb{R}^m, \quad (3)$$

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,

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

The raw snapshot data \mathbf{Y} is the collection of all measured QOIs,

$$\mathbf{Y} = (\mathbf{y}(t_1), \mathbf{y}(t_2), \dots, \mathbf{y}(t_m)) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \dots \times \mathbb{R}^{n_m}. \quad (5)$$

The goal of inference is to characterise the probability distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ representing heterogeneity in cellular processes. The numbers of cells sampled in typical experimental setups is large, and, following previous work, we represent snapshot data \mathbf{Y} using probability distributions [12, 15, 22, 23]. 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, $\mathbf{q} \in \mathbb{R}^m$. We suppose these kernel density estimates approximate a true distribution over the observed data, $p(\mathbf{q}|\Phi)$ and denote the estimated density as $p(\mathbf{q}|\hat{\Phi})$. After this initial fitting, this distribution – which we term the “target distribution” – becomes the object we seek to replicate in

237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255

our inference problem. We assume there are enough observational data that
the estimated probability distributions are approximate sufficient statistics
of the posterior distribution, meaning $p(\boldsymbol{\theta}|\hat{\Phi}) \approx p(\boldsymbol{\theta}|Y)$.
256
257
258

The aim of our inverse problem, hence, becomes to derive a “posterior”
parameter distribution, which, when fed through the deterministic trans-
formation described by the model, $\mathbf{q}(\boldsymbol{\theta})$, recapitulates the fitted output
density,
259
260
261
262

$$p(\boldsymbol{\theta}|\hat{\Phi}) \xrightarrow{\mathbf{q}(\boldsymbol{\theta})} p(\mathbf{q}|\hat{\Phi}). \quad (6)$$

In measure theoretic terms, the intrinsic measure $p(\mathbf{q}|\hat{\Phi})$ implied by $p(\boldsymbol{\theta}|\hat{\Phi})$
is known as the *push forward* of the measure with respect to the model [24].
263
264

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

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

3.2 Theoretical development of CMC

265

We consider the under-determined case where there are fewer QOIs than
model parameters ($m < p$). This means that, provided a given QOI can be
generated by the model, it can be produced from any member of a subset
of parameter space. Unlike the fully-determined case, these subsets (in
general) have non-zero “volume”, and we term them “iso-output contour
regions”. Symbolically, we represent the iso-output contour region for a
given quantity of interest $\tilde{\mathbf{q}}$ (say) by $\Omega(\tilde{\mathbf{q}}) = \{\boldsymbol{\theta} : \mathbf{q}(\boldsymbol{\theta}) = \tilde{\mathbf{q}}\}$.
266
267
268
269
270
271
272
273

In general, contour “volumes” $\mathcal{V}(\tilde{\mathbf{q}})$ depend on the chosen output value
 $\tilde{\mathbf{q}}$ (Figure 3). Further, the interpretation of these “volumes” depends upon
their dimensions. Considering cases with a single QOI: for a model with
two parameters, iso-output contour regions are one-dimensional lines, whose
size is a length; for a model with three parameters, the contour regions are
surfaces, whose size is an area; for four-dimensional parameter spaces, the
contour regions are three-dimensional and their size is a volume; and for
274
275
276
277
278
279

models with $p > 4$ parameters, the contour regions are $p - 1$ dimensional manifolds, whose size is a hypervolume. 280
281

MCMC methods aim to approximate a posterior parameter distribution by sampling from it. In this case, the resultant parameter samples, when pushed through the model, should approximate samples from the desired QOI distribution. “Vanilla” MCMC methods, like Random Walk Metropolis [25], work fine in more traditional Bayesian analyses but are biased for our inference problem. Such vanilla MCMC samplers choose where next to step based on the ratio of probability densities at the proposed parameter value and current position. Using a vanilla sampler for our case, unfortunately, does not work because the Markov chains are biased towards those regions of parameter space with the largest iso-output contour volumes. This bias means that the stationary parameter distribution obtained, when fed through the model, does not recapitulate the observed output distribution [26]. We stress again the difference between this problem and a traditional Bayesian analysis: here, uncertainty is due to the forward map being many-to-fewer meaning that the inverse map is indeterminant; in Bayesian inference, it comes from stochastic processes in the system itself. This difference means traditional inference methods cannot be used and motivates the method we introduce here. 282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308

Sampling algorithms, therefore, need to explicitly account for the differential volume of iso-output contours. In applied problems, however, we do not know the volumes of iso-output contours and they cannot be exactly calculated for all but the simplest models. Instead in CMC, we estimate them. The following analysis provides a brief introduction to a probabilistic formulation of under-determined inverse problems (see our companion paper [26] for a more comprehensive discussion). In doing so, this suggests a sampling based approach for estimating contour volumes, which are then exploited by our CMC algorithm. 300
301
302
303
304
305
306
307

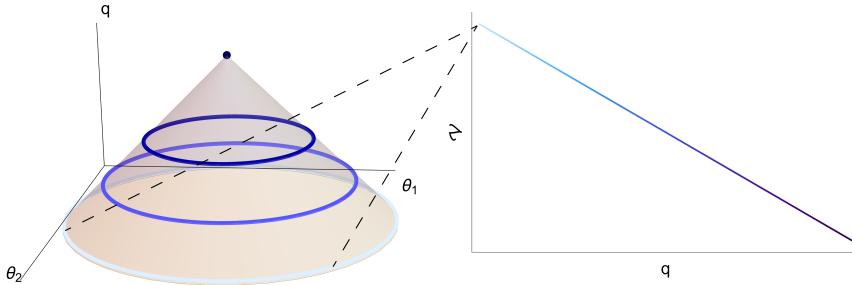


Figure 3: Left: An example output function $q(\theta_1, \theta_2)$ along with iso-output contours indicated (coloured lines). **Right:** The “volume” of output contours as a function of output value. Note that here, since parameter space is two dimensional, the “volume” of each output value corresponds 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, $p(q|\hat{\Phi})$. To derive the posterior parameter distribution, we consider the joint density of parameters and QOIs, $p(\boldsymbol{\theta}, q|\hat{\Phi})$. This can be decomposed in two ways using the law of 309
310
311
312
313

conditional probability,

314

$$p(\boldsymbol{\theta}, \mathbf{q}|\hat{\Phi}) = p(\boldsymbol{\theta}|\mathbf{q}, \hat{\Phi}) \times p(\mathbf{q}|\hat{\Phi}) = p(\mathbf{q}|\boldsymbol{\theta}, \hat{\Phi}) \times p(\boldsymbol{\theta}|\hat{\Phi}). \quad (7)$$

Rearranging eq. (7), we obtain the posterior parameter distribution,

315

$$p(\boldsymbol{\theta}|\hat{\Phi}) = \frac{p(\boldsymbol{\theta}|\mathbf{q}, \hat{\Phi}) \times p(\mathbf{q}|\hat{\Phi})}{p(\mathbf{q}|\boldsymbol{\theta}, \hat{\Phi})}. \quad (8)$$

For a deterministic map, eq. (8) is only well defined when $\mathbf{q} = \mathbf{q}(\boldsymbol{\theta})$. (Since the mapping from parameters to outputs is deterministic, $p(\mathbf{q}|\boldsymbol{\theta}, \hat{\Phi}) = \delta(\mathbf{q}(\boldsymbol{\theta}))$, i.e., the Dirac delta function centred at $\mathbf{q} = \mathbf{q}(\boldsymbol{\theta})$.) Thus eq. (8) becomes,

316

317

318

319

$$p(\boldsymbol{\theta}|\hat{\Phi}) = p(\boldsymbol{\theta}|\mathbf{q}(\boldsymbol{\theta}), \hat{\Phi}) \times p(\mathbf{q}(\boldsymbol{\theta})|\hat{\Phi}). \quad (9)$$

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 $p(\mathbf{q}|\hat{\Phi})$ can be caused by any member of a set of parameter distributions. To ensure uniqueness of the “posterior” parameter distributions, we must, therefore, specify “prior” distributions for the parameters, as in more traditional Bayesian inference. In what follows, we assume the conditional distribution $p(\boldsymbol{\theta}|\mathbf{q}, \hat{\Phi})$ is independent of the data, i.e., $p(\boldsymbol{\theta}|\mathbf{q}, \hat{\Phi}) = p(\boldsymbol{\theta}|\mathbf{q})$, and thus represents a conditional “prior” which can be manipulated using Bayes’ rule as,

320

321

322

323

324

325

326

327

$$p(\boldsymbol{\theta}|\mathbf{q}(\boldsymbol{\theta})) = \frac{p(\boldsymbol{\theta})}{p(\mathbf{q}(\boldsymbol{\theta}))}. \quad (10)$$

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

328

329

$$p(\boldsymbol{\theta}|\hat{\Phi}) = \frac{p(\boldsymbol{\theta})}{p(\mathbf{q}(\boldsymbol{\theta}))} p(\mathbf{q}(\boldsymbol{\theta})|\hat{\Phi}). \quad (11)$$

Again, we defer to our companion piece [26] for detailed explanation of eqs. (10) and (11) 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. (10) is the prior induced on output space by the prior over parameter space. For a uniform prior on parameter values, this is,

330

331

332

333

334

335

$$p(\boldsymbol{\theta}|\mathbf{q}(\boldsymbol{\theta})) = \frac{1}{\mathcal{V}(\mathbf{q}(\boldsymbol{\theta}))}, \quad (12)$$

where $\mathcal{V}(\mathbf{q}(\boldsymbol{\theta}))$ is the volume of parameter space occupied by the iso-output contour $\Omega(\mathbf{q}(\boldsymbol{\theta}))$ (see Fig. 3 for the meaning of this volume for a two parameter example). Therefore, a uniform prior over parameter space implies a prior structure where all parameter values producing the same output are given equal weighting.

336

337

338

339

340

3.3 Implementation of CMC

341

Except for some toy examples, the denominator of eq. (10) cannot be calculated, so exact sampling from the posterior parameter distribution of eq. (11) is not, in general, possible. We propose, instead, a computationally efficient sampling method to estimate $p(\mathbf{q}(\boldsymbol{\theta}))$, which forms the first step of our so-called “Contour Monte Carlo” (CMC) algorithm (Algorithm 1; Figure

342

343

344

345

346

4(ii)), where the volume of iso-output contours with each feasible output value is estimated. This step involves repeated independent sampling from the prior distribution of parameters, $\boldsymbol{\theta}^{[i]} \sim p(\boldsymbol{\theta}|\Xi)$, where Ξ parameterises the prior probability density. Each parameter sample is then mapped to an output value, $\mathbf{q}^{[i]} = \mathbf{q}(\boldsymbol{\theta}^{[i]})$. The collection of output samples is then fitted using a vine copula kernel density estimator (KDE) [27], $\hat{\Psi} = \arg \max_{\Psi} p((\mathbf{q}^{[1]}, \dots, \mathbf{q}^{[N_1]}) | \Psi)$. Throughout the course of development of CMC, we tested many KDE methods and found vine copula KDE best suited to approximating the higher dimensional probability distributions required in practice – other methods produced coarse estimates of the joint density and took substantially more computational resource. Indeed, the ability to do KDE in high dimensions was the motivation behind the creation of vine copula KDE in the first place [27].

The second step in our algorithm then uses MCMC to sample from an approximate version of eq. (11), where the estimated density, $p(\mathbf{q}(\boldsymbol{\theta})|\hat{\Psi})$ replaces its corresponding estimand (Algorithm 1; Figure 4(iii)),

$$p(\boldsymbol{\theta}|\hat{\Phi}, \Xi, \hat{\Psi}) = \frac{p(\boldsymbol{\theta}|\Xi)}{p(\mathbf{q}(\boldsymbol{\theta})|\hat{\Psi})} p(\mathbf{q}(\boldsymbol{\theta})|\hat{\Phi}). \quad (13)$$

The final step in CMC is to compare output samples generated by MCMC with the target distribution (Figure 4(iv)). As the sample size of both sampling steps (i.e. the contour volume estimation and MCMC steps) tends to infinity, CMC produces a sample of parameter values $(\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2]}, \dots)$ which, when mapped to the output space, corresponds to the target distribution $p(\mathbf{q}|\hat{\Psi})$. 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 good 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 in the contour volume estimation step to ensure reasonable approximation of the distribution of output values obtained by sampling the prior density.

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, Contour Monte Carlo is a pragmatic approach to approximate it by sampling if one should exist. A useful way to diagnose whether the target distribution can be produced from the model and chosen priors is to plot the output values from the contour volume estimation step of CMC: this is akin to visualising the prior predictive distribution in traditional Bayesian inference [25]. If the bulk of target probability mass does not overlap with the simulated output values, then the model and/or chosen prior are unlikely to be invertible to this particular target. In §4.2.2 and §4.4, we provide examples that illustrate this aspect of model checking.

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. 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 to improve sampling efficiency [28].

394
395

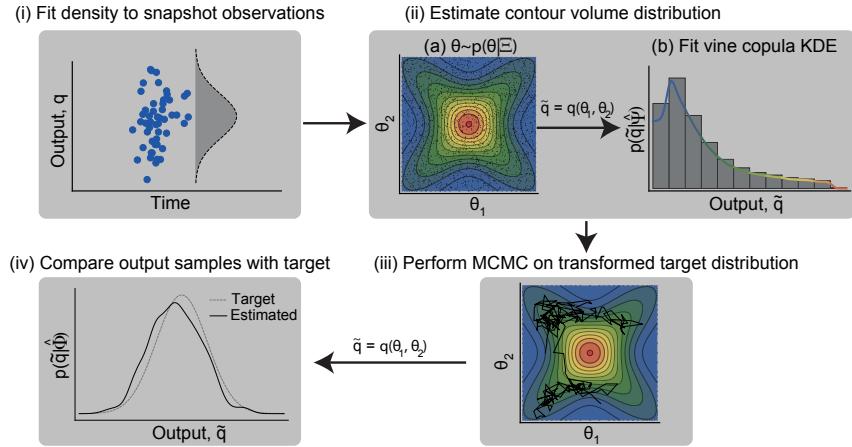


Figure 4: Workflow for Contour Monte Carlo to estimate cell population heterogeneity. The distribution targeted in (iii) is given by eq. (13). Here, \tilde{q} is used to represent an output value resultant from applying the functional q to parameter samples (θ_1, θ_2) .

Algorithm 1 Pseudocode for the Contour Monte Carlo algorithm for sampling from the posterior parameter distribution of eq. (13).

```

procedure CMC( $\mathbf{Y}, \Xi, N_1, N_2$ )       $\triangleright$  Sample from posterior parameter distribution
     $\hat{\Phi} = \text{SNAPSHOTESTIMATOR}(\mathbf{Y})$ 
     $\hat{\Psi} = \text{CONTOURVOLUMEESTIMATOR}(\Xi, N_1)$ 
     $(\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N_2]}) = \text{MCMC}(\hat{\Phi}, \Xi, \hat{\Psi}, N_2)$ 
    converged = COMPAREOUTPUTTOTARGET(( $\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N_2]}$ ),  $\hat{\Phi}$ )
    while converged ≠ 1 do  $\triangleright$  Rerun contour volume estimation (if necessary modify
    vine copula KDE hyperparameters) and/or MCMC, with larger sample sizes if required
         $\hat{\Psi} = \text{CONTOURVOLUMEESTIMATOR}(\Xi, N'_1), N'_1 \geq N_1$ 
         $(\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N'_2]}) = \text{MCMC}(\hat{\Phi}, \Xi, \hat{\Psi}, N'_2), N'_2 \geq N_2$ 
        converged = COMPAREOUTPUTTOTARGET(( $\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N'_2]}$ ),  $\hat{\Phi}$ )
         $N_1 \leftarrow N'_1, N_2 \leftarrow N'_2$ 
    end while
    return  $(\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N_2]})$ 
end procedure

procedure SNAPSHOTESTIMATOR( $\mathbf{Y}$ )  $\triangleright$  Fit snapshots with kernel density estimator
    (KDE)
     $\hat{\Phi} = \arg \max_{\Phi} p(\mathbf{Y} | \Phi)$ 
    return  $\hat{\Phi}$ 
end procedure

procedure CONTOURVOLUMEESTIMATOR( $\Xi, N_1$ )  $\triangleright$  Estimate volume of contours
    for  $i$  in  $1 : N_1$  do
         $\boldsymbol{\theta}^{[i]} \sim p(\boldsymbol{\theta} | \Xi)$   $\triangleright$  Sample from prior density
         $\mathbf{q}^{[i]} = \mathbf{q}(\boldsymbol{\theta}^{[i]})$   $\triangleright$  Calculate corresponding output value
    end for
     $\hat{\Psi} = \arg \max_{\Psi} p((\mathbf{q}^{[1]}, \dots, \mathbf{q}^{[N_1]}) | \Psi)$   $\triangleright$  Fit vine copula KDE
    return  $\hat{\Psi}$ 
end procedure

procedure MCMC( $\hat{\Phi}, \Xi, \hat{\Psi}, N_2$ )  $\triangleright$  Random Walk Metropolis algorithm targeting
    posterior parameter distribution
     $\boldsymbol{\theta}^{[0]} \sim \pi(\cdot)$   $\triangleright$  Sample from arbitrary initialisation distribution
    for  $i$  in  $1 : N_2$  do
         $\boldsymbol{\theta}^{[i]'} \sim \mathcal{N}(\boldsymbol{\theta}^{[i-1]}, \Sigma)$   $\triangleright$  Propose new parameter values
         $r = p(\boldsymbol{\theta}^{[i']} | \Xi) p(\mathbf{q}(\boldsymbol{\theta}^{[i-1]} | \hat{\Psi})) p(\mathbf{q}(\boldsymbol{\theta}^{[i]'} | \hat{\Phi})) / [p(\boldsymbol{\theta}^{[i-1]} | \Xi) p(\mathbf{q}(\boldsymbol{\theta}^{[i]'} | \hat{\Psi})) p(\mathbf{q}(\boldsymbol{\theta}^{[i-1]} | \hat{\Phi}))]$ 
         $u \sim U(0, 1)$   $\triangleright$  Sample from uniform distribution
        if  $r > u$  then
             $\boldsymbol{\theta}^{[i]} = \boldsymbol{\theta}^{[i']}$   $\triangleright$  Accept proposal
        else
             $\boldsymbol{\theta}^{[i]} = \boldsymbol{\theta}^{[i-1]}$   $\triangleright$  Reject proposal
        end if
    end for
    return  $(\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N_2]})$ 
end procedure

procedure COMPAREOUTPUTTOTARGET(( $\boldsymbol{\theta}^{[1]}, \dots, \boldsymbol{\theta}^{[N_2]}$ ),  $\hat{\Phi}$ )  $\triangleright$  Check output
    distribution close to target
    for  $i$  in  $1 : N_2$  do
         $\tilde{\mathbf{q}}^{[i]} = \mathbf{q}(\boldsymbol{\theta}^{[i]})$   $\triangleright$  Compute QOIs for each parameter sample
    end for
    if  $p(\tilde{\mathbf{q}}) \approx p(\tilde{\mathbf{q}} | \hat{\Phi})$ ? then
        return 1  $\triangleright$  If sufficiently close then converged
    else
        return 0
    end if
end procedure

```

To generate our results in §4, we assumed for the contour volume estimation step sample sizes were sufficient if the output samples from 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 [28]) to sample from the target distribution, as it typically provides a considerable speed-up over Random Walk Metropolis [25, 29]. We also used the Gelman-Rubin convergence statistic, \hat{R} , to diagnose convergence [25, 30], with a convergence threshold of $\hat{R} \leq \sim 1.1$.

To solve the forward model of each differential equation, we used Julia’s [31] “solve” method for ODE models from the “DifferentialEquations.jl” library [32], which automatically chooses an efficient inbuilt solver. To replicate the results in this section, we recommend readers execute the corresponding Julia scripts (one for each result section) at <https://github.com/ben18785/inverse-sensitivity/tree/master/examples>. Note that, these scripts use the “RCall” library for Julia [33], which calls R from Julia. This package was necessary to use the “kdevine” R package for vine copula kernel density estimation [34].

4 Results

In this section, we use CMC to estimate posterior parameter distributions for four biological systems. In two of the examples, we assume that the first step of CMC (“SnapshotEstimator” within Algorithm 1) has already been completed, 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,

$$\frac{dR}{dt} = R_T k_{deg} + k_1 L R(t) + k_{-1} P(t) - k_{deg} R(t) \quad (14)$$

$$\frac{dP}{dt} = k_1 L R(t) - k_{-1} P(t) - k_{deg}^* P(t), \quad (15)$$

with initial conditions,

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

and $\boldsymbol{\theta} = (R_T, k_1, k_{-1}, k_{deg}, k_{deg}^*)$ 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. (15) as $P(t; \boldsymbol{\theta}, L)$. Here we generate a target model by forward simulations of eq. (14); in each case recording $(P(10; \boldsymbol{\theta}, 2), P(10; \boldsymbol{\theta}, 10))$. In each forward simulation, we fix $(k_{-1}, k_{deg}, k_{deg}^*) = (8, 0.015, 0.25)$ and independently sample values $R_T \sim \mathcal{N}(6.5 \times 10^5, 0.6 \times 10^4)$ and $k_1 \sim \mathcal{N}(1.7, 0.05)$.

This generates an output distribution approximately given by,

$$\mathbf{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} P(10; \boldsymbol{\theta}, 2) \\ P(10; \boldsymbol{\theta}, 10) \end{pmatrix} \sim \mathcal{N} \left[\begin{pmatrix} 2 \times 10^4 \\ 3 \times 10^4 \end{pmatrix}, \begin{pmatrix} 1 \times 10^5 & 0 \\ 0 & 1 \times 10^5 \end{pmatrix} \right]. \quad (16)$$

We note that, whilst the parameters $(k_{-1}, k_{deg}, k_{deg}^*)$ are fixed during this step (to generate output distributions), they are allowed to vary in §4.1.1 and §4.1.2 (where we use CMC to perform inference).

4.1.1 Uniform prior

For an under-determined model, the number of QOIs, m , is less than the number of parameters, p , and there typically exists a non-singular set of parameter distributions mapping to the same target output distribution. To uniquely identify a posterior parameter distribution, it is, therefore, necessary to specify a prior parameter distribution. By incorporating priors, this allows pre-existing biological knowledge to be included, leading to reduced uncertainty in parameter estimates. CMC allows any prior with correct support to be used. Changes to priors affect both the “ContourVolumeEstimation” and “MCMC” steps of CMC (Algorithm 1), so that the (changed) posterior parameter distribution still maps to the target.

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.

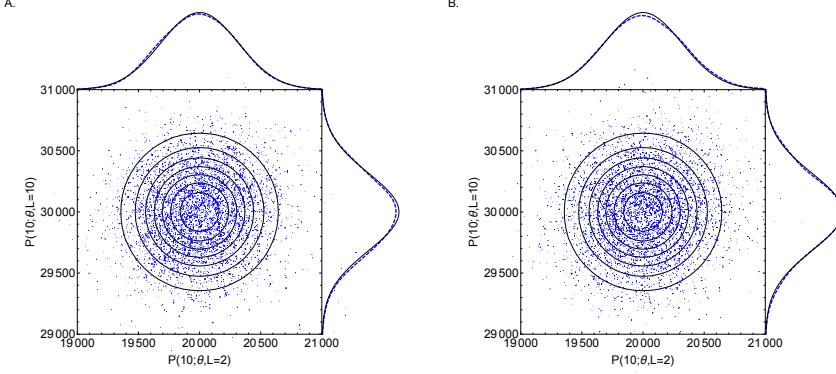


Figure 5: Growth factor model. Target joint output distribution (solid contour lines) and target marginal distributions (solid lines; above and to the right of each figure) versus outputs sampled by CMC (blue points) and reconstructed marginals (dashed lines). (A) uniform priors. (B) Gaussian priors. 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” [25]. For the reconstructed marginal densities in the plots, we use Mathematica’s “SmoothKernelDistribution” function specifying bandwidths of 100 with Gaussian kernels [35].

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 , the rate of introduction of ligand-free cell surface receptors. This plot shows more concentrated posterior mass than in Figure 6A.

Why do our measurements allow us to better resolve (k_{deg}, R_T) compared to (k_1, k_{-1}) ? To answer this, it is useful to calculate the sensitivity of $P(t; \theta, L)$ 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 [36], and these are shown in Figure 7. When the exogenous ligand is set at $L = 2$, these indicate the active ligand-bound receptor concentration is most elastic to changes in R_T and k_{deg} . This higher elasticity means that their range is more restricted by the output measurement than for k_1 and k_{-1} , which have much smaller elasticities 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

455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480

the magnitude of the elasticities for each parameter ($\rho = 0.95$, $t = -5.22$,
 $df = 3$, $p = 0.01$ for Pearson's product-moment correlation), indicating the
utility of sensitivity analyses for optimal experimental design, see e.g., [37].
We suggest, however, that CMC can also be used for this purpose. If an
experimenter generates synthetic data for various choices of QOIs, they can
use CMC to derive the posterior parameter distributions in each case. They
then, simply, select the particular QOI producing the narrowest posterior
for key parameters.

In both panels of Figure 6, we also plot the “*actual*” parameter values
as dashed lines: for k_{-1} and k_{deg} , these indicate the true (fixed) parameter
values, and, for k_1 and R_T , they show the mean of each Gaussian sampling
distribution (\pm two standard deviations shown by shaded rectangles). For
most parameters, these indicate that the area of highest posterior density is
close to the causative parameter values. This is reaffirmed in the top panel
of Table 2, where, in all cases, the actual parameter values lie within the
estimated 95% quantiles for each parameter – indicating that the parameters
were reasonably well identified.

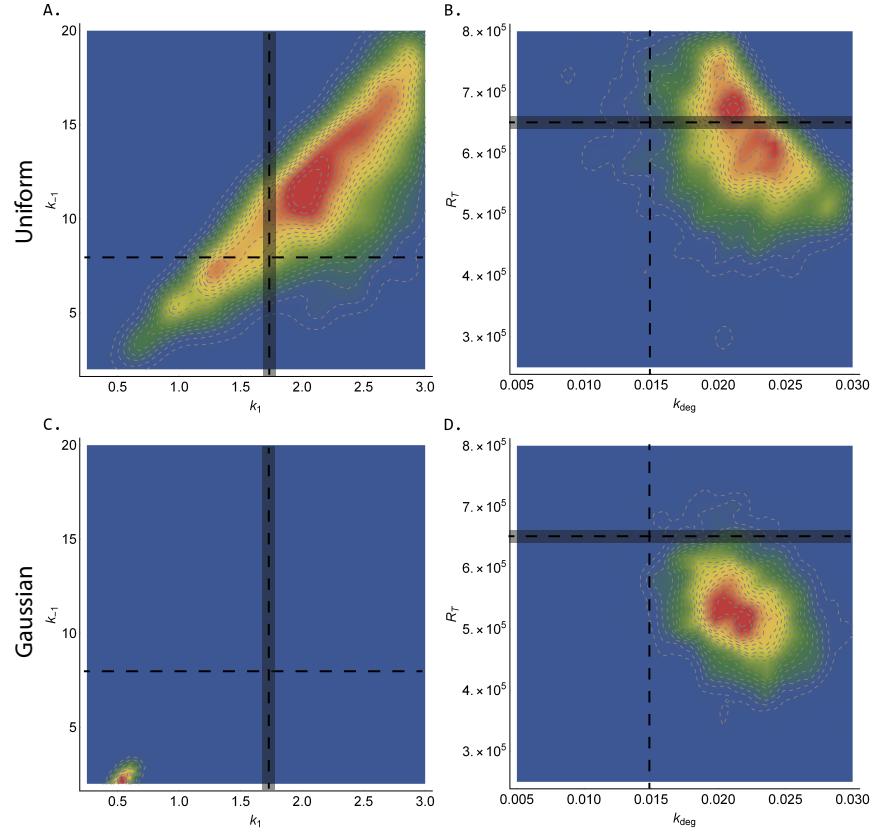


Figure 6: **Growth factor model.** Joint posterior distributions estimated by CMC. Top row (A-B): (k_1, k_{-1}) and (k_{deg}, R_T) using uniform priors. Bottom row (C-D): (k_1, k_{-1}) and (k_{deg}, R_T) using Gaussian priors. In all panels, dashed lines indicate the parameter set or distribution used to generate the target distribution given by eq. (16): for k_{-1} and k_{deg} , the dashed lines show true parameter values and for k_1 and R_T , they show the mean of each Gaussian sampling distributions (\pm two standard deviations shown by shaded rectangles). See Figure 5 caption for CMC details and Table 3 for the priors used. Red (blue) indicates areas of relatively high (low) probability density.

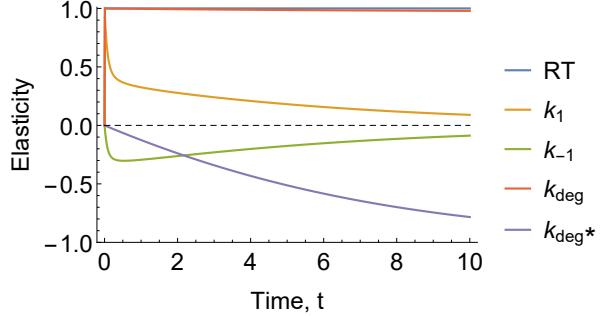


Figure 7: Growth factor model. Elasticities of the active ligand-bound receptors P with respect to each parameter as a function of time. When calculating the elasticities of each parameter, the other parameters were set to their posterior medians given in Table 2 and $L = 2$.

4.1.2 Gaussian prior

We now use CMC to estimate the posterior parameter distribution, when using Gaussian priors (prior hyperparameters shown in Table 3), which are more concentrated than the uniform priors used in §4.1.1. As desired, the target output distribution appears virtually unaffected by the change of priors (Figure 5B) although with substantial changes to the posterior parameter distribution (Figure 6C and 6D). In particular, the marginal posterior distributions obtained from the Gaussian prior are narrower compared to the uniform case (rightmost column of Table 2).

As in traditional Bayesian inference, prior choice has a greater influence on the posterior distribution when data provide less information on the underlying process. This is readily apparent in comparing the dramatic change from Figure 6A to 6C for (k_1, k_{-1}) , which have low sensitivities, with the more nuanced change from Figure 6B to 6D for (k_{deg}, R_T) , which have high sensitivities. The results also indicate the bias-variance trade-off inherent in Bayesian analysis: when relatively uninformative priors are specified (Figure 6A&B), the posterior distributions are wider but their centre lies, in general, closer to the true values (dashed lines) than when more information is included in the priors (Figure 6C&D).

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, [38]) - illustrating how CMC can determine resolve population substructure from a multimodal output distribution. The Michaelis-Menten model of enzyme kinetics describes the dynamics of concentrations of an enzyme, E , a substrate, S , an enzyme-substrate complex, C , and a product, P ,

$$\begin{aligned} \frac{dE}{dt} &= -k_f E(t)S(t) + k_r C(t) + k_{cat}C(t), \\ \frac{dS}{dt} &= -k_f E(t)S(t) + k_r C(t), \\ \frac{dC}{dt} &= k_f E(t)S(t) - k_r C(t) - k_{cat}C(t), \\ \frac{dP}{dt} &= k_{cat}C(t), \end{aligned} \tag{17}$$

Parameter							Posterior 25%-75% conc.
	2.5%	25%	Quantiles			True values	
	Uniform prior						
R_T	441,006	548,275	606,439	677,055	772,484	650,000	23%
k_1	0.90	1.69	2.17	2.56	2.95	1.70	32%
k_{-1}	4.35	8.35	11.23	14.23	18.71	8.00	33%
k_{deg}	0.013	0.019	0.021	0.024	0.029	0.015	20%
k_{deg}^*	0.20	0.34	0.40	0.44	0.49	0.25	27%
	Gaussian prior						
R_T	408,396	487,372	529,558	577,970	678,632	650,000	16%
k_1	0.39	0.49	0.54	0.60	0.70	1.70	4%
k_{-1}	1.39	1.92	2.26	2.63	3.35	8.00	4%
k_{deg}	0.016	0.020	0.022	0.024	0.027	0.015	16%
k_{deg}^*	0.22	0.29	0.33	0.38	0.46	0.25	21%

Table 2: **Growth factor model. Estimated quantiles from CMC samples 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.

with initial conditions,

524

$$E(0) = E_0, S(0) = S_0, C(0) = C_0, P(0) = P_0, \quad (18)$$

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

525
526
527

4.2.1 Bimodal output distribution

528
529
530
531
532
533
534
535
536

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, 23]. Before formal inference occurs, an output distribution with multiple modes 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,

$$\begin{aligned} \mathbf{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} &= \begin{pmatrix} E(2.0; \boldsymbol{\theta}) \\ S(1.0; \boldsymbol{\theta}) \end{pmatrix} \sim p(\mathbf{q}; \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\mu}_2, \Sigma_2) \\ &= \frac{1}{2} (\mathcal{N}(\mathbf{q}; \boldsymbol{\mu}_1, \Sigma_1) + \mathcal{N}(\mathbf{q}; \boldsymbol{\mu}_2, \Sigma_2)), \end{aligned} \quad (19)$$

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

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

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 provides a close approximation to the output target distribution (Figure 8A). Without providing *a priori* information on the subpopulations of cells, two distinct

537
538
539
540

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 using MCMC to sample multimodal distributions similarly apply here. So, whilst adaptive MCMC [28] sufficed to explore this posterior surface, it may be necessary to use MCMC methods more robust to such geometries in other cases (for example, population MCMC [39]).

541
542
543
544
545
546
547
548

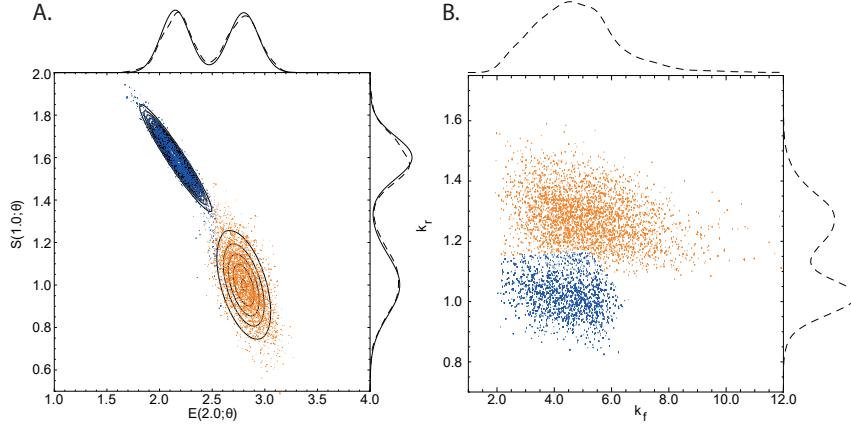


Figure 8: Michaelis-Menten model. (A) Bimodal target distribution q (solid contour lines) versus output samples (points). (B) posterior parameter samples (points). The solid and dashed lines above and to the side of panel A indicate the target and estimated marginal output distributions, respectively. In B, only estimated parameter marginals are shown as the exact solutions are unknown. 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 [35] 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 [35] was used to identify clusters in B.

4.2.2 Four-dimensional output distribution

549
550
551
552
553
554

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 $t = 2$,

$$\begin{aligned} \mathbf{q} &= \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \begin{pmatrix} E(1.0; \theta) \\ S(1.0; \theta) \\ E(2.0; \theta) \\ S(2.0; \theta) \end{pmatrix} \\ &\sim \mathcal{N} \left[\begin{pmatrix} 0.5 \\ 2.8 \\ 0.9 \\ 1.4 \end{pmatrix}, \begin{pmatrix} 0.02 & -0.05 & 0.04 & -0.05 \\ -0.05 & 0.30 & -0.15 & 0.20 \\ 0.04 & -0.15 & 0.12 & -0.17 \\ -0.05 & 0.20 & -0.17 & 0.30 \end{pmatrix} \right]. \end{aligned} \quad (20)$$

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. (20), we plotted the output measurements used to estimate contour volumes (obtained from 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.

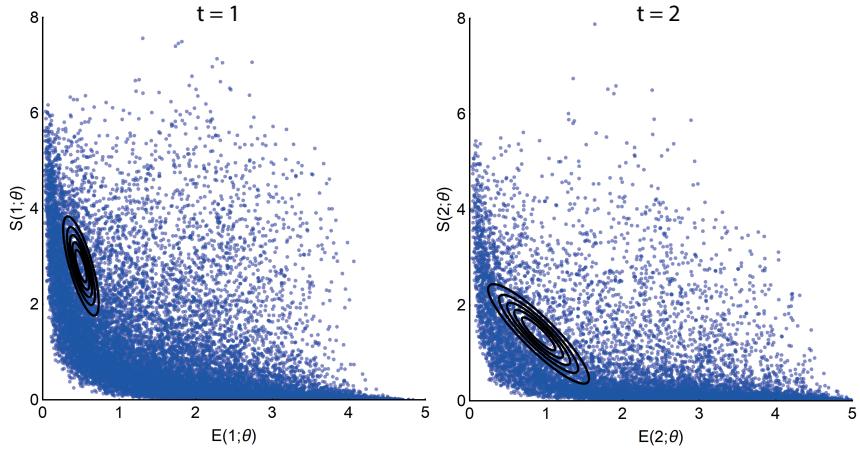


Figure 9: Michaelis-Menten model. QOIs (blue points) obtained by independently sampling the priors versus the target distribution (black solid contours). Left: (q_1, q_2) . Right: (q_3, q_4) . We show 20,000 output samples, where each set of four measurements was obtained from a single sample of all parameters. The output target distribution shown by the contours corresponds to the marginal densities of each pair of enzyme-substrate measurements given by eq. (20).

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. (20). The distribution of paired enzyme-substrate samples illustrates that the CMC output distribution closely approximates the target density, itself representing dynamic evolution of the covariance between enzyme and substrate measurements. Target marginal distributions (solid lines) along with their approximations from kernel density estimation (dashed lines) are also shown above and to the right of the main panel of Figure 10 and largely indicate correspondence.

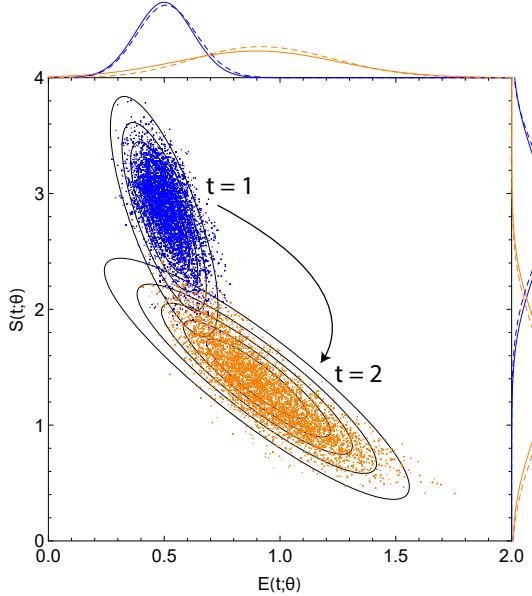


Figure 10: Michaelis-Menten model. Posterior output samples from CMC (coloured points) versus contour plots (black solid lines) of the joint marginal distributions of eq. (20). Enzyme and substrate measurements are given by the horizontal and vertical axes, respectively. Output functionals for (q_1, q_2) and (q_3, q_4) are given by blue and orange points, respectively. The solid and dashed coloured lines outside the panels indicate exact target marginals of eq. (20) and those estimated by CMC, respectively. In the “ContourVolumeEstimator” step, 200,000 independent samples were used, and in the MCMC step, 10,000 samples across each of 4 Markov chains were used, with the first half of the chains discarded as “warm-up” [25]. Mathematica’s “SmoothKernelDistribution” function, using Gaussian kernels [35] and bandwidths ranging from 0.1 to 0.4, was used to reconstruct marginal densities.

4.3 TNF signalling pathway

We now illustrate how CMC can be applied to another ODE system: the tumour necrosis factor (TNF) signalling pathway model introduced in [40] 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

$$\begin{aligned} \frac{dx_1}{dt} &= -x_1(t) + \frac{1}{2} [\beta_4(x_3(t))\alpha_1(u(t)) + \alpha_3(x_2(t))] \\ \frac{dx_2}{dt} &= -x_2(t) + \alpha_2(x_1(t))\beta_3(x_3(t)) \\ \frac{dx_3}{dt} &= -x_3(t) + \beta_2(x_2(t))\beta_5(x_4(t)) \\ \frac{dx_4}{dt} &= -x_4(t) + \frac{1}{2} [\beta_1(u(t)) + \alpha_4(x_3(t))] , \end{aligned} \tag{21}$$

578
579
580
581
582
583
584
585

with initial conditions,

586

$$x_1(0) = 0.0, \quad x_2(0) = 0.0, \quad x_3(0) = 0.29, \quad x_4(0) = 0.625. \quad (22)$$

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

587
588

$$\begin{aligned} \alpha_i(z) &= \frac{z^2}{a_i^2 + z^2}, \quad i = 1, \dots, 4, \\ \beta_j(z) &= \frac{b_j^2}{b_j^2 + z^2}, \quad j = 1, \dots, 5, \end{aligned} \quad (23)$$

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,

589
590
591

$$u(t) = \begin{cases} 1, & \text{if } t \in [0, 2], \\ 0, & \text{otherwise.} \end{cases} \quad (24)$$

4.3.1 Recovering parameter values in under-determined systems

592

In under-determined models, a set of parameters of non-zero volume can produce the same 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 (21) and performing inference on all nine system parameters, whilst collecting only two output measurements. Specifically, we randomly sample $a_1 \sim \mathcal{N}(0.6, 0.05)$ for each simulation of the forward model, 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 $q_2 = x_2(1.0)$ in each case. In doing so, we obtain an output distribution well-approximated by the bivariate Gaussian distribution,

593
594
595

$$\begin{aligned} \mathbf{q} &= \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} x_1(2.0) \\ x_2(1.0) \end{pmatrix} \\ &\sim \mathcal{N} \left[\begin{pmatrix} 0.26 \\ 0.07 \end{pmatrix}, \begin{pmatrix} 2.1 \times 10^{-4} & 5.9 \times 10^{-5} \\ 5.9 \times 10^{-5} & 1.8 \times 10^{-5} \end{pmatrix} \right]. \end{aligned} \quad (25)$$

We now apply CMC to the target output distribution given by eq. (25) to estimate a posterior distribution over all nine parameters of eq. (21). Apart from 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 how the recovered posterior distribution and data generating distribution differ. In Figure 11A, we plot the actual parameter values (horizontal axis) used to generate the data versus the estimated values (vertical axis). This illustrates that, due to the chosen priors, there is a disjunction between actual and estimated parameter values in all cases apart from a_1 . Though because the model is under-determined, the corresponding output distribution closely approximates the target despite these differences (Figure 11B).

596
597
598
599
600
601
602
603
604
605
606
607

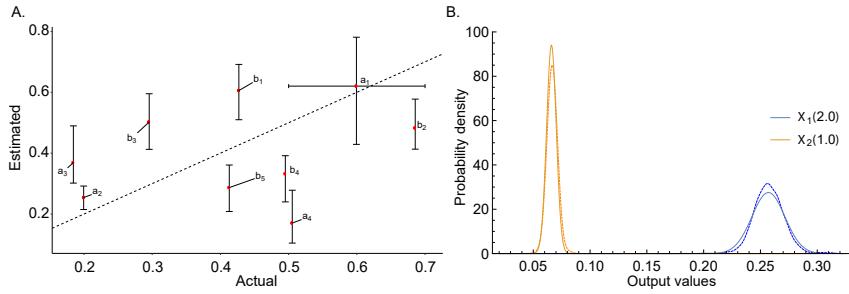


Figure 11: TNF signalling pathway model. (A) Actual parameter values versus estimated quantiles for the output distribution of eq. (25). (B) Marginal output targets (solid lines) and sampled output distributions (dashed lines). In A, in the vertical direction, red points indicate 50% posterior quantiles 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, with the first half of the chains discarded as ‘‘warm-up’’ [25]. Mathematica’s ‘‘SmoothKernelDistribution’’ function, using a Gaussian kernel [35] and a bandwidth of 0.003 was used to reconstruct marginal densities.

4.3.2 Bimodal output distribution

The dynamics of all cells can often be modelled by assuming cells exist in subpopulation clusters, which evolve differently over time. A hint that such subpopulation structure may exist is output distributions with multiple modes. 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,

$$\mathbf{q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}, \quad (26)$$

where,

$$\begin{aligned} q_1 &= \mathbf{x}_2(1.0) \sim \mathcal{N}(0.06, 0.01) \\ q_2 &= \mathbf{x}_2(2.0) \sim \frac{1}{2} (\mathcal{N}(0.1, 0.01) + \mathcal{N}(0.14, 0.01)) \\ q_3 &= \mathbf{x}_2(4.0) \sim \frac{1}{2} (\mathcal{N}(0.1, 0.01) + \mathcal{N}(0.20, 0.01)), \end{aligned} \quad (27)$$

where the target distributions for $\mathbf{q}_2(2.0)$ and $\mathbf{q}_2(4.0)$ indicate mixtures of univariate Gaussians, and the priors used are given 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 a bimodal output distribution causes CMC to sample clusters of parameter values, without the need for subpopulation information to be provided ahead of estimation.

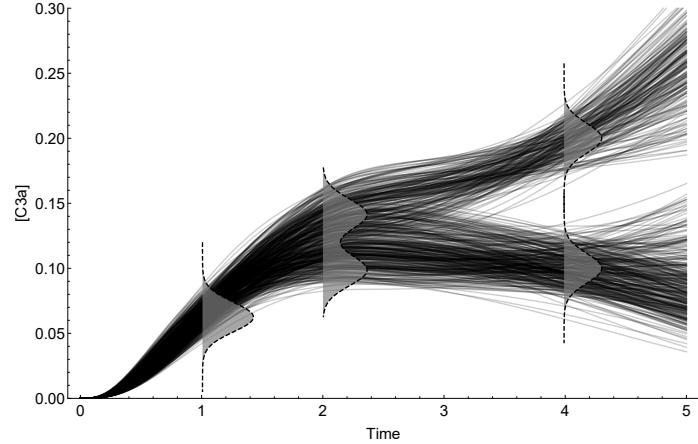


Figure 12: **TNF signalling pathway model.** Target output distribution (dashed plots with grey filling) and unique trajectories (black solid lines) obtained from the posterior parameter distribution. In CMC, 10,000 independent samples were used in the “ContourVolumeEstimator” step, and 5,000 MCMC samples across each of 4 Markov chains were used in the second, with the first half of the chains discarded as “warm-up” [25].

Model	Target density	Parameter	Prior density	Prior θ_1	Prior θ_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	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
hESC differentiation	1D-2D Gaussian	p_1	uniform	40.0	60.0
		p_2	uniform	2.0	10.0
		p_3	uniform	0.5	16
		p_4	uniform	0.0	0.7
		p_5	uniform	2.0	4.0
		p_6	uniform	2.0	20.0
		p_7	uniform	0.0	0.2

Table 3: **Priors used for each example in §4.** The parameters θ_1 and θ_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.

4.4 Embryonic stem cell differentiation

624
625
626
627
628
629
630
631

We now demonstrate how CMC can be applied to real data generated from experiments investigating human embryonic stem cell (hESC) differentiation. Specifically, we use a reaction kinetics-based model presented in [41], which seeks to explain regulation of three transcription factors involved in hESC fate: CDH1, ZEB1 and KLF8. The regulation of these three transcription factors was modelled by the following ODE system involving a number of Michaelis-Menten-type terms,

$$\begin{aligned}\frac{dC}{dt} &= \frac{k_1}{k_2 + Z^2} + \frac{k_3}{k_4 + K^2} - d_1 C \\ \frac{dZ}{dt} &= \frac{ak_5K^2}{k_6 + K^2} - d_2 Z \\ \frac{dK}{dt} &= \frac{rk_7}{k_8 + C^2} - d_3 K,\end{aligned}\tag{28}$$

where $C = [\text{CDH1}]$, $Z = [\text{ZEB1}]$ and $K = [\text{KLF8}]$, subject to initial conditions: $C(0) = C_0$, $Z(0) = Z_0$, $K(0) = K_0$, and $a = 1$ and $r = 1$ are nondimensional parameters. We recast this system, using nondimensional variables

$$y_1 = \frac{k_2 d_1}{k_1} C, \quad y_2 = \frac{k_6 d_1}{k_4 k_5} Z, \quad y_3 = \frac{1}{\sqrt{k_4}} K,\tag{29}$$

and time scale $\frac{1}{d_1}$, so that $\tau = d_1 t$, which results in the following system,

$$\begin{aligned}\frac{dy_1}{d\tau} &= \frac{1}{1 + p_1 y_2^2} + \frac{p_2}{1 + y_3^2} - y_1 \\ \frac{dy_2}{d\tau} &= \frac{y_3^2}{1 + p_3 y_3^2} - p_4 y_2 \\ \frac{dy_3}{d\tau} &= \frac{p_5}{p_6 + y_1^2} - p_7 y_3,\end{aligned}\tag{30}$$

with initial states $y_1(0) = y_{1,0}$, $y_2(0) = y_{2,0}$, $y_3(0) = y_{3,0}$. (See supplementary files for further details.)

In what follows, we perform parameter inference for eq. (30) on single-cell RNA-seq data obtained and processed as described in [41] from NCBI's Gene Expression Omnibus. The dataset has single-cell gene expression data for 758 cells collected at six times during the course of experiment (0 h, 12 h, 24 h, 36 h, 72h, 96 h) for [CDH1], [ZEB1] and [KLF8] with 92, 102, 66, 172, 138 and 188 measurements at each time point respectively.

Here, we consider estimating the posterior parameter distributions using data obtained at $t = 12\text{h}$ for CDH1 and $t = 72\text{h}$ for KLF8. Across the three cases described below, we use CMC with priors for parameters in eq. (30) as given in Table 3. We assume the initial values of each variable are given by: $y_1(0) = 1.5$, $y_2(0) = 0.0$, $y_3(0) = 0.0$.

We first consider the CDH1 data in isolation. To do so, we fit a Gaussian distribution to these data and obtain $q_1 = y_1(12) \sim \mathcal{N}(5.50, 1.05)$. CMC produces samples that closely approximate this distribution (Figure 13A; blue lines). The joint posterior distribution for two model parameters, (p_2, p_5) , is shown in the leftmost panel of Figure 14 and shows a concentrated distribution.

Next we consider the KLF8 data alone. We fit a Gaussian distribution to these data and obtain $q_2 = y_3(72) \sim \mathcal{N}(3.77, 1.37)$. Again, using CMC,

we obtain samples that closely approximates this target (Figure 13A; orange lines). The posterior distribution for (p_2, p_5) is, however, now quite different to previously (Figure 14 middle panel) hinting that it may be quite difficult to determine a posterior distribution where we target both q_1 and q_2 .
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677

Finally, we attempt to target the distribution described by both q_1 and q_2 . Here, we assume that there is no correlation between these targets because we have no cells with observations for both $t = 12$ and $t = 72$ since the measurement process is destructive. In Figure 13B, we plot the joint target distribution and samples from CMC. In this plot, it is clear that there is a disjunction between the target distribution and the samples. In particular, the target distribution for q_2 has a mean that is far below the target value.
670
671
672
673
674
675
676
677

The rightmost panel of Figure 14 shows the posterior parameter distribution for (p_2, p_5) when targeting this bivariate output distribution. In comparing it to the other panels in the same figure, it is clear that the posterior distribution when targeting (q_1, q_2) is somewhere between the distributions obtained when targeting q_1 and q_2 in isolation; unfortunately, this midway house is not suited to either case. Indeed, this failure to target both q_1 and q_2 simultaneously suggests that the model does not actually cohere with the data.
670
671
672
673
674
675
676
677

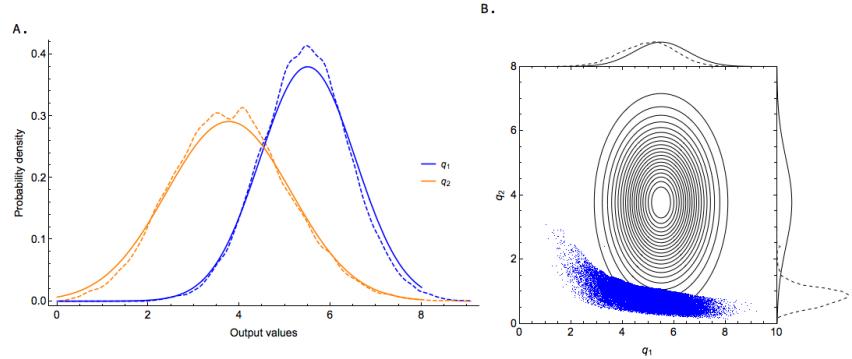


Figure 13: Embryonic stem cell differentiation model: output targets. In A., we show the output target distributions described in §4.4 and kernel density estimates of the distributions reconstructed from the CMC samples. In B., we show the joint target distribution (contour lines) for the case where we target both q_1 and q_2 simultaneously; above and to the right of the plot, we show the target marginals (solid lines) and the marginals reconstructed from the samples (dashed lines). In CMC, 50,000 independent samples were used in the “ContourVolumeEstimator” step, and 50,000 MCMC samples across each of 4 Markov chains were used in the second, with the first half of the chains discarded as “warm-up” [25]. Mathematica’s “SmoothKernelDistribution” function, using a Gaussian kernel [35] and default bandwidths were used to reconstruct marginal densities.

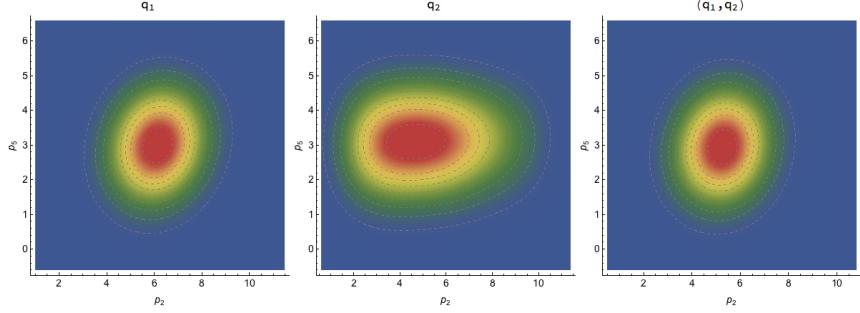


Figure 14: Embryonic stem cell differentiation model: posterior parameter distribution. In the left panel, we show the posterior distribution for (p_2, p_5) when targeting q_1 ; in the middle, we show the same when targeting q_2 ; and in the right panel, we show the same when targeting (q_1, q_2) . In CMC, 50,000 independent samples were used in the “ContourVolumeEstimator” step, and 50,000 MCMC samples across each of 4 Markov chains were used in the second, with the first half of the chains discarded as “warm-up” [25].

To investigate this further, we overlay samples (blue points) from the contour volume estimation step of Algorithm 1 on the joint target distribution of (q_1, q_2) in Fig. 15. In the main panel, the output samples are concentrated in a band that runs from the top left of the plot towards the lower right. Because of this, there is relatively low overlap between the joint target and the contour volumes, indicating that the model does not cohere well with the data and illustrates why inference struggles to find a posterior distribution consistent with the joint target. However, as illustrated in Fig. 13A, it is possible to find posterior distributions consistent with the target distributions for either q_1 or q_2 in isolation. Fig. 15 indicates why this is the case. Above the main panel, we plot these target distributions (black solid lines) and the corresponding contour volume distributions (blue dashed lines). There is considerably more overlap in these marginals than in the joint densities, which shows that finding a posterior consistent with each of these targets in isolation is possible but with them both is tricky. Overall, it appears that the model is not consistent with the data.

678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693

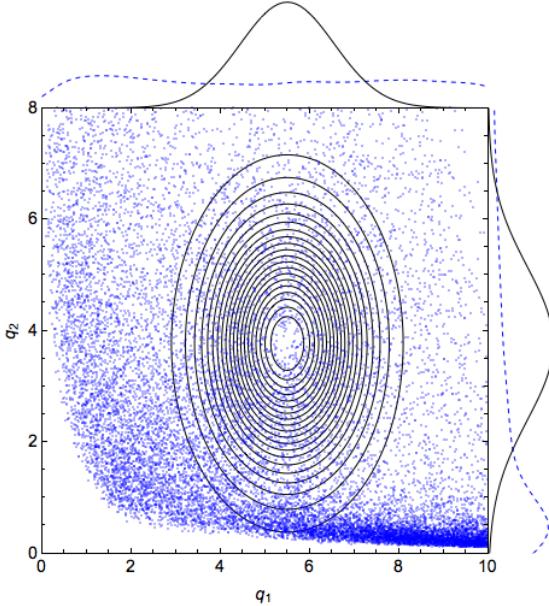


Figure 15: **Embryonic stem cell differentiation model: contour volume distribution.** In the main panel, we plot contours (solid black lines) for the joint target distribution of (q_1, q_2) and 20,000 samples from the contour volume estimation step of Algorithm 1 (blue points). Above and to the right of the main panel, we plot the marginal target distribution in each dimension (solid black lines) and the marginal contour volume distributions (dashed blue lines).

By “model” here, it could either be that the ODE system described in eq. (30) is inappropriate; it is also possible that this could be due to failure to include noise in the measurement process. Given the extent of the discrepancy between the ODE means and the target contours, we suggest that it’s most likely that the ODE model misses or misrepresents key processes. These results illustrate how CMC can be used to determine when a model is inconsistent with data and also suggest that extending CMC to handle noisy measurement is likely worthwhile.

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 during experiments. Our approach assumes what we call a “heterogeneous ordinary differential equation” (HODE) framework, in which biochemical processes in all cells are governed by a common ODE. In HODEs, each cell has different rate parameter values, causing a variety of measurements to be obtained across cells. In this framework, estimating heterogeneity in cellular processes amounts to determining the probability distributions of 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 the number of cell clusters

to be provided before estimation, unlike in other approaches. CMC can
716
be used to process high volumes of individual cellular measurements since
717
the framework involves fitting a kernel density estimator to raw exper-
718
imental data and using these distributions rather than data as the target
719
outcome. CMC can handle arbitrary multivariate structure in measured
720
outputs, meaning it can capture correlations between the same cellular
721
species at different timepoints or, for example, contemporaneous corre-
722
lations between different cellular compartments. Being a Bayesian approach,
723
CMC uses prior distributions over parameter values to ensure uniqueness
724
of the posterior distribution, allowing pre-experimental knowledge to be
725
used to improve estimation robustness. The flexible and robust framework
726
that CMC provides means it can be used to perform automatic inference
727
for wide-ranging systems of practical interest.
728

Our approach also provides a natural way to test that the process is
729
working satisfactorily. Feeding posterior parameter samples obtained by
730
CMC into forward model simulations results in a distribution of output
731
values which can be compared to the target. Indeed, we have found this
732
comparison indispensable in applying CMC in practice and include it as
733
the last step in the CMC algorithm (Algorithm 1). Discrepancies between
734
the target output distribution and its CMC approximation can occur either
735
as a result of poor estimates of the “contour volume distribution” in the
736
first stage of the algorithm or due to insufficient MCMC samples in the
737
second. Either of these issues are often easily addressed by increasing
738
sample sizes or changing hyperparameter settings for the kernel density
739
estimator. Although kernel density estimation in high dimensional spaces
740
remains an open research problem, we have found vine copula kernel density
741
estimation works well for the dimensionality of output measurements we
742
investigate here [27].
743

Failure to reproduce a given output distribution can also indicate that
744
the generating model (the priors and the forward model) are incongruent
745
with experimental results. This may either be due to misspecification of the
746
ODE system or because the assumption of a deterministic forward model
747
is inappropriate. Our approach currently assumes that output variation is
748
dominated by cellular variation in the parameter values of the underlying
749
ODE, with measurement noise making a negligible contribution. Whether
750
this is a reasonable assumption depends on the system under investiga-
751
tion and, more importantly, on experimental details. We recognise that
752
neglecting measurement noise when it is, in fact, important in determining
753
observed data means CMC will overstate cellular variation. It may also
754
mean that some output distributions cannot be obtained by our model
755
system (i.e. HODEs without noise). Future work incorporating a stochastic
756
noise process or, more generally, including stochastic cellular mechanisms
757
is thus likely to be worthwhile.
758

In Figure 4, we present the workflow for our approach, which includes
759
as its last step comparing output samples with the target distribution. As
760
discussed above, if output samples do not correspond with the target, this
761
may indicate that a model isn’t fit for purpose. Conversely, if there is
762
correspondence with the target distribution, it is possible that a simplified
763
model – with (say) one or more fewer parameters – could also recapitulate
764
the same results. Thus, a process of repeated rounds of model simplification
765
then CMC could be pursued to simplify a model until output samples no
766
longer correspond with the target. The most parsimonious model would
767
then be the simplest case where the output samples still match the target.
768

We note however, that such an approach may be dangerous if the most parsimonious model is then used to predict the distributions of other functionals. 769
770
771

Whilst we have illustrated our approach by fitting ODE models to data, 772
we recognise that our approach is applicable to deterministic forward models 773
in general. These include a large swathe of models used in computational 774
biology, such as partial differential equations and difference equations. 775
Similarly, whilst we have illustrated our approach by fitting to models 776
with time-invariant parameters, it could also be used to determine how 777
parameters vary throughout the course of an experiment - provided the 778
dynamic evolution of parameter values is itself parameterised. 779

We have labelled our approach as Bayesian since it involves explicit 780
estimation of probability distributions and requires priors. We recognise, 781
however, that it is not of the form used in traditional Bayesian inference. 782
This is because, rather than aiming to formulate a model that describes 783
output observations, our approach aims to recapitulate output *distributions*. 784
Others [24], (including us [26]), have considered similar problems before; 785
perhaps most notably by Albert Tarantola in his landmark work on inverse 786
problem theory (see, for example, [42]). In Tarantola's framework, a joint 787
input parameter and output space is considered, where prior knowledge and 788
experimental theory combine elegantly to produce a posterior distribution 789
whose marginal output distribution is a weighted "conjunction" of various 790
sources of information. This work has seen considerable interest in areas 791
such as the geosciences [43, 44], and we propose that Tarantola's approach 792
may prove useful for the biosciences. 793

The natural world is rife with variation, and mathematical models 794
represent frameworks for understanding its causes. Typically, the state of 795
biological knowledge is such that one effect – a given pattern of variation – 796
has many possible causes. Observational or experimental data can be used 797
to apportion weight to each cause, in a process that amounts to solving 798
an inverse problem. The approach we describe here follows the Bayesian 799
paradigm of inverse problem solving where uncertainty in potential causes 800
(i.e. parameter values) is described using probability distributions. Here, 801
we illustrate the worth of our method by using it to estimate cellular 802
heterogeneity in biochemical processes. However, it could equally be used to 803
invert other classes of under-determined systems arising elsewhere. Contour 804
Monte Carlo provides an automatic framework for performing inference on 805
such under-determined systems, and the use of priors allows for robust and 806
precise parameter estimation unattainable through the data alone. 807

6 Author contributions

BL, DJG and SJT conceived the study. BL carried out the analysis. All 808
authors helped to write and edit the manuscript. 809
810

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] M Karlsson, DL Janzén, L Durrieu, A Colman-Lerner, MC Kjellsson, and G Cedersund. Nonlinear mixed-effects modelling for single cell estimation: when, why, and how to use it. *BMC Systems Biology*, 9(1):52, 2015.

- [17] C Zechner, M Unger, S Pelet, M Peter, and H Koepll. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11(2):197, 2014.
- [18] L Dharmarajan, HM Kaltenbach, F Rudolf, and J Stelling. A simple and flexible computational framework for inferring sources of heterogeneity from single-cell dynamics. *Cell Systems*, 8(1):15–26, 2019.
- [19] 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.
- [20] 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.
- [21] YH Chan, J Intosalmi, S Rautio, and H Lähdesmäki. A subpopulation model to analyze heterogeneous cell differentiation dynamics. *Bioinformatics*, 32(21):3306–3313, 2016.
- [22] 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.
- [23] 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.
- [24] T Butler, J Jakeman, and T Wildey. Combining push forward measures and Bayes rule to construct consistent solutions to stochastic inverse problems. *SIAM J. Sci. Comput.*, 40(2):A984–A1011, 2018.
- [25] B Lambert. *A Student’s Guide to Bayesian Statistics*. Sage Publications Ltd., 2018.
- [26] B Lambert, D Gavaghan, and SJ Tavener. Inverse sensitivity analysis of mathematical models avoiding the curse of dimensionality. *BioRxiv*, page 432393, 2018.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] A Gelman and DB Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

- [31] J Bezanson, A Edelman, S Karpinski, and VB Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.
- [32] C Rackauckas and Q Nie. Differentialequations. jl—a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1), 2017.
- [33] D Bates, R Lai, Byrne S, and contributors. Rcall. <https://github.com/JuliaInterop/RCall.jl>, 2015.
- [34] T Nagler. *kdevine: Multivariate Kernel Density Estimation with Vine Copulas*, 2018. R package version 0.4.2.
- [35] Inc. Wolfram Research. Mathematica 8.0. <https://www.wolfram.com>.
- [36] 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.
- [37] HT Banks, K Holm, and F Kappel. Comparison of optimal design methods in inverse problems. *Inverse Problems*, 27:1–31, 2011.
- [38] JD Murray. *Mathematical biology: I. An Introduction (interdisciplinary applied mathematics)(Pt. 1)*. New York, Springer, 2007.
- [39] A Jasra, DA Stephens, and CC Holmes. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.
- [40] 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.
- [41] X Tu, Q Zhang, W Zhang, and X Zou. Single-cell data-driven mathematical model reveals possible molecular mechanisms of embryonic stem-cell differentiation. *Mathematical Biosciences and Engineering*, 16(5):5877–5896, 2019.
- [42] A Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. SIAM, 2005.
- [43] K Mosegaard and A Tarantola. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447, 1995.
- [44] 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.